Aplicación de modelos de conteo y espaciales en el estudio de la presencia de pupas del mosquito *Aedes aegypti* en el departamento del Cauca, Colombia.

Juan Sebastián García Rojas

Universidad El Bosque
Facultad de Ciencias
Departamento de Matématicas
Programa de Estadística
Bogotá D.C, Colombia
2023

Aplicación de modelos de conteo y espaciales en el estudio de la presencia de pupas del mosquito *Aedes aegypti* en el departamento del Cauca, Colombia.

Juan Sebastián García Rojas

Trabajo de grado como requisito parcial para optar al título de:
Estadístico

Director:
Emiliano Rodriguez Arango
Jesus David Ramos Montaña

Universidad El Bosque
Facultad de Ciencias
Departamento de Matematicas
Programa de Estadística
Bogotá D.C, Colombia
2023

Agradecimientos

Personas involucradas en el desarrollo de la investigación

Quiero expresar mi sincero agradecimiento a las personas que desempeñaron un papel fundamental en el desarrollo de mi tesis. En primer lugar, al profesor Jesús David Ramos, quien ha sido un apoyo constante, brindándome orientación y tranquilidad a lo largo de todo el proceso, así como su compañía en las exposiciones que he tenido.

De igual manera, quiero reconocer la labor del profesor Emiliano Rodríguez Arango, quien actuó como mi primer director, ofreciendo asesorías esenciales y comprometiéndose desde el inicio en este proyecto.

Mi gratitud también se extiende a otros profesores que contribuyeron significativamente a mi investigación. Al profesor Ricardo Alberto Borda, cuyo material dedicado en gran parte aportó valiosos elementos a mi trabajo. Al profesor Mario José Pacheco López y al profesor Danny Samuel Martínez, quienes brindaron asesorías sin ninguna obligación, y cuyas enseñanzas me ayudaron a identificar aspectos faltantes en mi trabajo. Asimismo, quiero agradecer al profesor Ramón Giraldo Henao por su perspectiva acerca de mi trabajo y al profesor Kenneth Roy Cabrera por las valiosas asesorías proporcionadas.

Amigos y familiares

A nivel personal, deseo expresar mi profundo agradecimiento a las personas que han sido un pilar en mi vida durante este recorrido académico. A mi madre, Constanza García, le agradezco de todo corazón por su apoyo inquebrantable a lo largo de mis años de estudio, así como por sobrellevar conmigo momentos difíciles que me han permitido forjar la fuerza necesaria para alcanzar logros tanto personales como profesionales.

Mi agradecimiento a mi pareja, Margareth Quispe Rodas, por su constante motivación y por inculcarme la importancia del estudio y la adquisición de conocimientos. Su apoyo incondicional ha sido invaluable.

Ademas, quiero dedicar este trabajo a mi abuelita, Mercedes Rojas Gutiérrez, que aunque ya no está presente para ver este proyecto, su deseo de que llegara este momento es una de las razones más poderosas que me impulsaron para llegar hasta acá y a presentar esta investigación.

Mi tía Marcela García que me abrio las puertas de su hogar al inicio de mi carrera, ha sido un motor para mi, siempre confio en mis conocimientos y habilidades y me ha impulsado en mi vida.

A mi padre, Evhard Mateus, le agradezco por su apoyo desde el inicio de mi carrera en estadística, por contribuir financieramente a mis estudios y por brindarme sabios consejos que han enriquecido mi vida durante todos estos años.

Mis tías, Martha García y Ana María García, merecen un agradecimiento especial por su apoyo tanto a nivel personal como profesional. Desde siempre, me enseñaron la importancia de ser un profesional.

A Margarita Rodas, le agradezco por enseñarme la importancia de ser un profesional y por darme la fortaleza para luchar por mi título durante estos años.

En otro ámbito, quiero expresar mi gratitud a María Fernanda Leyva, quien ha sido mi compañera a lo largo de mi carrera, y a Diego Felipe Cortés, por su apoyo constante a lo largo de estos años y por su participación activa en mi vida académica y personal.

A mis amigos de las prácticas profesionales, a Saray López Estrada por llegar a mi vida y darme ánimos para concluir este proyecto de la mejor manera, con su apoyo en los momentos difíciles que estaba pasando, a Maria Alejandra Sánchez León por otorgarme su apoyo en la etapa final de mis prácticas, a Juan Camilo Gómez Cano por su apoyo profesional y por acompañarme en momentos de dedicación a este trabajo.

Por otro lado, a mi círculo de compañeros y amigos en mi hobby principal que ha sido partícipe de mi vida en este proceso, de igual forma, a Sebastián Ortiz, Sebastián Munza, Juan Ricardo Ospina y Nicolas David Rojas, gracias por la compañía en el proceso, los consejos y el acompañamiento.

A todas estas personas, mi más sincero agradecimiento por formar parte de mi vida.

Colegas y colaboradores académicos

Mi agradecimiento se extiende a MSD por permitirme realizar mis prácticas profesionales en su empresa, y al Semillero EACD de la Universidad El Bosque por brindarme la oportunidad de adquirir valiosos conocimientos.

De nuevo, al profesor Jesús David Ramos, quien me tuvo en cuenta para participar en diversas investigaciones y artículos.

No sin antes olvidar el curso que ofrece el doctor Noam Ross, donde utilizo GAM, R y otras herramientas para comprender mejor las relaciones complejas entre los animales, sus enfermedades y cómo las infecciones pueden transmitirse de los animales a los humanos.

Instituciones académicas

Agradezco a la ingeniera Laura Cabezas Pinzón, perteneciente al Grupo de Investigación Ambiental de la Universidad El Bosque, por permitirme utilizar los datos

y trabajar con ellos para hacer realidad este proyecto. De igual manera, a Jesús David Ramos por considerarme como su primera opción para desarrollar este proyecto y por compartir su experiencia y artículos como base.

Mentores y profesores

A todas las personas que han sido mis mentores y profesores, agradezco de manera especial por elevar mis conocimientos tanto a nivel profesional como personal. A Carolina Rojas Celis, por inculcarme el amor por la investigación. Al profesor Ricardo Alberto Borda, por recibirme en la carrera de la mejor manera y por impartir los primeros conocimientos desde el primer semestre. Al profesor Jesús David Ramos, uno de mis más grandes mentores, por desafiarme, brindarme oportunidades y compartir su experiencia. Al profesor Danny Samuel Martínez, por exigirme y desafiarme en los momentos que más lo necesitaba, ayudándome a enfocar mi carrera. Al profesor Andrés Cardona, por la generosidad de brindarme la oportunidad de realizar mis prácticas profesionales en una gran empresa. Al profesor Mario José Pacheco, por impartir conocimientos de forma clara, sencilla y profesional. Por último, darle un gran agradecimiento a María Fernanda Rodríguez Beltrán por ser mi mentora en las prácticas profesionales, poder recibir de ella tanto conocimiento en mi primera experiencia laboral ha sido de gran ayuda tanto a nivel profesional como personal.

Declaración personal del autor

Juan Sebastián García Rojas con C.C. 1193109306, estudiante de Estadística en la Universidad El Bosque, como autor de este documento académico titulado "Aplicación de modelos estadísticos para el estudio de pupas del mosquito *Aedes aegypti* en el departamento del Cauca, Colombia.z presentado como Trabajo Final de Grado.

Declaro que:

Es un trabajo original, que no copio ni utilizo parte de obra alguna sin mencionar de forma clara y precisa su origen tanto en el cuerpo del texto como en su bibliografía y que no empleo datos de terceros sin la debida autorización, de acuerdo con la legislación vigente. Asimismo, declaro que soy plenamente consciente de que no respetar esta obligación podrá implicar la aplicación de sanciones académicas, sin perjuicio de otras actuaciones que pudieran iniciarse.

En Bogota D.C., a 30 de Octubre de 2023 .

Fdo:

Juan Sebastián García Rojas

Resumen

Los cambios en las condiciones ambientales tienen un impacto directo en el aumento de la cantidad y dispersión de vectores, así como en la incidencia de enfermedades transmitidas por ellos. Estas variaciones están estrechamente relacionadas con el aumento promedio de la temperatura superficial de la Tierra debido al calentamiento global. En el caso específico del mosquito *Aedes aegypti*, que es el transmisor de enfermedades virales como el dengue, su expansión geográfica aumentaría de forma paralela con el aumento de la temperatura global. Estas enfermedades son un problema de salud pública en Colombia debido al alto número de casos nuevos que se presentan.

La prevención de estas enfermedades se centra en el control del mosquito *Aedes aegypti*, y tradicionalmente se ha utilizado la vigilancia entomológica para este propósito. Sin embargo, se ha observado que esta estrategia puede resultar costosa en términos tanto humanos como económicos. Por lo tanto, hemos propuesto alternativas como los modelos predictivos de estadística clásica y espacial, los cuales permiten explicar el número de pupas del vector.

El insumo para la realización de los modelos propuestos correspondió a información recolectada en el año 2017: datos entomológicos, geográficos, climáticos y demográficos de 393 localidades ubicadas en 33 municipios del Cauca, Colombia.

Se ajustaron modelos de regresión para datos de conteo en presencia de sobredispersión y cero inflación, y se compararon por medio de medidas de bondad de ajuste, test de hipótesis para prueba de supuestos y parámetros de regresión estimados: regresión lineal generalizado Poisson (GLMP), regresión lineal generalizado binomial negativa (GLMNB), regresión binomial negativo cero inflado (ZINB), modelo de regresión de Hurdle y modelo de regresión de Tweedie. Se encontró que el ZINB es el que mejor modela el número de pupas.

Tomando en cuenta que también existe la componente espacial dentro del conjunto de datos, se realizó un análisis de tipo espacial para datos de area georeferenciados con pruebas de hipótesis y se ajustaron modelos espaciales: modelo de error espacial, modelo de retardo espacial, modelo de Durbin y modelo aditivo generalizado espacial. Se concluye que el modelo aditivo generalizado espacial es el que mejor modela el numero de pupas teniendo en cuenta la componente espacial.

Palabras clave: Vectores, *Aedes aegypti*, Vigilancia entomológica, Modelos predictivos, Modelos lineales generalizados, Poisson, Sobredispersión, Cero-inflación, Binomial negativo, Hurdle, Tweedie, Estadística espacial, Indice de Moran, Semivariograma, Geoestadística, Aditivo generalizado.

Abstract

Changes in environmental conditions have a direct impact on the increase in the number and dispersion of vectors, as well as on the incidence of vector-borne diseases. These variations are closely related to the average increase in the Earth's surface temperature due to global warming. In the specific case of the *Aedes aegypti* mosquito, which is the transmitter of viral diseases such as dengue fever, its geographic expansion would increase in parallel with the increase in global temperature. These diseases are a public health problem in Colombia due to the high number of new cases.

Prevention of these diseases focuses on the control of the *Aedes aegypti* mosquito, and entomological surveillance has traditionally been used for this purpose. However, it has been observed that this strategy can be costly in both human and economic terms. Therefore, we have proposed alternatives such as classical and spatial statistical predictive models, which can explain the number of vector pupae.

The input for the realization of the proposed models corresponded to information collected in 2017: entomological, geographical, climatic and demographic data from 393 localities located in 33 municipalities of Cauca, Colombia.

Four regression models were fitted for count data in the presence of overdispersion and zero inflation, and compared by goodness-of-fit measures, hypothesis test for assumption testing and estimated regression parameters: generalized linear Poisson regression (GLMP), generalized linear negative binomial regression (GLMNB), zero-inflated negative binomial regression (ZINB) and Hurdle regression model. GLMNB was found to best model the number of pupae.

Taking into account that there is also a spatial component within the data set, a spatial analysis was performed for georeferenced area data with hypothesis tests and spatial models were fitted: spatial error model, spatial lag model, Durbin model, semivariogram and spatial generalized additive model. It is concluded that the spatial generalized additive model is the one that best models the number of pupae taking into account the spatial component.

Keywords: Vectors, *Aedes aegypti*, Entomological surveillance, Predictive models, Generalized linear models, Poisson, Overdispersion, Zero-inflation, Negative binomial, Hurdle, Tweedie, Spatial statistics, Moran's index, Semivariogram, Geostatistics, Generalized additive.

Índice general

Índice de figuras

Índice de tablas

1. Introducción

Durante los últimos 300 años, ha habido aumento en la temperatura atmosférica, lo que ha incrementado en 0,7°C en los últimos cien años Yi et al. (2014). Colombia podría experimentar un aumento gradual alrededor de 2,14°C en su temperatura media anual, y la región del Pacífico podría experimentar un aumento de 2,15°C a finales del siglo, si las emisiones globales de gases de efecto invernadero continúan aumentando. El incremento de la temperatura también ha llevado a un aumento en la transmisión de enfermedades transmitidas por vectores, como las arbovirosis, que son virus generados por artrópodos. Este factor principal, entre otros, afectan tanto la abundancia como la distribución de los vectores y la dinámica de la transmisión de los virus Reiter (2001), Heinisch et al. (2019).

El aumento de la temperatura global puede provocar cambios en los factores climáticos que contribuyen a la distribución geográfica de los mosquitos vectores de arbovirus en áreas que anteriormente no presentaban riesgo de transmisión, lo que puede provocar un aumento en la incidencia de enfermedades como el dengue en regiones donde antes no se presentaban Padilla et al. (2012), Ramírez et al. (2013). Estas enfermedades son un problema de salud pública en Colombia debido al alto número de casos nuevos que se presentan Adin Urtasun et al. (2018).

El principal vector de arbovirosis es el mosquito *Aedes aegypti* (*L.*) Carvajal et al. (2016), Olano (2016). Según Singh (2013), el dengue es una enfermedad transmitida por arbovirus que, por medio de modelos predictivos, muestra expansión geográfica bajo escenarios de cambio climático. Durante los últimos 30 años, ha habido aumento en la incidencia del dengue en América, y aproximadamente 500 millones de personas en la región están en riesgo de contraer la enfermedad. Se estima que se producen alrededor de 390 millones de casos de dengue cada año, según modelos estadísticos Brady et al. (2012). La Organización Panamericana de la Salud ha enfatizado la prioridad de los programas de control, prevención y tratamiento de los arbovirosis OPS (2016). La prevención del dengue se enfoca principalmente en el control del mosquito *Aedes aegypti* a través de los tres índices de la vigilancia entomológica, sin embargo, estos no han traído buenas asociaciones en los resultados, es por eso que se intenta con otros parámetros como el número de pupas llegar a mejores estimaciones y correlaciones con respecto a la población adulta, lo que posibilita el modelamiento por medio de distribuciones matemáticas y estadísticas. Por lo anterior, es posible la estimación y comparación de modelos probabilísticos para el conteo de pupas de *Aedes aegypti* como posible herramienta de vigilancia entomológica para predecir la variación media poblacional bajo distintos escenarios de cambio climático tomado como referencia la componente espacial y demás variables explicativas. Estos modelos podrían ayudar a predecir el riesgo de transmisión del dengue y a desarrollar estrategias de prevención más efectivas. De esta forma... ¿Cuál modelo puede explicar bien el conteo de pupas del mosquito *Aedes aegypti*?

2. Justificación

En los últimos años, el aumento de la temperatura atmosférica ha generado un impacto significativo en la transmisión de enfermedades transmitidas por vectores, como el dengue. Yi et al. (2014). Colombia, al igual que muchas otras regiones del mundo, ha experimentado un incremento gradual en la temperatura media anual, lo que ha llevado a un aumento en la abundancia y distribución de mosquitos vectores, en particular el mosquito Aedes aegypti. Este vector es conocido por ser el principal transmisor del dengue, una enfermedad que representa un importante problema de salud pública en el país Ramírez et al. (2013).

La vigilancia entomológica ha sido una estrategia tradicional para controlar la población de mosquitos y prevenir la propagación del dengue. Sin embargo, esta metodología presenta diversas limitaciones que dificultan su eficacia y eficiencia. La recolección manual de datos entomológicos requiere una gran cantidad de recursos humanos, tiempo y recursos financieros García Pérez y Alfonso Aguilar (2013). Además, la obtención de resultados precisos y oportunos a partir de esta vigilancia tradicional puede ser un desafío, especialmente en áreas extensas y con recursos limitados.

Ante este panorama, se hace evidente la necesidad de buscar alternativas más eficientes y menos costosas para llevar a cabo la vigilancia entomológica. En este sentido, los modelos predictivos estadísticos emergen como una herramienta prometedora para predecir la abundancia y distribución de mosquitos vectores, así como el riesgo de transmisión de enfermedades como el dengue Rotela (2012). Estos modelos permiten analizar múltiples variables, como datos climáticos, geográficos y demográficos, y obtener estimaciones precisas sobre la población de mosquitos en diferentes regiones y bajo distintos escenarios de cambio climático.

La presente investigación tiene como objetivo principal comparar y evaluar diferentes modelos probabilísticos para el conteo de pupas de Aedes aegypti como una herramienta de vigilancia entomológica más efectiva y rentable. Se espera que esta investigación contribuya a mejorar la comprensión de la dinámica de la población de mosquitos y permita desarrollar estrategias de prevención y control más precisas y adaptadas a las condiciones específicas de cada región. Además, estos modelos podrían facilitar la predicción del riesgo de transmisión del dengue, lo que a su vez ayudaría a implementar medidas preventivas más eficaces y mitigar los impactos negativos en la salud pública.

3. Objetivos

3.1. Objetivo general

Aplicar una variedad de modelos estadísticos, tanto con componente espacial como sin esta, con el fin de analizar y comparar su capacidad predictiva e inferencial en situaciones de variabilidad climática con respecto al número de pupas del mosquito *Aedes aegypti*.

3.2. Objetivos específicos

1. Realizar un análisis exploratorio y descriptivo de los datos suministrados por medio de métricas y graficos adecuados.
2. Aplicar modelos de estadística para datos de conteo para el estudio del número de pupas.
3. Realizar análisis exploratorio y descriptivo espacial de los datos.
4. Aplicar modelos de estadística espacial para el estudio del número de pupas.
5. Comparar los modelos planteados mediante métricas y criterios adecuados, y definir cual es el que mejor se ajusta a los datos, el más sencillo y el que mejor capacidad predictiva tenga.

4. Marco Teórico

4.1. Antecedentes

En la literatura consultada, se encontraron varios estudios relacionados con el uso de modelos estadísticos para el análisis de datos de conteo, específicamente en el contexto de la presencia de pupas del mosquito Aedes aegypti en el departamento del Cauca, Colombia. Tomando como artículo base y primero en el que destacaremos, se trata del artículo de Laura Cabezas (2023), se exploraron diferentes métodos estadísticos para identificar las regiones geográficas con mayor riesgo de presencia del mosquito *Aedes aegypti* departamento del Cauca con el mismo insumo de datos con los que se realizara la presente investigación. En este artículo se muestra que el mejor modelo para explicar el conteo de pupas por unidad muestral es el modelo cero-inflado (ZINB). Sin embargo, se destacó que el número de pupas se deriva de un proceso aleatorio y no espacial.

Por otro lado, Atoche Calzada (2017) realizó un proyecto de grado que abordó el uso de modelos de regresión para datos de conteo, específicamente en el contexto de competiciones deportivas. Este estudio proporcionó una revisión estructurada de los posibles modelos y la teoría necesaria para modelar frecuencias, como el número de pupas; y en la misma línea está Calcaterra (2017) que llevo a cabo una revisión de los modelos de conteo con exceso de ceros en un informe de pasantías. Esta revisión proporcionará un punto de partida para abordar la asimetría en los datos con los que se trabajará en el estudio.

Entrando en la etapa del componente espacial esta Alcalá et al. (2020), ellos realizaron un análisis espacial de un índice pupal de *Aedes aegypti* en el municipio de Tena, Cundinamarca. Este estudio demostró la aplicabilidad de la información geográfica en la evaluación del riesgo de transmisión de arbovirosis y proporcionó información relevante para la gestión de recursos por parte de las entidades territoriales. De igual forma, en todos los trabajos de Pina et al. (2010) se realiza una revision metodologica de como realizar modelos espaciales para datos de área, incluso en el software en el que se trabaja.

Para terminar esta sección, esta la investigación realizada por Sothe et al. (2017) en el que usa un modelo aditivo generalizado espacial para analizar los deslizamiento y susceptibilidad en un territorio de Santa Catarina en Brasil, por otro lado esta el artículo de Pedersen et al. (2019) en el que habla de los contextos en los que se pueden utilizar los modelos GAMs en R. De la misma forma el curso creado por Ross et al. (2018).

4.2. Modelos y su relación con la realidad

Varios autores, entre ellos Fisher (1955), Neyman (1967) y Cox (1972) han abordado el problema de la modelización mediante la literatura estadística. Este problema se refiere a la representación matemática de la realidad, su variabilidad e incertidumbre, con el propósito de estudiarla, analizarla y comprenderla. El objetivo principal de la modelización es transformar la realidad, predecir su comportamiento futuro o simplemente adquirir conocimiento sobre ella.

De este objetivo se desprenden los modelos deterministas y los modelos probabilísticos, en el primero se busca que aquellos que tienen la capacidad de predecir o explicar un fenómeno específico mediante información suficiente para representar con exactitud la realidad que están modelando, sin presentar errores significativos en la representación. Por otro lado, los modelos probabilísticos trabajan con la variedad o incertidumbre causada porque el estudio presenta distintas fuentes de aleatoriedad, es por ello que los modelos probabilísticos tienen al final un error resultante de la desviación entre el fenómeno observado y su representación bajo un supuesto modelo matemático. A estos últimos también se les denomina *modelos estadísticos*.

El objetivo de esto es buscar explicar el comportamiento de una o varias características de los individuos o elementos de una población, utilizando las diferencias entre las características asociadas a los individuos.

La variable que se desea explicar se conoce como variable respuesta, mientras que las variables que se utilizan para explicar se denominan variables explicativas o covariables. Con lo anteriormente dicho, es necesario elegir un modelo que describa la forma de la relación entre las variables. La elección del modelo depende del tipo de variables que sean de conocimiento (continuas, de conteo, cualitativas, etc.) y de la clase de relaciones entre la variable objetivo y las variables explicativas.

Dependiendo de estas dos cosas nombradas en el anterior párrafo se pueden seleccionar diferentes modelos para explicar la realidad.

El modelo más comúnmente estudiado y utilizado en la modelización estadística es el modelo lineal. Este tipo de modelo busca expresar la relación entre las variables explicativas, a través de una combinación lineal de dichas variables.

4.3. Modelos Lineales: ajuste e inferencia

Para poder entender un poco más acerca del modelo que se trabajara en el proyecto se debe empezar explicando algunos conceptos básicos acerca de los modelos lineales.

Uno de los puntos a destacar en la estadística es el análisis de la relación que existe entre algunas variables debido a que en algunas situaciones es de interés conocer el efecto que puede causar una variable sobre otra o incluso predecir el valor de una variable a partir de otra.

Cuando esta relación de la que se habla se asume de forma lineal, es decir que la forma de cambio de una variable es constante con respecto a otra u otras, se utilizan los modelos lineales.

La regresión lineal es un método paramétrico que permite modelar la relación entre una variable numérica continua Y y un conjunto de variables X siempre y cuando se asuma que existe una relación lineal entre ellas.

El modelo de regresión lineal tiene la siguiente estructura:

$$f(X) = Y \approx \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \epsilon$$

donde los β_i $(i = 0, 1, 2, ..., p)$ son valores que a priori se desconocen y que son llamados coeficientes, los cuales hay que estimar posteriormente y ϵ es el término del error. En la anterior expresión lo que se logra evidenciar es que toda la influencia en la variable Y (variable respuesta) procede de dos grupos, uno que contiene al conjunto de variables explicativas X y otro que son todos factores que no se pueden controlar y que se atribuyen al error ϵ, el cual es aleatorio y que obliga a que no haya relación perfecta, sino con un cierto grado de incertidumbre Carmona (2005).

