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.

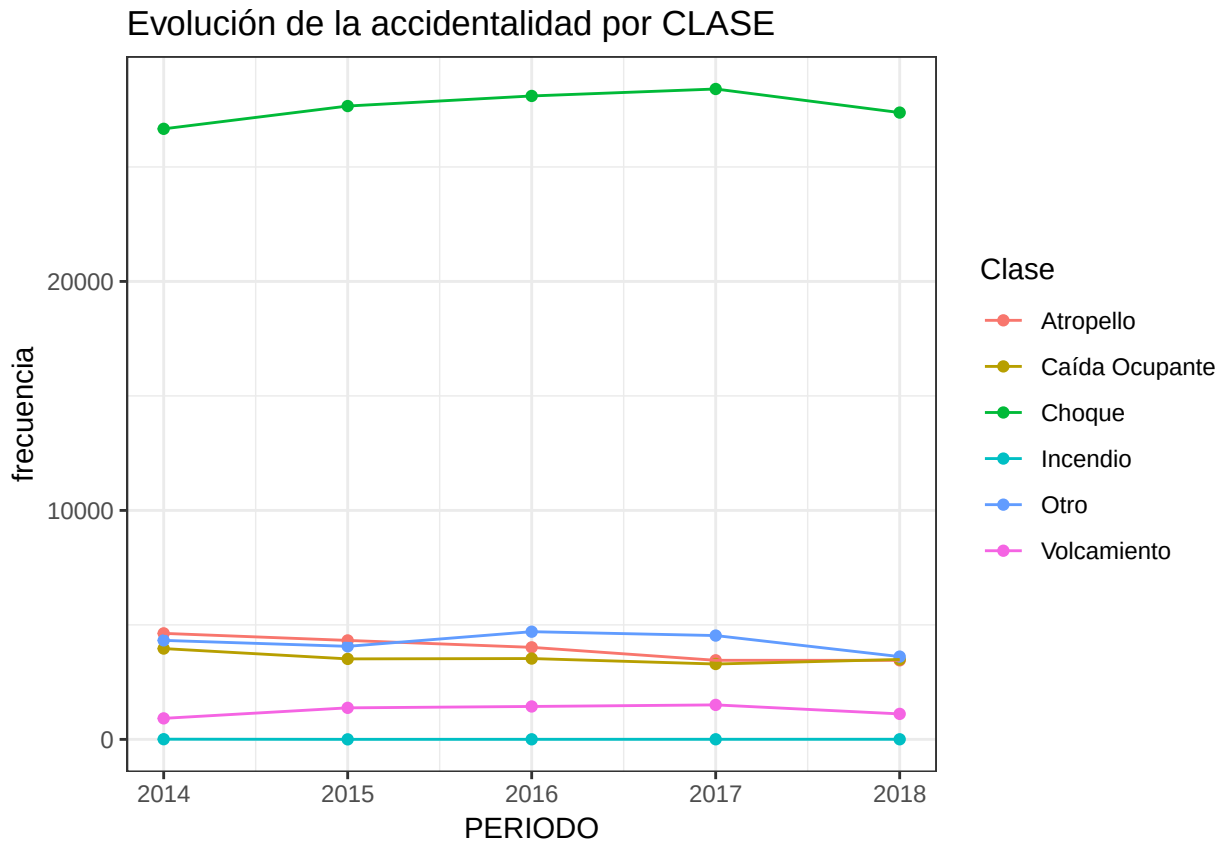
Tras las modificaciones mencionadas de los 209426 registros originales entre las bases de datos del 2014 al 2018, se utilizarán 203507 restantes en el análisis y en la cosntrucción de los modelos.

4.3 Análisis

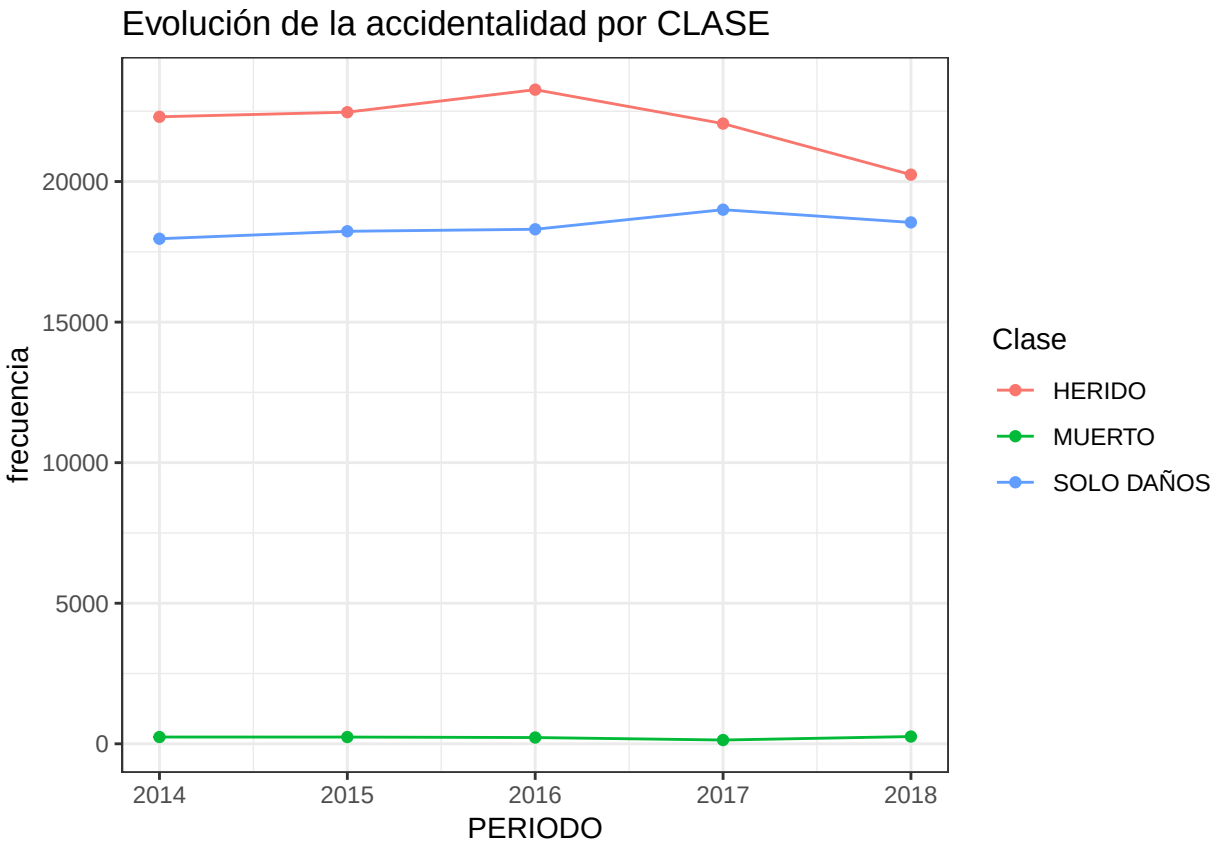
Se considera que las siguientes variables no son de interes para la construcción de los modelos:

