Master in Management + Analytics

Un enfoque de aprendizaje automático para la predicción del delito en la Ciudad Autónoma de Buenos Aires

Autoría: *De Antonio, Julieta*
Año académico: *2023*

UNIVERSIDAD TORCUATO DI TELLA

Master in Management + Analytics

Un enfoque de aprendizaje automático para la predicción del delito en la Ciudad Autónoma de Buenos Aires

TESIS

Julieta De Antonio

Mayo 2023

Tutores: María de los Ángeles Scetta & Ramiro Galvez

Resumen

El delito es sin lugar a dudas un problema que persigue a todos las naciones y gobiernos del planeta. La prevención del mismo, por consiguiente, forma parte de la agenda para cada uno de ellos. El objetivo de esta tesis es, a partir de un enfoque de aprendizaje automático, demostrar qué es posible estimar el lugar y momento en el que se dará un crimen en el futuro. Particularmente, se buscará determinar si los delitos son en realidad hechos aleatorios o si los mismos se ven afectados de manera simultánea por un conjunto de variables espacio - temporales para la Ciudad Autónoma de Buenos Aires. Un modelo con estas características, si resulta exitoso, permitiría asignar de manera más precisa, patrulleros y policías de las fuerzas de seguridad de CABA. Los resultados obtenidos sugieren que frente a un modelo ingenuo, los algoritmos de aprendizaje automático son ampliamente superadores y que es posible determinar la cantidad de delitos que se tendrán en el mes siguiente. En este trabajo se detallan los distintos conjuntos de datos que fueron utilizados para enriquecer los registros de delitos, así como también, el trabajo generado para realizar la grilla, que servirá como punto de partida para estimar los modelos. Asimismo, se explica el tradeoff que se genera al elegir un tamaño de cuadrícula para la misma.

Abstract

Crime is undoubtedly a problem that affects all nations and governments worldwide. Therefore, its prevention is part of the agenda for each of them. The objective of this thesis is to demonstrate, through a machine learning approach, that it is possible to estimate the place and time where a crime will occur in the future. Particularly, it aims to determine whether crimes are truly random or if they are simultaneously affected by a set of spatial-temporal variables in the Autonomous City of Buenos Aires (CABA). A model with these characteristics, if successful, would allow for a more precise allocation of patrol officers and police from CABA's security forces. The obtained results suggest that, compared to a naive model, machine learning algorithms are vastly superior, and it is possible to determine the number of crimes expected in the following month. This work details the different datasets used to enrich crime records, as well as the efforts made to create a grid that will serve as a starting point for estimating the models. Additionally, it explains the tradeoff generated when choosing a grid size for the analysis.

Índice

Índice de tablas

Índice de figuras

1 Introducción

1.1 Contexto

El delito, como un problema a largo plazo, es una grave amenaza para todas las comunidades y naciones de la Tierra. El impacto que genera en la sociedad es tanto significativo como negativo. 'La prevención del delito es "la totalidad del conjunto de políticas, medidas y técnicas, fuera de los límites del sistema de justicia penal, que tienen por objeto la reducción de los distintos tipos de daño causado por actos definidos como delitos por el Estado" [Van Dijk, 1990]. La prevención situacional se orienta a dificultar actividades que son consideradas delitos a través de la disuasión, como cámaras de seguridad, patrullas policiales, iluminación'[Appiolaza, 2010].

Los métodos tradicionales para la prevención de crímenes dependen en gran medida de la experiencia de las fuerzas policiales. Esto representa un gran desafío para generalizar procedimientos y compartir información al respecto, llevando a la pérdida de oportunidades para prevenir delitos.

El avance tecnológico y la creciente disponibilidad de datos en formato digital, se han vuelto una importante herramienta que permite investigar, prevenir e incluso frenar la delincuencia. Un gran desafío al que se enfrentan todas las organizaciones encargadas de hacer cumplir la ley es analizar de manera precisa y eficiente estos crecientes volúmenes de datos. El aprovechamiento de los mismos excede fronteras, por lo que su utilización se da tanto en organismos gubernamentales como en instituciones privadas en todo el mundo. Es por todo esto que, aprovechados correctamente, pueden ayudar a aumentar la eficacia y eficiencia en la toma de decisiones, permitiendo transformarlas de reactivas a proactivas e incluso predictivas. Permitiendo así, a partir de la anticipación de futuros eventos, un análisis prescriptivo que, basándose en datos, funcione de guía en la elaboración de políticas relacionadas con la seguridad ciudadana.

La minería de datos como herramienta para el análisis del crimen se ha reconocido como un área de investigación relativamente nueva pero muy utilizada. La inclusión de información espacio - temporal en las bases de datos utilizando sistemas de información geográfica ha revolucionado las predicciones de delitos, permitiendo a los investigadores obtener predicciones más precisas y confiables sobre los crímenes. Además, la combinación de estas con técnicas de pronóstico de series de tiempo o técnicas de aprendizaje profundo como las redes neuronales representan hoy el *estado del arte* en la previsión de delitos.

1.2 Problema

Actualmente, en la Ciudad Autónoma de Buenos Aires, las bases de datos del Ministerio de Justicia y Seguridad contienen una gran cantidad de datos sobre delitos que podrían usarse para entender las tendencias y patrones delictivos actuales y aquellos que se darán a futuro. Sin embargo, no es posible aprovecharlas con información y herramientas públicas en algunos aspectos. A pesar de poseer un Mapa del Delito, el mismo no tiene como objetivo la anticipación de futuros crímenes, sino el de ser un lugar de consulta para los

ciudadanos con información confiable y de calidad sobre los hechos delictivos históricos.

Si bien el Mapa del Delito ha sentado un precedente en materia de seguridad en el país, el mismo no conlleva un análisis predictivo ni prescriptivo que pueda ser consultado de manera abierta.

1.3 Objetivo

La pregunta que surge a partir de lo comentado previamente es si utilizando dichos datos es posible estimar dónde y cuándo ocurrirán los crímenes en el futuro. En realidad, lo que podríamos preguntarnos es si los delitos son en realidad hechos aleatorios o si los mismos se ven afectados por el entorno, el nivel socioeconómico y/o variables climáticas/temporales.

Teniendo todo esto en cuenta, el objetivo de esta tesis consiste, a partir de datos públicos sobre los registros de delitos en la Ciudad Autónoma de Buenos Aires (CABA), en relacionar un lugar y un momento en el tiempo con la ocurrencia de robos y hurtos. Particularmente, se implementarán modelos de *machine learning* para predecir la ocurrencia de estos crímenes en celdas de una grilla sobre el territorio de CABA con una resolución de 200 x 200 metros. Incorporando variables de entorno, temporales, climáticas y socioeconómicas en simultáneo. Los resultados obtenidos en la misma podrán ser utilizados para tomar decisiones en cuanto a cómo distribuir, tanto en dónde como cuándo, los recursos que poseen los servicios policiales de la Ciudad Autónoma de Buenos Aires. En otras palabras, a partir de conocer dónde y en qué momento es más probable que ocurra un delito, se podrán alocar de manera más precisa, patrulleros y policías. Logrando que los mismos puedan actuar más rápida y efectivamente a la hora de prevenir o proceder frente a un crimen.

Es importante destacar que al llevar a cabo el análisis utilizando datos completamente abiertos, se puede lograr replicarlo a un bajo costo para diferentes organismos gubernamentales. Por lo que su impacto, podría ser aún más significativo.

2 Datos

Los datos utilizados para el desarrollo de esta tesis provienen de distintas fuentes de información: el repositorio abierto del Gobierno de la Ciudad Autónoma de Buenos Aires [Gobierno de la Ciudad Autónoma de Buenos Aires, 2012] (principal fuente), OpenStreetMap [OpenStreetMap Contributors, 2017] y la API Dark Sky [DarkSky, 2023]. Algo a destacar es que todos los *datasets* son públicos y pueden ser descargados de forma gratuita.

Cabe mencionar que no solo se utilizaron los conjuntos de datos con los delitos, ya que se ha demostrado que la incorporación de variables tanto espaciales, como socioeconómicas y climáticas aumentan la precisión de los modelos de predicción de crímenes [Dash et al., 2018].

Por otra parte, todos los *datasets* corresponden al mismo territorio: la Ciudad Autónoma de Buenos Aires. La misma es la Capital Federal de Argentina desde 1880 y según el censo de 2022 tiene una superficie de 200 km^2 y una población de 3.120.612 habitantes, dando una densidad de 15.150,96 hab/km^2 [Argentina.gob.ar, 2020]. En el sitio oficial del Gobierno de la Ciudad de Buenos Aires se menciona: "El tejido urbano se asemeja a un abanico que limita al sur, oeste y norte con la provincia de Buenos Aires y al este con el río. Oficialmente, la ciudad se encuentra dividida en 48 barrios."[Buenos Aires Ciudad, 2023] Esta información será relevante a la hora de entender las condiciones de la región sobre la que se trabajará.

En las siguientes secciones se hablará más en detalle de los datos que fueron utilizados en este trabajo.

2.1 Datos de delitos

Este conjunto de datos contiene información correspondiente a los homicidios, hurtos, lesiones y robos que ocurrieron en la Ciudad de Buenos Aires desde 2016 a 2021, inclusive.

Algo importante a tener en cuenta es que se está trabajando con los **registros de los delitos** que no necesariamente representan al 100 % la cantidad de crímenes que realmente ocurrieron, dado que para que exista un registro, la persona o el grupo de personas que ha sufrido el hecho debe presentarse en una comisaria y realizar la denuncia. En consecuencia, se analizará la información disponible y no se hará suposiciones acerca de lo que no se registra.

En relación a las bases de datos, el Gobierno de la Ciudad, al comienzo de cada año, publica un archivo *.csv* con la información de los crímenes denunciados del año anterior. Cada fila en dichos documentos representa el registro de un delito con las siguientes variables: *ID, fecha, franja horaria, tipo y subtipo de delito, cantidad registrada, comuna, barrio, lat y long*. Sin embargo, dichas columnas varían para los años 2020 y 2021. Las variables de estos últimos dos años son: *id_mapa, anio, mes, dia, fecha, franja, tipo, subtipo, uso_armas, barrio, comuna, latitud, longitud y víctimas*. Por lo tanto, para poder hacer uso de los datos de una forma más eficiente, se unificó el criterio de las variables, dejándolas como las de los primeros años. Además, se decidió trabajar solo con los delitos

que corresponden a las categorías de hurtos y robos, dejando de lado los homicidios y lesiones. Esto se debe a que la naturaleza de estos primeros dos tipos es semejante y se suelen dar en contextos similares. A diferencia de los homicidios y lesiones que suelen estar acompañados por otras particularidades. De esta forma, se generó un único *dataset* con el registro de los delitos para todos los años utilizados en el análisis. El mismo será el conjunto de datos central en este trabajo.

2.2 Datos espaciales

2.2.1 Puntos de Interés

Estos conjuntos de datos contienen información de distintos puntos de interés. "Los puntos de interés (POI, por sus siglas en inglés) son ubicaciones que, históricamente, los cartógrafos han agregado a los mapas para comunicar un lugar interesante o relevante, utilizando símbolos y etiquetas cartográficas. Suelen incluir características visuales y culturalmente importantes. Los puntos de interés digitales contemporáneos son representaciones de ubicaciones del mundo real, generalmente representados como entidades de puntos geométricos."[Psyllidis et al., 2022] En otras palabras, los datos de puntos de interés se refieren a una amplia gama de ubicaciones con su latitud y longitud, a las cuales se las puede denominar como variables de entorno: comercios, escuelas, estaciones de policía, restaurantes, bares, entre muchas otras. Se ha demostrado que el incorporar puntos de interés (información geográfica) a modelos de aprendizaje automático y particularmente de predicción de delitos, tiene una alta utilidad predictiva [Cichosz, 2020]. Debido a que las condiciones del entorno de un lugar pueden influir a que sea más propicio para que se cometa un delito o al contrario, tener condiciones desfavorables. Por ejemplo, si se encuentra una comisaria cerca, es de esperarse que ocurran menos delitos que en una zona en donde no hay una. Puesto que los policías se encuentran en los alrededores y, por lo tanto, podrán actuar más rápido, habrá una mayor probabilidad de que los ladrones sean atrapados y, como consecuencia, decidirán cometer delitos lejos de las mismas. Caso contrario, en zonas donde hay una escapatoria rápida, ya sea un acceso a una autopista o muchos medios de transportes públicos, es de esperarse que los ladrones se sientan atraídos por la posibilidad de una huida más efectiva y delincan más en estas zonas.

Estos *datasets* se obtuvieron a partir de dos fuentes de información: en primer lugar, del repositorio de datos públicos del Gobierno de la Ciudad Autónoma de Buenos Aires (GCABA) [Gobierno de la Ciudad Autónoma de Buenos Aires, 2012] y para complementar, de OpenStreetMap [OpenStreetMap Contributors, 2017]. A continuación, se hablará más en detalle de estos conjuntos de datos, diferenciando por la fuente.

BA Data

Como se mencionó anteriormente, estos conjuntos de datos corresponden a múltiples archivos obtenidos a partir de los datos públicos del GCABA. En los mismos se encuentran distintos puntos de interés; siendo cada fila, en cada archivo, un POI con su ubicación.

La obtención de estos *datasets* fue bastante sencilla. Dado que Buenos Aires Data posee una API, es posible importar los archivos directamente en `RStudio` sin necesidad de descargarlos desde la web, facilitando la tarea de ingesta, ya que son una cantidad considerable. Particularmente, son 35 archivos *.csv* en donde cada uno de ellos posee información de los distintos puntos de interés dentro de cada tipo. En función de esos archivos, se generaron 39 variables de entorno. Lo que no fue tan sencillo, fue encontrar los distintos *feautures* y por consiguiente, documentos, que se iban a incorporar como variables de entorno. Esto se debe a que en la página de BA Data hay más de 400 *datasets* y por más que es posible realizar algunos filtros para encontrar los puntos de interés, fue necesario revisar una gran proporción de esos archivos para determinar cuáles podían ser relevantes para el análisis.

Algo a tener en cuenta es que la mayoría de estos archivos poseen distintos formatos y en la mayoría de los casos distintas variables. Es por esto que, para simplificar, se omitieron todas aquellas variables menos las que indicaban la latitud y longitud y se generó una nueva variable con la clase de cada POI para conformar un único archivo.

Algunos ejemplos de estos puntos de interés son:

- Bar
- Boca de subte
- Comisaria
- Hospital
- Iglesia
- Restaurante
- Universidad

En el Apéndice A se puede encontrar el listado completo.

OpenStreetMap

Por otro lado, para incorporar más puntos de interés que no se encuentran en los datos públicos y disponibles del GCABA, se utilizó OpenStreetMap.

Es importante mencionar que se decidió priorizar la información obtenida a través de BA Data, ya que los datos de OSM, al ser un proyecto colaborativo, requieren de un mayor preprocesamiento para poder ser utilizados. Por ejemplo, dentro de los datos se puede encontrar el mismo punto de interés duplicado con una diferencia de 20 metros. El cual, para el análisis, se debería contabilizar como un único POI. Esto implica que se deben quitar los puntos de interés repetidos que poseen el mismo nombre. Sin embargo, muchas veces éste está escrito de manera distinta, dificultando aún más la tarea de limpieza y requiriendo analizar manualmente los datos para evitar tener información repetida.

Para obtener estos puntos de interés se debe ingresar a un servidor de OpenStreetMap (en esta instancia se utilizó Geofabrik [Geofabrik GmbH & OSM Contributors, 2018]), seleccionar el continente y país y descargar el archivo *.shp* correspondiente, en este caso a Argentina. Con ese archivo se debe generar la intersección entre los datos de puntos de interés descargados y el límite de la Ciudad Autónoma de Buenos Aires para obtener los POI dentro de dicho territorio. A diferencia de los *datasets* obtenidos a partir de BA Data, los puntos de interés en este caso se encuentran todos en un mismo archivo. Cada fila representa un POI con las siguientes variables: *osm_id, code, fclass, name, geometry*. La primera variable representa el número único identificatorio de OSM, tanto la segunda como la tercera representan la categoría a la que pertenece cada POI, la cuarta es el nombre comercial y por último, tenemos las coordenadas de los puntos en formato espacial. Se hablará más en detalle en la sección de Archivos con datos espaciales al respecto del tipo de archivo *.shp* y como se trabajó con los mismos.

Algunos ejemplos de estos puntos de interés son:

- Atracciones
- Centro de deportes
- Local de indumentaria
- Museo
- Shopping
- Super e Hipermercado

En el Apéndice A se puede encontrar el listado completo de los puntos de interés pertenecientes a OSM que se utilizaron en este trabajo.

2.2.2 Regiones de Interés

Estos conjuntos de datos contienen información de distintas regiones de interés. A diferencia de los POI, de los que se habló en la sección anterior, en este caso los ROI, por sus siglas en inglés, se tratan de polígonos que representan áreas o regiones de interés. Al igual que con los POI, las fuentes de datos son tanto el repositorio de datos públicos del GCABA como OpenStreetMap. Nuevamente, se hablará más en detalle de estos conjuntos de datos diferenciando por la fuente.

BA Data

La principal diferencia con los puntos de interés obtenidos a partir de esta misma fuente de datos, más allá de tener áreas en lugar de puntos, es que los archivos que se utilizaron tienen un formato *.shp*. En este caso, se trata de 2 archivos con dos áreas de interés. Al igual que con los POI, se utilizó la API para importar los archivos sin necesidad de descargarlos desde la web. De la misma manera, al tener distintos formatos se decidió

simplificar y omitir todas las variables menos aquellas que indicaban el área (*geometry*) y se generó una variable con la clase de cada ROI para conformar un único archivo.

Estas regiones de interés son:

- Barrios populares [1]
- Espacios verdes

OpenStreetMap

Al igual que con los POI, para incorporar algunas áreas más de interés, se recurrió a OpenStreetMap. La forma de obtener los datos es igual a la de los puntos de interés y el formato del archivo es el mismo también. Cada fila representa un área de interés y la diferencia con el archivo de los POI, es que la ubicación en este caso no representa un punto, sino, como su nombre lo indica, un área en el mapa. Por lo tanto, no se tiene solo la latitud y longitud, sino que se tienen múltiples puntos que al unirlos conforman un polígono como variable *geometry*.

Algunas de estas regiones de interes son:

- Campo de golf
- Cementerio
- Cine
- Zoológico

En el Apéndice A se puede encontrar el listado completo de las áreas de interés.

2.3 Datos socioeconómicos

Estos conjunto de datos contienen información que puedan dar una noción acerca del nivel socioeconómico de los distintos sectores de la Ciudad Autónoma de Buenos Aires. Todos estos *datasets* se obtuvieron a partir de BA Data, y al igual que el resto de los archivos de esta fuente, fueron obtenidos utilizando la API. El formato de estos archivos es *.shp*, por lo que se tienen las coordenadas en formato espacial.

En primer lugar, se utilizó el precio de oferta en dólares de venta de terrenos, tomándose como referencia el precio del metro cuadrado. En el archivo, cada fila representa un terreno, con las siguientes variables: *direccion, propiedads, preciousd, preciopeso, dolarm2, pesosm2, cotizacion, trimestre, barrio, comuna, geometry*. Para el análisis se utilizó la variable *dolarm2*, ya que es la que permite una mejor comparación entre terrenos y en una moneda más estable a lo largo de los años de análisis, que la moneda de curso legal del país.

[1] "(...) la sociogénesis de la violencia en los asentamientos no se vincula a una forma de vida particular de los sectores populares, sino a un modo de relación del Estado con estos grupos (y la co-construcción de la vida sociopolítica)." [Cravino, 2016]

En línea con el archivo anterior, se hizo uso del precio de venta de departamentos. Las variables de este documento son similares a las del *dataset* de terrenos, con el agregado de la información de la cantidad de ambientes. Utilizando el mismo criterio que antes, la variable que se tomó para el análisis es la que indica el precio de venta en dólares por metro cuadrado.

En última instancia, se recurrió a información del Censo de 2010. Particularmente, a la información censal por radio. Cada fila en este archivo representa una unidad espacial de Radios con las siguientes variables: *ID, co_frac_ra, comuna, fraccion, radio, total_plob, t_varon, t_mujer, t_vivienda, v_particul, v_colectiv, t_hogar, h_con_nbi, h_sin_nbi, geometry.* En el análisis, se decidió utilizar el cociente entre las variables *h_con_nbi* y *t_hogar* para tener el porcentaje de hogares con Necesidades Básicas Insatisfechas sobre el total de hogares de dicha unidad espacial.

2.4 Datos climáticos

Por último, se utilizaron datos provenientes de la API Dark Sky para obtener información meteorológica. Para recabar estos datos, a partir de la ubicación de un punto en CABA, en este caso en particular de Parque Centenario, se utilizó la API para descargar variables relacionadas con cómo se encontraba el clima para cada día desde 2016 a 2021. Cada fila, por lo tanto, representa un día y se tiene información acerca de la temperatura, viento y precipitaciones de dicho día en CABA. [2]

2.5 Archivos con datos espaciales

Como se mencionó anteriormente, una proporción importante de los datos que se han utilizado en este trabajo tienen un formato *.shp* o *shapefile*. Es decir, se trata de datos georreferenciados o espaciales. Lo particular de estos *datasets* es que "poseen información sobre su ubicación en la Tierra"[Montane, 2020] y esto se debe tener en cuenta para poder trabajar con ellos. Es por esto que, en `R`, se deben manipular con la librería `sf` [Pebesma, 2018]. La misma hace referencia a *simple features* en su nombre y nos permite de manera estandarizada poder trabajar con objetos que tienen tanto información espacial como no espacial. Particularmente, en este tipo de datos, se cuenta con una columna llamada *geometry*; la misma contiene información sobre las coordenadas, el Sistema de Coordenadas Referenciado o CRS, por sus siglas en inglés, y el tipo de objeto geométrico, es decir, si se trata de un punto, una línea o un polígono, entre otras.

Esta librería, también permite cambiar las coordenadas de los datos a distintos tipos. En este trabajo se utilizó el sistema de coordenadas geográficas, basado en un modelo elipsoidal, en donde las coordenadas están expresadas en latitud y longitud, y el sistema de coordenadas cartográficas proyectadas UTM, en donde la proyección de la tierra es redonda sobre una superficie plana y las coordenadas están expresadas en metros. Algo

[2] Al momento de entrenar el modelo no se utilizaron las variables meteorológicas del mes actual. Esto se debe a que cuando se implemente en nuevos datos (futuros), las mismas no serán conocidas.

a tener en cuenta es que este último, solo es preciso dentro del área de la zona, es decir, dentro de un rango de longitudes que depende del número de la zona. Para la Ciudad Autónoma de Buenos Aires, el EPSG asociado a las coordenadas UTM es el 32720.

2.6 Valores faltantes (NAs)

Los valores faltantes son aquellas observaciones para las cuales no se posee información, ya sea para una variable en particular o para el total de las variables de dicha fila. Se las suele denominar como NA, por las siglas en inglés de *Not Available*, es decir, *No Disponible*. En la mayoría de los conjuntos de datos utilizados en el análisis no se poseen NAs. Además, en muchos casos en donde sí los hay, los mismos no generan un problema. Sin embargo, el tener valores faltantes en las variables *latitud* y *longitud* en el dataset de **delitos**, sí nos genera un inconveniente, ya que no es posible ubicar en el mapa dicho registro. Por lo tanto, se optó por eliminar las observaciones que no poseían información de la ubicación para los delitos. Para el total de años se tiene un total de 5834 observaciones con valores faltantes, representando un 0.11 % del total. Si lo analizamos por año:

NAs	2016	2017	2018	2019	2020	2021
FALSE	99.69	98.30	97.27	99.84	99.82	99.15
TRUE	0.31	1.70	2.73	0.16	0.18	0.85

En 2017 corresponde al 1.7 % de los datos y en 2018 al 2.7 %. El resto de los años es menor al 1 %. A partir de estos resultados, se puede concluir que son valores faltantes aleatorios y que no conciernen a un año en particular. Además, estos valores se encuentran muy por debajo del 15 % propuesto por Ratcliffe [Ratcliffe, 2004] como tasa mínima de geocodificación confiable. Geocodificación entendido como el proceso de convertir ubicaciones, como las direcciones de las víctimas de robos o hurtos, en coordenadas de la grilla.

3 Metodología

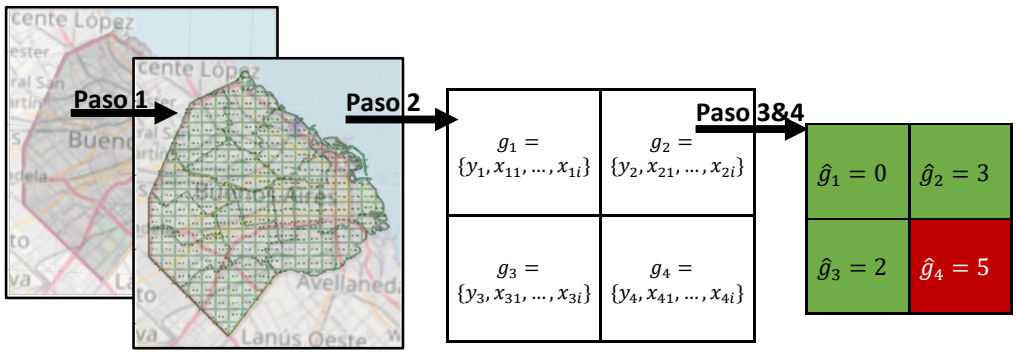
3.1 Esquema general

En esta primera sección metodológica se presentará de manera general el flujo de trabajo que se le aplicó a los datos hasta finalmente obtener la predicción de delitos.

Partiendo del trabajo presentado en [Lin et al., 2017] y [Lin et al., 2018] se generó la Figura 1, en donde se muestra el *workflow* que se realizó con los datos.

Como se puede observar, en primer lugar, se dividió el mapa de la Ciudad Autónoma de Buenos Aires en una grilla. En un principio se probaron diferentes tamaños para la misma hasta encontrar el que mejor se adaptaba al problema propuesto. Luego, se combinaron diferentes *features*, tanto espaciales como temporales y de distancias, para los dos tipos de delitos: hurto y robo. Después de haber calculado el valor de cada variable para cada celda de la grilla, se determinó como variable objetivo el nivel de delitos registrados del mes siguiente. Con el objetivo de predecirla, se preparó el conjunto de datos de entrenamiento con los meses previos, para ser utilizado en diferentes algoritmos como *XGBoost* y *Random Forest*. Por último, se evaluaron dichos modelos en función del Error Cuadrático Medio Logarítmico (RMSLE).

Figura 1. Flujo de trabajo de los datos



1. Generar una grilla a partir de los límites de la ciudad. Intersecar la grilla y el mapa M, eliminando las cuadrículas que no están en el mismo.
$M = \{g_1, g_2, g_3, \dots, g_n\}$.

2. Incorporar *features* en cada celda de la grilla $(x_1, x_2, x_3, \dots, x_i)$. Establecer los delitos que ocurren en el mes siguiente como variable objetivo (y). Si no ocurre ningún delito en la cuadrícula, $y = 0$. De lo contrario, $y \geq 1$. En conclusión, $g = \{y, x_1, x_2, x_3, \dots, x_i\}$.

3. Entrenar algoritmos de aprendizaje automático utilizando las variables creadas. Predecir los delitos del mes siguiente para cada cuadrado de la grilla.

4. Evaluar la *performance* con la métrica elegida.

3.2 Definición de la Grilla

En esta sección definiré qué se considera una grilla y el motivo por el que se la utilizó en este análisis. Luego, comentaré los pasos necesarios para crear la misma. Por último, discutiré el tamaño seleccionado en este trabajo y el *trade-off* que el mismo representa.

En primer lugar, para poder entender qué es una grilla, usaré la descripción de la función que se utilizó para generar la cuadrícula, `st_make_grid()`. La misma dice: "Crea una teselación regular sobre el cuadro delimitador de la geometría de un objeto `sf` o `sfc`"[CRAN, 2021]. En otras palabras, esta función se utiliza para crear una cuadrícula cuadrada o hexagonal que cubra el cuadro delimitador de la geometría de este tipo de objetos. Este cuadrado delimitador, también conocido como *bounding box*, es una matriz de las coordenadas mínimas y máximas de un objeto de esta clase. Es decir, es el área definida por las latitudes y longitudes máximas y mínimas. En la Tabla 1 se puede encontrar los datos para la Ciudad Autónoma de Buenos Aires.

Tabla 1. *Bounding Box* de CABA

	Longitud	Latitud
Mínima	-58.53152	-34.70529
Máxima	-58.33515	-34.52649

Por lo tanto, la grilla no es más que una cuadrícula cuadrada que está delimitada por los límites del territorio en el que se aplique.

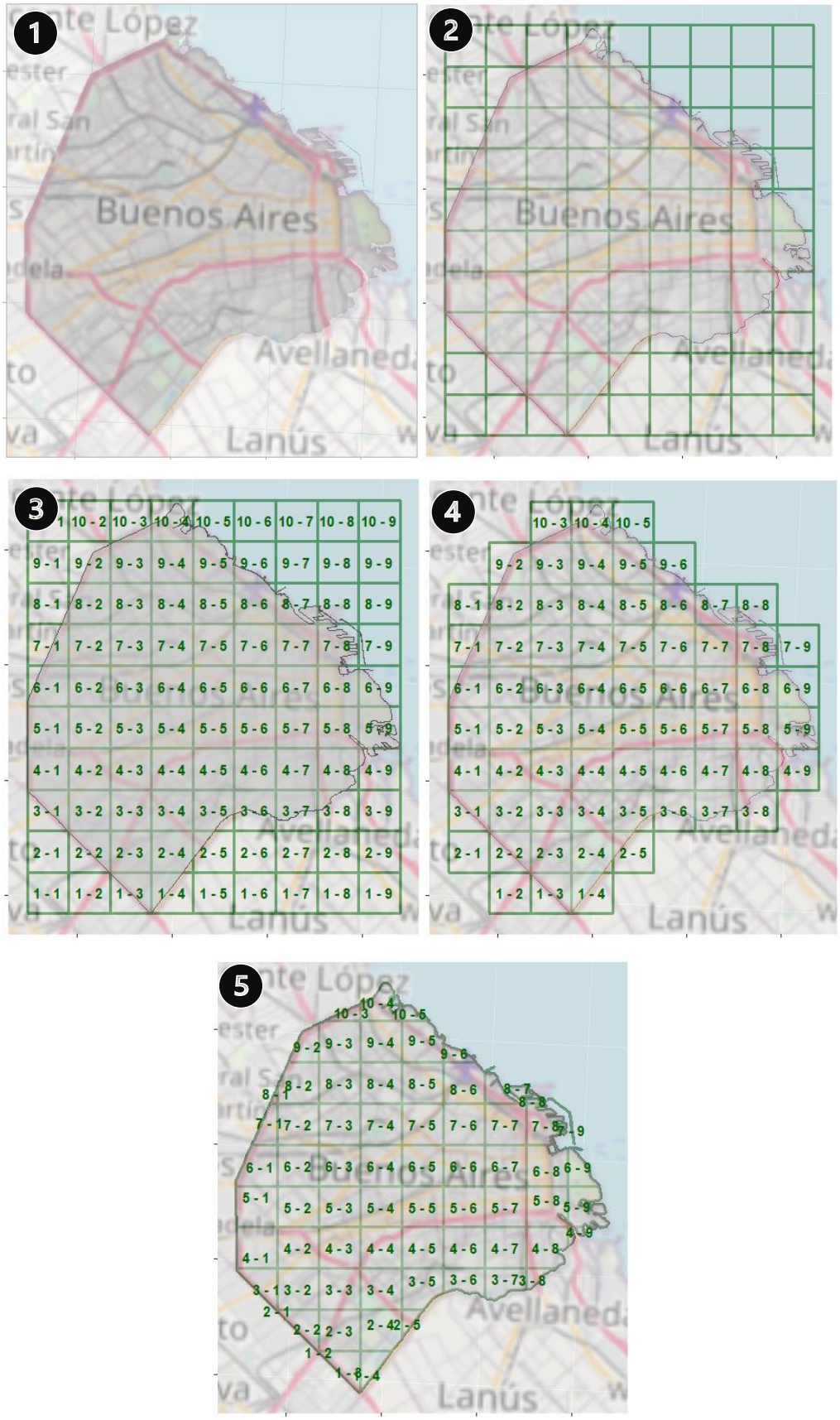
Este tipo de enfoques basados en grilla, permiten de una manera simple subdividir un territorio de manera tal de poder ubicar distintas características dentro de cada celda. Es por esto que, teniendo en cuenta que CABA cuenta con una alta densidad poblacional, se podrán identificar tanto variables espaciales como cantidad de delitos a lo largo y a lo ancho de la ciudad. De esta manera, es posible tener la granularidad que uno desee en cada porción de la cuadrícula según el tamaño que se elija para cada cuadrado.