4.3.1. Hipótesis

- Se desea que el término del error sea cercano a 0, es decir, $E[\varepsilon/X = x] = E[\varepsilon] = 0$
- Lo que en realidad se desea modelar es el valor esperado de una variable dado un conjunto de variables, es decir $E[Y/X = x] = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \epsilon$
- La relación entre la variable de respuesta y las variables explicativas debería ser lineal.
- El parámetro β_0 es el punto de corte con el eje Y y β_1 la pendiente, es decir el incremento en la variable de respuesta por cada unidad de la variable explicativa. Estos dos parámetros se deben estimar debido a que son desconocidos. Pontaque (2005)

4.3.2. Supuestos

- $Y \sim N(X\beta, \sigma^2)$, lo que quiere decir que $E[Y] = X\beta$ y $Var[Y] = \sigma^2$, siendo $X\beta$ una combinación lineal.
- $\epsilon \sim N(0, \sigma^2)$, lo que quiere decir que $E[\epsilon] = 0$ y $Var[\epsilon] = \sigma^2$ Cayuela (2010)

4.3.3. Estimación de los parámetros

Existen muchos métodos para realizar la estimación de los parámetros en un modelo de regresión lineal, sin embargo el método más común es el método de **mínimos cuadrados ordinarios**, este busca encontrar los valores para β_0 y β_1 que minimicen la suma de los errores al cuadrado que es de la siguiente forma:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} y_i - (\beta_0 + \beta_1 x_i)^2$$

Este método ofrece la solución realizando la derivada de SSE con respecto a cada parámetro que se quiere estimar para luego igualar a 0 esa derivada y obtener el valor que esté minimizando dicha SSE. Después de realizar este proceso se obtiene la estimación de los parámetros de regresión donde:

$$\begin{cases} \widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} \end{cases}$$

4.3.4. Inferencia

Asi como lo nombra Gomez Villegas (2005), debido a que los estimadores dependen de una muestra, entonces son variables aleatorias con una distribución de probabilidad, las cuales se pueden utilizar para construir intervalos de confianza. En el comienzo se hablaba de estudiar el efecto de las variables explicativas X sobre la variable respuesta Y lo que implicaría contrastar si el coeficiente estimado β_i es o no significativamente distinto de cero, pues un $\beta_i = 0$ implica la no existencia de relación lineal entre las variables. Para ello se realiza una prueba de hipótesis utilizando el estadístico de prueba t y su valor p en la prueba donde se desea contrastar:

$$H_o : \beta_i = 0$$

$$H_a : \beta_i \neq 0$$

Toda esta modelización de un modelo lineal no es nada más que un caso que se puede extender a una familia más general propuesta y ampliada por Nelder y Wedderburn (1972).

4.4. Modelos Lineales Generalizados

Los modelos lineales generalizados son un conjunto de modelos introducidos por Nelder y Wedderburn (1972) en donde se asume que el conjunto de variables aleatorias independientes o de respuesta $Y_1, Y_2, \cdots, Y_n$ provienen de una distribución que pertenece a la familia exponencial de distribuciones (EDM), esta familia nombrada permite que cada modelo pueda ajustarse a un tipo de datos en específico, por ejemplo, datos binarios, proporciones, datos de conteo, datos continuos y datos continuos cero inflados.

4.4.1. Propiedades

- La distribución de Y_i tiene la forma canónica y depende de un solo parámetro θ_i
- Las distribuciones de todos los Y_i son de la misma forma.

4.4.2. Componentes

Un modelo lineal generalizado segun Nelder y Wedderburn (1972) está caracterizado por tener tres componentes:

1. Componente aleatoria: Siendo la variable dependiente $Y = (Y_1, Y_2, \cdots, Y_n)^T$ de una distribución que pertenece a la familia exponencial, cada variable del vector de respuesta Y_i tiene función de densidad:

$$f(y_i, \theta_i, \phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right] = G(y_i, \phi) \cdot \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a(\phi)}\right]$$

donde:

- a, b y c son funciones específicas.

- θ_i es el parámetro canónico de la distribución.

- ϕ es un parámetro de dispersión.

- G es una función normalizadora que permite que esta sea una función de densidad de probabilidad, es decir que en un caso continuo, la integral de la función sea igual a uno, o en caso discreto la sumatoria de la función sea igual a uno.

Además, se puede verificar que:

$$E(Y) = \mu = b'(\theta) = \frac{\partial b(\theta)}{\partial \theta};\ Var(Y) = \sigma^2 = a(\phi)\frac{\partial^2 b(\theta)}{\partial \theta^2} = a(\phi)V(\mu)$$

Componente sistemática: Muestra una función lineal υ de valores fijos $x_{1i}, x_{2i}, \cdots, x_{pi}$ de las variables explicativas $X_1, X_2, \cdots, X_p$ de la siguiente forma:

$$\upsilon_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \quad i = 1, \cdots, n$$

donde:

- β_p son los parámetros del modelo lineal generalizado.

Reuniendo todos los valores observados de las variables explicativas en una matriz de diseño de tamaño $n \times (1 + p)$ dando como resultado:

$$C = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Otorgando un conjunto de parámetros β como $\beta = (\beta_0, \beta_1, \cdots, \beta_p)^T$ y los υ_i en el vector $\upsilon_i = (\upsilon_1, \upsilon_2, \cdots, \upsilon_n)^T$ para finalmente poderse escribir de forma vectorial como $\upsilon = C \cdot \beta$

Componente de enlace: Si se asume que $E(Y_i) = \mu_i$, siendo μ_i la misma función especifica θ_i, entonces el enlace esta dado por g que es una función monótona diferenciable denominada como *función de enlace*.

$$g(\mu_i) = \upsilon_i$$

de forma específica, su enlace canónico sería igual a tener $g(\mu_i) = \theta_i$, implica que si $\theta_i = \upsilon_i$, entonces, $\theta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$.

Para algunas de las distribuciones que utilizaremos después en algunos modelos, se tiene la función de enlace canónica correspondiente:

- **Binomial Negativa**: $\theta(\mu_i) = ln(\frac{\alpha\mu_i}{1+\alpha\mu_i});\ \theta^{-1}(\eta_i) = \frac{1}{\alpha(exp(-\eta_i)-1)}$
- **Poisson**: $\theta(\mu_i) = log(\mu_i);\ \theta^{-1}(\eta_i) = exp(\eta_i)$

En este trabajo nos centraremos en los modelos para datos de conteo. Aunque la distribución de Poisson es ampliamente utilizada en estos modelos, también abordaremos el modelo de Regresión Binomial Negativa como objeto de estudio para algunos casos.

- **Modelo de Poisson Log-linear**: Función de enlace canónica

$$log(\mu_i) = \eta_i = X_i\beta \ o\ \mu_i = exp(X_i\beta)$$

- **Modelo Binomial Negativa**: Función de enlace logaritmo

$$g(\mu_i) = log(\mu_i)$$

$$log(\mu_i) = \eta_i = X_i\beta \ o\ \mu_i = exp(X_i\beta)$$

Para cualquiera de los casos: $E(Y_i/X_i = x_i) = exp(x_i\beta)$

4.4.3. Estimación de los parámetros

Existen dos formas de realizar la estimación de parámetros de un modelo lineal generalizado, uno son los mínimos cuadrados ponderados que resulta ser una modificación que se le hace a los mínimos cuadrados ordinarios (método utilizado anteriormente para estimar los parámetros de un modelo lineal), el otro es el método de máxima verosimilitud, el cual es el más utilizado para realizar este procedimiento.

Se requiere encontrar el vector de parámetros β, para ello asumiremos una muestra $y_1, y_2, \cdots, y_p$ y un conjunto de variables explicativas $x_1, x_2, \cdots, x_n$ para maximizar su verosimilitud en el modelo $E[Y_i|X_i = x_i] = \mu_i = h(x_i\beta)$.

Asumiendo que ϕ es conocido y que puede ser igual a 1 sin perder generalidad, se reduce en la función de verosimilitud:

$$L(\theta; y) = f(y; \theta) = \prod_{i=1}^{n} f_i(y_i; \theta)$$

con observaciones independientes,

$$l(\theta, \phi, y) = \sum_{i=1}^{n} l_i(\theta_i, \phi, y_i) = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}$$

debido a que la función $c(y_i, \phi)$ no depende de θ_i se omite, haciendo que $\theta_i = \theta(\mu_i)$,

$$l(\mu_i, \phi, y) = \sum_{i=1}^{n} l_i(\mu_i, \phi, y_i) = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i(\mu_i) - b(\theta_i(\mu_i))}{a(\phi)} \right\}$$

como resultado de la relación existente entre la esperanza y el vector de parámetros $\mu_i = h(x_i^t\beta)$,

$$l(\beta,\phi,y)=\sum_{i=1}^{n} l_i(\beta,\phi,y_i)=\sum_{i=1}^{n}\left\{\frac{y_i\theta_i(h(x_i^t\beta))-b(\theta_i(h(x_i^t\beta))}{a(\phi)}\right\}$$

nace la primera derivada conocida como *función score*,

$$s(\beta)=\frac{\partial l}{\partial\beta}=\sum_i s_i\beta$$

$$s_i(\beta)=x_iD_i(\beta)\sigma_i^{-2}(\beta)[y_i-\mu_i(\beta)]$$

donde,

- $\mu_i(\beta)=h(x_i^t\beta)$
- $\sigma_i^2(\beta)=a(\phi)\varsigma(h(x_i^t\beta))$
- $V(\mu)=\frac{\partial^2 b(\theta)}{\partial\theta^2}$
- $D_i(\beta)=\frac{\partial h(x_i^t\beta)}{\partial v}$

aplicando la regla de la cadena para derivar,

$$\frac{\partial}{\partial\beta}\theta(h(x_i^t\beta))=\theta'(h(x_i^t\beta))h'(x_i^t\beta)x_i=x_iD_i(\beta)\theta'(h(x_i^t\beta))$$

$$\frac{\partial}{\partial\beta}b(\theta(h(x_i^t\beta)))=\frac{\partial}{\partial\theta}b(\theta(h(x_i^t\beta)))\frac{\partial}{\partial h}\theta(h(x_i^t\beta))\frac{\partial}{\partial v_i}h(x_i^t\beta)x_i=$$

$$\mu_i(\beta)\frac{\partial}{\partial h}\theta(h(x_i^t\beta))D_i(\beta)x_i=\mu_i(\beta)\left[\frac{\partial}{\partial\mu_i}\theta(\mu_i)\right]D_i(\beta)x_i$$

$$\mu(\theta)=b'(\theta)=\frac{\partial b(\theta)}{\partial\theta}$$

$$\frac{\partial\mu(\theta)}{\partial\theta}=b''(\theta)$$

por la función inversa,

$$\frac{\partial}{\partial\mu_i}\theta(\mu_i)=\frac{1}{b''(\theta(\mu_i))}=\frac{1}{V(\mu_i)}=a(\phi)\sigma_i^{-2}(\beta)$$

aplicando sustitución,

$$\frac{\partial}{\partial\beta}\theta(h(x_i^t\beta))=a(\phi)x_iD_i(\beta)\sigma_i^{-2}(\beta)$$

$$\frac{\partial}{\partial\beta}b(\theta(h(x_i^t\beta)))=a(\phi)\mu_i(\beta)\sigma_i^{-2}(\beta)D_i(\beta)x_i$$

para así obtener,

$$s_i(\beta)=\frac{\partial}{\partial\beta}l_i(\beta,\phi,y_i)=y_ix_iD_i(\beta)\sigma_i^{-2}(\beta)-\mu_i(\beta)x_iD_i(\beta)\sigma_i^{-2}(\beta)=$$
$$x_iD_i(\beta)\sigma_i^{-2}(\beta)[y_i-\mu_i(\beta)]$$

- **Matriz de información de Fisher esperada**:

  $F(\beta)=Cov\ s(\beta)=\sum_i F_i(\beta)$

  $F_i(\beta)=x_ix_i^tw_i(\beta)$, siendo, $w_i(\beta)=D_i^2(\beta)\sigma_i^{-2}(\beta)$

- **Matriz de información de Fisher observada**:

  $F_{obs}(\beta)=-\frac{\partial^2 l(\beta)}{\partial\beta\ \partial\beta^t}$, con, $F(\beta)=E(F_{obs}(\beta))$

Las soluciones de las ecuaciones de máxima verosimilitud $s(\hat{\beta})=0$ se debe dar bajo el manejo de un método iterativo como *Fisher Scoring* o *Newton Raphson*.

A través de estos métodos se obtienen las estimaciones de los parámetros del modelo $\hat{\beta}$ que presentan consistencia, eficiencia asintótica y normalidad asintótica.

Si el parámetro de dispersión es desconocido, se puede optar por utilizar el siguiente estimador, como lo nombra Atoche Calzada (2017):

$$\hat{\phi} = \frac{1}{n-p} \sum_i \frac{\left[y_i - \mu_i(\hat{\beta})\right]^2}{v(\mu_i(\hat{\beta}))}$$

4.4.4. Adecuación del Modelo

Para escoger el mejor modelo todo se basa en tener en cuenta el objetivo del modelo, hay modelos que se quieren solamente y tienen como objetivo principal hacer predicción, mientras que hay otros en donde se interesa hacer inferencia y para ello, es necesario un modelo muy interpretable.

Si utilizamos muchas variables para explicar un modelo, podríamos obtener un buen ajuste de los datos, pero sería difícil interpretar el modelo resultante. Por el contrario, si usamos pocas variables para explicar el modelo, el ajuste podría ser malo. Por lo tanto, lo ideal es buscar un modelo que tenga un número intermedio de variables que expliquen bien los datos y que sea fácil de interpretar. Durante el proceso de ajuste del modelo, se evalúan varios modelos diferentes para encontrar uno que esté entre el modelo saturado y el modelo nulo.

- **Modelo Saturado**: El número de parámetros estimados es igual al número de observaciones.
- **Modelo Nulo**: Contiene como único parámetro al valor esperado μ para todas las observaciones.

Después de encontrar ese modelo intermedio se procede con seguir algunos pasos:

1. Evaluar los parámetros a partir de la significancia estadística.
2. Analizar si la supresión de variables afectaría la precisión de las estimaciones o sesgara la estimación de los parámetros.
3. Valorar si las variables explicativas de interés deben permanecer en el modelo.

Buscando todo lo anteriormente dicho, es necesario utilizar los criterios de bondad de ajuste

4.4.5. Bondad de ajuste

- **El estadistico Deviance**: Es la distancia entre el logaritmo de la función de verosimilitud del modelo ajustado y el modelo saturado. Si la deviance es pequeña significa que el modelo en investigación proporciona un ajuste tan bueno como el modelo saturado, pero con menos parámetros.

$$D(y; \hat{\mu}) : 2\left[l(y; y) - l(\hat{\mu; y})\right]$$

Según Nelder y Wedderburn (1972), si el modelo es correcto, la deviance seguirá una distribución asintótica χ^2_{n-p}. Es decir, $D(y;\hat{\mu}) \sim \chi^2_{n-p}$

- **Pseudo R^2**:

$$R^2 = 1 - \frac{D(y;\hat{\mu})}{D(y;\hat{\mu_o})}$$

Es decir, el cociente entre la deviance entre el modelo ajustado y el modelo nulo. Además, se comprueba que $0 \leq R^2 \leq 1$

- **Estadístico χ^2 de Pearson**:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu_i})^2}{V(\hat{\mu_i})}$$

Con $V(\hat{\mu})$ como la función de varianza estimada para la distribución de la variable respuesta, como lo indica Atoche Calzada (2017).

4.4.6. Selección de variables

Al momento de hacer una comparación entre modelos, existen métricas que comparan las log-verosimilitudes de los modelos penalizando a aquel que tenga mayor cantidad de covariables, tal como se nombra en Kuha (2004).

- **Criterio de Información Akaike (AIC)**:

$$AIC = k - 2ln(\hat{L})$$

siendo k el número de parámetros del modelo y $\hat{L}$ es el valor que maximiza la función de verosimilitud del modelo. Siendo el modelo preferido el que tiene el menor valor del AIC.

- **Criterio de Información Bayesiano (BIC)**:

$$BIC = -2ln\hat{L} + kln(n)$$

Funciona de forma muy parecida al AIC siendo n el tamaño de la muestra y la penalización aun mayor en este criterio.

4.4.7. Inferencia

Se puede realizar inferencia acerca del vector β; en la mayoría de investigaciones se puede formular a través del siguiente sistema de hipótesis:

$$H_o : C\beta = \xi$$

$$H_a : C\beta \neq \xi$$

Entonces, el contraste de estas hipótesis se pueden construir por medio de varios métodos, el más común es denominado *estadístico de Wald*, que se basa en la distribución normal asintótica para $\hat{\beta}$, determinado por:

$$\xi_W = \left[C\hat{\beta} - \xi\right]^T \left[CF^{-1}(\hat{\beta})C'\right] \left[C\hat{\beta} - \xi\right]$$

Donde $F^{-1}(\hat{\beta})$ es la matriz de información de Fisher de $\hat{\beta}$, además, un intervalo de confianza para β con un nivel $(1-\alpha)$ por medio de:

$$\left\{\beta \in R^p \mid (\hat{\beta}-\beta)^T[\hat{Var}(\hat{\beta})]^{-1}(\hat{\beta}-\beta)\chi^2_{p,1-\alpha}\right\}$$

Este último estadístico se distribuye χ^2_s.

Supongamos que estamos considerando una situación especial en la que se evalúa la importancia estadística de un grupo específico de variables predictoras.

$$H_o : \beta_r = 0$$

$$H_a : \beta_r \neq 0$$

Entonces, el *estadístico de Wald* se determina por medio de la siguiente expresión:

$$\xi_W = (\hat{\beta})^T F_r^{-1}(\hat{\beta})\hat{\beta}_r$$

4.4.8. Residuos

Debido posiblemente a factores no controlados al momento de la toma de los datos, es posible que aunque se trate de ajustar el mejor modelo, existan algunas desviaciones que pueden indicar la presencia de valores influyentes que requieren de un análisis. Para ello, existen tres tipos de residuos:

- **Residuo Básico**: Es la diferencia entre el valor observado y el valor predicho por el modelo.

$$r_i^b = y_i - \hat{y}_i$$

- **Residuo de Pearson**: Dado por la expresión:

$$r_i^b = \frac{y_i - \hat{\mu_i}}{\hat{\phi} Var(\hat{\mu_i})}$$

y su versión del **Residuo de Pearson Estudentizado** que capta mejor la variabilidad porque mide la influencia de la i-ésima observación.

$$r_i^b = \frac{y_i - \hat{\mu_i}}{\hat{\phi} Var(\hat{\mu_i})(1-h_i)}$$

con $\hat{\phi}$ siendo un estimador del parámetro y h_i la diagonal de la matriz H, siendo:

$$H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$$

con W siendo una matriz diagonal que está dada por:

$$w_i = \frac{1}{Var(\mu_i)}\left(\frac{\partial \mu_i}{\partial \eta}\right)^2$$

4.4.9. Interpretación

Es importante tener en cuenta que al aplicar una función enlace en un modelo con algunas excepciones, la transformación resultante suele expresarse en términos multiplicativos.

La interpretación de los parámetros se realiza en términos del factor de cambio en el valor esperado para un incremento unitario en las variables explicativas.

Como se nombraba al principio, los modelos lineales generalizados están diseñados para poder tener una variedad de modelos que sirvan cuando la variable dependiente Y es de diferente tipo. Uno de los tipos más usuales en los que se puede encontrar dicha variable de respuesta son en datos discretos, o mejor nombrados, datos de conteo.

Los datos de conteo se trabajan cuando $Y \in (0, 1, 2, \cdots)$ con datos discretos y comúnmente estaría evaluando frecuencias.

4.5. Datos de Conteo

Mencionado por Salinas-Rodríguez et al. (2009), las variables de conteo son aquellas que miden el número de eventos que ocurren en una unidad de observación en un intervalo de tiempo o espacio. Se diferencian de las variables cuantitativas continuas en que son discretas y no negativas. La variable Y que representa el número de eventos toma valores infinitos y su probabilidad disminuye a medida que aumenta su valor.

Los modelos de datos de conteo no tienen un límite superior natural y pueden tomar valores desde cero hasta infinito, cabe destacar para la investigación que suelen tener una alta frecuencia de ceros. Analizar el $E(Y/X = x)$, donde las variables explicativas pueden ser de cualquier tipo.

La modelación de estas variables por medio de algunos modelos tradicionales puede provocar algunas irregularidades al momento de modelar, es por ello que hay modelos que son mejores aproximaciones para este tipo de datos, entre ellos se encuentra el modelo de regresión de Poisson o el modelo de regresión Binomial Negativa.

4.6. Modelo de Regresión de Poisson

Se asume que $Y \sim P(\mu)$ con función de densidad de probabilidad:

$$P(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!}$$

con $Y \in (0, 1, 2, \cdots)$ y μ_0, además de que $E(Y) = \mu$ y $Var(Y) = \mu$.

Se construye un modelo donde la variable de respuesta Y sigue una distribución Poisson donde μ_i es una forma de representar las variables explicativas X. Se utiliza la expresión:

$$E(Y_i|x_i) = exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})$$

Debido a lo anterior, la distribución de Poisson viene dada por:

$$P(Y_i = y_i|x_i) = \frac{e^{-\mu_i(x_i)}\mu_i(x_i)^{y_i}}{y_i!}$$

con,

$$E(Y_i|x_i) = \mu_i(x_i) = \mu(x_{i1}, x_{i2}, \cdots, x_{ip}) = exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})$$

Todo esto se realiza con el objetivo de modelar datos de tipo discreto y no negativo

4.6.1. Propiedades

1. Es un modelo heterocedástico, es decir que las varianzas de los errores no son constantes implicando que la variabilidad es diferente para cada observación.
2. Presenta la propiedad de equidispersión que muestra que $Var(Y_i|x_i) = E(Y_i|x_i)$, si esto no ocurre el modelo no estaría presentando un buen ajuste a los datos y sería viable escoger otro tipo de modelo que capture mejor la dispersión de los datos, es por eso que se hablara posteriormente del modelo binomial negativo que no necesita cumplir este supuesto de equidispersión.

Debido a que el *Modelo Poisson* es un modelo lineal generalizado tiene las métricas que se utilizan para este tipo de modelos.

- **Deviance**:

$$D(y;\hat{\mu}) = 2\left\{l(y;y) - l(\hat{\mu};y)\right\} = 2\sum_{i=1}^{n}\left\{y_i log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\right\}$$

- **Pseudo** R^2:

$$R^2 = 1 - \frac{D(y;\hat{\mu})}{D(y;\hat{\mu_o})} = \frac{\sum_{i=1}^{n}\{y_i log(\mu_i/y_i) - (\hat{\mu_i} - y_i)\}}{\sum_{i=1}^{n}\{y_i log(y_i/\bar{y_i})\}}$$

- **El estadístico** χ^2 **de Pearson**:

$$\chi^2 = \sum_{i=1}^{n}\frac{(y_i - \hat{\mu_i})^2}{V(\hat{\mu_i})} = \sum_{i=1}^{n}\frac{(y_i - \hat{\mu_i})^2}{\hat{\mu_i}}$$

- **Residuo de Pearson**:

$$r_i^p = \frac{y_i - \hat{\mu_i}}{\sqrt{\hat{\mu_i}}}$$

- **Residuo de Pearson Estudentizado**:

$$r_i^p = \frac{y_i - \hat{\mu_i}}{(\sqrt{\hat{\mu_i}})(1 - h_i)}$$

No menos importante está lo señalado por Winkelmann y Zimmermann (1995), cuando no se cumple la propiedad de equidispersión, es decir que $Var(Y_i|x_i) = E(Y_i|x_i)$, cuando esto no sucede podríamos empezar a hablar de un modelo de regresión Binomial Negativa.

4.7. Diagnóstico de la Sobredispersión

Para evaluar si hay sobredispersión en un modelo, hay varias pruebas que se pueden utilizar. Es por eso que es importante abordar algunas de ellas con las que trabajaremos en la aplicación. Se separan en tres grandes grupos:

- **Pruebas para modelos anidados**: Un modelo anidado es un tipo de modelo estadístico en el que uno de los modelos está completamente incluido en el otro.

La versión restringida es una simplificación del modelo no restringido, lo que permite evaluar fácilmente las diferencias entre ellos. En estos modelos las pruebas para diagnosticar sobredispersión se basan en la comparación de la varianza poissoniana con una función de varianza generalizada.

- **Pruebas para modelos no anidados**: Es aquel en el que los modelos que se están comparando no se encuentran completamente incluidos uno en el otro, es decir, no hay una versión restringida del modelo que pueda ser vista como una simplificación del modelo no restringido.
- **Pruebas basadas en regresión**: Se analizan los residuales de Poisson para indicar si se viola el supuesto de equidispersión Chib y Winkelmann (2001).

4.7.1. Pruebas para modelos anidados

La sobredispersión en modelos estadísticos, especialmente en el contexto de modelos anidados, es una situación en la que la variabilidad observada en los datos es mayor de lo esperado bajo la suposición de una distribución probabilística adecuada. Esto puede afectar la validez de las inferencias realizadas a partir del modelo, ya que las estimaciones de los parámetros y los intervalos de confianza pueden no ser precisos. Para diagnosticar en este tipo de modelos se utilizan diversas pruebas y técnicas estadísticas, algunas de las herramientas comunes incluyen:

Prueba de Razón de Verosimilitud (LR)

$$H_o : No\ hay\ sobredispersión : \alpha = 0$$

$$H_a : Si\ hay\ sobredispersión : \alpha \neq 0$$

entonces el estadístico calculado de la prueba está dado por:

$$LR = -2(\hat{l}_r - \hat{l}_{nr}) \sim \chi^2_{(k)}$$

Sin embargo, se realizan algún tipo de modificaciones gracias a Cameron y Trivedi (2013) que señalan que debido a que α no puede ser menor a 0, entonces los grados de libertad del estadístico χ^2 son la diferencia en el número de parámetros estimados entre los dos modelos, o dicho de mejor forma, el número de parámetros adicionales necesarios para el modelo con sobredispersión.