- 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

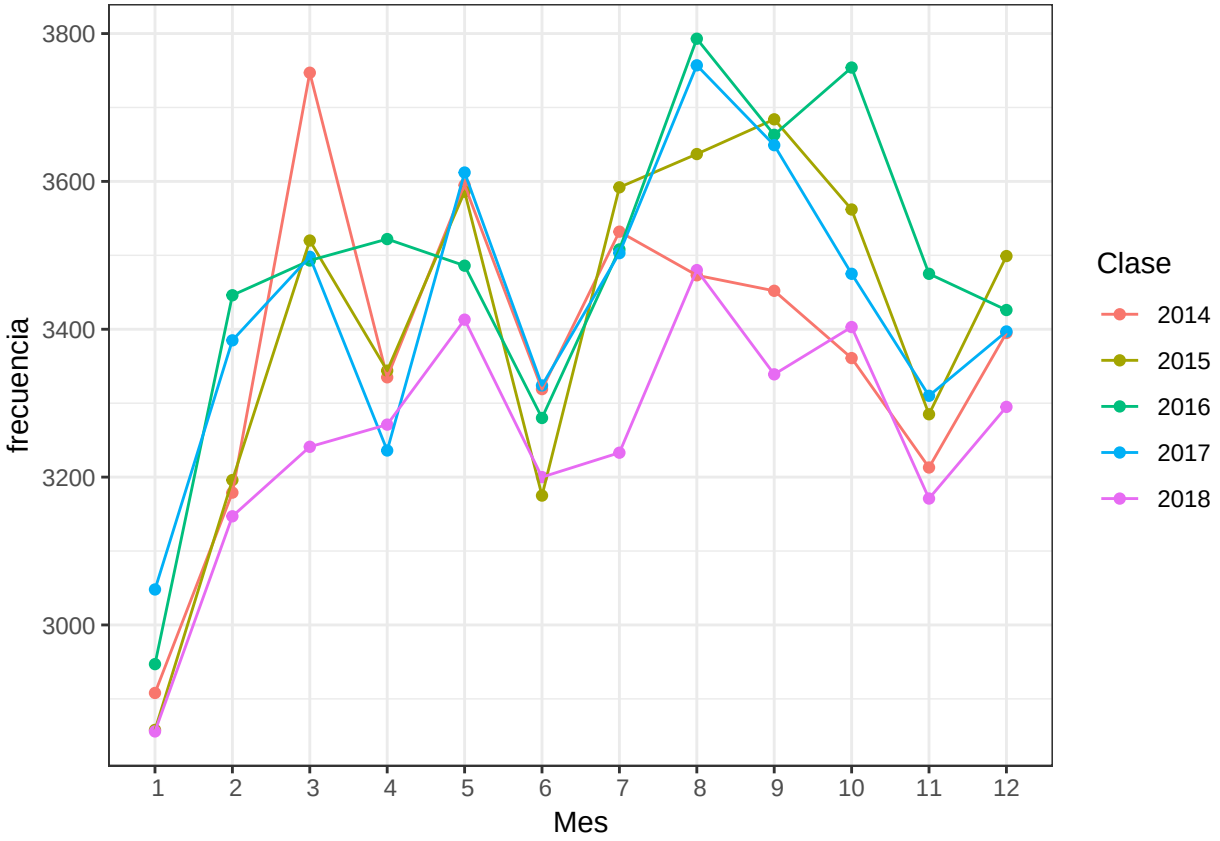


Se evidencia que la evolución de la accidentalidad por cada clase no tiene un patrón de dependencia claro con la Clase de los accidentes. Los choques son el tipo de accidente más frecuente y tuvieron un leve crecimiento entre el 2014 y 2017 y luego decrecieron para el 2018. Las otras clases de accidentes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



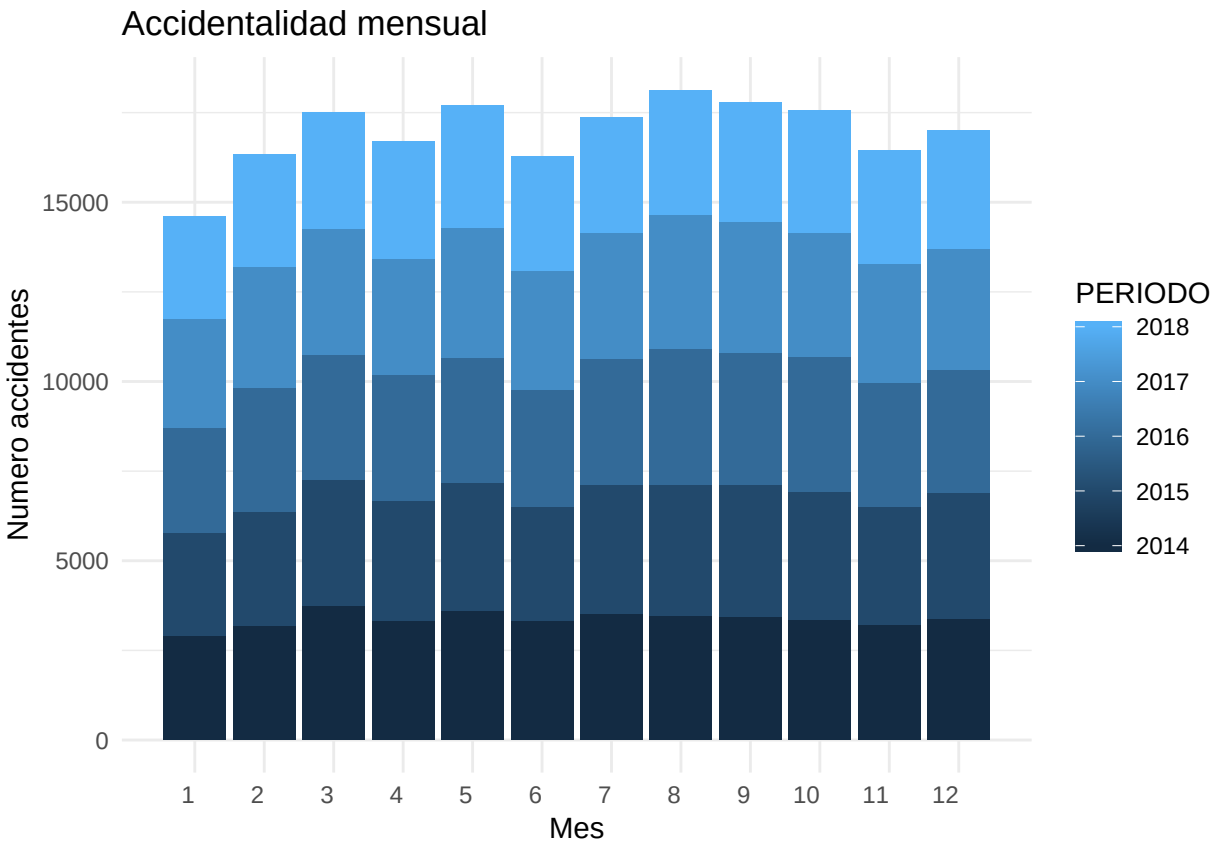
En general se denota una leve disminución en la cantidad de heridos desde el 2016 hasta el 2018, los accidentes con muertos y aquellos donde solo ocurrieron daños materiales no presentan cambios significativos.

