Estimación del tamaño del mercado de la telefonía móvil en Colombia a través de un modelo estadístico

Catalina Londoño Ceballos

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2023

Estimación del tamaño del mercado de la telefonía móvil en Colombia a través de un modelo estadístico

Catalina Londoño Ceballos

Trabajo final de maestría como requisito parcial para optar al título de:
Magister en Ciencias - Estadística

Director(a):
Juan Carlos Salazar Uribe Ph.D.
Profesor Titular, Escuela de Estadística
Universidad Nacional de Colombia

Líneas de Investigación:
Analítica, Análisis Multivariado de Datos

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2023

Dedicatoria

A todos los que me acompañaron y me apoyaron
en esta etapa de formación y crecimiento

Agradecimientos

Gracias al doctor Juan Carlos Salazar Uribe, profesor titular de la escuela de estadística de la Universidad Nacional de Colombia, director de esta tesis de profundización, por el aporte fundamental de sus conocimientos y su gran disposición para que esta fuera finalizada con éxito. Adicionalmente a Sandra Moreno magíster en estadística de la Universidad Nacional, quien me introdujo en el mundo de la estadística y sus palabras de apoyo fueron fundamentales para transitar esta etapa de mi vida.

Resumen

En este trabajo se presenta el ajuste de un modelo estadístico para pronosticar el tamaño del mercado móvil en Colombia (medido en cantidad de líneas) a partir de la utilización de los datos simulados en una de las compañías representativas del sector, con la finalidad de poder tener información oportuna para la toma de decisiones comerciales y tácticas que utilizan dicha información. Como resultado se logró obtener un modelo con márgenes mínimos de error mejorando respecto del modelo lineal normal de referencia, se obtuvo un error medio porcentual absoluto de 0.53 %.

Palabras clave: Modelos lineales generalizados, estimación de ecuaciones generalizadas, modelos estadísticos, pronóstico, telefonía móvil.

Estimation of the mobile telecommunication market size in Colombia through a statistical model

This paper presents the adjustment of a statistical model to forecast the size of the mobile market in Colombia (measured in number of lines) from the use of simulated data in one of the representative companies of the sector, with the purpose of being able to have timely information for making business decisions and tactics that use such information. As a result, it was possible to obtain a model with minimum margins of error, improving with respect to the reference normal linear model, an average absolute percentage error of 0.53 % was obtained.

Keywords: Generalized linear models, Generalized estimating equation, statistical models, forecasting, mobile telephony

Lista de Figuras

Lista de Tablas

1. Introducción

En Colombia el tamaño del mercado de la telefonía móvil se mide a través de la suma de la cantidad de abonados en cada operador prestador de este servicio, información que se disponibiliza en forma trimestral por parte del Ministerio de comunicaciones (Mintic). Es así como en la página del Ministerio se encuentran disponibles los datos sobre la cuota de mercado de líneas móviles por proveedor y por categoría (sí son de tipo prepago o pospago), información con la cual se puede calcular el índice de penetración de abonados en el servicio de telefonía móvil. Esta información es utilizada por las compañías de telecomunicaciones para definir tácticas comerciales, buscar oportunidades de negocio, saber cuál es su posicionamiento en el mercado, entre otros aspectos fundamentales en la definición de estrategia comercial. Uno de los grandes desafíos a los que se enfrentan al utilizar esta información es que Mintic la disponibiliza con un retraso de 5 meses.

Por lo anterior es de vital importancia poder contar con una estimación estable y precisa que permita tener la información más actualizada y así tomar decisiones basadas en datos confiables. Una de las características de esta industria consiste en que puede verse afectada por diversos factores como la entrada o salida de competidores, la desactivación temprana de las líneas, entre otros, por esta razón se hace necesaria la construcción de un modelo que pueda capturar estos efectos.

Utilizando la información y las herramientas disponibles en una de las compañías representativas del sector, en este trabajo se abordó esta necesidad buscando construir un modelo estadístico cuya variable respuesta es la cantidad total de abonados trimestral, en función de diversas variables explicativas provenientes de la información simulada de dicha compañía, como lo son: cantidad total de usuarios, cantidad total de llamadas salientes de los usuarios de este operador, cantidad de llamadas entrantes de otros operadores, entre otras, que permitirán incorporar de una manera más directa los movimientos de mercado.

Al realizar una búsqueda en la literatura, se encontró que los modelos más utilizados para abordar problemas similares correspondían a los modelos Bass, Gompertz y de crecimiento logístico, los cuales son modelos de crecimiento. Dichos modelos se caracterizan por ser univariados, permitiendo solo la inclusión de la variable respuesta y una serie de parámetros, como la tasa de crecimiento, el valor de la saturación y una serie de constantes particulares para cada modelo.

Autores como Chu et al. (2009), Gamboa and Otero (2009), Wu and Chu (2010), Herrera Giraldo (2012), Annafari (2013), Jha and Saha (2020), realizaron aplicaciones donde utilizaron algunos de los tres o los tres modelos anteriormente mencionados, para abordar problemas como la estimación de la penetración o difusión de la telefonía móvil en distintos países. Dado que en este tipo de modelos no es permitida la inclusión de covariables, se procedió a buscar en la literatura modelos más flexibles y que permitieran la inclusión de variables explicativas.

Dentro de esta búsqueda se encontraron los modelos lineales generalizados. De acuerdo con Agresti (2015), los modelos lineales generalizados extienden los modelos estándar de regresión lineal para abordar distribuciones de respuesta no normales y posibles funciones no lineales de la media. Es así como estos modelos permiten modelar variables discretas, como los conteos, y la implementación de splines.

De acuerdo con Li et al. (2013), los modelos lineales generalizados se han vuelto una de las herramientas favoritas para modelar datos longitudinales, particularmente para datos no normales repetidos o correlacionados, como aquellos en que la variable respuesta es de tipo binomial o Poisson, que es comúnmente encontrada en estudios longitudinales. Sin embargo, dado que los datos al estar correlacionados requieren una estructura distinta, Liang and Zeger (1986), introdujeron el método GEE (generalized estimating equations), que es una extensión muy útil de los modelos lineales generalizados a datos correlacionados; este método se ha convertido en un método de estimación muy popular.

Dado que los datos abordados en este trabajo corresponden a datos correlacionados, la metodología GEE fue la elegida para ajustar los modelos presentados, ya que esta permite la modelación de la estructura de correlación presente en los datos a través de la especificación de una matriz de correlación, conocida como la matriz de correlación de "trabajo". De acuerdo con Liang and Zeger (1986) el método GEE permite realizar estimaciones consistentes de los parámetros de regresión y de su varianza bajo suposiciones leves sobre la dependencia del tiempo. Estos supuestos son:

- Las observaciones repetidas de un sujeto son independientes entre sí.
- El promedio ponderado de las matrices de correlación estimadas converge a una matriz fija
- La estructura de covarianza a lo largo del tiempo es tratada como ruido.
- Se asume solo una forma funcional para la distribución marginal en cada momento.

Estas ecuaciones se derivan sin especificar la distribución conjunta de las observaciones de un sujeto, pero se reducen a las ecuaciones score para resultados gaussianos multivariados.

Como se mencionó anteriormente, en este método se debe de especificar el tipo de matriz de correlación de trabajo que será utilizada en la estimación de parámetros, teniendo la ventaja de ser robusto ante la especificación errónea de dicha matriz (Hardin and Hilbe, 2013).

En este trabajo, la primera parte corresponde a todo el proceso de generación, exploración y adaptación de las variables disponibles en la base de datos, en la segunda parte se abordan aspectos teóricos de los modelos estadísticos que fueron explorados e implementados, en la parte final se muestran los resultados de los modelos aplicados a datos reales, comparando su desempeño en términos de algunas de las métricas más utilizadas en la práctica.

La base de datos utilizada se obtuvo a partir de la simulación de las variables de una compañía de la industria de telecomunicaciones. De esta manera se obtuvieron datos simulados mensuales desde enero de 2018 hasta marzo de 2022. Las métricas elegidas para comparar los modelos fueron el ICC (intraclass correlation coefficient), el MAE, el QIC entre otras. Para calcular estas métricas se dividió la base de datos simulada en entrenamiento y prueba, dejando los últimos 6 meses para validación debido a que este es el horizonte habitual de pronóstico en la entidad propietaria de los datos y el tiempo de retraso para obtener la información de Mintic.

El modelo que presentó el mejor desempeño fue el modelo de regresión Poisson, superando al modelo normal en todas las medidas consideradas. Frente a los otros modelos su ICC y su QIC estuvieron muy cerca, siendo el modelo elegido para la estimación.

2. Creación y adecuación de las variables

Para la realización de este trabajo final de maestría, se inició con la construcción de la información de la variable respuesta y las covariables. Posteriormente se procedió a la adecuación de la variable respuesta y para el caso de las covariables se procedió con la simulación de éstas, esto con el fin de proteger y respetar la privacidad de la información de la entidad propietaria de los datos. Para la construcción de las covariables se hizo uso de las herramientas de big data disponibles en la entidad propietaria de los datos, concretamente de Oracle y SQL developer.

2.1. Adecuación de la variable respuesta

Como ya se mencionó anteriormente, el tamaño del mercado de la telefonía móvil en Colombia se mide a través de la suma del conteo de la cantidad de líneas en cada operador prestador de este servicio, esta es la variable respuesta que se buscó estimar en este trabajo. Esta información de la variable respuesta se encontraba en una periodicidad trimestral, mientras que la información de las covariables se encontraba en una periodicidad mensual. Debido a que la disponibilidad de las covariables es mensual, y los seguimientos y la planeación de la estrategia comercial se realizan en este mismo período de tiempo, se procedió a desagregar la variable respuesta en una periodicidad mensual. Este procedimiento es conocido como la desagregación temporal de una serie, de acuerdo con Islama (2014), este consiste en la estimación de series de alta frecuencia desagregadas, por ejemplo, trimestrales, mensuales, a partir de series disponibles de baja frecuencia agregadas, por ejemplo, anuales, trimestrales.

De acuerdo con Islama (2014), cuando solo las series que se encuentran en una baja frecuencia están disponibles, es decir no se tiene una serie indicadora o variable proxy[1] , se puede elegir entre métodos matemáticos, métodos de suavizado numérico y modelos ARIMA de series temporales. Los métodos de suavizado comúnmente utilizados son los desarrollados por Boot, Feibes y Lisman (1970), y por Denton (1971). Este mismo autor desarrolla una evaluación del método Denton y sus variantes, concluyendo que los métodos Denton propor-

[1]De acuerdo con Upton and Cook (2014), una variable proxy es una variable cuantificable que se utiliza en lugar de una variable que no se puede medir.

cional y Denton-Cholette son mucho más cercanos a la serie desagregada real, tanto en el caso en el que se cuenta con una serie indicadora disponible, como el caso en el que no.

De acuerdo con Dagum and Cholette (2006) el método Denton es un método clásico para la desagregración e interpolación de series de tiempo. Particularmente, los métodos basados en la regresión de Cholette-Dagum (también conocida como Denton-Cholette), corresponden a uno de los métodos más ampliamente aplicados en los organismos de estadística. Denton desarrolló métodos de evaluación comparativa basados en el principio de preservación del movimiento que todavía se aplican ampliamente. De acuerdo con este principio, la serie de referencia debe reproducir el movimiento de la serie original. Denton propuso una serie de definiciones de preservación del movimiento, cada una correspondiente a una variante de su método. Para más detalle de cada una de estas variantes ver Dagum and Cholette (2006).

De acuerdo con Dagum and Cholette (2006), el método original de Denton presenta dos fallas importantes:

- Introduce un movimiento transitorio al comienzo de la serie, lo que anula el principio establecido de preservación del movimiento.
- Implica un pronóstico implícito de la siguiente discrepancia al final de la serie, con base en las dos últimas discrepancias únicamente.

En este caso, el interés se centraba en pasar de una serie de baja frecuencia (trimestral) a una serie de alta frecuencia (mensual), para esto se utilizó la variante del método Denton correspondiente al método Denton-Cholette (Cholette-Dagum), ya que es uno de los métodos más ampliamente utilizados, se ha probado su efectividad y se encuentra disponible en el paquete Tempdisagg (Sax and Steiner, 2013) en el lenguaje de programación R (R Core Team, 2021). De acuerdo con Sax and Steiner (2013), el método Denton-Cholette permite desagregar las series de alta frecuencia sin necesidad de una serie indicadora de baja frecuencia, adicionalmente soluciona uno de los inconvenientes del enfoque original, eliminando el falso movimiento transitorio al comienzo de la serie resultante.

La serie mensual de Mintic se encuentra disponible desde el cuarto trimestre del 2015, de esta manera se pasó de disponer de 26 datos con periodicidad trimestral a 76 datos con una periodicidad mensual. En la figura 2-1 se presentan ambas series, donde la línea azul representa la serie original y la línea negra representa la serie desagregada mensual.

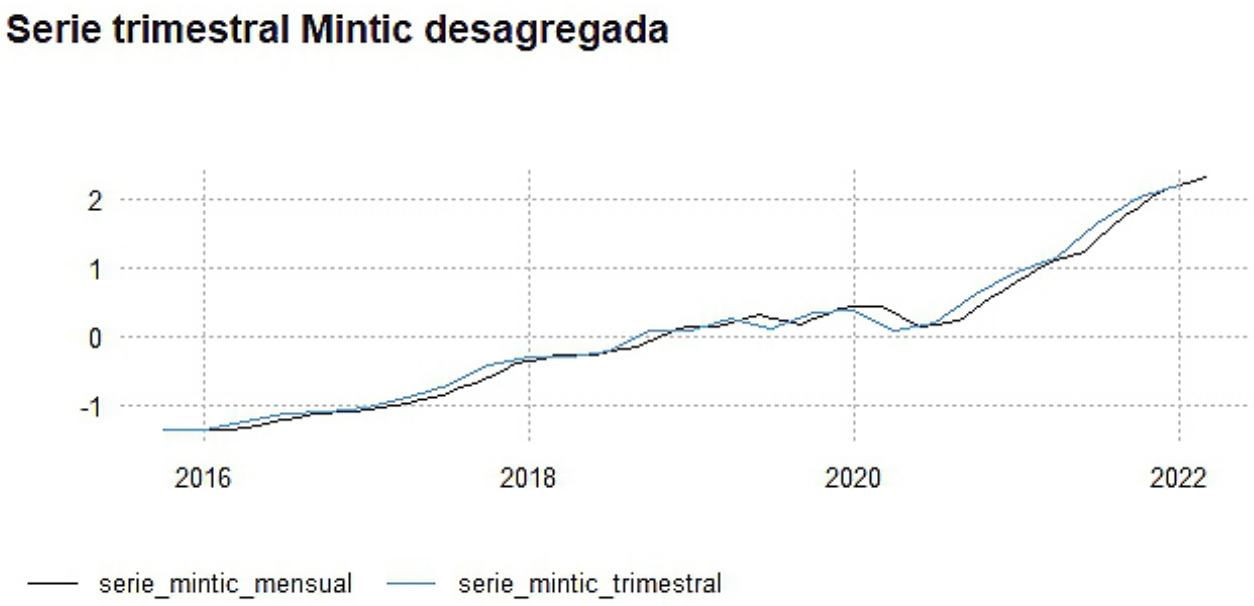

Figura 2-1.: Desagregación cantidad de líneas trimestrales a una periodicidad mensual.
Fuente: Elaboración propia.

2.2. Simulación de las covariables

Como se mencionó anteriormente, con el fin de proteger la privacidad de los datos de la compañía propietaria de estos, se procedió a simular las variables que contenían la información que se utilizó para conformar la base de datos de trabajo con la que se ajustó el modelo final.

Para simular dichas variables se usaron modelos lineales generalizados mixtos con intercepto aleatorios (modelos Poisson y gaussianos), tal y como se describen en Fitzmaurice et al. (2012)

En la Tabla 2-1 se presenta la descripción de cada una de estas covariables.

Tabla 2-1.: Descripción de las covariables

Variable	Descripción	Tipo de variable
Usuarios Pos	Cantidad de líneas pospago	Discreta
Usuarios Pre	Cantidad de líneas prepago	Discreta
Llamadas entrantes	Cantidad llamadas entrantes de otros operadores a los usuarios de la entidad propietaria de los datos	Discreta
Llamadas salientes	Cantidad llamadas salientes de los usuarios de la entidad propietaria de los datos a otros operadores	Discreta
Minutos entrantes	Duración llamadas entrantes de otros operadores a los usuarios de la entidad propietaria de los datos	Continua
Minutos salientes	Duración llamadas salientes de los usuarios de la entidad propietaria de los datos a otros operadores	Continua

El gráfico 2-2 presenta los gráficos de las variables originales en color azul y las variables simuladas en color violeta, como se puede observar los modelos utilizados capturaron adecuadamente el comportamiento de las series.

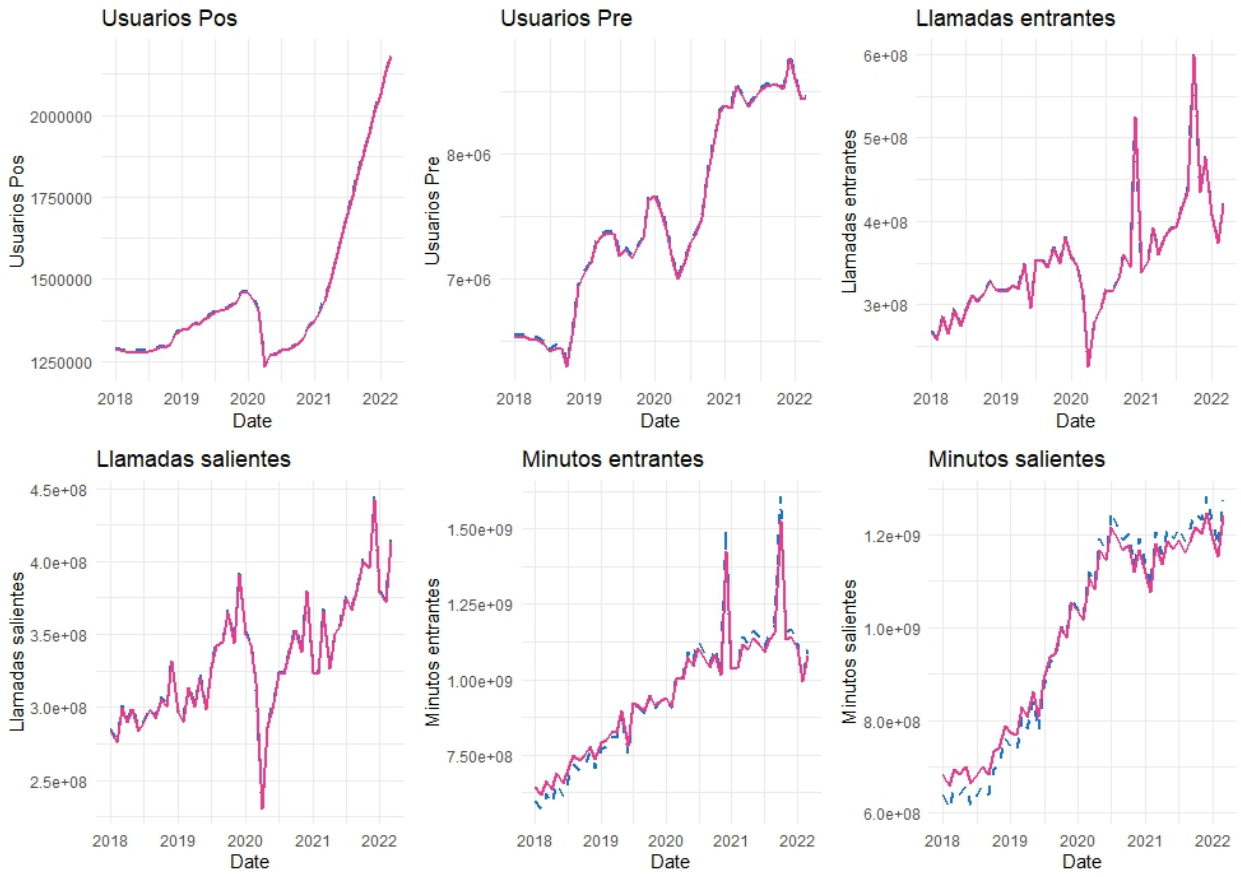

Figura 2-2.: Series de las covariables simuladas.
Fuente: Elaboración propia.

Adicionalmente para complementar el análisis gráfico, se calculó el coeficiente de correlación de Spearman, de acuerdo con Forthofer et al. (2007) con variables que no se distribuyen normalmente es mejor utilizar este coeficiente, que adicionalmente puede ser usado con datos ordinales. Entre más alto sea el valor absoluto del coeficiente de correlación de Spearman, más fuerte será la asociación entre las dos variables. Los valores positivos sugieren que los valores más altos de una variable se asocian con valores más altos de la otra variable, mientras que los valores negativos sugieren que los valores más altos de una se asocian con valores más bajos de la otra (Puth et al., 2015). Finalmente para complementar, se utilizó el coeficiente de correlación intra clase (ICC) para medir el grado de concordancia entre los valores simulados y los valores reales. El cálculo de este índice se explica con detalle en la Sección 5.6 en los criterios utilizados para la selección de modelos.

Tabla 2-2.: Comparación correlación variable respuesta con variables originales y simuladas

Variable	Corr. Spearman variables originales	valor p	Corr. Spearman variables simuladas	valor p
Usuarios Pospago	0.843	0	0.843	0
Usuarios Prepago	0.945	0	0.945	0
Minutos entrantes de otros operadores	0.845	0	0.845	0
Minutos salientes a otros operadores	0.827	0	0.827	0
Llamadas entrantes de otros operadores	0.85	0	0.85	0
Llamadas salientes a otros operadores	0.819	0	0.819	0

Tabla 2-3.: Comparación correlación variables originales con las variables simuladas

Variable	Corr Spearman	Valor p	ICC
Usuarios Pospago	1.00	0.00	1.00
Usuarios Prepago	1.00	0.00	1.00
Minutos entrantes de otros operadores	1.00	0.00	0.99
Minutos salientes a otros operadores	1.00	0.00	0.99
Llamadas entrantes de otros operadores	1.00	0.00	1.00
Llamadas salientes a otros operadores	1.00	0.00	1.00

Como se puede observar las variables simuladas conservan el valor de la correlación entre la variable respuesta y las covariables, adicionalmente al correlacionar las variables simuladas con las variables originales se puede observar que el grado de asociación es perfecto, y el ICC también es aproximadamente igual a 1 para todas las variables, por tanto las variables simuladas logran reflejar las variables originales. Una vez completada la base de datos con las nuevas variables simuladas y con la serie trimestral desagregada en periodos mensuales, se procedió a realizar el análisis exploratorio de las series. Dado que la información de las covariables se encontraba desde el 2018, se dispuso finalmente de una base de datos compuesta por 7 variables y 51 observaciones.

3. Análisis exploratorio de datos

El análisis exploratorio de datos, según Pearson (2018), se puede definir como el arte de observar un conjunto de datos, realizando un esfuerzo por comprender la estructura subyacente que hay en estos. Es así como éste es uno de los procedimientos para el entendimiento de un conjunto de datos presentado, siendo el primer paso para observar si un problema específico puede tener una solución o explicación a través de la utilización de dichos datos. Este debe ser realizado previo a la implementación de modelos estadísticos, ya que es en este paso donde se puede evidenciar cuál es la naturaleza de los datos que se están analizando, realizar una delimitación adecuada del problema y si los datos que se tienen disponibles pueden ser utilizados para abordar dicho problema.