Prueba de Wald

En esta prueba, se ajusta un modelo de regresión con una distribución Poisson y se evalúa la heterogeneidad en los datos.

$$H_o : No\ hay\ sobredispersión : \alpha = 0$$

$$H_a : Si\ hay\ sobredispersión : \alpha \neq 0$$

Su estadístico calculado está dado por:

$$Wald = \frac{(\hat{\alpha - 0})}{\sqrt{\hat{Var}(\hat{\alpha})}} \sim \chi^2_{(1)}$$

4.7.2. Pruebas para modelos no anidados

Los modelos no anidados también pueden enfrentar el problema de la sobredispersión, donde la variabilidad observada en los datos supera las expectativas del modelo estadístico. Aquí hay una breve introducción sobre las pruebas para diagnosticar la sobredispersión en modelos no anidados:

Prueba de Vuong

Vuong (1989) realizo una extensión de la prueba LR para modelos no anidados Chib y Winkelmann (2001), donde las hipótesis están dadas por:

$$H_o : Los\ modelos\ son\ equivalentes : E_0\left[l_f(\hat{\alpha}) - l_g(\hat{\beta})\right] = 0$$

$$H_a : Los\ modelos\ no\ son\ equivalentes : E_0\left[l_f(\hat{\alpha}) - l_g(\hat{\beta})\right] \neq 0$$

y el estadístico calculado viene de la expresión:

$$LR_{NA} = \frac{\frac{1}{\sqrt{n}}\left[l_f(\hat{\alpha}) - l_g(\hat{\alpha})\right]}{\omega} \sim N$$

donde, $\omega^2 = \frac{1}{n}\sum_{i=1}^{n}\left[l_f(y_i/x_i, \hat{\alpha}) - l_g(y_i/x_i, \hat{\beta})\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n}\left[l_f(y_i/x_i, \hat{\alpha}) - l_g(y_i/x_i, \hat{\beta})\right]\right]^2$

Cabe destacar que esta prueba se utiliza para comparar la capacidad predictiva de dos modelos de regresión que tienen diferentes distribuciones de error, a través de la comparación de sus funciones de verosimilitud ajustadas por un factor de penalización. No se utiliza específicamente para diagnosticar la sobredispersión en un modelo de regresión, sin embargo en alguna parte de la literatura como en Brosa (2002) funciona como una buena opción.

4.7.3. Pruebas basadas en regresión

Se evalúan los residuales del modelo de regresión de Poisson según Cameron y Trivedi (2013).

$$\frac{(y_i - \hat{\mu_i})^2 - y_i}{\hat{\mu_i}} = \alpha\hat{\mu_i} + u_i$$

con u_i siendo un término de error y el test en forma general proyecta las siguientes hipótesis:

$$H_o : No\ hay\ sobredispersión : \alpha = 0$$

$$H_a : Si\ hay\ sobredispersión : \alpha \neq 0$$

Evaluando el caso específico para el modelo de regresión binomial negativa I está dada por:

$$\frac{(y_i-\hat{\mu}_i)^2-y_i}{\hat{\mu}_i} = \alpha + u_i$$

Según los mismos autores es importante considerar el valor de la estimación de α.

- Para el *modelo de regresión binomial negativa I* con función de varianza $(1+\alpha)\mu_i$, entonces hay sobredispersión moderada cuando $0 < \alpha < 1$ y sobredispersión alta cuando $\alpha > 1$.

- Para el *modelo de regresión binomial negativa II* con función de varianza $\mu_i + \alpha\mu_i^2 = (1 + \alpha\mu_i)\mu_i$, entonces hay sobredispersión alta cuando $\alpha\mu_i > 1$.

4.8. Modelo de Regresión Binomial Negativa

Cuando sucede que $Var(Y_i|x_i) \geq E(Y_i|x_i)$, es un fenómeno que llamamos sobredispersión y es esta la característica que hace que se haya pensado en la incorporación de un término de perturbación en el modelo de Poisson, más específicamente, en el parámetro μ_i.

$$\mu_i^* = exp(x_i\beta + \varepsilon_i) = \mu_i exp(\varepsilon_i)$$

Donde $\varepsilon_i \sim Gamma$, es por ello que su función de densidad de probabilidad está dada por:

$$P(Y = y_i/x_i) = \frac{\Gamma(y_i+\upsilon_i)}{\Gamma(y_i+1)\Gamma(\upsilon_i)}\left(\frac{\upsilon_i}{\upsilon_i+\mu_i}\right)^{\upsilon}\left(\frac{\mu_i}{\upsilon_i+\mu_i}\right)^{\mu}$$

definiendo a $\mu_i = E\left[Y_i/x_i\right] = exp(x_i\beta)$ y $\upsilon_i = (\frac{1}{\alpha})\mu_i^t$.

El modelo de regresión Binomial Negativa depende de la definición de υ:

- **Modelo de Regresión Binomial Negativo I (NB1)**: Si $\upsilon = (\frac{1}{\alpha})$ entonces,

$$E(Y_i/x_i) = exp(x_i\beta)$$

$$Var(Y_i/x_i) = (1 + \alpha)exp(x_i\beta)$$

- **Modelo de Regresión Binomial Negativo II (NB2)**: Si $\upsilon = (\frac{1}{\alpha})\mu$ entonces,

$$E(Y_i/x_i) = exp(x_i\beta)$$

$$Var(Y_i/x_i) = exp(x_i\beta)(1 + \alpha exp(x_i\beta))$$

En cualquiera de los dos modelos se puede tomar que si $Var(Y_i|x_i) = E(Y_i|x_i)$ entonces $\alpha 0$ y hay sobredispersión.

Según Patil y Boswell (1970), los modelos de regresión *Binomial Negativa* dependen del tipo de problema en el que se vaya a utilizar, ya que existen más de 12 aproximaciones de este modelo.

La distribución binomial negativa se refiere a una variable que analiza la probabilidad de registrar un número específico de fracasos antes de obtener el r-ésimo éxito en una serie de experimentos Bernoulli independientes, r normalmente se considera como un entero positivo.

Entonces se puede decir que una variable aleatoria $Y_i \sim BN(r,p)$ tiene función de probabilidad:

$$P(Y_i = y_i) = \binom{y_i + r - 1}{r - 1} p^r (1-p)^{y_i}$$

con $E(Y_i) = \frac{r(1-p)}{p}$ y $V(Y_i) = \frac{r(1-p)}{p^2}$, es decir que $V(Y_i) = \frac{1}{p}E(Y_i)$, y por lo tanto, $V(Y_i)E(Y_i)$.

Por lo último, se indica que el modelo es apto para modelar la sobredispersión.

4.8.1. Enfoques del modelo

El modelo de regresión *binomial negativa* puede estimarse desde dos enfoques. El primero por Hilbe (1994), en donde se ve como un miembro de la familia de distribuciones exponenciales, lo que lo convierte en un modelo lineal generalizado si se introduce el parámetro de dispersión en la distribución como una constante. Esto permite aplicar técnicas como pruebas de bondad de ajuste y análisis de residuos.

El segundo enfoque es desarrollado por Cameron y Trivedi (2013), en donde el modelo de regresión *binomial negativa* se puede obtener como una distribución compuesta de Poisson y Gamma. En este enfoque, la distribución Gamma se utiliza para ajustar los datos de Poisson que presentan sobredispersión. De esta manera se puede obtener el modelo binomial negativo tradicional, que se representa a menudo como el modelo de regresión $BN2$.

Modelo de Regresión Binomial Negativo como MLG

- **Componente Aleatoria**:

$$Y \sim BN(r,p)$$

$$P(Y = y) = \binom{y + r - 1}{r - 1} p^r (1-p)^y \; con \; y = 0, 1, ...$$

que se expresa como miembro de la familia exponencial negativa así:

$$P(Y = y) = exp\left\{ yln(1-p) + r(ln(p)) + ln\left(\binom{y + r - 1}{r - 1}\right)\right\}$$

aplicando resultados:

$$\theta = ln(1-p) \rightarrow p = 1 - exp(\theta)$$

$$b(\theta) = -rln(p) \rightarrow -r(1 - exp(\theta))$$

$$a(\phi) = 1$$

Derivando por primera y segunda vez con respecto a θ:

$$b'(\theta) = \frac{\partial b}{\partial p}\frac{\partial p}{\partial \theta} = -\frac{r}{p}\left\{-(1-p)\right\} = \frac{r(1-p)}{p} = \mu$$

$$b''(\theta) = \frac{r}{p^2}(1-p)^2 - \frac{r}{p}(1-p) = \frac{r(1-p)}{p^2} = \sigma^2$$

Intercambiando p y r por μ y α:

$$\frac{(1-p)}{(\alpha p)} = \mu \rightarrow \frac{(1-p)}{p} = \alpha\mu \rightarrow p = \frac{1}{(1+\alpha\mu)}$$

con $\alpha = \frac{1}{r}$, entonces:

$$P(Y = y) = \binom{y+\frac{1}{\alpha}-1}{\frac{1}{\alpha}-1} \left(\frac{1}{1+\alpha\mu}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y}$$

entonces,

$$b'(\theta) = \frac{\partial b}{\partial \mu}\frac{\partial \mu}{\partial \theta} = \frac{1}{1+\alpha\mu}\mu(1+\alpha\mu) = \mu$$

$$b''(\theta) = V(\mu) = \mu + \alpha\mu^2$$

- **Función de Enlace**: Se parametriza la relación entre μ y las covariables.

$$g(\mu) = \theta = ln((\alpha\mu)/(1+\alpha\mu)) = -ln(1/\alpha\mu + 1)$$

$$g^{-1}(\mu) = \mu = \frac{1}{\left\{\alpha(e^{-\theta}-1)\right\}}$$

sustentándose como un modelo que tiene la siguiente forma en su función de enlace:

$$ln\left(\frac{1}{\alpha\mu+1}\right) = \eta = x\beta$$

aplicando un enlace logarítmico por cuestiones de facilidad:

$$ln(\mu) = \eta = x\beta$$

Obteniendo el modelo denominado *log-binomial negativa* o *NB2*.

4.8.2. Estimación de los parámetros

En la estimación de modelos de recuento, el método más utilizado es la máxima verosimilitud, para lo cual se recurre a procedimientos numéricos como el método de Newton Raphson o el método Fisher scoring. Estos métodos son comúnmente utilizados en el análisis de datos Poisson y binomiales negativos. Si el modelo es un modelo compuesto Poisson-Gamma, la estimación se realiza mediante el método de Newton Raphson para estimar tanto el parámetro media como el parámetro de dispersión binomial negativo, α. Si el modelo de regresión binomial negativa se considera como un miembro de la familia de modelos lineales generalizados, se utiliza el método Fisher scoring para estimar el parámetro de la media, mientras que el parámetro de dispersión debe ser incluido como una constante conocida en el algoritmo. La mayoría de los autores utilizan ambos métodos para el modelado global. En esta sección se describen brevemente ambos métodos, incluyendo la función log-verosimilitud, la función score y los elementos de la matriz hessiana de cada uno de los modelos binomiales negativos principales.

La expresión que define la función de log-verosimilitud para las dos formas principales del modelo de regresión *binomial negativa*:

$$BN1 : L(\beta, \alpha) = \sum_{i=1}^{n} \left[-ln(y_i!) + \sum_{j=1}^{y_i} ln(\alpha y_i + \mu_i - \alpha_j) - \left(\frac{\mu_i}{\alpha} + y_i\right) ln(1+\alpha)\right]$$

$$BN2 : L(\beta, \alpha) =$$
$$\sum_{i=1}^{n} \left[-ln(y_i!) + \sum_{j=1}^{y_i} ln(\alpha y_i + 1 - \alpha_j) + y_i ln(\mu_i) - \left(\frac{1}{\alpha} + y_i\right) ln(1 + \alpha\mu_i) \right]$$

En el *BN2*, los parámetros pueden ser estimados mediante el método de máxima verosimilitud, conservando las propiedades teóricas de los estimadores. Sin embargo, estimar simultáneamente α y β puede llevar a resultados inconsistentes si la distribución real de la variable respuesta no es *BN2*. Por lo tanto, el modelo *BN2* es más popular que el *BN1* porque el método de máxima verosimilitud solo produce estimadores consistentes de β, y no de α.

La función score está dada por:

$$s(\beta, \alpha) = \left\{ \frac{\partial L(\beta,\alpha)}{\partial \beta}, \frac{\partial L(\beta,\alpha)}{\partial \alpha} \right\}$$

Los estimadores de máxima verosimilitud para β y α son los valores que hacen que la verosimilitud $L(\beta, \alpha)$ sea máxima dentro del rango válido de los parámetros. Las ecuaciones de verosimilitud se utilizan para encontrar estos estimadores de máxima verosimilitud:

- **BN1**:

$$\frac{\partial L(\beta,\alpha)}{\partial \beta} = \sum_{i=1}^{n} x_i' \mu_i \left[\sum_{j=1}^{y_i} \frac{1}{\alpha y_i + \mu_i - \alpha_j} - \frac{ln(1+\alpha)}{\alpha} \right] = 0$$

$$\frac{\partial L(\beta,\alpha)}{\partial \alpha} = \sum_{i=1}^{n} x_i' \mu_i \left\{ \sum_{j=1}^{y_i} \frac{y_i - j}{\alpha y_i + \mu_i - \alpha_j} + \frac{1}{\alpha} \left[\frac{\mu_i ln(1+\alpha)}{\alpha} - \frac{\mu_i + \alpha y_i}{1+\alpha} \right] \right\}$$

con *matriz Hessiana*:

$$H(\beta, \alpha) = \begin{bmatrix} \frac{\partial^2 L(\beta,\alpha)}{\partial \beta \beta'} & \frac{\partial^2 L(\beta,\alpha)}{\partial \beta \partial \alpha} \\ \frac{\partial^2 L(\beta,\alpha)}{\partial \alpha \partial \beta} & \frac{\partial^2 L(\beta,\alpha)}{\partial \alpha^2} \end{bmatrix}$$

con *matriz de información de Fisher esperada*:

$$I_e(\beta, \alpha) = E\left[-H(\beta, \alpha)\right]$$

y con *matriz de información observada en la muestra*:

$$I_{Obs}(\beta, \alpha) = -H(\beta, \alpha)$$

obteniendo

$$\frac{\partial^2 L(\beta,\alpha)}{\partial \beta \beta'} = \sum_{i=1}^{n} x_i' x_i \mu_i \left\{ \sum_{j=1}^{y_i} \frac{\alpha(y_i - j)}{[\alpha y_i + \mu_i - \alpha_j]^2} - \frac{ln(1+\alpha)}{\alpha} \right\}$$

$$\frac{\partial^2 L(\beta,\alpha)}{\partial \beta \partial \alpha} = \sum_{i=1}^{n} x_i' \mu_i \left[\frac{ln(1+\alpha)}{\alpha^2} - \sum_{j=1}^{y_i} \frac{y_i - j}{[\alpha y_i + \mu_i - \alpha_j]^2} - \frac{1}{\alpha(1+\alpha)} \right]$$

$$\frac{\partial^2 L(\beta,\alpha)}{\partial \alpha^2} = \sum_{i=1}^{n} \left\{ - \sum_{j=1}^{y_i} \left(\frac{y_i - j}{\alpha y_i + \mu_i - \alpha_j} \right)^2 + \frac{1}{\alpha^2} \left[\frac{2\mu_i + 3\alpha\mu_i + \alpha^2 y_i}{(1+\alpha)^2} - \frac{2\mu_i ln(1+\alpha)}{\alpha} \right] \right\}$$

- **BN2**:

$$\frac{\partial L(\beta,\alpha)}{\partial \beta} = \sum_{i=1}^{n} \left(x_i' \frac{y_i - \mu_i}{1 + \alpha\mu_i} \right) = 0$$

$$\frac{\partial L(\beta,\alpha)}{\partial \alpha} = \sum_{i=1}^{n} x_i' \mu_i \left\{ \sum_{j=1}^{y_i} \frac{y_i - j}{\alpha y_i + 1 - \alpha_j} + \frac{1}{\alpha} \left[\frac{ln(1+\alpha\mu_i)}{\alpha} - \frac{\mu_i(1+\alpha y_i)}{1+\alpha\mu_i} \right] \right\} = 0$$

con la *matriz Hessiana*, la *matriz de información de Fisher esperada* y la *matriz de información observada en la muestra*, se obtiene:

$$\frac{\partial^2 L(\beta,\alpha)}{\partial\beta\beta'} = \sum_{i=1}^{n} x_i' x_i \mu_i \frac{1+\alpha y_i}{(1+\alpha y_i)^2}$$

$$\frac{\partial^2 L(\beta,\alpha)}{\partial\beta\partial\alpha} = \sum_{i=1}^{n} x_i' \mu_i \frac{y_i-\mu_i}{(1+\alpha\mu_i)^2}$$

$$\frac{\partial^2 L(\beta,\alpha)}{\partial\alpha^2} = \sum_{i=1}^{n} \left\{ -\sum_{j=1}^{y_i} \left(\frac{y_i-j}{\alpha y_i+\mu_i-\alpha_j} \right)^2 + \frac{1}{\alpha^2} \left[\frac{2\mu_i+3\alpha\mu_i^2+\alpha^2\mu_i^2 y_i}{(1+\alpha)^2} - \frac{2ln(1+\alpha\mu_i)}{\alpha} \right] \right\}$$

Para encontrar los estimadores de máxima verosimilitud usando el método de máxima verosimilitud, utilizaremos dos algoritmos iterativos:

- *Newton Raphson*: Este proceso algorítmico inicia con un valor θ_0 y produce una serie θ_k, donde $k = 1, 2, ...$ que se aproxima a $\hat{\theta}$ bajo ciertas condiciones adecuadas.

  Se calcula un estimador consistente de máxima verosimilitud de β y α, determinado por $\hat{\theta} = \binom{\hat{\beta}}{\hat{\alpha}}$.

  $$\theta_{k+1} = \theta_k + \frac{s(\theta)}{I_{Obs}(\theta_k)}$$

  donde $s(\theta) = \frac{\partial L(\theta)}{\partial\theta}$

- *Fisher Scoring*: Este método es el habitualmente utilizado para la evaluación de modelos lineales generalizados y su principal ventaja es que proporciona una estimación coherente y precisa del parámetro β. En este método, el parámetro de dispersión α se considera constante y conocido. Aunque requiere más iteraciones, los calculos son más sencillos.

  $$\beta_{k+1} = \beta_k + \frac{s(\beta_k)}{I_e(\beta_k)}$$

4.8.3. Estimación de otras métricas

Cuando consideramos el modelo de regresión BN2 se pueden establecer las mismas métricas que un modelo lineal generalizado debido a que este se puede estimar como composición de un poisson-gamma o también como un MLG, es por eso que existen ciertas métricas:

- **Deviance**:

  $$D = 2\left\{l(y;y) - l(\hat{\mu};y)\right\} =$$
  $$2\sum_{i=1}^{n} \left\{ y_i ln(y_i/\mu_i) - (y_i + 1/\hat{\alpha})/\hat{\alpha} \cdot ln((1+\hat{\alpha}y_i)/(1+\hat{\alpha}\hat{\mu}_i)) \right\}$$

- **Estadístico χ^2 de Pearson**:

  $$\chi^2 = \sum_{i=1}^{n} \frac{(y_i-\hat{\mu_i})^2}{V(\hat{\mu_i})} = \sum_{i=1}^{n} \frac{(y_i-\hat{\mu_i})^2}{(\hat{\mu_i}+\hat{\alpha}\hat{\mu_i})^2}$$

- **Residuo básico**:

  $$r_i^b = y_i - \hat{y}_i \; con \; i = 1, ..., n$$

- **Residuo de Pearson**:

  $$r_i^p = \frac{(y_i-\hat{\mu_i})}{\sqrt{\hat{\mu_i}+\hat{\alpha}\hat{\mu_i}^2}}$$

con su forma estudentizada,

$$r_i^{pt} = \frac{(y_i - \hat{\mu_i})}{\phi\sqrt{\hat{\mu_i} + \hat{\alpha}\hat{\mu_i}^2}}$$

con su forma estandarizada,

$$r_i^{pst} = \frac{(y_i - \hat{\mu_i})}{\sqrt{(1-h_i)(\hat{\mu_i} + \hat{\alpha}\hat{\mu_i}^2)}}$$

Además, existe un residuo especialmente utilizado para el modelo de regresión binomial negativa denominado *residuo Anscombe*, estos usan la función de varianza de esta forma:

$$Poisson : V(\mu) = \mu$$

$$BN2 : V(\mu) = \mu(1 + \alpha\mu)$$

definiendo,

$$r_i^A = \frac{A(y_i) - A(\hat{\mu_i})}{A'(\hat{\mu_i})\sqrt{V(\hat{\mu_i})}}$$

donde $A(.) = \int d\mu_i / V^{1/3}(\mu_i)$, entonces,

$$r^A = \frac{\left\{3/\hat{\alpha}\left\{(1+\hat{\alpha}y_i)^{2/3} - (1+\hat{\alpha}\mu_i)^{2/3}\right\} + 3(y_i^{2/3} - \hat{\mu_i}^{2/3})\right\}}{2(\hat{\alpha}\hat{\mu_i}^2 + \hat{\mu_i})^{1/6}}$$

4.8.4. Interpretación

Tal cual como en los modelos lineales generalizados, la interpretación de los parámetros se realiza en términos del factor de cambio en el valor esperado cuando una variable explicativa aumenta en una unidad.

El modelo tradicional viene dado por:

$$ln(E[Y]) = \hat{\beta}_0 + X_1\hat{\beta}_1 + ... + X_p\hat{\beta}_p$$

con función de enlace logarítmica, el efecto de cada variable explicativa sobre la variable dependiente puede ser positivo o negativo, dependiendo del valor de $\hat{\beta}$. En los casos en los que una variable explicativa es cualitativa, como en la regresión logística, es necesario crear variables dummy. Si la variable explicativa x_k es una variable dummy, se establece la siguiente relación:

$$\frac{E(Y_i/X_{ik}=1)}{E(Y_i/X_{ik}=0)} = exp(\hat{\beta}_k)$$

Si la variable explicativa X_k es continua, entonces cada incremento unitario en su valor se asocia con un aumento en β^j del logaritmo del valor esperado de la variable respuesta. Esta interpretación es adecuada para el análisis de datos de recuento, ya que implica una restricción de valores positivos para la variable respuesta.

4.9. Otros modelos para datos de conteo

4.9.1. Modelo Cero-Inflado

Propuesto por Lambert (1992), en el que tiene como objetivo poder modelar datos que tienen una cantidad de ceros mayor a los permitidos por alguna distribución.

El modelo de regresión de inflado de ceros es una combinación de dos componentes que da más peso a la probabilidad de que la variable tenga un valor de cero. Esto significa que la función de probabilidad de este modelo es una mezcla de una función de masa concentrada en cero y un modelo perteneciente a la familia exponencial. A diferencia del modelo de Hurdle, el primer componente solo produce recuentos de cero, mientras que el segundo produce toda la gama de recuentos, incluyendo los ceros.

Entonces, por una parte están los famosos "cero falsos" y por otra los ceros que aparecen de la distribución propuesta, entonces:

$$P(Y_i = y_i/y_i = 0) = g + (1-g)f(0)$$

$$P(Y_i = y_i/y_i 0) = (1-g)f(y_i)$$

con g como la probabilidad de encontrar un "cero falso" y $f(0)$ como la distribución escogida para datos de conteo en la que se encuentra un 0.

De esta forma, el modelo cero inflado viene dado por:

$$f_{ceroinf}(y; x, z, \beta, \gamma) = \begin{cases} f_{cero}(0; z, y) + (1 - f_{cero}(0; z, \gamma))f_{cont}(0; x, \beta) & si\ y = 0 \\ (1 - f_{cero}(0; z, y))f_{cont}(y; x, \beta) & si\ y \neq 0 \end{cases}$$

Esta puede ser una descripción de cómo se crean dos modelos y se combinan en uno solo.

Si el coeficiente del componente binario se estima positivamente, esto sugiere que si la variable de referencia se toma, la probabilidad de un conteo mayor que cero aumentará. Por otro lado, el componente de conteos se interpreta de manera similar a los modelos Poisson y Binomial Negativo al analizar los parámetros.

