Reporte técnico - Técnicas de Aprendizaje Estadístico - Accidentalidad

Andrés Felipe Aguilar
Juan Felipe Múnera
David Chaverra

Técnicas de Aprendizaje Estadístico
Universidad Nacional de Colombia Sede Medellín

Introducción

Los incidentes viales causan lesiones fatales y no fatales, con efectos en salud, bienestar y productividad (Espinosa López, Cabrera Arana & Velásquez Osorio, 2017), sumado a ello, la rápida urbanización, el cambio tecnológico y el crecimiento económico en los diferentes países han llevado a un crecimiento sustancial en la densidad vehicular y en la complejidad del tráfico en las vías de éstos (Hyder & Vecino-Ortiz, 2014). En este reporte se presenta una aplicación de los modelos GAMLSS (Generalized Additive Models for Location, Shape and Scale) que permiten asumir distribuciones estadísticas para la variable respuesta diferentes a la normal (Barajas, Torres, Arteaga & Castro, 2015) para analizar los datos de accidentalidad en Medellín: los datos utilizados están expuestos en las bases de datos del portal Geomedellín (Portal geográfico del municipio de Medellín) ("GEO medellín", 2019) y contienen información referente a múltiples siniestros, en los que se detalla el tipo de accidente, dónde y cuándo ocurrió. Este análisis fue realizado con el fin de construir dos sistemas, uno que prediga el número de accidentes tomando como entradas una ventana y una resolución temporal específicas, además de una zona espacial que puede ser un barrio o comuna de Medellín; y otro que agrupe barrios con características similares referentes a los tipos de accidentes que ocurren allí.

Justificación

La accidentalidad constituye una problemática común a la mayor parte de las sociedades modernas que anualmente genera grandes costos en vidas humanas, problema que no es ajeno a la ciudad de Medellín y que causa en promedio 220 muertes al año, 22071 Heridos de gravedad y 18410 casos de solo daños materiales. Lo anterior evidencia un problema que no solo pone en riesgo la vida de los habitantes, sino que además pone grandes cargas en el sistema de salud y causa pérdidas millonarias en daños e indemnizaciones. Peter Drucker señala que «Lo que no se mide, no se puede cambiar, lo que no se puede cambiar no se puede mejorar», por las cifras antes mencionadas se puede deducir que hemos logrado avances en la medición y aunque se evidencia en estudios como los realizados por Erdogan (Erdogan, Yilmaz, Baybura & Gullu, 2008) que la medición está lejos de ser ideal, esto no significa que no se pueda dar un primer paso a el cambio. Para dar los primeros pasos en el camino a la mejora es importante tomar los datos medidos y depurarlos con el objeto de entenderlos, es por esto que se procede en un inicio a realizar un estudio exploratorio de los datos y un análisis descriptivo de éstos, estudios que proveen las pautas para el diseño de los modelos predictivos que serán presentados más adelante. Los resultados encontrados por los modelos descriptivos permiten no solo entendendimiento a los autores de este trabajo, pero le permite a estos presentar con datos concretos una imagen de la accidentalidad que se ajusta a la realidad y de una manera simple expone los problemas de la misma en un medio simple de consumir como lo son: los mapas y los gráficos. Los modelos predictivos por otro lado van más allá de los dicho por Drucker, ya que la estadística nos permite usar datos del pasado para tener una idea de lo que puede llegar a ser, se extiende entonces la fase de medición al futuro con el objeto de generar una plataforma que pueda servir como palanca para la prevención y la mejora de la accidentalidad y de sus consecuencias.

Objetivos

El análisis y los sistemas que se pretenden construir sólo abarcan el área urbana de Medellín correspondiente a sus 16 comunas y sus respectivos barrios.

- Construir un sistema que prediga el número de accidentes tomando como entradas una ventana y una resolución temporal específicas, además de una zona espacial que puede ser un barrio o comuna de Medellín.
- Elaborar un sistema que agrupe barrios con características similares en razón de los tipos de accidentes que tienen lugar en estos.

4. Descripción de los datos

4.1. La base de datos

Los datos se obtienen de la página del portal GeoMedellín de la alcaldía de Medellín en su sección de datos abierto.
Se utilizan 5 bases de datos correspondientes a la Accidentalidad Georreferenciada de los años 2014 al 2018. Posee los siguientes atributos:

- OBJECTID: id de cada registro.
- X: coordenada.
- Y: coordenada.
- RADICADO: código emitido por la secretaría de movilidad de Medellín.
- FECHA.
- HORA.
- DIA.
- PERIODO: año del siniestro.
- CLASE: tipo de accidente
- DIRECCION.
- DIRECCION_ENC.
- CBML: Código de ubicación del predio en la ciudad.
- TIPO_GEOCOD.
- GRAVEDAD: repercusiones del accidente
- BARRIO.
- COMUNA.
- DISENO: clasificación del lugar del accidente.
- DIA_NOMBRE.
- MES.

4.2. Análisis y depuración

Dado que el presente trabajo se centrará sólo en las zonas urbanas de Medellín se procede trabajar unicamente con los registros cuyo atributo GEOCOD no corresponda a “ZONA RURAL” y su atributo COMUNA corresponda a alguna de las 16 comunas de Medellín.

Se consideran los siguientes errores de imputación y se realizan las respectivas correcciones:

- Atributo CLASE:
  - “Caida Ocupante” y “Caida de Ocupante” se considera equivalente a “Caída de Ocupante.
  - “*Choque*” equivalente a”Choque".
- Atributo GRAVEDAD:
  - “CON MUERTO” se considera equivalente a “MUERTO”.
- Atributo BARRIO: se realizan los siguientes reemplazos.
  - “Aures No. 2” por “Aures No.2”.
  - “Asomadera No. 1” por “Asomadera No.1”.
  - “Barrio de Jesús” por “Barrios de Jesús”.
  - “B. Cerro El Volador” por “B. Cerro El Volador”.
  - “Berlin” por “Berlín”.
  - “Bomboná No. 1” por “Bomboná No.1”.
  - “Campo Valdés No.2” por “Campo Valdés No. 2”.

- "Manrique Central No.1" por "Manrique Central No. 1".
- "Manrique Central No.2" por "Manrique Central No. 1".
- "Moscú No.1" por "Moscú No. 1".
- "Moscú No.2" por "Moscú No. 2".
- "Nueva Villa de La Iguaná" por "Nueva Villa de la Iguaná".
- "Santa María de Los Ángeles" por "Santa María de los Ángeles".
- "Santo Domingo Savio No.1" por "Santo Domingo Savio No. 1".
- "Versalles No.1" por "Versalles No. 1".
- "Versalles No.2" por "Versalles No. 2".
- "Villa Liliam" por "Villa Lilliam".

Además se encuentran valores nulos para los atributos DISEÑO y BARRIO. El primero se descarta como un atributo relevante para los análisis del presente documento, por lo que no será utilizado en ninguno de los modelos construidos en el trabajo y no se removerán las tuplas con DISEÑO nulo. Para la variable BARRIOS todas las tuplas correspondientes a valores faltantes o con valores "0" o "Sin Nombre" son descartadas.

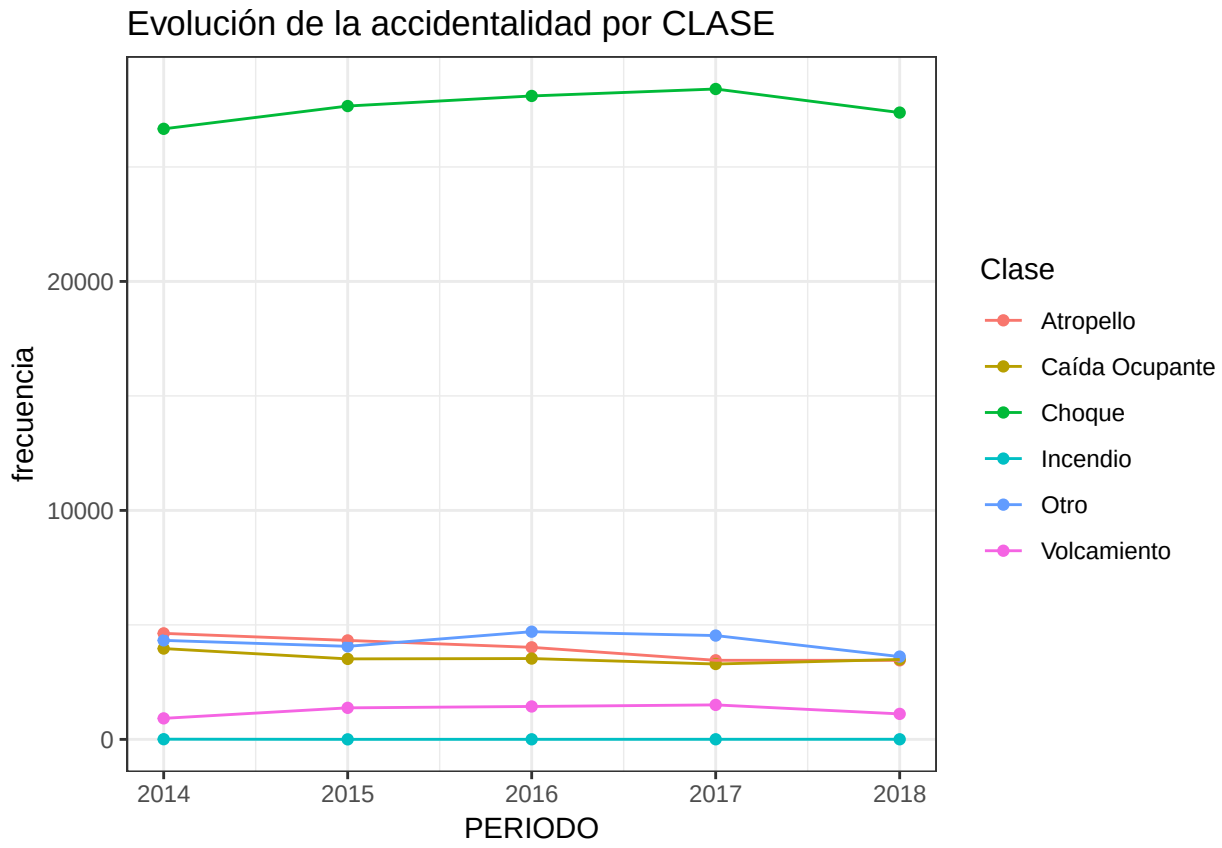
Tras las modificaciones mencionadas de los 209426 registros originales entre las bases de datos del 2014 al 2018, se utilizarán 203507 restantes en el análisis y en la cosntrucción de los modelos.

4.3 Análisis

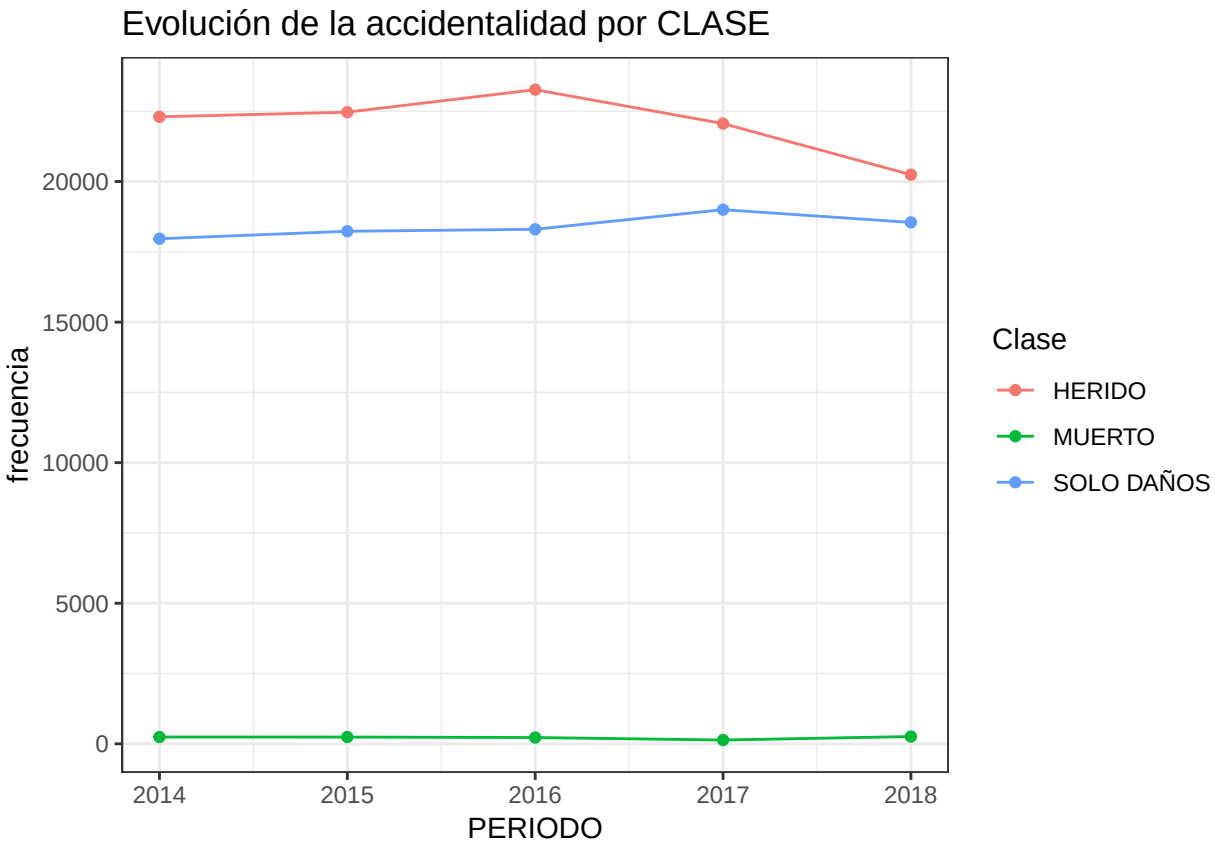
Se considera que las siguientes variables no son de interes para la construcción de los modelos:

- OBJECTID, CBML, RADICADO: son identificadores de los registros.
- DIRECCION, DIRECCION_ENC, X, Y: hacen referencia al posicionamiento geográfico específico del accidente. Ya que los modelos buscan predecir a nivel de barrio y comuna esta información no es relevante.
- TIPO_GEOCOD: no presenta información variable. Al parecer indica malla vial y la dirección del accidente.
- DISEÑO: presenta información relevante sobre la estructura vial de los accidentes. Esto podría ser relevante para un estudio sobre la GRAVEDAD del accidente, pero no se encontraron relaciones aparentes con otros atributos de la base que puedan ayudar a mejorar las predicciones.