Evolución accidentalidad por meses

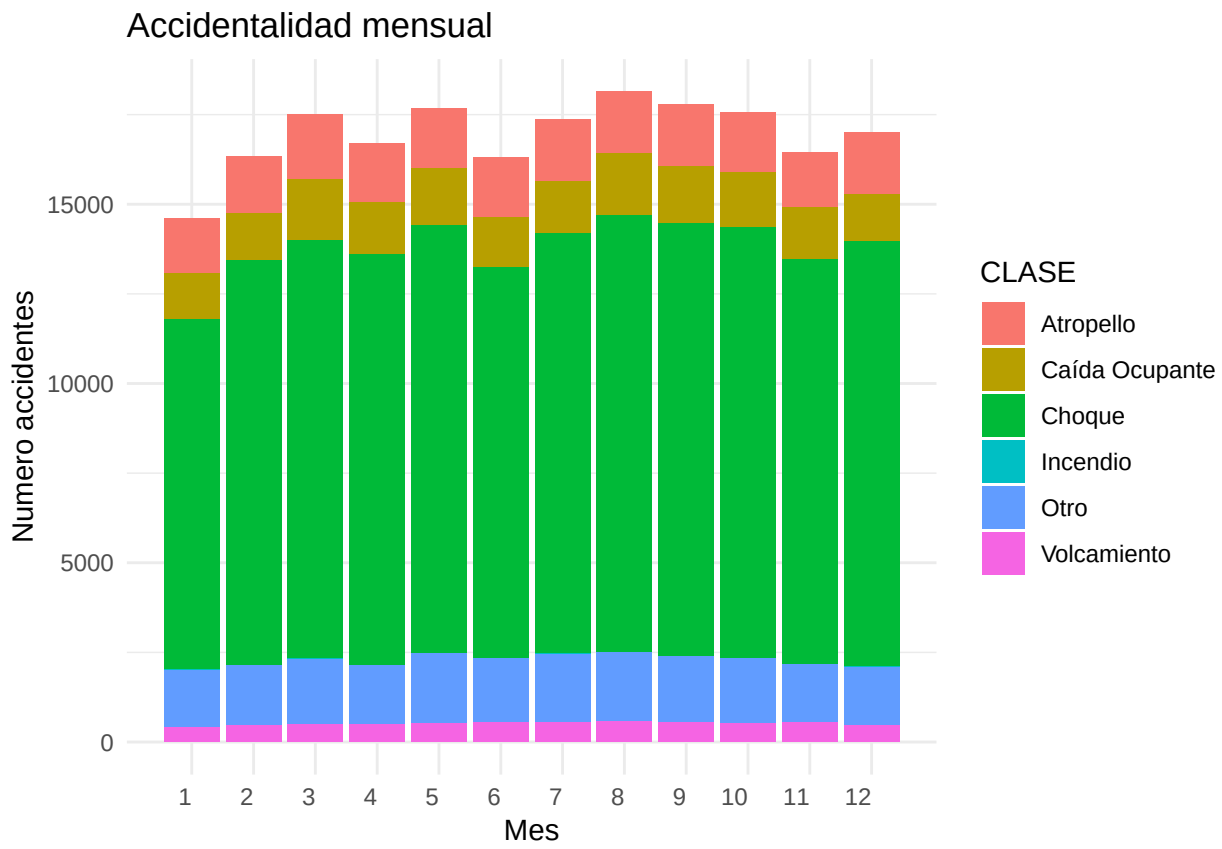


Se puede apreciar que en terminos generales la accidentalidad para el año 2018 es siempre una de las dos más pequeñas independiente del mes. Esto y las leves diferencias de accidentalidad anual nos indican que el PERIODO podría ser relevante para los modelos predictivos.

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.

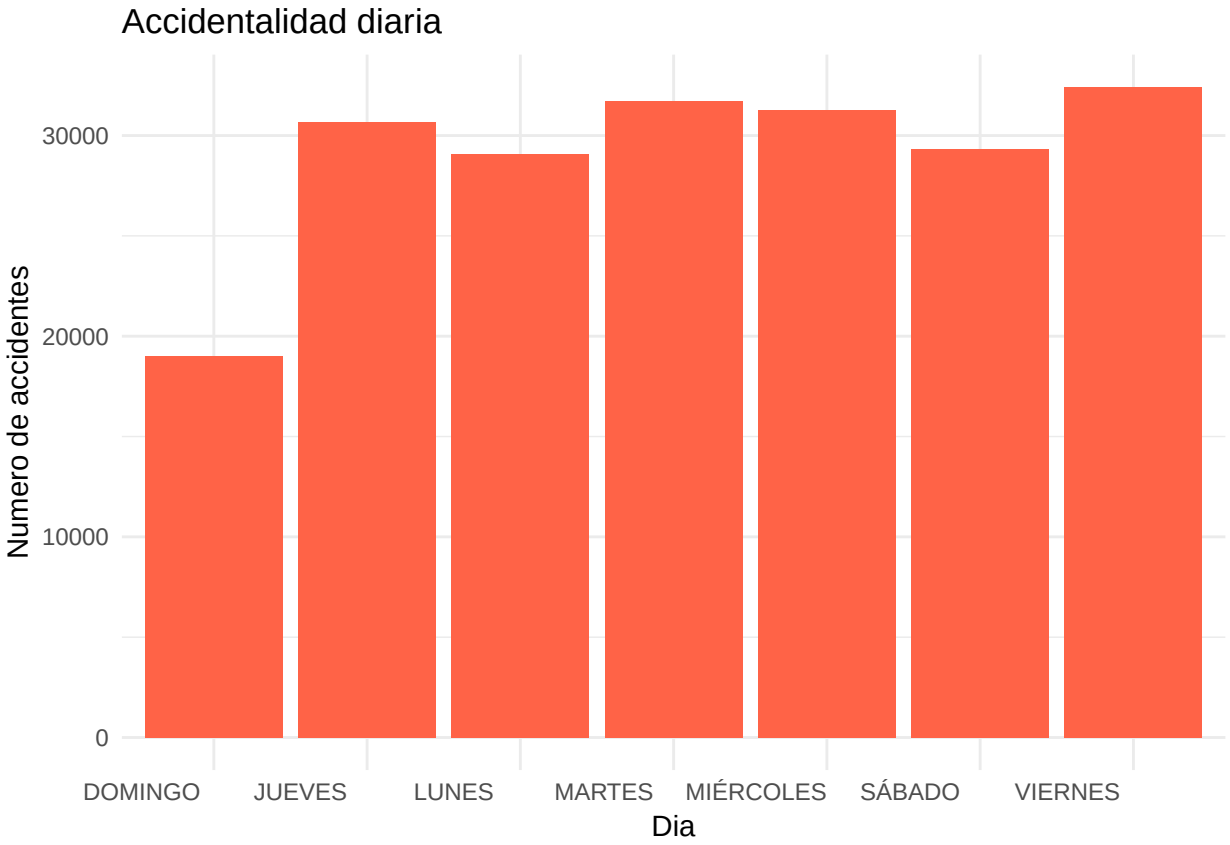


Este comportamiento se repite para todos los años (aunque se marca un poco más en algunos). Esto nos dice que el MES podría ser variable relevante para la construcción de los modelos, pero su consideración podría ser categórica.



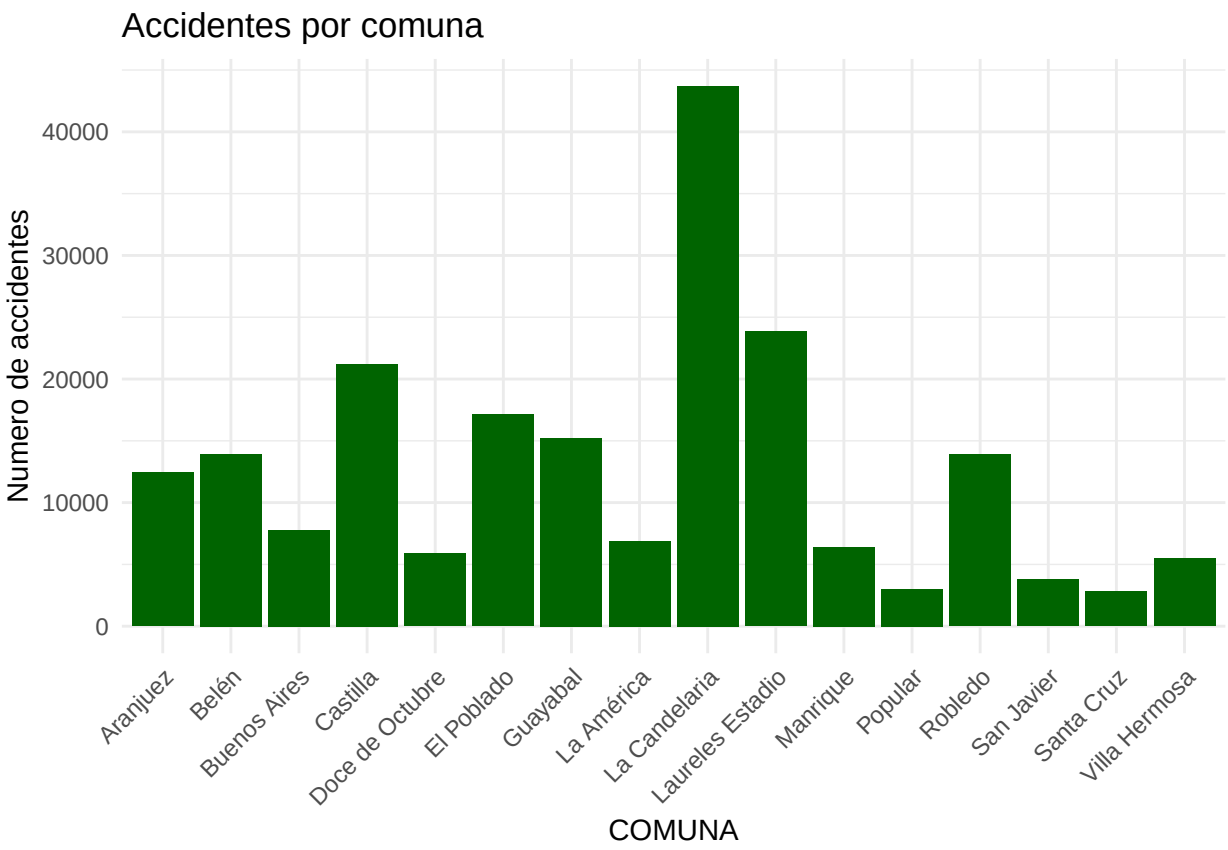
CLASE no parece tener una variación notable con los meses.