En el análisis exploratorio de datos se hace uso de las herramientas brindadas por la estadística descriptiva, donde a través de la generación de visualizaciones de los datos se busca entenderlos gráficamente. De acuerdo con Iliinsky and Steele (2011) hay diferentes tipos de visualización, las que se presentarán a continuación se pueden definir como visualizaciones exploratorias de los datos. De acuerdo con los autores estas son apropiadas cuando se necesita tener una idea de lo que hay dentro de un conjunto de datos, ya que al llevarlo a un medio visual puede ser útil para identificar rápidamente características de éste, como curvas interesantes, líneas, tendencias o valores atípicos o anómalos.

Como se mencionó en el capítulo anterior, la base de datos de este trabajo está compuesta por 7 variables y 51 observaciones. Estas 7 variables corresponden a:

- El tamaño del mercado móvil total de Colombia, medido en cantidad de líneas.
- La cantidad de líneas en el servicio pospago de una entidad representativa del sector.
- La cantidad de líneas en el servicio prepago de una entidad representativa del sector.
- La cantidad de minutos entrantes en llamadas de otros operadores a los usuarios de la entidad en mención.
- La cantidad de minutos salientes en llamadas a otros operadores de los usuarios de la entidad en mención.
- La cantidad de llamadas entrantes de otros operadores a los usuarios de la entidad en mención.

- La cantidad de llamadas salientes a otros operadores de los usuarios de la entidad en mención.

Las anteriores variables fueron renombradas con el fin de facilitar la elaboración de los gráficos y la distinción entre cuáles son las covariables y cuál es la variable respuesta, se renombraron de la siguiente manera:

- Y: Tamaño del mercado móvil total de Colombia.
- X_1: La cantidad de usuarios pospago.
- X_2: La cantidad de usuarios prepago.
- X_3: La cantidad de minutos entrantes de otros operadores.
- X_4: La cantidad de minutos salientes a otros operadores.
- X_5: La cantidad de llamadas entrantes de otros operadores.
- X_6: La cantidad de llamadas salientes a otros operadores.

3.1. Descripción de la variable respuesta

Para entender y visualizar la variable respuesta Y (tamaño del mercado móvil en Colombia), se realizó un gráfico de la serie de tiempo para ver la evolución de la cantidad de líneas a través de los años, adicionalmente se graficaron los valores de la serie por cada año contra los meses para detectar la posible presencia de estacionalidad.

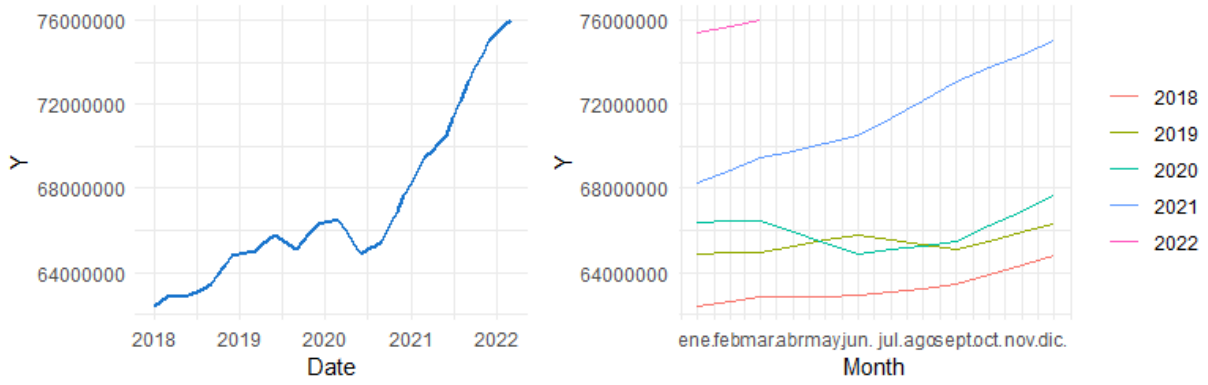

Figura 3-1.: Descripción de la variable respuesta.
Fuente: Elaboración propia.

En el gráfico de líneas se puede observar que el mercado móvil viene presentando una tendencia creciente en el tiempo, con caídas eventuales que corresponden a la dinámica de este mercado. Adicionalmente, en el gráfico de estacionalidad que se presenta en la parte superior derecha de la Figura 3-1 no se detecta que haya un patrón de comportamiento periódico que se repita de enero a diciembre a través de todos los años.

3.2. Descripción de las covariables

Al igual que para la variable respuesta, para las covariables se construyeron gráficos de las series de tiempo de cada una de las variables y se graficaron los valores de la serie por cada año contra los meses para detectar la posible presencia de estacionalidad, y adicionalmente, se presentan los correlogramas para aquellas series que parecen presentar comportamientos estacionales.

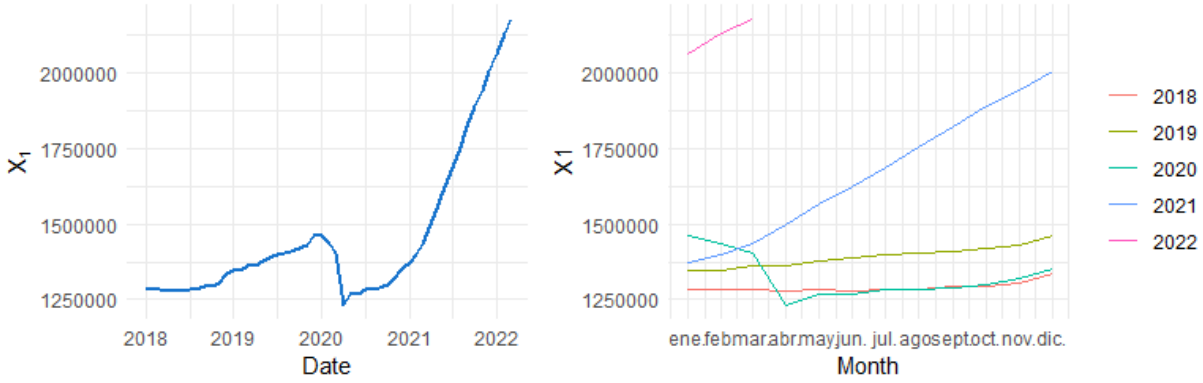

Figura 3-2.: Descripción de la variable X_1.
Fuente: Elaboración propia.

En la Figura 3-2 se puede observar que el comportamiento de la cantidad de líneas pospago presenta un crecimiento sostenido a través de los años, donde en el año 2020 se puede observar una caída fuerte que se corresponde con la época de cuarentena donde por Covid 19 muchos usuarios prescindieron de sus planes de telefonía móvil lo cual implica directamente una disminución de la cantidad de líneas pospago, esto a su vez generó un impacto en algunas de las variables que se presentarán a continuación. Adicionalmente se puede observar en el segundo gráfico que no existe una estacionalidad definida.

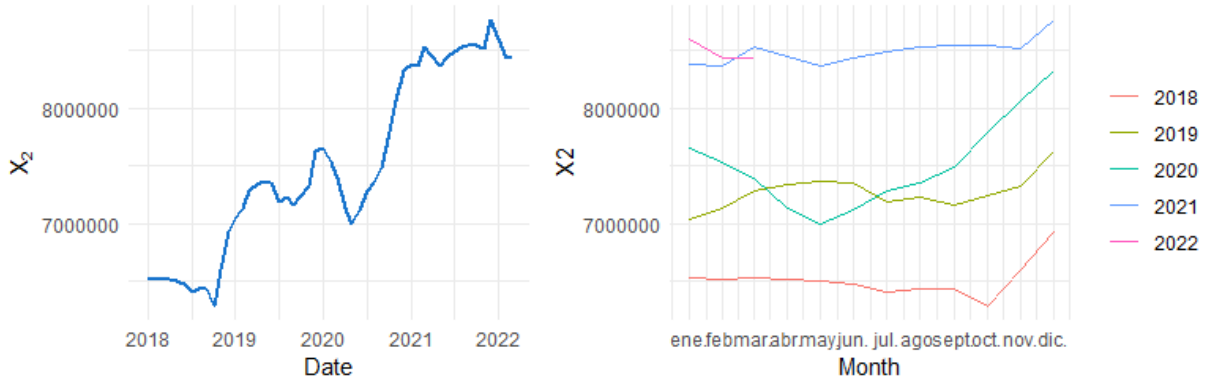

Figura 3-3.: Descripción de la variable X_2.
Fuente: Elaboración propia.

En la Figura 3-3 se puede observar que el comportamiento de la cantidad de líneas de prepago presenta una tendencia creciente a través de los años, sin embargo a diferencia de las líneas pospago se puede observar que cada cierto tiempo se presentan disminuciones en la cantidad de líneas, cabe destacar del segundo gráfico que dichas disminuciones no corresponden a un fenómeno estacional definido y que se corresponden con la dinámica del mercado, donde las líneas prepago suelen ser desechadas por una parte importante de los usuarios que las adquieren, esto se puede deber a que este tipo de líneas se suelen obsequiar con los teléfonos móviles nuevos, también para promocionar las entidades prestadoras de servicio, entre otras, lo que hace que estas líneas no sean necesariamente las principales de los usuarios y se desechan o se dejan de utilizar por diferentes motivos.

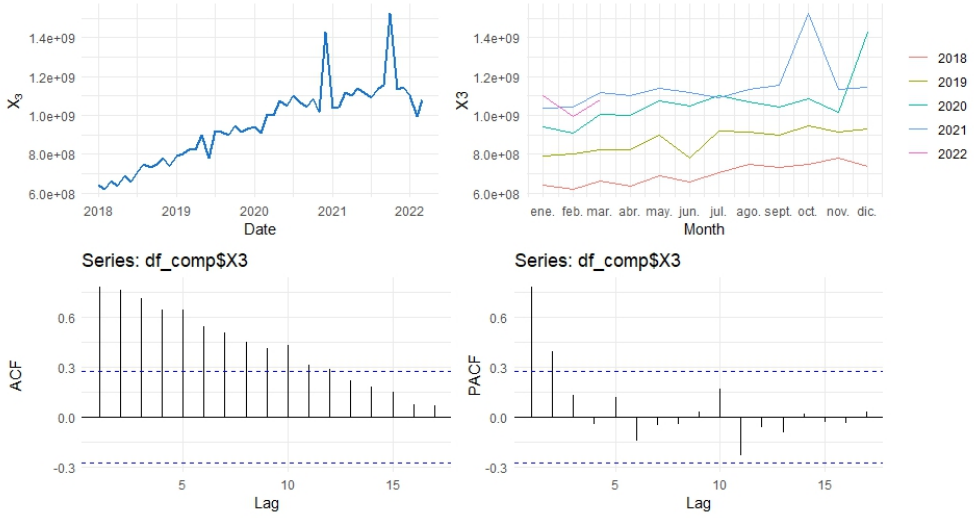

Figura 3-4.: Descripción de la variable X_3.
Fuente: Elaboración propia.

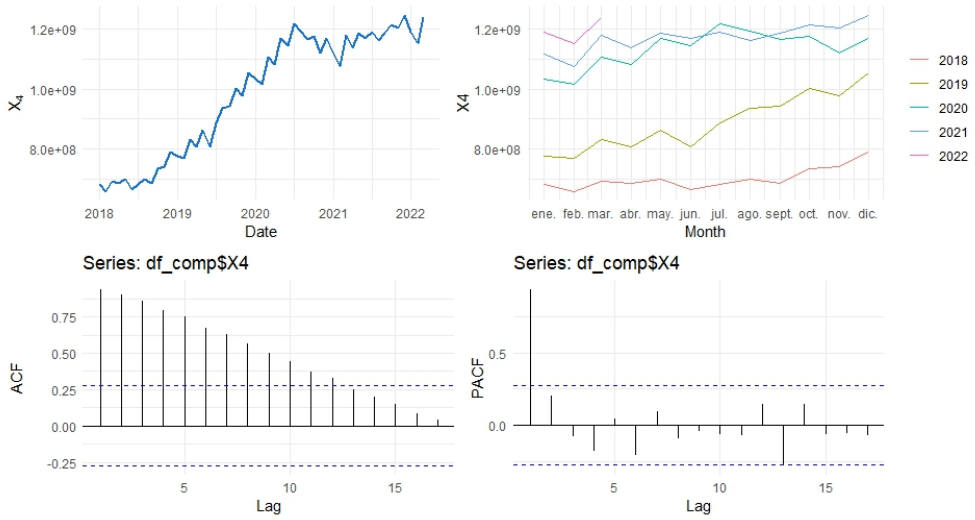

Figura 3-5.: Descripción de la variable X_4.
Fuente: Elaboración propia.

En el caso de las covariables X3 y X4, presentadas en las Figuras 3-4 y 3-5 respectivamente, las cuales corresponden a la cantidad total de minutos entrantes y salientes en llamadas a otros operadores, se puede observar que estas variables tienden a presentar aumentos en meses como octubre y mayo, que se corresponden con fechas como la celebración del día de la madre donde se suele presentar un mayor tráfico de llamadas debido a las celebraciones. Adicionalmente estas series también presentan un comportamiento creciente debido a que están directamente relacionadas con la cantidad de usuarios.

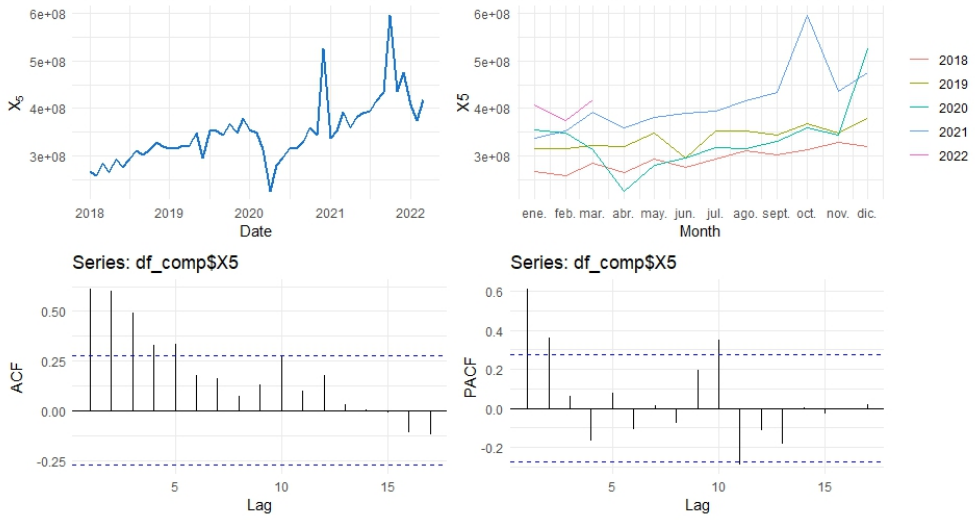

Figura 3-6.: Descripción de la variable X_5.
Fuente: Elaboración propia.

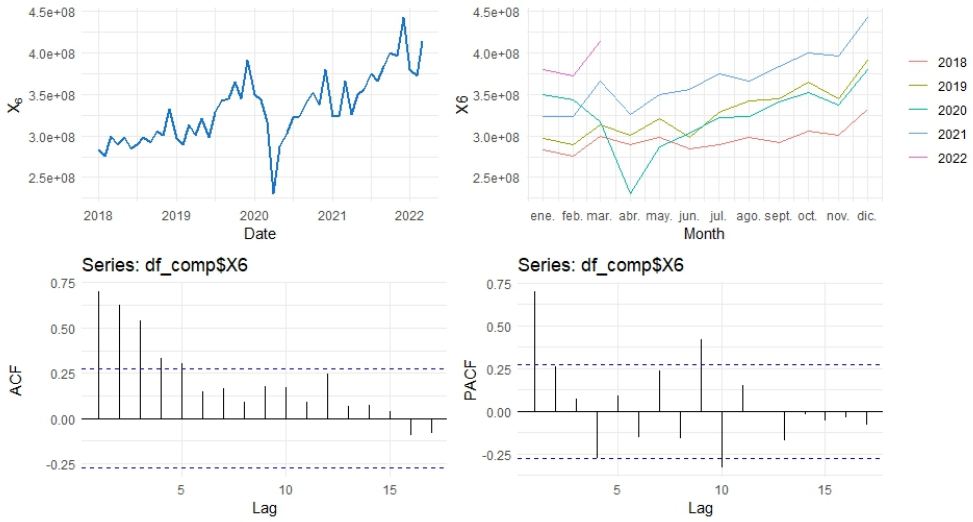

Figura 3-7.: Descripción de la variable X_6.
Fuente: Elaboración propia.

Finalmente en las Figuras 3-6 y 3-7 se presentan respectivamente la cantidad de llamadas entrantes de otros operadores y la cantidad de llamadas salientes a otros operadores. Para estas series también se puede observar una tendencia creciente en el tiempo, al observar los correlogramas, se puede evidenciar por el comportamiento de acf que existe un patrón estacional en los datos.

3.3. Relación entre la variable respuesta y las covariables

En el gráfico de la Figura 3-8 se presenta un gráfico que contiene los gráficos de dispersión entre todas las variables y el coeficiente de correlación de Spearman calculado. Adicionalmente en la Figura 3-9 se presenta una ampliación de los gráficos de dispersión de la variable respuesta contra cada una de las covariables.

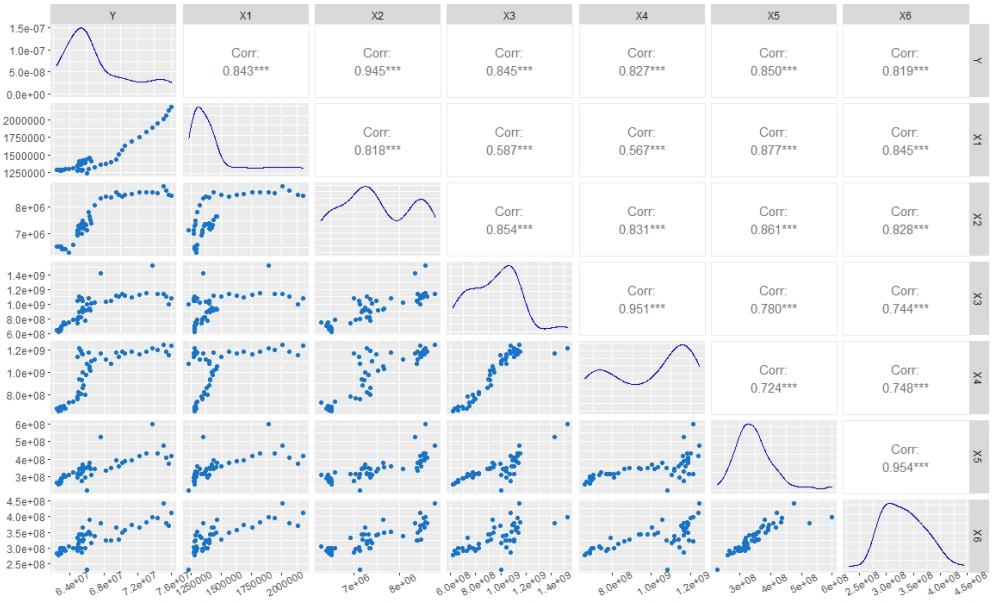

Figura 3-8.: Gráficos de dispersión y correlación entre la variables estudiadas.
Fuente: Elaboración propia.

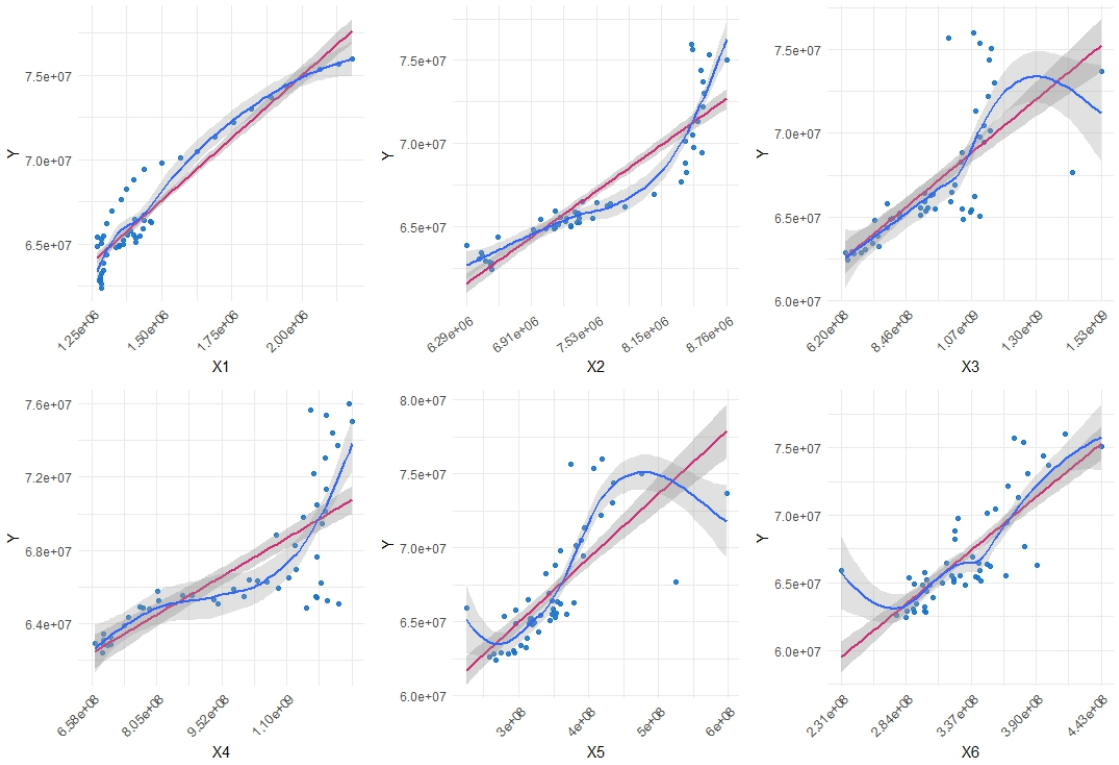

Figura 3-9.: Gráficos de dispersión de la variable respuesta contra todas las variables explicativas.
Fuente: Elaboración propia.

En la Figura 3-9, los puntos azules corresponden al diagrama de dispersión entre las variable Y y cada una de las covariables, la línea magenta corresponde a una recta ajustada por regresión lineal y la línea azul corresponde a la curva ajustada con una regresión loess, que es el método de suavizado por defecto de ggplot2, las bandas grises se corresponden a las bandas de confianza de las estimaciones de cada uno de los métodos. Como se puede observar en los gráficos 3-8 y 3-9, todas las covariables presentan relaciones no lineales con la variable respuesta, esto fue tenido en cuenta a la hora de plantear los modelos.