4.9.2. Modelo de Hurdle

Por otra parte esta lo mencionado por Lima (2018), el modelo de Hurdle es un tipo de modelo de regresión utilizado para modelar datos de conteo con exceso de ceros, además, es un modelo en dos partes:

1. Proceso binario para los valores que están por encima o por debajo del valor de selección, que se modela mediante un proceso *logit* para describir la probabilidad de cruzar el .ºbstáculo". Este proceso modela datos que toman dos valores: éxito o fracaso. Sin embargo, es importante destacar que este componente del modelo solo genera conteos cero. Entonces siendo $y_i \sim Ber(p_i)$ con $p_i = E(y_i/x_i)$, entonces el modelo logístico con transformación logit:

$$E(Y/X) = \pi_i = \frac{e^{X\beta}}{1+e^{X\beta}} \rightarrow \pi_i = \frac{1}{1+e^{-X\beta}} \rightarrow \frac{\pi_i}{1-\pi_i} = \frac{1+e^{X\beta}}{1+e^{-X\beta}} = e^{X\beta} \rightarrow log\left(\frac{\pi_i}{1-\pi_i}\right) = X\beta$$

siendo la última expresión $\left(\frac{\pi_i}{1-\pi_i}\right)$ conocida como el odds.

2. Se utiliza un proceso que solo genera conteos mayores que cero, lo que se logra a través de un modelo de Cero Truncado. Este componente puede ser modelado por medio de un modelo Poisson, Binomial Negativo o PIG.

Para terminar con el modelo de Hurdle que tiene la siguiente forma:

$$f_{hurdle}(y; x, z, \beta) = \begin{cases} f_{cero}(0; z, y) & si\ y = 0 \\ (1 - f_{cero}(0; z, y)) f_{cont}(y; x, \beta)/(1 - f_{cont}(0; x, \beta)) & si\ y \neq 0 \end{cases}$$

se postula que los datos se originan a partir de un proceso que produce conteos mayores que cero después de superar un obstáculo. Mientras la barrera no sea cruzada, el proceso solo produce conteos iguales a cero. Los parámetros β y γ del modelo se obtienen mediante el método de máxima verosimilitud y pueden ser optimizados de manera independiente.

La interpretación del modelo de Hurdle se enfoca en cómo los predictores influyen en la probabilidad de tener un valor cero y en la cantidad de eventos que ocurren después de superar ese obstáculo. En la primera parte, los coeficientes estimados representan el efecto de los predictores en la probabilidad de tener un valor cero. En la segunda parte, los coeficientes estimados representan el efecto de los predictores en la cantidad de eventos después de que se supera la barrera.

4.9.3. Modelo de Tweedie

Los Modelos Lineales Generalizados (GLM) se desarrollaron inicialmente para datos que no siguen supuestos tradicionales de normalidad, como los que forman parte de la familia de distribuciones exponenciales. Estas ideas se extienden a los Modelos de Dispersión Exponencial (MDE), que son una clase de distribuciones con parámetros de la familia exponencial lineal y un parámetro adicional de dispersión. Los MDE son importantes en estadísticas y se utilizan en modelos lineales generalizados. La varianza en los MDE se relaciona con la media mediante una función de varianza.

$$E(Y) = \mu$$

$$Var(Y) = \phi V(\mu)$$

La devianza es una herramienta de inferencia general para una amplia gama de datos y se basa en una generalización de la suma de cuadrados residual. Los MDE pueden modelar datos de diversos tipos, incluyendo datos discretos, continuos y mixtos, utilizando distribuciones como Binomial, Poisson, Normal, Inversa Gaussiana, Gamma y Tweedie.

La distribución Tweedie, propuesta por el físico y estadístico Maurice Tweedie en 1984, es un subconjunto de los Modelos de Dispersión Exponencial (MDE). Esta distribución es conocida por su capacidad para modelar datos de tipo discreto, continuo y mixto. Puede acomodar un conjunto de elementos de datos iguales a cero.

El estadístico danés Jørgensen (1987) consolidó el concepto de los modelos de dispersión exponencial y nombró la clase Tweedie en honor a Maurice Tweedie. Formalmente, una variable aleatoria Y que pertenece a los MDE sigue una distribución Tweedie si su función de varianza, que relaciona la media y la varianza, incluye un parámetro de potencia constante p. En términos más precisos, esto se expresa como $Y \sim \text{TW}_p(\mu, \phi)$, donde μ es la media, ϕ es el parámetro de dispersión y p es el parámetro de potencia. Estos parámetros cumplen con $\phi > 0$ y $p \in (-\infty, 0] \cup [1, \infty)$, lo que permite dar forma a la distribución Tweedie. Bonat y Kokonendji (2017)

La función de densidad de probabilidad tiene dos formas de ser escrita:

$$f(y; \mu, \phi) = a(y; \phi) exp\left[\frac{y\theta - k(\theta)}{\phi}\right], \ y \in R$$

$$f(y; \mu, \phi) = b(y; \phi) exp\left\{-\frac{d(y,\mu)}{2\phi}\right\}, \ y \in R$$

Identidad de reescalamiento

Esta propiedad permite que las funciones del modelo mantengan expresiones cerradas cuando se aplica una transformación de escala. Imaginemos una transformación $Z = cY$ para un valor positivo c, donde la variable Y sigue una distribución Tweedie con una media μ y una función de varianza $V(\mu) = \mu^p$. Cuando encontramos la función generadora acumulada para Z, descubrimos que sigue una distribución Tweedie con el mismo parámetro p, una media $c\mu$, y una dispersión $c^{2-p}\phi$.

Además, el Jacobiano de esta transformación es igual a $1/c$ para todos los valores de $y > 0$. La combinación de estos dos eventos resulta en una identidad de reescalamiento extremadamente útil. Esta identidad se expresa como $f(y; \mu, \theta) = cf(cy; c\mu, c^{2-p}\phi)$, lo que significa que podemos seleccionar convenientemente valores de y y parámetros para evaluaciones numéricas y obtener la densidad en otros valores mediante el reescalamiento.

$$f(y; \mu, \phi) = cf(cy; c\mu, c^{2-p}\phi)$$

Modelo

Tal como lo señala Bernales y Daniel (2018), el interes esta en la función de varianza $V(\mu) = \mu^p$ para algunos $p \geq 1$.

Dunn y Smyth (2005) muestra que la notación de este modelo viene dado por $Y \sim ED_p(\mu, \phi)$, que indica una variable aleatoria Y que se distribuye como un modelo de dispersión exponencial Tweedie con media μ, dispersión ϕ y parámetro de potencia $p \in R$ excepto para el intervalo (0,1).

Estimación de los parámetros

El método de estimación de máxima verosimilitud (EMV) se emplea para estimar los parámetros en modelos lineales y no lineales. Para obtener los estimadores de máxima verosimilitud de los parámetros (β_j), se requiere la función de probabilidad. Esta función de probabilidad generalmente se expresa de la siguiente manera.

$$L(\beta; y) = \prod_{i=1}^{n} f(y, \beta)$$

luego,

$$l(\beta; y) = \sum_{i=1}^{n} logf(y, \beta)$$

como MLG, se obtiene

$$l(\theta, \phi; y) = \sum_{i=1}^{n} a(\phi, y) + \frac{1}{\phi}[y\theta - k(\theta)]$$

La combinación de cuatro derivadas demuestra el *score*.

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{x_{ij}}{g'(\mu_i)}$$

4.10. Estadística Espacial

Tal y como lo explica Bohórquez et al. (2008), la estadística espacial abarca métodos estadísticos diseñados para examinar cómo se distribuyen los fenómenos en el espacio y para descubrir patrones y posibles relaciones de causa y efecto en un contexto geográfico. El análisis espacial involucra un conjunto de técnicas que expanden las capacidades del análisis estadístico convencional, especialmente cuando la disposición geográfica de los datos tiene un impacto en las variables que se están estudiando y se considera como un factor relevante en la investigación.

Cuando se observa un mapa, es común que automáticamente se extraiga información de él al identificar patrones, evaluar tendencias y tomar decisiones. Las estadísticas espaciales agregan una capa adicional de información que facilita un análisis más completo y conciso, permitiendo descubrir relaciones previamente desconocidas. Esto, a su vez, brinda la capacidad de responder de manera objetiva a preguntas sobre los datos y tomar decisiones significativas basadas en análisis más allá de la simple observación visual.

La estadística espacial tiene tres grandes enfoques: datos de área o por polígonos, patrones puntuales y geoestadística. En la presente investigación se realizará una agrupación de datos para trabajar como datos de área por lo cual se extenderá más la conceptualización bajo este enfoque.

4.10.1. Datos de área

Se refiere a la distribución de eventos cuya localización se asocia a zonas delimitadas por polígonos, además estos datos están diseñados de forma discreta.

$$\left\{ Z(s) : s \in D \subset R^P \right\}$$

Está dado por un proceso estocástico que tiene número de parámetros DR^P discreto. Por otro lado, el muestreo puede ser considerado de variadas formas y la selección de los sitios de medición dependen totalmente del investigador.

Problema de las áreas modificables (MAUP)

Segun Wong (2004) es un desafío común en la estadística espacial y el análisis de datos geoespaciales. Se refiere a la tendencia de que los resultados de un análisis estadístico puedan variar según cómo se definan o agrupen las unidades de área en un conjunto de datos geográficos. En otras palabras, los resultados de un estudio pueden

cambiar dependiendo de cómo se divida o agregue el área geográfica en unidades más pequeñas o más grandes.

El MAUP puede surgir debido a diferentes niveles de agregación espacial, lo que puede afectar la precisión y la interpretación de los resultados. Esto es especialmente relevante cuando se trabaja con datos geográficos que se han recopilado en diferentes unidades de área, como censos o encuestas que utilizan divisiones políticas o administrativas. El problema puede llevar a conclusiones incorrectas o sesgadas si no se aborda adecuadamente en el análisis.

Como propuesta a la solución de este problema, puede funcionar que la variable sea homogénea, agregar en áreas más pequeñas, cambiar la escala espacial.

Análisis exploratorio espacial

- Permite descubrir errores en la codificación
- Detectar datos atípicos
- Comprobar supuestos

Todo el análisis exploratorio tiene que ver con medidas de centralidad, dispersión y distribuciones de los datos a trabajar.

Por otro lado también puede ser una opción interesante agrupar por medio de un método multivariado estadístico como el dendograma los vecinos más cercanos.

Los mapas coropléticos pueden ayudar a determinar el comportamiento de una variable en los espacios determinados, de esta forma existen varios métodos para hallar estas divisiones:

- *Divisiones naturales:* Las categorías se establecen en función de agrupaciones naturales de los datos, y se pueden establecer diferentes divisiones basadas en puntos de quiebre significativos conocidos por su relevancia, los cuales buscan identificar cambios importantes en los valores de los datos.
- *Divisiones por cuantiles:* Cada categoría contiene una cantidad igual de observaciones, y en la práctica, las divisiones en cuartiles (cuatro categorías) son las más frecuentemente empleadas.
- *Divisiones por intervalos iguales:* Son adecuadas cuando las observaciones están distribuidas de forma relativamente uniforme en su rango, pero si los datos presentan un sesgo significativo podría tener problemas.
- *Divisiones según desviación estándar:* Se basa en divisiones alrededor de la media, en unidades de desviación estándar.

Autocorrelación espacial

- *Autocorrelación positiva:* Según Cliff y Ord (1970) se refiere a la tendencia de que valores similares de una variable tiendan a agruparse en el espacio. En otras palabras, áreas geográficas cercanas exhiben valores similares para la variable de interés.

- *Autocorrelación negativa:* Según Anselin (2020) se produce cuando valores similares de una variable están dispersos o separados en el espacio. En este caso, áreas cercanas exhiben valores que son notablemente diferentes entre sí.
- *Autocorrelación global:* Por otro lado, Cliff y Ord (1970) analizan la existencia de patrones generales de autocorrelación en un conjunto de datos espaciales, indica que se puede detectar si existe una tendencia general de similitud o diferencia en toda el área geográfica
- *Autocorrelación local:* De la misma forma, Anselin (2020) analiza la autocorrelación a nivel local, identificando clústeres o agrupamientos de valores similares y áreas de valores diferentes en un mapa espacial.

Matriz de ponderaciones espaciales

Con el objetivo de determinar la autocorrelación espacial, es necesario determinar la ubicación que rodea a un punto en específico y que además se podría considerar que esa ubicación es influyente. Es por eso, que existen tres alternativas que funcionan como criterio de una ubicación cercana.

Figura 4.1: Tipos de contigüidad. Bohórquez et al. (2008)

Estas ubicaciones se pueden representar por medio de una matriz de ponderaciones W, de tamaño $n \times n$ donde n representa el número de áreas identificadas con un centroide conocido donde cada elemento de dicha matriz caracteriza la relación entre dos áreas.

$$W = \begin{pmatrix} W_{11} & W_{12} & \cdots & W_{1n} \\ W_{21} & W_{21} & \cdots & W_{2n} \\ \vdots & \vdots & & \vdots \\ W_{n1} & W_{n2} & \cdots & W_{nn} \end{pmatrix}$$

Existen diferentes tipos de matrices de ponderaciones:

- *Matriz Binaria:* Los elementos diagonales son cero, además, los elementos W_{ij} $(i \neq j)$ son $\neq 0$, uno en dado caso de que i y j sean vecinos.
- *Matriz W:* Tambien llamada estandarizada por filas, en donde la suma de los elementos W_{ij} son proporcionales con el fin de que tenga un mismo peso y $\frac{W_{ij}}{Wj} = 1$
- *Matriz C:* $\sum_{i=1}^{n} X_{ij} = n$
- *Matriz U:* $\frac{\sum_{i=1}^{n} X_{ij}}{N} = n$
- *Matriz S:* También llamada matriz estandarizadora de varianza, contiene valores que indican la fuerza de la relación entre las observaciones. Si dos observaciones

son vecinas, la matriz S tendrá un valor negativo basado en la inversa de la suma de las distancias de la observación i a todas sus vecinas. Si dos observaciones no son vecinas o si estamos mirando la misma observación, el valor será cero.

Índice de Moran

Como se indica en Ramírez y Falcón (2015) es una herramienta de análisis de datos espaciales que se utiliza para evaluar la autocorrelación espacial. Esto significa que busca determinar si los valores de un atributo en un conjunto de entidades están agrupados, dispersos o distribuidos de manera aleatoria en el espacio. Se calcula el Índice I de Moran, junto con una puntuación z y un valor P para determinar la significancia de este índice.

El Índice Global de Moran, desarrollado por Alfred Pierce Moran, mide la presencia o ausencia de autocorrelación espacial en una variable. Los valores del índice oscilan entre +1 (autocorrelación positiva perfecta), -1 (autocorrelación negativa perfecta) y 0 (patrón completamente aleatorio). Se utiliza en el contexto de la hipótesis nula, que asume que los valores del atributo están distribuidos de manera aleatoria en el área de estudio.

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij})(z_i - \bar{z})(z_j - \bar{z})}{\sum_{i=1}^{n}(z_i - \bar{z})^2}$$

Donde n es el número de áreas, W_{ij} es una matriz de pesos, z_i es el valor de la variable z en el área i y z_j es el valor de la variable z en el área j, por último, $\bar{z}$ es la media aritmética de todos los valores de las variables en todas las áreas.

$$H_o : Es\ un\ proceso\ aleatorio : I = 0$$

$$H_a : Hay\ correlación\ espacial : I \neq 0$$

Índice C de Geary

También Ramírez y Falcón (2015) busca determinar si los valores de un atributo en un conjunto de entidades muestran algún tipo de patrón de agrupación, dispersión o aleatoriedad en el espacio geográfico.

El índice C de Geary varía entre $[0, \infty)$, si el índice C tiende a 0 presenta una autocorrelación positiva, si el índice C tiende a 1 quiere decir que corresponde a un proceso aleatorio y si el índice C supera a 1 todo va en desacuerdo con la ley de Tobler.

Nota: *Ley de Tobler:* Como Celemin (2020) cuenta en su artículo, "Todas las cosas están relacionadas entre sí, pero las cosas más próximas en el espacio tienen una relación mayor que las distantes." Tobler (2004).

$$c = \frac{(n-1)}{2\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(z_i - z_j)^2}{\sum_{i=1}^{n}(z_i - \bar{z})^2}$$

Las hipótesis de las pruebas se comprueban siendo planteadas de la siguiente forma:

$$H_o : Es\ un\ proceso\ aleatorio : C = 1$$

$$H_a : Hay\ correlación\ espacial : C \neq 1$$

Índice de asociación local (LISA)

Como lo expuso Anselin (2020) los gráficos LISA se basan en el estadístico I de Moran de asociación local, que es diferente al estadístico I de Moran global. Este estadístico se calcula de manera individual para cada observación en un mapa, lo que permite identificar puntos calientes (hot spots) o valores atípicos espaciales. La intensidad de estos puntos calientes varía según la significatividad de los estadísticos asociados a cada observación.

Variograma

De acuerdo con los artículos de Ver Hoef y Cressie (1993) e Isaaks y Srivastava (1989) un variograma es una herramienta importante para medir la correlación espacial o la estructura espacial de los datos. Representa cómo varía la covarianza o la semivarianza entre pares de puntos en función de la distancia entre ellos. La idea básica es que si los valores de dos puntos están fuertemente correlacionados espacialmente, la semivarianza entre ellos será baja a medida que la distancia entre los puntos aumenta, y viceversa. El cálculo de las semivarianzas esta dado por:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} (Z(x_i) - Z(x_i + h))^2$$

Con $\gamma(h)$ la semivarianza para una distancia h, $N(h)$ es el número de pares de puntos separados por una distancia h, $Z(x_i)$ es el valor del punto en x_i y $Z(x_i + h)$ es el valor de x_i desplazado por la distancia h.

Bondad de ajuste

- Criterio de información de Akaike (AIC)
- Analisis de residuales
- Comprobación de supuestos
- Selección de variables

4.11. Modelos de regresión espacial

Según Guyón (2010), la regresión espacial es una extensión de los modelos de regresión tradicionales que incorpora la influencia espacial en el análisis de relaciones entre variables. Estos modelos reconocen la posible dependencia espacial entre las observaciones, es decir, la idea de que las unidades geográficas cercanas pueden tener comportamientos o características similares debido a su proximidad espacial.

En un modelo de regresión espacial, se busca capturar y cuantificar la autocorrelación espacial, la tendencia de las observaciones cercanas a ser más similares entre sí que las observaciones más distantes. Esto se logra mediante la inclusión de términos de vecindad espacial en la especificación del modelo.

4.11.1. Modelos de regresión espacial lineales

Tal como supone Mur y Angulo (2006), que las observaciones sean independientes, el modelo se vuelve más simple, pero al analizar datos espaciales, esta simplicidad

puede llevar a resultados inconsistentes debido a la dependencia espacial. Esta dependencia puede manifestarse en las variables explicativas, la variable dependiente o los residuos. Los datos suelen seguir una distribución normal. En el caso de variables de interés que involucran recuentos o proporciones, se espera que los modelos se ajusten a una distribución de Poisson o Binomial. A medida que el número de áreas aumenta, estas distribuciones tienden gradualmente a una distribución normal. Los métodos de estimación de parámetros se realizan mediante el método de máxima verosimilitud.

El modelo estándar espacial está dado de la siguiente forma:

$$y_i = \rho W_1 y_i + \beta X_i + \theta W_2 X_i + \varepsilon_i$$

$$\varepsilon_i = \lambda W_3 \varepsilon_i + u_i$$

Con $u_i \sim N(0, \Omega)$ perteneciendo a las diagonales de $\Omega_{ij} = h_i(z\alpha)$ con h_{i0}, y_i es la variable de interés, X_i es una matriz de covariables, ε_i es el error que incorpora una estructura autorregresiva y W_1, W_2 y W_3 son matrices con ponderación espacial.

Modelo de error espacial (SEM)

Como bien menciona Shi et al. (2022) la dependencia espacial puede ocurrir en los errores y es necesario captar toda esa dependencia espacial. Viene dado por la siguiente expresión:

$$Y = X\beta + (I - \lambda W)^{-1} u$$

$$\varepsilon = (I - \lambda W)^{-1} u$$

los efectos de las interacciones entre los errores no necesitan que tengamos un modelo teórico específico para un proceso de interacción espacial o social. En su lugar, estos efectos son coherentes con dos situaciones:

1. Cuando los factores que no hemos incluido en nuestro modelo están relacionados entre sí en el espacio, es decir, están çerca".
2. Cuando los impactos no observados siguen un patrón espacial, lo que significa que hay una cierta estructura en cómo estos impactos se distribuyen en el espacio.

Modelo de retardo espacial (SLM)

En una aplicación realizada por Lam y Souza (2020) la expresión del modelo viene dada por:

$$y_i = \rho \sum_{j=1}^{n} W_{ij} y_j + \sum_{q=1}^{Q} x_{iq} \beta_q + \varepsilon_i, \quad i = 1, \cdots, n$$

con ρ estimando la relación autorregresiva espacial, por lo tanto, se puede utilizar cuando la dependencia espacial se encuentra en la variable de interés.

Modelo de Durbin espacial (SDM)

Como está propuesto en Mur y Angulo (2006) este modelo no solo considera la variable dependiente en relación con las variables explicativas, sino que también incorpora las variables explicativas que tienen una relación espacial, es decir, están relacionadas con la ubicación geográfica. Además, el modelo considera factores de la variable dependiente y de la matriz de datos en sí misma. Estos factores se promedian en las regiones circundantes a la región de interés. En resumen, el SDM es un modelo

que tiene en cuenta tanto la dependencia espacial entre las variables como la influencia de los factores vecinos en el análisis.

La expresión del modelo de Durbin es:

$$y = \rho W y + X\beta + WX\gamma + \varepsilon$$

4.12. Splines

Para modelar relaciones de datos no lineales como lo indica Toalombo Rojas et al. (2022) se puede utilizar otra alternativa como los splines, en donde se dividen los datos en intervalos y se ajusta un modelo polinómico en cada intervalo, lo que generaría discontinuidades en los puntos de corte, pero podrían añadirse restricciones adicionales para que sus derivadas sean continuas. Lo anteriormente dicho genera tener que transformar la predictora X.

$$1, x, \cdots, x^d, (x - z_1)^d_+, \cdots, (x - z_k)^d_+$$

para después realizar un ajuste lineal con $(x - z)_+ = max(0, x - z)$

El grado del polinomio y el número de nodos son factores clave que determinan la flexibilidad del modelo. La complejidad del modelo se puede medir considerando el número de parámetros en el ajuste lineal, es decir, los grados de libertad.

A medida que aumenta el grado del modelo polinómico, las predicciones tienden a volverse más variables, especialmente en los límites. Para abordar este problema, se utilizan comúnmente los splines naturales. Estos splines de regresión imponen restricciones adicionales, lo que hace que el ajuste sea lineal en los extremos de los intervalos. Esto generalmente resulta en estimaciones más estables en los límites y una mejor capacidad de extrapolación. Estas restricciones también reducen la complejidad del modelo, lo que es equivalente a considerar una nueva base en un ajuste sin restricciones como lo presenta Toalombo Rojas et al. (2022).

4.12.1. Smooth Splines

Wang (2011) muestra que los splines suavizados se obtienen por medio de una funcion dos veces diferenciable que minimiza la suma de cuadrados residual con un hiperparámetro de penalización $0 \leq \lambda < \infty$ y que puede tanto ajustarse a las observaciones ($\lambda = 0$) como ser una línea recta ($\lambda \to \infty$).

$$\sum_{i=1}^{n}(y_i - s(x_i))^2 + \lambda \int s''(x)^2 dx$$

El objetivo de esta forma de modelar es encontrar el hiperparámetro óptimo para el suavizado, para ello se puede utilizar validación cruzada.

4.13. Modelos aditivos generalizados (GAM)

Clemente García (2023) hace una revisión de este tipo de modelos y propuestos por Hastie y Tibshirani (1987) que indican que un (GAM) es un modelo lineal generalizado que incluye un predictor lineal compuesto por la suma de funciones de suavizado que capturan las relaciones entre las covariables. La estructura está dada por la expresión

$$g(\mu_i) = X_i\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \cdots + \varepsilon_i$$

Lo anterior se basa en que $\mu_i \equiv E(Y_i)$ y $Y_i \sim EFD$, además está relacionado con que X_i^* es una fila de la matriz de modelo para cualquier componente estrictamente paramétrico del modelo, θ es el vector parámetro correspondiente, y las f_j son funciones de suavizado de las covariables x_k. El modelo permite una especificación más flexible de la dependencia de la respuesta en las covariables, pero se especifica el modelo solo en términos de funciones de suavizado en lugar de relaciones paramétricas detalladas. Esta flexibilidad y conveniencia conllevan dos nuevos problemas teóricos: la elección de la función de suavizado y el ajuste de sus parámetros (suavidad). La desventaja del modelo estará en la inconsistencia de sus estimaciones cuando existe una fuerte interacción entre variables del tipo $f_{j,k}(x_j x_k)$ ó $f_{j,k}(x_j, x_k)$.