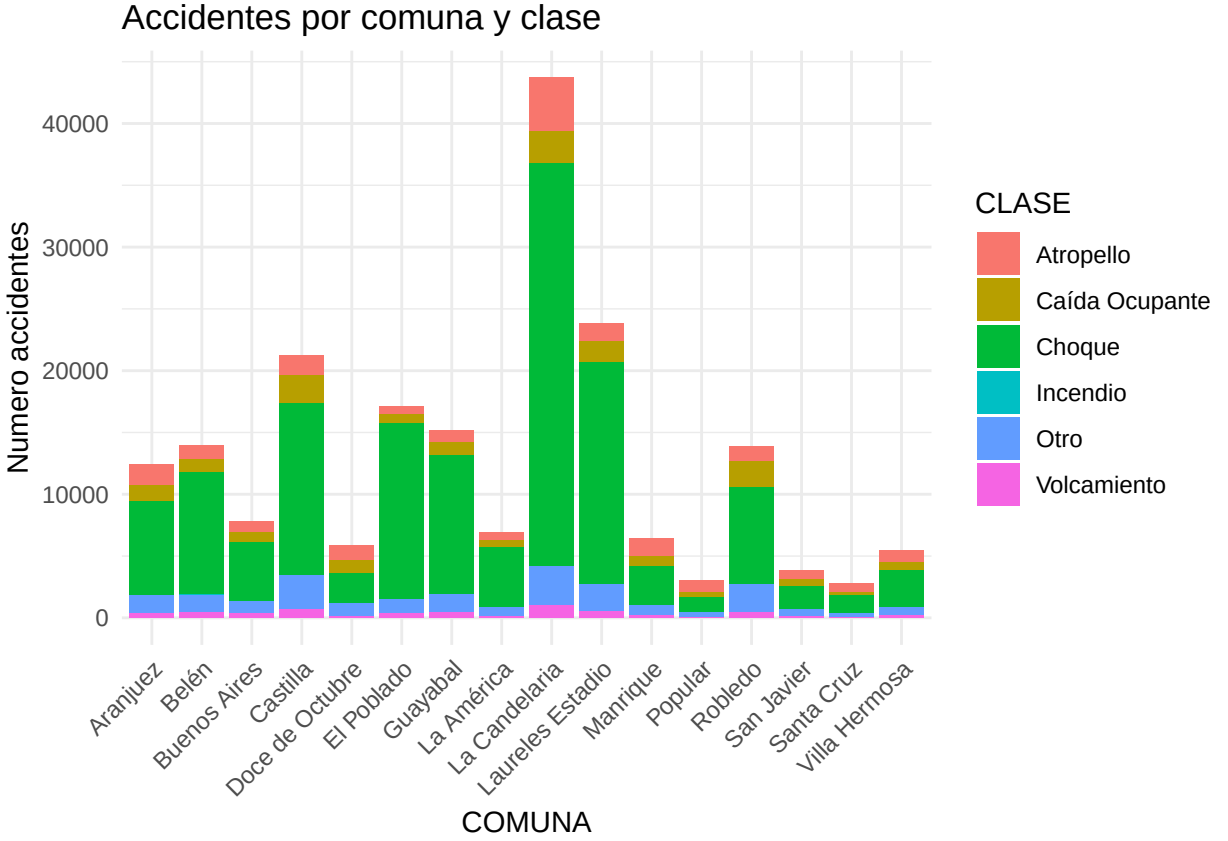
Análisis de accidentalidad por día de la semana



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





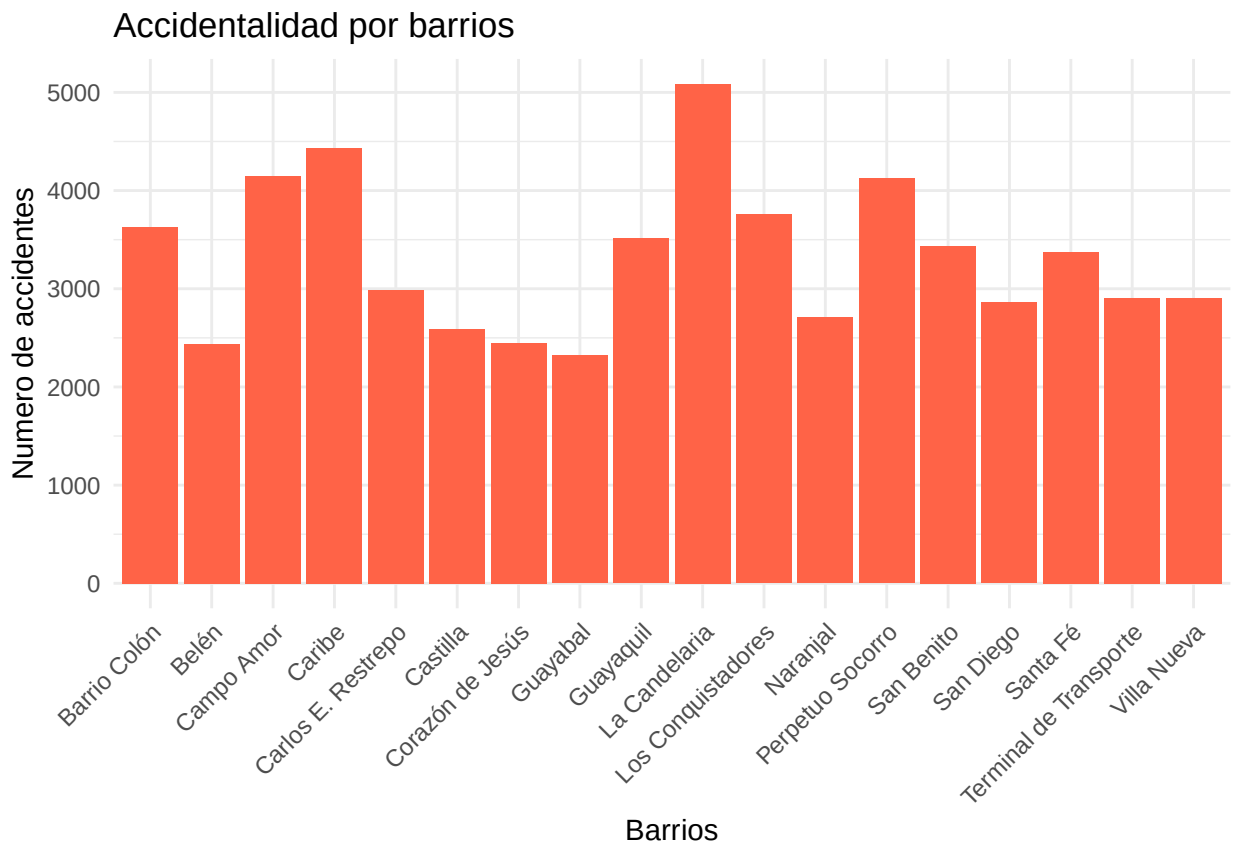
La Candelaria es donde ocurren la mayor cantidad de accidentes, con un número mayor a 40000. Las comunas en donde se presentan la menor cantidad de accidentes son las de Santa Cruz y Popular, seguidas por San Javier, con menos de 5000 accidentes en todos los casos. Otras comunas con valores de accidentalidad 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:

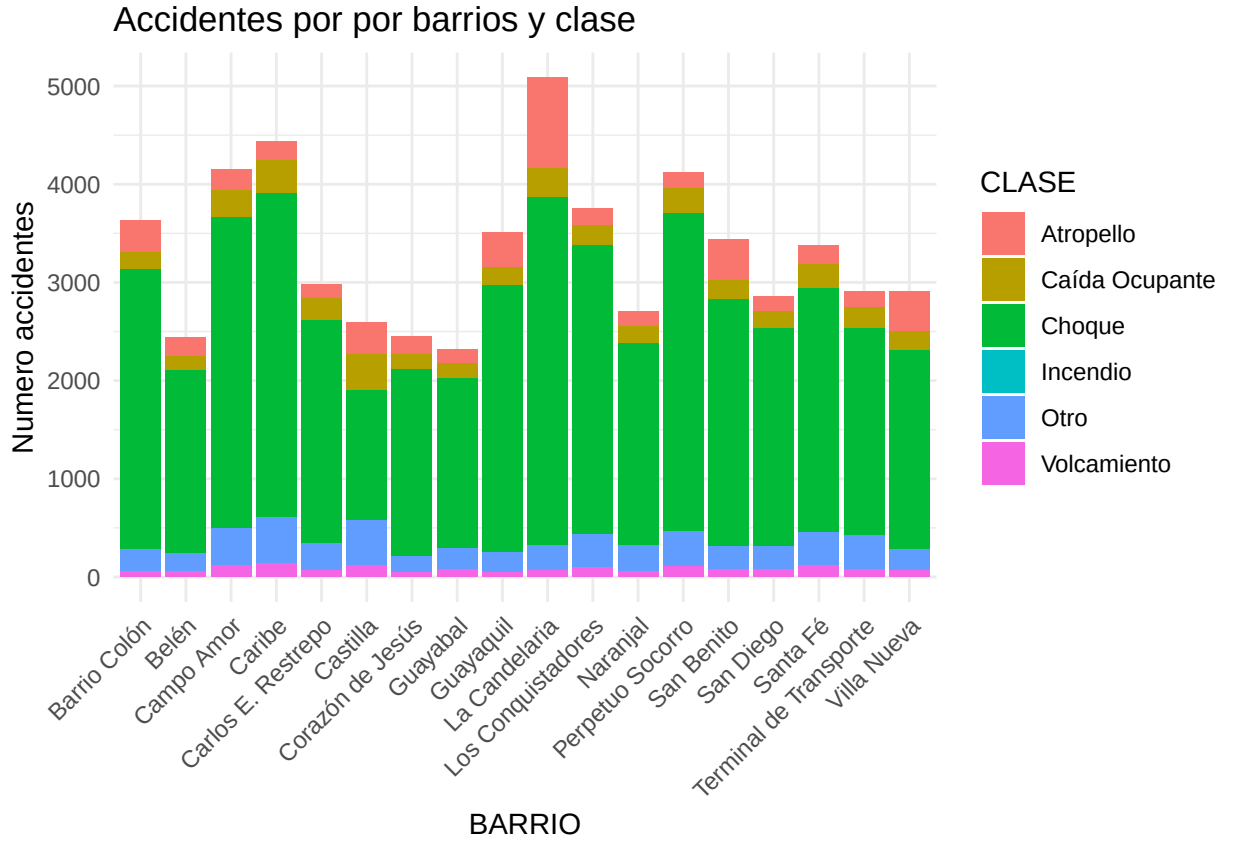
```
## Joining, by = "COMUNA"
```

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 variables que puedan estar relacionadas a los totales y se elegirá el que más disminuya el MSE para los datos de validación como el óptimo.

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.

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. 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.

Modelo 1:

Comuna	Family	Formula
Aranjuez	PO	total ~ DIA_NUM + especial + MES + PERIODO + I(PERIODOˆ2)
Belén	PO	total ~ DIA_NUM + especial + MES + DIA
Buenos Aires	PO	total ~ DIA_NUM + especial + DIA + MES + PERIODO + I(PERIODOˆ2)
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(PERIODOˆ2) + MES + PERIODO
Manrique	PO	total ~ DIA_NUM + especial + MES + PERIODO + I(PERIODOˆ2)
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(PERIODOˆ2)
Santa Cruz	NBI	total ~ DIA_NUM + MES + especial
Villa Hermosa	NBI	total ~ DIA_NUM + especial + DIA + MES + PERIODO + I(PERIODOˆ2)

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

Modelo 1:

Comuna	Family	Formula
Aranjuez	PO	total ~ SEMANA + I(PERIODO^2) + PERIODO
Belén	NBI	total ~ SEMANA + I(PERIODO^2) + PERIODO
Buenos Aires	PO	total ~ SEMANA + I(PERIODO^2) + PERIODO
Castilla	PO	total ~ SEMANA + MES
Doce de Octubre	NBI	total ~ SEMANA + I(PERIODO^2) + PERIODO
El Poblado	PO	total ~ SEMANA + MES
Guayabal	PO	total ~ SEMANA + MES
La América	PO	total ~ SEMANA + MES
La Candelaria	PO	total ~ SEMANA + MES + PERIODO
Laureles Estadio	NBI	total ~ SEMANA + I(PERIODO^2) + PERIODO
Manrique	PO	total ~ SEMANA + MES + 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	NBI	total ~ SEMANA + I(PERIODO^2) + PERIODO
Villa Hermosa	NBI	total ~ SEMANA + MES + PERIODO

Modelo 2:

Comuna	Family	Formula
Aranjuez	PO	total ~ SEMANA + I(PERIODO^2) + PERIODO
Belén	NBI	total ~ SEMANA + MES
Buenos Aires	PO	total ~ SEMANA + MES
Castilla	PO	total ~ SEMANA + MES + PERIODO
Doce de Octubre	PO	total ~ SEMANA + MES
El Poblado	PO	total ~ SEMANA + MES + 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	NBI	total ~ SEMANA + MES + PERIODO
Robledo	PO	total ~ SEMANA + I(PERIODO^2) + MES + PERIODO
San Javier	NBI	total ~ SEMANA + I(PERIODO^2) + PERIODO
Santa Cruz	NBI	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	142.92062	108.03291	139.299061	160.89786
Belén	170.96745	200.29695	178.568870	205.55046
Buenos Aires	62.69177	62.55174	60.861718	77.70311
Castilla	377.93167	419.45709	346.066245	486.89610
Doce de Octubre	33.91425	48.27624	31.696573	51.37677
El Poblado	311.30711	298.03937	263.991451	395.25863
Guayabal	222.21484	239.59201	231.017109	238.56952
La América	53.80159	59.02269	52.424195	62.86120
La Candelaria	1636.16195	1574.25469	1656.789516	1827.08571
Laureles Estadio	535.70764	441.97312	560.240573	569.94352
Manrique	41.85289	48.53983	41.551460	57.81366
Popular	14.52174	14.98573	14.138378	16.93478
Robledo	159.34183	134.13965	156.085539	172.57974
San Javier	20.02099	20.17392	19.714362	25.73240
Santa Cruz	10.50102	15.55964	9.851837	17.07452
Villa Hermosa	37.17984	40.36493	35.646330	51.69459

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 ~ MES + I(MESˆ2) + PERIODO
Belén	NBI	total ~ MES
Buenos Aires	NBI	total ~ I(PERIODOˆ2) + MES + PERIODO
Castilla	NBI	total ~ MES + I(MESˆ2)
Doce de Octubre	PO	total ~ MES + I(MESˆ2)
El Poblado	NBI	total ~ MES + I(MESˆ2)
Guayabal	NBI	total ~ MES + I(MESˆ2)
La América	NBI	total ~ MES
La Candelaria	NBI	total ~ MES + PERIODO
Laureles Estadio	PO	total ~ I(PERIODOˆ2) + MES + PERIODO
Manrique	NBI	total ~ MES + I(MESˆ2) + PERIODO
Popular	NBI	total ~ MES + I(MESˆ2)
Robledo	PO	total ~ MES + PERIODO
San Javier	PO	total ~ I(PERIODOˆ2) + MES + PERIODO
Santa Cruz	NBI	total ~ MES
Villa Hermosa	PO	total ~ MES + I(MESˆ2) + PERIODO

Modelo 2:

Comuna	Family	Formula
Aranjuez	PO	total ~ MES + I(MES^2)
Belén	PO	total ~ MES + I(MES^2)
Buenos Aires	NBI	total ~ MES + I(MES^2)
Castilla	PO	total ~ I(PERIODO^2) + MES + PERIODO
Doce de Octubre	NBI	total ~ MES + I(MES^2)
El Poblado	PO	total ~ MES + I(MES^2) + PERIODO
Guayabal	NBI	total ~ MES + I(MES^2)
La América	PO	total ~ MES + I(MES^2)
La Candelaria	PO	total ~ MES + I(MES^2)
Laureles Estadio	PO	total ~ MES + I(MES^2)
Manrique	NBI	total ~ I(PERIODO^2) + MES + PERIODO
Popular	NBI	total ~ MES + PERIODO
Robledo	PO	total ~ MES + I(MES^2) + PERIODO
San Javier	PO	total ~ MES + I(MES^2)
Santa Cruz	NBI	total ~ MES
Villa Hermosa	PO	total ~ MES + I(MES^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	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. Las tablas resumen se presentan solo para los 18 barrios con mayor número de accidentes.