Adicionalmente para complementar el coeficiente de correlación de Spearman, se procedió a buscar un coeficiente de correlación más robusto dada la no estacionariedad de las variables presentadas, en dicha búsqueda se encontró el coeficiente de correlación DCCA, este fue propuesto por Zebende (2011), de acuerdo con el autor hay variables que son observadas en el mismo intervalo de tiempo simultáneamente y tienen la misma longitud, donde la medida para medir correlación más popular es el coeficiente de correlación de Pearson, sin embargo este no es robusto y puede no ser tan acertado cuando hay datos atípicos presentes y altos grados de no estacionariedad. El autor muestra que el coeficiente de correlación DCCA es robusto y funciona para identificar componentes estacionales, en ambos tipos de correlación

cruzada positiva y negativa. El coeficiente de correlación DCCA se define como:

$$\sigma_{DCCA} \equiv \frac{F^2_{DCCA}}{F_{DFA}\{y_i\}} F_{DFA}\{y'_i\} \tag{3-1}$$

donde,

- F^2_{DCCA}: La función de covarianza sin tendencia.
- F_{DFA}: La función de varianza sin tendencia.

Para más detalle ver Zebende (2011).

Autores como Kristoufek (2014) y Shen (2015), realizaron aplicaciones en series de tiempo reales y simuladas, donde concluyen que el coeficiente de correlación DCCA es adecuado y robusto para medir correlación entre series de tiempo no estacionarias. Adicionalmente, Shen (2015) realiza un ejercicio donde rezaga una serie de variables para ver cómo los valores pasados de una serie pueden afectar a otra, calculando el coeficiente DCCA para cada rezago realizado. Este ejercicio se replicó en este trabajo.

Para el cálculo del coeficiente de correlación DCCA se utilizó la función rhodcca disponible en el software de programación R en el paquete DCCA (Prass and Pumi, 2020). El cálculo se realizó en 2 momentos, el primero para las variables en el mismo período de tiempo, es decir Y_i en función de X_t, este el que corresponde a la primera columna de la tabla 3-1, luego rezagando las covariables hasta tres meses, es decir Y_i en función de X_{t-j} para j desde 1 hasta 3, se rezagan solo las covariables ya que el interés se centra en conocer el valor futuro de la variable respuesta en función de los valores pasados de éstas, ya que la variable respuesta siempre se encuentra disponible después que las covariables.

Tabla 3-1.: Rezagando las covariables

Variable	Corr. valores contemporáneos	Corr. rezago 1	Corr. rezago 2	Corr. rezago 3
X1	0.75	0.69	0.67	0.62
X2	0.52	0.37	0.17	0.03
X3	0.03	0.04	0.01	-0.06
X4	0.05	0.01	0.09	0.16
X5	0.14	0.04	-0.02	-0.11
X6	0.17	-0.03	-0.03	-0.05

Como se puede observar en la Tabla 3-1, la única variable que parece aumentar su grado de asociación con el ejercicio de rezago es la variable X_4, sin embargo es a su vez la que presenta uno de los menores coficientes de asociación. Por tanto la correlación se observa que es mayor en el mismo período para cinco de las seis variables consideradas. De esta manera por el principio de parsimonia y dado que el grado de correlación es más fuerte cuando se considera el mismo período de tiempo, se utilizarán en los modelos estadísticos las covariables en el mismo período de tiempo.

Con el análisis exploratorio realizado anteriormente, se procedió a buscar cuál sería el modelo estadístico que mejor podría ajustarse para abordar el problema de la estimación del tamaño de mercado móvil en Colombia, en el siguiente capítulo se expone el marco teórico de los modelos encontrados en la literatura para abordar este tipo de problemas.

4. Modelos estadísticos

De acuerdo con Cameron and Trivedi (1999), en muchos casos se busca predecir una variable de interés, que es un número entero o recuento no negativo, en términos de un conjunto de variables explicativas. En este caso la variable respuesta es discreta con una distribución que coloca la masa de probabilidad solo en valores enteros no negativos, a diferencia del modelo de regresión clásico. Los modelos de regresión para conteos, como otros modelos de variables dependientes limitadas o discretas, son no lineales con propiedades y características especiales relacionadas con la discreción y la no linealidad de la variable.

La variable que se pronosticó en este trabajo corresponde al caso de una variable de tipo conteo. A lo largo de la literatura se encontró que uno de los modelos más utilizados para abordar este tipo de variables corresponde al modelo de regresión Poisson, donde la variable respuesta es discreta y se utiliza para el análisis de registros de la cantidad de veces que ocurre un evento en un determinado tiempo o espacio. Cuando se presenta un problema de sobredispersión en el modelo Poisson, una alternativa válida es la regresión binomial negativa, donde la variable respuesta también consiste en conteos (Agresti, 2015).

Otra opción apropiada encontrada en la literatura corresponde a los modelos de regresión spline, dado el comportamiento presentado por la serie correspondiente a la cantidad de suscriptores de la telefonía móvil. De acuerdo con Fitzmaurice et al. (2011) los modelos de spline lineales proporcionan una forma muy útil y flexible de adaptarse a las tendencias no lineales que no pueden aproximarse mediante polinomios simples en el tiempo.

Finalmente, la variable respuesta correspondía a una serie de tiempo, por lo tanto sus observaciones no eran independientes y se necesitaba un método que pudiera incorporar este hecho. En la búsqueda de la literatura se encontró el método de estimación de ecuaciones generalizadas (GEE, por sus siglas en inglés), que se describirá más adelante, con aplicaciones y amplios desarrollos para datos de tipo longitudinal. Frees et al. (2004), define el análisis de datos longitudinales como una combinación entre regresión y series de tiempo, donde los datos longitudinales están compuestos de la observación de variables de varios sujetos (n) a través del tiempo, es así como los datos longitudinales se pueden ver como una generalización de las series de tiempo donde n>1, y por tanto una serie de tiempo puede a su vez considerarse un dato longitudinal donde n= 1. Un ejemplo de lo anterior se puede ver en Ruppert et al. (2003), en uno de los ejemplos desarrollados en su libro utiliza una serie de

tiempo para chequear autocorrelación, específicamente utilizó una serie de tiempo de datos correspondientes al uso de electricidad, los autores literalmente dicen en su texto, "los datos de uso de electricidad son longitudinales. Es decir, los datos fueron recopilados a lo largo del tiempo para una sola residencia."(p. 27).

Dado lo anterior, las series de tiempo comparten características fundamentales con los datos longitudinales, de acuerdo con Fitzmaurice et al. (2011), una de las más importantes es que las medidas repetidas están correlacionadas. El método GEE se presenta como una alternativa para extender las ecuaciones de máxima verosimilitud de los modelos lineales generalizados a datos correlacionados, este está basado en el concepto de "estimación de ecuaciones" y provee una aproximación general y unificada para analizar respuestas correlacionadas que pueden ser discretas o continuas. Para lograr dicha extensión el método GEE incorpora la matriz de covarianza del vector de respuestas, Y_i (Fitzmaurice et al., 2011).

4.1. Modelo de regresión Poisson

De acuerdo con Fitzmaurice et al. (2011) cuando la variable es un conteo, es razonable asumir que Y_i tiene una distribución Poisson. La distribución Poisson describe la probabilidad de que un número específico de eventos y_i ocurra como:

$$Pr(Y_i = y_i) = \frac{e^{-\mu}\mu^{y_i}}{y_i!}\ ;\ y_i = 0, 1, 2, ..., \tag{4-1}$$

donde μ es la media de la distribución. Así, la media y la varianza de esta distribución están dadas por:

$$E[Y_i] = \mu = Var(Y_i) \tag{4-2}$$

El modelo de regresión Poisson se deriva de la distribución Poisson. De acuerdo con Fitzmaurice et al. (2011), este modelo pertenece a la familia de modelos de regresión log-lineales, y se puede expresar como:

$$log(E[Y_i|X_1, ..., X_p]) = \sum_{j=1}^{p} \beta_j x_{ij}, \tag{4-3}$$

donde:
β_1, β_2,..., β_p: Parámetros de la regresión que representan el efecto de cada covariable sobre la media de la variable respuesta.
X_1, X_2,.., X_p: Conjunto de variables explicativas

4.2. Modelo de regresión binomial negativa

De acuerdo con Agresti (2015) en la práctica las observaciones de conteo frecuentemente exceden la variabilidad predicha por el modelo Poisson, ya que la varianza no es igual a la media, lo que puede ocasionar la presencia de sobredispersión, un fenómeno muy común en variables de tipo conteo. Ante este escenario, la otra distribución que se puede ajustar a variables de tipo conteo es la distribución binomial negativa.

La función de masa de probabilidad de la distribución binomial negativa es producida marginalmente por la mezcla gamma de las distribuciones Poisson (Agresti, 2002), esta se expresa como:

$$f(y;\mu,\alpha) = \frac{\Gamma(y_i + 1/\alpha)}{\Gamma(y_i+1)\Gamma(1/\alpha)} \left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(1 - \frac{1}{1+\alpha\mu_i}\right)^{y_i} \tag{4-4}$$

Dado y, se cumple que $\Gamma(y+1) = y!$, $\Gamma(y+1/\alpha-1) = (y+1/\alpha)!$, y $\Gamma(1/\alpha) = (1/\alpha-1)!$, teniendo esto en cuenta, el término izquierdo con las funciones Gamma de la ecuación 4-4 se reestructura entonces en la forma de una combinación, y el término extremo de la derecha se puede convertir en una sola fracción, resultando en la siguiente expresión popular de la distribución de probabilidad binomial negativa.

$$f(y;\mu,\alpha) = \binom{y_i + \frac{1}{\alpha} - 1}{\frac{1}{\alpha} - 1} \left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \tag{4-5}$$

α>0 es un tipo de parámetro de dispersión. Entre más grande sea α, más grande es la sobredispersión relativa a la Poisson. Los modelos lineales generealizados con distribución binomial negativa usan comúnmente el logaritmo como función de enlace. Expresando la ecuación 4-5 en términos del parámetro de dispersión α, la función de verosimilitud para un modelo lineal generalizado con distribución binomial negativa es:

$$\begin{aligned} L(\mu;y,\alpha) = \prod_{i=1}^{n} exp\left\{ y_i \ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \frac{1}{\alpha}\ln(1+\alpha\mu_i) + \ln\Gamma\left(y_i + \frac{1}{\alpha}\right) \right. \\ \left. - \ln\Gamma(y_i+1) - \ln\Gamma\left(\frac{1}{\alpha}\right) \right\} \end{aligned} \tag{4-6}$$

donde:
α: Es el parámetro de dispersión

μ_i : Es una función de β a través de $\eta_i = g(\mu_i) = \sum_j \beta_j x_i j$, donde g es la función de enlace de un modelo lineal generalizado

El logaritmo de la función de verosimilitud es obtenido tomando logaritmo natural en ambas partes de la ecuación. Como con los modelos Poisson, la función se convierte en aditiva en lugar de multiplicativa.

$$\mathcal{L}(\mu; y, \alpha) = \sum_{i=1}^{n} y_i \ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \frac{1}{\alpha}\ln(1+\alpha\mu_i) + \ln\Gamma\left(y_i + \frac{1}{\alpha}\right) - \ln\Gamma(y_i+1) - \ln\Gamma\left(\frac{1}{\alpha}\right) \tag{4-7}$$

El logaritmo de la función de verosimilitud de la binomial negativa, parametrizado en términos de β, los coeficientes del modelo, puede ser expresado como:

$$\mathcal{L}(\beta; y, \alpha) = \sum_{i=1}^{n} y_i \ln\left(\frac{\alpha\exp(x_i'\beta)}{1+\alpha\exp(x_i'\beta)}\right) - \frac{1}{\alpha}\ln\left(1+\alpha\exp(x_i'\beta)\right) + \ln\Gamma\left(y_i + \frac{1}{\alpha}\right) - \ln\Gamma(y_i+1) - \ln\Gamma\left(\frac{1}{\alpha}\right) \tag{4-8}$$

Los principios de máxima verosimilitud definen las ecuaciones de estimación como las derivadas de la función de verosimilitud logarítmica. Por simplicidad, el parámetro α se asume constante para las n observaciones y se trata como desconocido al igual que la varianza en los modelos normales, adicionalmente, el vector de parámetros β y α, son ortogonales y por tanto, sus estimadores son asintóticamente independientes de modo que el error estándar de $\hat{\beta}$ es el mismo cuando α es conocido o estimado (Agresti, 2015).

4.3. Regresión spline

De acuerdo con Ruppert et al. (2003) el modelo de regresión spline es útil cuando se observa gráficamente que la serie que se busca predecir presenta un comportamiento lineal por partes, más concretamente, se observan dos o más rectas con pendientes diferentes que se unen en un punto o valor determinado, así antes de dicho valor la pendiente de la recta es una y después de dicho valor la pendiente de la recta cambia. A este valor donde la pendiente de la recta cambia se le conoce como nudo. Como plantean Ruppert et al. (2003) el modelo de regresión spline, con un solo regresor x, se puede especificar de la siguiente manera:

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} b_k (x - \kappa_k)_+, \tag{4-9}$$

donde:
β_0, β_1 : Parámetros de la regresión
x : Variable explicativa
κ: Nodos o Nudos.
b_k: Efecto de las funciones truncadas

De acuerdo con Montgomery et al. (2012) los splines son polinomios por partes de orden k, los puntos donde se unen estos polinomios son los nodos. Por lo general se requiere que los valores de la función y las primeras $k-1$ derivadas coincidan en los nodos, de modo tal que el spline sea una función continua con $k-1$ derivadas continuas. Los autores resaltan que los spline cúbicos ($k=3$) generalmente son adecuados para la mayoría de los problemas prácticos. Montgomery et al. (2012), citaron a Wold (1974), de acuerdo con este autor se debe tener el mínimo número de nodos posibles, con por lo menos 4 o 5 datos por nodo.

Para el caso del trabajo actual, el número de variables explicativas que se usó fue superior a 1, por tanto, el modelo se planteó como un modelo GAM (Generalized additive model). De acuerdo con Molnar (2020) este tipo de modelos proporcionan un marco que permite generalizar un modelo lineal estándar para que admita la utilización de funciones no lineales de cada una de las variables, manteniendo la aditividad. De acuerdo con el autor este modelo se puede plantear como:

$$g(E_Y(y|x)) = \beta_0 + f_1(x_1) + f_2(x_2) + ... + f_p(x_p), \tag{4-10}$$

De esta manera se usaron dentro de las $f(x)$ las funciones planteadas en la regresión spline. Es importante mencionar que los métodos de suavizado comparados con los modelos lineales generalizados tienen la desventaja de la pérdida de interpretabilidad para describir el efecto de una variable explicativa sobre la variable respuesta. De esta manera no es claro cómo aplicar los intervalos de confianza a los efectos en un GAM, por lo tanto, es más difícil juzgar cuándo un efecto es más relevante. Adicionalmente, debido a que cualquier método de suavizado tiene potencialmente una gran cantidad de parámetros, puede requerir un n grande para estimar la forma funcional con precisión. Finalmente dado el número de parámetros posibles el sobreajuste siempre es un riesgo, por tanto se debe buscar un equilibrio entre el sesgo y la varianza del modelo (Agresti, 2015).

Los datos de la variable respuesta además de estar correlacionados, corresponden a conteos de líneas, De acuerdo con Fitzmaurice et al. (2011) los métodos de suavizado (smoothing), donde se enmarca la regresión spline, pueden ser utilizados en datos de este tipo. Adicionalmente, los autores resaltan que la aplicación de estos métodos de suavizado puede extenderse de forma natural a modelos lineales generalizados para datos longitudinales discretos, lo que permite estimar adecuadamente modelos de regresión para datos binarios, ordinales y de

conteo. Para mayor detalle consultar Fitzmaurice et al. (2011).

4.4. Ecuaciones de estimación generalizadas

Dado el tipo de datos de la variable respuesta, se hacía necesario poder enmarcar los modelos dentro de una metodología que permitiera incluir el hecho de que estos se encontraban correlacionados entre sí. De acuerdo con Fitzmaurice et al. (2011) un enfoque adecuado para este tipo de datos es el de ecuaciones de estimación generalizadas, GEE por sus siglas en inglés.

El enfoque GEE se basa en el concepto de "ecuaciones de estimación" y brinda un marco general y unificado para analizar variables respuestas correlacionadas, de naturaleza discreta o continua. La idea esencial detrás del enfoque GEE es generalizar y extender las ecuaciones de probabilidad habituales de un modelo lineal generalizado para una variable respuesta univariada, como se mencionó anteriormente, con observaciones correlacionadas, mediante la incorporación de la matriz de covarianza de Y (Fitzmaurice et al., 2011).

Para el sujeto i, sea

$$y_i = (y_{i1}, ..., y_{iT_i})' \quad y \quad \mu_i = (\mu_{i1}, ..., \mu_{iT_i}), \tag{4-11}$$

donde:
$\mu_{iT} = E(Y_{it})$

El número T_i de respuestas puede variar por cluster. Sea x_{it} un vector $p \times 1$ de valores de variables explicativas para y_{it}. El predictor lineal del modelo es $\eta_{it} = g(\mu_{it}) = \mathbf{x}'_{it}\beta$ para la función de enlace g. El modelo se refiere a la distribución marginal para cada t en lugar de la distribución conjunta. Sea $\mathbf{X_i}$, la matriz $T_i \times p$ de valores predictores por grupo (o sujeto) i, para los cuales la fila t es $\mathbf{x}'_{ij}$. Se asume que y_{it} tiene una función de masa de probabilidad de la forma:

$$f(y_{it} : \theta_{it}, \phi) = \exp\{[y_{it}\theta_{it} - b(\theta_{it})]/\phi + c(y_{it}, \phi)\} \tag{4-12}$$

Cuando ϕ es conocido, esta es la familia exponencial natural, con parámetro natural θ_{it}, donde:

$$\mu_{it} = E(Y_{it}) = b'(\theta_{it}), \quad v(\mu_{it}) = var(Y_{it}) = b"('\theta_{it})\phi. \tag{4-13}$$

El método GEE asume una matriz de correlación de trabajo $\mathbf{R}(\alpha)$ para $\mathbf{Y_i}$, dependiendo de los parámetros α. De acuerdo con Ziegler (2011), las opciones estándar para trabajar con matrices de correlación en los paquetes de software más ampliamente utilizados son:

- Fija
- Independiente
- Exchangeable(Intercambiable)
- m-dependiente
- Autoregresiva de orden 1
- No estructurada

Al igual que Agresti (2002), Ziegler (2011), destaca que una de las estructuras de correlación más utilizadas es la estructura de correlación "exchangeable", la cual se define como :

$$corr(y_{it}, y_{it'}) = \begin{cases} 1, & si\ \ t = t', \\ \alpha, & si\ \ t \neq t'. \end{cases} \tag{4-14}$$

Para ver más detalle de las estructuras de correlación consultar Ziegler (2011). La matriz de correlación de trabajo "exchangeable" se compone de $corr(Y_{it}, Y_{is} = \alpha)$ para cada par en $\mathbf{Y_i}$. Sea $b_i(\theta) = (b(\theta_{i1}), ..., b(\theta_{iT_i}))$, y $\mathbf{B}_i$ una matriz diagonal con los elementos de la diagonal pricipal iguales a $b_i"(\theta)$. Entonces la matriz de covarianza de trabajo para $\mathbf{Y}_i$ es

$$\mathbf{V}_i = \mathbf{B}_i^{1/2}\mathbf{R}(\alpha)\mathbf{B_i^{1/2}}\phi \tag{4-15}$$

donde, $\mathbf{V_i} = cov(\mathbf{Y_i})$ si $\mathbf{R}$ es la matriz de correlación verdadera para $\mathbf{Y_i}$.

Ahora, sean:

- Δ_i la matriz diagonal con elementos $\partial\theta_{it}/\partial\eta_{it}$ en la diagonal principal para $t = 1, ..., T_i$ (para la función de enlace canónica, esta es la matriz identidad).
- $\mathbf{D}_i = \partial\mu_i/\partial\beta = \mathbf{B}_i\Delta_i\mathbf{X}_i$ una matriz $T_i \times p$ con la típica expresión de elementos $\partial\mu_{it}/\partial\beta_j$ en la forma $(\partial\mu_{it}/\partial\theta_{it})(\partial\theta_{it}/\partial\eta_{it})(\partial\eta_{it}/\partial\beta_j)$

Para los modelos lineales generalizados univariados las ecuaciones de estimación de la cuasi-verosimilitud tienen la forma:

$$\sum(\partial\mu_i/\partial\beta)v(\mu_i)^{-1}[y_i - \mu_i(\beta)] = 0, \tag{4-16}$$

donde $\mu_i = \mu_i(\beta) = g^{-1}(x_i\beta)$.

De acuerdo con Fitzmaurice et al. (2011), para el caso de GEE, el estimador de β para modelos marginales o modelos lineales generalizados para datos longitudinales proviene de minimizar la función objetivo:

$$\sum_{i=1}^{N} \{y_i - \mu_i(\beta)\}' V_i^{-1} \{y_i - \mu_i(\beta)\}, \tag{4-17}$$

Se puede demostrar que sí un mínimo de la función dada por (4-6) existe, entonces debe resolver el siguiente sistema de ecuaciones generalizadas:

$$\sum_{i=1}^{N} D_i' V_i^{-1} (y_i - \mu_i) = 0, \tag{4-18}$$

donde:
V_i: matriz de covarianza de "trabajo", aproxima la verdadera matriz de covarianza de Y_i
D_i: $\partial\mu_i/\partial\beta$, la matriz que contiene la derivada de μ_i con respecto a los componentes de β.

El estimador GEE $\hat{\beta}$ es la solución de estas ecuaciones. Sí se utiliza $R(\alpha) = I$, se tratan los pares de respuestas como independientes, de esta manera la solución de $\hat{\beta}$ es entonces la misma que un estimador ordinario para un modelo lineal generalizado con la función de enlace y la función de la varianza elegida, tratando $(y_{i1}, ..y_{iT_i})$ como observaciones independientes (Agresti, 2002).

Adicionalmente, de acuerdo con Fitzmaurice et al. (2011), cuando hay convergencia, la solución de las ecuaciones de estimación generalizadas tienen las siguientes propiedades:

1. $\hat{\beta}$ es un estimador consistente de β
2. En muestras grandes, la distribución muestral de β es normal multivariante.

Finalmente Fitzmaurice et al. (2011), concluyen que:

- En muchos estudios longitudinales la aproximación GEE para estimar β es casi tan precisa como el método de máxima verosimilitud.
- El estimador GEE $\hat{\beta}$ es un estimador consistente de β, incluso si la matriz de correlación ha sido mal especificada.
- La aproximación GEE, puede manejar fácilmente el desequilibrio debido a datos faltantes en las variables de respuesta.
- La aproximación GEE también puede ser utilizada en datos continuos.

5. Aplicación: estimación de modelos

Para estimar los parámetros de los distintos modelos considerados para pronosticar el tamaño del mercado móvil de Colombia, se utilizó la aproximación GEE (Generalized estimating equations), especificando la estructura de correlación denominada "exchangeable", que es una de las más utilizadas en la práctica, de acuerdo con Agresti (2002), esta estructura de correlación trata corr(Y_t, Y_s) como idéntica para todo s y t, de acuerdo con el autor es más flexible y realista que suponer independencia entre las observaciones, adicionalmente, recomienda su uso cuando no se esperan diferencias muy grandes en las correlaciones, que es el caso en este trabajo, ya que reconoce la dependencia a costa de un solo parámetro adicional.