Estimación de los parámetros

- Algoritmo backfitting: Como lo mostró de igual manera Hastie y Tibshirani (1987), sea $E[Y] = \beta_0 + \sum_{j=1}^{p} f_j(x_j)$ se obtiene la estimación maximizando la función de verosimilitud penalizada:

  $$l(\beta; y) - \sum_{j=1}^{p} \lambda_j \beta_j^t S_j \beta_j$$

  siendo los λ_j obtenidos mediante validación cruzada generalizada y las matrices S_j las correspondientes a X_j. Todo esto sigue un proceso iterativo de mínimos cuadrados penalizados.

5. Metodología

5.1. Diseño de la investigación

Se realiza una investigación de tipo correlacional y modelado estadístico debido a que se analizarán posibles relaciones entre variables relacionadas con las pupas del mosquito *Aedes aegypti* por medio de la aplicación de modelos estadísticos en las cuales se podría predecir el número de pupas en función de variables ambientales, o evaluar la eficacia de diferentes estrategias de control, aunque no sea este el objetivo principal del proyecto.

5.2. Variables

Los datos utilizados en el análisis son de un conjunto que describe 393 localidades en el departamento de Cauca. Estas localidades se caracterizan por diferentes variables explicativas, como la altitud sobre el nivel del mar, el número de personas presentes durante el muestreo, la temperatura media diaria, la precipitación total diaria y el porcentaje de humedad relativa. La variable de interés es el conteo de pupas por localidad.

Con el objetivo de incluir la componente espacial se realiza una agregación de datos por municipio, el cual tiene 42 municipios, sin embargo, cabe destacar que existen 9 municipios no muestreados, lo cual deja 33 municipios que cuentan con 5 variables; la altitud sobre el nivel del mar, la densidad de personas (número de personas muestreadas sobre área del municipio), la temperatura media diaria, la precipitación total diaria y el porcentaje de humedad relativa.

5.3. Métodos

Todo se realizó en el software estadístico **R** v4.2.3 Ihaka y Gentleman (1993). Se obtuvieron medidas descriptivas de todas las covariables e incluso de la variable de interés por medio del paquete **DataExplorer** Briers (2019) y el paquete base **stats**.

Para continuar, se comenzaron a ajustar los modelos y sus distintas medidas de bondad de ajuste. Para el ML se utilizó la función **lm()**, mientras que para el modelo GLMP se utilizó la función **glm()** Cleveland et al. (1992). Posteriormente, se realiza un test de sobredispersión utilizando la función **dispersiontest()** del paquete **AER** por Zeileis (2019). Como siguiente paso, se realiza el modelo GLMNB por medio de la función **glm.nb()** de la librería **MASS** Ripley et al. (2013). A los dos últimos modelos se les aplica un test para determinar cero-inflación junto con la función **testZeroInflation()** que pertenece a la librería **DHARMa** Hartig y Hartig (2017). Con la evaluación

de la última prueba se procede a ajustar un modelo cero-inflado con ayuda del paquete **pscl** Zeileis et al. (2008) y su función **zeroinfl()**. Además, se comparan los modelos mediante una prueba de Vuong para modelos anidados como lo son los dos últimos, mediante **vuong()** del mismo paquete **pscl**. Como última tarea se utilizan dos modelos más para intentar tener mejores medidas de bondad de ajuste, el modelo de Hurdle se ajustó por medio de la función **hurdle()** del paquete anterior. Para terminar la primera parte, el modelo de Tweedie utilizo la librería **tweedie** Duval y Tweedie (2000) para estimar los parámetros por medio de la función **tweedie.profile()** y sus medidas de bondad de ajuste con **AICtweedie()**.

Luego, se procedió a realizar la segunda parte que tiene que ver con todo el análisis espacial de los datos y la creación de los modelos con la componente espacial, para ello se realizó un análisis descriptivo espacial.

Para leer el archivo "shapefile"de extensión *.shp* se utiliza la librería **rgdal** Bivand et al. (2015) que tiene como función especializada para dicha tarea **readOGR()**, para observar la correlación espacial entre los datos se utilizaron las funciones **moran.test()**, **geary.test()** y **localmoran()** de la librería **spdep** Bivand et al. (2005). Para realización de los modelos se utilizaron las librerías **spatialreg** Bivand (2022) y **mcgv** Wood y Wood (2015) con sus funciones **lagsarlm()**, **errorsarlm()** y **gam()** respectivamente; con respecto a la evaluación de supuestos se utilizaron funciones de las mismas librerías como **bptest.Sarlm()** y **s()** para las funciones de suavizado de los modelos aditivos generalizados espaciales.

5.4. Flujo de trabajo

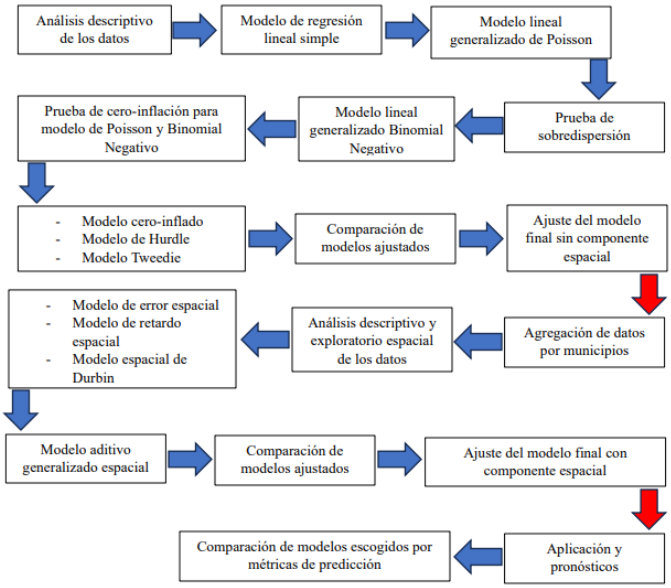


Figura 5.1: Flujo de trabajo

6. Resultados

6.1. Descriptivos

Se realizó la importación de la base de datos en el software. Además, como primer paso se llevó a cabo la verificación de valores perdidos dentro de la data correspondiente. Cabe destacar que la data está compuesta de 10 columnas, 4 de estas columnas son de componentes espaciales (*x, y, XGeo y YGeo*), 6 son variables climáticas y covariables (*msnm, personas, tmed, prec y hum*), por último, está la variable respuesta, que es la que se quiere explicar con la aplicación (*pupa*).

Se observa que no hay valores perdidos en ninguna de las columnas de la base de datos. Por otro lado, para trabajar los modelos estadísticos sin la componente espacial, se dejan por fuera las 4 variables espaciales y se analiza la estructura de las variables con las que se trabajaran los modelos de estadística clásica.

Nombre de la variable	**Naturaleza**	**Descripción**
pupa	Numérica (Discreta)	Pupas encontradas
msnm	Numérica (Discreta)	Metros sobre el nivel del mar
personas	Numérica (Discreta)	Personas por barrio
tmed	Numérica (Continua)	Temperatura media
prec	Numérica (Continua)	Precipitación
hum	Numérica (Continua)	Porcentaje de humedad relativa

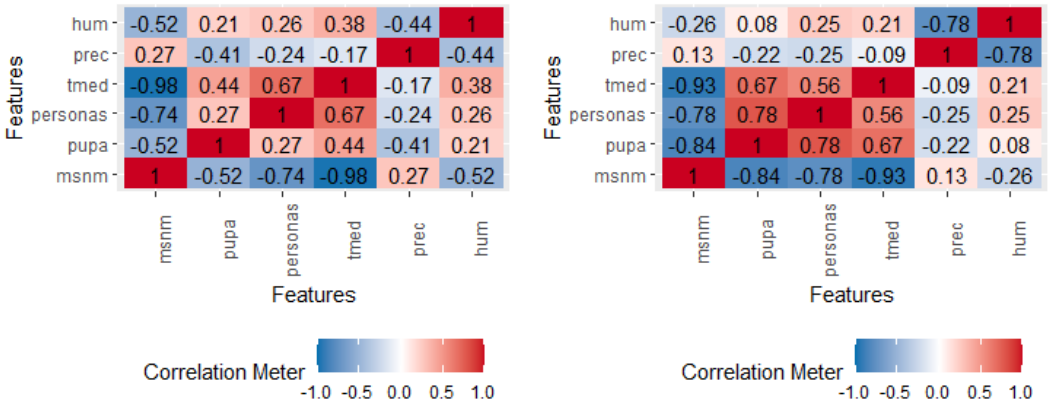
Tabla 6.1: Estructura de las variables

Se establece otro conjunto de variables a las que solo se incluirán a las numéricas con el objetivo de observar la correlación que existe entre estas. Se usará la correlación de Spearman y de Pearson.

Los dos gráficos muestran resultados muy parecidos. En la matriz de correlaciones de Pearson se ve como la variable (*tmed*) tiene una correlación inversa casi perfecta y (*personas*) tiene una correlación inversa alta, ambas con la variable (*msnm*), asimismo, (*tmed*) tiene una correlación directa alta con la variable (*personas*). Por otro lado, en la matriz de correlaciones de Spearman, (*msnm*) tiene una correlación alta inversa con (*tmed*), (*personas*) y la variable respuesta (*pupa*), mientras que (*tmed*) y (*personas*) tienen una correlación alta directa con la variable de respuesta.

Se muestra la distribución de frecuencias de cada variable con el objetivo de dar una descripción clara y concisa de cómo se distribuyen los valores de una variable en particular.

Se nota como la variable (*tmed*) tiene tendencias bimodales, al igual que la variable (*msnm*), la variable (*personas*) presenta asimetría positiva, (*prec*) tiene valores en una escala muy cercana a 0 por su naturaleza de variable climática y de condiciones



(a) Pearson (b) Spearman

Figura 6.1: Matriz de correlaciones

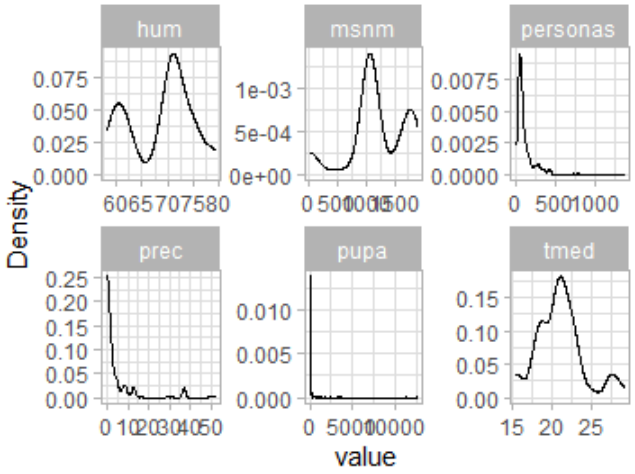


Figura 6.2: Densidad empírica por variable

ambientales y por último la variable de respuesta (*pupa*) muestra que tiene muchos valores de 0.

Además, se grafica la distribución conjunta de la variable (*pupa*) con respecto a las demás covariables numéricas, donde se muestran los siguientes resultados:

Para terminar con el análisis descriptivo se obtienen las medidas de tendencia central y posición de cada variable numérica.

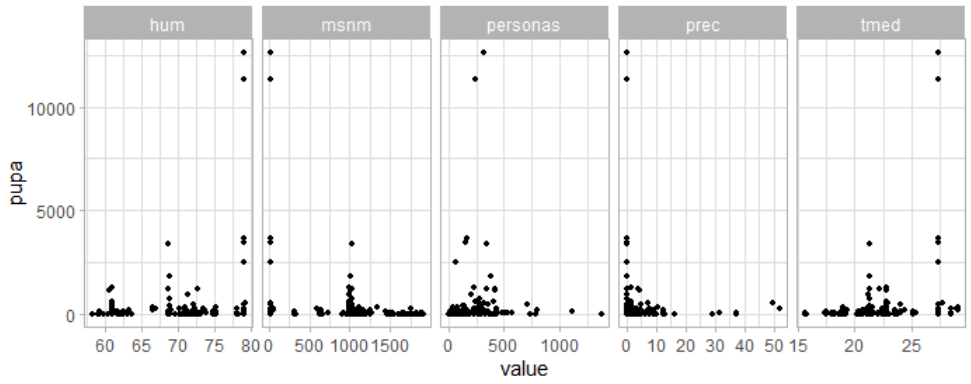


Figura 6.3: Distribución conjunta de pupas por variable

	Nombre de variables					
Medida	*msnm*	*pupa*	*personas*	*tmed*	*prec*	*hum*
Mínimo	17	0	0	15.61	0	58.25
1er Cuartil	1002	0	40	18.94	0	61.69
Mediana	1114	1	63	20.74	1.37	70.89
Media	1189	157	112	21.02	4.12	68.73
3er Cuartil	1691	48	123	22.40	3.60	72.66
Máximo	1904	12636	1371	29.08	51.87	79.09
Desviación Estándar	505.90	925.10	144.81	2.89	8.37	6.10
Coeficiente de Variación (%)	42.55	589.08	129.02	13.76	203.16	8.88

Tabla 6.2: Medidas de tendencia central y posición por variables numéricas

6.2. Modelos

Se ajustaron los modelos de estadística clásica adecuados, comenzando por el más sencillo de todos para ir analizando el porqué no se ajustan bien a los datos, y como toca ir avanzando en la complejidad de los modelos para lograr un buen ajuste.

6.2.1. Modelo de regresión lineal simple (LM)

Como primer paso se ajusta un modelo de regresión lineal simple para demostrar que no funciona por razones anteriormente expuestas, como la naturaleza de la variable *pupa*: número de pupas por unidad de muestreo, que funciona como una variable de conteo, y es por eso que en la tabla se puede ver como el coeficiente del R^2 ajustado es tan cercano a 0, pues es un modelo que no explica bien la variable respuesta. Cabe mencionar que los coeficientes estimados son todos estadísticamente iguales a cero, por lo que no son significativos dentro del modelo, todo esto debido a que no es un modelo propiamente ajustable para la naturaleza de la variable de aplicación.

Modelo de regresión lineal simple (LM)			
Coeficientes	***Estimado***	***Error Estándar***	***Valor-p***
Intercepto	-1236.54	948.66	0.19
msnm	-0.19	0.16	0.23
personas	0.44	0.34	0.19
tmed	35.78	27.48	0.19
prec	-4.5	5.64	0.42
hum	12.27	7.72	0.11
R^2 **Ajustado**	**0.06**		

Tabla 6.3: Modelo de regresión lineal simple (LM)

Por lo anteriormente dicho, y por la naturaleza de la variable, se escoge ajustar un modelo Poisson, que entra dentro de la gama de modelos lineales generalizados, se utiliza cuando la variable es de conteo, como en este caso.

6.2.2. Modelos para datos de conteo

Modelo de regresión de Poisson (GLMP)

Se ajusta un GLMP que intente explicar la variable *pupa* con las demás covariables climáticas.

Modelo de regresión Poisson (GLMP)			
Coeficientes	***Estimado***	***Error Estándar***	***Valor-p***
Intercepto	5.88	$1.017e^{-1}$	$<2e^{-16}$
msnm	$-2.13e^{-3}$	$1.93e^{-5}$	$<2e^{-16}$
personas	$6.45e^{-4}$	$1.54e^{-5}$	$<2e^{-16}$
tmed	$3.82e^{-4}$	$2.83e^{-3}$	0.89
prec	$-2.99e^{-2}$	$7.77e^{-4}$	$<2e^{-16}$
hum	$1.25e^{-2}$	$9.17e^{-4}$	$<2e^{-16}$
Pseudo R^2	**1**		
Deviance (Prueba)	**0**		
AIC	**205546.5**		
BIC	**205570.3**		

Tabla 6.4: Modelo de regresión Poisson (GLMP)

Como se puede ver en la tabla 6.4 el GLMP ajustado tiene un pseudo R^2 de 1, es decir que otorga una explicación perfecta de la variable respuesta. Además, realizando la prueba de la deviance, rechaza la hipótesis nula, es decir, que el modelo se ajusta bien a los datos, y, por otro lado, regala un AIC de 205546 y un BIC de 205570. Cabe aclarar que todos los coeficientes son estadísticamente significativos excepto el de la temperatura. Sin embargo, este modelo (GLMP) es sobredisperso, que puede deberse a que, por ejemplo, el promedio muestral de pupas es menor que su varianza muestral,

lo cual se confirma aplicando un test de sobredispersión (tabla 6.5) que rechaza la hipótesis nula de no sobredispersión (valor-p: 0.005). Se concluye que (GLMP) no es adecuado para explicar el conteo de pupas en función de las variables que se tienen en el estudio.

Test de sobredispersión para (GLMP)		
Estadistico Z	*Valor-p*	*Alpha*
2.55	0.005	1134.19

Tabla 6.5: Test de sobredispersión para el modelo de regresión de Poisson.

Por los resultados anteriores se busca ajustar un modelo que sea capaz de captar la sobredispersión de los datos, es por ello que se opta por el modelo de regresión binomial negativo (GLMNB).

Modelo de regresión binomial negativo (GLMNB)

Se procede a ajustar un modelo (GLMNB) que controle la sobredispersión. Como este modelo también hace parte de la gama de modelos lineales generalizados, se identificaran las mismas medidas para efectos de comparación que el modelo (GLMP).

Modelo de regresión binomial negativo (GLMNB)			
Coeficientes	***Estimado***	***Error Estándar***	***Valor-p***
Intercepto	3.29	2.71	0.22
msnm	-0.001	0.0004	$4.18e^{-5}$
personas	0.007	0.0009	$1.72e^{-15}$
tmed	0.09	0.07	0.21
prec	-0.02	0.01	0.14
hum	-0.003	0.02	0.87
Pseudo R^2	**0.30**		
Deviance (Prueba)	**0**		
AIC	**2848.73**		
BIC	**2876.54**		

Tabla 6.6: Modelo de regresión binomial negativo (GLMNB)

En este modelo se puede evidenciar que el Pseudo R^2 ya no presenta un ajuste que explique de tan buena manera el conteo de pupas, sin embargo, la prueba de la deviance rechaza H_o, es decir que el modelo si está ajustando a los datos. El AIC es mucho menor que él (GLMP), así que si se está modelando mejor la sobredispersión y por último cabe destacar que solo las variables *msnm* y *personas* son estadísticamente distintos de cero y funcionan como variables significativas.

No obstante, aún queda por corregir el tema de la cero inflación, y es que se observó en el análisis descriptivo que la variable de respuesta tiene muchos conteos en 0, es por eso que es necesario aplicar un test de cero inflación tanto para el modelo

(GLMP) como para el modelo (GLMNB), esto con el fin de observar si la sobredispersión obedece a un problema generado por la cero inflación o viceversa.

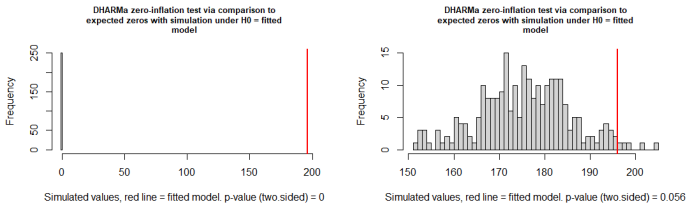
Prueba de cero-inflación	
Valor-p	$<2.2e^{-16}$

(a) Modelo GLMP

Prueba de cero-inflación	
Valor-p	0.056

(b) Modelo GLMNB

Tabla 6.7: Test de cero-inflación para modelos de conteo



(a) GLMP (b) GLMNB

Figura 6.4: Prueba de cero-inflación para modelos de conteo

Como se puede ver las pruebas realizadas, el modelo de poisson tiene cero-inflación y también sobredispersión, por lo que se busca una alternativa que sea capaz de modelar estas dos cosas mencionadas anteriormente. Es por lo mencionado anteriormente que se procede a ajustar el modelo cero-inflado.

6.2.3. Modelo Cero-Inflado (ZINB)

Se ajusta el modelo ZINB con el objetivo de observar si realiza la captura de la sobredispersión y la cero-inflación, y es capaz de modelar de mejor forma la cantidad de pupas por unidad de muestreo.

(ZINB) Modelo de conteo binomial negativo (link: log)			
Coeficientes	***Estimado***	***Error Estándar***	***Valor-p***
Intercepto	-2.56	2.5	0.30
msnm	-0.0002	0.0005	0.65
personas	0.006	0.0009	$1.24e^{-11}$
tmed	0.25	0.08	0.002
prec	-0.03	0.01	0.012
hum	0.01	0.01	0.28
log (theta)	-0.58	0.15	0.0001

Tabla 6.8: Modelo de conteo binomial negativo "link: log"(ZINB)

(ZINB) Modelo cero-inflado binomial (link: logit)			
Coeficientes	***Estimado***	***Error Estándar***	***Valor-p***
Intercepto	-1.42e	4.08	0.0004
msnm	$3.20e^{-3}$	$8.03e^{-4}$	$6.52e^{-5}$
personas	$-2.72e^{-3}$	$1.26e^{-3}$	0.03
tmed	$2.76e^{-1}$	$1.09e^{-1}$	0.01
prec	$6.69e^{-3}$	$1.87e^{-2}$	0.72
hum	$6.78e^{-2}$	$2.46e^{-2}$	0.005

Tabla 6.9: Modelo cero-inflado binomial "link: logit"(ZINB)

Medidas de bondad de ajuste (ZINB)	
Pseudo R^2	0.91
AIC	2758.32
BIC	2809.98
Deviance (Prueba)	0

Tabla 6.10: Medidas de bondad de ajuste del modelo cero-inflado (ZINB)

Prueba de Vuong (ZINB) y (GLMNB)			
Medida	***Estadístico Z-Vuong***	***Ha***	***Valor-p***
Raw	5.63	(ZINB) >(GLMNB)	$8.92e^{-9}$
AIC	4.97	(ZINB) >(GLMNB)	$3.31e^{-7}$
BIC	3.66	(ZINB) >(GLMNB)	0.0001

Tabla 6.11: Prueba de Vuong para modelos anidados (ZINB) y (GLMNB)

En las tablas 6.8 y 6.9 se muestra el ajuste del modelo cero-inflado se pueden obtener algunas interpretaciones con respecto a que tipo de variables tienen influencia en el conteo del número de pupas y que tipo de variables tienen influencia en la presencia o ausencia de pupas.

En la tabla 6.10 se puede observar que el modelo ZINB tiene un AIC mucho menor que los anteriores modelos, además la prueba de Deviance rechaza la hipótesis nula, es decir que el modelo se ajusta bien a los datos. Es decir que en conclusión, es un modelo que sí está capturando la sobredispersión y la cero inflación de los datos. Por otro lado, en la tabla 6.11 se realiza la prueba de Vuong para modelos anidados como lo son el ZINB y el GLMNB, en esta se resume que bajo distintas medidas de bondad de ajuste el modelo ZINB estadísticamente ajusta mucho mejor que el modelo de regresión binomial negativo.

6.2.4. Modelo de Hurdle

Otro modelo que de igual forma que el modelo ZINB intenta capturar esas dos problemáticas de los modelos es el modelo de Hurdle, por lo que se procede a ajustar dicho modelo.

Modelo de conteo binomial negativo truncado (link: log)			
Coeficientes	***Estimado***	***Error Estándar***	***Valor-p***
Intercepto	-2.14	2.40	0.37
msnm	-0.0003	0.0004	0.51
personas	0.006	0.0008	$1.39e^{-11}$
tmed	0.23	0.08	0.003
prec	-0.03	0.01	0.01
hum	0.01	0.01	0.25
log (theta)	0.55	0.14	0.0001

Tabla 6.12: Modelo de conteo binomial negativo truncado "link: log"(Hurdle)

Cabe destacar que los modelos de Hurdle y de cero inflado funcionan de forma muy parecida, ambos tienen un modelo de conteo con función de enlace (log); sin embargo, en el modelo ZINB es un binomial negativo y en el modelo de Hurdle se realiza un truncamiento a la binomial negativa. De otra forma, en el modelo de Hurdle también se plantea otro modelo para el exceso de ceros, a la cual se le realizan algunas modificaciones en la forma de la estimación de los parámetros.

Modelo cero-hurdle binomial (link: logit)			
Coeficientes	***Estimado***	***Error Estándar***	***Valor-p***
Intercepto	10.61	2.55	$3.33e^{-5}$
msnm	-0.002	0.0004	$1.37e^{-8}$
personas	0.003	0.001	0.006
tmed	-0.18	0.07	0.01
prec	-0.01	0.01	0.52
hum	-0.05	0.01	0.003

Tabla 6.13: Modelo cero-hurdle binomial "link: logit"(Hurdle)

Por otro lado, en la tabla 6.14 se muestra el cambio de las medidas de bondad de ajuste. La prueba de la deviance rechaza la hipótesis nula también, quiere decir que el modelo de Hurdle se ajusta bien a los datos. Sin embargo, el valor AIC es mínimamente más alto que el del modelo ZINB, se podría realizar el ajuste.