Para generar la grilla fue necesario seguir los siguientes pasos:

1. Seleccionar el territorio.
2. Generar una teselación regular cuadrada sobre el *bbox* de la ciudad.
3. Generar un ID para cada cuadrado de la grilla.
4. Quedarse solo con aquellos cuadrados que contienen territorio.
5. Generar la intersección con la superficie de CABA.

En la Figura 2 se puede ver el proceso que se realizó para llegar a la grilla final. En la misma, se puede observar una cuadrícula de menor resolución que la que se utilizó finalmente en el análisis, ya que permite visualmente entender mejor el proceso (cuadrados con un tamaño 2.000 x 2.000 metros):

Figura 2. *Workflow* de una grilla con resolución de 2.000 x 2.000 metros



En caso de querer entender con mayor detalle cómo se construye una grilla paso por paso ir al Apéndice B.

En esta etapa del *workflow*, la principal decisión está dada por el tamaño que tendrá la grilla. Particularmente, hay un *trade-off* entre la utilidad y la confiabilidad de los resultados. Por un lado, es importante que las áreas de predicción de los delitos sean lo suficientemente pequeñas como para obtener una mayor utilidad y por el otro, usar áreas más grandes da una mayor confiabilidad en los resultados de la predicción. En otras palabras, áreas más pequeñas, harán que los resultados sean más útiles para los organismos encargados de hacer cumplir la ley, ya que permitirían asignar los recursos de prevención con mayor precisión. Sin embargo, la poca cantidad de puntos de interés dentro de las celdas de la cuadrícula impedirían descubrir patrones generalizables que aporten valor predictivo. En cambio, una cuadrícula de menor resolución garantizaría que haya suficientes POI dentro de cada celda, pero las predicciones resultantes de los delitos seguramente resultarían de una utilidad práctica más cuestionable. [Cichosz, 2020]

Teniendo en cuenta lo anterior, se definió una grilla para subdividir la ciudad de 200 x 200 metros en cada cuadrado. Sabiendo que la mayoría de las manzanas en la ciudad tienen una dimensión de 100 x 100 metros, se decidió que los cuadrados de la grilla representaran 2 x 2 manzanas.

Como resultado, la cuadrícula cuenta con 5.387 celdas, con un total de 91 columnas y 101 filas. Lógicamente, como se mencionó antes, al dejar solo aquellos cuadrados que se intersecan con el límite de la ciudad, la cantidad total de celdas no corresponde a $filas * columnas$ sino que es menor.

3.3 Trade-Off Temporal

Al igual que en el caso de la cantidad de celdas de la grilla a nivel temporal también existe un *trade-off* entre cantidad de tiempo y utilidad de los resultados. Por un lado, a mayor cantidad de tiempo que se toma como variable objetivo, menor será la utilidad para los organismos policiales para tomar acciones ya que no sabrán en que momento de ese ciclo se dará. Por el otro lado, cuanto menor sea el período elegido menor será la cantida de delitos que se tiene, lo que podría generar que el modelo prediga de manera constante que no se dará ningún delito.

Teniendo esto en cuenta y dada la cantidad de delitos que se dan en CABA en distintos períodos de tiempo, se decidió elegir un mes como variable objetivo. En otras palabras, se buscará estimar la cantidad de delitos del mes siguiente.

3.4 Ingeniería de Atributos

El propósito de la ingeniería de atributos es extraer características de datos sin procesar, transformándolos para que se ajusten mejor al modelo de aprendizaje automático en el que se quieran usar como *input*. Lo que aprende el algoritmo dependerá en gran

medida de los atributos de los que disponga. Por lo tanto, es un paso fundamental dado que los *features* correctos pueden aliviar la dificultad del modelado y, por lo tanto, ayudar a que el modelo genere resultados de mejor calidad. Tanto es así que muchos profesionales de la industria coinciden en que la gran mayoría del tiempo del *pipeline* de creación de modelos predictivos se dedica a la ingeniería de atributos y la limpieza de datos. [Zheng and Casari, 2018]

En las siguientes subsecciones se mostrarán las distintas variables que se generaron para aportar valor predictivo a los modelos.

3.4.1 Variables espaciales

Las variables espaciales son *feautures* que poseen atributos geográficos. Cada observación representa una ubicación en la Tierra, con sus respectivas latitudes y longitudes. En cuanto a su forma, las mismas pueden ser polígonos, líneas o puntos.

La principal variable de este tipo son los robos y hurtos. Otros tipos incluyen los puntos de interés, las regiones de interés y los datos socioeconómicos. En esta misma línea, también se incorporaron variables que indican el barrio, área en metros cuadrados del mismo, comuna, latitud y longitud a la que pertenece cada celda de la grilla. Las variables de barrio y comuna se incorporan a la base de datos como variables categóricas.

3.4.2 Variables temporales

Las variables temporales son aquellas que hacen referencia a lo sucedido en una misma celda de la cuadrícula en diferentes momentos del tiempo. Particularmente, en este trabajo se utilizaron variables que indican el clima, como las precipitaciones, velocidad del viento y temperatura, así como también, los delitos.

Para las variables meteorológicas, los valores serán distintos para cada mes del análisis, pero serán iguales para todas las celdas de la grilla. Específicamente, dado que estas variables tienen una periodicidad diaria, se las agrupó de manera mensual para poder ser incorporadas. En el caso de la temperatura y velocidad del viento se calculó el promedio para cada mes. En cambio, para las precipitaciones, se computó la suma.

Por el contrario, para los delitos los valores cambiarán tanto para cada mes como para cada celda de la grilla. Dado que al tener la latitud y longitud de cada registro, es posible asignar cada uno de estos en una cuadrícula. Sin embargo, al igual que con las variables climáticas, la periodicidad que se tiene es diaria e incluso en el conjunto de datos original se encontraba la hora. Por consiguiente, también se procedió a agrupar la cantidad a nivel mensual realizando la suma.

A partir de utilizar un método conocido como *sliding window* o ventana deslizante se busca evitar que algún período que presente anomalías en el pasado tenga un impacto tan directo y significativo en el resultado actual, logrando así una base más confiable y estable para predecir delitos futuros [Chainey et al., 2008]. El algoritmo utilizado consiste en pararse en un mes del análisis y calcular el número de delitos acumulado en cada celda de la grilla para el mes previo, los últimos 3, 6, 9 y 12 meses y el valor del mismo mes pero del

año anterior. Lo particular de esta técnica es que el procedimiento es aplicado de manera móvil, moviendo el mes que se considera como actual hasta llegar al último período del análisis, diciembre 2021 [Zambrano, 2021][Lin et al., 2018]. Dado que la variable objetivo son los delitos en el mes de diciembre, y que se necesitan 12 meses de historia para poder calcular las distintas variables, se decidió utilizar como período inicial a diciembre de 2017. De esta forma, es posible calcular los últimos 12 meses para todos los períodos. Conformando así un total de 49 meses.

Este procedimiento también fue aplicado en las variables climáticas. Sin embargo, al momento de entrenar el modelo no se utilizaron las variables meteorológicas del mes actual. Dado que cuando el modelo se implemente en nuevos datos (futuros), las mismas no serán conocidas.

Por último, se incorporaron variables que indican el mes y año.

3.4.3 Variables de distancias

Como se menciona anteriormente, en muchos casos, la cercanía a ciertos puntos de interés puede propiciar, o por el contrario, dificultar un robo o hurto. Por lo tanto, se consideró importante incorporar variables de distancia. Las mismas fueron calculadas teniendo en cuenta la distancia Euclídea entre el punto de interés más cercano y el centroide de cada celda de la grilla.

En primer lugar, se calculó el centroide de cada cuadrado. Luego, utilizando la función `st_nearest_feature()` se determinó el punto de interés más cercano a cada centroide. Por último, se obtuvo la distancia en metros entre cada centroide y el POI o ROI más cercano.

Los puntos de interés que se consideraron relevantes para calcular las distancias fueron: comisarias, cuarteles de bomberos y hospitales. Así como también, se calculó la distancia a una región de interés: los barrios populares.

Por otro lado, a pesar de no haberlas utilizado como POIs, se computó la distancia a avenidas, túneles y subidas, bajadas y enlaces de autopistas. La referencia geográfica se obtuvo a partir de un *shapefile* de BA Data. Algo a tener en cuenta, es que no son puntos, sino que en algunos casos son líneas y en otros, como en el de las subidas, bajadas y enlaces de autopistas, son multi-líneas. Por esta razón, para calcular la distancia al centroide de cada cuadrado de la grilla, primero, se calculó el centroide de cada línea o multi-línea y a partir de ello se siguió el mismo proceso que con los puntos y región de interés.

Teniendo en cuenta todos los tipos de variables mencionados anteriormente el conjunto de datos utilizado posee un total de 163 variables. Las mismas son:

- Variable de interés: cantidad de delitos en el mes actual.
- Variables de interés de periodos anteriores: 5 variables con el acumulado en los últimos 1, 3, 6, 9 y 12 meses.
- Variable de interés un año atrás.

- 2 Variables de temporalidad: mes y año de los delitos actuales.
- 66 Variables territoriales: barrio, área del barrio, comuna, latitud y longitud de cada celda de la cuadrícula. [3]
- 60 Variables espaciales geográficas: Puntos o Regiones de Interés (POIs - ROIs).
- 3 Variables espaciales socioeconómicas.
- 3 Variables temporales climáticas de los periodos anteriores y de un año atrás (18 en total).
- 7 Variables de distancias.

3.5 Aprendizaje automático

En esta sección se describirán los distintos modelos que se utilizarán para predecir la cantidad de delitos del mes de diciembre de 2021.

En todos los modelos propuestos, a excepción del modelo *Naïve*, fue utilizada la librería `caret` [Kuhn, 2008]. La misma es una librería que facilita el proceso de creación de modelos predictivos en `R`. "El paquete `caret` (*Classification And REgression Training*) se creó para agilizar el proceso de creación y evaluación de modelos predictivos. Utilizándola, es posible evaluar rápidamente muchos tipos diferentes de modelos para encontrar la herramienta más adecuada para los datos"[Kuhn and Johnson, 2013]. En palabras simples, permite utilizar distintos algoritmos de *machine learning*, así como también ajustar los distintos hiperparámetros y el resto de los argumentos de cada modelo de una forma sencilla.

En las siguientes subsecciones se describirán cada uno de ellos.

3.5.1 *Naïve* Model

El objetivo de este modelo es el de ser un *benchmark*, un modelo base. En otras palabras, busca establecer el límite inferior en la métrica a utilizar, que con algoritmos más complejos buscara ser superada. Como su nombre lo indica, *Naïve* (ingenuo), implica que se dedicará poco esfuerzo de manipulación en los datos y en la elección de hiperparámetos, si es que los tiene, para entrenarlo.

En este caso, teniendo en cuenta que no se busca un modelo sofisticado, se decidió utilizar un modelo de series de tiempo conocido como *Naïve Forecasting Method*. Este modelo es uno más simples para pronosticar y consiste en utilizar la observación más reciente como predicción del modelo. Esto implica que en este caso, como se busca estimar la cantidad de delitos del mes de diciembre de 2021, se utilizará como *output* del modelo el valor que se tiene para el mes de noviembre del mismo año.

[3] Dado que no todos los algoritmos de aprendizaje automático incorporan formas de lidiar con atributos categóricos, se les aplicó a las variables de barrio y comuna *one-hot-encoding*.

3.5.2 XGBoost

eXtreme Gradiente Machine (XGBoost) [Chen and Guestrin, 2016] es un modelo de ensambles de la familia de *boosting*. En particular, *XGBoost* aprende ensambles de árboles del tipo CART. Que sea un modelo de ensambles implica que se construye combinando predicciones de modelos más pequeños y menos complejos. Además, al ser del tipo *boosting* se construye de manera secuencial y aditiva, es decir, cada nuevo árbol tiene información de los árboles que se construyeron previamente. Por otro lado, como es posible sumar muchos árboles al modelo, algo fundamental de este algoritmo es que el mismo aprende lento. De esta manera, se evita sobre ajustar los datos de entrenamiento. [James et al., 2021]

El proceso que realiza consiste en entrenar un árbol, ver el error de dicho modelo y construir un nuevo árbol que se enfoque en ese error. Por último, se toma como predicción final una combinación de las predicciones de cada modelo. Por lo tanto, la construcción de cada nuevo árbol va a ir dependiendo de los modelos anteriores.

Los hiperparámetros que se suelen modificar son:

- *nrounds*: Cantidad de árboles. $(0; +\infty]$
- *max_depth*: Profundidad máxima de los árboles, $(0; +\infty]$
- *eta*: Tasa de aprendizaje. $[0; 1]$
- *gamma*: Reducción mínima del error para generar una nueva partición. $[0; +\infty]$
- *colsample_bytree*: Porcentaje de columnas utilizadas para cada árbol. $(0; 1]$
- *subsample*: Porcentaje de observaciones utilizadas para cada árbol. $(0; 1]$
- *min_child_weight*: Cantidad mínima de observaciones por hoja. $(0; +\infty]$

3.5.3 Random Forest

Random Forest [Breiman, 2001] es un modelo de ensambles que pertenece a la familia de *bagging* o también conocido como agregación *bootstrap*. Al igual que en *XGBoost*, al ser un modelo de ensambles, se combinan varios árboles de decisión para construir el modelo final. Pero en este caso, al utilizar *bagging*, la forma en la que los errores se compensan entre sí, es entrenando cada árbol solo con un subconjunto de los datos de entrenamiento. Es decir, en cada modelo nuevo se elige una proporción de observaciones de los datos para entrenar.

En particular, *Random Forest* propone una mejora al algoritmo de *bagging*, ya que al usar *bootstraps* los árboles suelen estar muy correlacionados. A pesar de ver una submuestra de los datos, las variables siempre son las mismas. Por lo tanto, *Random Forest* propone un muestro doble: por un lado, observaciones y, por otro lado, *feautures*. Eligiendo en cada nodo, un subconjunto de variables también. De esta forma, se logra decorrelacionar los árboles. [James et al., 2021]

Los hiperparámetros que se suelen modificar son:

- *mtry*: Número de variables seleccionadas aleatoriamente en cada árbol.
- *min.node.size*: Tamaño mínimo que tiene que tener un nodo para poder ser dividido.
- *splitrule*: Criterio de división.

Además, se debe elegir la cantidad de árboles.

En este caso, se ha decidido utilizar estos dos algoritmos: *XGBoost* y *Random Forest* debido a su amplia aceptación en la industria y su destacado desempeño en problemas de aprendizaje automático, incluyendo dentro de estos, la predicción del crimen. Estos modelos fueron seleccionados considerando la amplia cantidad de variables independientes que se iban a utilizar y su capacidad para manejar grandes conjuntos de datos de manera eficiente, así como su facilidad para ser implementados. Además, ambos algoritmos demostraron una alta precisión en este contexto [Yuki et al., 2019][Bogomolov et al., 2014]. A pesar de que en la literatura previa también se han empleado algoritmos más complejos, como las redes neuronales, para esta tesis, *XGBoost* y *Random Forest* resultaron ser más prácticos y eficientes computacionalmente.

3.6 Optimización de hiperparámetros

Los hiperparámetros son aquellos parámetros que uno de antemano debe definir y tienen un impacto en la performance de los modelos. Son parámetros que los algoritmos no aprenden sino que deben estar dados. Además, son los que nos permiten, si están bien elegidos y en conjunto con el esquema de validación, evitar hacer *overfitting*. Este último concepto entendido como ajustar demasiado los datos de entrenamiento, captando en exceso las particularidades, lo que lleva a un error en el conjunto de entrenamiento mucho más bajo que en el de *test*. Si esto sucede, el modelo no será bueno para predecir en datos desconocidos.

En relación con la estrategia elegida para encontrar los mejores hiperparámetros para cada modelo, se decidió utilizar *random search*. Dado que computacionalmente no es factible probar todas las combinaciones posibles de hiperparámetros entre un valor mínimo y un máximo, como plantea *grid search*, este método propone definir cada hiperparámetro entre esos dos valores y al azar seleccionar distintas combinaciones. De esta forma, eligiendo una cantidad suficiente de puntos se lograría explorar bien el espacio de posibles valores.

3.7 Esquema de validación

En esta sección se describirá el esquema de validación que se utilizó para los modelos anteriormente mencionados. Como objetivo principal se tiene poder simular datos desconocidos para evaluar la performance de los mismos.

Se utilizó el esquema de *hold-out set*. Es una de las maneras mas simples de simular datos desconocidos, pero es la que menos computo requiere. Por lo que, dada la gran

cantidad de observaciones que se tiene, resulta ser la más adecuada. Además, al ser datos que siguen una línea temporal, lo que se hizo fue dejar el mes de noviembre como validación y el de diciembre como evaluación. De esta forma, imitamos lo que sucedería si realmente quisiéramos estimar el futuro. Estos dos conjuntos poseen 5.387 observaciones cada uno, ya que es la cantidad de celdas en la grilla. Por otro lado, los datos desde diciembre de 2017 a octubre de 2021 se utilizaron como datos de entrenamiento. Conformando un total de 253.189 observaciones.

3.8 Métrica de evaluación

Dado que estamos frente a un problema de regresión, se utilizará como métrica de evaluación el Error Cuadrático Medio Logarítmico. El mismo considera el error relativo entre el valor predicho y el real, y la escala del error no es significativa. Solo la diferencia porcentual importa.

La fórmula es:

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\Big(\log(p_i+1) - (log(a_i+1)\Big)^2} \tag{1}$$

Dónde:

n = es el número total de observaciones en el conjunto de datos,
a_i = es el verdadero valor de i (no puede ser negativo), y
p_i = es el valor predicho.

En particular, se decidió utilizar esta métrica de performance, ya que la variable objetivo tiene un rango relativamente amplio de valores que puede tomar. La misma va desde el 0 hasta el 91. Por lo tanto, al utilizar el RMSLE se penaliza de manera porcentual en lugar de en valor absoluto.

4 Análisis Exploratorio

En esta sección se presentará un análisis exploratorio de los datos. El mismo estará centrado en los distintos *datasets* que se presentaron anteriormente y dado que se trata de información georreferenciada, se analizarán tanto en función de las características de las variables como de su espacio en la Tierra.

El objetivo de esta sección es, por un lado, comprender los distintos conjuntos de datos y cómo se comportaron a lo largo de los años y al mismo tiempo en las distintas zonas del territorio del análisis; y por el otro, a partir de ese entendimiento, poder tomar decisiones mejor informados, como qué variables son más relevantes para así incluirlas en los modelos.

Para esto, se estudiarán los distintos conjuntos de datos que se mencionaron en la sección de Datos de manera individual para luego poder analizarlos de forma conjunta.

4.1 Análisis de Delitos

Teniendo en cuenta que el conjunto de datos de **delitos** es el centro de este análisis, es importante poder entenderlo en profundidad para poder sacar *insights* valiosos que permitan tomar mejores decisiones con relación a qué variables incluir en los modelos, cómo combinarlas y qué transformaciones aplicarles.

A lo largo de esta sección se presentarán tablas, gráficos y mapas generados a partir de este conjunto de datos en particular.

En la siguiente tabla se podrá observar la cantidad de delitos diferenciando por tipo a lo largo de los diferentes años del análisis.

Tabla 2. Cantidad de delitos diferenciando por tipo y año

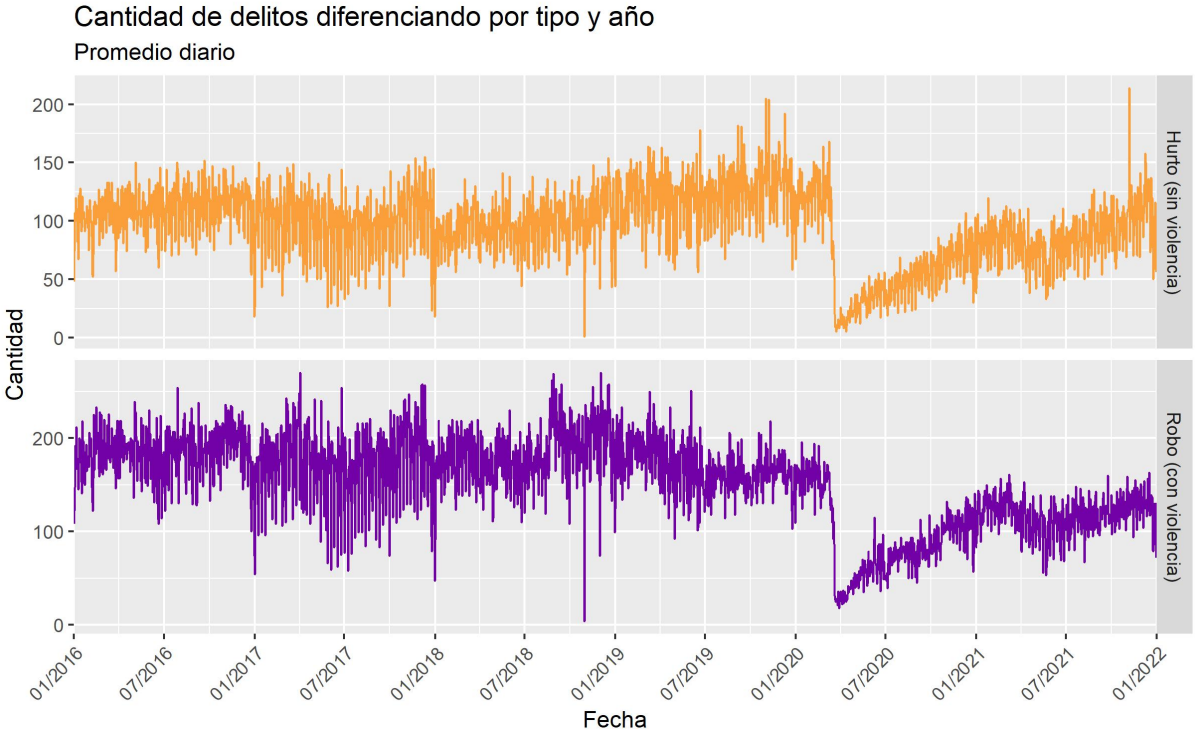
	2016	2017	2018	2019	2020	2021
Hurto (sin violencia)	39.945	36.098	35.222	44.618	22.735	31.707
Robo (con violencia)	67.511	63.124	66.574	61.310	33.511	43.074

En relación a los tipos de delitos, es posible observar que en todos los años los robos son superiores a los hurtos, representando casi un 60 % de las observaciones totales en cada año. Algo a destacar también es que en el año 2020, año en el que el territorio de CABA estuvo bajo cuarentena por la pandemia del COVID-19, la cantidad de delitos de los dos tipos disminuyó a la mitad prácticamente. En el caso del año 2021, los valores aumentaron, pero aún se encuentran por debajo de los que se tenían en los períodos pre-pandemia.

4.1.1 Análisis temporal

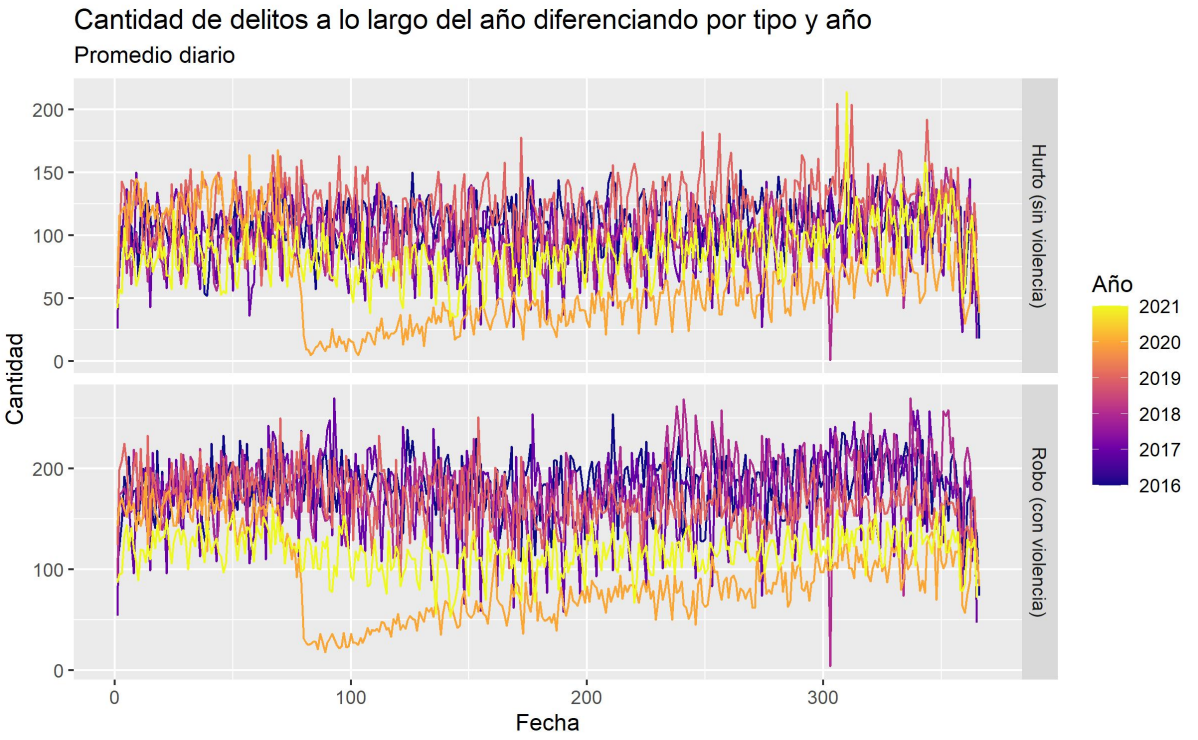
En el siguiente gráfico, se podrá observar la cantidad de delitos agrupada de manera diaria a lo largo de todos los años del análisis.

Figura 3. Tendencia de delitos diferenciando por tipo a lo largo de los años



En los años pre-pandemia se puede observar que no hay una tendencia muy clara a lo largo de los años. Por otro lado, sí es posible notar la gran baja de delitos que se tiene a comienzo del año 2020, en donde, como ya se mencionó, comenzó la cuarentena estricta en el territorio de CABA. En los meses siguientes, tanto para los hurtos como para los robos, se ve una tendencia creciente hasta diciembre de 2021. Sin embargo, como se puede observar en la siguiente imagen, dichos niveles nunca alcanzan las cantidades pre - pandemia.

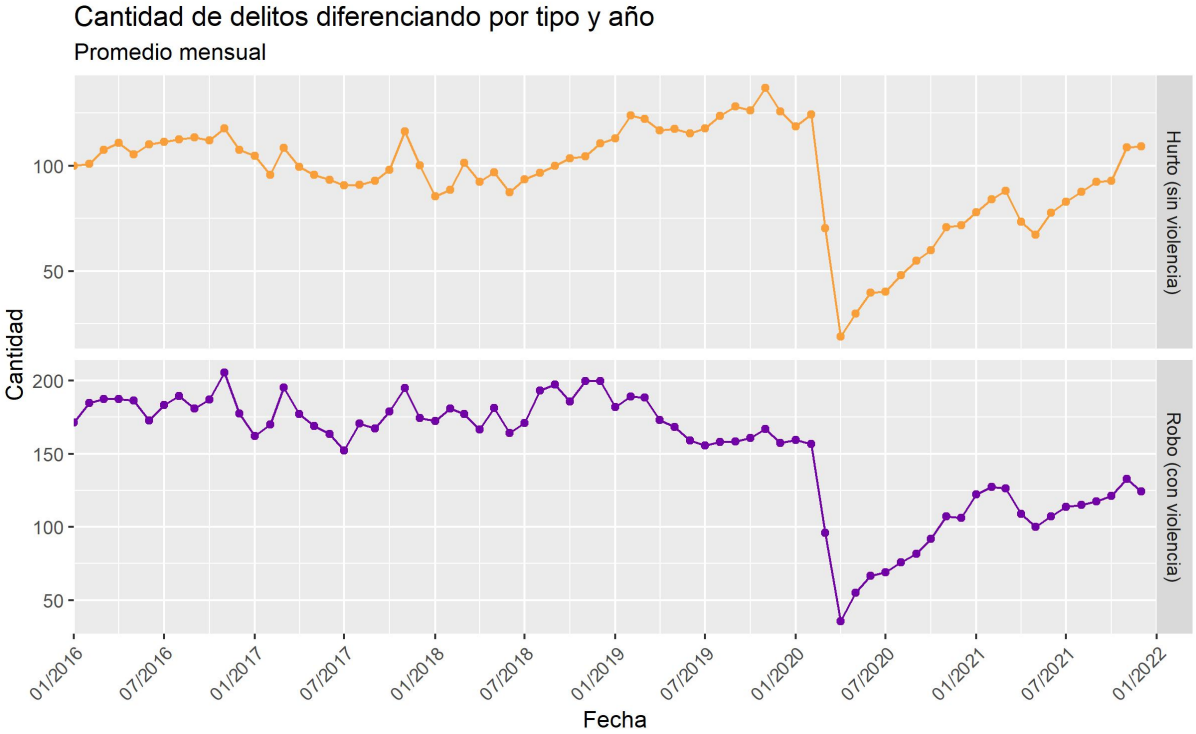
Figura 4. Distribución de los días del año por tipo y año



Al ver superpuestas las tendencias de todos los años queda aún más en evidencia que el comportamiento de los primeros meses del año 2020 seguía una forma similar a los períodos anteriores. Asimismo, para el 2021 se puede observar que, para ambos tipos, las cantidades son inferiores. Principalmente para los robos y, para la primera mitad del año, también para los hurtos. Sin lugar a dudas, estas anomalías serán algo que se deberá tener en cuenta a la hora de querer estimar la cantidad de delitos para el mes de diciembre de 2021.

En la siguiente figura se podrá observar la cantidad de delitos promediados de manera mensual.