En total se consideraron cinco modelos, el primero correspondió al modelo lineal normal, el cual se utilizó como modelo de referencia, el segundo fue el modelo de regresión Poisson, el tercero fue el modelo de regresión binomial negativa, el cuarto fue el modelo de regresión con smoothing splines, y finalmente se estimó un modelo de regresión Poisson con smoothing splines. Para cada uno de los modelos se reportó el valor estimado de cada uno de los coeficientes de regresión, el valor de la desviación estándar asociado, el estadístico de Wald y el valor p asociado a dicho estadístico, estos dos últimos se utilizarán para determinar la significancia de cada una de las variables de los modelos. Adicionalmente se reportó el criterio de información de cuasiverosimilitud (QIC).

Pan (2001) propuso el QIC, que es una extensión del AIC para modelos estimados a través de GEE, esto debido a que el criterio de información de Akaike es una medida utilizada para la selección de modelos pero este solo puede ser utilizado en modelos estimados por máxima verosimilitud, donde los errores deben de ser independientes. Pan (2001) destaca que el QIC puede ser utilizado para la selección de modelos, de la estructura de correlación de trabajo y de las covariables, al igual que el AIC, es preferible un modelo que tenga un menor valor de QIC, ya que esto implica un mejor ajuste a los datos. La fórmula del QIC se presenta en la ecuación 5-1:

$$QIC(R) \equiv -2Q(\hat{\beta}(R); I, \mathcal{D}) + 2trace(\hat{\Omega}_I \hat{V}_r) \tag{5-1}$$

donde de acuerdo con Tsai et al. (2011),

- $\hat{\beta}(R)$ es el estimador GEE, usando cualquier estructura de covarianza de trabajo general R.
- $Q(\hat{\beta}(R); I, \mathcal{D})$ es la cuasi-verosimilitud marginal bajo el modelo de independencia funcional con la sustitución del estimador $\hat{\beta}(R)$
- I esa una matriz identidad
- $\hat{\Omega}_I = -\frac{\partial^2 Q(\beta;I)}{\partial\beta\partial\beta^t}\Big|_{\beta=\hat{\beta}(R)}$
- $\hat{V}_R$ denota el estimador sándwich de la matriz de covarianza de $var(\hat{\beta})$.

Autores como Luo and Pan (2022),Tsai (2015) y Tsai et al. (2011) , han utilizado el QIC para seleccionar el mejor modelo entre un conjunto de modelos ajustados para abordar distintos problemas.

Finalmente, para poder determinar cuál de los 5 modelos considerados en este trabajo tuvo un mejor ajuste, la base de datos se dividió en un conjunto de entrenamiento y otro de prueba, de esta manera el conjunto de entrenamiento (Train) correspondió a los primeros 45 meses observados, que corresponde al 90 % de los datos, y se dejó el 10 % restante para validación, se tomaron 6 meses ya que este es el horizonte de pronóstico habitual. Finalmente la estimación del modelo de regresión binomial negativa se realizó en Python lenguaje de programación donde se encontraba disponible las ecuaciones de estimación generalizadas permitiendo la especificación de la distribución binomial negativa , donde se utilizaron principalmente las herramientas de la librería statsmodel (Seabold and Perktold, 2010), los demás modelos fueron estimados en R donde se utilizaron principalmente herramientas de las librerías “geepack”(Halekoh et al., 2006) y “gam”(Hastie, 2022).

5.1. Modelo de regresión lineal normal

El primer modelo utilizado correspondió al modelo de regresión lineal normal, este modelo se utilizó como modelo de referencia, haciendo uso, como ya se mencionó, del método GEE, especificando en la estructura de correlación “exchangeable” para la matriz de covarianza de trabajo, en R se especifica en el argumento corstr(que corresponde a correlation structure). La ecuación 5-2 corresponde al modelo con todas las variables, que fue utilizado inicialmente, y la ecuación 5-3 corresponde al modelo utilizado finalmente. En la Tabla 5-1 se pueden observar los resultados de la estimación de parámetros para el modelo con todas las variables, donde las únicas variables significativas fueron X_1 y X_2 y el intercepto, por esta razón se eliminó la variable X_3 que fue la menos significativa, al realizar este paso, se ajustó de nuevo el modelo y en este todas las variables fueron significativas como se puede observar

en la Tabla 5-2. Adicionalmente en la Tabla 5-3 se reportaron los valores obtenidos para el QIC del modelo con todas las variables y el modelo reducido respectivamente, en este caso la diferencia fue de 0.0603 % siendo aproximadamente iguales, por tanto se observa que el modelo reducido no empeoró la bondad de ajuste del modelo inicial.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \epsilon_i \tag{5-2}$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{4i} + \beta_4 X_{5i} + \beta_5 X_{6i} + \epsilon_i \tag{5-3}$$

Tabla 5-1.: Resultados estimación modelo de regresión lineal normal GEE con todas las variables

Coeficiente	Estimación	DE	Wald	P(|Z|>|W|)
β_0	4.10E+07	2.75E+06	221.84	< 2E-16
β_1	1.06E+01	6.90E-01	235.34	< 2E-16
β_2	1.71E+00	3.22E-01	28.18	1.1E-07
β_3	-8.95E-04	6.05E-03	0.02	0.88
β_4	3.64E-03	5.08E-03	0.51	0.47
β_5	9.84E-03	1.67E-02	0.35	0.56
β_6	-2.54E-02	1.72E-02	2.19	0.14

Tabla 5-2.: Resultados estimación modelo de regresión lineal normal reducido con GEE

Coeficiente	Estimación	DE	Wald	P(|Z|>|W|)
β_0	4.07E+07	1.04E+06	1525.67	< 2E-16
β_1	1.06E+01	6.69E-01	253.03	< 2E-16
β_2	1.73E+00	2.52E-01	47.02	7.0E-12
β_3	2.90E-03	5.20E-04	31.13	2.4E-08
β_4	7.50E-03	3.07E-03	5.98	0.014
β_5	-2.30E-02	4.05E-03	32.38	1.3E-08

Tabla 5-3.: QIC modelos con distribución de la variable respuesta normal

Modelo	Valor QIC
Modelo todas las variables	6.65E+12
Modelo reducido	6.65E+12

5.2. Modelo de regresión Poisson

El segundo modelo utilizado correspondió al modelo de regresión Poisson, como ya se mencionó anteriormente, se eligió la distribución Poisson ya que la variable respuesta corresponde a una serie asociada a conteos. También el método de estimación utilizado fue GEE, especificando en la estructura de correlación "exchangeable" para la matriz de covarianza de trabajo, en R se especifica en el argumento corstr(que corresponde a correlation structure). La ecuación 5-4 corresponde al modelo con todas las variables, que fue el utilizado inicialmente, y la ecuación 5-5 corresponde al modelo utilizado finalmente. En la Tabla 5-4 se pueden observar los resultados de la estimación de parámetros para el modelo con todas las variables, donde las únicas variables significativas fueron X_1 y X_2 y el intercepto, por esta razón se eliminó la variable X_3 que fue la menos significativa, al realizar este paso, se ajustó de nuevo el modelo y en este todas las variables fueron significativas como se puede observar en la Tabla 5-5. Adicionalmente en la Tabla 5-6, se presenta el QIC donde se puede observar que los valores son aproximadamente iguales para ambos modelos, por tanto al quitar la variable X_3 no se afectó el grado de ajuste del modelo con todas las variables.

$$log(\lambda_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} \tag{5-4}$$

$$log(\lambda_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{4i} + \beta_4 X_{5i} + \beta_5 X_{6i} \tag{5-5}$$

Tabla 5-4.: Resultados estimación modelo de regresión Poisson GEE con todas las variables

Coeficiente	Estimación	DE	Wald	P(|Z|>|W|)
β_0	1.76E+01	4.09E-02	1.86E+05	<2.00E-16
β_1	1.51E-07	9.50E-09	2.54E+02	<2.00E-16
β_2	2.66E-08	4.67E-09	3.23E+01	1.30E-08
β_3	-8.92E-12	8.80E-11	1.00E-02	0.92
β_4	5.11E-11	7.38E-11	4.80E-01	0.49
β_5	1.30E-10	2.44E-10	2.80E-01	0.59
β_6	-3.55E-10	2.52E-10	1.98E+00	0.16

Tabla 5-5.: Resultados estimación modelo de regresión Poisson reducido con GEE

Coeficiente	Estimación	DE	Wald	P(|Z|>|W|)
β_0	1.76E+01	1.54E-02	1.31E+06	<2E-16
β_1	1.52E-07	9.13E-09	2.77E+02	<2E-16
β_2	2.68E-08	3.62E-09	5.47E+01	1.4E-13
β_3	4.37E-11	7.77E-12	3.16E+01	1.9E-08
β_4	1.07E-10	4.59E-11	5.40E+00	0.02
β_5	-3.31E-10	5.98E-11	3.06E+01	3.2E-08

Tabla 5-6.: QIC modelos con distribución de la variable respuesta Poisson

Modelo	Valor QIC
Modelo todas las variables	-1.01E+11
Modelo reducido	-1.01E+11

5.3. Modelo de regresión binomial negativa

El tercer modelo utilizado correspondió al modelo de regresión binomial negativa, como ya se mencionó anteriormente, se elegió la distribución binomial negativa ya que la variable respuesta corresponde a una serie asociada a conteos, presentándose como una alternativa a la distribución Poisson, debido a la posible presencia de sobredispersión, ya que la media muestral de variable era igual a 6.71E+07, mientras que la varianza muestral correspondía a 1.47E+13, como se puede observar la media de la variable difiere considerablemente de la varianza. También el método de estimación utilizado fue GEE, especificando en la estructura de correlación la matriz "exchangeable". La ecuación 5-6 corresponde al modelo con todas las variables, que fue el utilizado inicialmente, y la ecuación 5-7 corresponde al modelo utilizado finalmente. En la Tabla 5-7 se pueden observar los resultados de la estimación de parámetros para el modelo con todas las variables, donde las únicas variables significativas fueron X_1 y X_2 y el intercepto, por esta razón se eliminó la variable X_3 que fue la menos significativa, al realizar este paso, se ajustó de nuevo el modelo y en este todas las variables fueron significativas como se puede observar en la Tabla 5-8. Adicionalmente se calculó el QIC, que se presenta en la Tabla 5-9, donde el modelo reducido presentó una disminución en el QIC, por tanto éste presentó un mejor ajuste que el modelo con todas las variables.

$$log(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} \tag{5-6}$$

$$log(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{4i} + \beta_4 X_{5i} + \beta_5 X_{6i} \tag{5-7}$$

donde como se mencionó en el marco teórico de este documento,
μ_i : Es una función de β a través de $\eta_i = g(\mu_i) = \sum_j \beta_j x_{ij}$, donde g es la función de enlace de un modelo lineal generalizado

Y,

$$\mathcal{L}(\beta; y, \alpha) = \sum_{i=1}^{n} y_i \ln\left(\frac{\alpha \exp(x_i'\beta)}{1 + \alpha \exp(x_i'\beta)}\right) - \frac{1}{\alpha}\ln\left(1 + \alpha \exp(x_i'\beta)\right) + \ln\Gamma\left(y_i + \frac{1}{\alpha}\right) - \ln\Gamma(y_i + 1) - \ln\Gamma\left(\frac{1}{\alpha}\right) \tag{5-8}$$

Tabla 5-7.: Resultados estimación modelo de regresión binomial negativa con GEE con todas las variables

Coeficiente	Estimación	DE	Wald	P(|Z|>|W|)
β_0	17.632	0.041	428.003	0
β_1	1.51E-07	9.61E-09	15.733	0
β_2	2.62E-08	4.77E-09	5.503	0
β_3	-1.08E-11	8.85E-11	-0.121	0.903
β_4	5.30E-11	7.43E-11	0.712	0.476
β_5	1.38E-10	2.46E-10	0.561	0.575
β_6	-3.61E-10	2.54E-10	-1.423	0.155

Tabla 5-8.: Resultados estimación modelo de regresión binomial negativa reducido con GEE

Coeficiente	Estimación	DE	Wald	P(|Z|>|W|)
β_0	17.628	0.016	1128.843	0
β_1	1.52E-07	9.31E-09	16.308	0
β_2	2.65E-08	3.70E-09	7.149	0
β_3	4.41E-11	7.88E-12	5.594	0
β_4	1.10E-10	4.75E-11	2.315	0.021
β_5	-3.33E-10	6.10E-11	-5.454	0

Tabla 5-9.: QIC modelos con respuesta binomial negativa

Modelo	Valor QIC
Modelo todas las variables	1.98E+12
Modelo reducido	1.65E+12

5.4. Modelo de regresión splines

El cuarto modelo utilizado correspondió al modelo de regresión splines, se consideró este modelo ya que como se observó en el análisis exploratorio, todas las variables parecían tener una relación no lineal fuerte con la variable respuesta. Como ya se mencionó, el método de estimación fue GEE, especificando en la estructura de correlación la matriz "exchangeable".

La ecuación 5-9 corresponde al modelo especificado inicialmente con todas las variables, donde se agregaron cubic splines con un nudo para todas éstas, la ecuación 5-10 corresponde al modelo utilizado finalmente. En la Tabla 5-10 se pueden observar los resultados de la estimación de parámetros para el modelo con todas las variables, donde la única variable que no fue significativa fue la variable X_6, por esta razón se eliminó y se estimó de nuevo sin ella, el QIC obtenido para este modelo fue levemente superior al del modelo inicial, lo que implica que hubo una pérdida en la bondad de ajuste, sin embargo dicha variación no fue significativa. Finalmente, en el modelo especificado en la ecuación 5-10 todas las variables fueron significativas en al menos uno de los coeficientes como se puede observar en la Tabla 5-11.

$$\begin{aligned} Y =&\beta_0 + \beta_{10}X_1 + \beta_{11}X_1^2 + \beta_{12}X_1^3 + \beta_1(X_1 - \sigma_0)_+^3 + \beta_{20}X_2 + \beta_{22}X_2^2 + \beta_{22}X_2^3 + \beta_2(X_2 - \epsilon_0)_+^3 + \\ &\beta_{30}X_3 + \beta_{31}X_3^2 + \beta_{32}X_3^3 + \beta_3(X_3 - \tau_0)_+^3 + \beta_{40}X_4 + \beta_{41}X_4^2 + \beta_{42}X_4^3 + \beta_4(X_4 - \mu_0)_+^3 + \\ &\beta_{50}X_5 + \beta_{51}X_5^2 + \beta_{52}X_5^3 + \beta_5(X_5 - \alpha_0)_+^3 + \beta_{60}X_6 + \beta_{61}X_6^2 + beta_{62}X_6^3 + \beta_6(X_6 - \gamma_0)_+^3 \end{aligned} \tag{5-9}$$

$$\begin{aligned} Y =&\beta_0 + \beta_{10}X_1 + \beta_{11}X_1^2 + \beta_{12}X_1^3 + \beta_1(X_1 - \sigma_0)_+^3 + \beta_{20}X_2 + \beta_{22}X_2^2 + \beta_{22}X_2^3 + \beta_2(X_2 - \epsilon_0)_+^3 + \\ &\beta_{30}X_3 + \beta_{31}X_3^2 + \beta_{32}X_3^3 + \beta_3(X_3 - \tau_0)_+^3 + \beta_{40}X_4 + \beta_{41}X_4^2 + \beta_{42}X_4^3 + \beta_4(X_4 - \mu_0)_+^3 + \\ &\beta_{50}X_5 + \beta_{51}X_5^2 + \beta_{52}X_5^3 + \beta_5(X_5 - \alpha_0)_+^3 \end{aligned} \tag{5-10}$$

Donde,
σ_0, ϵ_0, τ_0, μ_0, α_0 y γ_0 fueron estimados por el software.

Tabla 5-10.: Resultados estimación modelo de regresión splines GEE con todas las variables

Coeficiente	Estimación	DE	Wald	P(|Z|>|W|)
β_0	62098408	487494	16226.42	<2E-16
β_{10}	3122488	1046835	8.9	0.003
β_{11}	4468788	513544	75.72	<2E-16
β_{12}	6522824	1114136	34.28	0
β_1	6352617	1431020	19.71	0
β_{20}	211112	249824	0.71	0.398
β_{21}	-592968	602338	0.97	0.325
β_{22}	1705078	418693	16.58	0
β_2	2588810	519195	24.86	0
β_{30}	3044006	1073006	8.05	0.005
β_{31}	-1379805	2163679	0.41	0.524
β_{32}	6121588	3702741	2.73	0.098
β_3	-14297306	15337772	0.87	0.351
β_{40}	1286745	847610	2.3	0.129
β_{41}	1863414	1329251	1.97	0.161
β_{42}	2360004	1324348	3.18	0.075
β_4	2448566	1531457	2.56	0.11
β_{50}	-5298103	2681675	3.9	0.048
β_{51}	-5076517	2434546	4.35	0.037
β_{52}	-4412253	3871041	1.3	0.254
β_5	11393670	15365839	0.55	0.458
β_{60}	1586692	2602457	0.37	0.542
β_{61}	1747019	2678690	0.43	0.514
β_{62}	333020	2755376	0.01	0.904
β_6	172510	2781818	0	0.951

Tabla 5-11.: Resultados estimación modelo de regresión splines con GEE

Coeficiente	Estimación	DE	Wald	P(|Z|>|W|)
β_0	62292864	259184	57764.29	<2e-16
β_{10}	3078268	687783	20.03	0
β_{11}	4601569	487634	89.05	<2e-16
β_{12}	6203737	1065521	33.9	0
β_1	7357960	1118913	43.24	0
β_{20}	170200	264381	0.41	0.52
β_{21}	-704096	506749	1.93	0.165
β_{22}	1781816	415671	18.37	0
β_2	2796125	406973	47.2	0
β_{30}	2483477	407082	37.22	0
β_{31}	439229	1272669	0.12	0.73
β_{32}	6847071	3623231	3.57	0.059
β_3	-4652753	13183236	0.12	0.724
β_{40}	1403057	511654	7.52	0.006
β_{41}	821387	729220	1.27	0.26
β_{42}	1154127	928129	1.55	0.214
β_4	778397	816628	0.91	0.341
β_{50}	-3409782	516538	43.58	0
β_{51}	-3972856	807047	24.23	0
β_{52}	-5552093	1282499	18.74	0
β_5	3092247	13727894	0.05	0.822

Tabla 5-12.: QIC modelos splines

Modelo	Valor QIC
Modelo todas las variables	1.16E+12
Modelo reducido	1.26E+12

5.5. Modelo de regresión Poisson con splines

El último modelo utilizado correspondió al modelo de regresión splines, adicionalmente especificando la distribución Poisson para la variable respuesta, se consideró este modelo ya que como se observó en el análisis exploratorio, todas las variables parecían tener una relación no lineal fuerte con la variable respuesta. Como ya se mencionó, el método de estimación fue GEE, especificando en la estructura de correlación la matriz "exchangeable".

La ecuación 5-11 corresponde al modelo con todas las variables, donde se agregaron cubic splines con un nudo para todas las variables, que fue el utilizado inicialmente, y la ecuación 5-12 corresponde al modelo utilizado finalmente. En la Tabla 5-13 se pueden observar los resultados de la estimación de parámetros para el modelo saturado, donde las única variable que no fue significativa fue X_6, por esta razón se eliminó, por tanto se eliminó y se utilizó el modelo especificado en la ecuación 5-12, donde como se puede observar en la Tabla 5-14 todas las variables fueron significativas en al menos 1 de los coeficientes. El QIC se sostuvo, como se puede observar en la tabla 5-15, por tanto en este caso la eliminación de la variable X_6 no afectó la bondad de ajuste del modelo.

$$\begin{aligned}&\beta_0 + \beta_{10}X_1 + \beta_{11}X_1^2 + \beta_{12}X_1^3 + \beta_1(X_1 - \sigma_0)_+^3 + \beta_{20}X_2 + \beta_{22}X_2^2 + \beta_{22}X_2^3 + \beta_2(X_2 - \epsilon_0)_+^3 + \\ &\beta_{30}X_3 + \beta_{31}X_3^2 + \beta_{32}X_3^3 + \beta_3(X_3 - \tau_0)_+^3 + \beta_{40}X_4 + \beta_{41}X_4^2 + \beta_{42}X_4^3 + \beta_4(X_4 - \mu_0)_+^3 + \\ &\beta_{50}X_5 + \beta_{51}X_5^2 + \beta_{52}X_5^3 + \beta_5(X_5 - \alpha_0)_+^3 + \beta_{60}X_6 + \beta_{61}X_6^2 + beta_{62}X_6^3 + \beta_6(X_6 - \gamma_0)_+^3\end{aligned} \tag{5-11}$$

$$\begin{aligned}log(\lambda_i) =&\beta_0 + \beta_{10}X_1 + \beta_{11}X_1^2 + \beta_{12}X_1^3 + \beta_1(X_1 - \sigma_0)_+^3 + \beta_{20}X_2 + \beta_{22}X_2^2 + \beta_{22}X_2^3 + \beta_2(X_2 - \epsilon_0)_+^3 + \\ &\beta_{30}X_3 + \beta_{31}X_3^2 + \beta_{32}X_3^3 + \beta_3(X_3 - \tau_0)_+^3 + \beta_{40}X_4 + \beta_{41}X_4^2 + \beta_{42}X_4^3 + \beta_4(X_4 - \mu_0)_+^3 + \\ &\beta_{50}X_5 + \beta_{51}X_5^2 + \beta_{52}X_5^3 + \beta_5(X_5 - \alpha_0)_+^3\end{aligned} \tag{5-12}$$

Donde,
τ_0, μ_0, α_0 y γ_0 fueron estimados por el software.