Medidas de bondad de ajuste (Hurdle)	
Pseudo R^2	0.92
AIC	2761.60
BIC	2813.26
Deviance (Prueba)	0

Tabla 6.14: Medidas de bondad de ajuste del modelo de Hurdle

6.2.5. Modelo de Tweedie

Para continuar con el análisis de modelo, se probará un modelo más que tiene que ver con la distribución Tweedie. Esta distribución cabe dentro de un conjunto de modelos de dispersión exponencial, y tiene la posibilidad de modelar tanto la sobredispersión, pero además tiene una gran masa de ceros al principio de la distribución, por lo que permite modelar los inconvenientes de los datos.

Parametros optimos (Modelo Tweedie)	
Parametro	***Estimado***
p	1.68
ϕ	0.4

Tabla 6.15: Parametros optimos de estimación para el modelo de regresión Tweedie

Para ajustar un modelo de Tweedie primero hay que encontrar el parámetro de potencia y de dispersión óptimos, es por eso que en la tabla 6.15 se muestra que los parámetros óptimos dan lugar a un $p = 1{,}68$ y un $\phi = 0{,}4$.

Modelo de regresión de Tweedie			
Coeficientes	***Estimado***	***Error Estándar***	***Valor-p***
Intercepto	-7.68	4.44	0.08
msnm	-0.002	0.0006	$4.41e^{-5}$
personas	0.02	0.003	$8.89e^{-10}$
tmed	0.33	0.11	0.003
prec	-0.03	0.01	0.05
hum	0.10	0.02	0.0003
Dispersión de Tweedie	**24.69**		
Deviance (Prueba)	**0**		
AIC	**2818.19**		

Tabla 6.16: Modelo de regresión Tweedie

En el modelo de Tweedie nos damos cuenta que tiene un buen ajuste, aunque un valor AIC un poco más elevado que los anteriores dos modelos, su prueba de la Deviance rechaza Ho, quiere decir que el modelo se ajusta a los datos.

6.3. Ajuste del modelo final (Sin componente espacial)

Con la elección del modelo ZINB, se realiza la el método de selección de variables y se presentan los resultados adecuados en la tabla 6.18 y 6.19.

(ZINB) Modelo de conteo binomial negativo (link: log)			
Coeficientes	***Estimado***	***Error Estándar***	***Valor-p***
Intercepto	-2.56	2.5	0.30
msnm	-0.0002	0.0005	0.65
personas	0.006	0.0009	$1.24e^{-11}$
tmed	0.25	0.08	0.002
prec	-0.03	0.01	0.012
hum	0.01	0.01	0.28
log (theta)	-0.58	0.15	0.0001

Tabla 6.17: Modelo de conteo binomial negativo "link: log"(ZINB)

(ZINB) Modelo cero-inflado binomial (link: logit)			
Coeficientes	***Estimado***	***Error Estándar***	***Valor-p***
Intercepto	-1.42e	4.08	0.0004
msnm	$3.20e^{-3}$	$8.03e^{-4}$	$6.52e^{-5}$
personas	$-2.72e^{-3}$	$1.26e^{-3}$	0.03
tmed	$2.76e^{-1}$	$1.09e^{-1}$	0.01
prec	$6.69e^{-3}$	$1.87e^{-2}$	0.72
hum	$6.78e^{-2}$	$2.46e^{-2}$	0.005

Tabla 6.18: Modelo cero-inflado binomial "link: logit"(ZINB)

6.3.1. Significancia estadistica de los coeficientes

- El número de personas influye sobre la cantidad de pupas por unidad de muestreo.
- La temperatura media del lugar de muestreo influye sobre los conteos de pupas.
- Las precipitaciones en el lugar de muestreo influyen sobre los conteos de pupas.
- La altitud o el número de metros sobre el nivel del mar influye en sí se encuentran pupas o no.
- La temperatura media del lugar de muestreo influye sobre si es posible encontrar pupas o no se encuentra ninguna.
- El número de personas influyen directamente en la presencia de pupas.
- La humedad del lugar de muestreo tiene influencia directa sobre la posibilidad de encontrar pupas o no.

6.3.2. Interpretación de los coeficientes estimados

Conteos

- Por cada metro más de altura sobre el nivel del mar, el número de pupas disminuye a razón de 1.0002.

- Por cada persona más, el número de pupas aumenta a razón de 1.006.
- Por cada grado de temperatura más, el número de pupas en la unidad de muestreo aumenta a razón de 1.29.
- Por cada unidad más en el porcentaje relativo de humedad en el lugar de muestreo, el número de pupas aumenta a razón de 1.01
- Por cada precipitación en el lugar de muestreo, el número de pupas disminuye a razón de 1.03

Ceros

- La oportunidad de encontrar pupas o no, aumenta a razón de 0.50 por cada aumento en los metros sobre el nivel del mar.
- La oportunidad de encontrar pupas o no, disminuye a razón de 0.49 por cada persona.
- La oportunidad de encontrar pupas o no, aumenta a razón de 0.56 por cada aumento en la temperatura media de la unidad de muestreo.
- La oportunidad de encontrar pupas o no, aumenta a razón de 0.51 por cada aumento en el porcentaje relativo de humedad media de la unidad de muestreo.
- La oportunidad de encontrar pupas aumenta a razón de 0.50 por cada precipitación en la unidad de muestreo.

6.4. Agregación de datos

Con el objetivo de añadir la componente espacial dentro del análisis y construcción de modelos se realiza la agregación de los datos por municipios muestreados en el departamento del Cauca. Obteniendo 33 municipios con las mismas variables que se trabajaron anteriormente. Se realizaron algunas transformaciones para que todo sea medible y comparable:

- Ya no se trabajará con el número de personas en la unidad de muestreo, en cambio, con la densidad de personas en un municipio.
- Ya no se trabajará con el porcentaje de humedad relativa, en cambio, con la probabilidad de humedad relativa (entre 0 y 1)

Con lo anteriormente dicho, el mapa de los departamentos de Cauca con las unidades muestreadas se ve de la siguiente manera:

Figura 6.5: Mapa de Cauca con datos agregados

6.5. Análisis Descriptivo Espacial

Además también es relevante saber después de la agregación como quedaron las distribuciones de las variables para tener una noción de que relación existente entre el conteo de pupas y sus variables explicativas:

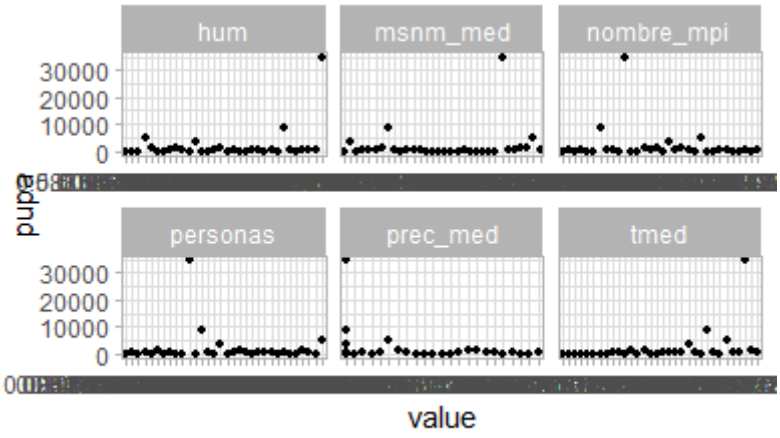


Figura 6.6: Scatterplot de distribución conjunta de pupas agregadas por municipios

Para realizar el análisis descriptivo espacial de los datos por municipio se realiza un agrupamiento utilizando la matriz de distancias con las unidades de área y coordenadas de cada municipio, dando como resultado:

Cabe recordar que el objetivo de este análisis descriptivo espacial es poder dar una idea de cuáles son los municipios más cercanos y lejanos.

También se realiza un mapa coroplético de cuantiles porque organiza los grupos para que tengan la misma cantidad. Entonces, cuenta la cantidad de cada grupo y lo coloca lo más cerca posible al promedio. Teniendo en cuenta lo anteriormente dicho,

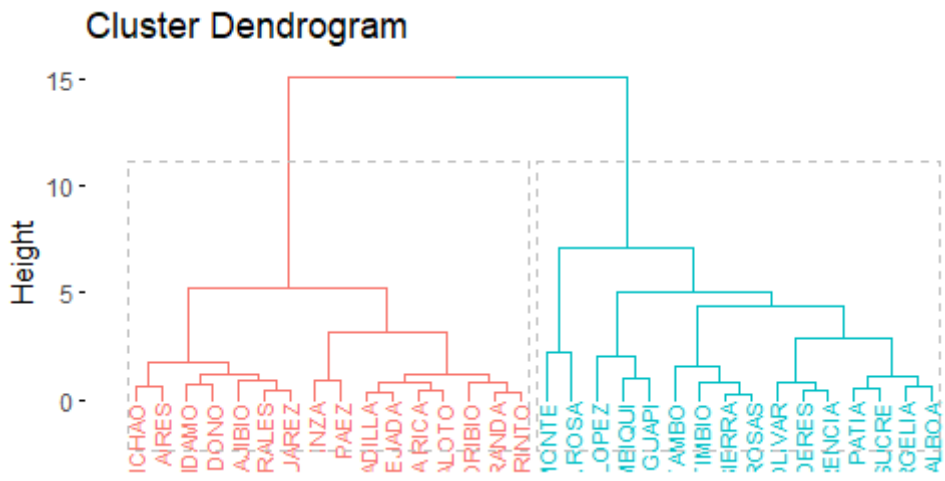


Figura 6.7: Dendograma de municipios

el mapa coroplético utilizando sd e intervalos iguales se utiliza de mejor forma cuando hay homogeneidad en los datos, sin embargo, en este caso por la heterogeneidad se utiliza el mapa de cuantiles que se puede ver de esta forma, cabe destacar que aquí si el conteo de pupas es la variable de interés:

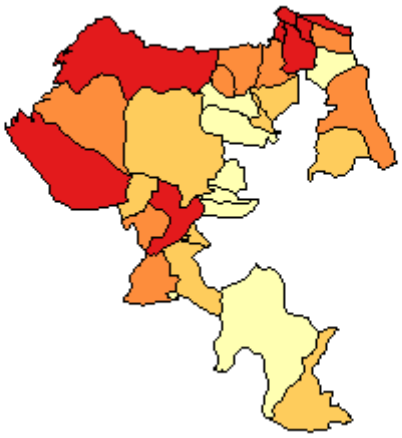


Figura 6.8: Mapa coroplético de municipios

Por otro lado se puede nombrar como datos un poco menos relevantes por la afectación que podrían tener de valores extremos o atípicos que el promedio de pupas es de 1929 pupas promedio por municipio y una desviación estándar de 6245 pupas por municipio.

Como siguiente paso, se realiza la medición de la correlación espacial por medio del índice de moran y el índice de geary para el conteo de pupas.

Conteo de Pupas						
Matriz	*Moran*	*Moran Var*	*Moran Valor-p*	*Geary*	*Geary Var*	*Geary Valor-p*
Reina (W)	-0.036	0.003	0.5	0.881	0.07	0.3
Reina (B)	-0.039	0.003	0.6	0.873	0.22	0.4
Reina (S)	-0.039	0.002	0.6	0.889	0.13	0.4
Torre (W)	-0.032	0.003	0.5	0.873	0.07	0.3
Torre (B)	-0.034	0.003	0.5	0.872	0.22	0.4
Torre (S)	0.035	0.003	0.5	0.884	0.12	0.4

Tabla 6.19: Indice de Morán y Geary para conteo de pupas

Aunque en ningún valor-p de ninguna prueba se rechaza H_o, es decir que no hay evidencia estadística para decir que el conteo viene de un proceso espacial, se escoge la matriz con efecto reina por ser la que tiene un índice de moran más alejado de 0 y un índice de geary más alejado de 1 para poder tener un mejor control de la poca o mínima correlación espacial existente.

De la misma forma, se realizó un análisis de índice local LISA utilizando la matriz binaria tipo reina para determinar en que sitios puede existir la correlación espacial.

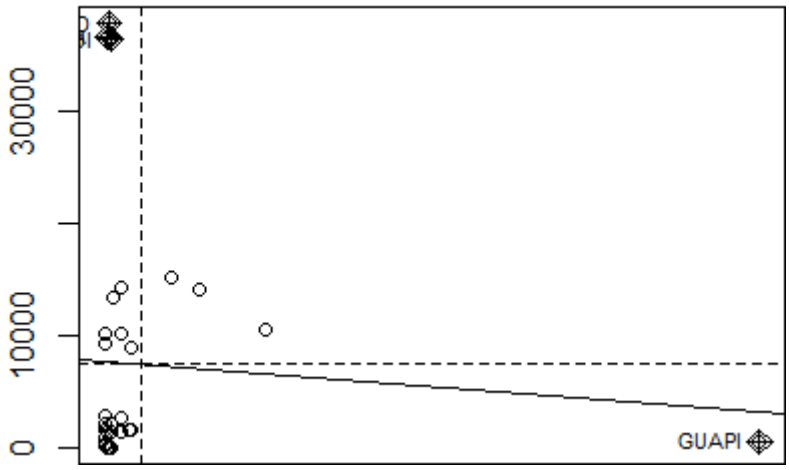


Figura 6.9: Grafico de Moran

Indice LISA (Conteo de Pupas)				
Municipio	***Lisa***	***Esperado***	***Varianza***	***Valor-p***
Timbiqui	-1.34	-0.006	0.20	0.03
Argelia	-1.41	-0.01	0.33	0.01
Guapi	-4.50	-2.76	6.45	0.49
El Tambo	-0.83	-0.02	0.54	0.27

Tabla 6.20: Indice LISA para conteo de pupas

Como se puede ver en la tabla y el gráfico, en Timbiqui y Argelia se rechaza Ho, es decir que hay una correlación espacial local significativa en esos dos municipios, por lo tanto, hay una agrupación espacial de valores altos o bajos de pupas.

6.6. Modelos Espaciales

Se comienza ajustando los modelos espaciales para analizar sus medidas de bondad de ajuste y selección de variables, con el objetivo de escoger el que mejor ajuste presenta.

6.6.1. Modelo Retardo Espacial

Se realiza el primer modelo espacial básico saturado con el objetivo de empezar a explicar la variable pupas por medio de la componente espacial.

Modelo Retardo Espacial			
Coeficientes	***Estimado***	***Error Estandar***	***Valor-p***
Intercepto	-23607.84	26278.74	0.36
msnm	-1.34	3.96	0.73
personas	-53.72	214.16	0.80
tmed	681.49	651.21	0.29
prec	-145.71	128.74	0.25
hum	20956.38	19361.004	0.27
rho	-0.07	0.06	0.28
AIC	**654**		
BIC	**666**		

Tabla 6.21: Modelo de Retardo Espacial

En este modelo se puede ver que aparece un parámetro que controla el retardo de la componente espacial como lo es el *rho*, por otro lado, cabe destacar que el AIC y BIC es bajo, pero que no es comparable con los modelos que hemos presentado anteriormente, pues este tiene los datos agregados para poder representar la componente espacial.

6.6.2. Modelo de Error Espacial

Para continuar con la presentación de modelos básicos espaciales, se presenta un modelo de error espacial que otorga los siguientes resultados:

Modelo Error Espacial			
Coeficientes	***Estimado***	***Error Estandar***	***Valor-p***
Intercepto	-18497.05	23618-47	0.43
msnm	-1.85	3.58	0.60
personas	-47.53	201.59	0.81
tmed	548-45	574.88	0.34
prec	-151.63	124.52	0.22
hum	17739.43	18669.22	0.34
lambda	-0.09	0.06	0.21
AIC	**653**		
BIC	**665**		

Tabla 6.22: Modelo de Error Espacial

Como se puede ver el anterior modelo también trae un parámetro que representa el error espacial, en este caso, *lambda*, sin embargo, las medidas de bondad de ajuste solo traer una unidad menos en comparación con las del anterior modelo.

6.6.3. Modelo Espacial de Durbin

Para terminar con los modelos básicos, se presenta el modelo espacial de Durbin que muestra los siguientes resultados explicando el conteo de pupas en cada municipio del Cauca:

Modelo Espacial de Durbin			
Coeficientes	***Estimado***	***Error Estandar***	***Valor-p***
Intercepto	$-2.11e^{4}$	$3.09e^{4}$	0.49
msnm	-1.27	4.23	0.76
personas	$-1.49e^{2}$	$2.45e^{2}$	0.54
tmed	$6.50e^{2}$	$7.35e^{2}$	0.37
prec	$-1.5e^{2}$	$1.38e^{2}$	0.27
hum	$1.77e^{4}$	$2.31e^{4}$	0.44
lag (intercepto)	$5.69e^{3}$	$1.31e^{4}$	0.66
lag (msnm)	$-6.67e^{-1}$	1.81	0.71
lag (personas)	-5.69	$1.49e^{2}$	0.96
lag (tmed)	$6.90e$	$3.59e^{2}$	0.84
lag (prec)	$-8.06e$	$1.18e^{2}$	0.49
lag (hum)	$-8.35e^{3}$	$1.45e^{4}$	0.56
rho	-0.09	0.06	0.23
AIC	**664**		
BIC	**685**		

Tabla 6.23: Modelo Espacial de Durbin

Como se puede ver en la tabla 6.23, los resultados del ajuste del último modelo no son tan buenos respecto a las medidas de bondad de ajuste y es por esto último,

entendiendo que estos modelos asumen linealidad, se tomó la consideración de aplicar suavizamientos que permitan determinar con mayor exactitud la explicación de la variable de interés.

6.6.4. Modelo Aditivo Generalizado Espacial (SGAM)

Para comenzar, se realiza una prueba de sobredispersión y cero inflación para el conjunto de datos agregados sobre un modelo de Poisson, con el fin de determinar que familia a usar en el modelo aditivo generalizado. Otorgando los siguientes resultados:

Test de sobredispersión para (GLMP) de datos agregados		
Estadistico Z	*Valor-p*	*Alpha*
1	0.1	13586

Tabla 6.24: Test de sobredispersión para GLMP de datos agregados.

Prueba de cero-inflación	
Valor-p	$<2.2e^{-16}$

Tabla 6.25: Modelo GLMP

Las tablas 6.24 y 6.25 presentadas muestran como no hay presencia de sobredispersión, pero si hay presencia de cero-inflación y es por eso que se decide utilizar la familia Tweedie de la cual ya hicimos revisión para incluir dentro de los GAMs.

Como bien hay que destacar se pueden presentar cantidad de suavizados tanto como el número de coeficientes no superen el número de datos, es por eso que se realizó una prueba de modelos aditivos utilizando las posibles combinaciones de suavizados para las covariables como se muestra en la tabla 6.26:

Modelo Aditivo Generalizado Espacial (Tweedie)					
msnm	*personas*	*tmed*	*prec*	*hum*	***AIC***
					448
s					443
	s				448
		s			439
			s		441
				s	448
s	s				443
s		s			439
s			s		437
s				s	444
	s	s			439
	s		s		441
	s			s	448
		s	s		432
		s		s	439
			s	s	441
s	s	s			439
s	s		s		437
s	s			s	444
s		s	s		432
s		s		s	439
s			s	s	437
	s	s	s		431
	s	s		s	439
	s		s	s	441
		s	s	s	432

Tabla 6.26: Suavizamientos de SGAMs

Después de realizar este análisis entendemos que el modelo que mejor se ajusta con un AIC de 431 es el que presenta suavizamientos en las covariables *personas, tmed y prec*. Es por ello que se realiza el ajuste del modelo saturado junto con sus medidas de bondad de ajuste:

Modelo Aditivo Generalizado Espacial (Tweedie)			
Coeficientes	***Estimado***	***Error Estandar***	***Valor-p***
Intercepto	-1.78e^{2}	6.59e	0.014
msnm	-5.08e^{-3}	1.17e^{-3}	0.0003
hum	-1.73	5.21	0.74
spatial coords (lat)	-2.40	8.53e^{-1}	0.01
spatial coords (lon)	2.46	6.11e^{-1}	0.0007
R^2 ***Ajustado***	**0.81**		
Deviance	**83.8 %**		
AIC	**431**		
BIC	**455**		

Tabla 6.27: Modelo Espacial Aditivo Generalizado Tweedie (SGAM)

Modelo Aditivo Generalizado Espacial (Tweedie)		
Coeficientes	***F***	***Valor-p***
s (personas)	1.74	0.18
s (tmed)	4.36	0.01
s (prec)	4.53	0.01

Tabla 6.28: Significancia aproximada de terminos suavizados

En las anteriores tablas 6.27 y 6.28 se muestra como las medidas de bondad de ajuste del AIC y BIC mejoran significativamente con respecto a los modelos lineales, por otro lado, el modelo explica el conteo de pupas en un 81 %.

Significancia estadística de los coeficientes

- Los metros sobre el nivel del mar influyen negativamente sobre los conteos de pupas por municipio
- La latitud del municipio influyen negativamente sobre el conteo de pupas.
- La longitud del municipio influyen positivamente sobre el conteo de pupas.
- El suavizamiento o la no linealidad de la temperatura influye sobre el conteo de pupas por municipio.
- La no linealidad de la precipitación media influye sobre el conteo de pupas por municipio.

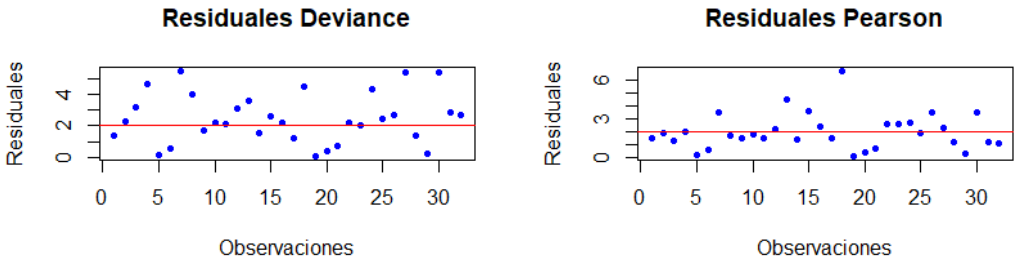
Interpretación de los coeficientes estimados

- Un aumento de una unidad en los metros sobre el nivel del mar está asociado con una disminución de aproximadamente 0.005 unidades en el logaritmo de los conteos de pupas.

- Un aumento de una unidad en la humedad está asociado con una disminución de 1.73 unidades en el logaritmo de los conteos de pupas.
- Un aumento de un grado angular en el sistema de coordinadas geográficas en la latitud está asociado con una disminución de 2.40 unidades en el logaritmo de los conteos de pupas por municipio.
- Un aumento de un grado angular en el sistema de coordinadas geográficas en la longitud está asociado con un aumento de 2.46 unidades en el logaritmo de los conteos de pupas.
- La temperatura media y las precipitaciones medias tienen efectos no lineales en el conteo de pupas.

Análisis de residuales

Para terminar se realiza el análisis de los residuales con el objetivo de detectar datos atípicos e influyentes dentro del modelo que probablemente toque eliminar para que el modelo tenga un mejor ajuste. De esta manera los residuales de la deviance y de pearson otorgan los siguientes resultados:



(a) Deviance (b) Pearson

Figura 6.10: Residuales del SGAM

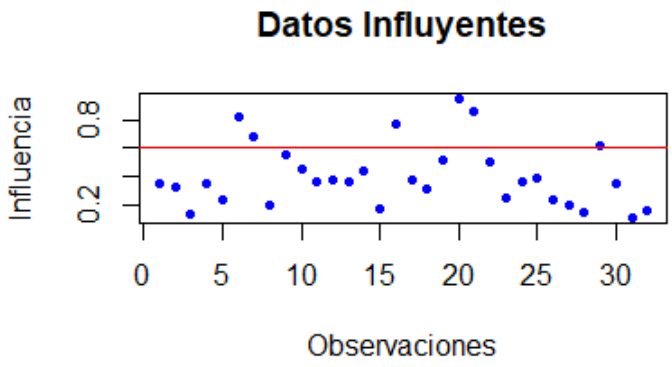


Figura 6.11: Datos influyentes SGAM

En las anteriores gráficas nos podemos dar cuenta que las observaciones 7 y 16 son tanto atípicas como influyentes, es por eso que se deciden eliminar del modelo final.