Figura 5. Tendencia de delitos diferenciando por tipo - Promedio mensual

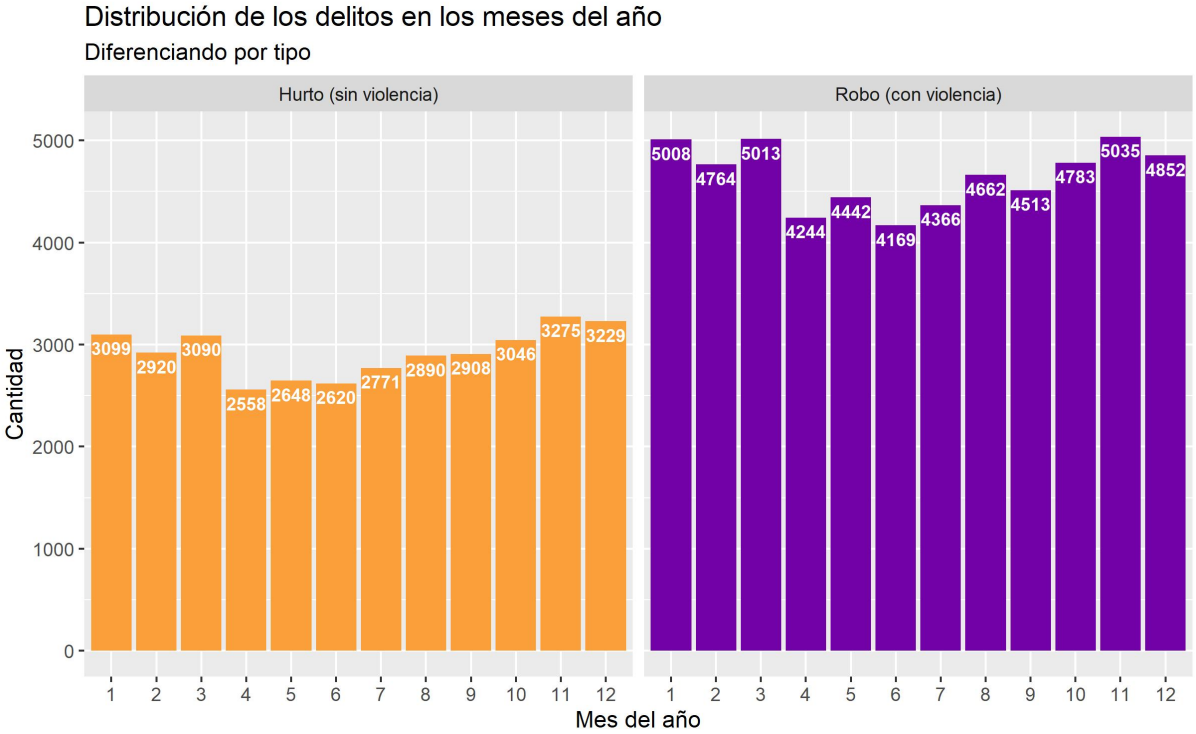


Las diferencias y similitudes se mantienen con las mencionadas para el gráfico anterior. Al igual que en la figura previa, en ambos tipos es posible observar una baja a comienzos de 2020 y una tendencia creciente a partir de ese entonces, hasta fines de 2021. En cambio, la escala, al ser un promedio mensual, genera que sea aún más notaria la diferencia, mostrando que los robos son en promedio casi el doble que los hurtos.

Por estos motivos, sin lugar a dudas, la variable de fecha será importante a la hora de predecir la cantidad de delitos.

Por último, para concluir con el análisis de los tipos de delitos en variables que indican temporalidad, se analizarán las cantidades de ambos tipos de delitos por separado para cada mes del año.

Figura 6. Distribución de los meses del año por tipo



En este caso, se puede observar que los meses con menor cantidad tanto de robos como de hurtos son abril y junio. En cambio, los de mayor cantidad son noviembre, diciembre, enero y marzo. Sin embargo, la diferencia porcentual entre el mayor y el menor mes no es tan grande. Para los hurtos es del 28 % y para los robos de un 21 %.

Esta última no es una variable que parezca a priori tener tanto poder predictivo, pero de todas formas será interesante incorporarla en los modelos para ver su impacto.

4.1.2 Análisis espacial

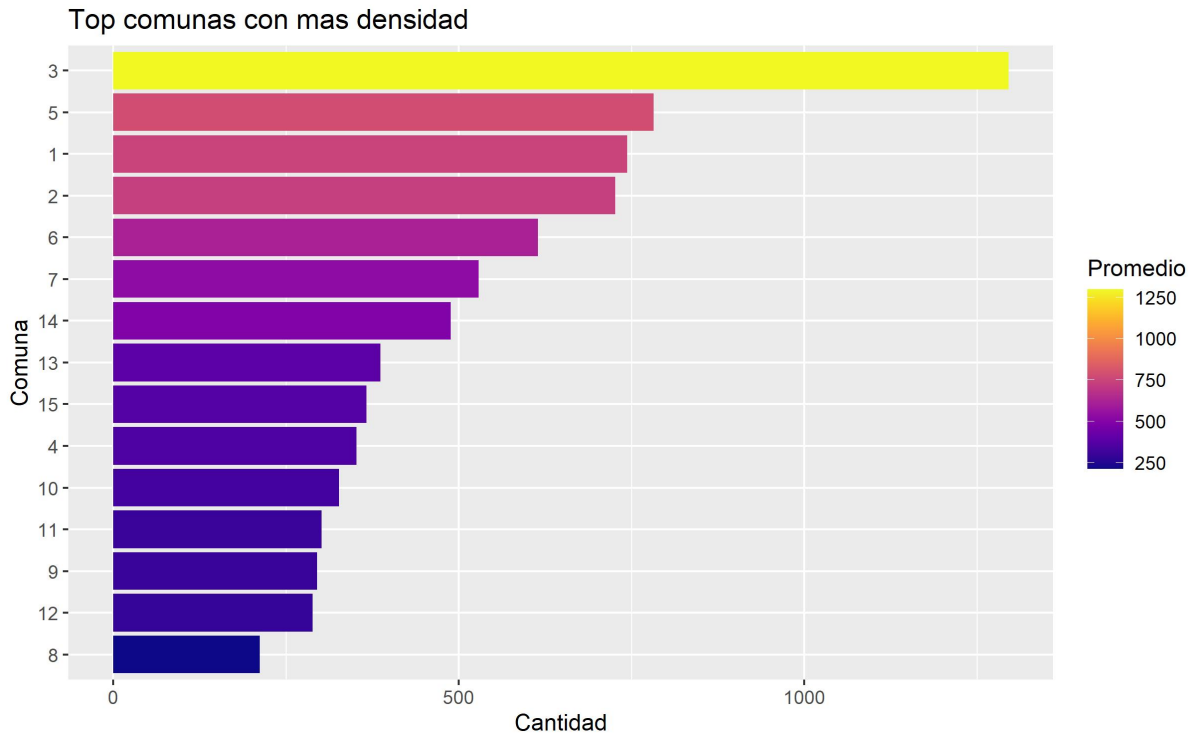
Para poder analizar la cantidad de delitos en cada barrio o comuna de la Ciudad Autónoma de Buenos Aires es importante tener en cuenta la superficie de cada uno. Esto se debe a que si se trata de un barrio o comuna con una gran superficie, probablemente tenga una mayor cantidad de registros de delitos. Pero la conclusión de que podría tratarse de un barrio "peligroso"podría ser errónea, ya que la cantidad de hurtos y robos está relacionada también con el tamaño del territorio. Por este motivo, se decidió calcular la densidad. Para esto, se dividió la cantidad de delitos por el área de cada barrio o comuna. De esta forma, se vuelven comparables los valores. [4]

En el siguiente gráfico se podrán observar las comunas ordenadas en función de mayor a menor según la densidad de delitos. La misma fue calculada utilizando el promedio de crímenes por comuna divido el área en km^2 de cada una.

[4]Cabe mencionar que se podría haber calculado la cantidad de habitantes de cada barrio o comuna, en lugar de haber utilizado los kilómetros cuadrados.

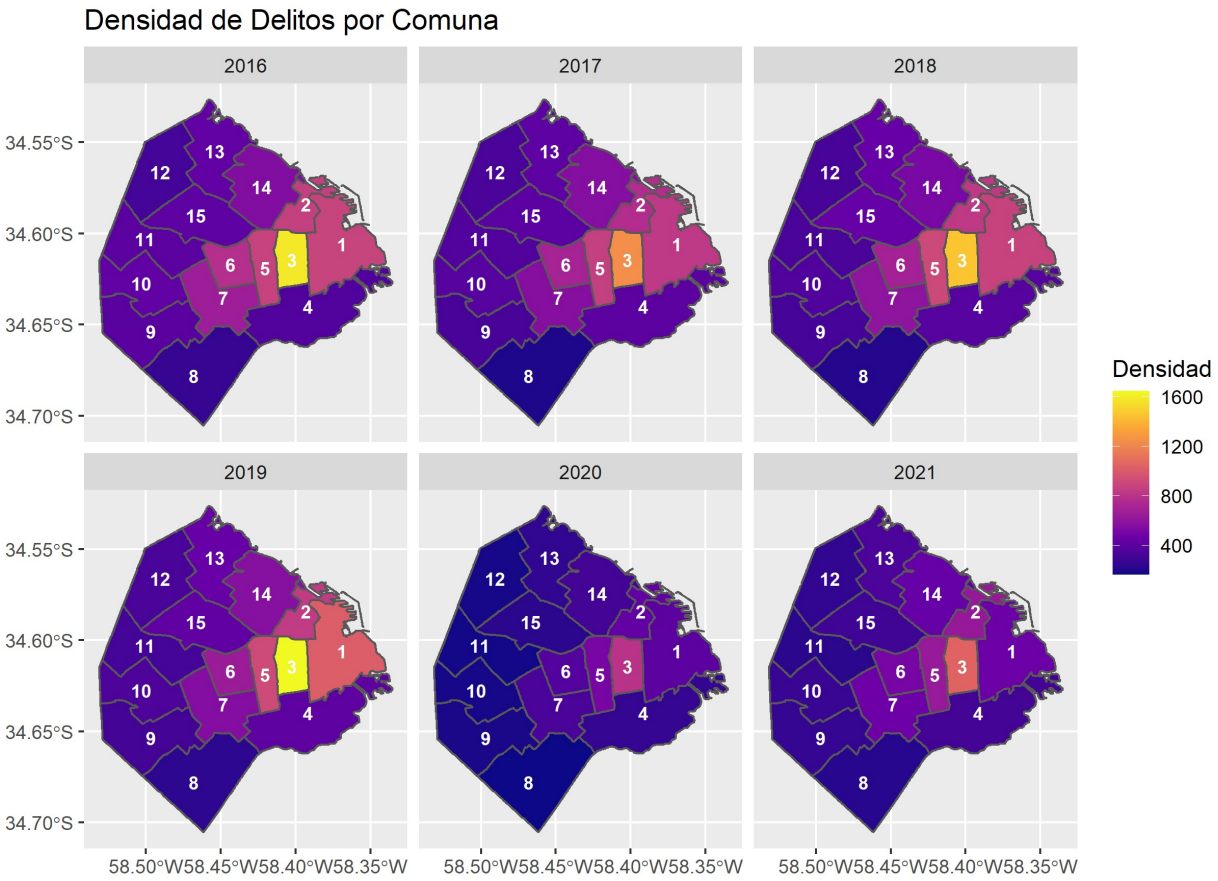
Como se puede observar, por una amplia diferencia, la comuna con mayor densidad de registros de delitos en promedio es la 3 con 1.295. Seguida de la 5 y 1 con 782 y 744, respectivamente. En contraste, la comuna con menor cantidad es la 8 con 212. Seguida por la 12, 9 y 11 con 289, 295 y 301, respectivamente.

Figura 7. Top comunas con más densidad de registros



Para continuar con el análisis, se realizó un mapa coroplético distinguiendo por comuna y año.

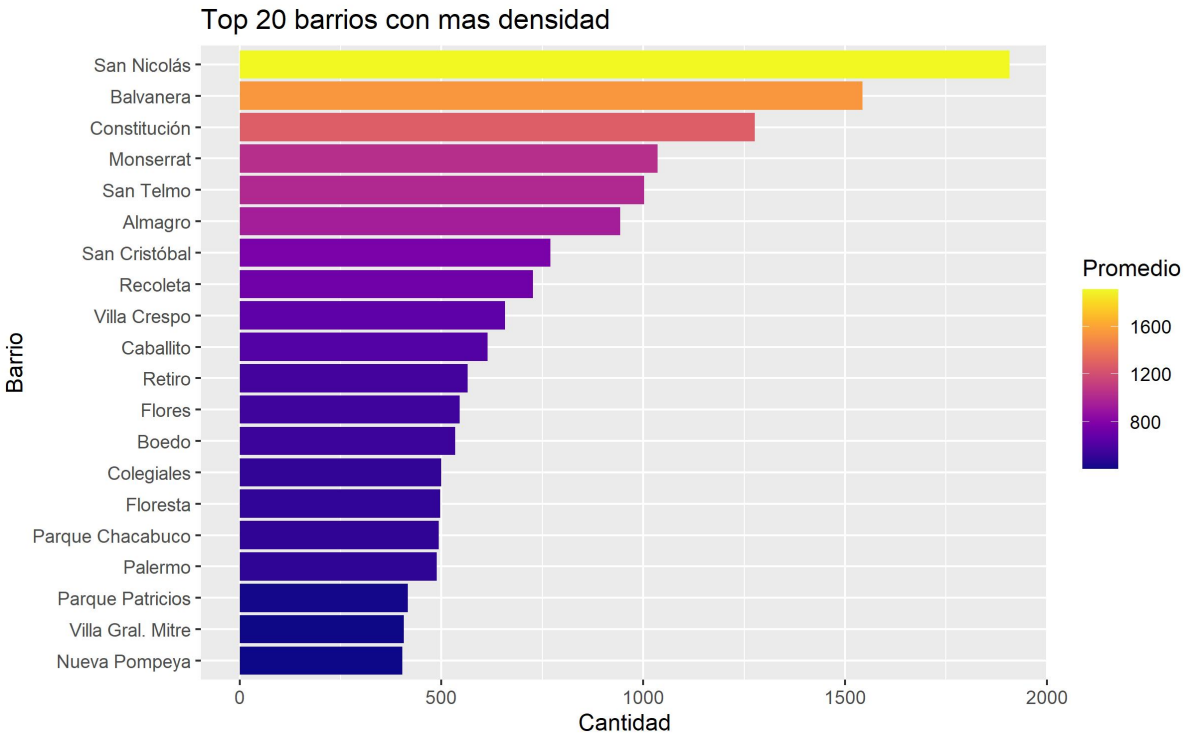
Figura 8. Cantidad de delitos por comuna por año



Al igual que en los análisis anteriores, se puede ver que la cantidad de delitos es menor para los últimos dos años del análisis. Por otro lado, para todos los años es posible observar que la comuna 3 es la que mayor cantidad de delitos tiene. 2019 es el año con mayor cantidad, con más de 1600 en promedio. Por el contrario, la comuna con menor cantidad de delitos para todos los años es la 8, con un mínimo en el año 2020 de menos de 165 delitos en promedio. Seguida de la 11, 10 y 12 para ese mismo año. Todas con menos de 180 delitos.

El mismo análisis se realizó por barrio. En el siguiente gráfico se podrá observar el top 20 de barrios con más densidad de delitos. Al igual que con las comunas, la misma fue calculada utilizando el promedio de delitos por barrio, divido el área en km^2 de cada uno.

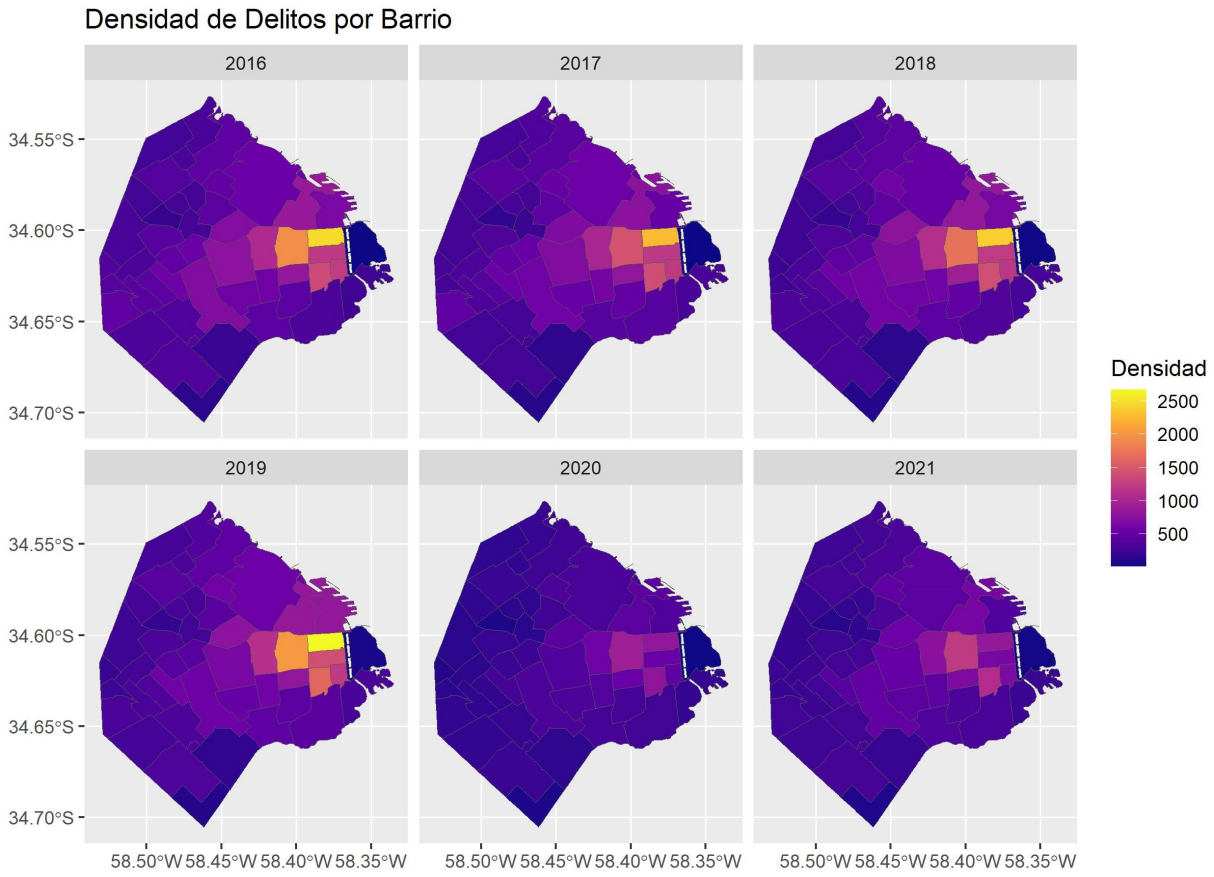
Figura 9. Top 20 barrios con más densidad de registros



A partir de la figura, se puede decir que San Nicolás es el barrio con mayor cantidad de registros de delitos, con un promedio de 1.908, seguido de Balvanera con 1.543 y Constitución con 1.276. A pesar de que en el gráfico no se pueda observar, el mínimo se encuentra en el barrio de Puerto Madero con 25 registros. Luego, se encuentra Villa Riachuelo con 97 y Agronomía con 149 en promedio.

También, al igual que con las comunas, se realizó el mapa coroplético distinguiendo por barrio y año. Para permitir una mejor lectura, se realizó un gráfico extra para el 2021 en donde se pueden leer los barrios y comunas. El mismo se puede encontrar en el Apéndice C

Figura 10. Cantidad de delitos por barrio por año



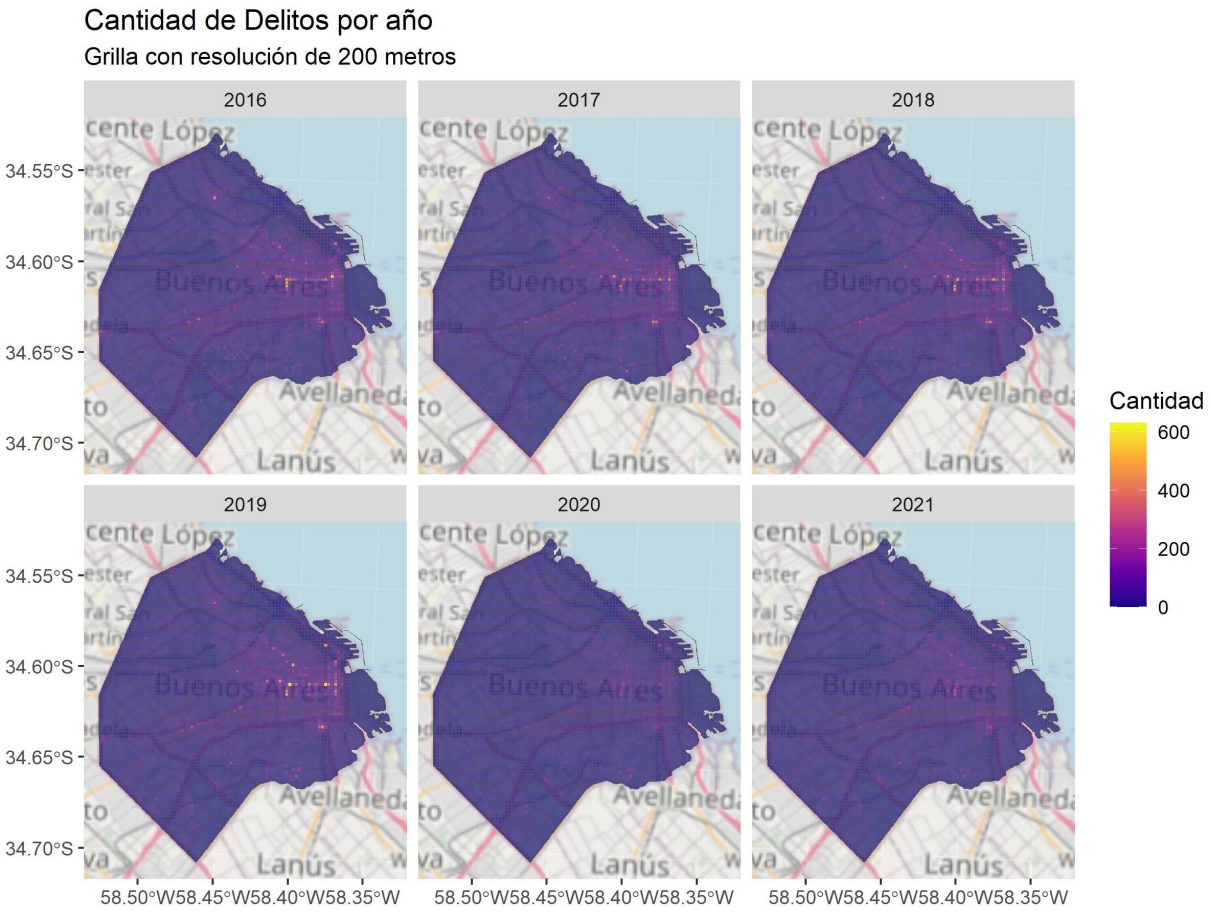
En la figura se puede observar que a pesar de que la comuna 3 tenía la mayor cantidad de delitos para todos los años, el barrio que particularmente tiene el mayor valor es San Nicolás, con más de 2.660 registros en el año 2019. El mismo pertenece a la comuna 1. Sin embargo, en los años 2020 y 2021 el barrio con mayor cantidad es Balvanera. El cual sí pertenece a la comuna 3. Por el contrario, para todos los años los barrios con menor cantidad son Puerto Madero y Villa Riachuelo. Con un mínimo en 2017 en Puerto Madero de 14 registros y en Villa Riachuelo en 2020 con 65. Los mismos pertenecen a la comuna 1 y 8 respectivamente. Esto demuestra, que como en el caso de la comuna 1, se puede tener el barrio con mayor cantidad de registros y el de menor cantidad contiguos. [5] Por lo que si se hubiera analizado el volumen de registros a nivel comuna, se habría pasado esta información por alto.

Teniendo lo anteriormente mencionado en cuenta, potencialmente la variable de barrio podría ser relevante para explicar los delitos en mayor medida que la que indica las comunas.

Agregando al análisis anterior, en el siguiente gráfico se podrá observar para cada celda de la grilla la cantidad de delitos que se tienen, distinguiendo por los años del análisis.

[5] Por este motivo, el promedio de dicha comuna es de aproximadamente 750, los valores del barrio con mayor y menor cantidad de registros se compensan.

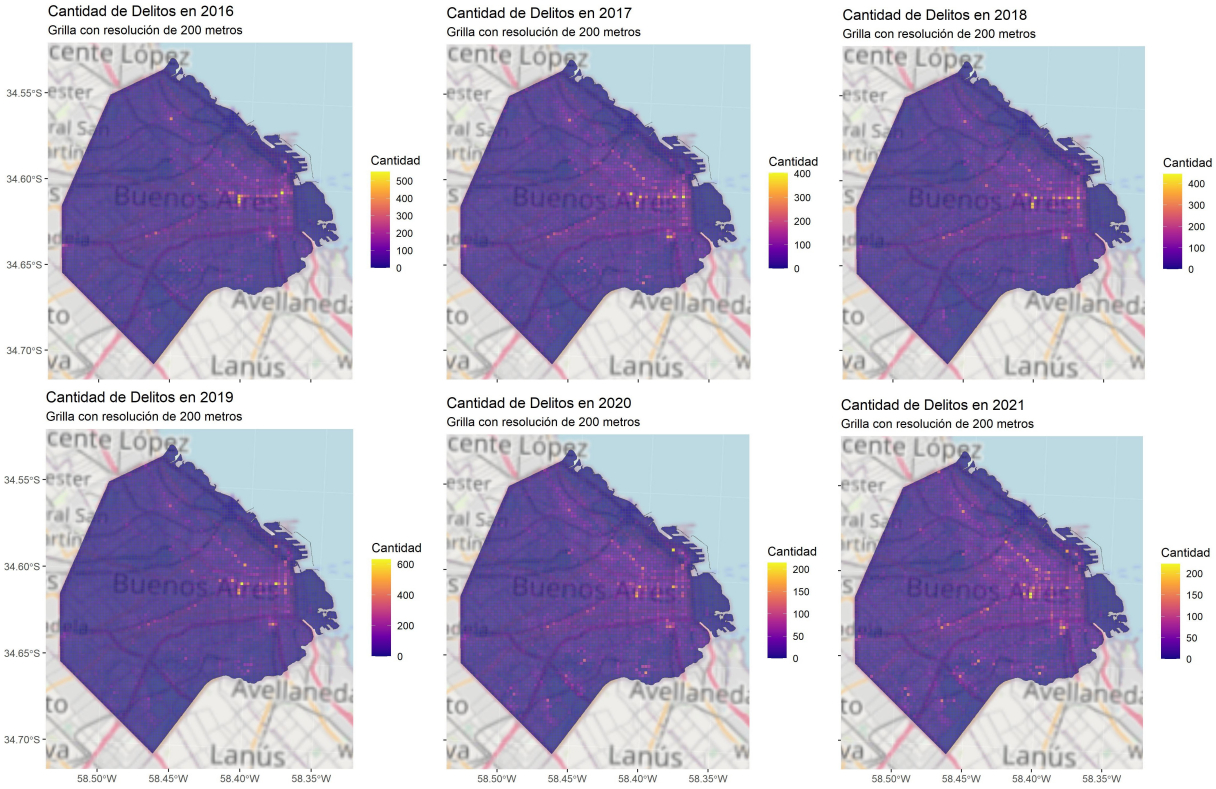
Figura 11. Cantidad de delitos en la grilla por año



Como se mencionó anteriormente, en los primeros cuatro años del análisis se tiene una mayor cantidad de delitos y eso se replica a nivel celda de la grilla. Por lo tanto, es posible observar que los mayores valores, aquellas celdas en amarillo, se encuentran desde el 2016 al 2019, inclusive. Particularmente, dichos cuadrados de la grilla se encuentra en el centro este de la Ciudad Autónoma de Buenos Aires. Con un máximo de 630 registros de delitos en 2019 para una misma celda y un mínimo de 0 para varias celdas y años del análisis.

Para poder observar más en detalle cada año, se realizó un gráfico para cada año de forma tal que cada uno tenga su propia escala.

Figura 12. Cantidad de delitos en la grilla por año



A partir de liberar la escala para cada año, es posible observar que tanto para el 2020 como para el 2021 el máximo en una celda es de un poco más de 200. Luego, para el 2018 y 2017 es de 400. Por último, los máximos se encuentra en 2019 y 2016 con 600 y 500, respectivamente. Sin embargo, lo que sí tienen en común todos los años es la zona con cuadrados de la grilla con mayor cantidad de registros; el centro - este. Esto coincide tanto con los barrios como con las comunas con mayor cantidad de delitos. Además, para los últimos dos años del análisis también se pueden observar celdas amarillas en el sur de la ciudad.

4.2 Análisis de Puntos y Regiones de Interés

En esta sección, se analizarán los conjuntos de datos relacionados a los puntos y regiones de interés obtenidos a partir tanto de BA Data como de OSM.

En primer lugar, presentaré un ranking con el top 10 por cantidad de tipo de establecimiento. Luego, presentaré mapas coropléticos con la cantidad de puntos y regiones en cada celda de la grilla. Así como también, mapas indicando la ubicación de distintos puntos y áreas relevantes para el análisis.

4.2.1 Puntos de Interés

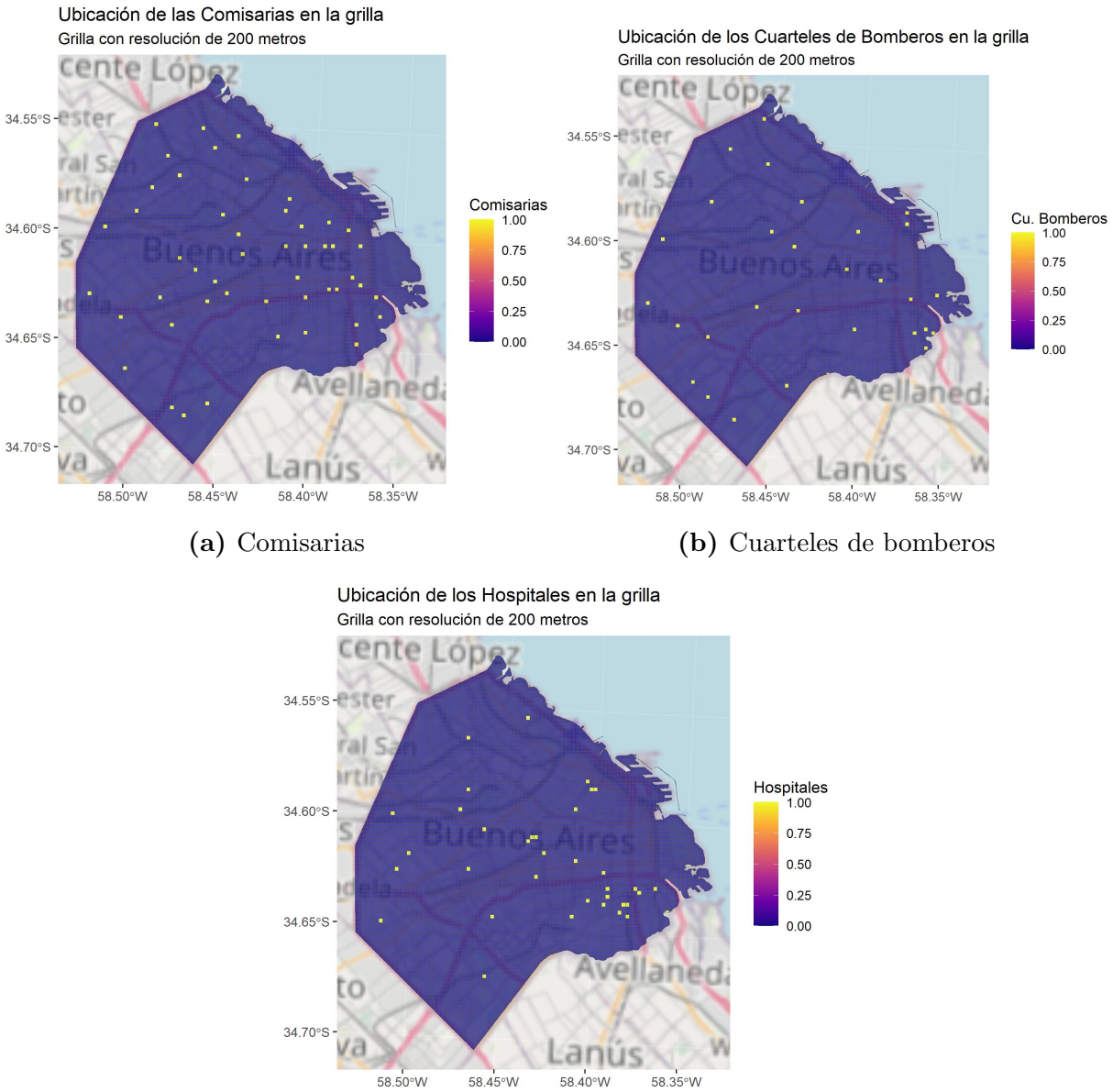
Tabla 3. Top 10 de puntos de interés BA Data

	Tipo	Cantidad
1	Garaje comercial	2.759
2	Restaurante	2.134
3	Establecimiento educativo	1.892
4	Farmacia	1.376
5	Cajero automatico	1.279
6	Organizacion social	721
7	Libreria	602
8	Cafe	389
9	Boca de sube	379
10	Parada de taxi	333

Tal como se muestra en la tabla, los garajes comerciales y restaurantes predominan en la grilla con más de 2000 puntos cada uno. Algo a tener en cuenta es que el valor de la columna cantidad no representa los cuadrados de la grilla que poseen el punto de interés, sino la suma, ya que en una misma celda puede haber más de un mismo POI de la misma clase. Con más de 1000 puntos se encuentran los establecimientos educativos, tanto públicos como privados, así como también, farmacias y cajeros automáticos. Continuando con el ranking se tienen las organizaciones sociales, librerías, cafés, bocas de subte y paradas de taxi.