Tabla 5-13.: Resultados estimación modelo de regresión spline Poisson saturado con GEE

Coeficiente	Estimación	DE	Wald	P(|Z|>|W|)
β_0	17.944	0.008	5600000	<2E-16
β_{10}	0.048	0.016	9.3	0
β_{11}	0.069	0.008	76.9	<2E-16
β_{12}	0.099	0.017	35.4	0
β_1	0.093	0.022	18.5	0
β_{20}	0.003	0.004	0.74	0.507
β_{21}	-0.01	0.009	1.09	0.156
β_{22}	0.026	0.006	17	0
β_2	0.038	0.008	23	0
β_{30}	0.047	0.017	7.96	0
β_{31}	-0.02	0.033	0.38	0.684
β_{32}	0.093	0.056	2.81	0.056
β_3	-0.226	0.229	0.98	0.702
β_{40}	0.021	0.013	2.45	0.006
β_{41}	0.03	0.02	2.15	0.218
β_{42}	0.037	0.02	3.3	0.191
β_4	0.039	0.024	2.76	0.281
β_{50}	-0.082	0.042	3.89	0
β_{51}	-0.078	0.038	4.26	0
β_{52}	-0.07	0.059	1.38	0
β_5	0.181	0.229	0.62	0.802
β_{60}	0.025	0.04	0.4	0.528
β_{61}	0.027	0.042	0.41	0.521
β_{62}	0.006	0.043	0.02	0.89
β_6	0.003	0.043	0.01	0.938

Tabla 5-14.: Resultados estimación modelo de regresión spline Poisson con GEE

Coeficiente	Estimación	DE	Wald	P(|Z|>|W|)
β_0	17.947	0.004	20400000	<2e-16
β_{10}	0.048	0.011	21.1	0
β_{11}	0.071	0.008	89.1	<2e-16
β_{12}	0.095	0.016	35.5	0
β_1	0.108	0.017	40.2	0
β_{20}	0.003	0.004	0.44	0.507
β_{21}	-0.011	0.008	2.01	0.156
β_{22}	0.027	0.006	18.6	0
β_2	0.041	0.006	44.5	0
β_{30}	0.039	0.006	38.7	0
β_{31}	0.008	0.019	0.17	0.684
β_{32}	0.104	0.054	3.66	0.056
β_3	-0.076	0.199	0.15	0.702
β_{40}	0.022	0.008	7.64	0.006
β_{41}	0.014	0.011	1.52	0.218
β_{42}	0.018	0.014	1.71	0.191
β_4	0.013	0.012	1.16	0.281
β_{50}	-0.053	0.008	43.9	0
β_{51}	-0.062	0.012	25.3	0
β_{52}	-0.085	0.019	19.4	0
β_5	0.052	0.207	0.06	0.802

Tabla 5-15.: QIC modelos con respuesta poisson

Modelo	Valor QIC
Modelo todas las variables	-1.01E+11
Modelo reducido	-1.01E+11

Adicionalmente, en la Figura 5-1, se presentan los valores estimados por cada uno de los modelos ajustados versús los valores reales, se puede observar que de manera general los modelos capturaron bien la tendencia y gráficamente parecen ajustarse bien a los datos, sin embargo en el caso de los splines en la primera estimación por fuera de la muestra se observa que el valor fue significativamente superior al real, los modelos que presentaron el mejor ajuste fueron el modelo de regresión Poisson y el modelo de regresión Binomial Negativa.

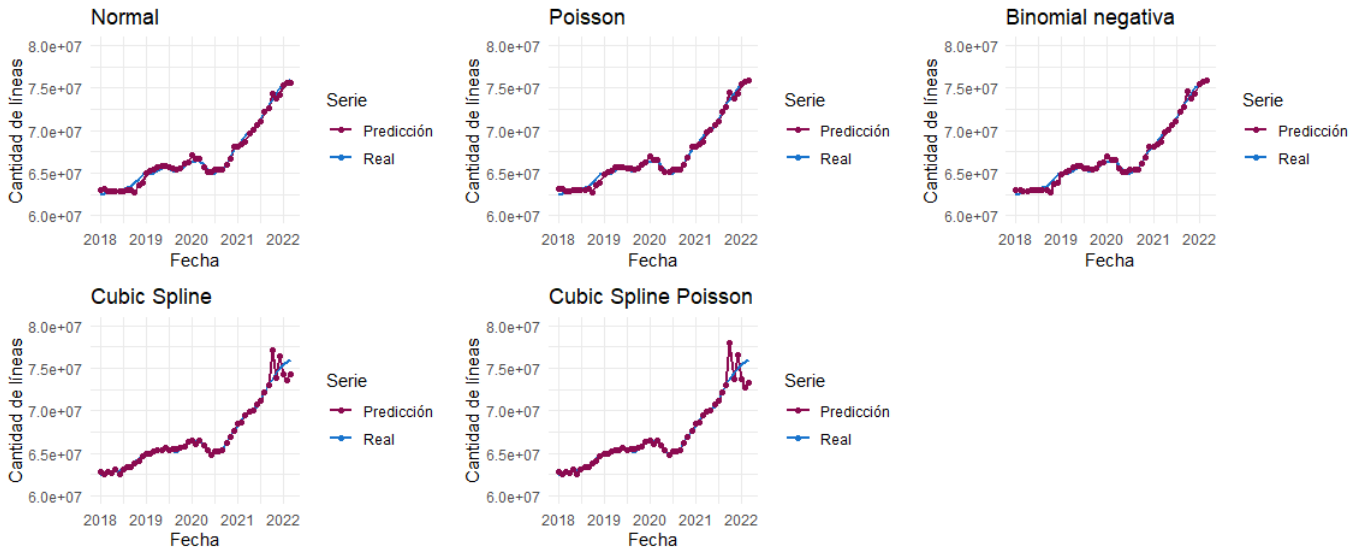

Figura 5-1.: Valores ajustados por todos los modelos versus valores reales de la variable respuesta.
Fuente: Elaboración propia.

Finalmente, en la Figura 5-2 se presentan los gráficos de los residuales de Pearson para los 5 modelos. De acuerdo con Park et al. (1998), en el método GEE la estimación de los parámetros de correlación se basa en los residuales de Pearson, que implícitamente se supone distibuyen asintóticamente de manera normal, conservando los supuestos de media 0 y varianza constante, los autores demuestran a través de estudios de simulación que la elección de los residuales tiene poco o ningún efecto sobre las propiedades de las estimaciones resultantes, por lo tanto, el uso de los residuales de Pearson también puede ser usado en respuestas discretas como la Poisson. Como se puede observar los residuales no parecen presentar ningún patrón definido.

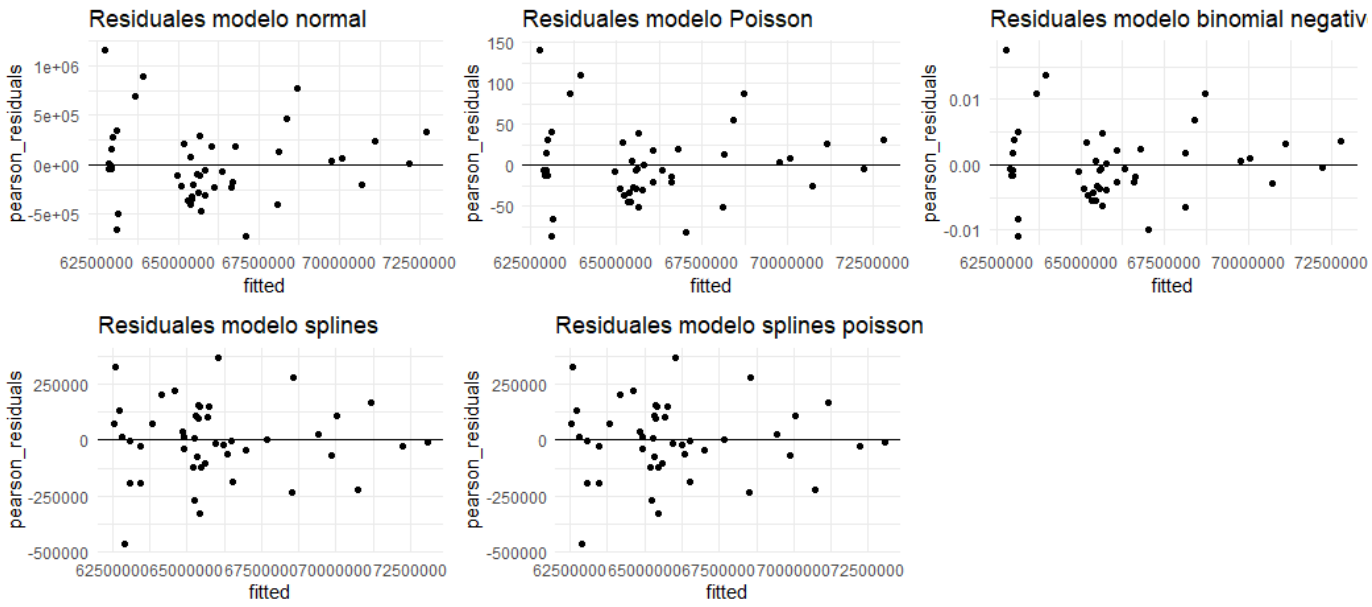

Figura 5-2.: Gráficos residuales de Pearson.
Fuente: Elaboración propia.

En la siguiente sección se observarán las métricas calculadas para corroborar cuál fue efectivamente el modelo que presentó los menores márgenes de error y el mejor grado de ajuste y que será elegido para pronosticar.

5.6. Selección del modelo final

De acuerdo con Wooldridge (2015) para elegir un método de pronóstico, se necesita una manera de comparar los modelos y de esta manera poder decidir cuál es el más adecuado. De acuerdo con el autor, de modo general, existen dos tipos de criterios de elección, criterios dentro de la muestra y criterios fuera de la muestra. En un contexto de regresión clásica, los criterios dentro de la muestra incluyen R-cuadrado y el R-cuadrado ajustado, MSE y criterios de información.

Para el caso en el que se desea utilizar un modelo para pronosticar, es mejor usar cristerios fuera de la muestra (teniendo en cuenta también la calidad del ajuste) , ya que el pronóstico es esencialmente un problema de este tipo. Un modelo podría proporcionar un buen ajuste en la muestra utilizada para estimar los parámetros, pero esto no implica necesariamente un buen desempeño en los pronósticos realizados. De esta manera, para realizar una comparación fuera de muestra es necesario utilizar la primera parte de la muestra de datos disponibles para estimar los parámetros del modelo y guardar la última parte de la muestra para medir sus capacidades de pronóstico (Wooldridge, 2015). Tal como se mencionó anteriormente que

se hizo en este trabajo.

De acuerdo con Wooldridge (2015), las dos medidas más comunes para medir la calidad del pronóstico son error cuadrático medio y el error absoluto medio. La raíz del error cuadrático medio (RMSE) se define como se muestra en la ecuación 5-12, y el error medio absoluto (MAE) como se muestra en la ecuación 5-13:

$$RMSE = \left(m^{-1} \sum_{h=0}^{m-1} \hat{e^2}_{n+h+1} \right)^{1/2} \tag{5-13}$$

$$MAE = m^{-1} \sum_{h=0}^{m-1} |\hat{e^2}_{n+h+1}| \tag{5-14}$$

donde,

$\hat{e}_{n+h+1} = y_{n+h+1} - \hat{y}_{n+h+1}$
m= Cantidad de observaciones utilizadas para pronosticar.

Adicionalmente para complementar las dos medidas anteriores, se encontró en la literatura el Error Porcentual Absoluto Medio (MAPE). De acuerdo con de Myttenaere et al. (2016) se usa a menudo en la práctica debido a que su interpretación es muy intuitiva en términos del error relativo. El MAPE se define como se muestra en la ecuación 5-14:

$$MAPE = \frac{1}{m} \sum_{h=0}^{m-1} \left| \frac{\hat{e^2}_{n+h+1}}{y_{n+h+1}} \right| \times 100 \tag{5-15}$$

En las 3 medidas, se busca el mínimo valor posible, de esta manera el mejor modelo es aquel que tenga un menor MAE, RMSE y MAPE. Las anteriores medidas fueron calculadas en este trabajo, como se mencionó anteriormente, entrenando el modelo con 41 datos y dejando los 6 meses restantes para validación (se elegieron 6 meses debido a que este es el horizonte habitual de pronóstico).

Adicionalmente, se realizó el cálculo del coeficiente de correlación intraclase (ICC). El coeficiente de correlación intraclase (ICC) es un número que se encuentra entre 0 y 1, de acuerdo con Koo and Li (2016) es una medida que refleja tanto el grado de correlación como el grado de concordancia entre mediciones, los valores cercanos a 1 implican consistencia y concordanica, y los valores cercanos a cero todo lo contrario. Hay diferentes tipos de ICC, en este

caso la fórmula del utilizado correspondió a:

$$ICC = \frac{MS_R - MS_W}{MS_R + (k+1)MS_W}, \tag{5-16}$$

donde:

- MS_R= Media cuadrática por filas, se calcula la media por cada una de las observaciones, en este caso para comparar el valor real versus el valor arrojado por el modelo.
- MS_W= Media cuadrática para los residuales y varianza.
- k= número de evaluadores/mediciones, en este caso el número de evaluadores se toma igual a 1, donde el modelo se toma como un evaluador.

El ICC se calculó usando la función icc del paquete irr (Gamer et al., 2019).

Finalmente, como medida de ajuste de los modelos se eligió el criterio de información de cuasi-verosimilitud, QIC, cuya definición se realizó al inicio de esta sección.

En la Tabla 5-16 se presentan las anteriores métricas calculadas para cada uno de los modelos expuestos anteriormente, adicionalmente en la última columna se específica cuál fue el mejor modelo para cada una de las medidas calculadas. En el caso del MSE, este corresponde al cuadrado del RMSE. Las tres primeras métricas corresponden a la calidad del pronóstico, donde es preferido siempre el modelo que presente un menor valor en éstas, el ICC se refiere a la consistencia de las estimaciones, es decir la cercanía que hay entre Y_i y $\hat{Y}_i$, por tanto es preferido un modelo con un mayor ICC y el QIC es para comparar la bondad de ajuste de los modelos, donde el mejor modelo es aquel que tenga un menor QIC.

Tabla 5-16.: Métricas desempeño de los modelos ajustados

Modelo/Métrica	MAE	MSE	MAPE	ICC	QIC
Normal	4.4E+05	2.8E+11	0.594	0.994	6.7E+12
Poisson	3.9E+05	2.4E+11	0.530	0.995	-1.01E+11
Binomial Negativa	4.0.E+05	2.5.E+11	0.536	0.995	1.65E+12
Splines	1.71E+06	3.75E+12	2.28	0.984	1.26E+12
Splines Poisson	2.29E+06	6.56E+12	3.06	0.973	-1.01E+11
Mejor modelo	**Poisson**	**Poisson**	**Poisson**	**Poisson**	**Splines Poisson**

5.7. Modelo final

Como se observó en la sección anterior el modelo que mejor se ajustó a los datos corresponde al modelo de regresión Poisson, pese a la posible presencia de sobre dispersión, el modelo

Poisson continuó presentando mejor desempeño que el modelo de regresión Binomial Negativa. En la Tabla 5-17 se pueden observar los coeficientes estimados, donde todos fueron significativos a un nivel del 5 %. En la Figura 5-3 se pueden observar los valores ajustados por el modelo y los valores reales, la línea azul corresponde a los valores reales y la línea magenta punteada corresponde a los valores ajustados. Se puede observar que el modelo presenta un muy buen grado de ajuste. Los últimos 3 puntos corresponden al pronóstico con los 3 meses que estaban disponibles para pronosticar.

Tabla 5-17.: Resultados estimación final modelo de regresión Poisson con GEE

Coeficiente	Estimación	DE	Wald	P(|Z|>|W|)
β_0	1.76E+01	1.41E-02	1.57E+06	<2.00E-16
β_1	1.50E-07	4.48E-09	1.12E+03	<2.00E-16
β_2	2.80E-08	2.81E-09	9.88E+01	<2.00E-16
β_3	4.05E-11	7.06E-12	3.29E+01	9.70E-09
β_4	6.40E-11	2.35E-11	7.40E+00	0.006
β_5	-2.60E-10	4.91E-11	2.81E+01	1.20E-07

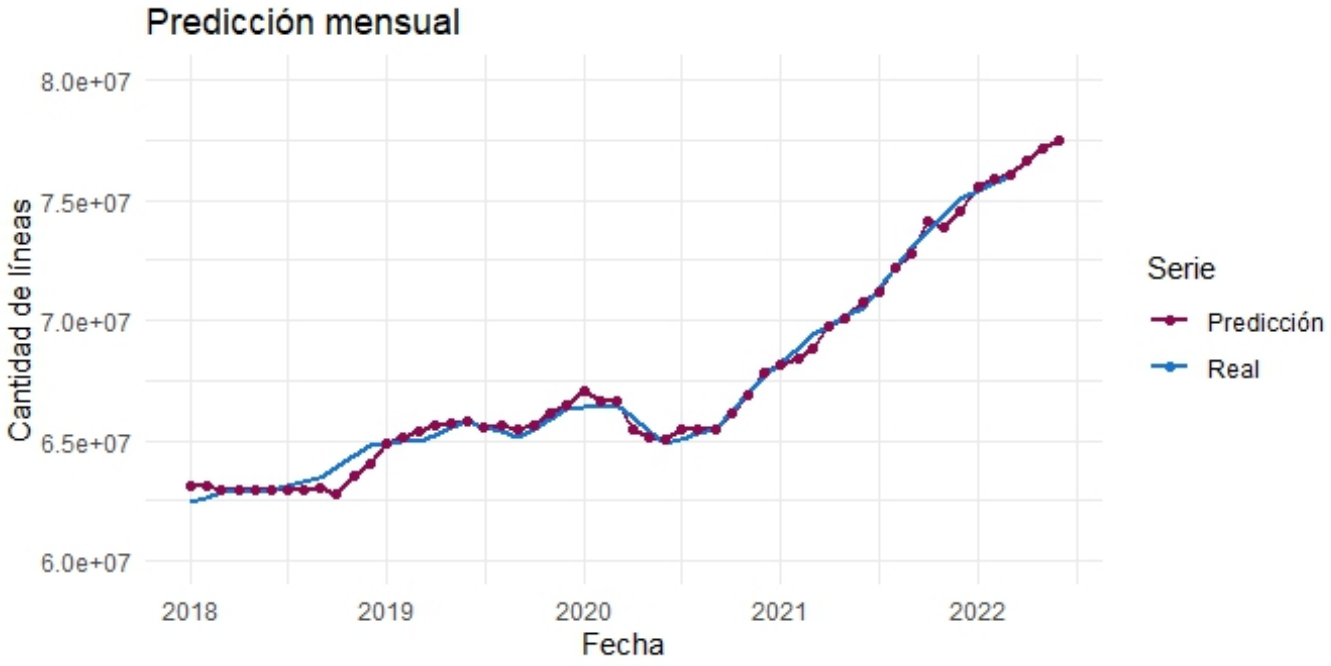

Figura 5-3.: Valores ajustados por el modelo versús valores reales.
Fuente: Elaboración propia.

6. Conclusiones y recomendaciones

6.1. Conclusiones

A pesar de las diferencias en las frecuencias de los datos de la variable respuesta y las variables predictoras, el método de desagregación temporal permitió construir una base de datos en la que todas las variables fueron llevadas a un período mensual, simulando los datos de ésta, logrando obtener valores congruentes con los valores reales, pudiendo de esta manera ser utilizados para todo el entrenamiento, evaluación y selección del modelo.

El método GEE se presenta como una alternativa flexible que permite ajustar modelos lineales generalizados en datos correlacionados, pudiendo así extender su uso a series de tiempo, habilitando la generación de modelos multivariados haciendo uso de este tipo de datos. Adicionalmente, al ser un modelo multivariado se puede utilizar para pronosticar y también para entender relaciones entre variables. Una de las desventajas del método es que la literatura sobre la medición de los residuales y la bondad de ajuste se encuentra aún limitada.

La división de la base de datos en entrenamiento y prueba, permitió comparar estadísticamente los distintos modelos entrenados, pudiendo así sustentar válidamente la elección del modelo final, utilizando una forma moderna y ampliamente aceptada en el aprendizaje estadístico.

Los modelos entrenados presentaron buenas medidas de pronóstico, pudiendo capturar y seguir la tendencia de los datos. El modelo elegido correspondió al modelo de regresión Poisson, el cual mejoró el MAE en un 11 % y el MSE en un 15 % respecto del modelo con respuesta normal, adicionalmente el QIC mejoró significativamente.

6.2. Recomendaciones

Para futuras líneas de trabajo, una de las variables importantes que podría ser incluida es la cantidad de personas que cada año pueden empezar a adquirir líneas telefónicas a su nombre, de esta manera se tendría la inclusión de un índice poblacional que podría dar cuenta del crecimiento esperado por aumento de población activa en este mercado.

Adicionalmente, otra variable que podría entrar en consideración es la cantidad de líneas distintas de otros operadores a las que los usuarios de la entidad propietaria de la información realizan llamadas. Este dato es complejo de obtener por la cantidad de registros contenidos en las bases de datos, donde diariamente se almacenan millones de llamadas, para lograr obtenerlo, se requeriría una optimización de la base de datos en cuanto al almacenamiento y procesamiento de la información, ya que debido a la cantidad de registros, periódicamente se realiza un barrido de la información y no se almacena un histórico amplio.