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
20	Barrio Colón	total ~ DIA_NUM + MES + especial
25	Belén	total ~ DIA_NUM + MES + especial
44	Campo Amor	total ~ DIA_NUM + MES + especial
47	Caribe	total ~ DIA_NUM + MES + especial
48	Carlos E. Restrepo	total ~ DIA_NUM + MES + especial
50	Castilla	total ~ DIA_NUM + DIA + MES + especial
56	Corazón de Jesús	total ~ DIA_NUM + DIA + MES + especial
103	Guayabal	total ~ DIA_NUM + MES + especial
104	Guayaquil	total ~ DIA_NUM + MES + especial
117	La Candelaria	total ~ DIA_NUM + MES + especial
169	Los Conquistadores	total ~ DIA_NUM + MES + especial
187	Naranjal	total ~ DIA_NUM + MES + especial
203	Perpetuo Socorro	total ~ DIA_NUM + DIA + MES + especial
213	San Benito	total ~ DIA_NUM + MES + especial
215	San Diego	total ~ DIA_NUM + MES + especial
229	Santa Fé	total ~ DIA_NUM + MES + especial
246	Terminal de Transporte	total ~ DIA_NUM + MES + especial
265	Villa Nueva	total ~ DIA_NUM + MES + especial

Modelo 2:

	Barrio	Formula
20	Barrio Colón	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
25	Belén	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
44	Campo Amor	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
47	Caribe	total ~ DIA_NUM + DIA + MES + especial
48	Carlos E. Restrepo	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
50	Castilla	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
56	Corazón de Jesús	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
103	Guayabal	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
104	Guayaquil	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
116	La Candelaria	total ~ DIA_NUM + DIA + MES + especial
168	Los Conquistadores	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
186	Naranjal	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
202	Perpetuo Socorro	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
212	San Benito	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
214	San Diego	total ~ DIA_NUM + DIA + I(PERIODO^2) + MES + PERIODO + especial
228	Santa Fé	total ~ DIA_NUM + MES + especial
245	Terminal de Transporte	total ~ DIA_NUM + DIA + MES + especial
264	Villa Nueva	total ~ DIA_NUM + MES + especial

Tabla comparativa de los MSE

	BARRIO	MSE.tr Modelo 1	MSE.va Modelo 1	MSE.tr Modelo 2	MSE.va Modelo 2
20	Barrio Colón	2.047703	2.732915	1.920300	2.402876
25	Belén	1.443062	1.191471	1.415435	1.447549
44	Campo Amor	2.171203	4.493221	2.279527	3.239849
47	Caribe	2.783056	3.024432	2.891871	3.332214
48	Carlos E. Restrepo	1.936766	1.562133	1.952837	2.144549
50	Castilla	1.530420	1.332672	1.473985	1.465265
56	Corazón de Jesús	1.502920	1.382666	1.546759	1.733098
103	Guayabal	1.523742	1.108161	1.450713	1.562287
104	Guayaquil	2.180048	1.721141	2.308487	2.567718
117	La Candelaria	2.965210	3.767631	3.143826	3.901973
169	Los Conquistadores	2.132929	2.297820	2.200779	2.458142
187	Naranjal	1.652811	1.177441	1.688694	1.780053
203	Perpetuo Socorro	2.506419	3.002656	2.396855	2.906635
213	San Benito	2.149751	1.643656	2.238767	2.487640
215	San Diego	1.664843	1.739516	1.652691	1.814914
229	Santa Fé	1.969714	2.357621	1.961192	2.389669
246	Terminal de Transporte	1.830518	1.128955	1.770036	1.722406
265	Villa Nueva	1.695237	1.617376	1.793082	1.965342

Para los 18 barrios con más datos la suma de los MSE fueron: - Modelo 1 fue de 35.68 para entrenamiento y 37.28 para validación, - Modelo 2 fueron 36.08 y 41.32. Sin embargo en general la suma de los MSE de entrenamiento son para entrenamiento y validación respectivamente: - Modelo 1: 124.2207 y 116.711. - Modelo 2: 261.9699 y 133.3192. 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	MSE.va
20	Barrio Colón	20.00913
25	Belén	12.25126
44	Campo Amor	42.26181
47	Caribe	28.98595
48	Carlos E. Restrepo	13.77403
50	Castilla	10.05497
56	Corazón de Jesús	14.30758
103	Guayabal	12.32849
104	Guayaquil	15.62545
117	La Candelaria	23.95969
169	Los Conquistadores	15.86485
187	Naranjal	14.04216
203	Perpetuo Socorro	27.05927
213	San Benito	17.58839
215	San Diego	14.57450
229	Santa Fé	26.28801
246	Terminal de Transporte	12.87974
265	Villa Nueva	12.87068

Modelo 2:

	Barrio	MSE.va
20	Barrio Colón	19.30665
25	Belén	12.46124
44	Campo Amor	27.41907
47	Caribe	30.57918
48	Carlos E. Restrepo	17.23969
50	Castilla	11.70528
56	Corazón de Jesús	12.15262
103	Guayabal	11.58832
104	Guayaquil	27.26174
116	La Candelaria	33.10529
168	Los Conquistadores	24.51609
186	Naranjal	14.95982
202	Perpetuo Socorro	24.83009
212	San Benito	23.75135
214	San Diego	14.18868
228	Santa Fé	19.28488
245	Terminal de Transporte	13.97170
264	Villa Nueva	16.54225

Tabla comparativa de los MSE

	BARRIO	MSE.tr Modelo 1	MSE.va Modelo 1	MSE.tr Modelo 2	MSE.va Modelo 2
20	Barrio Colón	19.78984	20.00913	18.56732	19.30665
25	Belén	10.94420	12.25126	12.26499	12.46124
44	Campo Amor	22.52052	42.26181	24.21322	27.41907
47	Caribe	30.03329	28.98595	30.56780	30.57918
48	Carlos E. Restrepo	17.18976	13.77403	17.63256	17.23969
50	Castilla	12.24970	10.05497	12.19272	11.70528
56	Corazón de Jesús	17.87865	14.30758	10.09313	12.15262
103	Guayabal	13.16793	12.32849	11.48595	11.58832
104	Guayaquil	25.54649	15.62545	27.29958	27.26174
117	La Candelaria	34.43271	23.95969	35.77845	33.10529
169	Los Conquistadores	24.81185	15.86485	26.91525	24.51609
187	Naranjal	16.28088	14.04216	15.07944	14.95982
203	Perpetuo Socorro	23.95649	27.05927	23.66267	24.83009
213	San Benito	21.57873	17.58839	22.70300	23.75135
215	San Diego	13.55031	14.57450	14.12526	14.18868
229	Santa Fé	16.92266	26.28801	16.83540	19.28488
246	Terminal de Transporte	16.05378	12.87974	14.03888	13.97170
265	Villa Nueva	16.52475	12.87068	17.15056	16.54225

En este caso se encuentra que la suma de los MSE de entrenamiento y validación son respectivamente para los 18 barrios con mayor accidentalidad son: - Modelo 1: 353.4326,334.726. - Modelo 2: 350.6062, 354.8639. Usando todos los barrios son mayor accidentalidad son: - Modelo 1: 939.4501, 910.3487. - Modelo 2: 793.9533, 856.0634.