Accidentalidad Anual

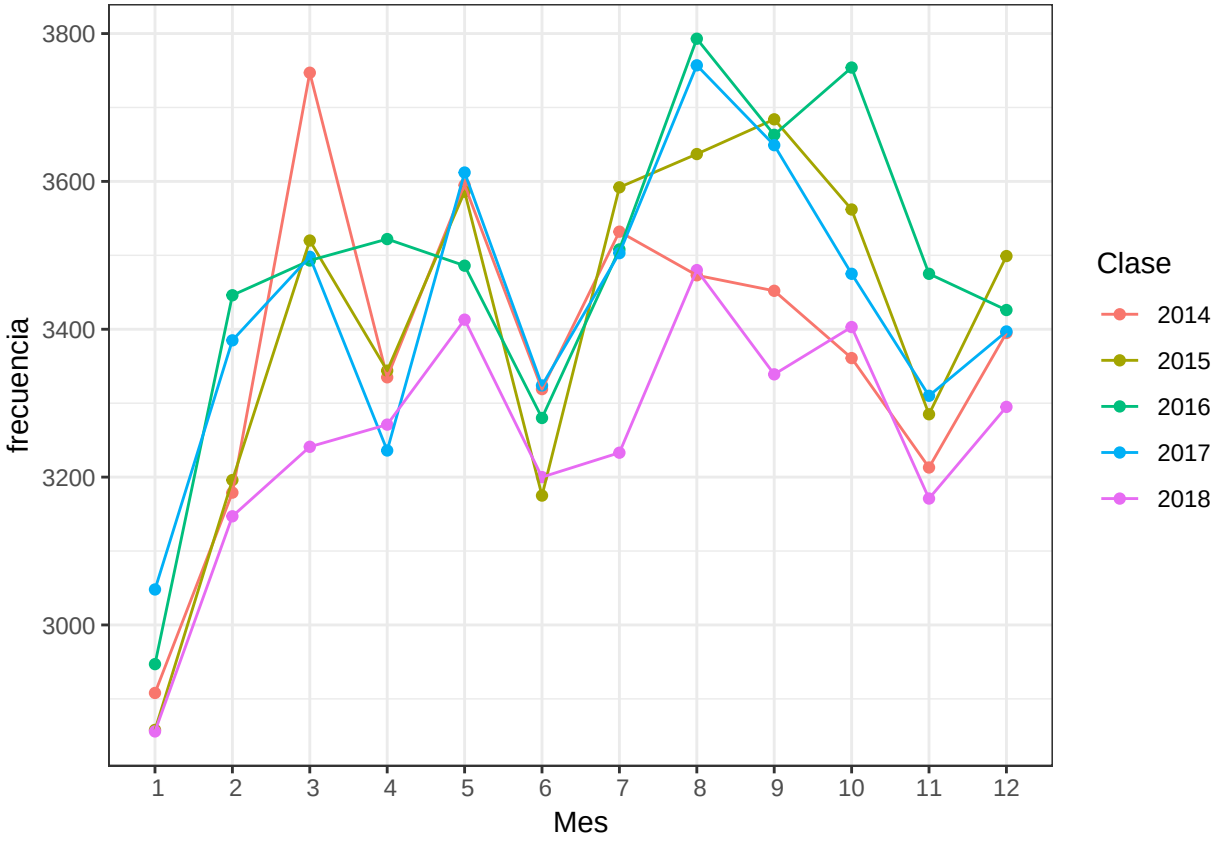


Se evidencia que la evolución de la accidentalidad por cada clase no tiene un patrón de dependencia claro con la Clase de los accidentes. Los choques son el tipo de accidente más frecuente y tuvieron un leve crecimiento entre el 2014 y 2017 y luego decrecieron para el 2018. Las otras clases de accidentes no presentaron variaciones significativas con el paso de los años a excepción de talvez "atropello" y "otro" los que tuvieron mayores variaciones.



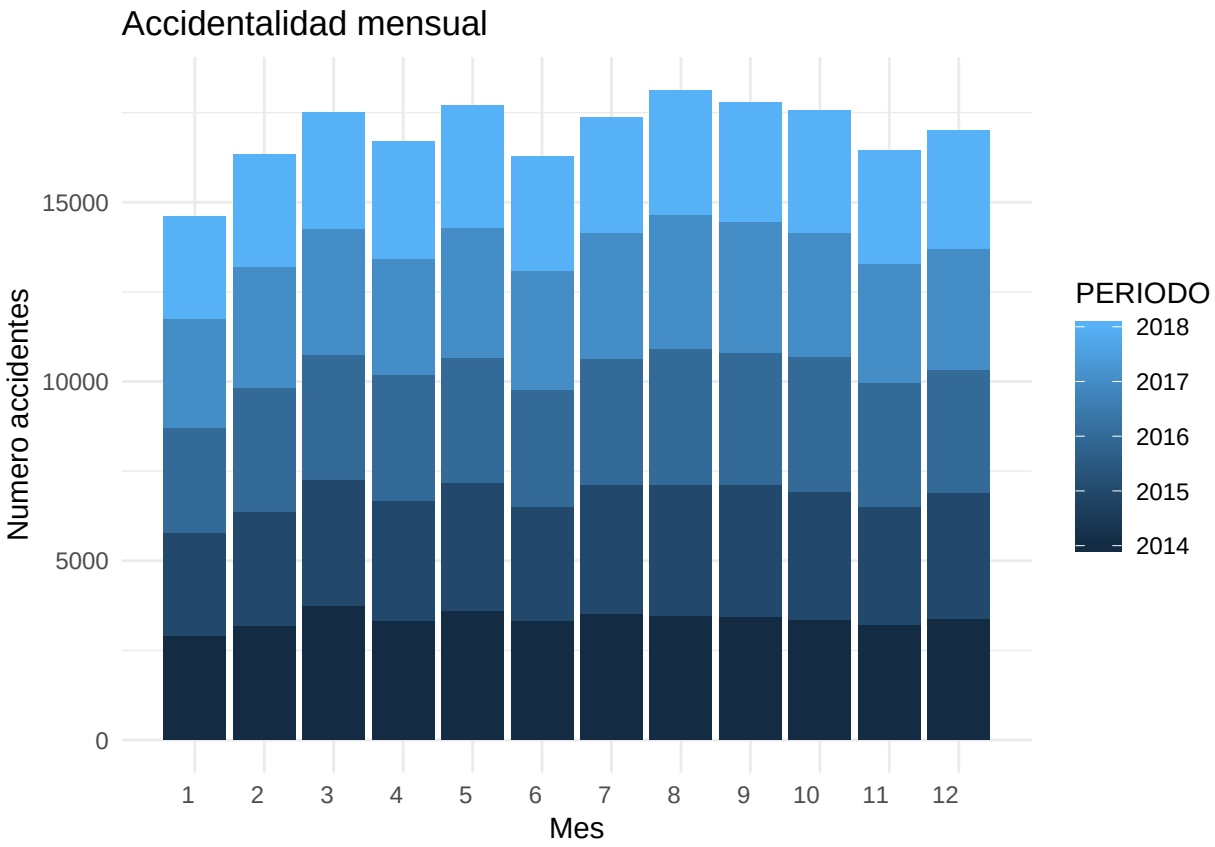
En general se denota una leve disminución en la cantidad de heridos desde el 2016 hasta el 2018, los accidentes con muertos y aquellos donde solo ocurrieron daños materiales no presentan cambios significativos.

Evolución accidentalidad por meses

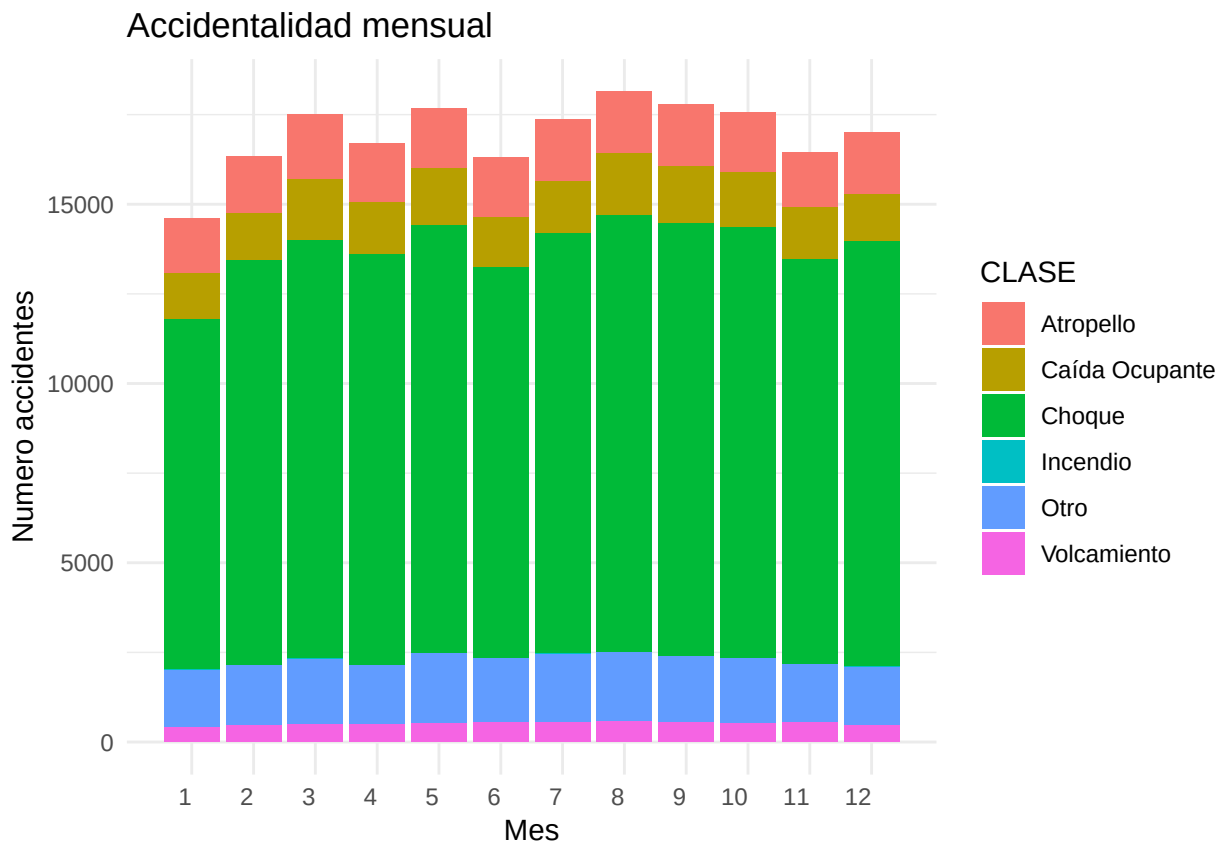


Se puede apreciar que en terminos generales la accidentalidad para el año 2018 es siempre una de las dos más pequeñas independiente del mes. Esto y las leves diferencias de accidentalidad anual nos indican que el PERIODO podría ser relevante para los modelos predictivos.

Otra cosa a notar es que la accidentalidad parece tener un patron NO LINEAL para los tipos de accidentes, siendo muy bajo en Enero, creciendo hasta Marzo, disminuyendo nuevamente en Abril y Junio con un aumento en Mayo. Des Agosto a Octubre parece ser el punto para la accidentalidad, disminuye en el mes de Noviembre y aumenta nuevamente el Diciembre.