A. Anexo: Script de R utilizado para la simulación de los datos

```
#####################################
###### SCRIPT SIMULACION DATOS#####
#####################################


library(glme)
library(lme4)
library(ggplot2)
library(janitor)
library(dplyr)
library(readxl)
library(gridExtra)

#Simulating data
#POS
#April 28th 2022


df<-read_xlsx("C:/Users/USUARIO/Documents/Maestria_Estadistica/Tesis/Datos
    /Datos_tesis_series.xlsx")
df$Date<-(1:54)
#pos$lndate<-log(pos$Date)

lambda<-mean(df$USRS_POS)

ggplot(df, aes(x=Date,y=USRS_POS))+
  geom_line(col="dodgerblue3", size=1)

### Fitting a Poisson model with random intercepts

fit_1<-glmer(USRS_POS ~ 1+(1 | Date), family=poisson, data = df)
summary(fit_1)

df$pred<-predict(fit_1, newdata=df, type="response")

ggplot(df, aes(x=Date,y=USRS_POS))+
  geom_line(col="dodgerblue3", size=1)+
```

```
37   geom_line(aes(x=Date,y=pred),col="violetred3", size=1, alpha=0.5)
38
39 df<-cbind(df_comp,df_orig)
40 df$date<-seq(from=as.Date("2018-01-01"), by="month", length.out=51)
41 df<-df%>%dplyr::rename(new_pos=X1, new_pre=X2,
42                                    new_MIN_LLAM_ENT_OTROS_OPER=X3,new_MIN_
     LLAM_SAL_OTROS_OPER=X4,
43                                    new_LLAM_ENT_OTROS_OPER=X5, new_LLAM_SAL_
     OTROS_OPER=X6)
44
45
46 ### What if we used those predicted values to generate a new dataset?
47
48 set.seed(1234)
49
50 #pos$new_pos<-(rpois(24,pos$pred)+rnorm(24,0,20000))
51 df$new_pos<-round((rpois(54,df$pred)+rnorm(1,0,50000)),0)
52
53 plot_pos<-ggplot(df, aes(x=Date,y=USRS_POS))+
54           geom_line(col="dodgerblue3", size=1, linetype="dashed" )+
55           geom_line(aes(x=date,y=new_pos),col="violetred2", size=1)+
56           ggtitle("Usuarios Pos") +
57           theme_minimal()+
58           xlab("Date") + ylab("Usuarios Pos")
59 plot_pos
60
61 ## PRE
62 lambda<-mean(df$USRS_PRE)
63
64 ggplot(df, aes(x=Date,y=USRS_PRE))+
65   geom_line(col="blue", size=1)
66
67 ### Fitting a Poisson model with random intercepts
68
69 fit_1<-glmer(USRS_PRE ~ 1+(1 | Date), family=poisson, data = df)
70 summary(fit_1)
71
72 df$pred<-predict(fit_1, newdata=df, type="response")
73
74 ggplot(df, aes(x=Date,y=USRS_PRE))+
75   geom_line(col="blue", size=1)+
76   geom_line(aes(x=Date,y=pred),col="red", size=1, alpha=0.5)
77
78 ### What if we used those predicted values to generate a new dataset?
79 4
80 set.seed(1234)
81
82 #pos$new_pos<-(rpois(24,pos$pred)+rnorm(24,0,20000))
```

```
df$new_pre<-round((rpois(54,df$USRS_PRE)+rnorm(1,0,300000)),0)
# (rpois(48,df$pred)+rnorm(1,0,50000))
# (rpois(48,df$pred)+rnorm(1,0,50000))
# (rpois(48,bd_completa$pred_usrs_pre)+rnorm(1,0,300000))
plot_pre<-ggplot(df, aes(x=Date,y=USRS_PRE))+
          geom_line(col="dodgerblue3", size=1, linetype="dashed" )+
          geom_line(aes(x=Date,y=new_pre),col="violetred2", size=1)+
          ggtitle("Usuarios Pre") +
          theme_minimal()+
          xlab("Date") + ylab("Usuarios Pre")

plot_pre

##MINUTOS LLAMADAS ENTRANTES OTROS OPERADORES
summary(df$MIN_LLAM_ENT_OTROS_OPER)

df$date<-seq(1:54)

ggplot(df, aes(x=date,y=MIN_LLAM_ENT_OTROS_OPER))+
  geom_line(col="blue", size=1)

# ggplot(df, aes(x=REPORT_MONTH,y=MIN_LLAM_ENT_OTROS_OPER))+
#   geom_line(col="blue", size=1)


### Fitting a normal model with random intercepts

fit_1<-lme(MIN_LLAM_ENT_OTROS_OPER ~ 1,
           random = ~ 1 |  date, data = df,
           control=lmeControl(opt="optim"))

summary(fit_1)

df$pred<-predict(fit_1, newdata=df, type="response")

ggplot(df, aes(x=date,y=MIN_LLAM_ENT_OTROS_OPER))+
  geom_line(col="blue", size=1)+
  geom_line(aes(x=date,y=pred),col="red", size=1, alpha=0.5)


### What if we used those predicted values to generate a new dataset?

set.seed(1234)

df$new_MIN_LLAM_ENT_OTROS_OPER<-(rpois(54,df$pred)+rnorm(1,0,50000))
df$new_MIN_LLAM_ENT_OTROS_OPER<-df$new_MIN_LLAM_ENT_OTROS_OPER+runif(54) #
   para que quede con decimales al ser minutos
summary(df$new_MIN_LLAM_ENT_OTROS_OPER)
```

```
130
131 plot_min_ent<-ggplot(df, aes(x=date,y=MIN_LLAM_ENT_OTROS_OPER))+
132               geom_line(col="dodgerblue3", size=1, linetype="dashed" )+
133               geom_line(aes(x=date,y=new_MIN_LLAM_ENT_OTROS_OPER),col="
    violetred2", size=1)+
134               ggtitle("Minutos entrantes") +
135               theme_minimal()+
136               xlab("Date") + ylab("Minutos entrantes")
137 plot_min_ent
138 ##MINUTOS LLAMADAS SALIENTES OTROS OPERADORES
139
140 summary(df$MIN_LLAM_SAL_OTROS_OPER)
141
142 ggplot(df, aes(x=date,y=MIN_LLAM_SAL_OTROS_OPER))+
143   geom_line(col="blue", size=1)
144
145 ### Fitting a normal model with random intercepts
146
147 fit_1<-lme(MIN_LLAM_SAL_OTROS_OPER ~ 1,
148            random = ~ 1 |  date, data = df,
149            control=lmeControl(opt="optim"))
150 summary(fit_1)
151
152 df$pred<-predict(fit_1, newdata=df, type="response")
153
154 ggplot(df, aes(x=date,y=MIN_LLAM_SAL_OTROS_OPER))+
155   geom_line(col="blue", size=1)+
156   geom_line(aes(x=date,y=pred),col="red", size=1, alpha=0.5)
157
158 ### What if we used those predicted values to generate a new dataset?
159
160 set.seed(1234)
161
162 df$new_MIN_LLAM_SAL_OTROS_OPER<-(rpois(54,df$pred)+rnorm(1,0,2000000))
163 df$new_MIN_LLAM_SAL_OTROS_OPER<-df$new_MIN_LLAM_SAL_OTROS_OPER+runif(54)
164
165 plot_min_sal<-ggplot(df, aes(x=date,y=MIN_LLAM_SAL_OTROS_OPER))+
166               geom_line(col="dodgerblue3", size=1, linetype="dashed" )+
167               geom_line(aes(x=date,y=new_MIN_LLAM_SAL_OTROS_OPER),col="
    violetred2", size=1)+
168               theme_minimal()+
169               ggtitle("Minutos salientes") +
170               xlab("Date") + ylab("Minutos salientes")
171 plot_min_sal
172 ##LLAMADAS SALIENTES OTROS OPERADORES
173
174 lambda<-mean(df$LLAM_SAL_OTROS_OPER)
175 summary(df$LLAM_SAL_OTROS_OPER)
```

```
lambda

ggplot(df, aes(x=date,y=LLAM_SAL_OTROS_OPER))+
  geom_line(col="blue", size=1)

### Fitting a Poisson model with random intercepts

# fit_1<-lme(LLAM_SAL_OTROS_OPER ~ 1,
#            random = ~ 1 |  date, data = df,
#             control=lmeControl(opt="optim"))

fit_1<-glmer(LLAM_SAL_OTROS_OPER ~ 1+(1 | date), family=poisson(link=sqrt)
    , data = df,
             control = glmerControl(optimizer = "bobyqa" ,
                                    optCtrl = list(maxfun= 100000)),
             nAGQ = 0,
             verbose = 1)

summary(fit_1)

df$pred<-predict(fit_1, newdata=df, type="response")

ggplot(df, aes(x=date,y=LLAM_SAL_OTROS_OPER))+
  geom_line(col="blue", size=1)+
  geom_line(aes(x=date,y=pred),col="red", size=1, alpha=0.5)

### What if we used those predicted values to generate a new dataset?

set.seed(1234)

df$new_LLAM_SAL_OTROS_OPER<-round((rpois(54,df$pred)+rnorm(1,0,20000000))
    ,0)

plot_llam_sal<-ggplot(df, aes(x=date,y=LLAM_SAL_OTROS_OPER))+
               geom_line(col="dodgerblue3", size=1, linetype="dashed" )+
               geom_line(aes(x=date,y=new_LLAM_SAL_OTROS_OPER),col="
    violetred2", size=1)+
               ggtitle("Llamadas salientes") +
               theme_minimal()+
               xlab("Date") + ylab("Llamadas salientes")


plot_llam_sal

##LLAMADAS ENTRANTES OTROS OPERADORES

lambda<-mean(df$LLAM_ENT_OTROS_OPER)
summary(df$LLAM_ENT_OTROS_OPER)
```

```
lambda

ggplot(df, aes(x=date,y=LLAM_ENT_OTROS_OPER))+
  geom_line(col="blue", size=1)

### Fitting a Poisson model with random intercepts

# fit_1<-lme(LLAM_SAL_OTROS_OPER ~ 1,
#            random = ~ 1 |  date, data = df,
#             control=lmeControl(opt="optim"))

fit_1<-glmer(LLAM_ENT_OTROS_OPER ~ 1+(1 | date), family=poisson(link=sqrt)
    , data = df,
             control = glmerControl(optimizer = "bobyqa" ,
                                    optCtrl = list(maxfun= 100000)),
             nAGQ = 0,
             verbose = 1)

summary(fit_1)

df$pred<-predict(fit_1, newdata=df, type="response")

ggplot(df, aes(x=date,y=LLAM_ENT_OTROS_OPER))+
  geom_line(col="blue", size=1)+
  geom_line(aes(x=date,y=pred),col="red", size=1, alpha=0.5)

### What if we used those predicted values to generate a new dataset?

set.seed(1234)

df$new_LLAM_ENT_OTROS_OPER<-round((rpois(54,df$pred)+rnorm(1,0,20000000))
    ,0)

plot_llam_ent<-ggplot(df, aes(x=date,y=LLAM_ENT_OTROS_OPER))+
               geom_line(col="dodgerblue3", size=1, linetype="dashed" )+
               geom_line(aes(x=date,y=new_LLAM_ENT_OTROS_OPER),col="
    violetred2", size=1)+
               ggtitle("Llamadas entrantes") +
               theme_minimal()+
               xlab("Date") + ylab("Llamadas entrantes")
plot_llam_ent

grid.arrange(plot_pos,plot_pre,plot_llam_ent,plot_llam_sal, plot_min_ent,
             plot_min_sal, ncol = 3, nrow=2
             )

grid.arrange(plot_pos,plot_pre, ncol = 2, nrow=1
)
```

```
# layout_matrix = cbind(c(1,1,1), c(2,3,4))
unique(df$LLAM_ENT_OTROS_OPER==df$new_LLAM_ENT_OTROS_OPER)
unique(df$LLAM_SAL_OTROS_OPER==df$new_LLAM_SAL_OTROS_OPER)
unique(df$MIN_LLAM_ENT_OTROS_OPER==df$new_MIN_LLAM_ENT_OTROS_OPER)
unique(df$new_MIN_LLAM_SAL_OTROS_OPER==df$MIN_LLAM_SAL_OTROS_OPER)
unique(df$USRS_PRE==df$new_pre)
unique(df$new_pos==df$USRS_POS)

##############################################
#########GUARDANDO LA BASE FINAL##############
##############################################

#Variables Mintic
df_mintic<-read.csv('C:/Users/USUARIO/Documents/Maestria_Estadistica/Tesis
    /Datos/220806_Mintic_m vil_mensual.csv', sep=";")
df_mintic<-data.frame(Movil_total=df_mintic[3:78,])
df_mintic$MES<-seq(from=as.Date("2015-12-01"), by="month", length.out=76)
df_mintic<-df_mintic[26:76,]

rownames(df_mintic)<-NULL
##Guardando las variables simuladas

df_sim<-df%>%dplyr::select(REPORT_MONTH,new_pos,new_pre,new_MIN_LLAM_ENT_
    OTROS_OPER,
                           new_MIN_LLAM_SAL_OTROS_OPER,new_LLAM_ENT_OTROS_
    OPER,new_LLAM_SAL_OTROS_OPER)

##TOMANDO LA BASE COMPLETA
df_comp<-cbind(Movil_total=round(df_mintic$Movil_total,0),df[1:51,])

df_comp<-df_comp%>%dplyr::relocate(REPORT_MONTH)
df_comp<-clean_names(df_comp)
write.table(df_comp,'C:/Users/USUARIO/Documents/Maestria_Estadistica/Tesis
    /Datos/BD_FIN_MOD_TESIS.csv', sep=";", dec=',',  row.names = FALSE)

df_sim_pron<-df_sim[52:54,]
df_sim_pron<-clean_names(df_sim_pron)
write.table(df_sim_pron,'C:/Users/USUARIO/Documents/Maestria_Estadistica/
    Tesis/Datos/220807_bd_pron.csv', sep=";", dec=',', row.names = FALSE)


write.table(df_orig,'C:/Users/USUARIO/Documents/Maestria_Estadistica/Tesis
    /Datos/230312_BD_VBLES_ORIG.csv', sep=";", dec=',', row.names = FALSE)

#se cargan los archivos con las variables simuladas
df_comp<-read.csv('C:/Users/USUARIO/Documents/Maestria_Estadistica/Tesis/
    Datos/BD_FIN_MOD_TESIS.csv', sep=";", dec=',')

```

```
df_comp<-df_comp%>%dplyr::rename(Y=movil_total, X1=new_pos, X2=new_pre,
                                 X3=new_min_llam_ent_otros_oper,X4=new_min_llam_sal_otros_oper,
                                 X5= new_llam_ent_otros_oper, X6=new_llam_sal_otros_oper )
#se cargan los datos originales
df_orig<-read.csv('C:/Users/USUARIO/Documents/Maestria_Estadistica/Tesis/Datos/230312_BD_VBLES_ORIG.csv', sep=";", dec=',')

#Se calculan las correlaciones para complementar el an lisis gr fico
cor_1<-cor.test(df_comp$Y, df_comp$X1, method = "spearman", alternative = "g")
cor_11<-cor.test(df_orig$Movil_total, df_orig$USRS_POS, method = "spearman", alternative = "g")

cor_2<-cor.test(df_comp$Y, df_comp$X2, method = "spearman", alternative = "g")
cor_22<-cor.test(df_orig$Movil_total, df_orig$USRS_PRE, method = "spearman", alternative = "g")

cor_3<-cor.test(df_comp$Y, df_comp$X3, method = "spearman", alternative = "g")
cor_33<-cor.test(df_orig$Movil_total, df_orig$MIN_LLAM_ENT_OTROS_OPER, method = "spearman", alternative = "g")

cor_4<-cor.test(df_comp$Y, df_comp$X4, method = "spearman", alternative = "g")
cor_44<-cor.test(df_orig$Movil_total, df_orig$MIN_LLAM_SAL_OTROS_OPER, method = "spearman", alternative = "g")

cor_5<-cor.test(df_comp$Y, df_comp$X5, method = "spearman", alternative = "g")
cor_55<-cor.test(df_orig$Movil_total, df_orig$LLAM_ENT_OTROS_OPER, method = "spearman", alternative = "g")

cor_6<-cor.test(df_comp$Y, df_comp$X6, method = "spearman", alternative = "g")
cor_66<-cor.test(df_orig$Movil_total, df_orig$LLAM_SAL_OTROS_OPER, method = "spearman", alternative = "g")

df_corr<-data.frame(Variable='Usuarios Pospago', Correlaci n_vble_original=cor_11[4],p_valor_asociado=cor_11[3],
                    correlaci n_vble_simulada= cor_1[4],p_valor_asociado=cor_1[3])
unique(df_comp$X1==df_orig$USRS_POS)
df_corr<-rbind(df_corr, data.frame(Variable='Usuarios Prepago', estimate=cor_22[4],p.value=cor_22[3],
                                   estimate.1=cor_2[4],p.value.1=cor_2[3])
```

```
    )
#Se comprueban que las series no son exactamente iguales
unique(df_comp$X2==df_orig$USRS_PRE)
df_corr<-rbind(df_corr, data.frame(Variable='Minutos entrantes de otros
    operadores', estimate=cor_33[4],p.value=cor_33[3],
                                    estimate.1=cor_3[4],p.value.1=cor_3[3])
    )
unique(df_comp$X3==df_orig$MIN_LLAM_ENT_OTROS_OPER)
df_corr<-rbind(df_corr, data.frame(Variable='Minutos salientes a otros
    operadores', estimate=cor_44[4],p.value=cor_44[3],
                                    estimate.1=cor_4[4],p.value.1=cor_4[3])
    )
unique(df_comp$X4==df_orig$MIN_LLAM_SAL_OTROS_OPER)
df_corr<-rbind(df_corr, data.frame(Variable='Llamadas entrantes de otros
    operadores', estimate=cor_55[4],p.value=cor_55[3],
                                    estimate.1=cor_5[4],p.value.1=cor_5[3])
    )
unique(df_comp$X5==df_orig$LLAM_ENT_OTROS_OPER)
df_corr<-rbind(df_corr, data.frame(Variable='Llamadas salientes a otros
    operadores', estimate=cor_66[4],p.value=cor_66[3],
                                    estimate.1=cor_6[4],p.value.1=cor_6[3])
    )
unique(df_comp$X6==df_orig$LLAM_SAL_OTROS_OPER)
df_comp<-cbind(df_comp, df_orig)
write.csv(df_corr,'C:/Users/USUARIO/Documents/Maestria_Estadistica/Tesis/
    Datos/Comp_corr.csv')

#Se agrega el ICC para calcular la cercania entre las series simuladas y
    las reales
irr_pos<-irr::icc( df_comp[,c("X1", "USRS_POS")], model = "oneway",
         type = "consistency", unit = "single")
irr_pos<-irr_pos$value
irr_pre<-irr::icc( df_comp[,c("X2", "USRS_PRE")], model = "oneway",
                   type = "consistency", unit = "single")
irr_pre<-irr_pre$value
irr_min_ent<-irr::icc( df_comp[,c("X3", "MIN_LLAM_ENT_OTROS_OPER")], model
    = "oneway",
                   type = "consistency", unit = "single")
irr_min_ent<-irr_min_ent$value
irr_min_sal<-irr::icc( df_comp[,c("X4", "MIN_LLAM_SAL_OTROS_OPER")], model
    = "oneway",
                       type = "consistency", unit = "single")
irr_min_sal<-irr_min_sal$value

irr_llam_ent<-irr::icc( df_comp[,c("X5", "LLAM_ENT_OTROS_OPER")], model =
    "oneway",
                       type = "consistency", unit = "single")
irr_llam_ent<-irr_llam_ent$value
```

```
371
372 irr_llam_sal<-irr::icc( df_comp[,c("X6", "LLAM_SAL_OTROS_OPER")], model =
    "oneway",
373                          type = "consistency", unit = "single")
374 irr_llam_sal<-irr_llam_sal$value
375 vec_irr<-c(irr_pos,irr_pre,irr_min_ent,irr_min_sal,irr_llam_ent, irr_llam_
    sal)
376
377 cor_v1<-cor.test(df_comp$X1, df_orig$USRS_POS, method = "spearman",
    alternative = "g")
378 cor_v2<-cor.test(df_comp$X2, df_orig$USRS_PRE, method = "spearman",
    alternative = "g")
379 cor_v3<-cor.test(df_comp$X3, df_orig$MIN_LLAM_ENT_OTROS_OPER, method = "
    spearman", alternative = "g")
380 cor_v4<-cor.test(df_comp$X4, df_orig$MIN_LLAM_SAL_OTROS_OPER, method = "
    spearman", alternative = "g")
381 cor_v5<-cor.test(df_comp$X5, df_orig$LLAM_ENT_OTROS_OPER, method = "
    spearman", alternative = "g")
382 cor_v6<-cor.test(df_comp$X6, df_orig$LLAM_SAL_OTROS_OPER, method = "
    spearman", alternative = "g")
383
384 df_corr_vbles<-data.frame(Variable='Usuarios Pospago', Valor_corr_sp=cor_
    v1[4], p_valor=cor_v1[3])
385 df_corr_vbles<-rbind(df_corr_vbles, data.frame(Variable='Usuarios Prepago'
    , Valor_corr_sp=cor_v2[4], p_valor=cor_v2[3]))
386 df_corr_vbles<-rbind(df_corr_vbles, data.frame(Variable='Minutos entrantes
     de otros operadores', Valor_corr_sp=cor_v3[4], p_valor=cor_v3[3]))
387 df_corr_vbles<-rbind(df_corr_vbles, data.frame(Variable='Minutos salientes
     a otros operadores', Valor_corr_sp=cor_v4[4], p_valor=cor_v4[3]))
388 df_corr_vbles<-rbind(df_corr_vbles, data.frame(Variable='Llamadas
    entrantes de otros operadores', Valor_corr_sp=cor_v5[4], p_valor=cor_v5
    [3]))
389 df_corr_vbles<-rbind(df_corr_vbles, data.frame(Variable='Llamadas
    salientes a otros operadores', Valor_corr_sp=cor_v6[4], p_valor=cor_v6
    [3]))
390 df_corr_vbles$ICC<-vec_irr
391 print(xtable(df_corr_vbles), include.rownames = FALSE)
```

Listing A.1: Script simulación datos

B. Anexo:Script de R para desagregación temporal

```
##################################################
######DESAGREGACION VARIABLE RESPUESTA###########
##################################################

library(tempdisagg)
library(readxl)
library(rtweet)
library(tsbox)
proy<- read_excel("C:/Users/USUARIO/Documents/Maestria_Estadistica/Tesis/
    Datos/Mintic_movil_trim.xlsx")
serie_mintic_trimestral <- ts((proy$Total_movil),start=c(2015,4),frequency
    = 4)
mod3 <- td(serie_mintic_trimestral ~ 1,conversion = "last", to = "monthly"
    , method = "denton-cholette")
serie_mintic_mensual<-predict(mod3)

ts_plot(
  ts_scale(
    ts_c(serie_mintic_mensual, serie_mintic_trimestral)
  ),
  title = "Serie trimestral Mintic desagregada"
)
write.xlsx(x, file, sheetName="Sheet1", col.names=TRUE, row.names=TRUE,
    append=FALSE)
y<-data.frame(proy_d)
write.table(y, "C:/Users/USUARIO/Documents/Maestria_Estadistica/Tesis/
    Datos/220806_Mintic_movil_mensual.csv",sep = ";", dec = ",",row.names =
    FALSE)
```