En la siguiente figura se podrán observar 3 puntos de interés que considero relevantes para el análisis y que como máximo poseen una observación por celda.

Figura 13. POIs de BA Data en una grilla de $200m^2$



(a) Comisarias

(b) Cuarteles de bomberos

(c) Hospitales

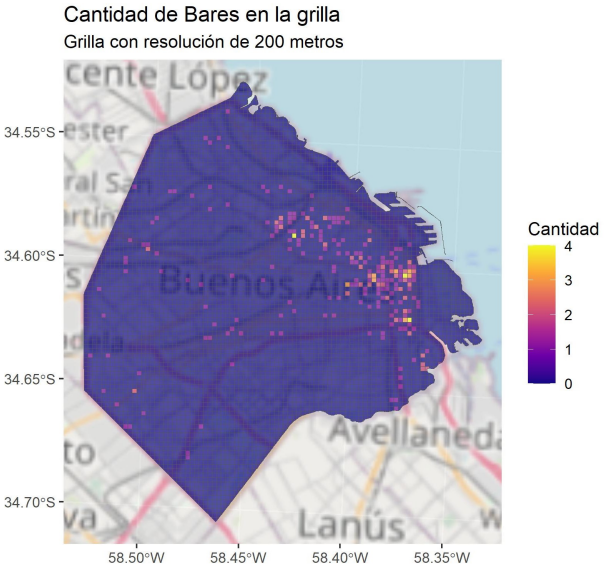
Particularmente, en el territorio de CABA se pueden encontrar 49 comisarias[6], 30 cuarteles de bomberos y un total de 36 hospitales [7]. En estos mapas, es posible observar que en los primeros dos casos los puntos se encuentran distribuidos equitativamente en el mapa. En el caso de los hospitales se puede observar una mayor presencia de los mismos en el sureste de la ciudad.

A continuación se exhibirán 5 POIs que también me parecen importantes para el análisis: bares, restaurantes, cajeros automáticos, bocas de subte y escuelas.

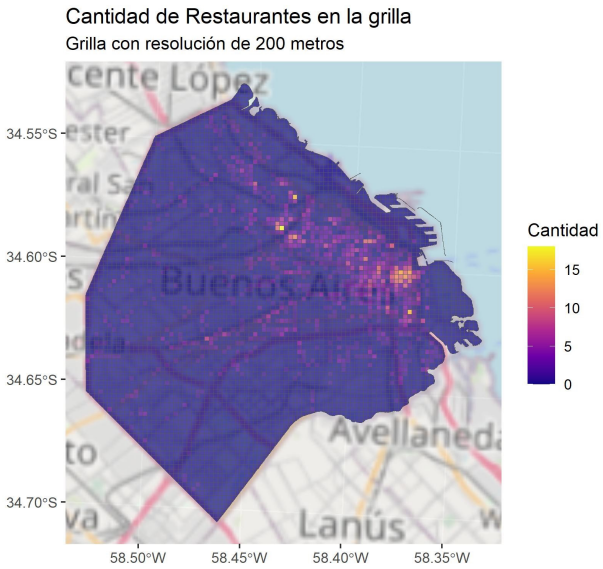
[6]Incluye tanto las comisarias vecinales como las comunales.

[7]Incluye solo los hospitales públicos.

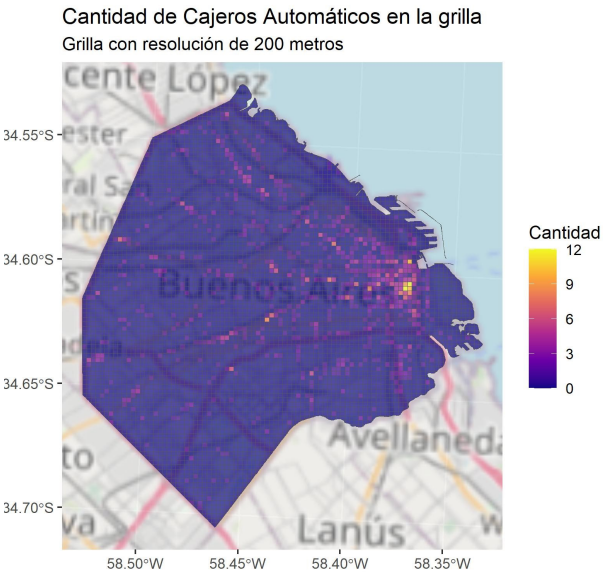
Figura 14. POIs de BA Data en una grilla de $200m^2$



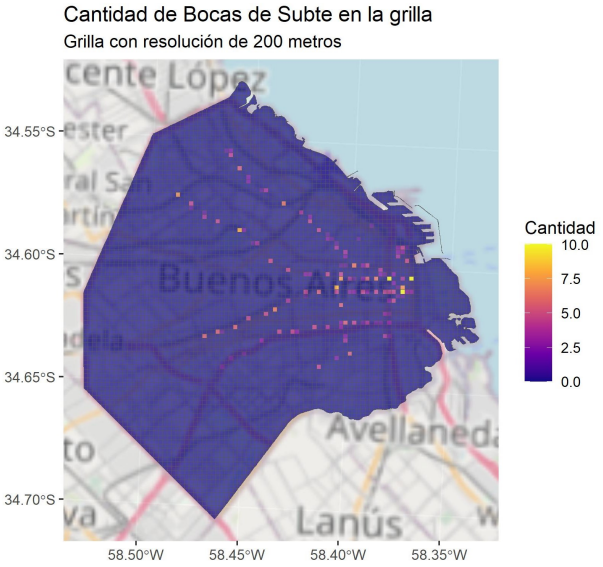
(a) Bares



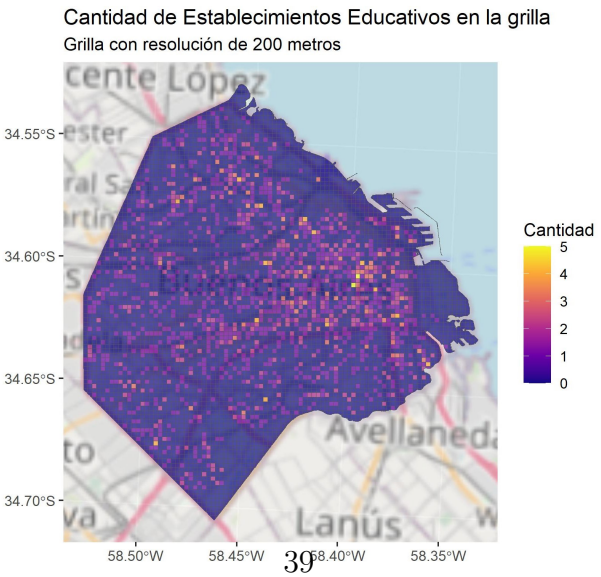
(b) Restaurantes



(c) Cajeros Automáticos



(d) Bocas de Subte



(e) Escuelas

Como puede observarse en los mapas, tanto los restaurantes, las bocas de subte como los cajeros automáticos están muy concentrados, llegando, en algunos casos, a valores mayores a 10 por celda. Algo también a destacar es que en el caso del mapa de las bocas de subte, es posible ver el recorrido de las distintas líneas con los puntos de interés distintos de 0. Por otro lado, a pesar de tener un máximo de 5 por cuadrado, es posible observar que los establecimientos educativos se encuentran distribuidos por toda la ciudad y que en gran parte de las celdas es posible encontrar al menos uno.

Adicionalmente, se realizó este mismo análisis para los POIs obtenidos a partir de OpenStreetMap. En este caso, solo se dejaron aquellos que poseían nombre, es decir, estaban identificados en la base de datos en la variable *name*. Con esto, se buscó minimizar la cantidad de observaciones que pueden estar repetidas. Dado que como se mencionó anteriormente, al ser un proyecto colaborativo puede haber puntos de interés duplicados.

A diferencia de los POIs de BA Data, ninguno en el top 10 tiene más de 1000 en cantidad. Encabezando el ranking se encuentran los locales de indumentaria, seguidos de locales de comidas rápidas, almacenes o supermercados pequeños, con más de 900 el primero y 600 los siguientes dos.

Tabla 4. Top 10 de puntos de interes OSM

	Tipo	Cantidad
1	Local de indumentaria	936
2	Local de comidas rápidas	672
3	Almacén/Supermercado pequeño	630
4	Super e Hipermercado	533
5	Peluquería	368
6	Arte (estatuas, murales)	255
7	Museo	104
8	Centros de deporte	62
9	Atracciones	49
10	Monumentos	36

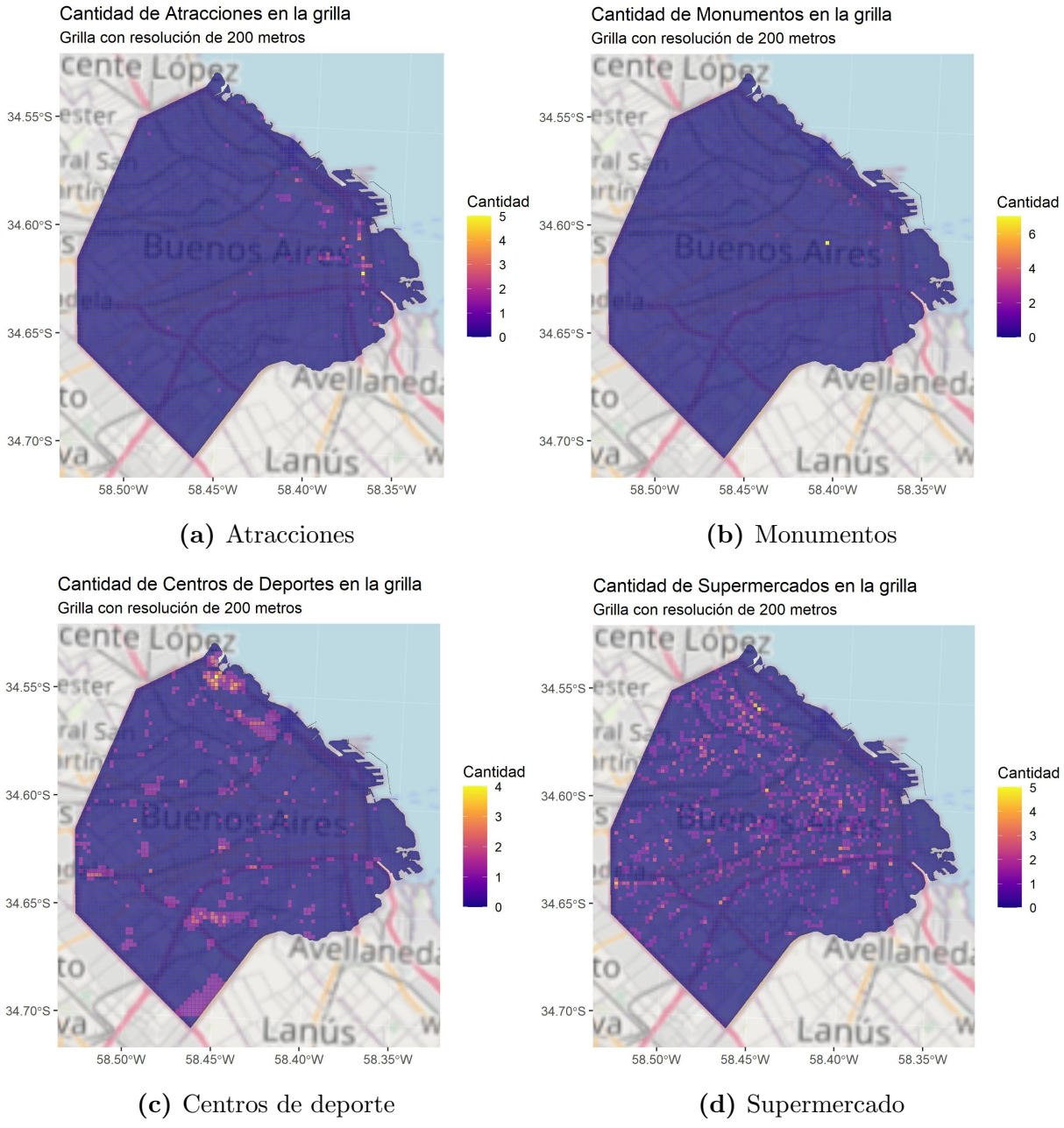
Algo a considerar es que el top 5 nos da una noción de dónde se encuentran las zonas comerciales y por lo tanto, zonas que probablemente tengan una mayor circulación de personas. Algo que puede interferir en la cantidad de robos y hurtos.

A continuación, se podrán observar 4 mapas de POIs del ranking. En todos los casos, se tiene un máximo mayor a 1.

Algo que es posible observar en el caso de los centros de deporte es que en la mayoría de los casos, por más que se encuentren como POIs, son áreas, ya que se pueden observar celdas contiguas con un 1 en cantidad. Por otro lado, en el caso de los supermercados es posible observar que se encuentran distribuidos por toda la ciudad, sin predominar particularmente en ninguna zona. A diferencia de estos, en el caso de las atracciones y

monumentos, es posible ver que se encuentran principalmente en el este. Algo que tiene sentido, ya que se encuentran las zonas más turísticas de CABA.

Figura 15. POIs de OSM en una grilla de $200m^2$



(a) Atracciones

(b) Monumentos

(c) Centros de deporte

(d) Supermercado

4.2.2 Regiones de Interés

Al igual que se analizaron los rankings y mapas coropléticos de los POIs, se analizaran los de las regiones de interés.

En primer lugar, el ranking de las dos regiones de interés que se utilizaron del repositorio público de la Ciudad Autónoma de Buenos Aires.

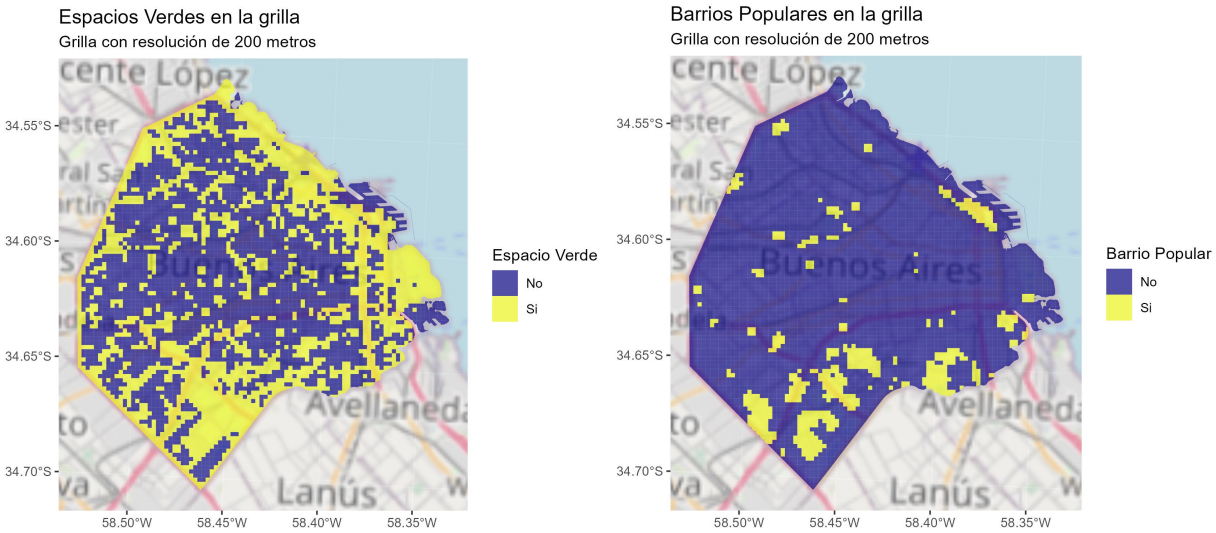
Tabla 5. Ranking de regiones de interés BA Data

	Tipo	Cantidad
1	Espacio verde	2.363
2	Barrio popular	533

Es posible observar que una gran proporción de las celdas poseen espacios verdes, siendo la misma un 44 % y en el caso de los barrios populares un 10 %.

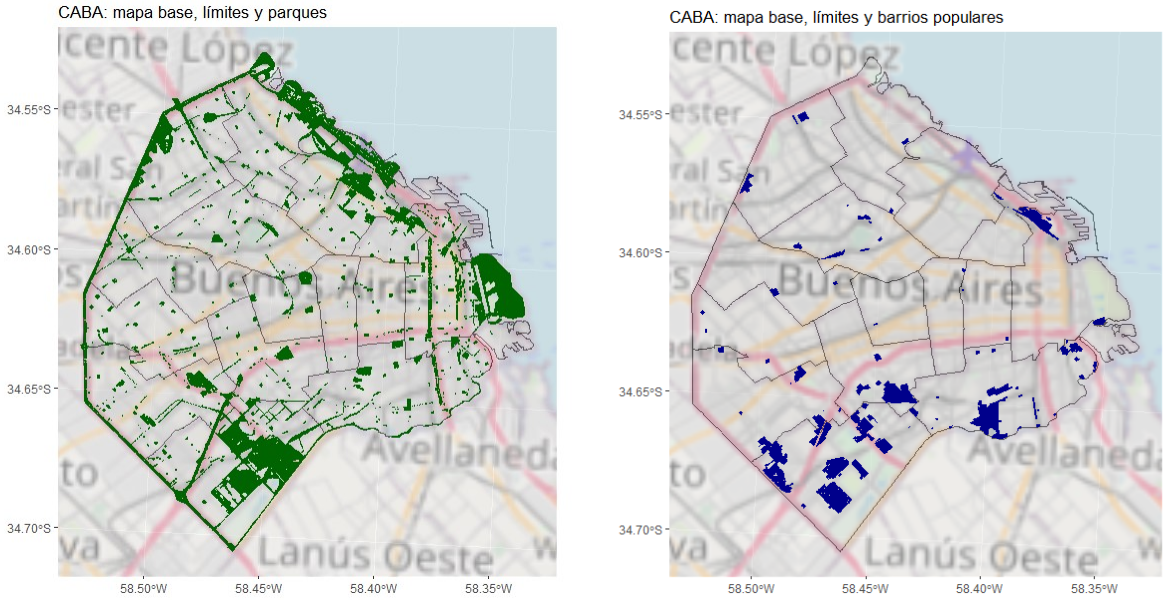
A continuación, se muestran los mapas coropléticos de ambos tipos de regiones. Así como también, dos mapas con dichas áreas resaltadas.

Figura 16. Regiones de Interés de BA Data en una grilla de $200m^2$



(a) Espacios verdes **(b)** Barrios populares

Figura 17. Regiones de Interés de BA Data



(a) Espacios verdes **(b)** Barrios populares

Ambas variables son binarias, es decir, indican si la celda de la grilla tiene o no tiene dicha región. Los barrios populares se localizan mayormente en el sur de la ciudad y en menor medida en el centro - este. En cambio, los espacios verdes, se encuentran distribuidos por toda la ciudad.

Al igual que en el caso de los POIs de OSM, solo se dejaron aquellos que poseían nombre en la base.

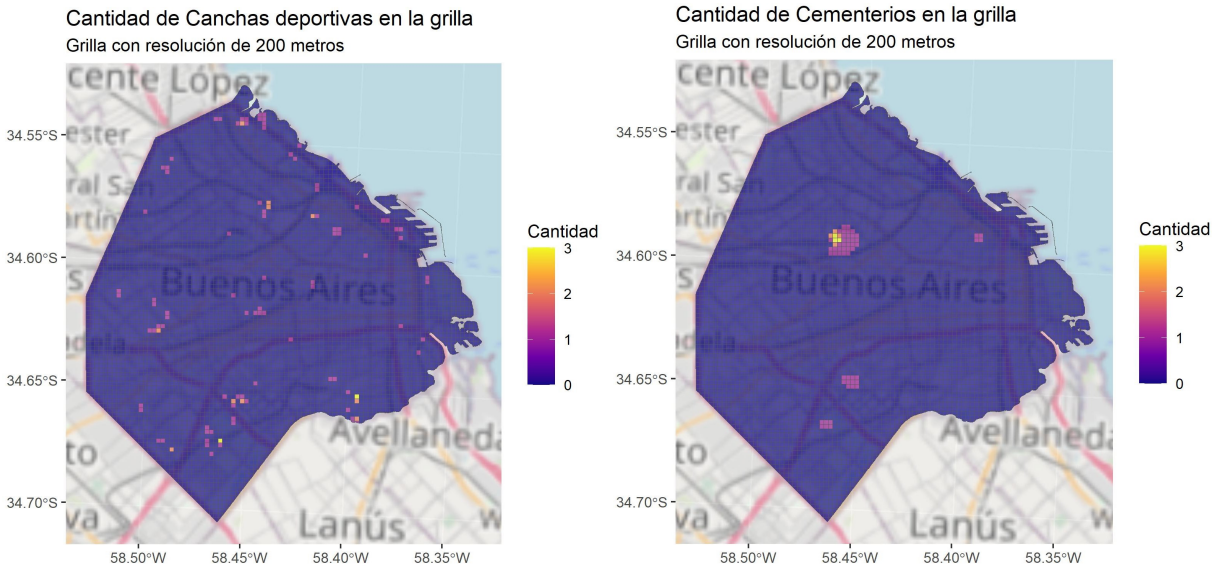
En el top se encuentran los centros deportivos con más de 500, seguidos de las canchas de deportes con más de 100 en cantidad. Luego, los cementerios con 69, los estadios con 59 y las atracciones con 54. Las siguientes 4 regiones del ranking tienen una cantidad menor a 50.

Tabla 6. Ranking de regiones de interes OSM

	Tipo	Cantidad
1	Centros deportivos	506
2	Cancha de deportes	115
3	Cementerio	69
4	Estadio	59
5	Atracciones	54
6	Campo de golf	39
7	Memorial	15
8	Zoológico	12
9	Cine	10

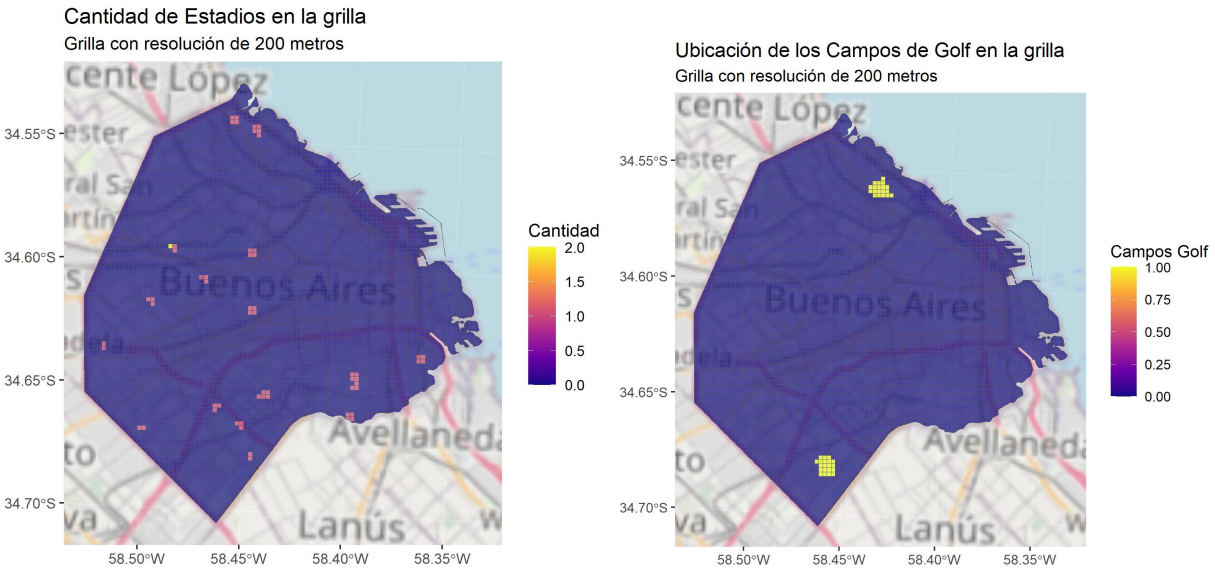
En los mapas coropléticos se puede observar que las canchas de deportes están distribuidas por toda la ciudad, al igual que los estadios. Aunque de estos últimos hay una mayor proporción principalmente en el suroeste. En el caso de los cementerios se pueden distinguir 4 y en el de los campos de golf 2.

Figura 18. POIs de OSM en una grilla de $200m^2$ (1)



(a) Canchas de deportes **(b)** Cementerios

Figura 19. POIs de OSM en una grilla de $200m^2$ (2)



(a) Estadios **(b)** Campos de golf

4.3 Análisis de datos socioeconómicos

En esta sección se analizarán los conjuntos de datos relacionados con las variables socioeconómicas: precio de los terrenos y departamentos en dólares por metro cuadrado y porcentaje de necesidades básicas insatisfechas (NBI). Al igual que en la sección anterior, se presentaran rankings y mapas coropléticos.

Algo a tener en cuenta a la hora de designar los barrios para cada cuadrado de la grilla es que en muchos casos una celda comparte barrio. Es decir, la misma contiene el límite entre dos o más barrios. Es por esto, que lo que se decidió en estos casos es quedarse con el barrio que le corresponde a la celda anterior. Por ejemplo, si la celda está en el límite entre Puerto Madero y San Telmo y el anterior cuadrado, que no se encontraba en el límite, pertenecía a San Telmo, se le asigna este barrio.

En relación con cómo se calculan los valores promedios para cada celda, en primer lugar, para los terrenos y departamentos a cada cuadrícula se le asigna el valor promedio de las observaciones que se encuentran en cada celda. En caso de que no haya ningún terreno o departamento, se le asigna el valor promedio de todas las observaciones del barrio al que pertenece dicha celda. En segundo lugar, para el porcentaje de necesidades básicas, dado que la información del mismo se encuentra por radio censal, se le asigna un valor por celda según el radio censal y en caso de que una celda contenga dos o más radios se calcula el valor promedio.

4.3.1 Precio de los terrenos en dólares por metro cuadrado

A continuación se mostrarán el top 10 de barrios más caros y baratos, según el precio promedio en dólares por metro cuadrado de los terrenos en venta en el 2020.

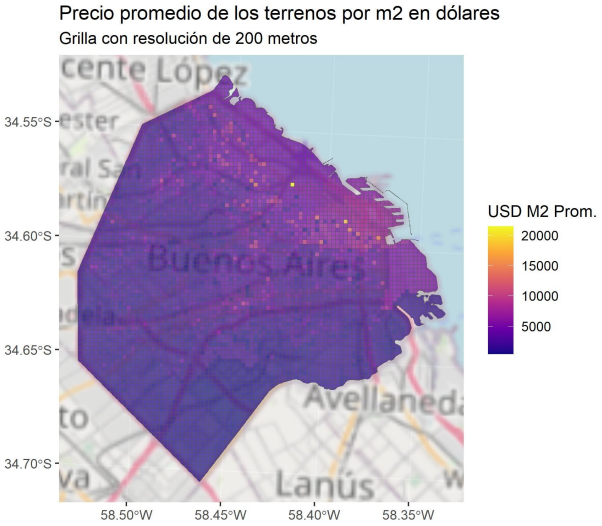
Tabla 7. Top 10 Barrios más caros en terrenos - BA Data

	Barrios	Precio promedio(USD x m^2)
1	Retiro	6449,08
2	Recoleta	5302,96
3	Puerto Madero	5085,54
4	Belgrano	4359,25
5	Palermo	4159,94
6	Nuñez	3877,99
7	San Nicolás	3658,70
8	Colegiales	3169,44
9	Caballito	3052,36
10	Villa Crespo	2829,05

Tabla 8. Top 10 Barrios más baratos en terrenos
BA Data

	Barrios	Precio promedio(USD x m^2)
1	Villa Soldati	597,90
2	Villa Riachuelo	677,93
3	Villa Lugano	902,25
4	Mataderos	1147,47
5	Villa Real	1172,03
6	Versalles	1178,46
7	Boca	1349,02
8	Nueva Pompeya	1399,18
9	Barracas	1433,35
10	Liniers	1458,05

Figura 20. Precio promedio de los terrenos por m^2 en dólares



Tanto a partir de los rankings como del mapa coroplético, es posible notar que los barrios con precios más altos en los terrenos se encuentran en el noreste de la Ciudad Autónoma de Buenos Aires.[8] Particularmente, los terrenos más caros se encuentran en el barrio de Retiro con un promedio de casi 6.500 dólares por metro cuadrado, seguido de Recoleta y Puerto Madero con un valor promedio mayor a 5.000.

Por el contrario, los terrenos más baratos en promedio se encuentran en el sur. Los mismos se ubican en Villa Soldati, Villa Riachuelo y Villa Lugano con un precio promedio en dólares por metro cuadrado menor a mil.

[8]Para ver el mapa coroplético con las delimitaciones y nombres de los barrios ir al Apéndice C

4.3.2 Precio de los departamentos en dólares por metro cuadrado

A continuación se mostrará el top 10 de barrios más caros y baratos, según el precio promedio en dólares por metro cuadrado de los departamentos en venta en el 2020.

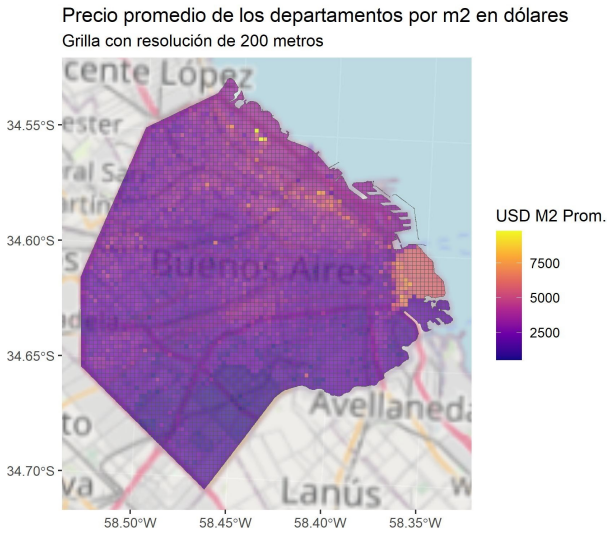
Tabla 9. Top 10 Barrios más caros en departamentos - BA Data

	Barrios	Precio promedio(USD xm^2)
1	Puerto Madero	6323,31
2	Palermo	3625,10
3	Belgrano	3541,08
4	Nuñez	3445,13
5	Recoleta	3300,81
6	Retiro	3251,98
7	Colegiales	3180,00
8	Villa Urquiza	3091,47
9	Coghlan	2950,62
10	Villa Ortuzar	2896,06

Tabla 10. Top 10 Barrios más baratos en departamentos - BA Data

	Barrios	Precio promedio(USD xm^2)
1	Villa Soldati	1097,04
2	Villa Lugano	1506,71
3	Villa Riachuelo	1856,04
4	Nueva Pompeya	1867,75
5	Constitución	1982,75
6	Parque Avellaneda	2041,91
7	Floresta	2071,60
8	Parque Patricios	2076,20
9	Boca	2078,32
10	San Cristobal	2124,76

Figura 21. Precio promedio de los departamentos por m^2 en dólares



A diferencia del mapa coroplético de los terrenos, en el de departamentos no son tan claras las zonas que poseen un precio promedio más alto. Sin embargo, sí es posible notar que el sur posee los precios más bajos y que en el noreste, al igual que con los terrenos, se encuentran las celdas de la grilla con precios más altos. [9]

Con respecto al top 10 de barrios más caros de terrenos, el de departamentos comparte con este 7 de los 10. En relación al ranking de barrios más baratos, comparten solo la mitad del mismo. Sin embargo, en el top 3 se encuentran los mismos barrios aunque invertidos en el orden. En el caso del precio promedio en dólares de departamentos el orden es Villa Soldati, Villa Lugano y Villa Riachuelo. En cambio, para los terrenos, los últimos dos están al revés.