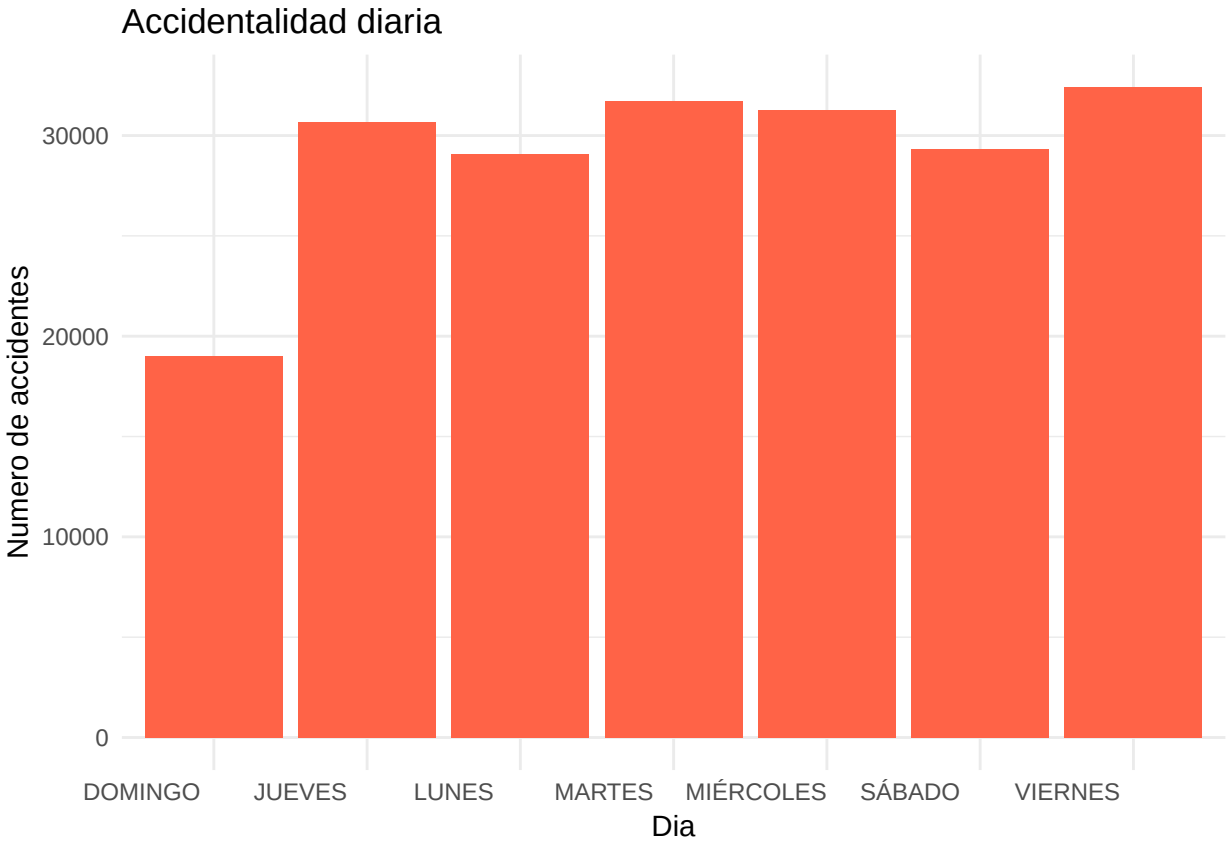


Este comportamiento se repite para todos los años (aunque se marca un poco más en algunos). Esto nos dice que el MES podría ser variable relevante para la construcción de los modelos, pero su consideración podría ser categórica.



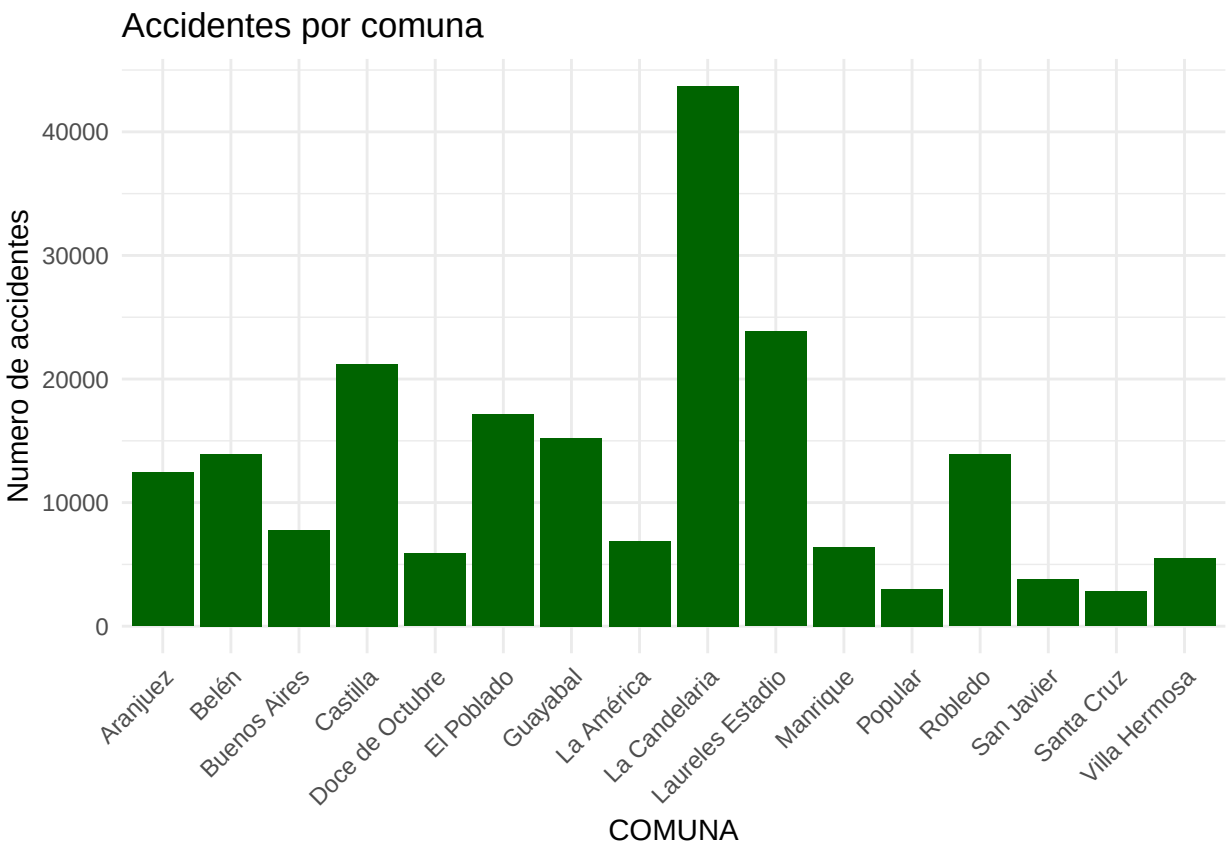
CLASE no parece tener una variación notable con los meses.

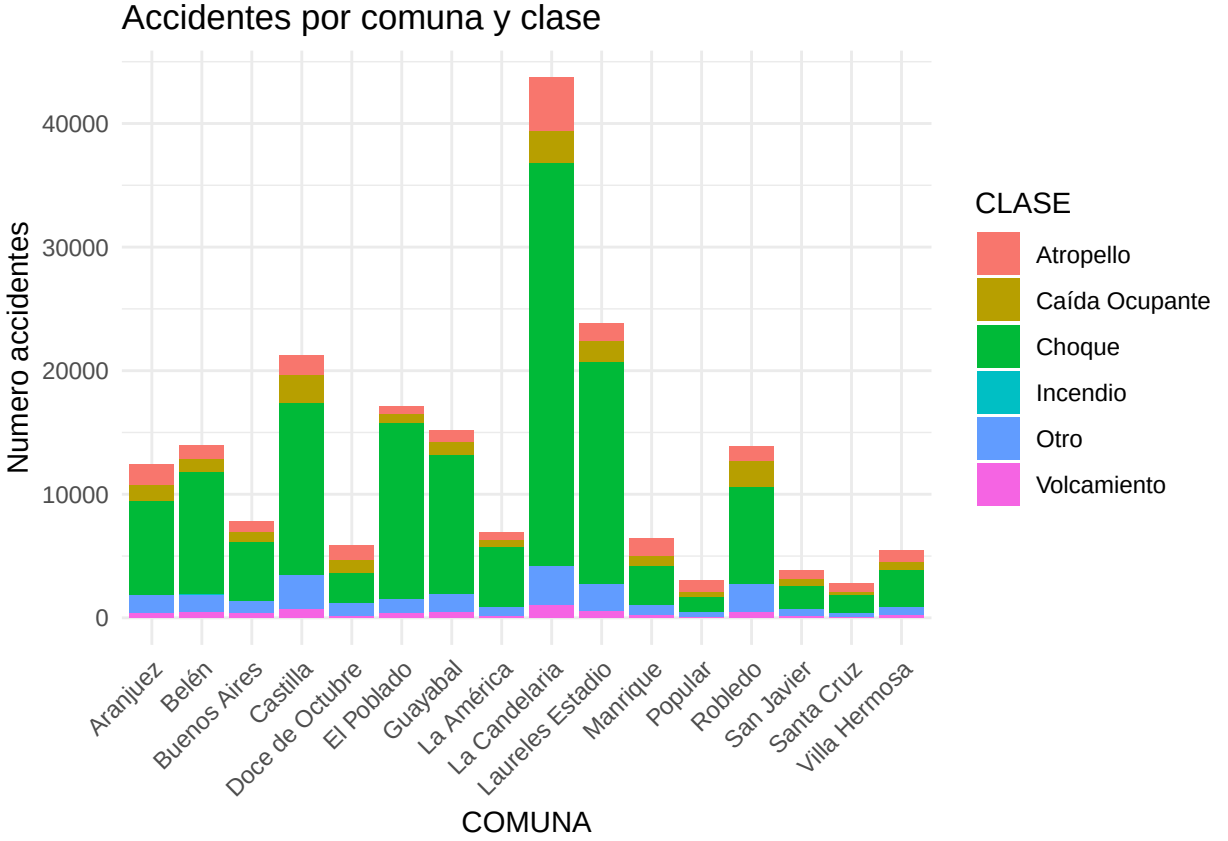
Análisis de accidentalidad por día de la semana



Existe una clara diferenciación entre la cantidad de accidentes según el día de la semana. El día domingo tiene una clara reducción de accidentes, lo cuál es acorde a lo esperable, ya que el día domingo al ser generalmente libre se suele ver menos circulacción por el area urbana de Medellín.

Accidentalidad para las Comunas





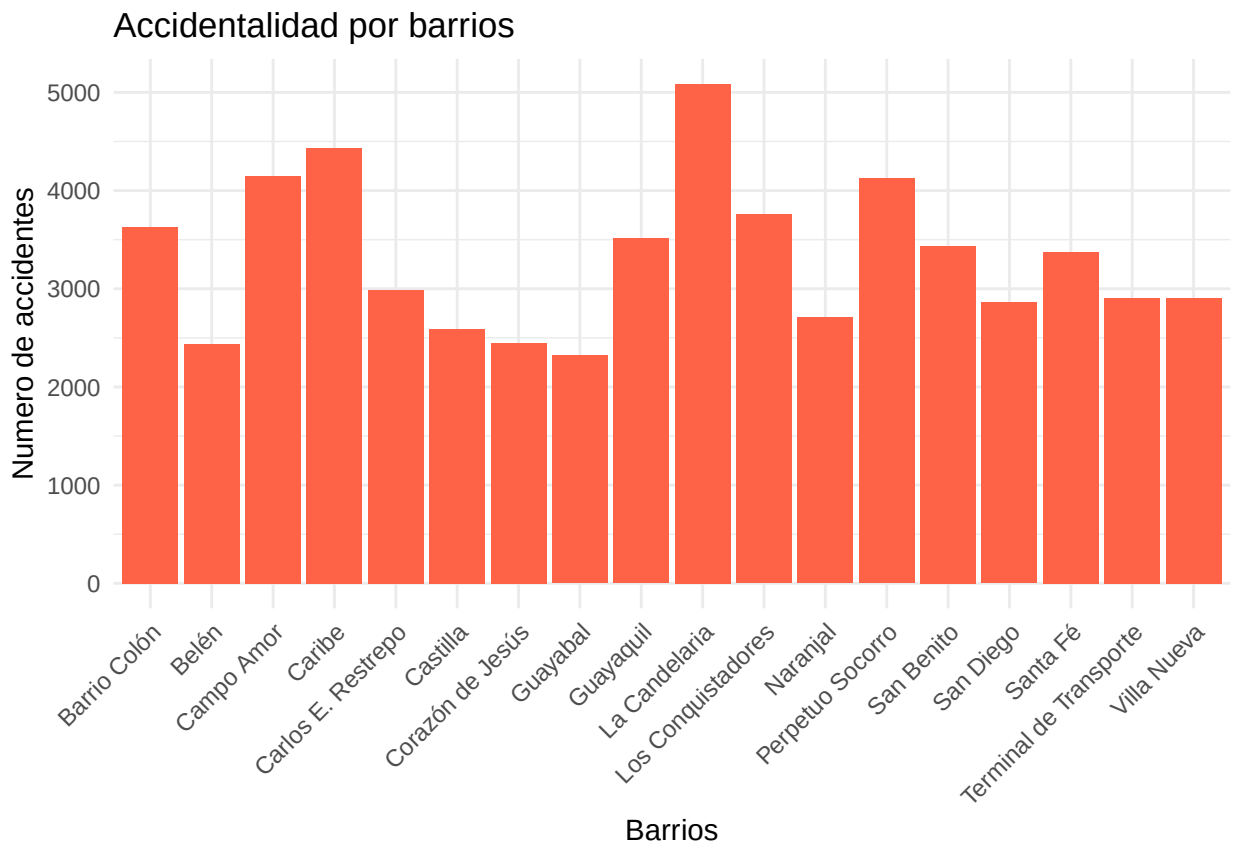
La Candelaria es donde ocurren la mayor cantidad de accidentes, con un número mayor a 40000. Las comunas en donde se presentan la menor cantidad de accidentes son las de Santa Cruz y Popular, seguidas por San Javier, con menos de 5000 accidentes en todos los casos. Otras comunas con valores de accidentalidad altos son las de Castilla y la de Laureles Estadio, por encima de los 20000 accidentes.
Se puede apreciar que la cantidad de accidentes difiere significativamente de una comuna a otra, con variaciones notables para cada clase de accidente. Se muestra una tabla de proporciones de estas a continuación:

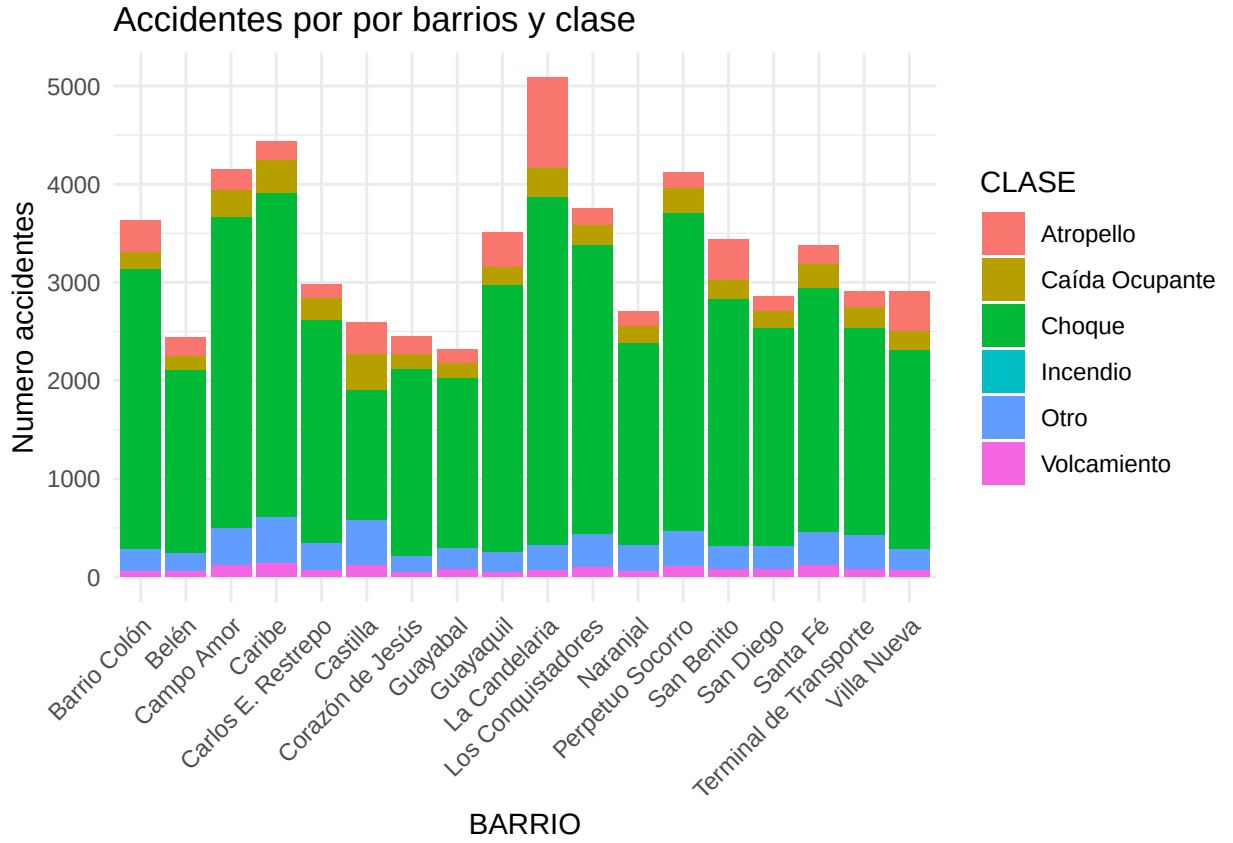
```
## Joining, by = "COMUNA"
```

COMUNA	Atropello	Caída Ocupante	Choque	Otro	Volcamiento	Incendio
Aranjuez	0.136	0.099	0.613	0.118	0.034	0
Belén	0.077	0.075	0.711	0.102	0.035	0
Buenos Aires	0.108	0.096	0.616	0.131	0.049	0
Castilla	0.074	0.104	0.659	0.127	0.036	NA
Doce de Octubre	0.196	0.190	0.411	0.169	0.033	NA
El Poblado	0.034	0.046	0.830	0.065	0.024	0
Guayabal	0.064	0.067	0.742	0.094	0.034	0
La América	0.084	0.086	0.699	0.105	0.026	0
La Candelaria	0.098	0.060	0.745	0.073	0.023	0
Laureles Estadio	0.061	0.068	0.755	0.092	0.024	0
Manrique	0.216	0.125	0.490	0.128	0.041	0
Popular	0.302	0.138	0.394	0.128	0.038	0
Robledo	0.084	0.152	0.569	0.159	0.036	NA
San Javier	0.173	0.141	0.498	0.142	0.046	NA
Santa Cruz	0.235	0.103	0.503	0.121	0.037	NA
Villa Hermosa	0.163	0.119	0.546	0.126	0.045	0

Se puede ver que hay variaciones de hasta el ~10% para Atropellos, del ~7% para “Otros” accidentes y del ~30% para los choques. Estas diferenccias para las comunas es esperable se repita para los barrios e incluso se marque más profundamente, esto indica diferencias entre la cantidad y proporciones de la CLASE de accidentes significativas lo que nos dice que un agrupamiento por este atributo es relevante.

Accidentalidad por barrios





En este gráfico se presentan los datos para 18 barrios de Medellín en donde hay mayor accidentalidad. El barrio que tiene una mayor accidentalidad es “La Candelaria”, superando los 5000 accidentes. Otros barrios con una alta accidentalidad son: “Caribe” y “Campor Amor” y “Naranjal”.

Se confirma que hay variaciones significativas en la accidentalidad por BARRIO y CLASE.

A continuación se muestra una tabla con las variaciones de las proporciones de accidentes según la clase y el tipo de Accidente.

Se denota que dato el bajo número de accidentes con Incendio, esta no es una categoría relevante para diferencias los barrios por grupos (Se expandirá en la sección 6).

```
## Joining, by = "BARRIO"
```

BARRIO	Atropello	Caída Ocupante	Choque	Otro	Incendio
Barrio Colón	0.088	0.047	0.785	0.064	NA
Belén	0.075	0.060	0.761	0.077	NA
Campo Amor	0.049	0.067	0.762	0.091	NA
Caribe	0.044	0.074	0.745	0.105	NA
Carlos E. Restrepo	0.047	0.078	0.758	0.092	NA
Castilla	0.124	0.140	0.512	0.177	NA
Corazón de Jesús	0.074	0.059	0.777	0.069	NA
Guayabal	0.059	0.068	0.746	0.092	NA
Guayaquil	0.099	0.052	0.776	0.058	0.000
La Candelaria	0.181	0.058	0.696	0.051	NA
Los Conquistadores	0.044	0.054	0.785	0.089	NA
Naranjal	0.054	0.067	0.758	0.099	NA
Perpetuo Socorro	0.039	0.060	0.788	0.086	NA
San Benito	0.119	0.058	0.730	0.069	0.000
San Diego	0.052	0.062	0.774	0.085	NA
Santa Fé	0.057	0.071	0.735	0.100	0.001
Terminal de Transporte	0.055	0.073	0.725	0.119	NA
Villa Nueva	0.138	0.068	0.695	0.073	0.000

5. Modelos predictivos

Como notamos en el análisis descriptivo existe una variabilidad notable a nivel de accidentalidad para cada columna y para cada barrio. Por ello se considera que crear un modelo predictivo específico para cada columna/barrio proveerá las mejores estimaciones de accidentalidad.

Dado que el objetivo es predecir la accidentalidad a nivel diario, mensual o semanal; se construirán modelos que predigan el total de accidentes en cada uno de los rangos temporales es decir se crearán modelos específicos para cada comuna/barrio según la categoría temporal.

En cada uno de estos casos se crearán dos modelos: el **Modelo 1** será entrenado con los datos correspondientes a los años 2014 - 2017 y los del año 2018 serán utilizados para su validación. El **Modelo 2** se construirá utilizando los datos de los años 2014 - 2016 y se validará con los correspondientes al 2017 - 2018.

La medida para evaluar la adecuación de los modelos será el error cuadrático medio en las predicciones tanto para los conjuntos de entrenamiento como para los de validación.

$$MSE = \frac{\sum_{n=1}^{N} (y_i - \hat{y}_i)^2}{N} = \frac{\sum_{n=1}^{N} e_i^2}{N}$$

Para elegir un modelo óptimo se involucrarán múltiples formulas con diferentes variables que puedan estar relacionadas a la accidentalidad y se elegirá la formula que mejor disminuya el MSE para los datos de validación como la óptima para el modelo.

En este reporte se mostraran los resultados de los MSE para los datos de entrenamiento y validación y la formula del modelo final siempre que sea posible, ya que en ciertos modelos al considerarse variables cualitativas, el número de parámetros impide su escritura. Este es el caso de los modelos por días y los modelos por semanas.

Los módelos completos pueden encontrarse en https://github.com/Nef997/Modelos-predictivos-de-accidentalidad-Medellin, en archivos tipo RData que contienen bases de datos con un resumen de los resultados y en su última columna el modelo ajustado. También se encuentran scripts para la realización de predicciones con cada modelo.

5.1. Modelos predictivos para Comunas

Para estos modelos se decidio ajustar una Regresión Poisson y una regresión Binomial negativa. Se considera que son óptimas porque el número de accidentes es un conteo y la cantidad de accidentes en las Comunas siempre es un numero positivo independiente de la escala de tiempo.

En los resultados se muestran tres tablas para cada modelo. Las dos primeras correspondientes al modelo 1 y al modelo 2 con tres columnas: 'Comuna' el nombre de la comuna, 'Family' el tipo de regresión ajustada (binomial negativa o poisson) y 'Formula' equivalente al parámetro *formula* ingresado a la función *gamlss* para crear el modelo final en los modelos por dia y semana mientras que en el modelo por año se muestra la ecuación final de la regresión. La tercera tabla incluye los resultados de los MSE para los datos de validación y entrenamiento de cada modelo.

5.1.1. Modelo por días

En base al análisis descriptivo encontramos que el total de accidentes cambia significativamente de un día de la semana a otro (especialmente el domingo). Adicionalmente vimos que los días especiales correspondientes a los días festivos de Colombia (tomados de Calendario Colombia) en los periodos de la base de datos, tenían un nivel de accidentalidad mucho menores a los de un día corriente. Por ende se considera que las

variables 'DIA_NUM' (1 = Lunes, 2=Martes, ... , 7=Domingo) y 'especial' (0 = normal, 1 = festivo). Adicionalmente se involucran las variables DIA, MES y AÑO.

Para la construcción de los modelos se consideran las siguientes relaciones:

- total~DIA_NUM+DIA+MES+PERIODO+I(PERIODOˆ2).
- total ~ DIA_NUM + especial + MES + DIA
- total ~ DIA_NUM + MES + especial
- total~DIA_NUM+especial+MES+PERIODO+I(PERIODOˆ2)
- total~DIA_NUM+especial+DIA+MES+PERIODO+I(PERIODOˆ2)

El término cuadrático "I(PERIODOˆ2)" se añade considerando que la accidentalidad anual ha cambiado de crecimiento a decrecimiento respecto al número de accidentes para algunas comunas.

Se construyen los modelos usando a MES como una variable cualitativa, debido a que al considerarla númerica y al considerar distintos tipos de relaciones con la variable respuesta: lineal,con termino cuadrático, termino cúbico y usando splines con hasta 4 nodos se obtenian por norma general errores mayores que al considerar la variable cualitativa.

Modelo 1:

Comuna	Family	Formula
Aranjuez	PO	total ~ DIA_NUM + especial + MES + PERIODO + I(PERIODO^2)
Belén	PO	total ~ DIA_NUM + especial + MES + DIA
Buenos Aires	PO	total ~ DIA_NUM + especial + DIA + MES + PERIODO + I(PERIODO^2)
Castilla	NBI	total ~ DIA_NUM + MES + especial
Doce de Octubre	NBI	total ~ DIA_NUM + MES + especial
El Poblado	PO	total ~ DIA_NUM + MES
Guayabal	PO	total ~ DIA_NUM + MES
La América	NBI	total ~ DIA_NUM + especial + MES + DIA
La Candelaria	PO	total ~ DIA_NUM + DIA + MES + PERIODO
Laureles Estadio	PO	total ~ DIA_NUM + I(PERIODO^2) + MES + PERIODO
Manrique	PO	total ~ DIA_NUM + especial + MES + PERIODO + I(PERIODO^2)
Popular	PO	total ~ DIA_NUM + MES + especial
Robledo	NBI	total ~ DIA_NUM + especial + MES + DIA
San Javier	PO	total ~ DIA_NUM + especial + MES + PERIODO + I(PERIODO^2)
Santa Cruz	NBI	total ~ DIA_NUM + MES + especial
Villa Hermosa	NBI	total ~ DIA_NUM + especial + DIA + MES + PERIODO + I(PERIODO^2)

Modelo 2:

Comuna	Family	Formula
Aranjuez	NBI	total ~ DIA_NUM + especial + MES + DIA
Belén	PO	total ~ DIA_NUM + especial + MES + DIA
Buenos Aires	NBI	total ~ DIA_NUM + especial + MES + DIA
Castilla	PO	total ~ DIA_NUM + MES + especial
Doce de Octubre	PO	total ~ DIA_NUM + MES + especial
El Poblado	PO	total ~ DIA_NUM + I(PERIODO^2) + MES + PERIODO
Guayabal	PO	total ~ DIA_NUM + MES
La América	PO	total ~ DIA_NUM + especial + MES + DIA
La Candelaria	NBI	total ~ DIA_NUM + I(PERIODO^2) + MES + PERIODO
Laureles Estadio	NBI	total ~ DIA_NUM + MES
Manrique	PO	total ~ DIA_NUM + especial + MES + PERIODO + I(PERIODO^2)
Popular	PO	total ~ DIA_NUM + MES + especial
Robledo	NBI	total ~ DIA_NUM + especial + MES + DIA
San Javier	NBI	total ~ DIA_NUM + especial + MES + DIA
Santa Cruz	PO	total ~ DIA_NUM + especial + MES + DIA
Villa Hermosa	PO	total ~ DIA_NUM + MES + especial

Tabla comparativa de los MSE

COMUNA	MSE.tr Modelo 1	MSE.va Modelo 1	MSE.tr Modelo 2	MSE.va Modelo 2
Aranjuez	7.243218	7.015269	7.346912	7.247382
Belén	9.227104	8.420567	9.001628	9.238715
Buenos Aires	4.444957	3.899620	4.395635	4.297949
Castilla	14.663598	13.600295	14.095461	15.336002
Doce de Octubre	3.012239	3.292249	2.925724	3.322395
El Poblado	13.359851	13.293340	11.683206	15.322878
Guayabal	11.325403	11.228509	11.209231	11.575311
La América	3.924999	3.497456	3.961746	3.652710
La Candelaria	47.138191	49.338104	47.269067	50.420493
Laureles Estadio	20.757111	19.398865	20.794055	20.386536
Manrique	3.383927	3.076613	3.441353	3.170467
Popular	1.293486	1.278194	1.342323	1.223310
Robledo	9.156078	8.111985	8.915954	9.037855
San Javier	1.800398	1.399231	1.791631	1.674507
Santa Cruz	1.302286	1.300221	1.308131	1.297059
Villa Hermosa	2.784805	2.595623	2.878685	2.646016

Es apreciable que el modelo 1 obtienes MSE más bajos para todas las predicciones al conjunto de validación y en general también ajusta mejor el conjunto de entrenamiento, salvo algunas comunas específicas.