Listing B.1: Script desagregación temporal

C. Anexo: Script de R para análisis exploratorio de datos

```r
####################################
########### SCRIPT EDA ############
####################################

#Graficos
library(GGally)
library(ggplot2)
library("gridExtra")
library(gamlss)
library("fitdistrplus")
library(ggpubr)
library(mgcv)
library(DescTools)
library(tsibble)
library(dplyr)
library(feasts)
library(DCCA)
library(forecast)
options(scipen = 0)
df_comp<-read.csv('C:/Users/USUARIO/Documents/Maestria_Estadistica/Tesis/Datos/BD_FIN_MOD_TESIS.csv', sep=";", dec=',')
names(df_comp)
df_comp$Date<-seq(from=as.Date("2018-01-01"), by="month", length.out=51)
#Se renombran las variables como Y y X's
# asi:
# Y="movil_total"
# X1="new_pos"
# X2= "new_pre"
# X3= "new_min_llam_ent_otros_oper"
# X4= "new_min_llam_sal_otros_oper"
# X5= "new_llam_ent_otros_oper"
# X6="new_llam_sal_otros_oper"

df_comp<-df_comp%>%dplyr::rename(Y=movil_total, X1=new_pos, X2=new_pre,
                                 X3=new_min_llam_ent_otros_oper,X4=new_min_llam_sal_otros_oper,
                                 X5= new_llam_ent_otros_oper, X6=new_llam_
```

```
    sal_otros_oper )

df_comp$Year<-as.numeric(format(df_comp$Date,'%Y'))

##Analizando la distribucion de la variable respuesta

graph_y<-ggplot(df_comp, aes( y=Y, x=Date)) +
  geom_line(color="dodgerblue3", size=1) +
  theme_minimal()

ggpairs(df_comp[,2:8],
        lower = list(continuous = wrap(ggally_points, color = "dodgerblue3
    ")),
        upper = list(continuous = wrap("cor", method = "spearman")),
        diag = list(continuous = wrap('densityDiag', colour = "blue"))) +
        theme(axis.text.x = element_text(angle = 30))

df_comp%>%
  gg_season(Y)
gg_season(as_tsibble(df_comp$Y))


df_graph<- df_comp %>%
  mutate(Month = tsibble::yearmonth(Date)) %>%
  as_tsibble(index = Month) %>%
  dplyr::select(Month,Y,X1, X2, X3, X4, X5, X6)

print(df_graph)
ggseason_y<-df_graph %>% gg_season(Y)+
  theme_minimal()

acf_y<-ggAcf(
  df_comp$Y,
  lag.max = NULL,
  type = c("correlation"),
  plot = TRUE,
  na.action = na.contiguous,
  demean = TRUE
)+theme_minimal()


pacf_y<-ggPacf(
  df_comp$Y,
  lag.max = NULL,
  plot = TRUE,
  na.action = na.contiguous,
  demean = TRUE
)+theme_minimal()
```

```
82
83
84
85 library(patchwork)
86 (graph_y+ggseason_y) / ((acf_y | pacf_y))
87 ##Analizando el comportamiento de las variables individualmente
88 ##Preguntar al profe s? el boxplot es adecuado para series de tiempo
89 bxp1 <- ggplot(df_comp, aes( y=Y, x=as.factor(Year))) +
90   geom_boxplot(color="dodgerblue3", fill = "#3366FF", alpha = 0.5) +
91   theme_minimal()+labs(x = 'Year')
92 bxp2 <- ggplot(df_comp, aes( y=X1,x=as.factor(Year))) +
93   geom_boxplot(color="dodgerblue3", fill = "#3366FF", alpha = 0.5) +
94   labs(y =expression(X["1"]))+
95   theme_minimal()+labs(x = 'Year')
96 bxp3 <- ggplot(df_comp, aes( y=X2,x=as.factor(Year))) +
97   geom_boxplot(color="dodgerblue3", fill = "#3366FF", alpha = 0.5) +
98   labs(y =expression(X["2"]))+
99   theme_minimal()+labs(x = 'Year')
100 bxp4 <- ggplot(df_comp, aes( y=X3,x=as.factor(Year))) +
101   geom_boxplot(color="dodgerblue3", fill = "#3366FF", alpha = 0.5) +
102   labs(y =expression(X["3"]))+
103   theme_minimal()+labs(x = 'Year')
104 bxp5 <- ggplot(df_comp, aes( y=X4,x=as.factor(Year))) +
105   geom_boxplot(color="dodgerblue3", fill = "#3366FF", alpha = 0.5) +
106   labs(y =expression(X["4"]))+
107   theme_minimal()+labs(x = 'Year')
108 bxp6 <- ggplot(df_comp, aes( y=X5,x=as.factor(Year))) +
109   geom_boxplot(color="dodgerblue3", fill = "#3366FF", alpha = 0.5) +
110   labs(y =expression(X["5"]))+
111   theme_minimal()+labs(x = 'Year')
112 bxp7 <- ggplot(df_comp, aes( y=X6,x=as.factor(Year))) +
113   geom_boxplot(color="dodgerblue3", fill = "#3366FF", alpha = 0.5) +
114   labs(y =expression(X["6"]))+
115   theme_minimal()+labs(x = 'Year')
116
117 ggarrange( bxp1,bxp2,bxp3,bxp4,bxp5,bxp6,bxp7,
118            ncol = 3, nrow = 3)
119 ggarrange(bxp2,bxp3,bxp4,bxp5,bxp6,bxp7, ncol = 3, # Second row with box
    and dot plots
120           nrow = 2
121                                       # Labels of the scatter plot
122 )
123
124 grid.arrange(bxp1,                                # First row with one plot
    spaning over 2 columns
125              arrangeGrob(bxp2,bxp3,bxp4,bxp5,bxp6,bxp7, ncol = 3), #
    Second row with 2 plots in 2 different columns
126              nrow = 2)
```

```
127
128 ##Graficando cada una de las covariables contra el tiempo
129 graf_x_1<-ggplot(df_comp, aes( y=X1, x=Date)) +
130           geom_line(color="dodgerblue3", size=1) +
131           labs(y =expression(X["1"]))+
132           theme_minimal()
133 graf_x_2<-ggplot(df_comp, aes( y=X2, x=Date)) +
134           geom_line(color="dodgerblue3", size=1) +
135           labs(y =expression(X["2"]))+
136           theme_minimal()
137 graf_x_3<-ggplot(df_comp, aes( y=X3, x=Date)) +
138           geom_line(color="dodgerblue3", size=1) +
139           labs(y =expression(X["3"]))+
140           theme_minimal()
141 graf_x_4<-ggplot(df_comp, aes( y=X4, x=Date)) +
142           geom_line(color="dodgerblue3", size=1) +
143           labs(y =expression(X["4"]))+
144           theme_minimal()
145 graf_x_5<-ggplot(df_comp, aes( y=X5, x=Date)) +
146           geom_line(color="dodgerblue3", size=1) +
147           labs(y =expression(X["5"]))+
148           theme_minimal()
149 graf_x_6<-ggplot(df_comp, aes( y=X6, x=Date)) +
150           geom_line(color="dodgerblue3", size=1) +
151           labs(y =expression(X["6"]))+
152           theme_minimal()
153
154 grid.arrange(graf_x_1,graf_x_2,graf_x_3,graf_x_4,graf_x_5,graf_x_6, #
    Second row with 2 plots in 2 different columns
155              ncol=3, nrow=2)
156 #Se construyen los correlogramas y las lineas para cada serie
157 ggseason_y<-df_graph %>% gg_season(X1)+
158   theme_minimal()
159
160 acf_x1<-ggAcf(
161   df_comp$X1,
162   lag.max = NULL,
163   type = c("correlation"),
164   plot = TRUE,
165   na.action = na.contiguous,
166   demean = TRUE
167 )+theme_minimal()
168
169 pacf_x1<-ggPacf(
170   df_comp$X1,
171   lag.max = NULL,
172   plot = TRUE,
173   na.action = na.contiguous,
```

```
  demean = TRUE
)+theme_minimal()

ggseason_x2<-df_graph %>% gg_season(X2)+
  theme_minimal()

acf_x2<-ggAcf(
  df_comp$X2,
  lag.max = NULL,
  type = c("correlation"),
  plot = TRUE,
  na.action = na.contiguous,
  demean = TRUE
)+theme_minimal()

pacf_x2<-ggPacf(
  df_comp$X2,
  lag.max = NULL,
  plot = TRUE,
  na.action = na.contiguous,
  demean = TRUE
)+theme_minimal()

ggseason_x3<-df_graph %>% gg_season(X3)+
  theme_minimal()

acf_x3<-ggAcf(
  df_comp$X3,
  lag.max = NULL,
  type = c("correlation"),
  plot = TRUE,
  na.action = na.contiguous,
  demean = TRUE
)+theme_minimal()

pacf_x3<-ggPacf(
  df_comp$X3,
  lag.max = NULL,
  plot = TRUE,
  na.action = na.contiguous,
  demean = TRUE
)+theme_minimal()


ggseason_x4<-df_graph %>% gg_season(X4)+
  theme_minimal()

acf_x4<-ggAcf(
```

```
  df_comp$X4,
  lag.max = NULL,
  type = c("correlation"),
  plot = TRUE,
  na.action = na.contiguous,
  demean = TRUE
)+theme_minimal()

pacf_x4<-ggPacf(
  df_comp$X4,
  lag.max = NULL,
  plot = TRUE,
  na.action = na.contiguous,
  demean = TRUE
)+theme_minimal()

ggseason_x5<-df_graph %>% gg_season(X5)+
  theme_minimal()

acf_x5<-ggAcf(
  df_comp$X5,
  lag.max = NULL,
  type = c("correlation"),
  plot = TRUE,
  na.action = na.contiguous,
  demean = TRUE
)+theme_minimal()

pacf_x5<-ggPacf(
  df_comp$X5,
  lag.max = NULL,
  plot = TRUE,
  na.action = na.contiguous,
  demean = TRUE
)+theme_minimal()

ggseason_x6<-df_graph %>% gg_season(X6)+
  theme_minimal()

acf_x6<-ggAcf(
  df_comp$X6,
  lag.max = NULL,
  type = c("correlation"),
  plot = TRUE,
  na.action = na.contiguous,
  demean = TRUE
)+theme_minimal()

```

```
pacf_x6<-ggPacf(
  df_comp$X6,
  lag.max = NULL,
  plot = TRUE,
  na.action = na.contiguous,
  demean = TRUE
)+theme_minimal()

(graf_x_1+ggseason_x1) / ((acf_x1 | pacf_x1))
(graf_x_2+ggseason_x2) / ((acf_x2 | pacf_x2))
(graf_x_3+ggseason_x3) / ((acf_x3 | pacf_x3))
(graf_x_4+ggseason_x4) / ((acf_x4 | pacf_x4))
(graf_x_5+ggseason_x5) / ((acf_x5 | pacf_x5))
(graf_x_6+ggseason_x6) / ((acf_x6 | pacf_x6))



##Al igual que la variable respuesta, las covariables vienen presentando un comportamiento creciente en el tiempo

##SCATTER PLOTS
scatt_1<-ggplot(df_comp, aes(x=X1, y=Y)) +
  geom_point(color="dodgerblue3", fill = "#3366FF", alpha = 0.9) +
  geom_smooth(method = "lm", color='violetred3',level = 0.8)+
  geom_smooth(level = 0.8, alpha = 0.3)+
  theme_minimal()+
  scale_x_continuous(breaks = seq(1233285, 2178743, 236365),
                     limits=c(1233285, 2178743),labels = function(x) format(x, scientific = TRUE))+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
scatt_2<-ggplot(df_comp, aes(x=X2, y=Y)) +
  geom_point(color="dodgerblue3", fill = "#3366FF", alpha = 0.9) +
  geom_smooth(method = "lm", color='violetred3',level = 0.8)+
  geom_smooth(level = 0.8, alpha = 0.3)+
  theme_minimal()+
  scale_x_continuous(breaks = seq( 6289978, 8763811, 618458),
                     limits=c(6289978, 8763811),labels = function(x) format(x, scientific = TRUE))+
  theme(axis.text.x = element_text(angle = 35, hjust = 1))
scatt_3<-ggplot(df_comp, aes(x=X3, y=Y)) +
  geom_point(color="dodgerblue3", fill = "#3366FF", alpha = 0.9) +
  geom_smooth(method = "lm", color='violetred3',level = 0.8)+
  geom_smooth(level = 0.8, alpha = 0.3)+
  theme_minimal()+
  scale_x_continuous(breaks = seq( 619719153, 1526676350, 226739299),
                     limits=c(619719153, 1526676350),labels = function(x) format(x, scientific = TRUE))+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
scatt_4<-ggplot(df_comp, aes(x=X4, y=Y)) +
  geom_point(color="dodgerblue3", fill = "#3366FF", alpha = 0.9) +
  geom_smooth(method = "lm", color='violetred3',level = 0.8)+
  geom_smooth(level = 0.8, alpha = 0.3)+
  theme_minimal()+
  scale_x_continuous(breaks = seq( 657722661, 1247176199, 147363385),
                     limits=c(657722661, 1247176199),labels = function(x)
    format(x, scientific = TRUE))+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
scatt_5<-ggplot(df_comp, aes(x=X5, y=Y)) +
  geom_point(color="dodgerblue3", fill = "#3366FF", alpha = 0.9) +
  geom_smooth(method = "lm", color='violetred3',level = 0.8)+
  geom_smooth(level = 0.8, alpha = 0.3)+
  theme_minimal()+
  scale_x_continuous(breaks = seq( 225249010, 597570038, 93080257),
                     limits=c(225249010, 597570038),labels = function(x)
    format(x, scientific = TRUE))+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
scatt_6<-ggplot(df_comp, aes(x=X6, y=Y)) +
  geom_point(color="dodgerblue3", fill = "#3366FF", alpha = 0.9) +
  geom_smooth(method = "lm", color='violetred3',level = 0.8)+
  geom_smooth(level = 0.8, alpha = 0.3)+
  theme_minimal()+
  scale_x_continuous(breaks = seq( 230715767, 442842323, 53031639),
                     limits=c(230715767, 442842323),labels = function(x)
    format(x, scientific = TRUE))+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

ggarrange( scatt_1,scatt_2,scatt_3,scatt_4,scatt_5,scatt_6,
           ncol = 3, nrow = 2)

cor(df_comp[,2:8], method = 'spearman')[,1]
# df_corr_1<-data.frame(lag(df_comp$X1),lag(df_comp$X2),lag(df_comp$X3),
#                       lag(df_comp$X4),lag(df_comp$X5),lag(df_comp$X6))
# df_corr_1<-na.omit(df_corr_1)
df_corr_1<-df_comp[2:51,2:8]
df_corr_1$Y<-df_comp$Y[1:50]
df_corr_1<-df_corr_1%>%dplyr::relocate(Y)
cor(df_corr_1, method = 'spearman')[,1]
#creando rezagos para analizar la correlacion

df_corr_2<-df_comp[3:51,2:8]
df_corr_2$Y<-df_comp$Y[1:49]
df_corr_2<-df_corr_2%>%dplyr::relocate(Y)
cor(df_corr_2, method = 'spearman')[,1]

df_corr_3<-df_comp[4:51,2:8]
#df_corr_3<-na.omit(df_corr_3)
```

```
df_corr_3$Y<-df_comp$Y[1:48]
df_corr_3<-df_corr_3%>%dplyr::relocate(Y)


df_corr_all<-data.frame(cor(df_comp[,2:8], method = 'spearman')[,1],
                        cor(df_corr_1, method = 'spearman')[,1],
                        cor(df_corr_2, method = 'spearman')[,1],
                        cor(df_corr_3, method = 'spearman')[,1]
                        )

names(df_corr_all)<-c('Corr_cont', 'Corr_rez1','Corr_rez2','Corr_rez3')
write.csv(df_corr_all,'C:/Users/USUARIO/Documents/Maestria_Estadistica/
    Tesis/Datos/corr_rez.csv')


list_variables<-c('X1','X2','X3','X4','X5','X6')
list_cor<-list()
for (i in list_variables ) {
  coef_dcca<-rhodcca(df_comp$Y/1000000, df_comp[, i]/1000000)[4]
  list_cor<-append(list_cor, coef_dcca)
}

df_dcca_cor<-as.data.frame(unlist(list_cor))

list_cor1<-list()
for (i in list_variables )
        {
  coef_dcca<-rhodcca(df_corr_1$Y/1000000, df_corr_1[, i]/1000000)[4]
  list_cor1<-append(list_cor1, coef_dcca)
}

df_dcca_cor_1<-as.data.frame(unlist(list_cor1))

list_cor2<-list()
for (i in list_variables ) {
  coef_dcca<-rhodcca(df_corr_2$Y/1000000, df_corr_2[, i]/1000000)[4]
  list_cor2<-append(list_cor2, coef_dcca)
}

df_dcca_cor_2<-as.data.frame(unlist(list_cor2))

list_cor3<-list()
for (i in list_variables ) {
  coef_dcca<-rhodcca(df_corr_3$Y/1000000, df_corr_3[, i]/1000000)[4]
  list_cor3<-append(list_cor3, coef_dcca)
}

df_dcca_cor_3<-as.data.frame(unlist(list_cor3))
```

```
406
407 df_rez_cov_fin<-cbind(df_dcca_cor,df_dcca_cor_1,df_dcca_cor_2,df_dcca_cor_
    3)
408 df_rez_cov_fin$Variable<-list_variables
409 df_rez_cov_fin<-df_rez_cov_fin%>%dplyr::relocate(Variable)
410
411 xtable::xtable(df_rez_cov_fin)
```