6.7. Ajuste del modelo final (Con componente espacial)

Se pudo observar que el modelo SGAM es el que mejor ajuste tiene a los datos agregados, sin embargo se realiza la eliminación de las observaciones atípicas e influyentes con el fin de mejorar el ajuste, sin embargo, se volvió a realizar la prueba de suavizamiento por covariable, ya que cabe destacar que el modelo no puede tener más coeficientes que datos, como se puede ver en la tabla:

Modelo Aditivo Generalizado Espacial Final (Tweedie)					
msnm	*personas*	*tmed*	*prec*	*hum*	***AIC***
					402.97
s					392.50
	s				402.98
		s			394.61
			s		384.22
				s	402.98
s	s				392.50
s		s			392.50
s			s		385.12
s				s	392.51
	s	s			394.61
	s		s		384.22
	s			s	402.98
		s	s		384.22
		s		s	394.61
			s	s	384.23

Tabla 6.29: Suavizamientos del modelo final SGAM

De esta manera se puede ver que el término suavizado que minimiza el AIC son las precipitaciones. Obteniendo como modelo final:

Modelo Aditivo Generalizado Espacial Final (Tweedie ρ : 1,64)			
Coeficientes	***Estimado***	***Error Estandar***	***Valor-p***
Intercepto	-1.51e^{2}	8.35e	0.08
msnm	-4.15e^{-3}	1.12e^{-3}	0.001
personas	-9.03e^{-2}	5.56e^{-2}	0.12
tmed	1.32e^{-2}	1.80e	0.94
hum	-3.27	4.77	0.50
spatial coords (lat)	-2.057	1.08	0.07
spatial coords (lon)	2.74	5.50e^{-1}	7.91eˆ-5
R ˆ2 Ajustado	**0.29**		
Deviance	**75.8 %**		
AIC	**384.22**		
BIC	**403.16**		

Tabla 6.30: Modelo Final Espacial Aditivo Generalizado Tweedie (SGAM)

Modelo Aditivo Generalizado Espacial Final (Tweedie ρ : 1,64)		
Coeficientes	***F***	***Valor-p***
s (prec)	6.14	0.002

Tabla 6.31: Significancia aproximada de terminos suavizados del modelo final

En las anteriores tablas se muestra como las medidas de bondad de ajuste del AIC y BIC mejoran significativamente con respecto a los modelos lineales, e incluso frente al primer modelo (SGAM), por otro lado, el modelo explica el conteo de pupas en un 29 %, lo cual es menos adecuado que el primer modelo (SGAM).

Significancia estadística de los coeficientes

- Los metros sobre el nivel del mar influyen de forma negativa sobre los conteos de pupas por municipio
- La coordenadas geograficas de longitud del municipio influyen positivamente sobre el conteo de pupas.
- La no linealidad de la precipitación media influye sobre el conteo de pupas por municipio.

Interpretación de los coeficientes estimados

- Manteniendo todas las demás covariables constantes, un aumento de un grado en la temperatura se espera que esté asociado con un aumento del 0.013 en el logaritmo del número de pupas.
- Un aumento de un grado en la longitud está asociado con un aumento del 2.74 en el logartimo del número de pupas.

- Un aumento de un metro en la altitud sobre el nivel del mar está asociado con una disminución del 0.0041 en el logaritmo del número de pupas por municipio.

- Por cada aumento en la tasa de personas por área, el logaritmo del número de pupas disminuye en 0.09.

- Por cada g/m^3 que aumenta la humedad, el logaritmo del número de pupas disminuye en 3.27.

- Un aumento de un grado en la latitud está asociado con una disminución del 2.05 en el logaritmo del número de pupas.

- A medida que te desplazas hacia el norte (aumento de la latitud), el número de pupas disminuye, y a medida que te desplazas hacia el este (aumento de la longitud), el número de pupas aumenta.

6.8. Aplicación

Se realizaron pronósticos acerca del número de pupas que existirían en el departamento de Cauca si se presentara un aumento en la temperatura media, en las precipitaciones y en la humedad, debido a que estas son las covariables que podrían variar en un posible suceso del calentamiento global del que se habló al principio de la investigación.

6.8.1. Pronósticos

El primer pronóstico se realizó aumentando la temperatura en 2,5 grados en todos los municipios muestreados, con el fin de determinar que pasaría con el número de pupas en cada municipio. Cabe destacar que si se aumenta la temperatura, la humedad y las precipitaciones posiblemente bajen, es por eso que se disminuye en 0,1 y 0,25 respectivamente en cada variable.

El segundo pronóstico se obtuvo duplicando los valores del anterior pronóstico, es decir que se tiene un aumento de temperatura de cinco grados en los municipios, una caída en la humedad de 0,2 y las precipitaciones a la baja de 0,5. En la siguiente tabla se pueden ver los resultados bajo el modelo ajustado SGAM:

Número de pupas predichas por el SGAM (Tweedie)			
Municipio	***Número de pupas***	***Predicciones I***	***Predicciones II***
Argelia	85	138.69	232.67
Balboa	352	363.02	426.48
Bolivar	52	383.98	574.50
Buenos Aires	810	3593.24	5560.56
Cajibio	0	2.05	3.40
Caldono	26	148.09	220.88
Caloto	8475	4256.97	6102.94
Corinto	408	4197.66	6017.91
Patia	1298	1736.89	2542.46
El Tambo	240	214.23	311.37
Florencia	0	5.41	8.76
Inza	41	31.80	45.59
La Sierra	0	76.47	109.63
Lopez	1402	2253.88	3163.93
Mercaderes	833	340.82	521.39
Miranda	1373	1210.26	1854.06
Morales	0	12.96	21.74
Padilla	3482	3809.42	5461.31
Paez	334	188.00	269.53
Piamonte	213	582.32	834.83
Piendamo	94	90.93	130.37
Puerto Tejada	4992	10010.39	15341.25
Rosas	5	13.96	22.64
Santa Rosa	0	5.41	7.87
Santander de Quilichao	850	2503.70	3831.13
Suarez	340	873.60	1454.59
Sucre	91	291.69	418.17
Timbio	6	57.93	83.05
Toribio	0	221.16	317.07
Villa Rica	853	1758.28	2621.75

Tabla 6.32: Pronósticos para cambio de condiciones climáticas

6.8.2. Mapa de valores observados y valores predichos

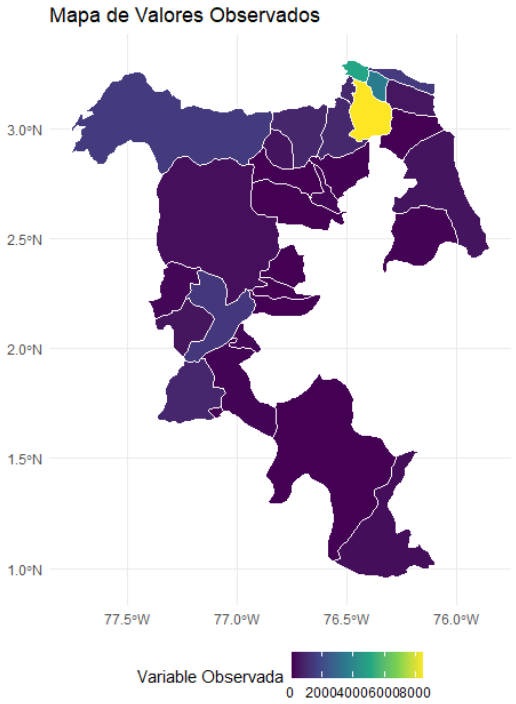
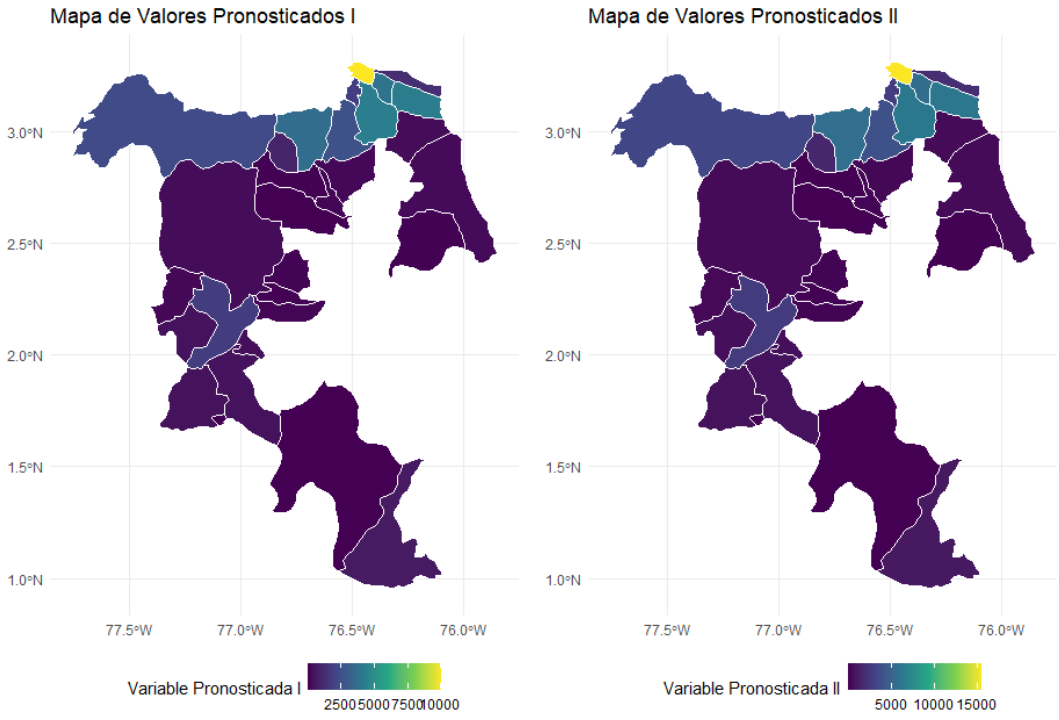


Figura 6.12: Valores observados



(a) Pronosticos I (b) Pronosticos II

Figura 6.13: Mapas de valores predichos

7. Discusión

7.1. Comparación de modelos

Para terminar con el análisis de los modelos estadísticos ajustados, se realizará una comparación de las medidas de bondad de ajuste de cada modelo, tanto para datos desagregados como para datos agregados, en la tabla 7.1. Se puede ver la comparación de los modelos para datos desagregados:

Comparación de los modelos ajustados para datos desagregados			
Nombre del modelo	***AIC***	***BIC***	***Deviance***
Modelo de Poisson (GLMP)	205546.5	205570.3	0
Modelo Binomial Negativo (GLMNB)	2848.73	2876.54	0
Modelo Cero-Inflado (ZINB)	2758.32	2809.98	0
Modelo de Hurdle	2761.60	2813.26	0
Modelo de Tweedie	2818.19		0

Tabla 7.1: Comparación de modelos para datos desagregados

Para concluir, hay 3 modelos que funcionan bien, sin embargo, se toma el AIC como métrica de elección, en este caso, el modelo sin la componente espacial que mejor está ajustando el conteo de pupas por unidad de muestreo es el modelo cero-inflado (ZINB).

Por otro lado están los modelos con la componente espacial para datos agregados, de los cuales se realiza la comparación de medidas de bondad de ajuste en la posterior tabla:

Comparación de los modelos ajustados para datos agregados		
Nombre del modelo	***AIC***	***BIC***
Modelo de Retardo Espacial	653.83	665.55
Modelo de Error Espacial	653.45	665.18
Modelo Espacial de Durbin	664.49	685.01
Modelo Aditivo Generalizado Espacial (SGAM)	430.66	455.07

Tabla 7.2: Comparación de modelos para datos agregados

Por la anterior comparación, se puede observar que el modelo que minimiza AIC y el BIC es el modelo aditivo generalizado espacial (SGAM), y todo esto se explica a que posiblemente haya covariables que no tienen un comportamiento lineal, que es un supuesto importante para los otros modelos espaciales.

Para terminar, se realiza la comparación de los modelos que han tenido mejor ajuste, bajo las métricas de predicciones pues la única forma de comparar todos los

modelos debido a sus diferentes funciones de enlace que tienen a la hora de ajustar los datos:

Comparación predictiva de los modelos ajustados			
Nombre del modelo	***RMSE***	***MSE***	***MAE***
Modelo cero-inflado (ZINB)	884.97 %	7831.85 %	49.43 %
Modelo Aditivo Generalizado Espacial (SGAM)	14.26 %	2.03 %	6.29 %

Tabla 7.3: Medidas de predicción para los modelos finales

De esta forma, nos damos cuenta que si se quiere realizar predicción de acuerdo a las condiciones climáticas, lo mejor es usar el modelo (SGAM) pues tiene unas medidas de error en cuanto a la predicción bajas. Es importante entender que además, la componente espacial si trae buenas contribuciones en cuanto al ajuste del modelo.

7.2. Perspectivas para futuros estudios

- Se plantea la posibilidad de emplear los datos desagregados para la construcción de un modelo aditivo generalizado espacial, aprovechando las coordenadas geográficas como una variable clave en el análisis.
- Explorar el desarrollo de un modelo mixto, incorporando la ubicación como un efecto aleatorio, tanto para los datos agregados como para los desagregados, abre nuevas oportunidades para comprender mejor la influencia de la geolocalización en los resultados.
- La omisión de la componente espacial podría considerarse en casos donde la correlación espacial entre los datos es mínima, permitiendo una simplificación del modelo sin perder información esencial.
- Considerar la posibilidad de tratar los datos desagregados como patrones puntuales, mediante ajustes en el muestreo realizado, podría ofrecer una perspectiva alternativa y enriquecedora para futuros análisis.
- Dejar abierta la puerta al desarrollo de un modelo lineal generalizado espacial mediante estadística bayesiana proporciona una metodología robusta y flexible para abordar complejidades espaciales.
- La exploración de un método de agrupamiento espacial puede enriquecer el análisis exploratorio y descriptivo, ofreciendo insights adicionales sobre la distribución espacial de los datos.
- Realizar un análisis de datos influyentes en el modelo cero-inflado final, particularmente en datos desagregados, permite abordar la coincidencia entre datos atípicos e influyentes, mejorando la robustez del modelo.
- Establecer un criterio para determinar los centroides de los municipios y evaluar cómo este ajuste puede afectar el modelo (SGAM) proporciona una perspectiva para la mejora de la precisión espacial.

- La aplicación de métodos de imputación espacial podría ser explorada para integrar datos adicionales de municipios en el Cauca que no fueron incluidos en el muestreo original, ampliando así la base de datos.
- Ampliar el análisis espacial y los modelos a áreas más pequeñas, como veredas, podría ofrecer una visión más detallada y precisa de la correlación espacial presente en el conjunto de datos.
- Considerar el desarrollo de un semivariograma como un modelo adicional para el análisis exploratorio espacial ofrece una herramienta valiosa para comprender la variabilidad espacial de los datos.
- Explorar modelos lineales generalizados espaciales que incluyan enfoques bayesianos podría proporcionar una perspectiva adicional en términos de robustez y flexibilidad del modelo.

8. Conclusiones

Las diferentes opciones de modelado analizadas revelan aspectos significativos sobre la relación entre las variables climáticas y los conteos de pupas por municipio en el contexto espacial estudiado.

El modelo aditivo generalizado espacial (SGAM) destaca como la opción más efectiva para realizar pronósticos en presencia de variabilidad climática. Sin embargo, es crucial tener en cuenta que este modelo presenta cierta complejidad en la interpretación de las precipitaciones y se basa en datos agregados por municipio.

Por otro lado, el modelo cero inflado (ZINB), aunque presenta errores de predicción superiores al SGAM, ofrece una interpretación más sencilla de los coeficientes, facilitando la inferencia. Tanto el modelo de Hurdle como el modelo de Tweedie se posicionan como opciones válidas para verificar la presencia de cero inflación en los datos.

La ausencia de correlación espacial significativa en las agregaciones por municipio sugiere que la componente espacial puede no aportar significativamente en este nivel de granularidad. La consideración de agrupaciones por unidades de área más pequeñas podría modificar esta correlación y ofrecer una perspectiva más detallada.

La agregación de datos revela que, a nivel lineal, muchas variables presentan relaciones entre sí, excepto en el caso de la precipitación media, que muestra una dinámica diferente.

La falta de claridad en el desarrollo del muestreo subraya la importancia de considerar factores que podrían ser relevantes pero que no fueron inicialmente contemplados.

A pesar de que la agregación de datos es una solución válida en el contexto espacial para trabajar como datos de área, se reconoce la pérdida de información asociada con este enfoque.

La posibilidad de mejorar la explicación del modelo mediante la inclusión de otras covariables ambientales influyentes sugiere una dirección futura de investigación para perfeccionar la comprensión de la relación entre las variables climáticas y los conteos de pupas por municipio.

Sobre la predicción cabe destacar que a medida que aumenta la temperatura y por consecuencia disminuye la humedad y las precipitaciones, el número de pupas aumenta en consecuencia de ello. Cabe destacar que aunque un aumento de temperatura de cinco grados centígrados podría ser inusual, en este contexto, el Cauca experimenta un clima diverso debido a su topografía variada y su posición geográfica cercana al ecuador, lo que podría explicar en algun momento dicho aumento.

9. Bibliografía

Adin Urtasun, A., Martínez Bello, D. A., López Quílez, A., y Ugarte Martínez, M. D. (2018). Two-level resolution of relative risk of dengue disease in a hyperendemic city of colombia. *PLoS ONE, 2018, 13 (9): e0203382.*

Alcalá, L., Niño, L., Morales, J. A., y Castro-Salas, M. (2020). Análisis espacial de un índice pupal de aedes aegypti: una configuración del riesgo de transmisión de arbovirosis. *Investigaciones Geográficas (Esp)*, (74):183–195.

Anselin, L. (2020). Local spatial autocorrelation. *Other Local Spatial Autocorrelation Statistics.*

Atoche Calzada, P. (2017). Modelos de regresión con datos de conteo: aplicación a competiciones deportivas.

Bernales, J. y Daniel, F. (2018). Aplicación del modelo tweedie a datos de conteos sobre consultas médicas del sector privado de chile realizadas en el año 2015.

Bivand, R. (2022). R packages for analyzing spatial data: A comparative case study with areal data. *Geographical Analysis*, 54(3):488–518.

Bivand, R., Bernat, A., Carvalho, M., Chun, Y., Dormann, C., Dray, S., Halbersma, R., Lewin-Koh, N., Ma, J., Millo, G., et al. (2005). The spdep package. *Comprehensive R Archive Network, Version*, pages 05–83.

Bivand, R., Keitt, T., Rowlingson, B., Pebesma, E., Sumner, M., Hijmans, R., Rouault, E., y Bivand, M. R. (2015). Package 'rgdal'. *Bindings for the Geospatial Data Abstraction Library. Available online: https://cran. r-project. org/web/packages/rgdal/index. html (accessed on 15 October 2017)*, page 172.

Bohórquez, I. A., Ceballos, H. V., et al. (2008). Algunos conceptos de la econometría espacial y el análisis exploratorio de datos espaciales. *Ecos de Economia: A Latin American journal of applied economics*, 12(27):9–2.

Bonat, W. H. y Kokonendji, C. C. (2017). Flexible tweedie regression models for continuous data. *Journal of Statistical Computation and Simulation*, 87(11):2138–2152.

Brady, O. J., Gething, P. W., Bhatt, S., Messina, J. P., Brownstein, J. S., Hoen, A. G., Moyes, C. L., Farlow, A. W., Scott, T. W., y Hay, S. I. (2012). Refining the global spatial limits of dengue virus transmission by evidence-based consensus.

Briers, R. (2019). cde-r package to retrieve data from the environment agency catchment data explorer site. *Journal of Open Source Software*, 4(39).

Brosa, J. V. (2002). *El diagnóstico de la sobredispersión en modelos de análisis de datos de recuento.* PhD thesis, Universitat Autònoma de Barcelona.

Calcaterra, E. (2017). Una revisión de los modelos de conteo con excesos de ceros.

Cameron, A. C. y Trivedi, P. K. (2013). *Regression analysis of count data*, volume 53. Cambridge university press.

Carmona, F. (2005). Modelos lineales. *Pub. Univ. de Barcelona, Barcelona*.

Carvajal, J. J., Honorio, N. A., Díaz, S. P., Ruiz, E. R., Asprilla, J., Ardila, S., y Parra-Henao, G. (2016). Detection of aedes albopictus (skuse)(diptera: Culicidae) in the municipality of istmina, chocó, colombia. *Biomédica*, 36(3):438–446.

Cayuela, L. (2010). Modelos lineales: Regresión, anova y ancova. *Eco Lab, Centro Andaluz de Medio Ambiente, Universidad de Granada. Notas de clase*, pages 1–57.

Celemin, J. P. (2020). Cincuenta años de la primera ley de tobler: revisión de sus aportes teóricos y prácticos a la ciencia geográfica.

Chib, S. y Winkelmann, R. (2001). Markov chain monte carlo analysis of correlated count data. *Journal of Business & Economic Statistics*, 19(4):428–435.

Clemente García, F. J. (2023). Modelos lineales generalizados y aditivos generalizados. B.S. thesis.

Cleveland, W. S., Grosse, E., Shyu, W. M., Chambers, J. M., y Hastie, T. J. (1992). Statistical models in s. *Local regression models*, pages Chapter–8.

Cliff, A. D. y Ord, K. (1970). Spatial autocorrelation: a review of existing and new measures with applications. *Economic Geography*, 46(sup1):269–292.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.

Dunn, P. K. y Smyth, G. K. (2005). Series evaluation of tweedie exponential dispersion model densities. *Statistics and Computing*, 15:267–280.

Duval, S. y Tweedie, R. (2000). Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56(2):455–463.

Fisher, R. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society: Series B (Methodological)*, 17(1):69–78.

García Pérez, C. y Alfonso Aguilar, P. (2013). Vigilancia epidemiológica en salud. *Revista Archivo Médico de Camagüey*, 17(6):121–128.

Gomez Villegas, M. A. (2005). *Inferencia estadística*. Ediciones Díaz de Santos.

Guyón, X. (2010). Modelacion para la estadistica espacial. *Revista de investigación Operacional*, 31(1):1–33.

Hartig, F. y Hartig, M. F. (2017). Package 'dharma'. *Vienna, Austria: R Development Core Team*.

Hastie, T. y Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386.

Heinisch, M., Diaz-Quijano, F. A., Chiaravalloti-Neto, F., Pancetti, F. G. M., Coelho, R. R., dos Santos Andrade, P., Urbinatti, P. R., de Almeida, R. M. M. S., y Lima-Camara, T. N. (2019). Seasonal and spatial distribution of aedes aegypti and aedes albopictus in a municipal urban park in são paulo, sp, brazil. *Acta tropica*, 189:104–113.

Hilbe, J. M. (1994). Generalized linear models. *The American Statistician*, 48(3):255–265.

Ihaka, R. y Gentleman, R. (1993). R project. *URL http://www. r-project. org*.

Isaaks, E. H. y Srivastava, R. M. (1989). Applied geostatistics. *(No Title)*.

Jørgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 49(2):127–145.

Kuha, J. (2004). Aic and bic: Comparisons of assumptions and performance. *Sociological methods & research*, 33(2):188–229.

Lam, C. y Souza, P. C. (2020). Estimation and selection of spatial weight matrix in a spatial lag model. *Journal of Business & Economic Statistics*, 38(3):693–710.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Laura Cabezas, Jesús-David Ramos, e. a. (2023). Exploración de modelos estadísticos para identificar regiones geográficas con mayor riesgo para la presencia del mosquito aedes aegypti en el departamento del cauca, colombia.

Lima, C. R. C. (2018). Utilização da estatística gradiente nos modelos hurdle.

Mur, J. y Angulo, A. (2006). The spatial durbin model and the common factor tests. *Spatial Economic Analysis*, 1(2):207–226.

Nelder, J. A. y Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

Neyman, J. (1967). *A selection of early statistical papers of J. Neyman*. University of California Press.

Olano, V. A. (2016). Aedes aegypti en el área rural: implicaciones en salud pública. *Biomédica*, 36(2):169–173.

OPS (2016). *Dengue: Guías para la atención de enfermos en la Región de las Américas*. OPS.

Padilla, J. C., Rojas, D. P., y Gómez, R. S. (2012). *Dengue en Colombia: epidemiología de la reemergencia a la hiperendemia*. Guías de Impresión Ltda.

Patil, G. y Boswell, M. (1970). A characteristic property of the multivariate normal density function and some of its applications. *The Annals of Mathematical Statistics*, 41(6):1970–1977.

Pedersen, E. J., Miller, D. L., Simpson, G. L., y Ross, N. (2019). Hierarchical generalized additive models in ecology: an introduction with mgcv. *PeerJ*, 7:e6876.

Pina, M. F., Alves, S. F., Ribeiro, A. I. C., y Olhero, A. C. (2010). Epidemiología espacial: nuevos enfoques para viejas preguntas. *Universitas Odontológica*, 29(63):47–65.

Pontaque, F. C. (2005). *Modelos lineales*, volume 3. Edicions Universitat Barcelona.

Ramírez, A. I., Torres, P., Fabro, G., Tosolini, L., y Ferreira, M. (2013). Epidemias y salud pública.

Ramírez, L. y Falcón, V. (2015). Autocorrelación espacial: Analogías y diferencias entre el índice de moran y el índice getis y ord. aplicaciones con indicadores de acceso al agua en el norte argentino. *Ponencia presentada en las Jornadas Argentinas de Geotecnologías, Universidad Nacional de San Luis*, 2.