5.1.2. Modelo por semanas

Se consideran relevantes los atributos SEMANA, MES y PERIODO.

Se construyen los modelos utilizando las siguientes relaciones:

- total~SEMANA+I(PERIODO^2)+MES+PERIODO
- total~SEMANA+MES
- total~SEMANA+MES+PERIODO
- total~SEMANA+I(PERIODO^2)+PERIODO

SEMANA es considerado como un atributo cualitativo. Al considerarlo número los MSE de validación siempre fueron superiores independiente de si se consideraba una relación lineal, cuadrática, cúbica o con splineas de hasta 6 nodos. Esto es equivalente para el modelo por semanas de Barrios.

Modelo 1:

Comuna	Family	Formula
Aranjuez	PO	total ~ SEMANA + I(PERIODO^2) + PERIODO
Belén	PO	total ~ SEMANA + I(PERIODO^2) + PERIODO
Buenos Aires	NBI	total ~ SEMANA + I(PERIODO^2) + PERIODO
Castilla	PO	total ~ SEMANA + MES
Doce de Octubre	PO	total ~ poly(SEMANA, degree = 3) + poly(MES, degree = 6) + PERIODO
El Poblado	PO	total ~ SEMANA + MES
Guayabal	PO	total ~ SEMANA + I(PERIODO^2) + PERIODO
La América	PO	total ~ SEMANA + I(PERIODO^2) + PERIODO
La Candelaria	NBI	total ~ SEMANA + I(PERIODO^2) + PERIODO
Laureles Estadio	PO	total ~ SEMANA + I(PERIODO^2) + PERIODO
Manrique	PO	total ~ poly(SEMANA, degree = 3) + poly(MES, degree = 5) + PERIODO
Popular	PO	total ~ SEMANA + MES
Robledo	PO	total ~ SEMANA + I(PERIODO^2) + PERIODO
San Javier	NBI	total ~ SEMANA + I(PERIODO^2) + PERIODO
Santa Cruz	PO	total ~ SEMANA + I(PERIODO^2) + PERIODO
Villa Hermosa	PO	total ~ poly(SEMANA, degree = 3) + poly(MES, degree = 5) + PERIODO

Modelo 2:

Comuna	Family	Formula
Aranjuez	PO	total ~ SEMANA + MES
Belén	PO	total ~ SEMANA + MES
Buenos Aires	PO	total ~ SEMANA + MES
Castilla	PO	total ~ SEMANA + I(PERIODO^2) + PERIODO
Doce de Octubre	PO	total ~ poly(SEMANA, degree = 3) + poly(MES, degree = 6) + PERIODO
El Poblado	NBI	total ~ SEMANA + I(PERIODO^2) + PERIODO
Guayabal	PO	total ~ SEMANA + MES
La América	PO	total ~ SEMANA + MES
La Candelaria	PO	total ~ SEMANA + MES
Laureles Estadio	PO	total ~ SEMANA + MES
Manrique	PO	total ~ SEMANA + I(PERIODO^2) + PERIODO
Popular	PO	total ~ SEMANA + MES + PERIODO
Robledo	PO	total ~ SEMANA + I(PERIODO^2) + PERIODO
San Javier	PO	total ~ poly(SEMANA, degree = 3) + poly(MES, degree = 5) + PERIODO
Santa Cruz	PO	total ~ SEMANA + MES + PERIODO
Villa Hermosa	PO	total ~ SEMANA + I(PERIODO^2) + PERIODO

Tabla comparativa de los MSE

COMUNA	MSE.tr Modelo 1	MSE.va Modelo 1	MSE.tr Modelo 2	MSE.va Modelo 2
Aranjuez	234.67365	204.97121	240.80234	219.16199
Belén	313.49910	302.49635	321.79669	299.09493
Buenos Aires	104.14199	96.45771	108.89066	94.48117
Castilla	650.92000	734.30055	634.42127	685.59023
Doce de Octubre	58.83977	69.54097	58.63815	64.43284
El Poblado	521.45450	535.86996	460.37566	580.00610
Guayabal	378.31304	408.20565	397.61707	354.10552
La América	90.69097	92.60792	89.78025	89.36114
La Candelaria	2811.16428	2773.84146	2899.04894	2688.13712
Laureles Estadio	936.11788	801.35216	957.00124	859.36394
Manrique	64.32883	54.94785	71.65620	53.11913
Popular	20.38913	19.61579	21.03340	18.23203
Robledo	275.35248	241.53196	284.13567	242.90620
San Javier	32.20650	25.99152	31.79782	30.59591
Santa Cruz	15.61891	17.55640	16.79652	14.73118
Villa Hermosa	57.72729	48.42322	62.56995	48.66858

En general los modelos tipo 1 vuelven a ser mejores para las predicciones de validación, con MSE considerablemente inferiores a los de los modelos tipo 2. Sin embargo cabe destacar que a la hora de ajustarse al modelo de entrenamiento se puede apreciar que ambos modelos son muy similares.

5.1.3. Modelo por meses

Se consideran relevantes los atributosMES y PERIODO.

Se construyen los modelos utilizando las siguientes relaciones:

- total~I(PERIODOˆ2)+MES+PERIODO
- total~MES
- total~MES+PERIODO
- total~MES+I(MESˆ2)+PERIODO
- total~MES+I(MESˆ2)

Modelo 1:

Comuna	Family	Formula
Aranjuez	NBI	total = 5773.01+5.72*MES-0.44*(MESˆ2)-2.77*PERIODO
Belén	NBI	total = 225.77+1.44*MES
Buenos Aires	NBI	total = -4995794.82-1.23*(PERIODOˆ2)+0.72*MES+4956.39*PERIODO
Castilla	NBI	total = 323.71+7.58*MES-0.37*(MESˆ2)
Doce de Octubre	PO	total = 76.92+8.04*MES-0.57*(MESˆ2)
El Poblado	NBI	total = 239.8+16.04*MES-1.11*(MESˆ2)
Guayabal	NBI	total = 219.97+10.66*MES-0.63*(MESˆ2)
La América	NBI	total = 119.27-0.25*MES
La Candelaria	NBI	total = 25622.24+10.11*MES-12.38*PERIODO
Laureles Estadio	PO	total = -18848808.76-4.64*(PERIODOˆ2)+1.72*MES+18698.23*PERIODO
Manrique	NBI	total = 13292.24+5.7*MES-0.37*(MESˆ2)-6.55*PERIODO
Popular	NBI	total = 41.9+2.64*MES-0.15*(MESˆ2)
Robledo	PO	total = 8137.43+1.31*MES-3.93*PERIODO
San Javier	PO	total = -383116.82-0.09*(PERIODOˆ2)+0.43*MES+380.68*PERIODO
Santa Cruz	NBI	total = 41.34+0.86*MES
Villa Hermosa	PO	total = 7091.2+2.54*MES-0.15*(MESˆ2)-3.48*PERIODO

Modelo 2:

Comuna	Family	Formula
Aranjuez	PO	total = 199.32+4.69*MES-0.32*(MES^2)
Belén	PO	total = 197.53+11.9*MES-0.79*(MES^2)
Buenos Aires	NBI	total = 112.28+7.04*MES-0.49*(MES^2)
Castilla	PO	total = -3680.01+0*(PERIODO^2)+3.39*MES+NA*PERIODO
Doce de Octubre	NBI	total = 79.64+7.6*MES-0.58*(MES^2)
El Poblado	PO	total = -22913.46+13.85*MES-0.93*(MES^2)+11.49*PERIODO
Guayabal	NBI	total = 216.4+9.46*MES-0.48*(MES^2)
La América	PO	total = 107.53+4.97*MES-0.43*(MES^2)
La Candelaria	PO	total = 607.99+42.91*MES-2.55*(MES^2)
Laureles Estadio	PO	total = 325.02+28.86*MES-2.03*(MES^2)
Manrique	NBI	total = 4153.6+0*(PERIODO^2)+1.3*MES+NA*PERIODO
Popular	NBI	total = 3462.3+0.29*MES-1.69*PERIODO
Robledo	PO	total = 7625.12+15.12*MES-1.1*(MES^2)-3.69*PERIODO
San Javier	PO	total = 58.19+2.68*MES-0.17*(MES^2)
Santa Cruz	NBI	total = 42.82+0.87*MES
Villa Hermosa	PO	total = 3889.91+3.51*MES-0.22*(MES^2)-1.89*PERIODO

Tabla comparativa de los MSE

COMUNA	MSE.tr Modelo 1	MSE.va Modelo 1	MSE.tr Modelo 2	MSE.va Modelo 2
Aranjuez	380.58204	365.68714	325.96359	608.64205
Belén	698.75055	361.96767	674.50574	404.02299
Buenos Aires	183.91018	205.73752	185.24970	157.94599
Castilla	880.66665	507.31641	760.04842	848.17376
Doce de Octubre	91.81969	225.77939	74.05467	184.01959
El Poblado	938.02139	920.62848	420.50662	1068.39054
Guayabal	734.11880	473.60893	835.01733	483.81730
La América	201.64521	282.95045	141.34489	283.78492
La Candelaria	2475.88978	1349.23659	1774.76313	3827.60815
Laureles Estadio	1299.73356	1694.64125	1052.26996	1607.20919
Manrique	142.14837	104.62895	178.35929	166.83340
Popular	71.13725	25.32347	66.07074	53.64801
Robledo	462.70803	243.17826	318.73479	328.52202
San Javier	65.41339	216.48104	59.87266	170.78959
Santa Cruz	47.98617	50.70220	45.07561	59.21788
Villa Hermosa	88.92000	187.54331	76.44104	171.98468

En las predicciones a nivel mensual se denota que ambos modelos tienden a ser relativamente igual de acertados con sus predicciones, aunque se denotan MES muy grandes para algunos barrios para conjuntos de validación comparados con los de entrenamiento y visceversa. Esto es probablemente debido a la falta de variables disponibles para predecir el número de accidentes mensualmente, contandose solo con MES y PERIODO. Pese a ello el modelo 1 sigue siendo mejor, obteniendo mejores MSE para 10 de las 16 columnas.

5.2. Modelos predictivos para Barrios

Al agrupar la base de datos por Barrios y las unidades temporales, se encuentra que para algunos barrios existe una sobrepoblación de ceros (no se registraron ocurrencias accidentes en la unidad temporal) para las agrupaciones por semana y por día.

Inicialmente se pensó en ajustar una regresión ZIP (Zero inflated Poisson), pero debido a problemas presentados se tuvo que descartar y se optó por una regresión de la familia Gaussiana. En general los barrios con una cantidad pequeña de accidentes ajustan modelos deficientes lo que los hace poco fiables. Las tablas resumen se presentan solo para los 40 barrios con mayor número de accidentes.

Las tablas mostradas en los resultados de barrios son analogas a las mostradas para las Comunas. Nuevamente solo se muestra la formula con el modelo final obtenido para el modelo de años.

5.2.1. Modelo por días

La lógica expuesta para la elección de variables en el modelo por Columnas es similar al de al modelo para Barrios. Por lo que los atributos elegidos son muy similares:

Se usaron los siguientes atibutos y relaciones:

- total~DIA_NUM+DIA+I(PERIODOˆ2)+MES+PERIODO
- total~DIA_NUM+DIA+MES+PERIODO
- total~DIA_NUM+I(PERIODOˆ2)+MES+PERIODO
- total~DIA_NUM+MES+DIA
- total~DIA_NUM+MES

Modelo 1:

Barrio	Formula
La Candelaria	total ~ DIA_NUM + MES + especial
Caribe	total ~ DIA_NUM + MES + especial
Campo Amor	total ~ DIA_NUM + MES + especial
Perpetuo Socorro	total ~ DIA_NUM + DIA + MES + especial
Los Conquistadores	total ~ DIA_NUM + MES + especial
Barrio Colón	total ~ DIA_NUM + MES + especial
Guayaquil	total ~ DIA_NUM + MES + especial
San Benito	total ~ DIA_NUM + MES + especial
Santa Fé	total ~ DIA_NUM + MES + especial
Carlos E. Restrepo	total ~ DIA_NUM + MES + especial
Villa Nueva	total ~ DIA_NUM + MES + especial
Terminal de Transporte	total ~ DIA_NUM + MES + especial
San Diego	total ~ DIA_NUM + MES + especial
Naranjal	total ~ DIA_NUM + MES + especial
Castilla	total ~ DIA_NUM + DIA + MES + especial
Corazón de Jesús	total ~ DIA_NUM + DIA + MES + especial
Belén	total ~ DIA_NUM + MES + especial
Guayabal	total ~ DIA_NUM + MES + especial
Boston	total ~ DIA_NUM + MES + especial
Villa Carlota	total ~ DIA_NUM + MES + especial
El Chagualo	total ~ DIA_NUM + MES + especial
Prado	total ~ DIA_NUM + MES + especial
Jesús Nazareno	total ~ DIA_NUM + MES + especial
Los Colores	total ~ DIA_NUM + MES + especial
Cristo Rey	total ~ DIA_NUM + MES + especial
Universidad Nacional	total ~ DIA_NUM + MES + especial
La Aguacatala	total ~ DIA_NUM + MES + especial