Nuevamente esto muestra que el modelo 1 se ajusta mejor siempre que haya un número considerable de accidentes. Mientras que el modelo 2 tiene mejores predicciones en general.

5.2.3. Modelo por meses

formula("total$_{\text{I(PERIODOˆ2)+MES+PERIODO"), formula("total}}$MES"), formula("total~MES+PERIODO") Se utilizan:

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

Modelo 1:

	Barrio	MSE.tr	MSE.va	Formula
20	Barrio Colón	75.84976	66.05773	total ~ MES + PERIODO
25	Belén	56.26410	35.51127	total ~ I(PERIODO^2) + MES + PERIODO
44	Campo Amor	109.63360	396.12140	total ~ MES + PERIODO
47	Caribe	108.33147	85.64718	total ~ MES + PERIODO
48	Carlos E. Restrepo	88.62193	73.25802	total ~ MES
50	Castilla	74.24531	27.59511	total ~ MES
56	Corazón de Jesús	80.45399	79.12091	total ~ MES + PERIODO
103	Guayabal	84.01376	171.36653	total ~ MES
104	Guayaquil	124.99830	105.31692	total ~ MES + PERIODO
119	La Candelaria	182.58056	63.49340	total ~ MES
171	Los Conquistadores	98.96794	124.98901	total ~ MES + PERIODO
189	Naranjal	78.33684	58.69759	total ~ MES + PERIODO
205	Perpetuo Socorro	68.87113	77.98177	total ~ MES
215	San Benito	148.55556	73.20893	total ~ MES + PERIODO
217	San Diego	53.35934	50.28521	total ~ I(PERIODO^2) + MES + PERIODO
231	Santa Fé	79.90687	84.89121	total ~ MES + PERIODO
248	Terminal de Transporte	70.09154	88.32271	total ~ MES + PERIODO
267	Villa Nueva	89.28820	51.47742	total ~ MES + PERIODO

Modelo 2:

	Barrio	MSE.tr	MSE.va	Formula
20	Barrio Colón	64.11223	91.36388	total ~ MES + PERIODO
25	Belén	72.52841	41.65055	total ~ MES
44	Campo Amor	130.09194	156.50047	total ~ MES + PERIODO
47	Caribe	111.01616	146.22889	total ~ MES
48	Carlos E. Restrepo	92.65186	73.16992	total ~ MES
50	Castilla	71.59201	62.05562	total ~ MES
56	Corazón de Jesús	73.73730	136.21534	total ~ MES
103	Guayabal	76.85233	123.91655	total ~ MES
104	Guayaquil	111.66815	435.65910	total ~ MES
116	La Candelaria	186.72203	144.68007	total ~ MES
168	Los Conquistadores	90.59925	146.20488	total ~ MES
186	Naranjal	88.00128	54.04089	total ~ MES + PERIODO
202	Perpetuo Socorro	73.36221	66.38507	total ~ MES
212	San Benito	129.09605	347.49749	total ~ MES
214	San Diego	59.24416	51.21527	total ~ MES
227	Santa Fé	75.96621	95.21253	total ~ MES + PERIODO
244	Terminal de Transporte	64.95463	99.53177	total ~ MES
263	Villa Nueva	97.15532	62.89775	total ~ MES + PERIODO

Tabla comparativa de los MSE

	BARRIO	MSE.tr Modelo 1	MSE.va Modelo 1	MSE.tr Modelo 2	MSE.va Modelo 2
20	Barrio Colón	75.84976	66.05773	64.11223	91.36388
25	Belén	56.26410	35.51127	72.52841	41.65055
44	Campo Amor	109.63360	396.12140	130.09194	156.50047
47	Caribe	108.33147	85.64718	111.01616	146.22889
48	Carlos E. Restrepo	88.62193	73.25802	92.65186	73.16992
50	Castilla	74.24531	27.59511	71.59201	62.05562
56	Corazón de Jesús	80.45399	79.12091	73.73730	136.21534
103	Guayabal	84.01376	171.36653	76.85233	123.91655
104	Guayaquil	124.99830	105.31692	111.66815	435.65910
119	La Candelaria	182.58056	63.49340	186.72203	144.68007
171	Los Conquistadores	98.96794	124.98901	90.59925	146.20488
189	Naranjal	78.33684	58.69759	88.00128	54.04089
205	Perpetuo Socorro	68.87113	77.98177	73.36221	66.38507
215	San Benito	148.55556	73.20893	129.09605	347.49749
217	San Diego	53.35934	50.28521	59.24416	51.21527
231	Santa Fé	79.90687	84.89121	75.96621	95.21253
248	Terminal de Transporte	70.09154	88.32271	64.95463	99.53177
267	Villa Nueva	89.28820	51.47742	97.15532	62.89775

En general el modelo 1 presenta mejores estimaciones para los datos de validación, para los 18 barrios con mayor accidentalidad como para el conjuntos global.

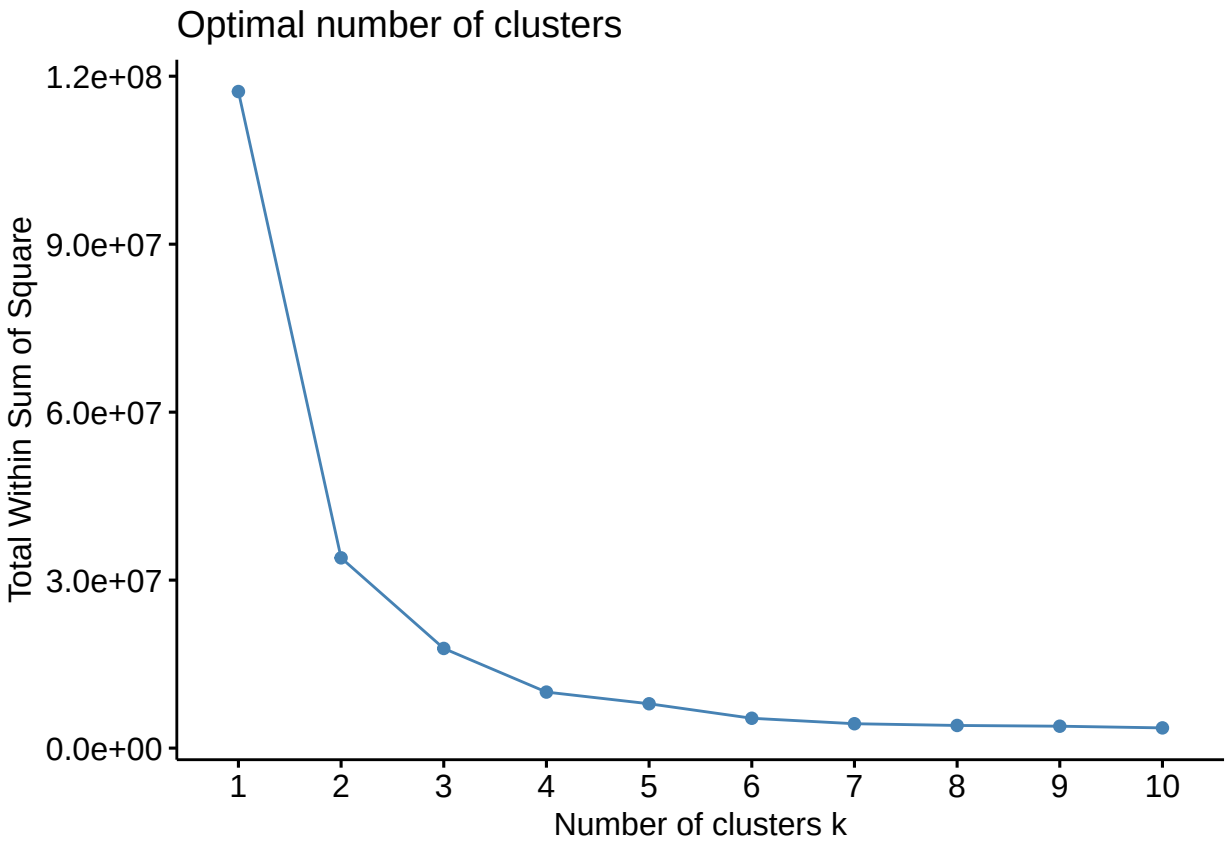
- Modelo 1: suma MSE.tr = 4600.85, suma MSE.va = 5915.225. Para los 18 barrios con mayor accidentalidad: MSE.tr = 1672.37, suma MSE.va = 1713.34.
- Modelo 2: suma MSE.tr = 4255.079, suma MSE.va = 6979.177. Para los 18 barrios con mayor accidentalidad: MSE.tr = 1669.352, suma MSE.va = 2334.426.