Reiter, P. (2001). Climate change and mosquito-borne disease. *Environmental health perspectives*, 109(suppl 1):141–161.

Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., Firth, D., y Ripley, M. B. (2013). Package 'mass'. *Cran r*, 538:113–120.

Ross, N., Miller, D. L., Simpson, G. L., y Pedersen, E. J. (2018). Hierarchical generalized additive models: an introduction with mgcv. Technical report, PeerJ Preprints.

Rotela, C. H. (2012). Desarrollo de modelos e indicadores remotos de riesgo epidemiológico de dengue en argentina.

Salinas-Rodríguez, A., Manrique-Espinoza, B., y Sosa-Rubí, S. G. (2009). Análisis estadístico para datos de conteo: aplicaciones para el uso de los servicios de salud. *salud pública de méxico*, 51:397–406.

Shi, D., DiStefano, C., Maydeu-Olivares, A., y Lee, T. (2022). Evaluating sem model fit with small degrees of freedom. *Multivariate behavioral research*, 57(2-3):179–207.

Singh, S. K. (2013). *Viral infections and global change*. John Wiley & Sons.

Sothe, C., Camargo, E. C. G., Gerente, J., Rennó, C. D., y Monteiro, A. M. V. (2017). Uso de modelo aditivo generalizado para análise espacial da suscetibilidade a movimentos de massa. *Revista do Departamento de Geografia*, 34:68–81.

Toalombo Rojas, B., Meneses Freire, A., Zúñiga Lema, L., y Espin Guerrero, R. (2022). Modelos de regresión b-splines paramétricos y no paramétricos polinómicos. una aplicación de ingeniería.

Tobler, W. (2004). On the first law of geography: A reply. *Annals of the Association of American Geographers*, 94(2):304–310.

Ver Hoef, J. M. y Cressie, N. (1993). Multivariable spatial prediction. *Mathematical Geology*, 25:219–240.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: journal of the Econometric Society*, pages 307–333.

Wang, Y. (2011). *Smoothing splines: methods and applications*. CRC press.

Winkelmann, R. y Zimmermann, K. F. (1995). Recent developments in count data modelling: theory and application. *Journal of economic surveys*, 9(1):1–24.

Wong, D. W. (2004). The modifiable areal unit problem (maup). In *WorldMinds: geographical perspectives on 100 problems: commemorating the 100th anniversary of the association of American geographers 1904–2004*, pages 571–575. Springer.

Wood, S. y Wood, M. S. (2015). Package 'mgcv'. *R package version*, 1(29):729.

Yi, H., Devkota, B. R., Yu, J.-s., Oh, K.-c., Kim, J., y Kim, H.-J. (2014). Effects of global warming on mosquitoes & mosquito-borne diseases and the new strategies for mosquito control. *Entomological Research*, 44(6):215–235.

Zeileis, A. (2019). Kleiber c. countreg: count data regression. 2018.

Zeileis, A., Kleiber, C., y Jackman, S. (2008). Regression models for count data in r. *Journal of statistical software*, 27(8):1–25.

Anexos

A. Anexo I: Codigo utilizado en el software R Project

```
# Librerias
library(readxl)
library(dplyr)
library(tidyverse)
library(DataExplorer)
library(DescTools)
library(ltm)
library(car)
library(MASS)
library(rcompanion)
library(pscl)
library(robust)
library(sf)
library(ggplot2)
library(gnm)
library(AER)
library(VGAM)
library(sp)
library(rgdal)
library(sf)
library(ggplot2)
library(tmap)
library(mapview)
library(tidyverse)
library(gdtools)
library(RColorBrewer)
library(readxl)
library(dplyr)
library(raster)
library(spdep)
library(spData)
library(spatialreg)
library(tseries)
library(classInt)
library(spatialreg)
library(splm)
library(DescTools)
library(AER)
library(DHARMa)
library(pscl)
```

```
library(mgcv)
library(gam)
library(tweedie)
library(statmod)

# Cargar la base de datos y directorio
datos <- read_excel("C:\\Users\\sebas\\Downloads\\V02_BaseDatosEneroPupa
Persona1.xlsx")

# Estructura de los datos
glimpse(datos)

# Datos desagregados con las coordenadas espaciales
espacial <- dplyr::select(datos, -Municipi, -FechaDefin, -nombre
, -pup_pers, -tmax, -tmin, -tmax_5, -tmax_7, -tmin_5, -tmin_7, -tmed_5,
-tmed_7, -prec_5, -prec_7, -hum_5, -hum_7, -areaHa, -lang, -Tipo)

# Datos desagregados sin las coordenadas espaciales
regresion <- espacial %>% dplyr::select(-x, -y, -XGeo, -YGeo)

# Presencia de datos faltantes
profile_missing(espacial)
plot_missing(espacial)

# Estructura de las variables
str(espacial)

# Matriz de correlaciones de Spearman
correlacion <- regresion
S <- cor(correlacion, method = "s")
round(S,2)
plot_correlation(S)

# Matriz de correlaciones de Pearson
P <- cor(correlacion, method = "p")
round(P,2)
plot_correlation(P)

# Distribucion de frecuencias
DataExplorer::plot_density(regresion,
                           ggtheme = theme_light(), nrow=2, ncol=2)

# Distribucion conjunta de pupas
DataExplorer::plot_scatterplot(regresion,
                               by = "pupa",
                               geom_point_args = list(size=1L),
                               nrow = 1,
```

```
                                ncol = 5,
                                ggtheme = theme_light())

# Resumen de variables
summary(regresion)

# Modelo de regresion lineal simple
modlineal <- lm(pupa~msnm+personas+tmed+prec+hum, data=regresion)
summary(modlineal)

# Modelo Poisson
modpoisson <- glm(pupa~msnm+personas+tmed+prec+hum, data = regresion,
family = poisson)
summary(modpoisson)
1-pchisq(322209-204376, 392-387) # Prueba de la Deviance
PseudoR2(modpoisson, "Nagelkerke") # Pseudo R2
AIC(modpoisson) # AIC
BIC(modpoisson) # BIC

# Prueba de sobredispersion
dispersiontest(modpoisson, trafo = 1)

# Modelo de regresion Binomial Negativa
modbn0 <- glm.nb(pupa~msnm+personas+tmed+prec+hum, data=regresion)
summary(modbn0)
1-pchisq(539.58-359.75, 392-387) # Prueba de la deviance
PseudoR2(modbn0, "Nagelkerke") # Pseudo R2
AIC(modbn0) # AIC
BIC(modbn0) # BIC

# Analisis de varianza modelo poisson y el modelo binomial negativo
anova(modpoisson, modbn0)

# Prueba de cero inflacion
testZeroInflation(modpoisson) # Modelo Poisson
testZeroInflation(modbn0) # Modelo Binomial Negativo

# Modelo Cero Inflado
mod_ceroinf <- zeroinfl(pupa~msnm+personas+tmed+prec+hum, data=regresion,
dist="negbin")
summary(mod_ceroinf)
AIC(mod_ceroinf) # AIC
BIC(mod_ceroinf) # BIC
1 - pchisq(2*abs(logLik(mod_ceroinf)[1] - logLik(modbn0)[1]), 17-9)
# Prueba de la deviance
null_model <- zeroinfl(pupa~1, data=regresion, dist="negbin") # Modelo Nulo
logLik(mod_ceroinf) / logLik(null_model) # Modelo Saturado / Modelo Nulo
```

```
# Prueba de Vuong
vuong(mod_ceroinf, modbn0)

# Modelo Hurdle
mod_hurdle <- hurdle(pupa~msnm+personas+tmed+prec+hum, data=regresion,
dist = "negbin")
null_hurdle <- hurdle(pupa~1, data=regresion, dist="negbin") # Modelo Nulo
logLik(mod_hurdle) / logLik(null_hurdle) # Modelo Saturado / Modelo Nulo
summary(mod_hurdle)
AIC(mod_hurdle) # AIC
BIC(mod_hurdle) # BIC
1 - pchisq(2*abs(logLik(mod_ceroinf)[1] - logLik(mod_hurdle)[1]), 17-17)
# Prueba de la deviance

# Modelo Tweedie
pow3 <- tweedie.profile(pupa~msnm+personas+tmed+prec+hum,
p.vec=seq(1.1, 1.9, by=0.1), do.plot=TRUE, data=regresion)
pow3$phi # Parametro de dispersión

model_tweedie <- glm(pupa~msnm+personas+tmed+prec+hum, data = regresion,
family=tweedie(var.power=pow3$p.max,link.power=0.4))
AICtweedie(model_tweedie) # AIC
1-pchisq(10924.2-5596.3, 392-385) # Prueba de la Deviance
summary(model_tweedie)

model_tweedie2 <- glm(pupa~1, data = regresion,
family=tweedie(var.power=pow3$p.max,link.power=0.4)) # Modelo Nulo
logLik(model_tweedie) / logLik(model_tweedie2) # Modelo Saturado / Modelo Nulo

# Modelo Final para datos desagregados (Cero-Inflado)

mod0 <- step(mod_ceroinf) # Seleccion de variables
summary(mod0)
AIC(mod0) # AIC

# Interpretacion de coeficientes
exp(mod0$coefficients$count)
exp(mod0$coefficients$zero)/(1 + exp(mod0$coefficients$zero))
1/0.99977360

# Establecer el diseño de la ventana gráfica
par(mfrow = c(1, 1))

# Gráfico de deviance residuals
plot(abs(residuals(mod0)),
     main = "Residuales Deviance",
```

```
     xlab = "Observaciones",
     ylab = "Residuales",
     pch = 20, col = "blue")
abline(h = 2, col = "red")
which(abs(residuals(mod0)) > 2)

# Gráfico de Pearson residuals
plot(abs(residuals(mod0, type = "pearson")),
     main = "Residuales Pearson",
     xlab = "Observaciones",
     ylab = "Residuales",
     pch = 20, col = "blue")
abline(h = 2, col = "red")
which(abs(residuals(mod0, type="pearson")) > 2)

# Directorio
setwd("C:/Users/sebas/Desktop/Noveno Semestre U/Tesis/Cauca")
dir()

# Datos de pupas y covariables
pupas <- read_excel("Datos_agregados.xlsx", sheet = 2)

# Eliminacion de lugares sin informacion de muestreo
pupas <- pupas[-1,] # Almaguer
pupas <- pupas[-14,] # Jambalo
pupas <- pupas[-15,] # La Vega
pupas <- pupas[-23,] # Popayan
pupas <- pupas[-24,] # Purace
pupas <- pupas[-25,] # San Sebastian
pupas <- pupas[1:32,] # Silvia, Sotara y Totoro

# Lectura del shapefile
cauca <- readOGR("shapes.shp", layer = "shapes")

# Eliminacion de lugares sin informacion en shape file
cauca <- cauca[-40,] # San Sebastian
cauca <- cauca[-39,] # La Vega
cauca <- cauca[-38,] # Silvia
cauca <- cauca[-37,] # Isla Gorgona
cauca <- cauca[-34,] # Almaguer
cauca <- cauca[-33,] # Jambalo
cauca <- cauca[-27,] # Purace
cauca <- cauca[-21,] # Sotara
cauca <- cauca[-12,] # Popayan
cauca <- cauca[-11,] # Totara

# Union de las bases de datos
```

```
base <- merge(cauca, pupas, by="nombre_mpi")

# Eliminacion de columnas irrelevantes
base <- base[,c("nombre_dpt","mpio","nombre_mpi","pupa", "msnm_med",
"personas", "tmed","prec_med", "hum")]
names(base)
head(base)

# Modelo poisson para pupas
lm <- lm(pupa~msnm_med+personas+tmed+prec_med+hum, data=base)
summary(lm)

# Ressiduales del modelo
lm$residuals

# Shapefile
plot(base)

# Mapa de cauca
xy=coordinates(base)
par(mai=c(0,0,0.3,0))
plot(base, main="Cauca por municipio")
text(xy, base$nombre_mpi, cex=0.5, col="blue")

mapview(base)
mapview(base, zcol = c("pupa"),col.regions = brewer.pal(9, "Purples"))

# Calculo de matriz de distancia y dendograma
d=as.matrix(round(dist(xy),2));d
colnames(d)=base$nombre_mpi
rownames(d)=base$nombre_mpi;d
par(mai=c(0,0.4,0.2,0))
plot(hclust(dist(d)), cex=0.5)

# Mapa coropletico por cuantiles (Buena opcion por heterogeneidad)
ncol=4
ploclr=brewer.pal(ncol,"YlOrRd")
class=classIntervals(base$pupa,ncol, style ="quantile")
colcode=findColours(class,ploclr);colcode
plot(base,col=colcode,border="black", cex=3)
title (main="Mapa usando cuantiles")
text(xy, base$nombre_mpi, cex=0.4, col="black")

# Media y desviacion de pupas en Cauca
mean(na.omit(base$pupa))
sd(na.omit(base$pupa))
```

```
# Matriz binaria efecto reina
bin_rei=poly2nb(base,row.names=base$nombre_mpi, queen=TRUE);bin_rei
queen=nb2mat(bin_rei, style='W'); queen
colnames(queen)=base@data$nombre_mpi;queen

# Matriz binaria efecto torre
bin_torre=poly2nb(base,row.names=base$nombre_mpi, queen=FALSE);bin_torre
tower=nb2mat(bin_torre, style='S'); tower
colnames(tower)=base@data$nombre_mpi;tower

# Mapa de vecindad con efecto reina
plot(base,col="gray", border="blue",lwd=0.2, main="Con efecto reina")
plot(bin_rei,coordinates(cauca), col="red", lwd=0.2, add=T)
text(xy, base$nombre_mpi, cex=0.5, col="black")

# Mapa de vecindad con efecto torre
plot(base,col="gray", border="blue",lwd=0.2, main="Con efecto torre")
plot(bin_torre,coordinates(base), col="red", lwd=0.2, add=T)
text(xy, base$nombre_mpi, cex=0.5, col="black")

# Residuales
## Matrices con efecto reina
rei1bin=poly2nb(base,row.names=base$nombre_mpi, queen=TRUE)
binrei1=nb2listw(rei1bin, style='B') # Matriz binaria
moran.test(lm$residuals, binrei1) # No hay correlacion espacial (-0.136)

wrei1=nb2listw(rei1bin, style="W") # Matriz W
moran.test(lm$residuals, wrei1) # No hay correlacion espacial (-0.185)

srei1=nb2listw(rei1bin, style="S") # Matriz S
moran.test(lm$residuals, srei1) # No hay correlacion espacial (-0.163)

## Matrices con efecto torre
torre1bin <- poly2nb(base,row.names=base$nombre_mpi, queen=FALSE)
bintorre1 <- nb2listw(torre1bin, style="B") # Matriz binaria
moran.test(lm$residuals, bintorre1) # No hay correlacion espacial (-0.136)

wtorre1 = nb2listw(torre1bin, style="W") # Matriz W
moran.test(lm$residuals, wtorre1) # No hay correlacion espacial (-0.183)

storre1 = nb2listw(torre1bin, style="S") # Matriz S
moran.test(lm$residuals, storre1) # No hay correlacion espacial (-0.162)

# Pupas
## Matriz binaria con efecto reina
moran.test(base$pupa, binrei1) # No hay correlacion espacial (-0.039)
geary.test(base$pupa, binrei1) # No hay correlacion espacial (0.872)
```

```
## Matriz W con efecto reina
moran.test(base$pupa, wrei1) # No hay correlacion espacial (-0.036)
geary.test(base$pupa, wrei1) # No hay correlacional espacial (0.880)
## Matriz S con efecto reina
moran.test(base$pupa, srei1) # No hay correlacion espacial (-0.039)
geary.test(base$pupa, srei1) # No hay correlacion espacial (0.888)
## Matriz binaria con efecto torre
moran.test(base$pupa, bintorre1) # No hay correlacion espacial (-0.034)
geary.test(base$pupa, bintorre1) # No hay correlacion espacial (0.872)
## Matriz W con efecto torre
moran.test(base$pupa, wtorre1) # No hay correlacion espacial (-0.032)
geary.test(base$pupa, wtorre1) # No hay correlacion espacial (0.873)
## Matriz S con efecto torre
moran.test(base$pupa, storre1) # No hay correlacion espacial (-0,035)
geary.test(base$pupa, storre1) # No hay correlacion espacial (0.884)

# Grafico de di-dispersion (Residuales)
moran.plot(lm$residuals, wrei1)

# Grafico de di-dispersion (pupa)
moran.plot(base$pupa, binrei1)

# Indice Lisa (pupa)
localmoran(base$pupa, binrei1) # En Timbiqui se rechaza Ho, es decir
# que hay una correlacion espacial local significativa en esos dos municipios,
# por lo tanto, hay una arupacion espacial de valores altos o bajos de pupas.

# Indice Lisa (residuales)
localmoran(lm$residuals, wrei1) # En Timbiqui, Argelia y Guapi existe
# la presencia de agrupaciones locales de valores residuales que no pueden
# ser explicadas por el modelo. En otras palabras, indica que hay patrones
espaciales locales
# de autocorrelación espacial en los residuos del modelo. Esto significa que
# los valores residuales son significativamente más altos o más bajos de lo
# que se esperaría si no hubiera patrones espaciales y es lo que contribuye
# a la variablidad en los residuos.

# Modelo de retardo espacial
modeloRE1 <- lagsarlm(pupa~msnm_med+personas+tmed+prec_med+hum, data=base,
                      listw = binrei1, method = "eigen", tol.solve = 2e-17)
# Modelo saturado
summary(modeloRE1)
BIC(modeloRE1) # BIC
AIC(modeloRE1) # AIC
moran.test(modeloRE1$residuals, wrei1) # No hay correlacion espacial
shapiro.test(modeloRE1$residuals) # No hay normalidad
bptest.Sarlm(modeloRE1) # No hay varianza constante
```

```
# Modelo de error espacial
modeloEE1 <- errorsarlm(pupa~msnm_med+personas+tmed+prec_med+hum, data=base,
                        listw = binrei1, method = "eigen", tol.solve = 2e-17)
# Modelo saturado
summary(modeloEE1)
AIC(modeloEE1) # AIC
BIC(modeloEE1) # BIC
moran.test(modeloEE1$residuals, wrei1) # No hay correlacion espacial
shapiro.test(modeloEE1$residuals) # No hay normalidad
bptest.Sarlm(modeloEE1) # No hay varianza constante

# Modelo Espacial de Durbin
modeloED1 <- lagsarlm(pupa~msnm_med+personas+tmed+prec_med+hum, data=base,
                listw = binrei1, Durbin=T ,method = "eigen", tol.solve = 2e-17)
# Modelo saturado
summary(modeloED1)
AIC(modeloED1)
BIC(modeloED1)
moran.test(modeloED1$residuals, wrei1) # No hay correlacion espacial
shapiro.test(modeloED1$residuals) # No hay normalidad
bptest.Sarlm(modeloED1) # Hay varianza constante

# Modelo poisson para datos agrupados
pois <- glm(pupa~msnm_med+personas+tmed+prec_med+hum+xy, family=poisson(),
            data=base)
summary(pois)

# Modelo binomial negativo para datos agrupados
nb <- glm.nb(pupa~msnm_med+personas+tmed+prec_med+hum+xy, data=base)

# Evaluar sobredispersion y cero inflacion
dispersiontest(pois)
testZeroInflation(pois)

# Modelo SGAM
Primer_GAM <- mgcv::gam(pupa~msnm_med+s(personas)+s(tmed)+s(prec_med)+hum+xy,
family = tw(), data=base)
summary(Primer_GAM)
AIC(Primer_GAM) # AIC
BIC(Primer_GAM) # BIC

# Establecer el diseño de la ventana gráfica
par(mfrow = c(1, 1))

# Gráfico de deviance residuals
plot(abs(residuals(model_GAM)),
```

```
     main = "Residuales Deviance",
     xlab = "Observaciones",
     ylab = "Residuales",
     pch = 20, col = "blue")
abline(h = 2, col = "red")

# Gráfico de Pearson residuals
plot(abs(residuals(model_GAM, type = "pearson")),
     main = "Residuales Pearson",
     xlab = "Observaciones",
     ylab = "Residuales",
     pch = 20, col = "blue")
abline(h = 2, col = "red")

# Volver al diseño de ventana gráfica original
par(mfrow = c(1, 1))

# Cálculo de influencia
infl <- influence.gam(model_GAM)

# Gráfico de valores Cook's D
plot(infl,
     main = "Datos Influyentes",
     xlab = "Observaciones",
     ylab = "Influencia",
     pch = 20, col = "blue")
abline(h = 0.6, col = "green")

# Se realizo el SGAM saturado y arrojo en el analisis de residuales que existen
# ciertos datos atipicos e influyentes (Eliminacion de observaciones)
base <- base[-16,]
base <- base[-7,]
xy <- xy[-16,]
xy <- xy[-7,]

# Modelo SGAM Tweedie final
model_GAM <- mgcv::gam(pupa~msnm_med+personas+tmed+s(prec_med)+hum+xy, family = tw(),
                       data=base)
summary(model_GAM)
AIC(model_GAM) # AIC
BIC(model_GAM) # BIC

# Directorio
setwd("C:/Users/sebas/Desktop/Noveno Semestre U/Tesis/Cauca")
dir()

# Datos de pupas y covariables
```

```
prono1 <- read_excel("Pronosticos.xlsx", sheet = 1)
prono2 <- read_excel("Pronosticos.xlsx", sheet = 2)

# Eliminacion de lugares sin informacion de muestreo
prono1 <- prono1[-1,] # Almaguer
prono1 <- prono1[-14,] # Jambalo
prono1 <- prono1[-15,] # La Vega
prono1 <- prono1[-23,] # Popayan
prono1 <- prono1[-24,] # Purace
prono1 <- prono1[-25,] # San Sebastian
prono1 <- prono1[1:32,] # Silvia, Sotara y Totoro
prono2 <- prono2[-1,] # Almaguer
prono2 <- prono2[-14,] # Jambalo
prono2 <- prono2[-15,] # La Vega
prono2 <- prono2[-23,] # Popayan
prono2 <- prono2[-24,] # Purace
prono2 <- prono2[-25,] # San Sebastian
prono2 <- prono2[1:32,] # Silvia, Sotara y Totoro

# Lectura del shapefile
cauca <- readOGR("shapes.shp", layer = "shapes")

# Eliminacion de lugares sin informacion en shape file
cauca <- cauca[-40,] # San Sebastian
cauca <- cauca[-39,] # La Vega
cauca <- cauca[-38,] # Silvia
cauca <- cauca[-37,] # Isla Gorgona
cauca <- cauca[-34,] # Almaguer
cauca <- cauca[-33,] # Jambalo
cauca <- cauca[-27,] # Purace
cauca <- cauca[-21,] # Sotara
cauca <- cauca[-12,] # Popayan
cauca <- cauca[-11,] # Totara

# Union de las bases de datos
prono1 <- merge(cauca, prono1, by="nombre_mpi")
prono2 <- merge(cauca, prono2, by="nombre_mpi")

# Eliminacion de columnas irrelevantes
prono1 <- prono1[,c("nombre_dpt","mpio","nombre_mpi", "msnm_med", "personas",
"tmed","prec_med", "hum")]
prono1 <- prono1[-16,]
prono1 <- prono1[-7,]
prono2 <- prono2[,c("nombre_dpt","mpio","nombre_mpi", "msnm_med", "personas",
"tmed","prec_med", "hum")]
prono2 <- prono2[-16,]
prono2 <- prono2[-7,]
```

```
# Pronostico 1
predict(model_GAM, prono1, type="response")

# Pronostico 2
predict(model_GAM, prono2, type="response")

# Rango de datos
ran_desag <- max(regresion$pupa)-min(regresion$pupa)
ran_agre <- max(base$pupa)-min(base$pupa)

# Metricas de pronosticos
## Cero Inflado
prono_zi <- predict(mod0, regresion, type="response")
RMSE_zi<- RMSE(prono_zi, regresion$pupa)/ran_desag; RMSE_zi
MSE_zi <- MSE(prono_zi, regresion$pupa)/ran_desag^2; MSE_zi
MAE_zi <- MAE(prono_zi, regresion$pupa)/ran_desag; MAE_zi
## SGAM
newdata <- data.frame(base$msnm_med, base$personas, base$tmed, base$prec_med,
                      base$hum, xy)
colnames(newdata) <- c("msnm_med", "personas","tmed","prec_med", "hum","xy1",
"xy2")
newdata
prono_gam <- predict(model_GAM, newdata, type="response")
RMSE_gam <- RMSE(prono_gam, base$pupa)/ran_agre; RMSE_gam
MSE_gam <- MSE(prono_gam, base$pupa)/ran_agre^2; MSE_gam
MAE_gam <- MAE(prono_gam, base$pupa)/ran_agre; MAE_gam
```