Manila	total ~ DIA_NUM + MES + especial
El Progreso	total ~ DIA_NUM + DIA + MES + especial
Suramericana	total ~ DIA_NUM + DIA + MES + especial
Calle Nueva	total ~ DIA_NUM + MES + especial
Las Acacias	total ~ DIA_NUM + DIA + MES + especial
Laureles	total ~ DIA_NUM + DIA + MES + especial
Rosales	total ~ DIA_NUM + MES + especial
El Estadio	total ~ DIA_NUM + MES + especial
Moravia	total ~ DIA_NUM + MES + especial
El Poblado	total ~ DIA_NUM + DIA + MES + especial
La Alpujarra	total ~ DIA_NUM + DIA + MES + especial
Sevilla	total ~ DIA_NUM + MES + especial
Bomboná No.1	total ~ DIA_NUM + MES + especial

Modelo 2:

Barrio	Formula
La Candelaria	total ~ DIA_NUM + DIA + MES + especial
Caribe	total ~ DIA_NUM + DIA + MES + especial
Campo Amor	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
Perpetuo Socorro	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
Los Conquistadores	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
Barrio Colón	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
Guayaquil	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
San Benito	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
Santa Fé	total ~ DIA_NUM + MES + especial
Carlos E. Restrepo	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
Villa Nueva	total ~ DIA_NUM + MES + especial
Terminal de Transporte	total ~ DIA_NUM + DIA + MES + especial
San Diego	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
Naranjal	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
Castilla	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
Corazón de Jesús	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
Belén	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
Guayabal	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
Boston	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
Villa Carlota	total ~ DIA_NUM + MES + especial
El Chagualo	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
Prado	total ~ DIA_NUM + DIA + MES + especial
Jesús Nazareno	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
Los Colores	total ~ DIA_NUM + MES + especial
Cristo Rey	total ~ DIA_NUM + MES + especial
Universidad Nacional	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
La Aguacatala	total ~ DIA_NUM + MES + especial
Manila	total ~ DIA_NUM + MES + especial
El Progreso	total ~ DIA_NUM + MES + especial
Suramericana	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
Calle Nueva	total ~ DIA_NUM + DIA + MES + especial
Las Acacias	total ~ DIA_NUM + DIA + MES + especial
Laureles	total ~ DIA_NUM + DIA + MES + especial

Rosales	total ~ DIA_NUM + DIA + MES + especial
El Estadio	total ~ DIA_NUM + DIA + MES + especial
Moravia	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
El Poblado	total ~ DIA_NUM + MES + especial
La Alpujarra	total ~ DIA_NUM + MES + especial
Sevilla	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
Bomboná No.1	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial

Tabla comparativa de los MSE

BARRIO	MSE.tr Modelo 1	MSE.va Modelo 1	MSE.tr Modelo 2	MSE.va Modelo 2
La Candelaria	2.9652103	3.7676305	3.1438256	3.9019730
Caribe	2.7830556	3.0244318	2.8918706	3.3322142
Campo Amor	2.1712030	4.4932215	2.2795269	3.2398490
Perpetuo Socorro	2.5064192	3.0026557	2.3968548	2.9066348
Los Conquistadores	2.1329290	2.2978201	2.2007787	2.4581423
Barrio Colón	2.0477035	2.7329148	1.9202998	2.4028761
Guayaquil	2.1800482	1.7211409	2.3084873	2.5677182
San Benito	2.1497510	1.6436564	2.2387672	2.4876397
Santa Fé	1.9697142	2.3576213	1.9611922	2.3896691
Carlos E. Restrepo	1.9367662	1.5621328	1.9528365	2.1445487
Villa Nueva	1.6952375	1.6173755	1.7930817	1.9653424
Terminal de Transporte	1.8305184	1.1289549	1.7700363	1.7224057
San Diego	1.6648429	1.7395165	1.6526912	1.8149144
Naranjal	1.6528106	1.1774410	1.6886942	1.7800532
Castilla	1.5304201	1.3326725	1.4739846	1.4652651
Corazón de Jesús	1.5029203	1.3826660	1.5467593	1.7330981
Belén	1.4430624	1.1914711	1.4154351	1.4475489
Guayabal	1.5237423	1.1081610	1.4507126	1.5622872
Boston	1.3607330	0.9831765	1.4562067	1.4202155
Villa Carlota	1.2084123	1.8495426	1.1918090	1.5777962
El Chagualo	1.2595178	0.8791213	1.3425871	1.3474929
Prado	1.2009588	1.0074807	1.2218122	1.2694544
Jesús Nazareno	1.1319222	1.2411610	1.1315364	1.2684708
Los Colores	1.1874051	1.7313699	0.9740225	1.2152734
Cristo Rey	1.1524780	0.8876266	1.1659016	1.1871923
Universidad Nacional	1.0875948	0.4724672	1.1494065	1.1057108
La Aguacatala	1.1159167	1.3070687	0.9334035	1.0969495
Manila	1.0784858	1.2406472	1.0441109	1.1982499
El Progreso	0.9094306	1.3993791	0.9000442	1.0684598
Suramericana	0.9434180	0.9921720	0.8735702	0.9422069
Calle Nueva	0.9365317	1.1949486	0.8558637	0.9802825
Las Acacias	0.8974175	0.9576906	0.9262887	0.9741000
Laureles	0.8469085	0.8851195	0.8106741	0.8804678
Rosales	0.9000540	1.2441023	0.8763946	0.9702912
El Estadio	0.9321658	0.4580227	0.9588564	0.9344847
Moravia	0.8637407	0.8419141	0.9202784	0.9168848
El Poblado	0.7713812	1.5266600	0.7142018	0.8872605
La Alpujarra	0.7997639	3.4071964	0.4250899	1.1291404

Sevilla	0.7323164	0.8390259	0.6541200	0.7107994
Bomboná No.1	0.7132065	0.6142187	0.7334932	0.7446466

Para los 18 barrios con más datos la media de los MSE son:

- Modelo 1 fue de 1.4429 para entrenamiento y 1.5810 para validación,
- Modelo 2 fueron 1.4336 y 1.6287.

Sin embargo, la media para entrenamiento y validación usando todos los datos son respectivamente:

- Modelo 1: 0.4652 y 0.9811.
- Modelo 2: 0.4387 y 0.5012.

Esto nos dice que el modelo 1 ajusta muy bien para cuando hay una cantidad de accidentalidad considerable, pero para barrios con poca accidentalidad comete sobre estimaciones respecto al modelo 2.

5.2.2. Modelo por semanas

Se utilizan:

- total~I(PERIODOˆ2)+MES+PERIODO+SEMANA
- total~MES+ SEMANA
- total~MES
- total~MES+PERIODO+SEMANA

Modelo 1:

Barrio	Formula
La Candelaria	total ~ MES + SEMANA
Caribe	total ~ MES + PERIODO + SEMANA
Campo Amor	total ~ MES + PERIODO + SEMANA
Perpetuo Socorro	total ~ MES + SEMANA
Los Conquistadores	total ~ MES + SEMANA
Barrio Colón	total ~ I(PERIODO^2) + MES + PERIODO + SEMANA
Guayaquil	total ~ MES + PERIODO + SEMANA
San Benito	total ~ MES + PERIODO + SEMANA
Santa Fé	total ~ MES + PERIODO + SEMANA
Carlos E. Restrepo	total ~ MES + SEMANA
Villa Nueva	total ~ MES + PERIODO + SEMANA
Terminal de Transporte	total ~ MES + SEMANA
San Diego	total ~ I(PERIODO^2) + MES + PERIODO + SEMANA
Naranjal	total ~ MES + PERIODO + SEMANA
Castilla	total ~ MES + SEMANA
Corazón de Jesús	total ~ MES
Belén	total ~ I(PERIODO^2) + MES + PERIODO + SEMANA
Guayabal	total ~ MES + SEMANA
Boston	total ~ MES + SEMANA
Villa Carlota	total ~ MES + PERIODO + SEMANA
El Chagualo	total ~ MES + PERIODO + SEMANA

Prado	total ~ MES + PERIODO + SEMANA
Jesús Nazareno	total ~ MES
Los Colores	total ~ MES + SEMANA
Cristo Rey	total ~ MES + SEMANA
Universidad Nacional	total ~ I(PERIODO^2) + MES + PERIODO + SEMANA
La Aguacatala	total ~ MES + SEMANA
Manila	total ~ MES + SEMANA
El Progreso	total ~ MES + PERIODO + SEMANA
Suramericana	total ~ MES + SEMANA
Calle Nueva	total ~ MES + SEMANA
Las Acacias	total ~ MES
Laureles	total ~ MES + SEMANA
Rosales	total ~ I(PERIODO^2) + MES + PERIODO + SEMANA
El Estadio	total ~ MES
Moravia	total ~ MES + SEMANA
El Poblado	total ~ MES + PERIODO + SEMANA
La Alpujarra	total ~ MES + PERIODO + SEMANA
Sevilla	total ~ MES + SEMANA
Bomboná No.1	total ~ MES + PERIODO + SEMANA

Modelo 2:

Barrio	Formula
La Candelaria	total ~ MES + SEMANA
Caribe	total ~ MES + PERIODO + SEMANA
Campo Amor	total ~ MES + PERIODO + SEMANA
Perpetuo Socorro	total ~ MES + SEMANA
Los Conquistadores	total ~ MES + SEMANA
Barrio Colón	total ~ MES + PERIODO + SEMANA
Guayaquil	total ~ MES + SEMANA
San Benito	total ~ MES + SEMANA
Santa Fé	total ~ MES + PERIODO + SEMANA
Carlos E. Restrepo	total ~ MES + SEMANA
Villa Nueva	total ~ MES + PERIODO + SEMANA
Terminal de Transporte	total ~ MES + PERIODO + SEMANA
San Diego	total ~ MES + SEMANA
Naranjal	total ~ MES + SEMANA
Castilla	total ~ MES + SEMANA
Corazón de Jesús	total ~ I(PERIODO^2) + MES + PERIODO + SEMANA
Belén	total ~ MES + SEMANA
Guayabal	total ~ MES + SEMANA
Boston	total ~ MES + SEMANA
Villa Carlota	total ~ MES + PERIODO + SEMANA
El Chagualo	total ~ MES + PERIODO + SEMANA
Prado	total ~ MES + PERIODO + SEMANA
Jesús Nazareno	total ~ MES + SEMANA
Los Colores	total ~ MES + PERIODO + SEMANA
Cristo Rey	total ~ MES + SEMANA
Universidad Nacional	total ~ MES + SEMANA
La Aguacatala	total ~ MES + SEMANA

Manila	total ~ MES + SEMANA
El Progreso	total ~ MES + PERIODO + SEMANA
Suramericana	total ~ MES + SEMANA
Calle Nueva	total ~ MES + SEMANA
Las Acacias	total ~ MES + SEMANA
Laureles	total ~ I(PERIODO^2) + MES + PERIODO + SEMANA
Rosales	total ~ MES + SEMANA
El Estadio	total ~ MES + SEMANA
Moravia	total ~ MES + PERIODO + SEMANA
El Poblado	total ~ I(PERIODO^2) + MES + PERIODO + SEMANA
La Alpujarra	total ~ MES + PERIODO + SEMANA
Sevilla	total ~ MES + PERIODO + SEMANA
Bomboná No.1	total ~ MES + SEMANA

Tabla comparativa de los MSE

BARRIO	MSE.tr Modelo 1	MSE.va Modelo 1	MSE.tr Modelo 2	MSE.va Modelo 2
La Candelaria	34.432712	23.959693	35.778451	33.105287
Caribe	30.033295	28.985951	30.567802	30.579182
Campo Amor	22.520516	42.261805	24.213225	27.419071
Perpetuo Socorro	23.956492	27.059272	23.662667	24.830089
Los Conquistadores	24.811853	15.864850	26.915253	24.516095
Barrio Colón	19.789844	20.009128	18.567317	19.306652
Guayaquil	25.546492	15.625453	27.299582	27.261738
San Benito	21.578732	17.588393	22.702996	23.751350
Santa Fé	16.922659	26.288013	16.835404	19.284876
Carlos E. Restrepo	17.189757	13.774031	17.632557	17.239686
Villa Nueva	16.524747	12.870685	17.150559	16.542254
Terminal de Transporte	16.053775	12.879737	14.038885	13.971701
San Diego	13.550313	14.574496	14.125264	14.188680
Naranjal	16.280879	14.042158	15.079438	14.959816
Castilla	12.249699	10.054970	12.192719	11.705280
Corazón de Jesús	17.878650	14.307582	10.093126	12.152621
Belén	10.944205	12.251257	12.264992	12.461238
Guayabal	13.167931	12.328489	11.485948	11.588316
Boston	11.916936	11.186711	12.706286	12.472931
Villa Carlota	10.742259	13.156852	10.468707	11.370427
El Chagualo	10.584019	8.869177	10.691293	10.728469
Prado	8.493779	7.885624	8.017098	8.206012
Jesús Nazareno	12.650501	9.161176	7.625590	8.406815
Los Colores	9.432541	12.266063	5.811801	7.134921
Cristo Rey	8.137703	9.436208	8.417689	8.629352
Universidad Nacional	8.495150	4.419754	8.942377	8.934574
La Aguacatala	8.591015	9.083827	6.332145	6.792777
Manila	7.142643	7.434296	6.527917	6.944439
El Progreso	7.436884	9.677457	7.228611	8.202095
Suramericana	6.552280	5.799942	6.104762	6.080501
Calle Nueva	5.131012	8.805541	4.234979	5.482098
Las Acacias	8.656902	6.052381	5.431008	6.237160

Laureles	5.982416	5.376812	5.490406	5.739213
Rosales	5.180873	8.412735	5.029903	6.088451
El Estadio	9.613167	5.207645	6.082661	6.313755
Moravia	4.384770	4.617647	4.058222	4.228921
El Poblado	4.165821	8.717593	3.563673	5.165281
La Alpujarra	5.533512	22.979350	2.096717	12.445590
Sevilla	4.698957	4.382465	3.894587	3.994200
Bomboná No.1	4.308870	4.396731	4.587666	4.821926

En este caso se encuentra que la media de los MSE de entrenamiento y validación son respectivamente para los 40 barrios con mayor accidentalidad son:

- Modelo 1: 13.0316 y 13.0513.
- Modelo 2: 12.3487 y 12.9821.