4.3.3 Porcentaje de Necesidad Básicas Insatisfechas

A continuación se mostrará el top 10 de barrios con mayor y menor porcentaje de necesidades básicas insatisfechas según la información censal por radio del Censo de 2010.

[9] Para ver el mapa coroplético con las delimitaciones y nombres de los barrios ir al Apéndice C

Tabla 11. Top 10 Barrios con mayor porcentaje de NBI - BA Data

	Barrios	Porcentaje(%)
1	Boca	34,34
2	Villa Soldati	26,27
3	Constitución	25,31
4	Monserrat	18,58
5	Retiro	16,73
6	Puerto Madero	14,97
7	Barracas	14,27
8	San Cristobal	13,79
9	Paternal	12,54
10	Balvanera	12,16

Tabla 12. Top 10 Barrios con menor porcentaje de NBI - BA Data

	Barrios	Porcentaje(%)
1	Versalles	1,06
2	Villa del Parque	1,08
3	Villa Real	1,27
4	Villa Devoto	1,32
5	Belgrano	1,38
6	Parque Chas	1,39
7	Coghlan	1,46
8	Saavedra	1,47
9	Agronomía	1,73
10	Villa Urquiza	1,85

En el caso de las Necesidades Básicas Insatisfechas, cuanto más alto es el porcentaje indica un menor nivel socioeconómico. Algo a destacar es que en el top 10 de barrios con menor porcentaje todos poseen un valor menor al 2 %. El top 3 está compuesto por Versalles y Villa del Parque con menos de un 1,1 % y Villa Real con menos de un 1,3 %.

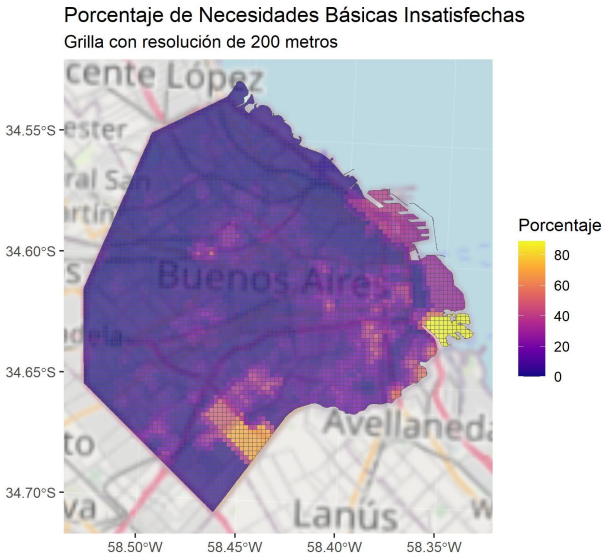
En el caso del top con mayor porcentaje, todos tienen más de un 12 %. Particularmente, La Boca posee más de un 34 %, seguido de Villa Soldati y Constitución con más de un 25 %. Siendo el barrio de La Boca un 3.000 % más alto que el de Versalles.

En los mapas coropléticos, se puede observar que las celdas de la grilla del norte poseen el menor porcentaje y las del sureste los mayores porcentajes.[10] Esto difiere un poco del patrón que se observaba en el precio promedio en dólares por metro cuadrado de los terrenos, principalmente, y de los departamentos. Incluso algunos barrios se encuentran en el top 10 contrario. Por ejemplo, Retiro y Puerto Madero se encuentran en el top 10 de barrios más caros tanto para terrenos como para departamentos, pero se encuentran en el

[10]Para ver el mapa coroplético con las delimitaciones y nombres de los barrios ir al Apéndice C

top 10 de barrios con mayor porcentaje de necesidades básicas insatisfechas. Sin embargo, sí comparten barrios como La Boca y Villa Soldati en el top 10 de barrios más baratos y a su vez con mayor porcentaje de NBI y Belgrano en el top 10 de barrios más caros y con menor porcentaje.

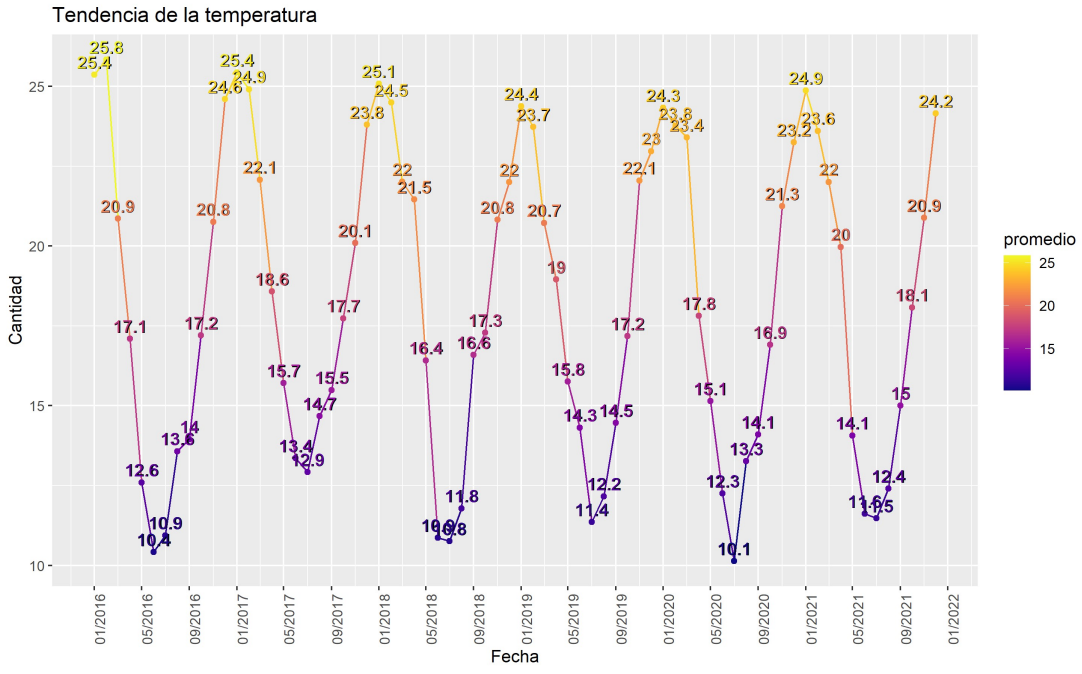
Figura 22. Porcentaje de Necesidades Básicas Insatisfechas



4.4 Análisis de variables climáticas

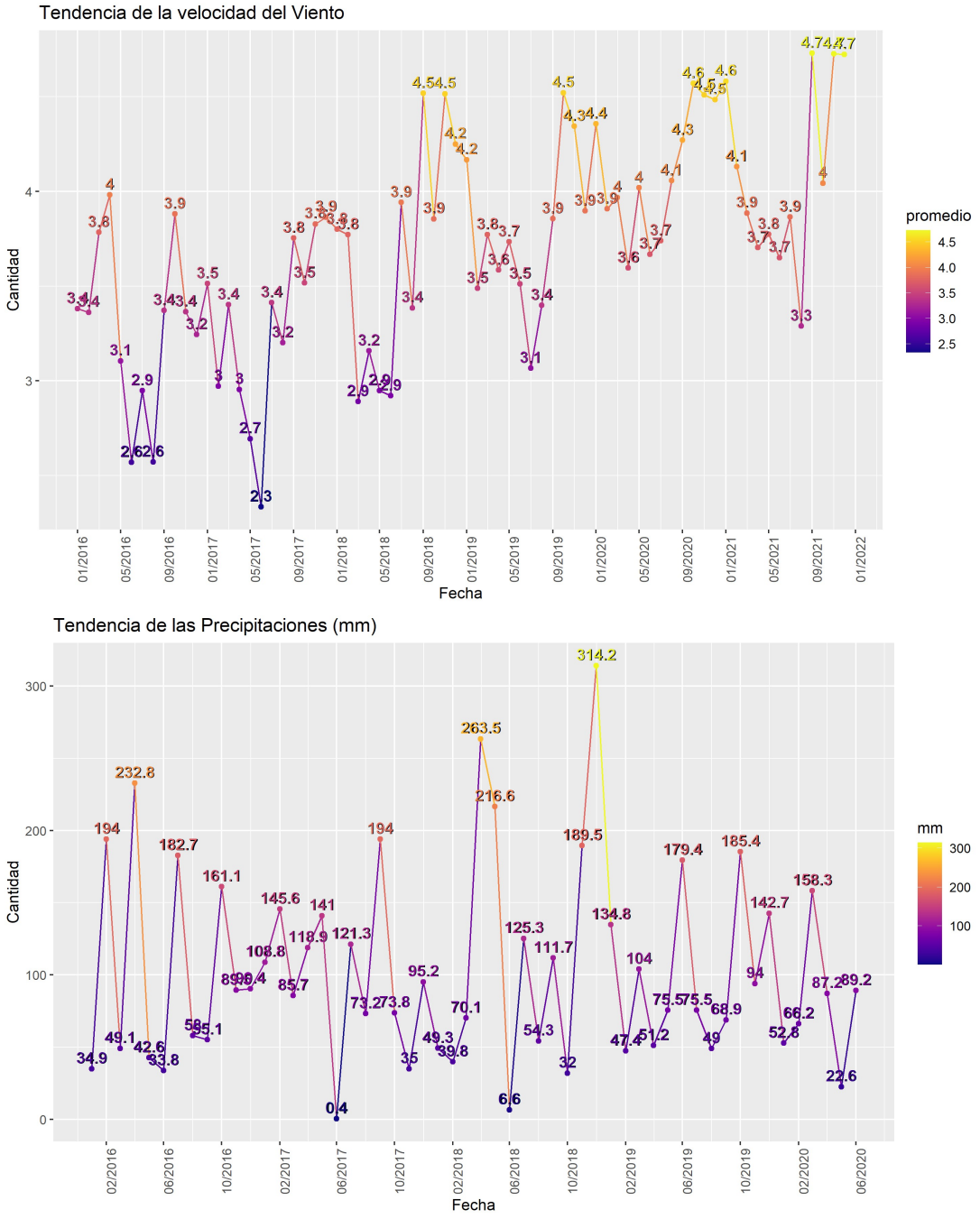
En esta sección, se analizarán las variables climáticas: temperatura, precipitaciones y velocidad del viento. Para esto, se calculó el promedio de cada una de las variables para cada mes y año en el análisis.

Figura 23. Tendencia de la temperatura



Al ver el gráfico de la temperatura vemos, como era de esperarse, que tiene variaciones bastante cíclicas. Las temperaturas máximas se encuentran en los meses de verano y las mínimas en los de invierno. El valor promedio más alto se da en 2016 con 25,8 grados y el mínimo en 2020 con 10,1^o.

Figura 24. Tendencia de la velocidad del viento y precipitaciones

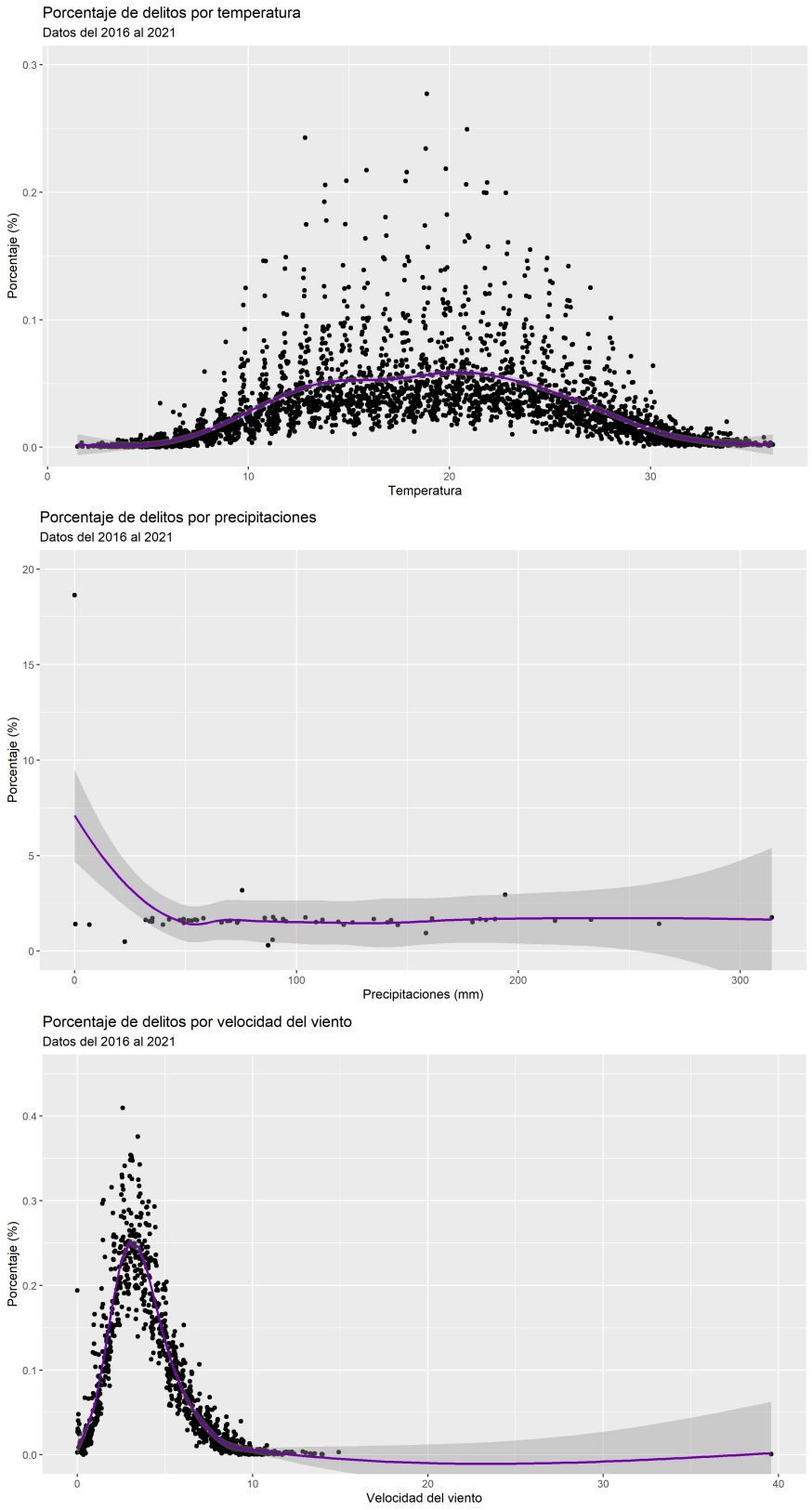


Para las variables de lluvia y velocidad del viento a priori no hay un análisis que se pueda realizar. Sin embargo, en el caso de la velocidad del viento se puede observar que la tendencia es creciente a lo largo de los años.

No obstante, para poder ver si las variables climáticas son relevantes, deben analizarse

en conjunto con la cantidad de delitos. Es por esto que se decidió realizar un gráfico para cada una de las variables en donde se calcula el porcentaje de delitos para cada valor promedio.

Figura 25. Porcentaje de delitos según las variables climáticas



Al observar estas imágenes, es posible notar que para las temperaturas bajas y altas la cantidad de delitos disminuye. Llegando al máximo en casi 19^o. Al pasar al gráfico de precipitaciones, sin lugar a dudas, el máximo en cantidad de crímenes se encuentra cuando las mismas son 0, es decir, no llueve. También a medida que aumentan las precipitaciones va disminuyendo el porcentaje de delitos. Algo a destacar, es que casi un 19 % de los crímenes se dan cuando no hay lluvias. Por último, cuando la velocidad del viento se encuentra entre 0 y 10 km/h se da el porcentaje máximo. Disminuyendo considerablemente a medida que aumentan los kilómetros por hora.

A partir de esto es posible pensar que el clima influye en la frecuencia de los delitos. Por lo tanto, estas podrían ser variables relevantes para explicarlos.

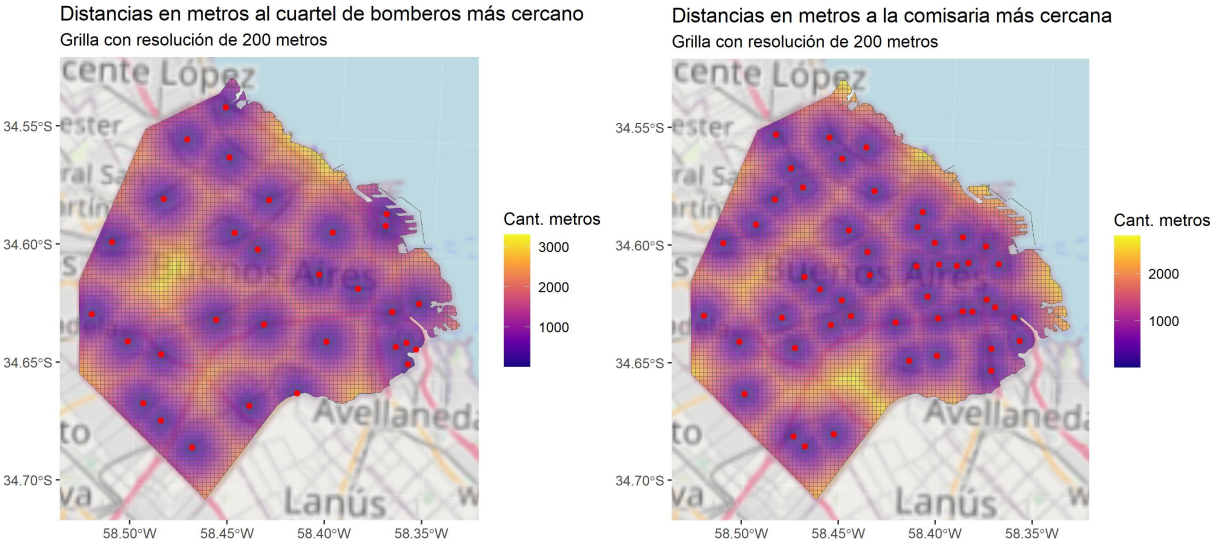
4.5 Análisis de variables de distancia

En esta sección se analizarán las variables relacionadas con la distancia a ciertos puntos y regiones de interés y tipos de calles.

4.5.1 Puntos de Interés

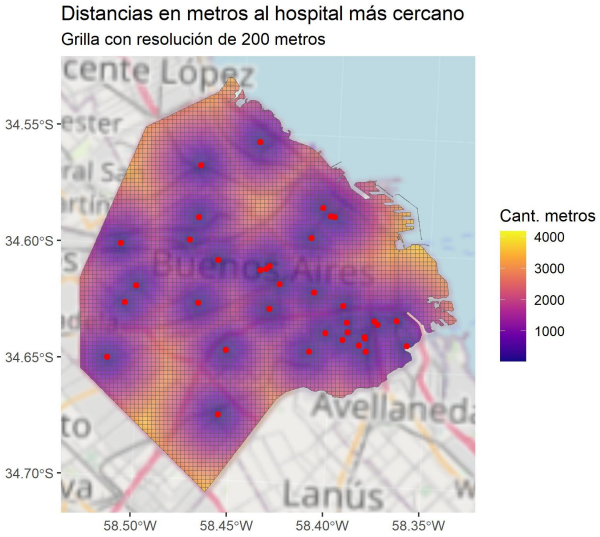
En primer lugar, se puede observar un mapa coroplético que indica la distancia Euclídea entre cada punto de interés más cercano y el centroide de cada una de las celdas. En este caso, se graficaron los tres POIs sobre los que se calcularon las variables de distancias: comisarias, cuarteles de bomberos y hospitales.

Figura 26. Distancias a POIs de BA Data en una grilla de $200m^2$



(a) Cuarteles de Bomberos

(b) Comisarias



(c) Hospitales

En el caso de las comisarias, la distancia máxima entre alguna celda y una delegación policial es de 2.800 metros. Mientras que para los cuarteles de bomberos es de 3.300 metros y para los hospitales de 4.200 metros. Esto es algo que se puede observar en los mapas, ya que los puntos colorados indican cada uno de los POIs.

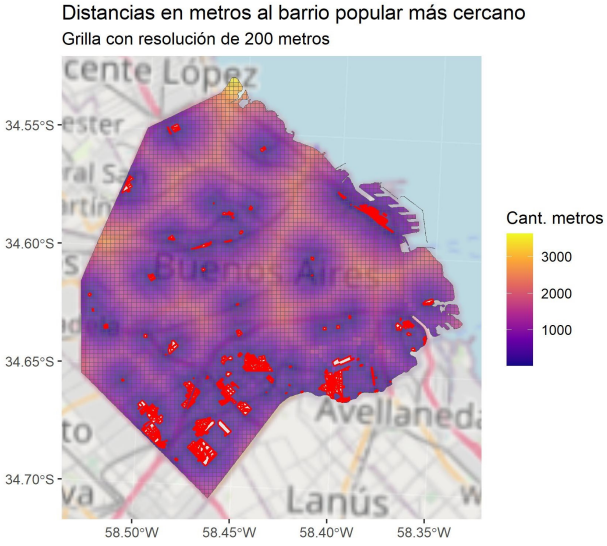
Como se menciona en el análisis de los puntos de interés, al tener en el territorio de CABA 49 comisarias, 30 cuarteles de bomberos y un total de 36 hospitales, es lógico que la menor distancia máxima entre cualquier celda y un POI sea una comisaria. Sin embargo, a pesar de haber más hospitales que cuarteles de bomberos, al estar estos últimos mejor distribuidos en la superficie de la ciudad, tienen una distancia máxima menor que los hospitales.

4.5.2 Región de Interés

Al igual que con los POIs a continuación vemos un mapa coroplético con los barrios populares en colorado y la distancia Euclídea entre los centroides de cada uno y el de cada una de las celdas de la grilla.

Es claro que los barrios populares predominan principalmente en el sur.

Figura 27. Distancia a Regiones de Interés de BA Data en una grilla de $200m^2$

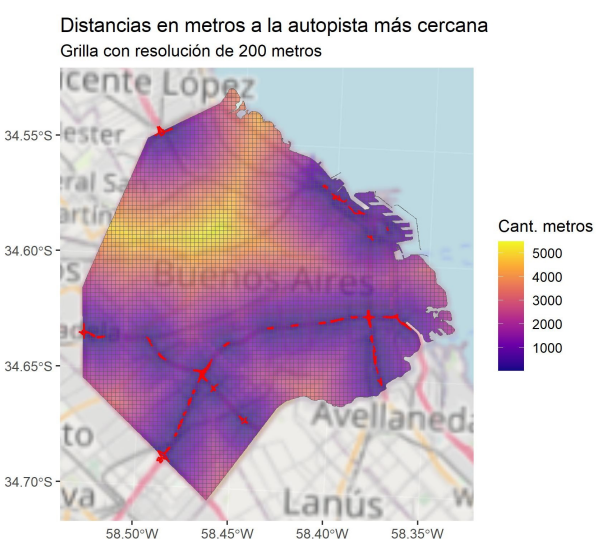


4.5.3 Tipos de calles

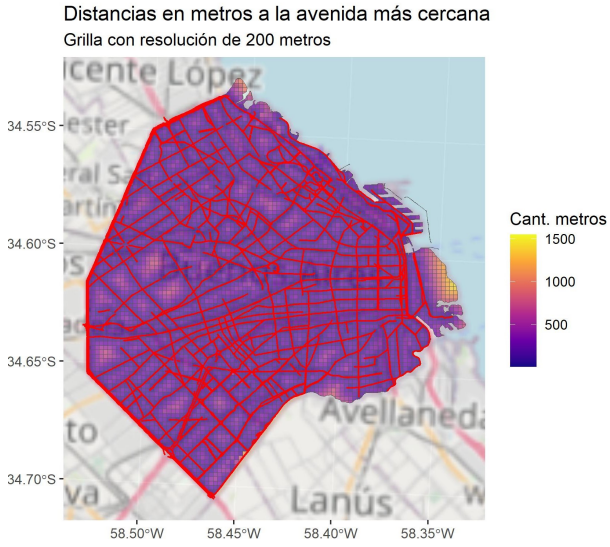
En relación con las distancias a los distintos tipos de calles, mostraremos tres mapas coropléticos, uno para cada tipo: autopistas, avenidas y túneles.

En este caso, la distancia máxima más baja la tienen las avenidas con un valor de 1.500 metros. Algo que resulta lógico, ya que se encuentran por toda la ciudad. En el otro extremo, tenemos a los túneles con una distancia máxima de 10.600 metros. Localizándose la gran mayoría en el norte. Las autopistas, por otro lado, tienen un máximo de 5.500 metros.

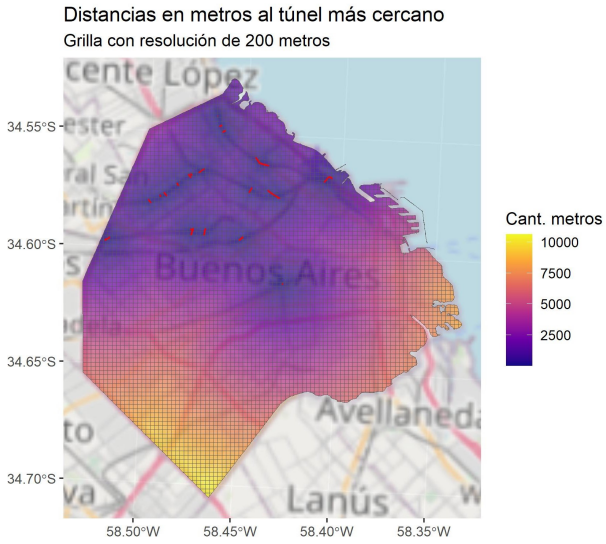
Figura 28. Distancias a tipos de calles de BA Data en una grilla de $200m^2$



(a) Autopistas



(b) Avenidas



(c) Túneles

4.6 Análisis multivariado

El objetivo de esta sección es hacer un análisis en conjunto de todos los *datasets* utilizados en este trabajo. En particular, se analizará la relación de los delitos con el resto de las variables. Así como también, la relación entre ellas.

Se realizará una distinción entre variables espaciales y temporales, dado que para las primeras se incluirá solo el último período para todas las celdas de la grilla. Mientras que para las segundas se incluirá la información de todos los períodos del análisis. Esto se debe a que en el caso de las espaciales, al tener información estática, las mismas no varían a lo largo del tiempo, pero sí entre celdas. Por el contrario, las temporales varían en el tiempo, pero no entre las cuadrículas. La única variable que cambia tanto temporal como

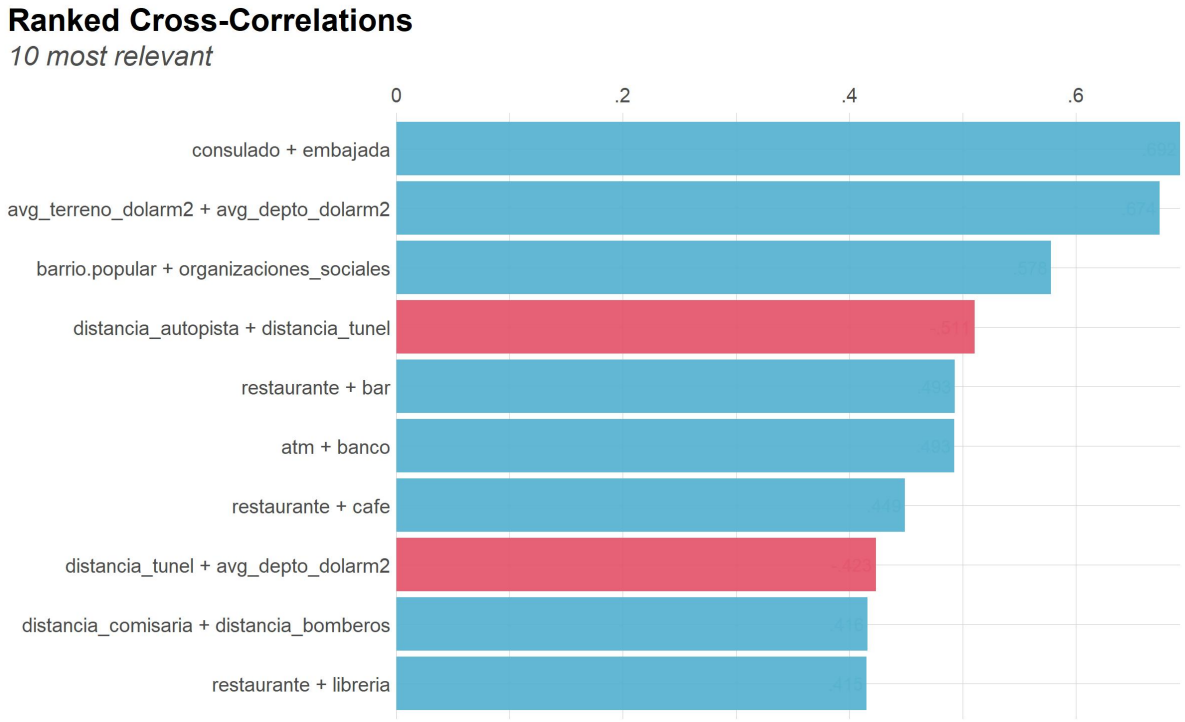
espacialmente es la de delitos.

4.6.1 Espacial

En esta sección se incluirán las variables que varían entre celdas de la grilla: delitos, barrios, comunas, POIs, ROIs, distancias y socioeconómicas. En el caso de los barrios y comunas, se decidió incorporarlas como variables *dummy*, indicadoras de las categorías.