Listing C.1: Script Análisis exploratorio

D. Anexo:Script de R para ajuste de modelos

```r
####################################
###########MODELOS#################
####################################

library(rsq)
library(gee)
library(geepack)
library(MESS)
library(gam)
library(DescTools)
library(splines)
library(mgcv)
library(ggplot2)
library(reshape)
library(dplyr)
library("irr")
library(ggpubr)
library(patchwork)

##FUNCIONES PARA CALCULAR EL ERROR.

MAE <- function(real, pred){
  mean( abs (real-pred) )
}
MSE <- function(real, pred){
  mean( (real-pred)^2 )
}
MAPE <- function(real, pred){
  100 * mean( abs( (real-pred)/real ) )
}


#options(scipen = 0)
df_comp<-read.csv('C:/Users/USUARIO/Documents/Maestria_Estadistica/Tesis/
    Datos/BD_FIN_MOD_TESIS.csv', sep=";", dec=',')
names(df_comp)
df_comp$Date<-seq(from=as.Date("2018-01-01"), by="month", length.out=51)
```

```r
#Se renombran las variables como Y y X's
# asi:
# Y="movil_total"
# X1="new_pos"
# X2= "new_pre"
# X3= "new_min_llam_ent_otros_oper"
# X4= "new_min_llam_sal_otros_oper"
# X5= "new_llam_ent_otros_oper"
# X6="new_llam_sal_otros_oper"
df_comp<-df_comp%>%dplyr::rename(Y=movil_total, X1=new_pos, X2=new_pre,
                                 X3=new_min_llam_ent_otros_oper,X4=new_min_llam_sal_otros_oper,
                                 X5= new_llam_ent_otros_oper, X6=new_llam_sal_otros_oper )
# En los graficos del EDA se puede observar que la distribucion de la variable respuesta no es necesariamente normal
# Adicionalmente la relacion entre las covariables y la variable respuesta no es estrictamente lineal
# Por tanto se entrenan 3 modelos, buscando comparar y obtener el mejor resultado posible


set.seed(1234)
#Dado que el horizonte de pronostico es habitualmente 6 meses se dejan 6 meses para test (aproximadamente el 12% de los datos)

Train<-df_comp[1:45,]
Test<-df_comp[46:51,]

##MODELO LINEAL NORMAL
mod_normal_gee_sat<-geeglm(Y ~ X1+X2+X3+X4+X5+X6,family =gaussian,id=report_month , corstr ="exchangeable",  data=Train)
summary(mod_normal_gee_sat)
format(QIC(mod_normal_gee_sat)[1], scientific=FALSE)
mod_normal_gee<-geeglm(Y ~ X1+X2+X4+X5+X6,family =gaussian,id=report_month , corstr ="exchangeable",  data=Train)
summary(mod_normal_gee)
format(QIC(mod_normal_gee)[1], scientific=FALSE)

pred_normal_gee<-predict(mod_normal_gee, newdata=Test, type = "response")
rsq_normal_mod<-rsq(mod_normal_gee)
mae_normal_mod<-MAE(Test$Y,pred_normal_gee)
mse_normal_mod<-MSE(Test$Y,pred_normal_gee)
mape_normal_mod<-MAPE(Test$Y,pred_normal_gee)
theil_normal_mod<-TheilU(Test$Y, pred_normal_gee, type = c(2, 1), na.rm = FALSE)
QIC_normal_mod<-QIC(mod_normal_gee)[1]
dfpred1<-data.frame(Y=as.numeric(df_comp[1:51,c(2)]))
```

```
dfpred1$Fecha<-seq(from=as.Date("2018-01-01"), by="month", length.out=51)
dfpred1$prediction<-''
dfpred1$prediction[1:45]<-as.numeric(round(fitted(mod_normal_gee),0))
dfpred1$prediction[46:51]<-as.numeric(round(pred_normal_gee,0))
dfpred1$prediction<-as.numeric(dfpred1$prediction)
irr::icc(dfpred1[,c(1,3)], model = "oneway",
    type = "consistency", unit = "single")
bd_pron11 <- melt(dfpred1, id.vars = "Fecha")
bd_pron11<-bd_pron11%>%dplyr::filter(value!='')
names(bd_pron11)[2]<-'Serie'
names(bd_pron11)[3]<-'Cantidad de l neas'
bd_pron11$`Cantidad de l neas `<-as.numeric(bd_pron11$`Cantidad de l neas
    `)
bd_pron11$Serie<-ifelse(bd_pron11$Serie=='Y', 'Real', 'Predicci n')

graf_11<-ggplot(bd_pron11, aes(x=Fecha, y=`Cantidad de l neas `, colour =
    Serie))  +
  scale_y_continuous(limits = c(60000000, 80000000))+
  geom_line( size=1)+
  geom_point(data=bd_pron11[bd_pron11$Serie=='Predicci n', ]) + scale_
    color_manual(values=c("#8B0A50", "#1874CD"))+
  labs(title='Normal')+
  scale_linetype_manual(values = c(1,2))+
  labs(fill = "Serie")+
  theme_minimal()
graf_11


##MODELO LINEAL POISSON
mod_poisson_gee_sat<-geeglm(Y ~ X1+X2+X3+X4+X5+X6,family =poisson,id=
    report_month , corstr ="exchangeable",  data=Train)
summary(mod_poisson_gee_sat)
format(QIC(mod_poisson_gee_sat)[1], scientific=FALSE)
mod_poisson_gee<-geeglm(Y ~ X1+X2+X4+X5+X6,family =poisson,id=report_month
    , corstr ="exchangeable",  data=Train)
summary(mod_poisson_gee)
format(QIC(mod_poisson_gee)[1], scientific=FALSE)
pred_poisson_gee<-predict(mod_poisson_gee, newdata=Test, type = "response"
    )
rsq_poisson_mod<-rsq(mod_poisson_gee)
mae_poisson_mod<-MAE(Test$Y,pred_poisson_gee)
mse_poisson_mod<-MSE(Test$Y,pred_poisson_gee)
mape_poisson_mod<-MAPE(Test$Y,pred_poisson_gee)
theil_poisson_mod<-TheilU(Test$Y, pred_poisson_gee, type = c(2, 1), na.rm
    = FALSE)
QIC_poisson_mod<-QIC(mod_poisson_gee)[1]
dfpred2<-data.frame(Y=as.numeric(df_comp[1:51,c(2)]))
dfpred2$Fecha<-seq(from=as.Date("2018-01-01"), by="month", length.out=51)
```

```
dfpred2$prediction<-''
dfpred2$prediction[1:45]<-as.numeric(round(fitted(mod_poisson_gee),0))
dfpred2$prediction[46:51]<-as.numeric(round(pred_poisson_gee,0))
dfpred2$prediction<-as.numeric(dfpred2$prediction)
irr::icc(dfpred2[,c(1,3)], model = "oneway",
    type = "consistency", unit = "single")
bd_pron12 <- melt(dfpred2, id.vars = "Fecha")
bd_pron12<-bd_pron12%>%dplyr::filter(value!='')
names(bd_pron12)[2]<-'Serie'
names(bd_pron12)[3]<-'Cantidad de l neas'
bd_pron12$`Cantidad de l neas `<-as.numeric(bd_pron12$`Cantidad de l neas
    `)
bd_pron12$Serie<-ifelse(bd_pron12$Serie=='Y', 'Real', 'Predicci n')

graf_12<-ggplot(bd_pron12, aes(x=Fecha, y=`Cantidad de l neas `, colour =
    Serie))  +
  scale_y_continuous(limits = c(60000000, 80000000))+
  geom_line( size=1)+
  geom_point(data=bd_pron12[bd_pron12$Serie=='Predicci n', ]) + scale_
    color_manual(values=c("#8B0A50", "#1874CD"))+
  labs(title='Poisson')+
  scale_linetype_manual(values = c(1,2))+
  labs(fill = "Serie")+
  theme_minimal()
graf_12



##MODELO SMOOTHING SPLINES
mod_spline_sat<-geeglm(Y ~ bs(X1, degree = 3, df = 4 )+bs(X2, degree = 3,
    df = 4 )+bs(X3, degree = 3, df = 4 )+bs(X4, degree = 3, df = 4 )+bs(X5,
     degree = 3, df = 4 )+bs(X6, degree = 3, df = 4 ),id=report_month ,
    corstr ="exchangeable",data=as.data.frame(Train))
#cuando se pone df=4, se est  especificando que se debe poner un nudo
summary(mod_spline_sat)
format(QIC(mod_spline_sat)[1], scientific=FALSE)
#Se quitan los spline menos significativos que son los de la variable X6
mod_spline<-geeglm(Y ~  bs(X1, degree = 3, df = 4 )+bs(X2, degree = 3, df
    = 4 )+bs(X3, degree = 3, df = 4 )+bs(X4, degree = 3, df = 4)+bs(X5,
    degree = 3, df = 4),id=report_month , corstr ="exchangeable",data=as.
    data.frame(Train))
format(QIC(mod_spline)[1], scientific=FALSE)
summary(mod_spline)
pred_mod_spline<-predict(mod_spline, newdata=Test,type="response")
rsq_spline_mod<-rsq(mod_spline)
mae_spline_mod<-MAE(Test$Y,pred_mod_spline)
mse_spline_mod<-MSE(Test$Y,pred_mod_spline)
mape_spline_mod<-MAPE(Test$Y,pred_mod_spline)
```

```
theil_spline_mod<-TheilU(Test$Y, pred_mod_spline, type = c(2, 1), na.rm = FALSE)
QIC_spline_mod<-QIC(mod_spline)[1]

dfpred3<-data.frame(Y=as.numeric(df_comp[1:51,c(2)]))
dfpred3$Fecha<-seq(from=as.Date("2018-01-01"), by="month", length.out=51)
dfpred3$prediction<-''
dfpred3$prediction[1:45]<-as.numeric(round(fitted(mod_spline),0))
dfpred3$prediction[46:51]<-as.numeric(round(pred_mod_spline,0))
dfpred3$prediction<-as.numeric(dfpred3$prediction)
irr::icc(dfpred3[,c(1,3)], model = "oneway",
    type = "consistency", unit = "single")
bd_pron13 <- melt(dfpred3, id.vars = "Fecha")
bd_pron13<-bd_pron13%>%dplyr::filter(value!='')
names(bd_pron13)[2]<-'Serie'
names(bd_pron13)[3]<-'Cantidad de l neas'
bd_pron13$`Cantidad de l neas `<-as.numeric(bd_pron13$`Cantidad de l neas `)
bd_pron13$Serie<-ifelse(bd_pron13$Serie=='Y', 'Real', 'Predicci n')

graf_13<-ggplot(bd_pron13, aes(x=Fecha, y=`Cantidad de l neas `, colour = Serie))  +
  scale_y_continuous(limits = c(60000000, 80000000))+
  geom_line( size=1)+
  geom_point(data=bd_pron13[bd_pron13$Serie=='Predicci n', ]) + scale_color_manual(values=c("#8B0A50", "#1874CD"))+
  labs(title='Cubic Spline')+
  scale_linetype_manual(values = c(1,2))+
  labs(fill = "Serie")+
  theme_minimal()
graf_13


##MODELO SMOOTHING SPLINES WITH POISSON
mod_spline_poisson_sat<-geeglm(Y ~ bs(X1, degree = 3, df = 4 )+bs(X2, degree = 3, df = 4 )+bs(X3,degree = 3, df = 4)+bs(X4,degree = 3, df = 4)+bs(X5, degree = 3, df = 4)+bs(X6, degree = 3, df = 4),family = poisson,id=report_month , data=Train,corstr="exchangeable")
summary(mod_spline_poisson_sat)
format(QIC(mod_spline_poisson_sat)[1], scientific=FALSE)
mod_spline_poisson<-geeglm(Y ~ bs(X1, degree = 3, df = 4 )+bs(X2, degree = 3, df = 4 )+bs(X3, degree = 3, df = 4 )+bs(X4, degree = 3, df = 4 )+bs(X5, degree = 3, df = 4 ),family =poisson,id=report_month , data=Train, corstr="exchangeable")
summary(mod_spline_poisson)
format(QIC(mod_spline_poisson)[1], scientific=FALSE)
pred_mod_spline_poisson<-predict(mod_spline_poisson, newdata=Test, type = "response")
```

```
rsq_spline_pois_mod<-rsq(mod_spline_poisson)
mae_spline_pois_mod<-MAE(Test$Y,pred_mod_spline_poisson)
mse_spline_pois_mod<-MSE(Test$Y,pred_mod_spline_poisson)
mape_spline_pois_mod<-MAPE(Test$Y,pred_mod_spline_poisson)
theil_spline_pois_mod<-TheilU(Test$Y, pred_mod_spline_poisson, na.rm =
    FALSE)
QIC_spline_pois_mod<-QIC(mod_spline_poisson)[1]

dfpred4<-data.frame(Y=as.numeric(df_comp[1:51,c(2)]))
dfpred4$Fecha<-seq(from=as.Date("2018-01-01"), by="month", length.out=51)
dfpred4$prediction<-''
dfpred4$prediction[1:45]<-as.numeric(round(fitted(mod_spline_poisson),0))
dfpred4$prediction[46:51]<-as.numeric(round(pred_mod_spline_poisson,0))
dfpred4$prediction<-as.numeric(dfpred4$prediction)
irr::icc(dfpred4[,c(1,3)], model = "oneway",
     type = "consistency", unit = "single")

bd_pron14 <- melt(dfpred4, id.vars = "Fecha")
bd_pron14<-bd_pron14%>%dplyr::filter(value!='')
names(bd_pron14)[2]<-'Serie'
names(bd_pron14)[3]<-'Cantidad de l neas'
bd_pron14$`Cantidad de l neas `<-as.numeric(bd_pron14$`Cantidad de l neas
    `)
bd_pron14$Serie<-ifelse(bd_pron14$Serie=='Y', 'Real', 'Predicci n')

graf_14<-ggplot(bd_pron14, aes(x=Fecha, y=`Cantidad de l neas `, colour =
    Serie))  +
  scale_y_continuous(limits = c(60000000, 80000000))+
  geom_line( size=1)+
  geom_point(data=bd_pron14[bd_pron14$Serie=='Predicci n', ]) + scale_
    color_manual(values=c("#8B0A50", "#1874CD"))+
  labs(title='Cubic Spline Poisson')+
  scale_linetype_manual(values = c(1,2))+
  labs(fill = "Serie")+
  theme_minimal()
graf_14

##binomial negativa
fi_bn<-read.csv('C:/Users/USUARIO/Documents/Maestria_Estadistica/Tesis/
    Datos/Fitted_bn.csv', sep=";", dec=',', header=FALSE)
dfpred5<-data.frame(prediction=as.numeric(fi_bn$V1))
dfpred5$Fecha<-seq(from=as.Date("2018-01-01"), by="month", length.out=51)
dfpred5$Y<-df_comp$Y
irr::icc(dfpred5[,c(1,3)], model = "oneway",
     type = "consistency", unit = "single")
ICC(dfpred5[,c(1,3)])
cor(dfpred5[,c(1,3)])
bd_pron15 <- melt(dfpred5, id.vars = "Fecha")
```

```
bd_pron15<-bd_pron15%>%dplyr::filter(value!='')
names(bd_pron15)[2]<-'Serie'
names(bd_pron15)[3]<-'Cantidad de l neas'
bd_pron15$`Cantidad de l neas`<-as.numeric(bd_pron15$`Cantidad de l neas`)
bd_pron15$Serie<-ifelse(bd_pron15$Serie=='Y', 'Real', 'Predicci n')

graf_15<-ggplot(bd_pron15, aes(x=Fecha, y=`Cantidad de l neas`, colour = Serie))  +
  scale_y_continuous(limits = c(60000000, 80000000))+
  geom_line( size=1)+
  geom_point(data=bd_pron15[bd_pron15$Serie=='Predicci n', ]) + scale_color_manual(values=c("#8B0A50", "#1874CD"))+
  labs(title='Binomial negativa')+
  scale_linetype_manual(values = c(1,2))+
  labs(fill = "Serie")+
  theme_minimal()
graf_15


(graf_11+graf_13+graf_14) / ((graf_12 | graf_15))

##AGRUPANDO LAS MEDIDAS PARA COMPARAR
vec1<-c('R Squared',round(rsq_normal_mod,3),round(rsq_poisson_mod,3),round(rsq_spline_mod,3),round(rsq_spline_pois_mod,3))
vec2<-c('MAE',round(mae_normal_mod,3),round(mae_poisson_mod,3),round(mae_spline_mod,3),round(mae_spline_pois_mod,3))
vec3<-c('MSE',round(mse_normal_mod,3),round(mse_poisson_mod,3),round(mse_spline_mod,3),round(mse_spline_pois_mod,3))
vec4<-c('MAPE',round(mape_normal_mod,3),round(mape_poisson_mod,3),round(mape_spline_mod,3),round(mape_spline_pois_mod,3))
# vec5<-c('Theil',round(theil_normal_mod,3),round(theil_poisson_mod,3),round(theil_spline_mod,3),round(theil_spline_pois_mod,3))
vec6<-c('QIC',round(QIC_normal_mod,3),round(QIC_poisson_mod,3),round(QIC_spline_mod,3),round(QIC_spline_pois_mod,3))

# df_results<-data.frame(rbind(vec1, vec2, vec3,vec4, vec5, vec6))
df_results<-data.frame(rbind(vec1, vec2, vec3,vec4, vec6))

names(df_results)<-c("M trica",'Modelo_normal', 'Modelo_poisson', 'Modelo_splines', 'Modelo_splines_poisson')
df_results$Mejor_modelo<-''
df_results$Mejor_modelo[1]<-names(df_results)[apply(df_results[1,(2:5)], 1, function (x) {which.max(x)})+1]
df_results$Mejor_modelo[2]<-names(df_results)[apply(df_results[2,(2:5)], 1, function (x) {which.min(x)})+1]
df_results$Mejor_modelo[3]<-names(df_results)[apply(df_results[3,(2:5)], 1, function (x) {which.min(x)})+1]
```

```
df_results$Mejor_modelo[4]<-names(df_results)[apply(df_results[4,(2:5)], 1, function (x) {which.min(x)})+1]
# df_results$MEJOR_MODELO[5]<-names(df_results)[apply(df_results[5,(2:5)], 1, function (x) {which.min(x)})+1]
df_results$Mejor_modelo[5]<-names(df_results)[apply(df_results[5,(2:5)], 1, function (x) {which.min(x)})+1]
#print(xtable(df_results), include.rownames = FALSE)
#write.table(df_results, 'C:/Users/USUARIO/Documents/Maestria_Estadistica/Tesis/Datos/df_results.csv',sep="|")
#De acuerdo a lo anterior se selecciona el modelo poisson
mod_poisson_gee<-geeglm(Y ~ X1+X2+X4+X5+X6,family =poisson,id=report_month , corstr ="exchangeable",  data=df_comp)
summary(mod_poisson_gee)
#Creando gr ficos de los residuales.
par(mfrow=c(2,2))
r1<-plot(mod_normal_gee, main='modelo normal gee')
r2<-plot(mod_poisson_gee)
r3<-plot(mod_spline)
r4<-plot(mod_spline_poisson)
scatter.smooth(sqrt(predict(mod_normal_gee, type='response')), qresid(mod_normal_gee), col='gray')

df_resid_plot<-data.frame(pearson_residuals=resid(mod_normal_gee, type='pearson'), fitted=mod_normal_gee$fitted.values)
plot_res_norm<-ggplot(df_resid_plot, aes(x=fitted, y=pearson_residuals)) + geom_point()+geom_hline(yintercept=0)+geom_smooth( color='violetred3', span=1,level = 0)+
                ggtitle("Residuales modelo normal")+
                theme_minimal()

plot(mod_normal_gee)

df_resid_plot1<-data.frame(pearson_residuals=resid(mod_poisson_gee, type='pearson'), fitted=mod_poisson_gee$fitted.values)
plot_res_pois<-ggplot(df_resid_plot1, aes(x=fitted, y=pearson_residuals)) + geom_point()+geom_hline(yintercept=0)+geom_smooth( color='violetred3',span=1,level = 0)+
                  ggtitle("Residuales modelo Poisson")+
                  theme_minimal()


df_resid_plot2<-data.frame(pearson_residuals=resid(mod_spline, type='pearson'), fitted=mod_spline$fitted.values)
plot_res_spl<-ggplot(df_resid_plot2, aes(x=fitted, y=pearson_residuals)) + geom_point()+geom_hline(yintercept=0)+geom_smooth( color='violetred3', span=1,level = 0)+
                ggtitle("Residuales modelo splines")+
                theme_minimal()
```

```
304
305
306 df_resid_plot3<-data.frame(pearson_residuals=resid(mod_spline_poisson,
    type='pearson'), fitted=mod_spline_poisson$fitted.values)
307 plot_res_spl_pois<-ggplot(df_resid_plot2, aes(x=fitted, y=pearson_
    residuals)) + geom_point()+geom_hline(yintercept=0)+geom_smooth( color=
    'violetred3',span=1,level = 0)+
308                   ggtitle("Residuales modelo splines poisson")+
309                   theme_minimal()
310
311 df_resid_plot4<-read.csv('C:/Users/USUARIO/Documents/Maestria_Estadistica/
    Tesis/Datos/df_resid_bn.csv')
312 names(df_resid_plot4)<-c('fitted', 'pearson_residuals')
313 plot_res_nb<-ggplot(df_resid_plot4, aes(x=fitted, y=pearson_residuals)) +
    geom_point()+geom_hline(yintercept=0)+geom_smooth( color='violetred3',
    span=1,level = 0)+
314             ggtitle("Residuales modelo binomial negativo")+
315             theme_minimal()
316 (plot_res_norm+plot_res_spl+plot_res_spl_pois) / ((plot_res_pois | plot_
    res_nb))
317
318
319
320 ##PRONOSTICO
321
322 bd_pron<-read.csv('C:/Users/USUARIO/Documents/Maestria_Estadistica/Tesis/
    Datos/220807_bd_pron.csv', sep=";", dec=',')
323 names(bd_pron)[2:7]<- c('X1', 'X2', 'X3', 'X4','X5', 'X6')
324 pred<-predict(mod_poisson_gee, newdata=bd_pron, type = "response")
325 dfpred1<-data.frame(Y=as.numeric(df_comp[,c(2)]))
326 dfpred1$Fecha<-seq(from=as.Date("2018-01-01"), by="month", length.out=51)
327
328 dfpred<-data.frame(Y=as.numeric(df_comp[,c(2)]))
329 new_rows<-data.frame(Y=c('','',''))
330 dfpred<-rbind(dfpred,new_rows)
331 dfpred$Fecha<-seq(from=as.Date("2018-01-01"), by="month", length.out=54)
332 dfpred$prediction<-''
333 dfpred$prediction[1:51]<-as.numeric(round(fitted(mod_poisson_gee),0))
334 #dfpred$prediction[1:51]<-df_comp$Y
335 dfpred$prediction[52:54]<-as.numeric(round(pred,0))
336 bd_pron4 <- melt(dfpred, id.vars = "Fecha")
337 bd_pron4<-bd_pron4%>%dplyr::filter(value!='')
338 names(bd_pron4)[2]<-'Serie'
339 names(bd_pron4)[3]<-'Cantidad de l neas'
340 bd_pron4$`Cantidad de l neas `<-as.numeric(bd_pron4$`Cantidad de l neas `)
341 bd_pron4$Serie<-ifelse(bd_pron4$Serie=='Y', 'Real', 'Predicci n')
342
343 graf_2<-ggplot(bd_pron4, aes(x=Fecha, y=`Cantidad de l neas `, colour =
```

```
    Serie))  +
344   scale_y_continuous(limits = c(60000000, 80000000))+
345   geom_line( size=1)+
346   geom_point(data=bd_pron4[bd_pron4$Serie=='Predicci n', ]) + scale_color
       _manual(values=c("#8B0A50", "#1874CD"))+
347   scale_linetype_manual(values = c(1,2))+
348   labs(fill = "Serie")+
349   #labs(y =expression(X["5"]))+
350   theme_minimal()
351 graf_2
```

Listing D.1: Script ajuste de modelos

E. Anexo:Script de Python para ajuste de regresión binomial negativa

```
import pandas as pd
import statsmodels.api as sm
import statsmodels.formula.api as smf
import numpy as np
import math
import pingouin as pg

df_comp=pd.read_csv('C:/Users/USUARIO/Documents/Maestria_Estadistica/Tesis/Datos/BD_FIN_MOD_TESIS_REN.csv', sep=";", decimal=',')
df_comp.head()
Train=df_comp.head(45)
Test=df_comp.tail(6)
print(Train.shape)
print(Test.shape)
fam3 = sm.families.NegativeBinomial()
mod3 = smf.gee("Y ~ X1+X2+X3+X4+X5+X6", "report_month", Train,cov_struct=ind, family=fam3)
res3 = mod3.fit(maxiter=10000)
print(res3.summary())
print(res3.qic(1.000))
print(res3.pseudo_rsquared('mcf'))
fam3 = sm.families.NegativeBinomial()
mod3 = smf.gee("Y ~ X1+X2+X4+X5+X6", "report_month", Train,cov_struct=ind, family=fam3)
res3 = mod3.fit(maxiter=10000)
print(res3.summary())
print(res3.qic(1.000))
print(res3.pseudo_rsquared('mcf'))
ypred2 = res3.predict(Test)
print(ypred2)
def MAE(Y_actual,Y_Predicted):
    mae = np.mean(np.abs(Y_actual - Y_Predicted))
    return mae

def MSE(Y_actual,Y_Predicted):
```

```
    mse = np.mean((Y_actual - Y_Predicted)**2)
    return mse

def MAPE(Y_actual,Y_Predicted):
    mape = np.mean(np.abs((Y_actual - Y_Predicted)/Y_actual))*100
    return mape

print(MAE(Test['Y'],ypred2))
print(MSE(Test['Y'],ypred2))
print(MAPE(Test['Y'],ypred2))
vec_fin=res3.fittedvalues
vec_fin=vec_fin.append(ypred2)
print(vec_fin)
#Obteniendo los residuales de pearson
df_residuales_bn=pd.DataFrame({'Fitted_values':res3.fittedvalues, '
   Pearson_resid':res3.resid_pearson})
```

Listing E.1: Script ajuste modelo de regresión binomial negativa

Referencias

Agresti, A. (2002). *Categorical data analysis*. John Wiley & Sons.

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

Annafari, M. T. (2013). Multiple subscriptions of mobile telephony: Explaining the diffusion pattern using sampling data. *Telecommunications Policy*, 37(10):930–939. Regulating and investment in new communications infrastructure Understanding ICT adoption and market trends: Papers from recent European ITS regional conferences.

Cameron, A. and Trivedi, P. (1999). Essentials of count data regression. *A Companion to Theoretical Econometrics. Malden, MA: Blackwell Publishing Ltd.*

Chu, W.-L., Wu, F.-S., Kao, K.-S., and Yen, D. C. (2009). Diffusion of mobile telephony: An empirical study in Taiwan. *Telecommunications Policy*, 33(9):506–520.

Dagum, E. B. and Cholette, P. A. (2006). Benchmarking, temporal distribution, and reconciliation methods for time series.

de Myttenaere, A., Golden, B., Le Grand, B., and Rossi, F. (2016). Mean absolute percentage error for regression models. *Neurocomputing*, 192:38–48. Advances in artificial neural networks, machine learning and computational intelligence.

Fitzmaurice, G., Laird, N., and Ware, J. (2011). *Applied Longitudinal Analysis*. Wiley Series in Probability and Statistics. Wiley.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2012). *Applied longitudinal analysis*, volume 998. John Wiley & Sons.

Forthofer, R. N., Lee, E. S., and Hernandez, M. (2007). 3 - descriptive methods. In Forthofer, R. N., Lee, E. S., and Hernandez, M., editors, *Biostatistics (Second Edition)*, pages 21–69. Academic Press, San Diego, second edition edition.

Frees, E. W. et al. (2004). *Longitudinal and panel data: analysis and applications in the social sciences*. Cambridge University Press.

Gamboa, L. F. and Otero, J. (2009). An estimation of the pattern of diffusion of mobile phones: The case of Colombia. *Telecommunications Policy*, 33(10-11):611–620.

Gamer, M., Lemon, J., and <puspendra.pusp22@gmail.com>, I. F. P. S. (2019). *irr: Various Coefficients of Interrater Reliability and Agreement*. R package version 0.84.1.

Halekoh, U., Højsgaard, S., and Yan, J. (2006). The r package geepack for generalized estimating equations. *Journal of Statistical Software*, 15/2:1–11.

Hardin, J. and Hilbe, J. (2013). Generalized estimating equations (second edition).

Hastie, T. (2022). *gam: Generalized Additive Models*. R package version 1.20.2.

Herrera Giraldo, M. F. (2012). Difusión de la telefonía móvil en Colombia. Master's thesis.

Iliinsky, N. and Steele, J. (2011). *Designing data visualizations: Representing informational Relationships*. O' Reilly Media, Inc.

Islama, M. R. (2014). R program for temporal disaggregation: Denton's method.

Jha, A. and Saha, D. (2020). Forecasting and analysing the characteristics of 3G and 4G mobile broadband diffusion in India: A comparative evaluation of Bass, Norton-Bass, Gompertz, and logistic growth models. *Technological Forecasting and Social Change*, 152:119885.

Koo, T. K. and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

Kristoufek, L. (2014). Measuring correlations between non-stationary series with dcca coefficient. *Physica A: Statistical Mechanics and its Applications*, 402:291–298.

Li, G., Lian, H., Feng, S., and Zhu, L. (2013). Automatic variable selection for longitudinal generalized linear models. *Computational Statistics & Data Analysis*, 61:174–186.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Luo, R. and Pan, J. (2022). Conditional generalized estimating equations of mean-variance-correlation for clustered data. *Computational Statistics Data Analysis*, 168:107386.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). *Introduction to linear regression analysis*. John Wiley & Sons.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, 57(1):120–125.

Park, T., Davis, C. S., and Li, N. (1998). Alternative gee estimation procedures for discrete longitudinal data. *Computational Statistics Data Analysis*, 28(3):243–256.

Pearson, R. K. (2018). *Exploratory data analysis using R*. Chapman and Hall/CRC.

Prass, T. S. and Pumi, G. (2020). *DCCA: Detrended Fluctuation and Detrended Cross-Correlation Analysis*. R package version 0.1.1.

Puth, M.-T., Neuhäuser, M., and Ruxton, G. D. (2015). Effective use of spearman's and kendall's correlation coefficients for association between two measured traits. *Animal Behaviour*, 102:77–84.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Cambridge University Press.

Sax, C. and Steiner, P. (2013). Temporal disaggregation of time series. *The R Journal*, 5(2):80–87.

Seabold, S. and Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Shen, C. (2015). Analysis of detrended time-lagged cross-correlation between two nonstationary time series. *Physics Letters A*, 379(7):680–687.

Tsai, M.-Y. (2015). Comparison of concordance correlation coefficient via variance components, generalized estimating equations and weighted approaches with model selection. *Computational Statistics Data Analysis*, 82:47–58.

Tsai, M.-Y., Wang, J.-F., and Wu, J.-L. (2011). Generalized estimating equations with model selection for comparing dependent categorical agreement data. *Computational Statistics Data Analysis*, 55(7):2354–2362.

Upton, G. and Cook, I. (2014). *A dictionary of statistics 3e*. Oxford university press.

Wold, S. (1974). Spline functions in data analysis. *Technometrics*, 16(1):1–11.

Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Cengage learning.

Wu, F.-S. and Chu, W.-L. (2010). Diffusion models of mobile telephony. *Journal of Business Research*, 63(5):497–501. TECHNOLOGY MANAGEMENT.

Zebende, G. (2011). Dcca cross-correlation coefficient: Quantifying level of cross-correlation. *Physica A: Statistical Mechanics and its Applications*, 390(4):614–618.

Ziegler, A. (2011). *Generalized estimating equations*, volume 204. Springer Science & Business Media.