Usando todos los barrios:

- Modelo 1: 3.5185, 3.4095.
- Modelo 2: 2.9847, 3.2182.

Esto muestra que el modelo1 y el modelo2 tienen errores muy similares cuando la cantidad de accidentes es alta en los barrios. Sin embargo el modelo 2 tiene mejores predicciones en general.

5.2.3. Modelo por meses

Se utilizan:

- total~I(PERIODOˆ2)+MES+PERIODO
- total~MES
- total~MES+PERIODO

Modelo 1:

Barrio	Formula
La Candelaria	total = 79.41+0.92*MES
Caribe	total = 5766.8+0.37*MES-2.83*PERIODO
Campo Amor	total = -1028.99+0.44*MES+0.54*PERIODO
Perpetuo Socorro	total = 62.99+0.97*MES
Los Conquistadores	total = 597.33+0.57*MES-0.27*PERIODO
Barrio Colón	total = -2376.78+0.22*MES+1.21*PERIODO
Guayaquil	total = 8890.45+0.95*MES-4.38*PERIODO
San Benito	total = 8968.16+1.54*MES-4.43*PERIODO
Santa Fé	total = -4146.09+0.39*MES+2.08*PERIODO
Carlos E. Restrepo	total = 49.58+0.2*MES
Villa Nueva	total = 3203.33+0.53*MES-1.57*PERIODO
Terminal de Transporte	total = 1910.51+0.6*MES-0.93*PERIODO
San Diego	total = -3895117.37-0.96*(PERIODOˆ2)+0.48*MES+3864.13*PERIODO
Naranjal	total = 2111.38+0.11*MES-1.03*PERIODO
Castilla	total = 39.39+0.56*MES

Corazón de Jesús	total = 3328.88+0.77*MES-1.63*PERIODO
Belén	total = -7874159.9-1.94*(PERIODO^2)+0.32*MES+7811.85*PERIODO
Guayabal	total = 36.38+0.73*MES
Boston	total = 71.72+0.12*MES-0.02*PERIODO
Villa Carlota	total = -2350.39+0.24*MES+1.18*PERIODO
El Chagualo	total = 5809.14+0.73*MES-2.87*PERIODO
Prado	total = 4755.95-0.21*MES-2.34*PERIODO
Jesús Nazareno	total = 1256.43+0.52*MES-0.61*PERIODO
Los Colores	total = 26.86+0.71*MES
Cristo Rey	total = 30.75+0.49*MES
Universidad Nacional	total = -7020611.43-1.73*(PERIODO^2)+0.1*MES+6968.46*PERIODO
La Aguacatala	total = 31.83-0.03*MES
Manila	total = 27.42+0.38*MES
El Progreso	total = -2590.98-0.1*MES+1.3*PERIODO
Suramericana	total = 28.08-0.01*MES
Calle Nueva	total = 22.97+0.62*MES
Las Acacias	total = -3045072.16-0.75*(PERIODO^2)-0.31*MES+3022.47*PERIODO
Laureles	total = 23.48+0.36*MES
Rosales	total = 22.67+0.33*MES
El Estadio	total = 3538.07-0.48*MES-1.74*PERIODO
Moravia	total = 22.07+0.13*MES
El Poblado	total = -3272.44+0.26*MES+1.63*PERIODO
La Alpujarra	total = -13880.34+1.15*MES+6.89*PERIODO
Sevilla	total = 21.86-0.08*MES
Bomboná No.1	total = 4405.27-0.02*MES-2.18*PERIODO

Modelo 2:

Barrio	Formula
La Candelaria	total = 83.41+0.71*MES
Caribe	total = 73.45+0.4*MES
Campo Amor	total = -4891.55+0.61*MES+2.46*PERIODO
Perpetuo Socorro	total = 60.73+1.14*MES
Los Conquistadores	total = 58.84+0.94*MES
Barrio Colón	total = -1452.04+0.02*MES+0.75*PERIODO
Guayaquil	total = 56.4+1.5*MES
San Benito	total = 52.43+1.74*MES
Santa Fé	total = -1880.5+0.53*MES+0.96*PERIODO
Carlos E. Restrepo	total = 49.95+0.04*MES
Villa Nueva	total = 4244.71+0.46*MES-2.08*PERIODO
Terminal de Transporte	total = 46.53+0.41*MES
San Diego	total = 45.08+0.45*MES
Naranjal	total = 2227.84+0.28*MES-1.08*PERIODO
Castilla	total = 37.29+0.67*MES
Corazón de Jesús	total = 38.06+0.94*MES
Belén	total = 38.63+0.41*MES
Guayabal	total = 34.22+0.74*MES
Boston	total = 40.07+0.01*MES
Villa Carlota	total = 32.79+0.51*MES
El Chagualo	total = 4146.61+0.8*MES-2.04*PERIODO

Prado	total = 5156.72+0.13*MES-2.54*PERIODO
Jesús Nazareno	total = 2213.45+0.51*MES-1.08*PERIODO
Los Colores	total = -3671.07+0.55*MES+1.83*PERIODO
Cristo Rey	total = 28+0.82*MES
Universidad Nacional	total = 32.37+0.31*MES
La Aguacatala	total = -3581.67-0.01*MES+1.79*PERIODO
Manila	total = 25.21+0.62*MES
El Progreso	total = -559.83-0.05*MES+0.29*PERIODO
Suramericana	total = 27.35+0.03*MES
Calle Nueva	total = 23.88+0.39*MES
Las Acacias	total = 30.23-0.48*MES
Laureles	total = 22.86+0.48*MES
Rosales	total = 21.9+0.41*MES
El Estadio	total = 26.49-0.05*MES
Moravia	total = 3465.88+0.2*MES-1.71*PERIODO
El Poblado	total = -989.43+0.31*MES+0.5*PERIODO
La Alpujarra	total = -1838.27+0.46*MES+0.92*PERIODO
Sevilla	total = 20.3+0*MES
Bomboná No.1	total = 1870.21-0.03*MES-0.92*PERIODO

Tabla comparativa de los MSE

BARRIO	MSE.tr Modelo 1	MSE.va Modelo 1	MSE.tr Modelo 2	MSE.va Modelo 2
La Candelaria	182.58056	63.49340	186.72203	144.68007
Caribe	108.33147	85.64718	111.01616	146.22889
Campo Amor	109.63360	396.12140	130.09194	156.50047
Perpetuo Socorro	68.87113	77.98177	73.36221	66.38507
Los Conquistadores	98.96794	124.98901	90.59925	146.20488
Barrio Colón	75.84976	66.05773	64.11223	91.36388
Guayaquil	124.99830	105.31692	111.66815	435.65910
San Benito	148.55556	73.20893	129.09605	347.49749
Santa Fé	79.90687	84.89121	75.96621	95.21253
Carlos E. Restrepo	88.62193	73.25802	92.65186	73.16992
Villa Nueva	89.28820	51.47742	97.15532	62.89775
Terminal de Transporte	70.09154	88.32271	64.95463	99.53177
San Diego	53.35934	50.28521	59.24416	51.21527
Naranjal	78.33684	58.69759	88.00128	54.04089
Castilla	74.24531	27.59511	71.59201	62.05562
Corazón de Jesús	80.45399	79.12091	73.73730	136.21534
Belén	56.26410	35.51127	72.52841	41.65055
Guayabal	84.01376	171.36653	76.85233	123.91655
Boston	42.81930	53.15844	46.43152	50.18413
Villa Carlota	50.86738	58.39180	47.21551	78.17098
El Chagualo	49.89004	42.40024	58.73097	44.37698
Prado	42.19637	40.05016	38.24402	46.63212
Jesús Nazareno	23.94689	44.16587	23.16819	35.64685
Los Colores	118.69083	67.29043	42.26359	181.06910
Cristo Rey	38.81791	124.11639	41.42165	72.40954
Universidad Nacional	26.44580	18.03701	27.42562	107.03770

La Aguacatala	68.05591	77.84590	32.49676	117.39608
Manila	40.57674	25.69030	32.12577	46.08007
El Progreso	36.23392	33.81277	34.38387	54.92383
Suramericana	32.78960	31.04315	33.57400	30.25386
Calle Nueva	29.70160	25.55074	23.67536	38.88019
Las Acacias	27.63666	50.57375	31.51072	46.94509
Laureles	31.69757	35.96618	30.39219	37.56002
Rosales	28.72463	17.86298	24.98617	29.61051
El Estadio	46.14909	22.76338	37.79868	101.20216
Moravia	28.48521	32.77829	16.78538	28.65716
El Poblado	12.15828	24.04987	12.44129	33.79275
La Alpujarra	78.63108	102.77237	22.90343	647.76145
Sevilla	33.54203	35.30063	25.21218	47.83736
Bomboná No.1	24.03713	17.21512	27.31629	26.52744

En este caso para los 40 barrios con mayor accidentalidad se encuentra que la media de los MSE de entrenamiento y validación son respectivamente:

- Modelo 1: media MSE.tr = 64.6116, media MSE.va = 67.3544.
- Modelo2: media MSE.tr = 59.4963, media MSE.va = 105.9345.

Usando todos los barrios:

- Modelo 1: media MSE.tr = 17.1035, media MSE.va = 21.98968.
- Modelo2: media MSE.tr = 16.0569, media MSE.va = 26.3365.

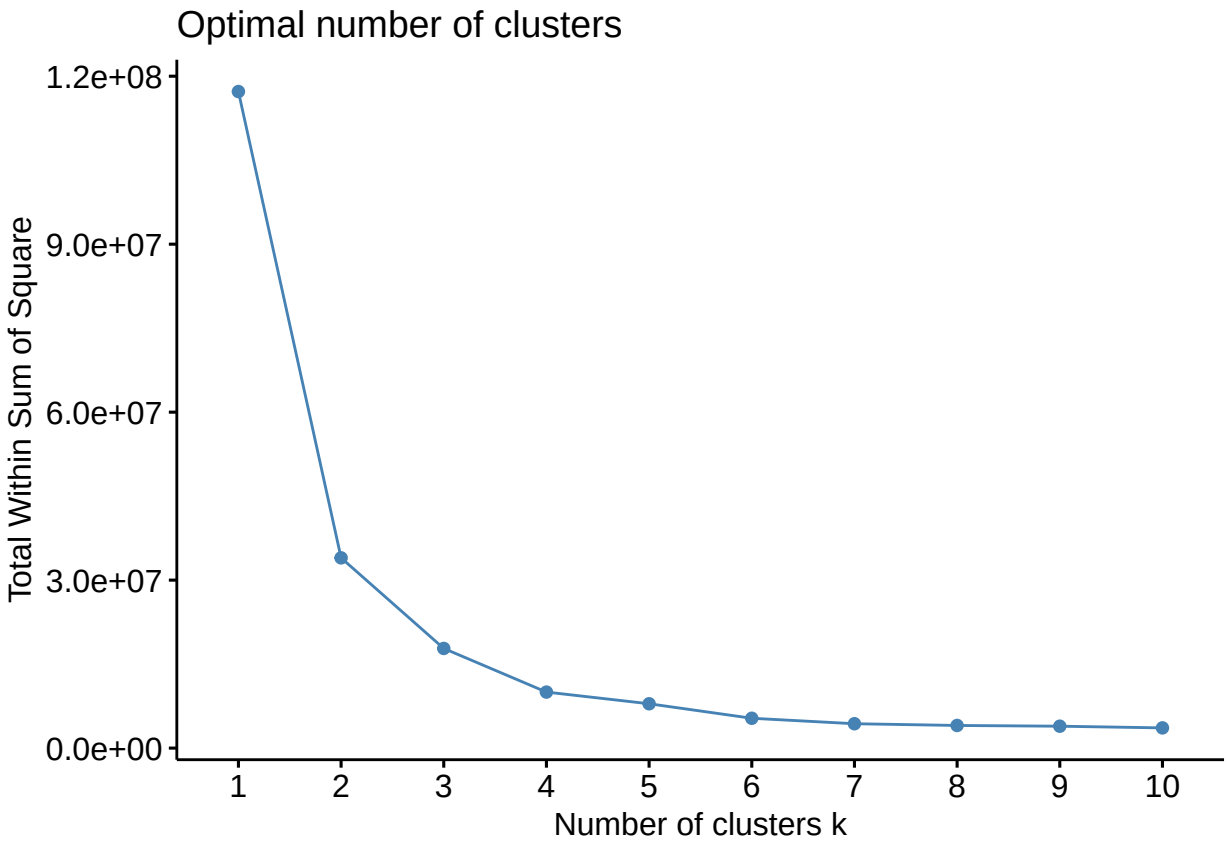
Se puede apreciar que el modelo1 es mejor prediciendo la accidentalidad de los barrios, independiente de la cantidad de accidentes de estos.