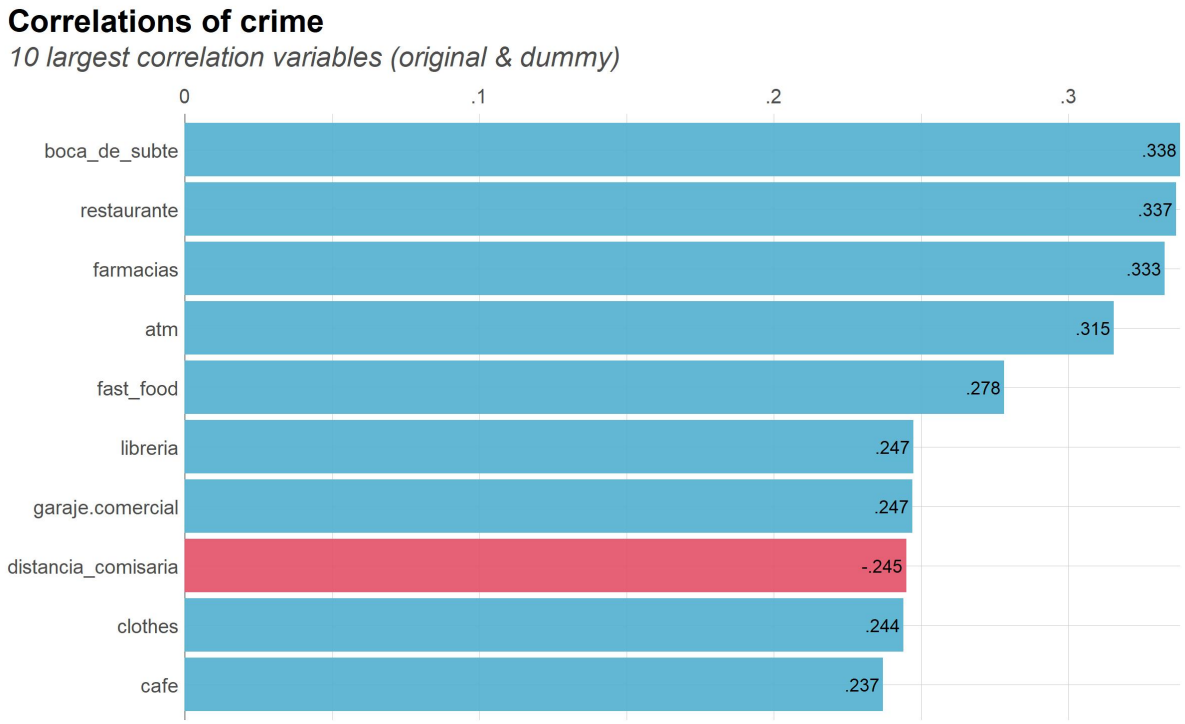
En el siguiente gráfico se puede observar el top 10 de variables con mayor correlación.

Figura 29. Top 10 Correlaciones Cruzadas - Espaciales



En primer lugar, tenemos las variables de consulado y embajada. Luego, las variables de precio promedio en dólares por metro cuadrado de terrenos y departamentos y en tercer lugar, las de barrio popular y organizaciones sociales. Algo a destacar es que las variables de distancia autopista y distancia a túnel correlacionan de forma negativa, al igual que las de distancia túnel y el precio promedio en dólares por metro cuadrado de los departamentos. Sin embargo, la variable de crímenes del mes actual, la variable objetivo, no es parte de este ranking. Por lo tanto, en el siguiente gráfico se tomó el top 10 de las variables que más correlacionan con nuestra variable de interés.

Figura 30. Top 10 Correlaciones con Delitos - Espaciales



Todas las variables correlacionan de forma positiva, a excepción de la distancia a comisarias que lo hace negativamente. Algo que a priori, parecería contradictorio. Después, en el top 3 se encuentran las bocas de subte, los restaurantes y las farmacias. Lo que podría darnos a entender que los lugares con mayor concentración de personas, son los que más cantidad de delitos denunciados tienen.

Por último, a pesar de no encontrarse entre las primeras 10, se decidió analizar las variables socioeconómicas de forma individual en relación a los delitos.

Figura 31. Correlación de delitos y las variables socioeconómicas (1)

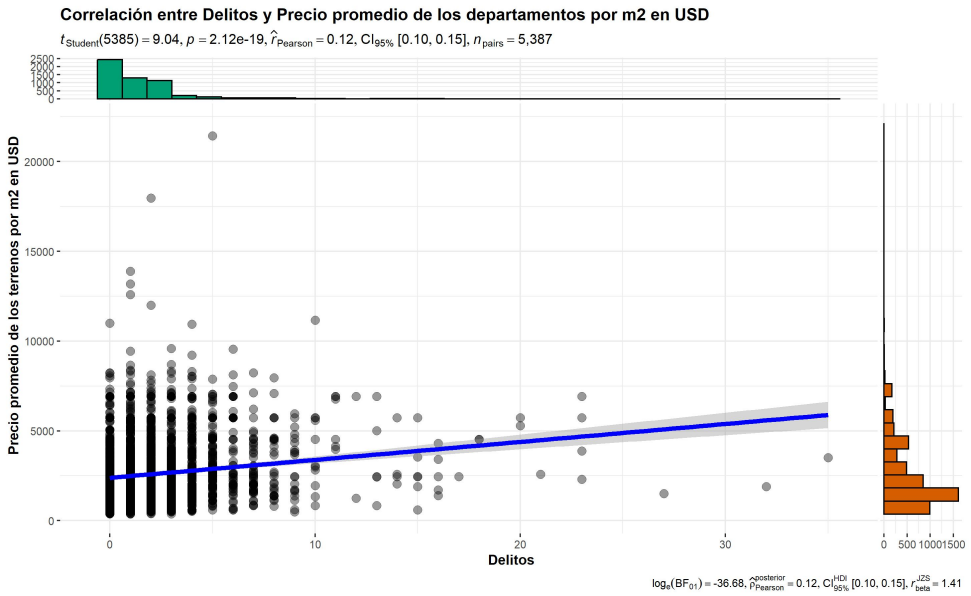
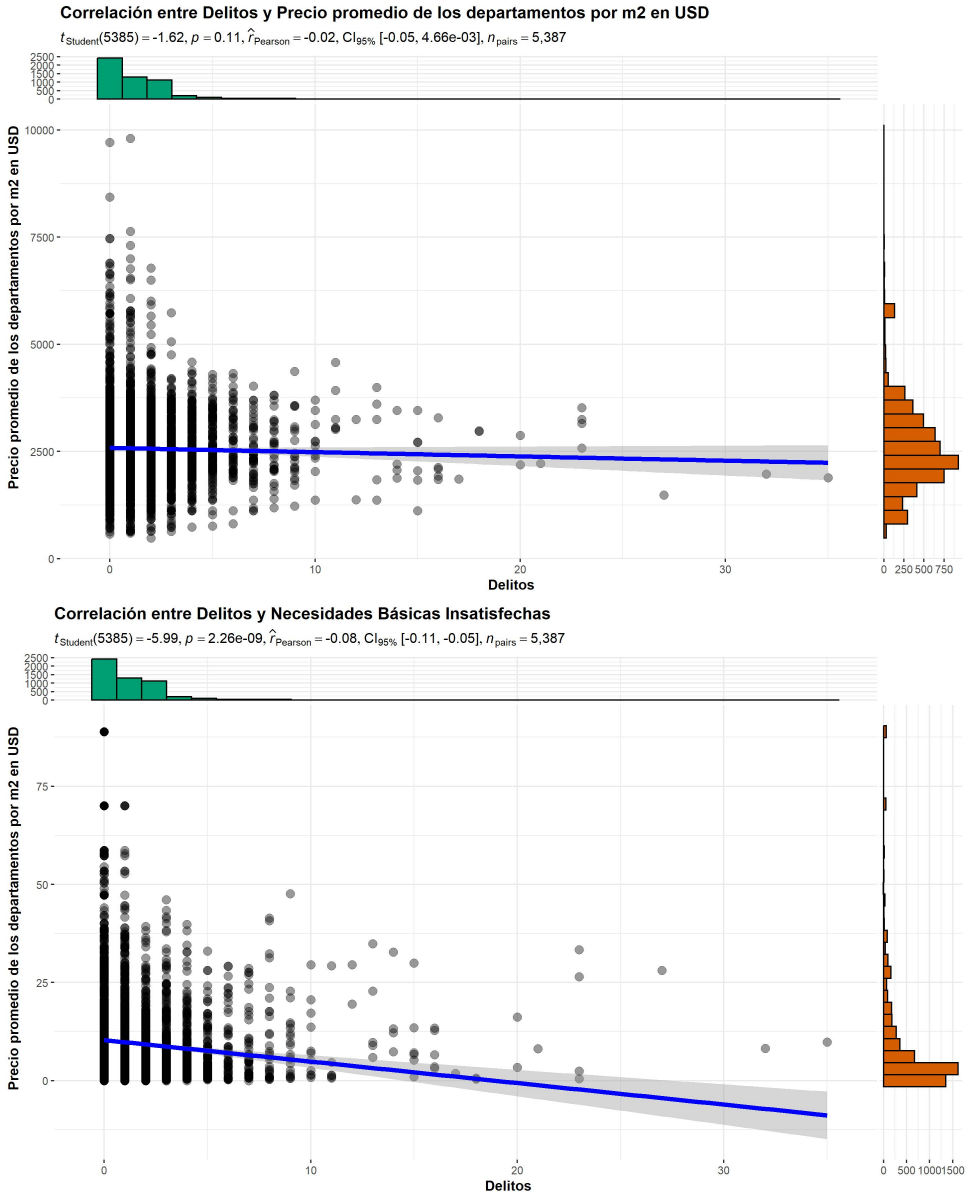


Figura 32. Correlación de delitos y las variables socioeconómicas (2)



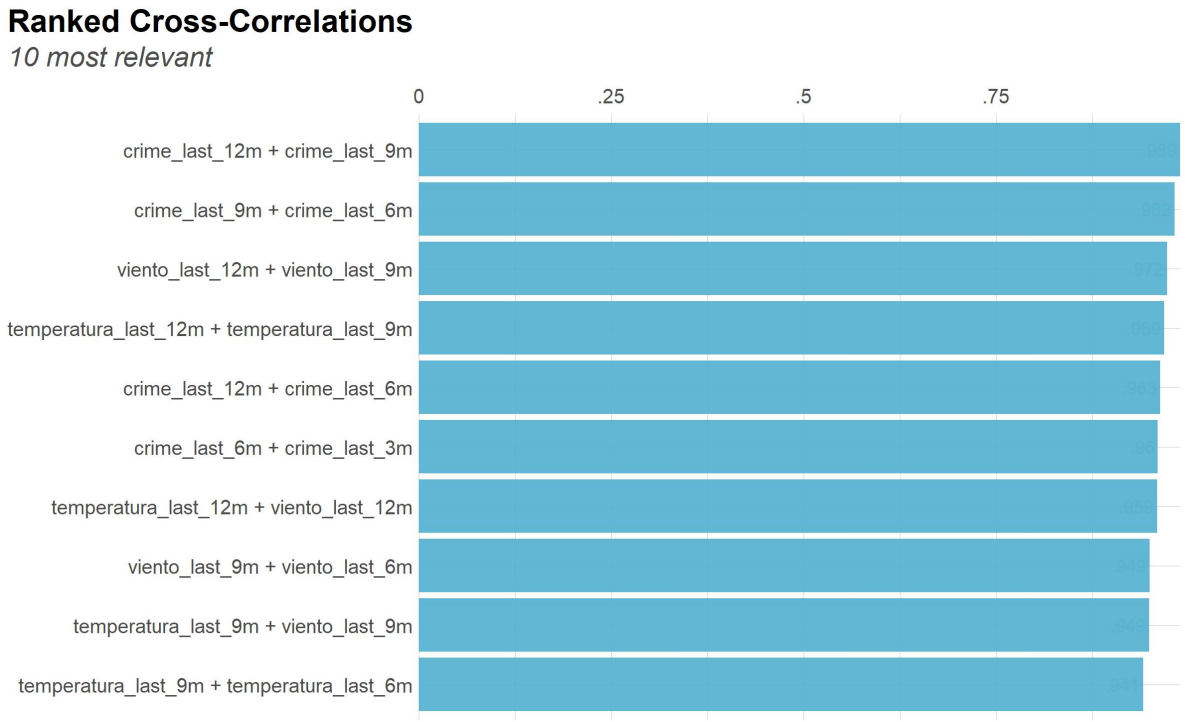
Lo primero para mencionar es que las dos variables de precio promedio del metro cuadrado en dólares tienen una correlación positiva y la del porcentaje de NBI negativa. Esto implica que en las celdas de la grilla en donde los precios son más elevados, la cantidad de delitos es más alta también. En el caso del NBI es al revés, cuando el porcentaje es más bajo, hay una mayor cantidad de delitos. No obstante, todas las variables reflejan la misma situación: frente a un mayor nivel socioeconómico, hay una mayor cantidad de delitos. Sin embargo, cabe destacar que la variable de departamentos no tiene una correlación estadísticamente significativa, mientras que las otras dos sí. De todas formas, no es una correlación muy fuerte dado que los coeficientes son inferiores a 0.15 en valor absoluto.

4.6.2 Temporal

En esta sección se incluirán aquellas variables que varían a lo largo del tiempo: delitos, temperatura, precipitaciones y viento. Todas en su versión del mes actual y sus valores pasados.

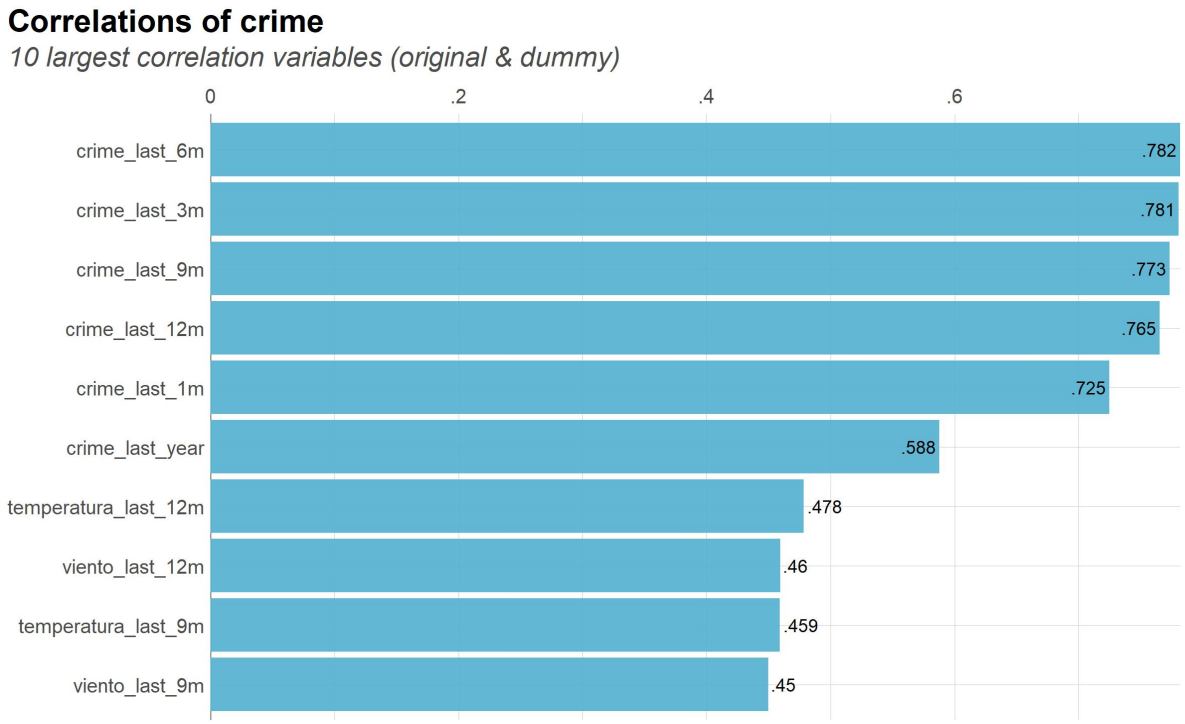
Al igual que con las variables espaciales, el primer ranking indica las 10 variables que más correlacionan.

Figura 33. Top 10 Correlaciones Cruzadas - Temporales



Encabezando el top 3 se encuentran las variables de delitos de hace 12 y 9 meses, seguidas de las de 9 y 6 y las de viento de hace 12 y 9 meses. Nuevamente, la variable de crimen no se encuentra en el ranking, por lo que se decidió hacer un top 10 de las que más correlacionan con nuestra variable objetivo.

Figura 34. Top 10 Correlaciones con Delitos - Temporales



En primer lugar, cabe mencionar que todas son positivas. Por otro lado, podemos observar que las primeras 6 son todas variables de crímenes pasados. Lo cual, tiene sentido, ya que es posible pensar que el mayor impacto en los delitos actuales son los que ya ocurrieron en esa misma celda de la grilla en el pasado. Luego, está la temperatura y el viento de hace 12 meses, seguido de las mismas variables de hace 9. Algo que en los gráficos de variables climáticas ya se había podido notar, que la temperatura y el viento tienen mayor incidencia que las precipitaciones.

Para ver el total de correlaciones, tanto de variables temporales como espaciales ir al Apéndice C.

5 Resultados

5.1 Modelos

Tal cual pudo observarse en la sección de Análisis Exploratorio, como consecuencia de la cuarentena (2020 y 21) la cantidad de registros de robos y hurtos disminuyó considerablemente en relación a los años anteriores (2016, 2017, 2018 y 2019). Por este motivo, se decidió probar tres modelos distintos:

- Modelo 1: utilizando todos los años del análisis para entrenar y predecir el mes de diciembre de 2021.
- Modelo 2: utilizando todos los años del análisis hasta noviembre de 2019 para entrenar y predecir el mes de diciembre de ese mismo año.
- Modelo 3: utilizando todos los años del análisis excepto 2020 y 2021 (hasta noviembre 2019) para entrenar y predecir el mes de diciembre de 2021.

De esta manera se busca comparar la *performance* de los modelos en la grilla y ver si las anomalías en las series, por la disminución en los casos, tienen realmente un impacto en las predicciones.

5.2 Experimentación con Modelos

Con el objetivo de predecir la cantidad de delitos que se darán el mes siguiente, se ha experimentado con diferentes conjuntos de variables y los algoritmos de aprendizaje supervisado mencionados en la sección de Aprendizaje automático. Para estos experimentos se han utilizado hiperparámetros por *default* para ambos algoritmos. Los mismos son:

- *XGBoost*: {*nrounds*: 100, *max_depth*: 6, *eta*: 0.3, *gamma*: 0, *colsample_bytree*: 1, *subsample*: 1, *min_child_weight*: 1}.
- *Random Forest*: {*mtry*: $\sqrt{variables}$, *min.node.size*: 250, *splitrule*: variance, *num.trees*: 100}.

Los resultados de estos experimentos para el Modelo 1 se presentan en la tabla Experimentos Modelo 1. Teniendo en cuenta que los Modelos 2 y 3 utilizan los mismos datos para entrenar y validar, los resultados de estos experimentos para ambos se presentan en la tabla Experimentos Modelos 2 y 3. [11]

[11] La forma de presentar los resultados de Experimentación con Modelos se basa en la propuesta de presentación establecida en [Dalla Via Monti, 2020].

Tabla 13. Experimentos Modelo 1

# de Grupo	Descripción del Modelo	Algoritmo	Cantidad de Variables	RMSLE Entrenamiento	RMSLE Validación	Overfitting
1	Delitos	Naïve Model	1	-	0.5978	-
1	Delitos	XGBoost	6	0.4705	0.4565	-3.0 %.
1	Delitos	Random Forest	6	0.4686	0.4568	-2.5 %.
2	Delitos + Fechas	XGBoost	8	0.4556	0.4864	6.8 %
2	Delitos + Fechas	Random Forest	8	0.4543	0.4620	1.7 %
3	Var. Previas + Territoriales	XGBoost	74	0.4508	0.4755	5.5 %
3	Var. Previas + Territoriales	Random Forest	74	0.4542	0.4541	0.0 %
4	Var. Previas + Espaciales	XGBoost	134	0.4483	0.4558	1.7 %
4	Var. Previas + Espaciales	Random Forest	134	0.4479	0.4578	2.2 %
5	Var. Previas + Socioeconomicas	XGBoost	137	0.4474	0.4565	2.0 %
5	Var. Previas + Socioeconomicas	Random Forest	137	0.4480	0.4490	0.2 %
6	Var. Previas + Clima	XGBoost	155	0.4452	0.4609	3.5 %
6	Var. Previas + Clima	Random Forest	155	0.4401	0.4559	3.6 %
7	Var. Previas + Distancias	Random Forest	162	0.4446	0.4591	3.3 %
7	Var. Previas + Distancias	XGBoost	162	0.4391	0.4549	3.6 %

Tabla 14. Experimentos Modelos 2 y 3

# de Grupo	Descripción del Modelo	Algoritmo	Cantidad de Variables	RMSLE Entrenamiento	RMSLE Validación	Overfitting
1	Delitos	Naïve Model	1	-	0.6102	-
1	Delitos	XGBoost	6	0.4700	0.4689	-0.2 %.
1	Delitos	Random Forest	6	0.4681	0.4677	-0.1 %.
2	Delitos + Fechas	XGBoost	8	0.4671	0.4662	-0.2 %
2	Delitos + Fechas	Random Forest	8	0.4652	0.4684	0.7 %
3	Var. Previas + Territoriales	XGBoost	74	0.4604	0.4645	0.9 %
3	Var. Previas + Territoriales	Random Forest	74	0.4610	0.4662	1.1 %
4	Var. Previas + Espaciales	XGBoost	134	0.4568	0.4649	1.8 %
4	Var. Previas + Espaciales	Random Forest	134	0.4519	0.4649	2.9 %
5	Var. Previas + Socioeconomicas	XGBoost	137	0.4561	0.4655	2.1 %
5	Var. Previas + Socioeconomicas	Random Forest	137	0.4518	0.4643	2.8 %
6	Var. Previas + Clima	XGBoost	155	0.4502	0.4695	4.3 %
6	Var. Previas + Clima	Random Forest	155	0.4464	0.4670	4.6 %
7	Var. Previas + Distancias	Random Forest	162	0.4501	0.4657	3.5 %
7	Var. Previas + Distancias	XGBoost	162	0.4453	0.4667	4.8 %

Como se puede observar en las tablas, para realizar los experimentos las variables fueron asociadas en diferentes grupos. La selección de los mismos se basó en el orden en que las variables se incorporaron al conjunto de datos.

- Grupo #1: Delitos

  En primer lugar, se estimó la performance del Modelo *Naïve* utilizando las cantidades de delitos del periodo pasado como predicción. Como se menciona en la sección *Naïve* Model el principal objetivo de éste es el de validar la metodología planteada y establecer el límite inferior en la métrica elegida, que buscará ser superada con las combinaciones de los siguientes grupos y algoritmos. Teniendo en cuenta que este modelo no se entrena, la performance se obtiene directamente para el conjunto de validación.

  En segundo lugar, dentro de este grupo podemos encontrar a los algoritmos *Random Forest* y *XGBoost* que sí fueron entrenados, tanto con aquellas variables que indican los delitos en el período como en el pasado.

  Los resultados para el Modelo 1 en validación fueron 0.4565 y 0.4568 para *XGBoost* y *Random Forest*, respectivamente. Para el Modelo 2 y 3 0.4689 y 0.4677.

  Si bien la *performance* en entrenamiento es similar para los dos modelos, lo que más difiere son los resultados en el conjunto de validación. Para los tres modelos, el porcentaje de *overfitting* tiene un valor negativo, es decir, el RMSLE en validación es inferior al de entrenamiento. Además, en el caso del Modelo 1 es un porcentaje relativamente alto, con un promedio de -2.7 % para ambos algoritmo. En cambio, para los otros Modelos es menor a 0.3 % en valor absoluto.

- Grupo #2: Fechas

  En este grupo se incorporaron aquellas variables que indican fechas: mes y año. Las mismas fueron probadas de manera individual y en conjunto con las variables del grupo anterior. El objetivo de esto, es demostrar si las variables por si solas, sin utilizar los delitos pasados, tienen poder predictivo.

  La performance en validación utilizando las variables de fechas únicamente se encuentran entre 0.67 y 0.78 para los 3 modelos. Indicando que por si solas no poseen gran poder predictivo si se las compara con el Grupo de Delitos (#1). En cambio, al incluirlas con las variables del conjunto anterior, los resultados mejoran en todos los casos en entrenamiento, pero en validación lo hacen marginalmente en algunos casos y en otros empeoran. Alcanzando la mayor diferencia para el Modelo 1 con *XGBoost* con un 6.5 % más que en el grupo anterior en el RMSLE en validación.

  El *overfitting* para el Modelo 1 es positivo en ambos algoritmos, mientras que en el Modelo 2 y 3 es negativo para *XGBoost* y positivo para *Random Forest*. Además, al igual que con el grupo anterior, el porcentaje es superior en valor absoluto para el Modelo 1 en ambos algoritmos, siendo mayor a 1.7 % para *Random Forest* y 6.8 %

para *XGBoost*. En cambio, para los Modelos 2 y 3 se encuentra entre 0.2 % y 0.7 % para ambos algoritmos respectivamente en valor absoluto.

- Grupo #3: Territoriales

  En este grupo se incluyeron las variables territoriales: barrio, área del barrio, comuna, latitud y longitud. Tanto la variable de barrio como la de comuna fueron incorporadas como variables categóricas aplicándoles el método de codificación *one-hot-encoding*. De esta forma, se incorporaron un total de 66 variables; 15 comunas y 48 barrios, más las otras tres variables.

  Los resultados en validación para estas variables individualmente son mejores para *XGBoost* que para *Random Forest*. Para los 3 modelos, el RMSLE se encuentra entre el 0.47 y 0.49 y el 0.57 y 0.6, respectivamente. Al incorporar las variables anteriores, los resultados mejoran para los 3 modelos en ambos algoritmos, a excepción del Modelo 1 con *XGBoost*. Para el cual el mejor resultado continúa siendo el Grupo #1. Sin embargo, para el conjunto de entrenamiento mejoran todos.

  El porcentaje de *overfitting* para los Modelos 2 y 3 es del 0.9 % para *XGBoost* y 1.1 % para *Random Forest*. En cambio, los valores para el Modelo 1 son 5.5 % y 0.0 % para ambos algoritmos.

- Grupo #4: Espaciales Geográficas

  En este grupo se incluyeron las variables espaciales geográficas, es decir, los POIs y ROIs. Las mismas conforman 60 *feautures*. Por este motivo, se decidió utilizar todas y también probar los algoritmos y modelos utilizando solo el top 20 de variables que más correlacionan con la variable de interés.

  Los resultados de manera individual se encuentran entre 0.52 y 0.53 en todos los modelos y algoritmos para el RMSLE en validación. Al utilizar solo el top 20 la *perfomance* empeora en todos los casos.

  Al sumar estas variables en su totalidad al grupo anterior, los resultados mejoran para los 3 modelos con *Random Forest*, pero no así con *XGBoost*. En donde para el Modelo 1 el mejor modelo continúa siendo el #1 y para los Modelos 2 y 3 el #3.

  Los porcentajes de *overfitting* se incrementan en casi todos los casos en relación con el grupo anterior, menos el Modelo 1 con *XGBoost* que es inferior.

- Grupo #5: Socioeconómicas

  En este grupo se incorporaron las tres variables que dan noción del nivel socioeconómico: precio promedio de los terrenos por metro cuadrado en dólares, precio promedio de los terrenos por metro cuadrado en dólares y porcentaje de Necesidad Básicas Insatisfechas.

  De manera individual, para los tres modelos la *performance* en validación se encuentra entre 0.47 y 0.53. Al incorporarlas al resto de las variables del grupo anterior,

los resultados mejoran para todos los modelos, a excepción de los Modelos 2 y 3 para *XGBoost* que sigue siendo mejor el Grupo #3.

Los porcentajes de *overfitting* se encuentran entre 0.2 % y 2.8 % para todos los modelos y algoritmos.

- Grupo #6: Clima

  En este grupo se incluyeron las variables climáticas: temperatura, viento y precipitaciones de los periodos anteriores y de un año atrás.

  De manera individual, para todos los modelos y algoritmos los resultados se encuentran entre 0.5 y 0.53. Al incorporar este conjunto a las variables anteriores, la performance no supera a los mejores grupos anteriores.

  Los porcentajes de *overfitting* se encuentran entre 3.5 % y 4.6 % para todos los modelos y algoritmos. Encontrando los valores más altos en los Modelos 2 y 3 para *Random Forest*.

- Grupo #7: Distancias

  Por último, en este grupo se incorporaron las variables que indican distancias a distintos POIs o tipos de calles.

  La *perfomance* de las variables de manera individual se encuentra entre 0.46 y 0.50. Al incorporarlas al grupo anterior, los resultados mejoran en relación al grupo 6, pero no son superadores con respecto a los de los grupos previos con un mejor resultado.

  Los porcentajes de *overfitting* se encuentran entre 3.3 % y 4.8 % para todos los modelos y algoritmos.

En conclusión, los mejores resultados para el Modelo 1 en ambos algoritmos se encuentran al utilizar el Grupo #5 con las variables de los conjuntos anteriores y la incorporación de las socioeconómicas. Para los Modelos 2 y 3 con *XGBoost* es el Grupo #3 y para *Random Forest* el Grupo #5.

5.3 Selección de Modelos

Después de elegir los conjuntos de variables con mejores resultados, se optimizaron los parámetros utilizando lo mencionado en Optimización de hiperparámetros. La siguiente tabla resume la mejor combinación para ambos algoritmos en los tres modelos. Asimismo, compara la diferencia entre los resultados en entrenamiento, validación y *overfitting* con los hiperparámetros por *default* y los valores optimizados.

Tabla 15. Selección de Modelos

Modelo	Algoritmo	Cant. Var.	Hiperparametros	RMSLE Entrenamiento	RMSLE Validación	Overfitting
Modelo 1	XGBoost	137	{nrounds: 2400, max_depth: 12, eta: 0.005, gamma: 5.5, colsample_bytree: 0.8, subsample: 0.8, min_child_weight: 18}	Default: 0.4474 Optimizado: 0.4288 Diferencia:-4.2 %	Default: 0.4565 Optimizado: 0.4517 Diferencia: -1.1 %	Default: 2.0 % Optimizado: 5.4 % Diferencia: 3.32pp
Modelo 1	Random Forest	137	{mtry: 15, min.node.size: 150, splitrule: extratrees, num.trees: 300}	Default: 0.4480 Optimizado: 0.4477 Diferencia:-0.1 %	Default: 0.4490 Optimizado: 0.4477 Diferencia:-0.3 %	Default: 0.2 % Optimizado: 0.0 % Diferencia:-0.24pp
Modelo 2 y Modelo 3	XGBoost	74	{nrounds: 2000, max_depth: 5, eta: 0.01, gamma: 5, colsample_bytree: 0.6, subsample: 0.7, min_child_weight: 26}	Default: 0.4604 Optimizado: 0.4541 Diferencia:-1.4 %	Default: 0.4645 Optimizado: 0.4642 Diferencia:-0.1 %	Default: 0.9 % Optimizado: 2.2 % Diferencia:1.34pp
Modelo 2 y Modelo 3	Random Forest	137	{mtry: 110, min.node.size: 300, splitrule: extratrees, num.trees: 1000}	Default: 0.4518 Optimizado: 0.4459 Diferencia:-1.3 %	Default: 0.4643 Optimizado: 0.4641 Diferencia:-0.05 %	Default: 2.8 % Optimizado: 4.1 % Diferencia:1.3pp

Para el Modelo 1 con *XGBoost*, se ha conseguido una disminución tanto en el RMSLE de entrenamiento como en el de validación. En el primer caso, es una reducción de un 4.2 %, mientras que en el segundo es de un 1.1 %. Sin embargo, el porcentaje de *overfitting* se ha incrementado en 3.32 puntos porcentuales. Para este mismo modelo con *Random Forest*, los porcentajes de mejora son inferiores. Para la métrica en entrenamiento es de un 0.1 %, mientras que para validación es de un 0.3 %. En cambio, se logró una disminución en el porcentaje de *overfitting* de 0.24 puntos porcentuales.

Para los Modelos 2 y 3 con *XGBoost* también se ha conseguido una disminución tanto en el RMSLE de entrenamiento como en el de validación. En el primer caso, es una reducción de un 1.4 %, mientras que en el segundo es de un 0.1 %. Sin embargo, el porcentaje de *overfitting* se ha incrementado en 1.34 puntos porcentuales. Al igual que con el Modelo 1, para este modelo con *Random Forest*, los porcentajes de mejora son inferiores. En entrenamiento es de un 1.3 %, mientras que para validación es de un 0.05 %. El porcentaje de *overfitting* se vió incrementado en 1.3 puntos porcentuales.

En líneas generales, se puede decir que el optimizar hiperparámetros mejora más la *performance* con el algoritmo *XGBoost*, pero consigue una mayor diferencia entre el RMSLE de entrenamiento y validación que *Random Forest*.