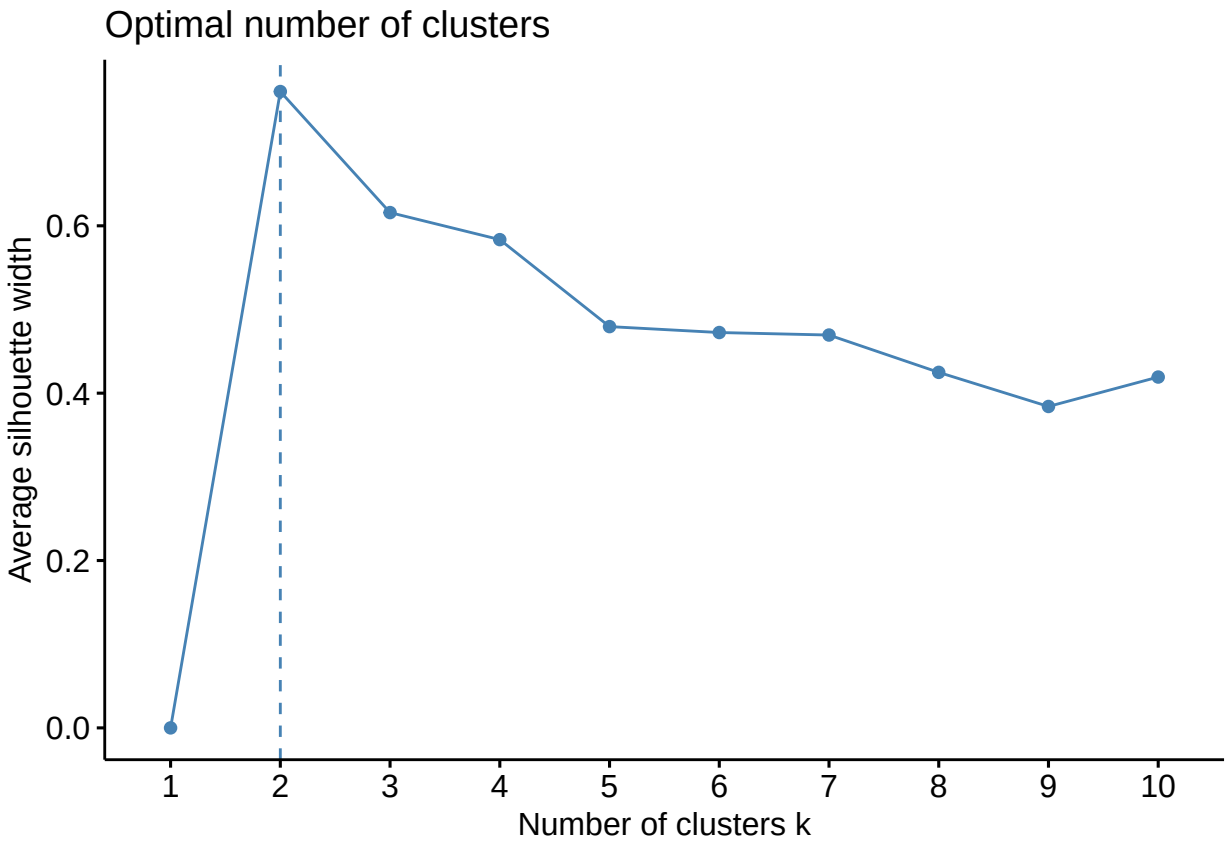
Se considera que tanto para barrios como para columnas en el modelo anual, la superioridad muy marcada del modelo 1 para predecir los datos de validación frente al modelo 2, es debida a que posee datos del año previo a sus datos de validación, al solo tener MES y PERIODO involucradas, el peso de esta 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:

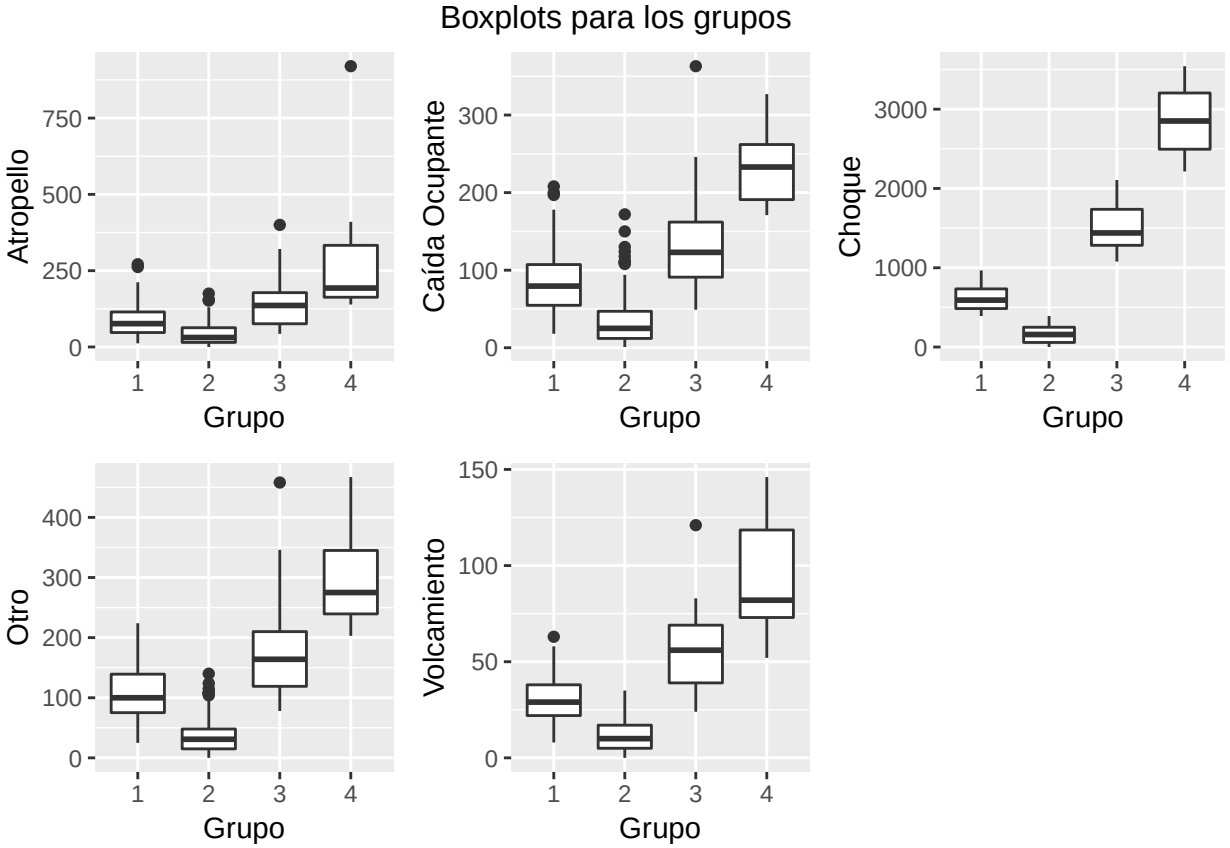
C4G	elementos	Atropellos	Caidas	Choques	Volcamientos	Otros
1	70	87.57143	86.97143	613.1714	30.28571	105.77143
2	161	43.69565	35.37267	163.7640	11.19876	37.39752
3	25	140.52000	138.40000	1506.6000	56.04000	179.80000
4	11	290.90909	231.81818	2840.7273	93.36364	301.72727

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.