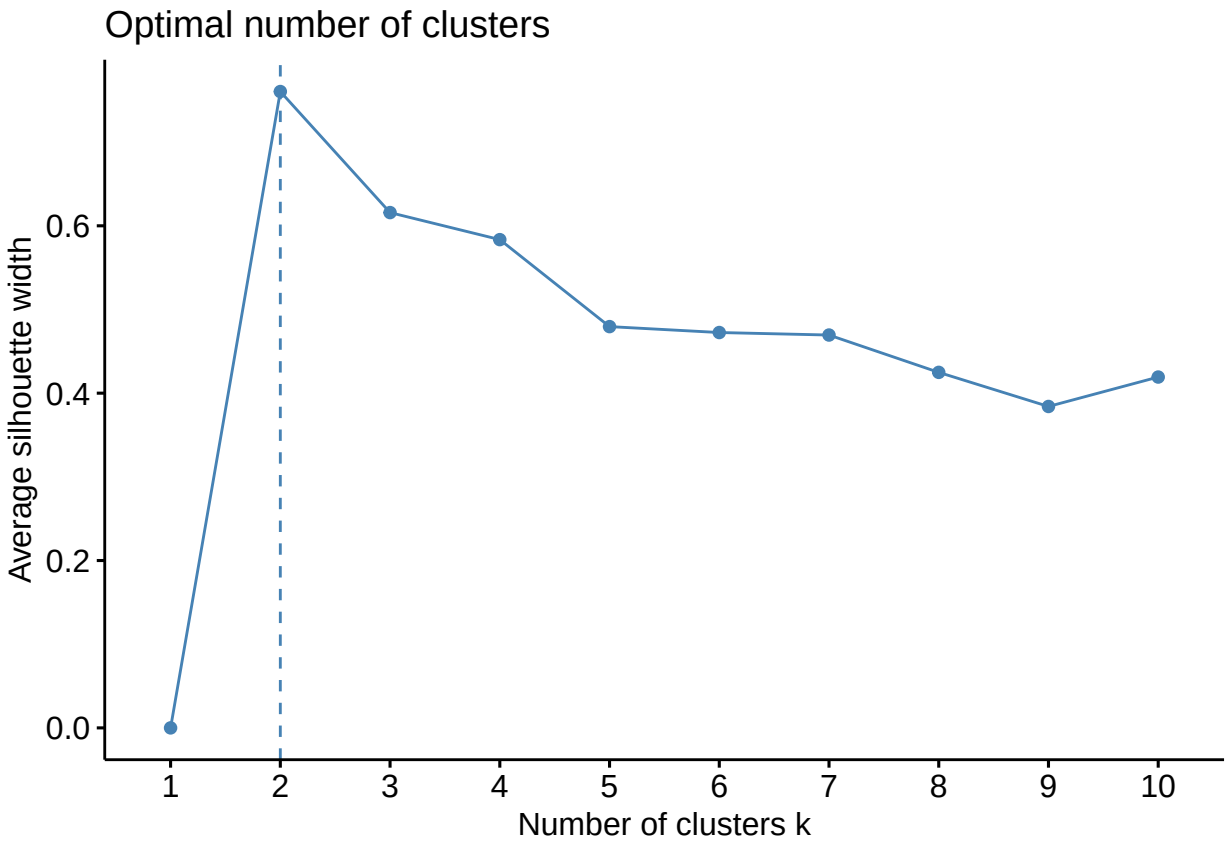
Se considera que tanto para barrios como para columnas en el modelo anual, la superioridad muy marcada del modelo 1 para predecir los datos de validación frente al modelo 2, es debida a que posee datos del año previo a sus datos de validación, al solo tener MES y PERIODO involucradas, el peso de esta última variable es mayor en las estimaciones y el modelo entrenado con solo los datos del 2014 al 2016 probablemente falla en notar la relación para los años y fracasa en predecir los datos correspondientes al 2018.

6. Agrupamiento

Se considera que el tipo de clase “Incendio” es poco relevante para la diferenciación de los grupos dada su baja ocurrencia para todos los barrios y se excluye del análisis. Para realizar la agrupación por CLASE de los barrios se decide utilizar el método kmeans.

Se construyen los gráfico wss y silhouette para determinar el número óptimo de clusters a realizar, se obtiene como resultado que para grupos de 2 a 6 existen valores significativos que soportan su escogencia.


Optimal number of clusters
Total Within Sum of Square
1.2e+08
9.0e+07
6.0e+07
3.0e+07
0.0e+00
1
2
3
4
5
6
7
8
9
10
Number of clusters k



Para la consideración del trabajo se decide probar con 3, 4 y 5 clusters para la elección del más óptimo. Tras un análisis se concluyo que el cluster con 3 grupos presentaba diferencias notables entre los grupos pero las desviaciones intragrupos eran muy grandes, especialmetende un grupo donde encasillaba a la mayoria de los barrios. El cluster con 5 se descartó en consideraciión a que no aportaba mucha información adicional respecto al cluster con 4.

Medias de los grupos formados:

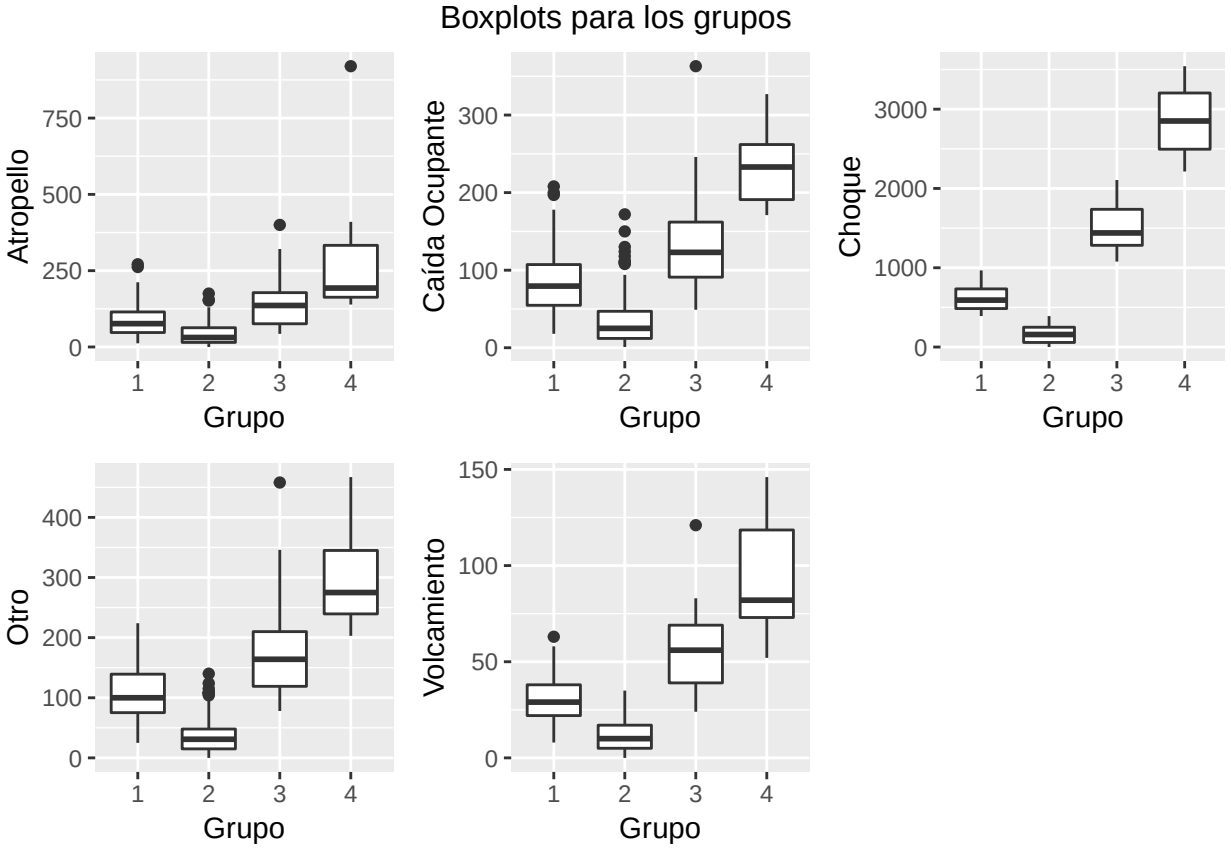
C4G	elementos	Atropellos	Caidas	Choques	Volcamientos	Otros
1	70	87.57143	86.97143	613.1714	30.28571	105.77143
2	161	43.69565	35.37267	163.7640	11.19876	37.39752
3	25	140.52000	138.40000	1506.6000	56.04000	179.80000
4	11	290.90909	231.81818	2840.7273	93.36364	301.72727

Desviaciones estandar:

C4G	Atropellos	Caidas	Choques	Volcamientos	Otros
1	55.99135	45.41116	157.5781	12.366555	44.64455
2	36.24139	31.18249	107.8792	8.200473	28.68564
3	84.99712	69.03622	313.7730	22.227011	85.77296
4	227.30264	50.96826	442.8619	30.738486	79.51615

Para el clustering en 4 grupos se tiene la siguiente distribución de datos por grupo: 26% en el grupo 1, 60% en el grupo 2, 9% en el grupo 3 y 5% en el grupo 4. Se nota una intercección entre los cuantiles de todos

los grupos en los accidentes por atropello, pero las diferencias entre estas siguen claras en los otros tipos de accidentalidad, en especial en choques donde las diferencias son más marcadas y la desviación estándar es menor. Es notable que el grupo 4 no se sobreponga a otros en los diferentes tipos de accidentalidad, con excepción en atropellos donde su desviación estándar se mantiene más baja. Se mantiene una diferencia considerable entre el grupo 4 y los otros grupos. Cabe notar que los grupos 1 y 3 suelen interceptarse sobre todos los tipos de accidentalidad excepto en choques donde todos los grupos son significativamente diferentes los unos de los otros. La principal diferencia entre los grupos yace en la media de la cantidad de accidentes que tienen, siendo el primer grupo donde menos accidentes se tienen y el 4 grupo donde los accidentes son mayores.



Los barrios pertenecientes al tercer grupo son los que tienen los mayores niveles de accidentalidad para cada uno de los tipos. Especialmente son aquellos con una marcada superioridad en la cantidad de accidentes por choques. Los grupos 1 y 2 tienen barrios similares para accidentes de clase "Otro", "Caída ocupante" y "Atrpello", pero se diferencian notablemente en su cantidad de volcamientos y choques. Además los barrios del grupo 2 podrían considerarse de baja accidentalidad. El grupo 4 correspondería a un grupo intermedio entre 2 y 3 respecto a la cantidad de choques, con un número de elementos algo bajo su dispersión es pequeña.

7. Conclusiones

- Los modelos ajustados logran la predicción de la accidentalidad, con errores relativamente bajos especialmente para comunas.
- Se presentan estrategias concretas de mejora basadas en la literatura al momento de la realización del mismo
- Los modelos predictivos que utilizan el ~80% de los datos tienden a ser superiores a los modelos que utilizan tan solo un ~60% de los datos disponibles, especialmente cuando se enfrentan a datos

desconocidos (predicen mucho mejor los datos de validación), como se vio en el ajuste de los modelos predictivos de la sección 5.
- La extrapolación para la predicción de accidentalidad anualmente disminuye la precisión de los modelos.

8. Recomendaciones

- Un gran problema a la hora de predecir la accidentalidad es la falta de datos concretos a cada tipo de accidente particular; estudios han demostrado que la velocidad límite de la vía en donde ocurre el accidente es un buen factor de predicción de la letalidad en accidentes de peatones (Nishimoto, Kubota & Ponte, 2019), lo que nos lleva de nuevo al problema de falta de datos de características propias de la locación concreta del accidente, como una cuantificación del número de intersecciones o vías con alta velocidad.

- Se propone hacer modelos a menor escala de localidad (calles, intersecciones, avenidas), esto podría mejor la identificación de factores únicos relacionados con la accidentalidad y mejorar los modelos (Zhang & Shi 2019).

9. Bibliografía

- Barajas, F., Torres, M., Arteaga, L., & Castro, C. (2015). GAMLSS models applied in the treatment of agro-industrial waste. Comunicaciones En Estadística, 8(2), 245. doi: 10.15332/s2027-3355.2015.0002.07
- Espinosa López, A., Cabrera Arana, G., & Velásquez Osorio, N. (2017). Epidemiología de incidentes viales Medellín-Colombia, 2010-2015. Revista Facultad Nacional De Salud Pública, 35(1), 7-15. doi: 10.17533/udea.rfnsp.v35n1a02
- Hyder, A., & Vecino-Ortiz, A. (2014). BRICS: opportunities to improve road safety. Bulletin Of The World Health Organization, 92(6), 423-428. doi: 10.2471/blt.13.132613
- Erdogan, S., Yilmaz, I., Baybura, T., & Gullu, M. (2008). Geographical information systems aided traffic accident analysis system case study: city of Afyonkarahisar. Accident Analysis & Prevention, 40(1), 174-181. doi: 10.1016/j.aap.2007.05.004
- OpenData Alcaldía de Medellín. (2019). Retrieved from https://geomedellin-m-medellin.opendata.arcgis.com/search?tags=movilidad
- Zhang, J., & Shi, T. (2019). Spatial analysis of traffic accidents based on WaveCluster and vehicle communication system data. Eurasip Journal on Wireless Communications and Networking, 2019(1) doi:10.1186/s13638-019-1450-0
- Nishimoto, T., Kubota, K., & Ponte, G. (2019). A pedestrian serious injury risk prediction method based on posted speed limit. Accident Analysis & Prevention, 129, 84-93. doi: 10.1016/j.aap.2019.04.021
- Lehmann, E. L.; Casella, George (1998). Theory of Point Estimation (2nd ed.). New York: Springer
- Rigby R.A. and Stasinopoulos D.M. (2005). Generalized additive models for location, scale and shape,(with discussion), Appl. Statist., 54, part 3, pp 507-554.
- Maechler, M.., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2019). cluster: Cluster Analysis Basics and Extensions. R package version 2.1.0.
- Mundt, F., Kassambara, A. (2017). factoextra: Extract and Visualize the Results of Multivariate Data Analyses (1.0.5).
- Friedrich, L., A (2006.). Toolbox for K-Centroids Cluster Analysis. Computational Statistics and Data Analysis, 51 (2), 526-544.