5.4 Modelo Final

Por último, habiendo seleccionado la mejor combinación de hiperparámetros y variables para cada modelo, se procede a predecir el conjunto de *test*. Para esto se vuelve a entrenar el modelo utilizando tanto el conjunto de entrenamiento como el de validación. A continuación, se darán a conocer los resultados de los distintos modelos para cada uno de los algoritmos[12].

Tabla 16. Resultados de Modelos

Modelo	Algoritmo	RMSLE Entrenamiento	RMSLE Evaluación	Overfitting	RMSE Evaluación	MAE Evaluación
Modelo 1	Naïve	-	0.5948		1.8849	1.0488
Modelo 1	XGBoost	0.4287	0.4552	6.2 %	1.3832	0.8493
Modelo 1	Random Forest	0.4437	0.4489	0.4 %	1.3960	0.8380
Modelo 2	Naïve	-	0.6002		1.9423	1.1518
Modelo 2	XGBoost	0.4538	0.4616	1.7 %	1.4845	0.9313
Modelo 2	Random Forest	0.4457	0.4597	3.1 %	1.4518	0.9157
Modelo 3	Naïve	-	0.6394		2.4209	1.2578
Modelo 3	XGBoost	0.4538	0.4543	0.1 %	1.3559	0.8401
Modelo 3	Random Forest	0.4457	0.4570	2.5 %	1.3872	0.8444

En primer lugar, al comparar los algoritmos de *machine learning* se puede notar que la mayoría presentan resultados similares en la métrica de interés. Sin embargo, al compararlos con el modelo *Naïve*, la mejoría que presentan los mismos es notable. En promedio, el RMSLE en evaluación de los algoritmos de aprendizaje automático tienen una mejora de un 25 % en los resultados con respecto a este. Siendo el Modelo 3 el que mayor proporción de reducción del error presenta con un 29 % en promedio.

En segundo lugar, al analizarlos de manera individual, podemos observar que la mejor performance en la métrica de interés la obtuvo el Modelo 1 con el algoritmo *Random Forest*. También, podemos destacar el Modelo 3 con *XGBoost*, el cual presenta el RMSLE en evaluación más bajo para este algoritmo. En relación al porcentaje de *overfitting*, el valor más alto lo presenta el Modelo 1 con *XGBoost* y el más bajo el Modelo 3 con el mismo algoritmo. Sin embargo, todos presenta porcentajes relativamente bajos, menores al 6.2 %.

Por último, si analizamos dos métricas más: RMSE y MAE, los mejores resultados los obtuvo el Modelo 3 con *XGBoost* para la primera y el Modelo 1 con *Random Forest*. De todas formas, al igual que con el RMSLE, no hay una gran diferencia entre la *performance* de los algoritmos de aprendizaje automático, pero sí en relación al modelo *Naïve*.

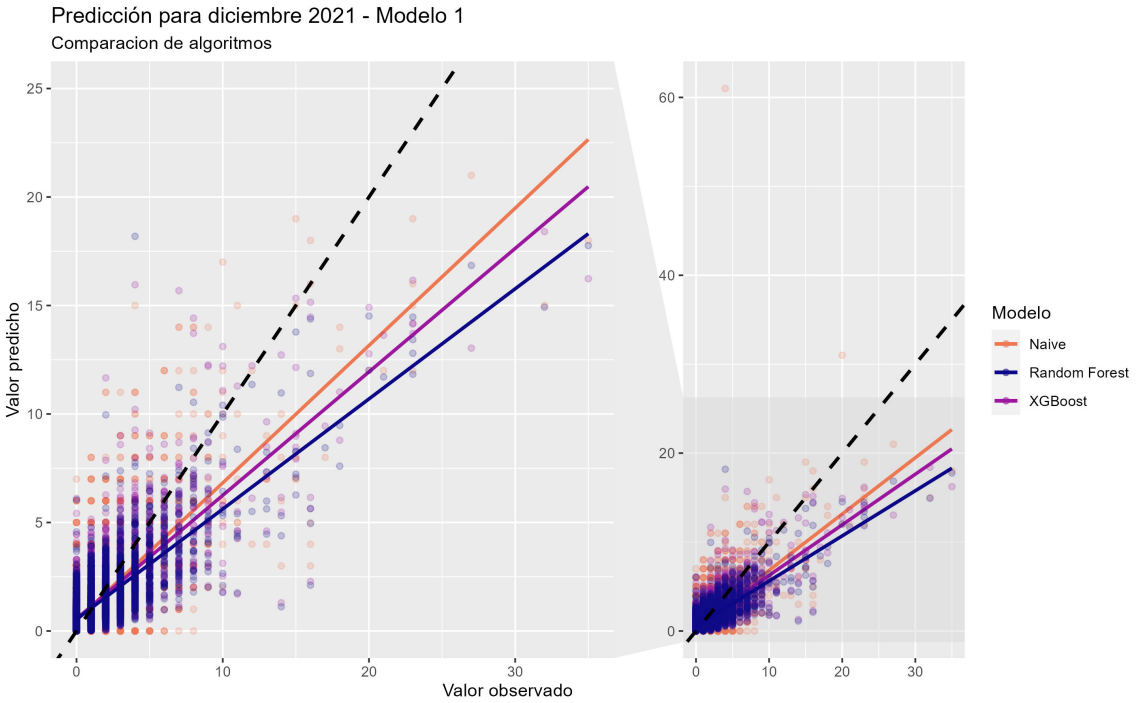
[12]Modelo 1 Naïve: Se utilizó noviembre de 2021 como predicción para diciembre de 2021. Modelo 2 Naïve: Se utilizó noviembre de 2019 como predicción para diciembre de 2019. Modelo 3 Naïve: Se utilizó noviembre de 2019 como predicción para diciembre de 2021.

5.5 Análisis de Resultados

En las siguientes figuras, se mostrarán los gráficos comparando los registros observados con los predichos a través de los algoritmos *Naïve*, *XGBoost* y *Random Forest* en los distintos modelos. La línea punteada representa dónde caerían los puntos si todos los valores predichos coincidieran perfectamente con los observados.

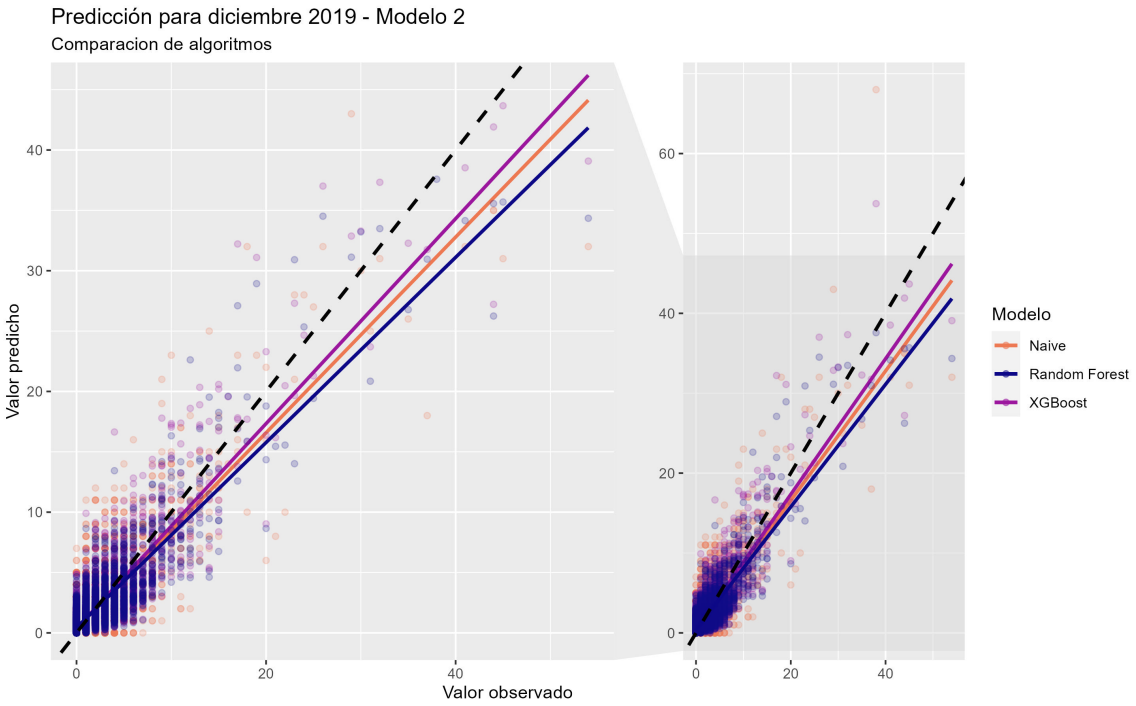
Al comparar los registros observados con los predichos para el Modelo 1 se puede observar que el modelo *Naïve* es el que más se asemeja a la línea punteada, seguido de *XGBoost* y por último *Random Forest*. El valor máximo predicho para el modelo ingenuo es de 61, mientras que los algoritmos de aprendizaje automático es 18. Adicionalmente, los valores observados tienen un máximo de 35. En cuanto al valor mínimo, todos poseen al 0.

Figura 35. Valores Predichos vs. Observados - Modelo 1



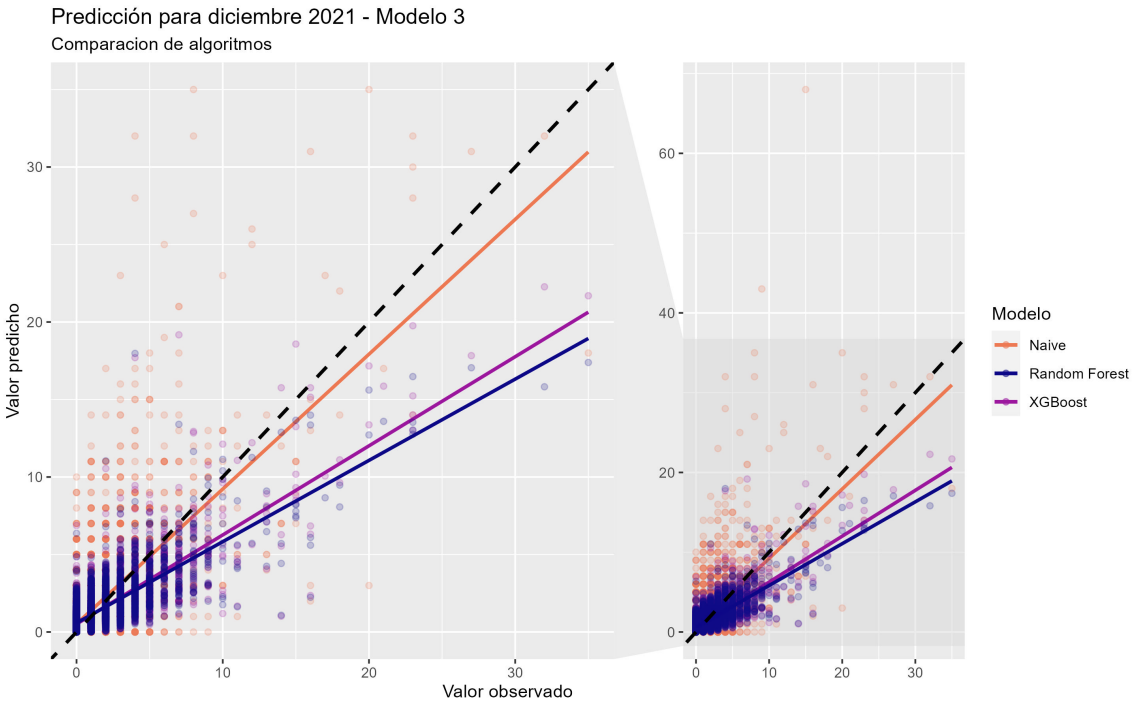
Por otro lado, al hacer esta misma comparación para el Modelo 2, el modelo que más se asemeja a la línea punteada es *XGBoost*, seguido del modelo *Naïve* y por último, *Random Forest*. Al verlo en relación con el Modelo 1, se puede ver que los tres algoritmos tienen una pendiente más similar a la de la línea punteada. Los valores más altos nuevamente los presenta el modelo ingenuo con 68, 54 para *XGBoost* y 38 *Random Forest*. Al igual que el modelo anterior, todos poseen como mínimo al 0.

Figura 36. Valores Predichos vs. Observados - Modelo 2



Por último, el Modelo 3 presenta la mayor diferencia entre algoritmos. El modelo *Naïve* es más similar a la línea punteada y los otros dos algoritmos presentan una pendiente con un menor ángulo. Mostrando una gran diferencia. Los valores más altos los presenta el modelo ingenuo con 68, seguido de *XGBoost* con 22 y *Random Forest* con 18. Como mínimo nuevamente todos los algoritmos tienen al 0.

Figura 37. Valores Predichos vs. Observados - Modelo 3



En los siguientes gráficos se presentará la distribución de los valores predichos para cada uno de los algoritmos en los distintos modelos.

Para los tres modelos se puede observar que las distribuciones de los valores observados y el modelo *Naïve* son muy similares y de la misma forma sucede con *XGBoost* y *Random Forest*. Siendo el Modelo 3 el que mayor diferencia presenta entre estos ultimos dos algoritmos. Al comparar las distribuciones, las del modelo *Naïve* y de los valores observados tienen un pico más alto en el 0. Mientras que para *XGBoost* y *Random Forest* se encuentra más cercano al 1 el pico. Algo que cabe destacar es que los resultados de los modelos de *machine learning* no son números enteros y, por lo tanto, se puede observar el pico más alto entre el 0.5 y 1.

Figura 38. Distribución de Valores Predichos - Modelo 1

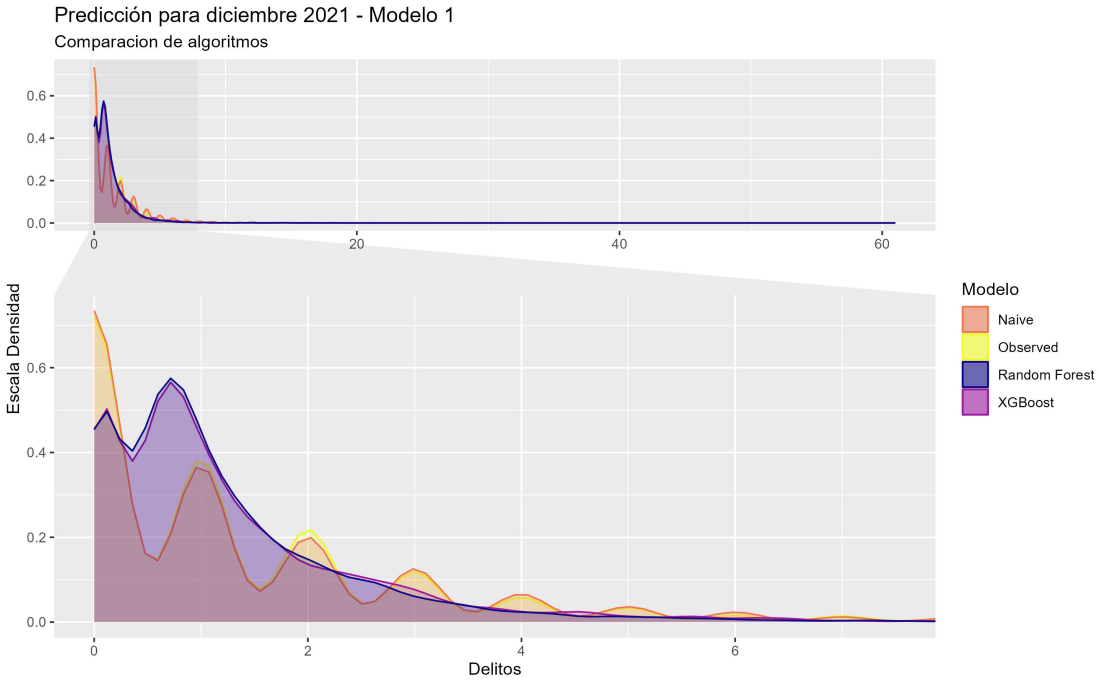


Figura 39. Distribución de Valores Predichos - Modelo 2

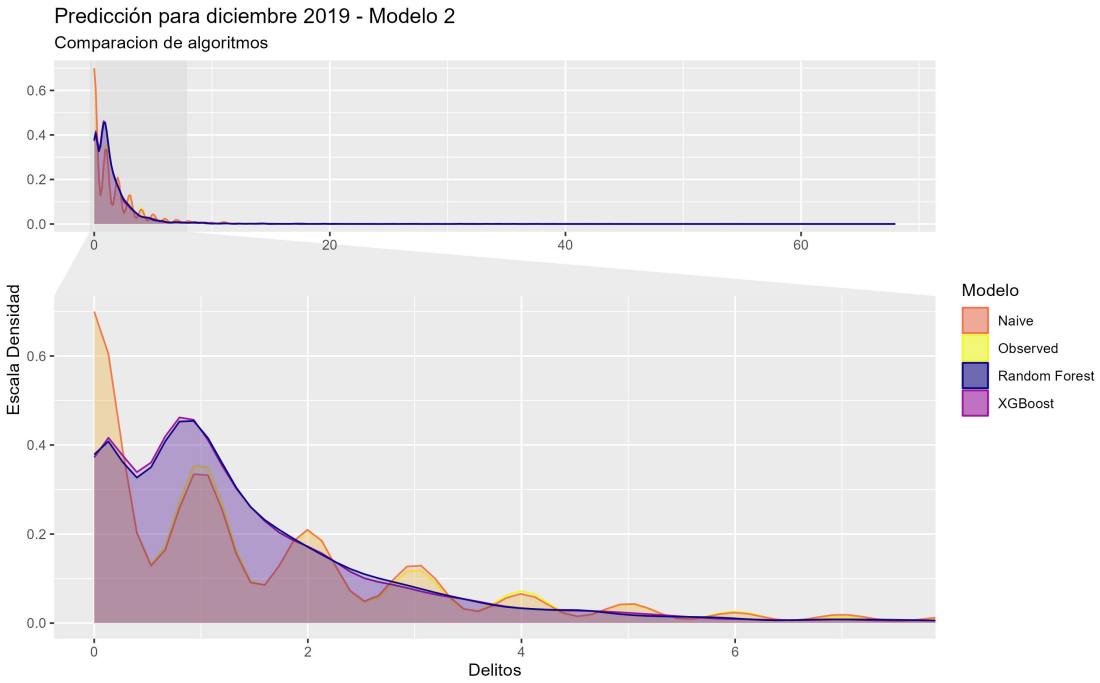
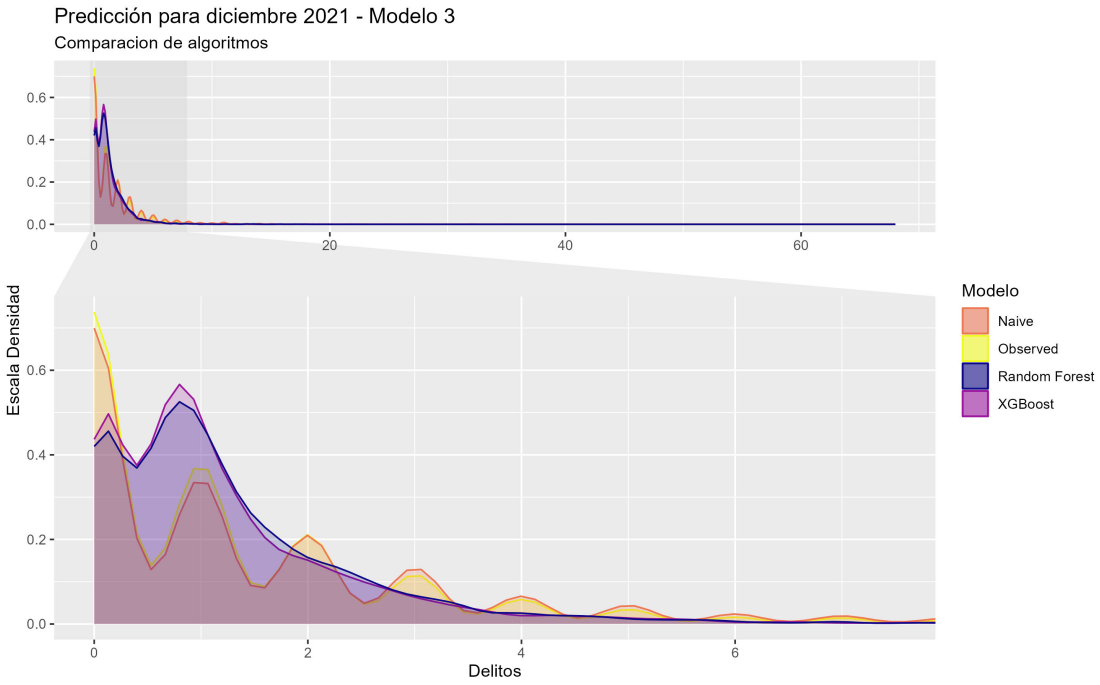


Figura 40. Distribución de Valores Predichos - Modelo 3

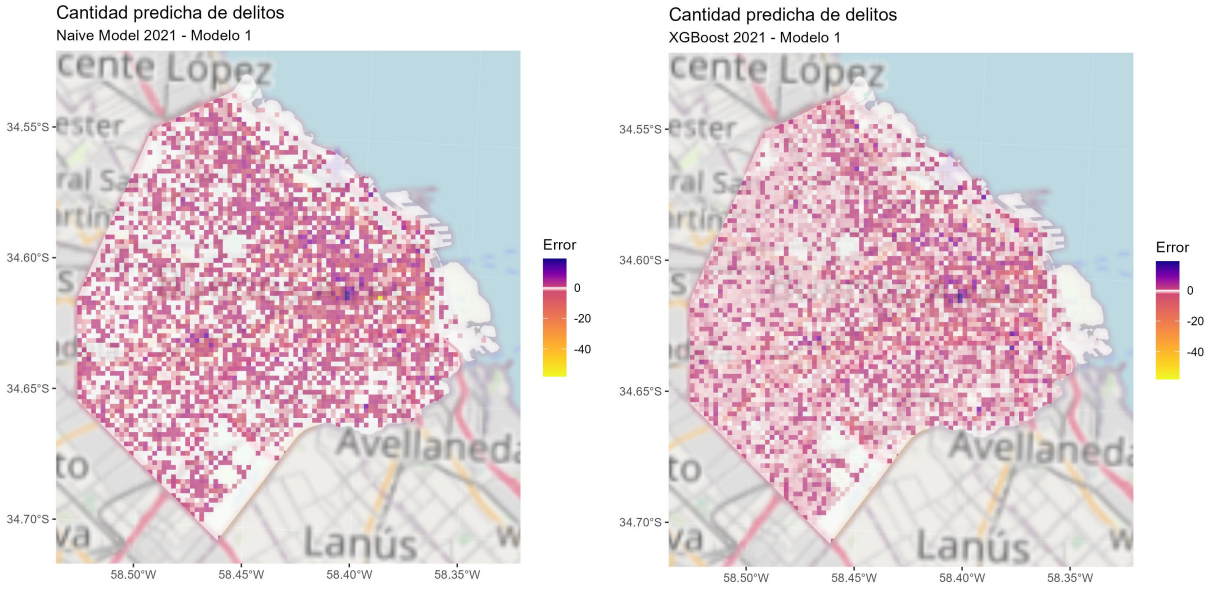


Sin embargo, al ser modelos en los que no solo se busca acertar con la cantidad de delitos que ocurrirán, sino que uno de los principales objetivos es conocer su ubicación, analizaremos mapas coropléticos para cada modelo - algoritmo en donde se graficaron los errores para cada celda de la grilla. En estos, los colores más similares al blanco representan un error de 0 o cercano al 0, mientras que los colores más rosas/violetas

representan valores positivos y los colores más naranjas/amarillos valores negativos. La escala es igual para todos los algoritmos de un mismo modelo.

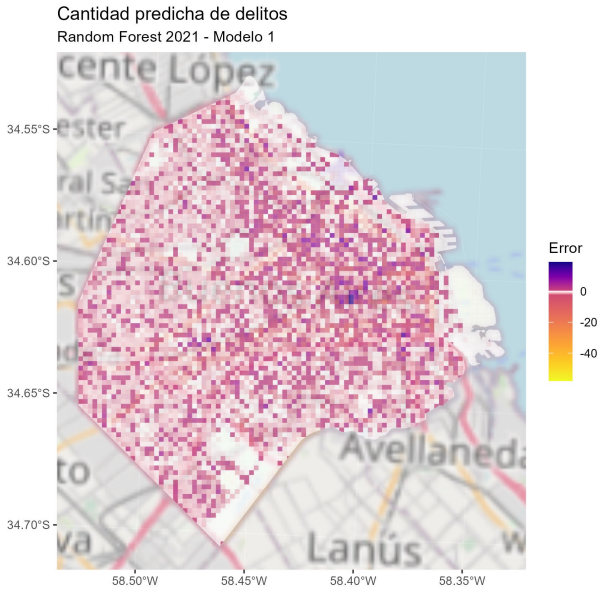
Para el Modelo 1 se puede observar que tanto *XGBoost* como *Random Forest* poseen muchas más celdas de la grilla con colores más claros, más similares al blanco y por lo tanto, con un menor error. Esto explica por qué a pesar de haber visto los gráficos anteriores, los algoritmos de *machine learning* poseen una mejor *performance* en la métrica de interés. Dado que si se analiza el error para cada cuadrado de la grilla de manera individual, se obtiene un valor más bajo. En particular, esto es más visible del lado oeste del mapa.

Figura 41. Errores de los algoritmos en la grilla - Modelo 1



(a) Naïve Error

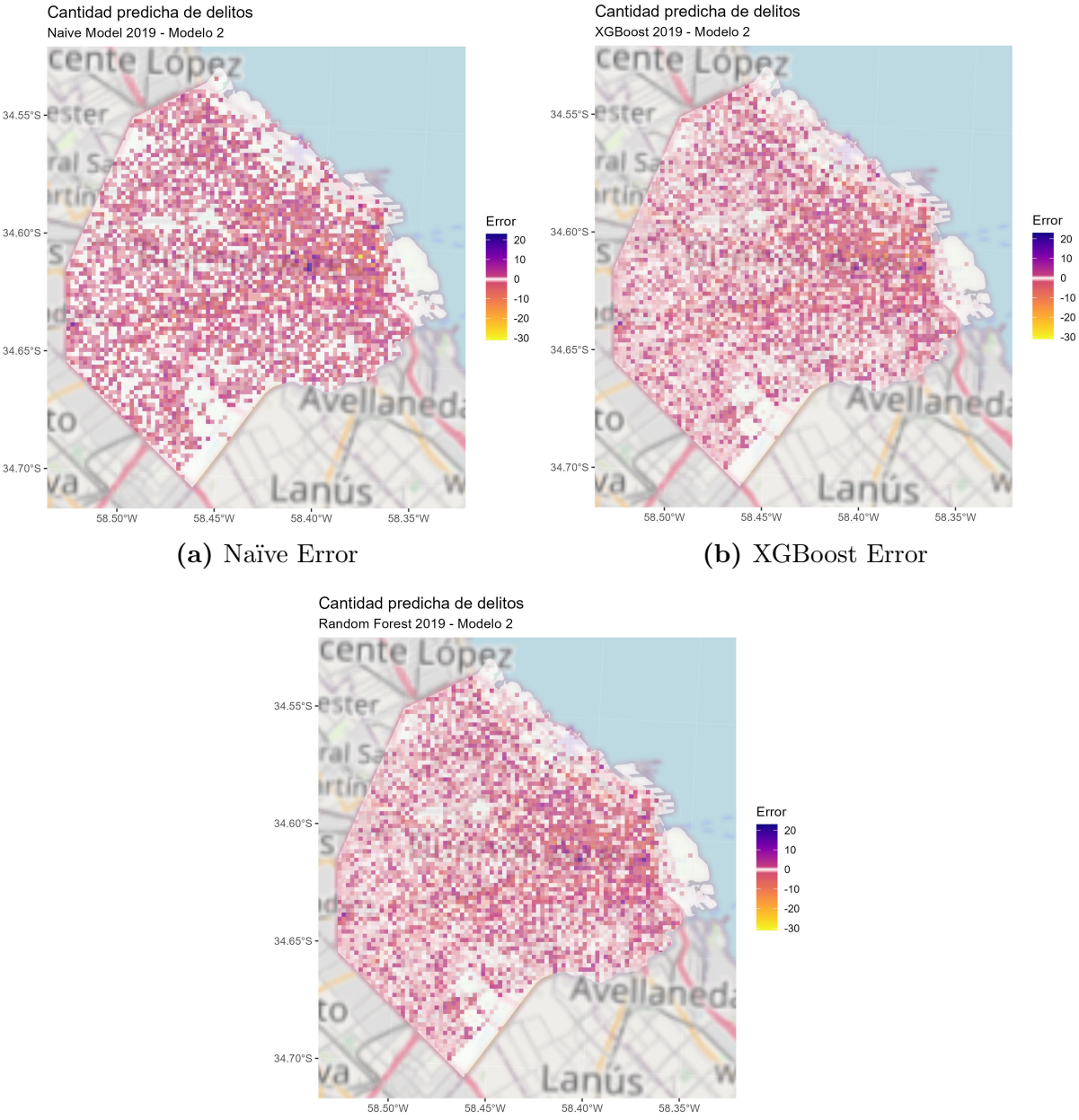
(b) XGBoost Error



(c) Random Forest Error

Para el Modelo 2, el análisis es similar al del anterior. Se puede observar más celdas de la grilla con colores más claros en los modelos de aprendizaje automático que en el modelo ingenuo. Sin embargo, esta diferencia es menos notoria que en el Modelo 1. Algo que se puede destacar también es que en el lado centro - este se pueden apreciar celdas de la grilla con un color más anaranjado, indicando que los valores de error en esas cuadrículas son negativos.

Figura 42. Errores de los algoritmos en la grilla - Modelo 2



(a) Naïve Error

(b) XGBoost Error

(c) Random Forest Error

Por último, los mapas del Modelo 3 son bastante similares a los del 2. En donde los colores anaranjados en el centro - este son aún más notorios en el modelo ingenuo y *Random Forest*.

Figura 43. Errores de los algoritmos en la grilla - Modelo 3

(a) Naïve Error

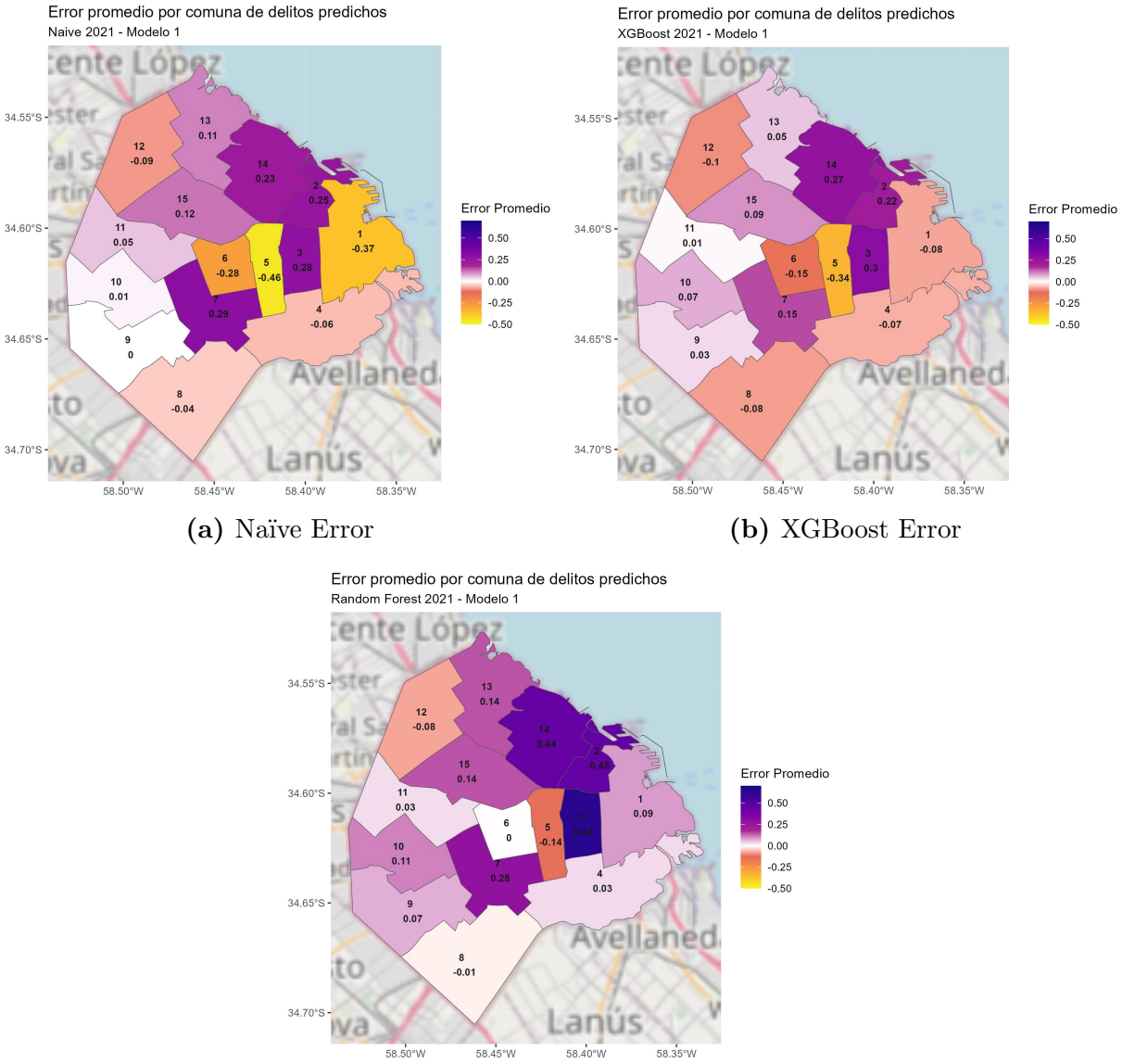
(b) XGBoost Error

(c) Random Forest Error

Para continuar con el análisis de los resultados a nivel espacial, se graficaron los errores promedio para cada comuna en cada algoritmo y modelo.

En el Modelo 1 se puede observar que 8 de las 15 comunas tanto para *XGBoost* como para *Random Forest* tienen un error más bajo que el modelo *Naïve*. Para *XGBoost* estas comunas son: 1, 2, 5, 6, 7, 11, 13 y 15. Para *Random Forest* son: 1, 4, 5, 6, 7, 8, 11 y 12. Estando el error más alto en términos absolutos en la comuna 3 para *Random Forest* y el más bajo en términos absolutos en la comuna 9 para el modelo *Naïve*. Al igual que con los mapas de grilla, se puede observar que el error de menor magnitud se encuentra del lado oeste del mapa. En más de la mitad de las comunas para los tres algoritmos el error promedio es positivo.

Figura 44. Errores promedio de los algoritmos por comuna - Modelo 1



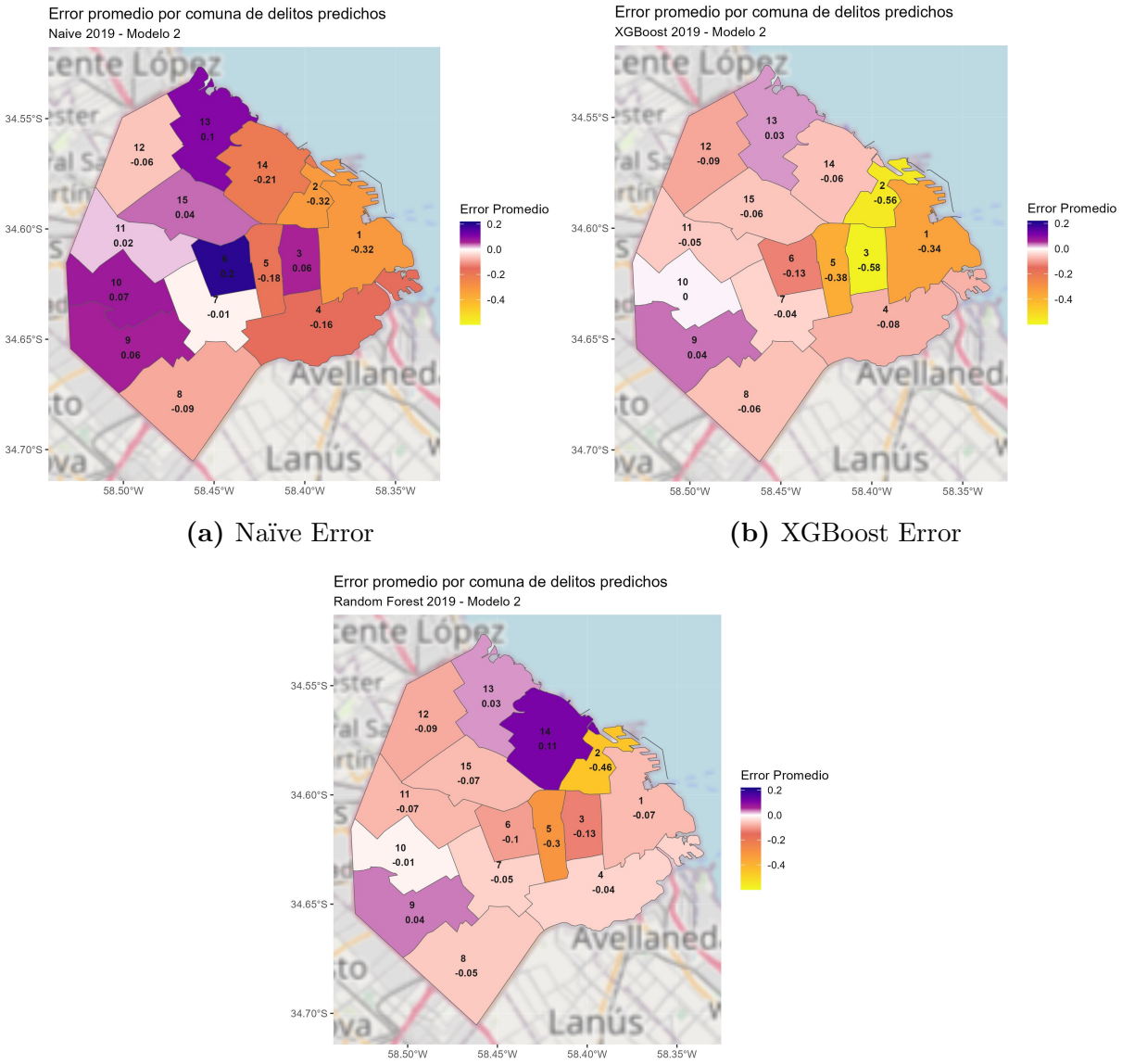
(a) Naïve Error

(b) XGBoost Error

(c) Random Forest Error

Al realizar este mismo análisis para el Modelo 2, se puede observar que el modelo *Naïve* tiene casi el 50 % de las más comunas con errores promedios positivos, mientras que para los algoritmos de aprendizaje automático la mayoría poseen errores negativos. Esto se puede observar a simple vista al ver que los colores que predominan son los naranjas/amarillos en lugar de los rosas/violetas. En este caso, la cantidad de comunas con error más bajo en términos absolutos que el modelo *Naïve* son 7 para *XGBoost* y 8 para *Random Forest*. Para *XGBoost* estas comunas son: 4, 6, 8, 9, 10, 13 y 14. Para *Random Forest* coinciden las mismas comunas y además la 1. Estando el error más alto en términos absolutos en la comuna 3 para *XGBoost* y el más bajo en la comuna 10 para el mismo algoritmo.

Figura 45. Errores promedio de los algoritmos por comuna - Modelo 2



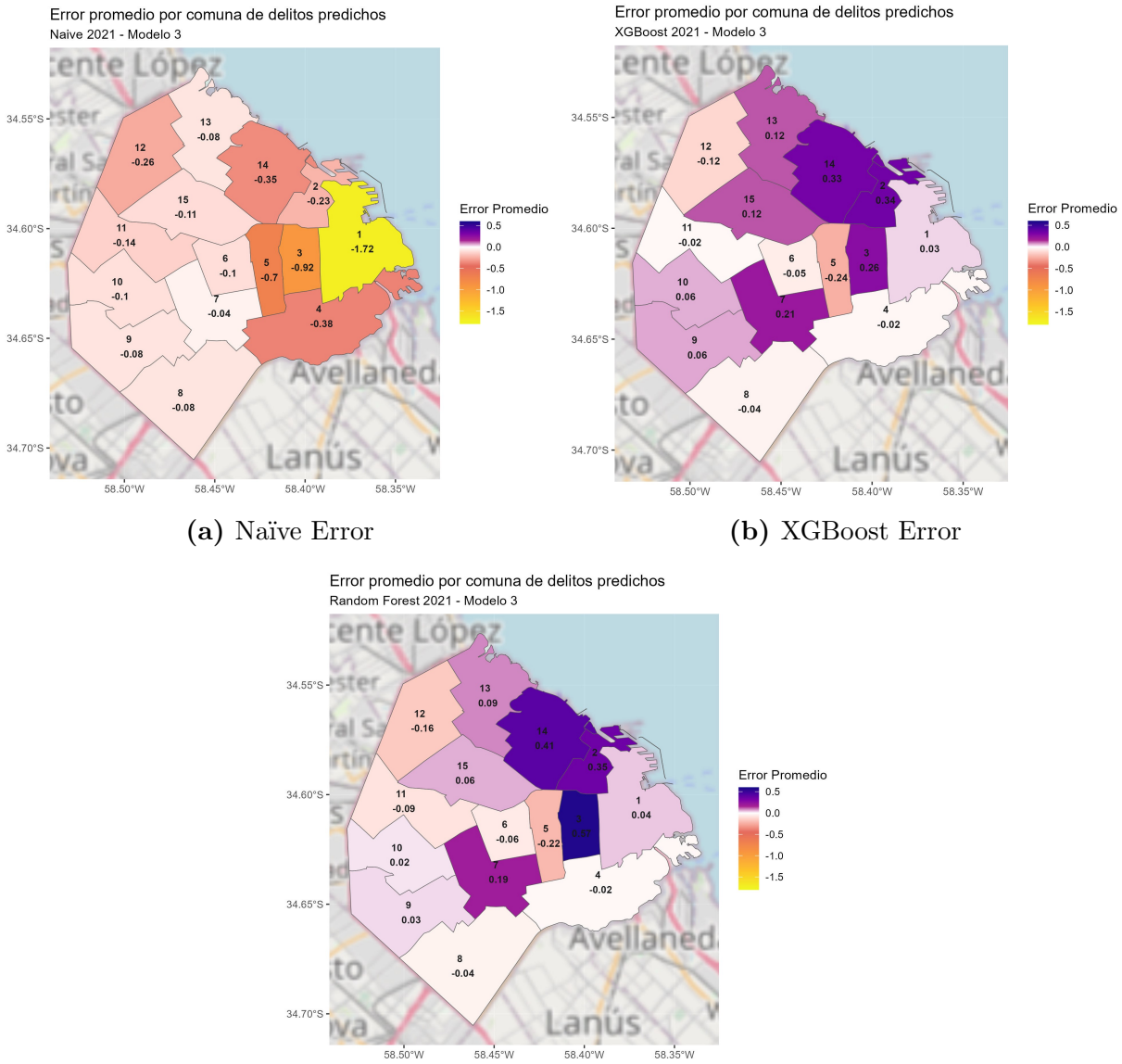
(a) Naïve Error

(b) XGBoost Error

(c) Random Forest Error

Por último, al analizar el mapa para el Modelo 3, se puede observar que el modelo *Naïve* posee todos los errores promedios negativos, mientras que para los algoritmos de aprendizaje automático más de la mitad de las comunas poseen errores positivos. En este caso la cantidad de comunas con error más bajo en términos absolutos que el modelo *Naïve* son 11 tanto para *XGBoost* como para *Random Forest*. Para *XGBoost* estas comunas son: 1, 3, 4, 5, 6, 8, 9, 10, 11, 12 y 14. Para *Random Forest* son las mismas, a excepción de la 14 y sí la 15. Estando el error más alto en términos absolutos en la comuna 1 para *Naïve* y el más bajo en la comuna 10 para el modelo *XGBoost* y en la 11 para *Random Forest*.

Figura 46. Errores promedio de los algoritmos por comuna - Modelo 3



(a) Naïve Error

(b) XGBoost Error

(c) Random Forest Error

Si se analizan en relación al modelo ingenuo, la mejor *performance* la obtuvo el Modelo 3 con 11 comunas en cada algoritmo con un menor error en términos absolutos.

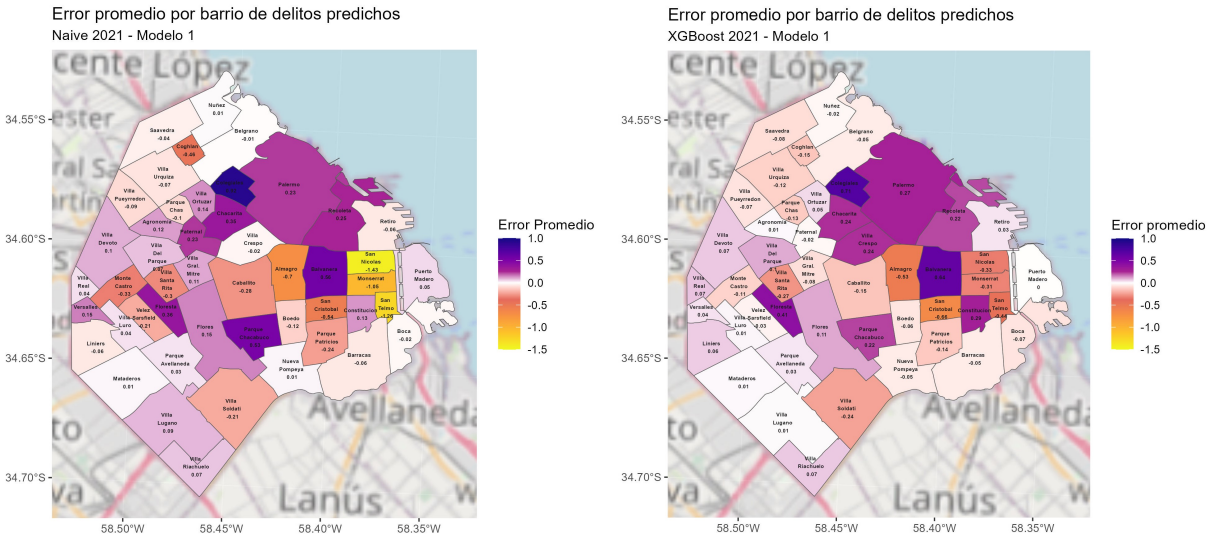
En último término se buscó comparar esto a nivel barrio de la ciudad. [13]

En el Modelo 1 se puede observar que 31 (64.6 %) de los 48 barrios para *XGBoost* tienen un error más bajo que el modelo *Naïve* y para *Random Forest* son 26 (54.2 %). Compartiendo entre ambos algoritmos 23 barrios con mejor performance que el modelo ingenuo. El error promedio más alto en términos absolutos se encuentra en el barrio San Nicolás para el modelo *Naïve* y el más bajo en el barrio Caballito para *Random Forest*. Los barrios con menor error se pueden apreciar del lado oeste en el mapa de *XGBoost* con los colores más claros. Los errores más altos positivos (más violetas) se observan en

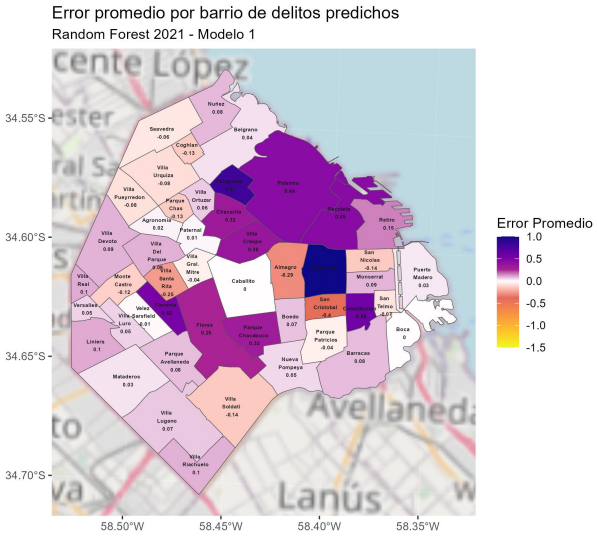
[13] Mapas en un tamaño más grande Mapas error promedio por barrio

el centro - este y de la misma forma sucede con los errores negativos. En particular, los colores más naranjas/amarillos (más negativos) se observan en el mapa del modelo *Naïve*.

Figura 47. Errores promedio de los algoritmos por barrio - Modelo 1



(a) Naïve Error **(b)** XGBoost Error

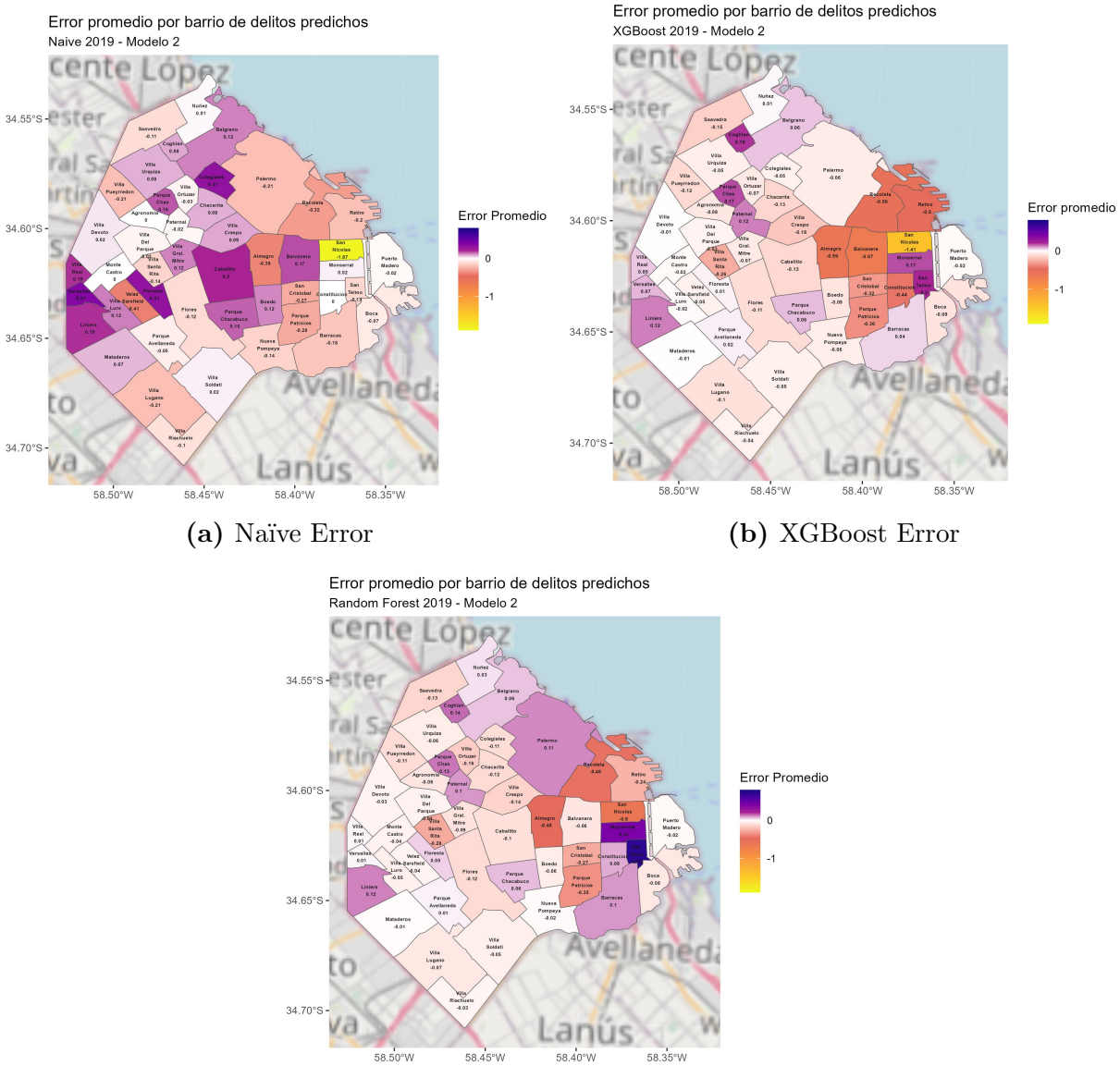


(c) Random Forest Error

En el Modelo 2 se puede observar que 25 (52.1 %) de los 48 barrios tanto para *XGBoost* como para *Random Forest* tienen un error más bajo que el modelo *Naïve*. Compartiendo entre ambos algoritmos 23 barrios con mejor performance que el modelo ingenuo. El error promedio más alto en términos absolutos se encuentra, al igual que el modelo anterior, en el barrio San Nicolás para el modelo *Naïve* y el más bajo en los barrios Monte Castro, Constitución y Agronomía para el mismo algoritmo. Los barrios con menor error se pueden apreciar del lado oeste en los mapas de *XGBoost* y *Random Forest* con los colores más claros. Los errores más altos positivos y negativos se observan en el centro - este. En particular, los colores más amarillos (más negativos) se observan en los mapas del modelo

Naïve y *XGBoost* y los más violetas (más positivos) en *Random Forest*.

Figura 48. Errores promedio de los algoritmos por barrio - Modelo 2



(a) Naïve Error

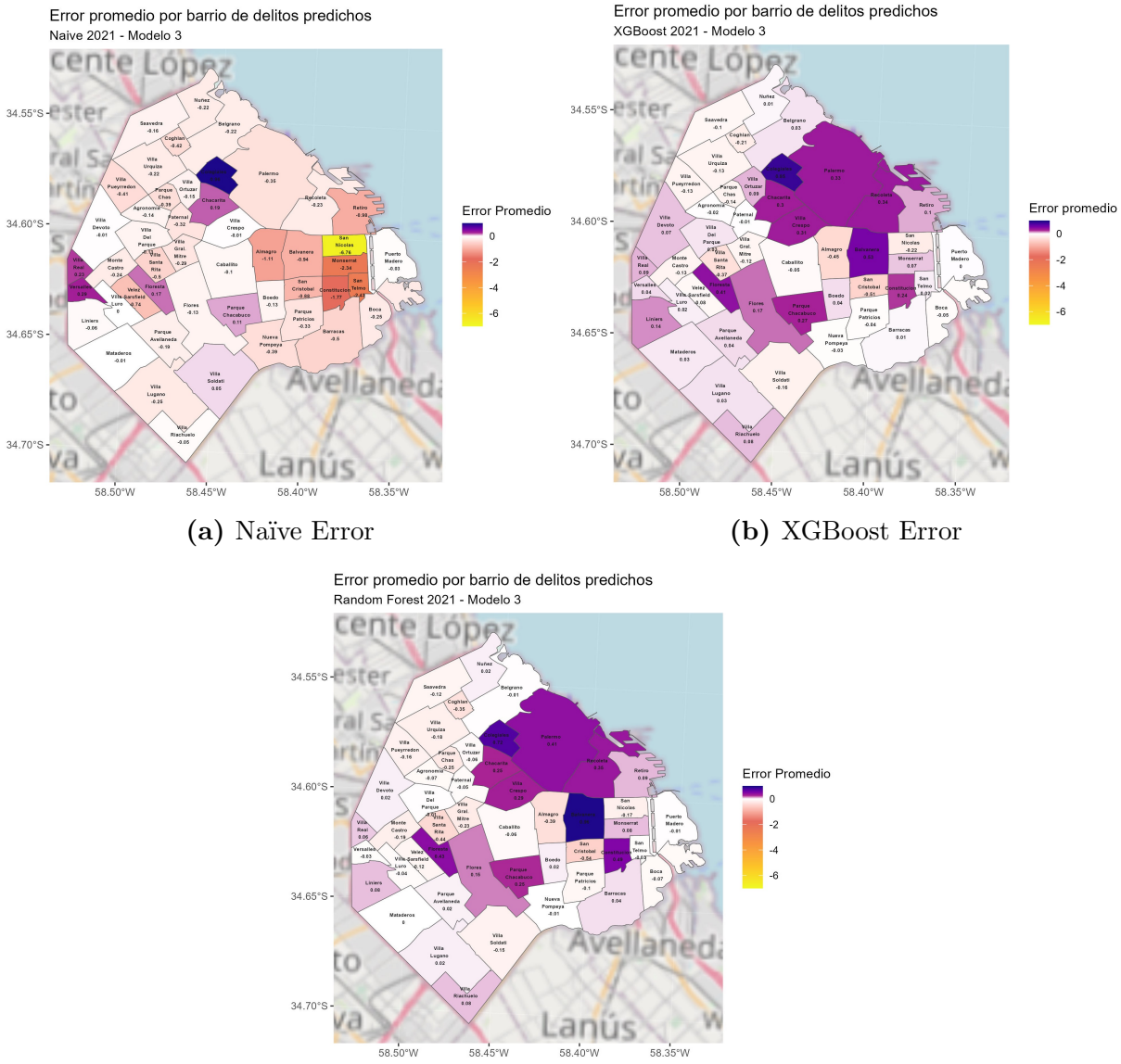
(b) XGBoost Error

(c) Random Forest Error

Por último, al analizar el Modelo 3 se puede observar que 36 (75 %) de los 48 barrios para *XGBoost* tienen un error más bajo que el modelo *Naïve* y para *Random Forest* son 35 (72.9 %). Compartiendo entre ambos algoritmos 34 barrios con mejor performance que el modelo ingenuo. El error promedio más alto en términos absolutos se encuentra, al igual que en los dos modelos anteriores, en el barrio San Nicolás para el modelo *Naïve* y el más bajo en el barrio Villa Luro para el mismo algoritmo. Los barrios con menor error se pueden apreciar del lado oeste en los mapas de *XGBoost* y *Random Forest* con los colores más claros. Los errores más altos positivos (más violetas) se observan en el centro - este de *XGBoost* y *Random Forest* y de la misma forma sucede con los errores negativos (más amarillos) en el modelo *Naïve*. Además, este último posee los errores más

altos positivos en el oeste.

Figura 49. Errores promedio de los algoritmos por barrio - Modelo 3



(a) Naïve Error

(b) XGBoost Error

(c) Random Forest Error

Para los barrios, si se analizan en relación al modelo ingenuo, la mejor *performance* la obtuvo el Modelo 3 con 36 barrios para *XGBoost* y 35 para *Random Forest* con un menor error en términos absolutos.

5.6 Importancia de variables

En términos de las variables más importantes para cada modelo, se puede destacar que los tres coinciden en que son las variables de los delitos acumulados en los meses

pasados.[14] En el caso del algoritmo *Random Forest* luego se encuentran los delitos de hace un año para los tres modelos y en el de *XGBoost* varía entre modelos; para el Modelo 1 se encuentran las variables de fechas, seguidas de la longitud, porcentaje promedio del NBI, latitud y los crímenes del año pasado. En cambio, para el Modelo 2 y 3 están la latitud y longitud, los crímenes del año pasado y luego, las variables de fechas.

Cabe aclarar que el Modelo 1 con ambos algoritmos fueron entrenados con 137 variables, incluyendo los delitos acumulados pasados y los de un año atrás, las variables de fechas, territoriales, espaciales y socioeconómicas. Asimismo, los Modelos 2 y 3 con *XGBoost* fueron entrenados con las mismas variables. Por último, los Modelos 2 y 3 con *Random Forest* incluyeron las mismas variables a excepción de las espaciales y socioeconómicas.

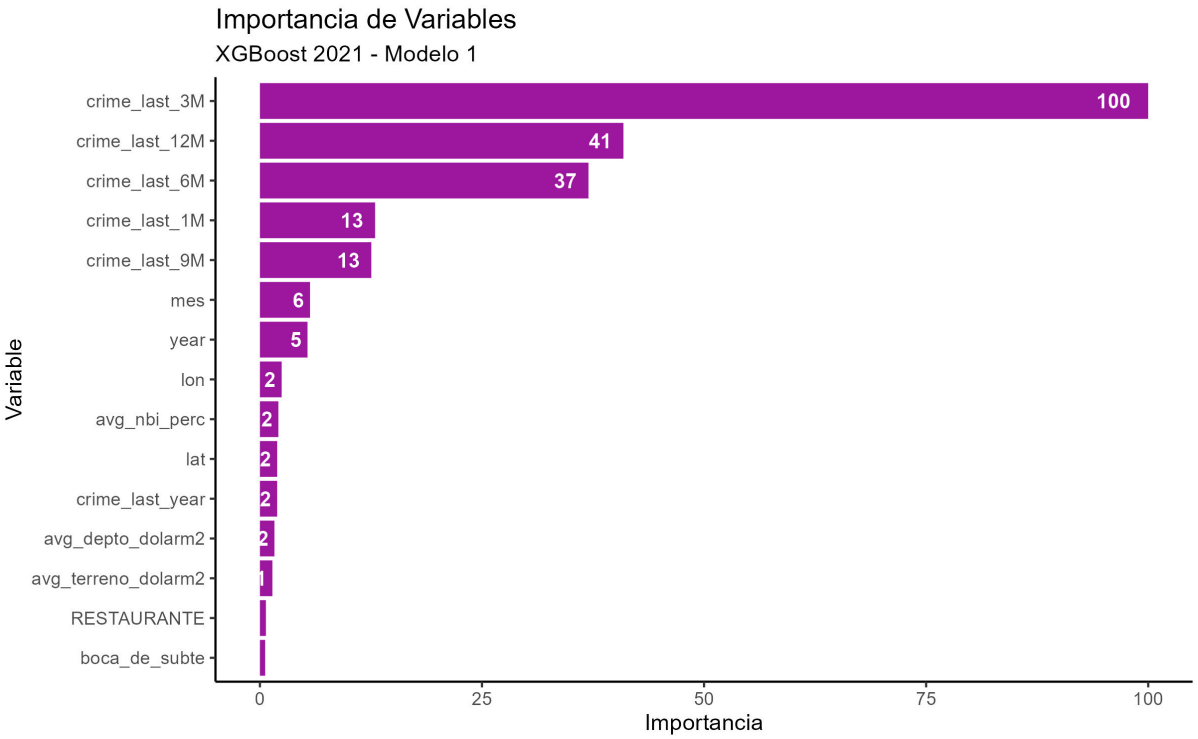
Por otro lado, los modelos con *Random Forest* poseen en el top 15 a las variables que indican a los restaurantes, cajeros automáticos, bocas de subte y farmacias. Además, en el caso de *Random Forest* para el Modelo 1 y *XGBoost* para los Modelos 2 y 3, tienen a la Comuna 3 y a los barrios de San Nicolás y Balvanera. Algo a destacar también es que para *XGBoost* en el Modelo 1 las 3 variables socioeconómicas tienen importancia. En el caso de los Modelos 2 y 3 para *Random Forest* la variable de porcentaje de NBI aparece en el ranking.

Otro punto es que el Modelo 1 para ambos algoritmos tiene en el top 3 de variables más importantes a las de 3, 6 y 12 meses. Mientras que los otros Modelos a las de 6, 9 y 12 meses. Coincidiendo todos que medio año (6 meses) y un año (12 meses) son importantes para todos los modelos - algoritmos a la hora de predecir los delitos del próximo mes. De este modo, parecería que al utilizar todos los datos para entrenar y predecir diciembre de 2021 tienen mayor poder predictivo los delitos más recientes y al utilizar los datos de 2016 a 2019 para entrenar tienen más valor los delitos acumulados de mayor tiempo.

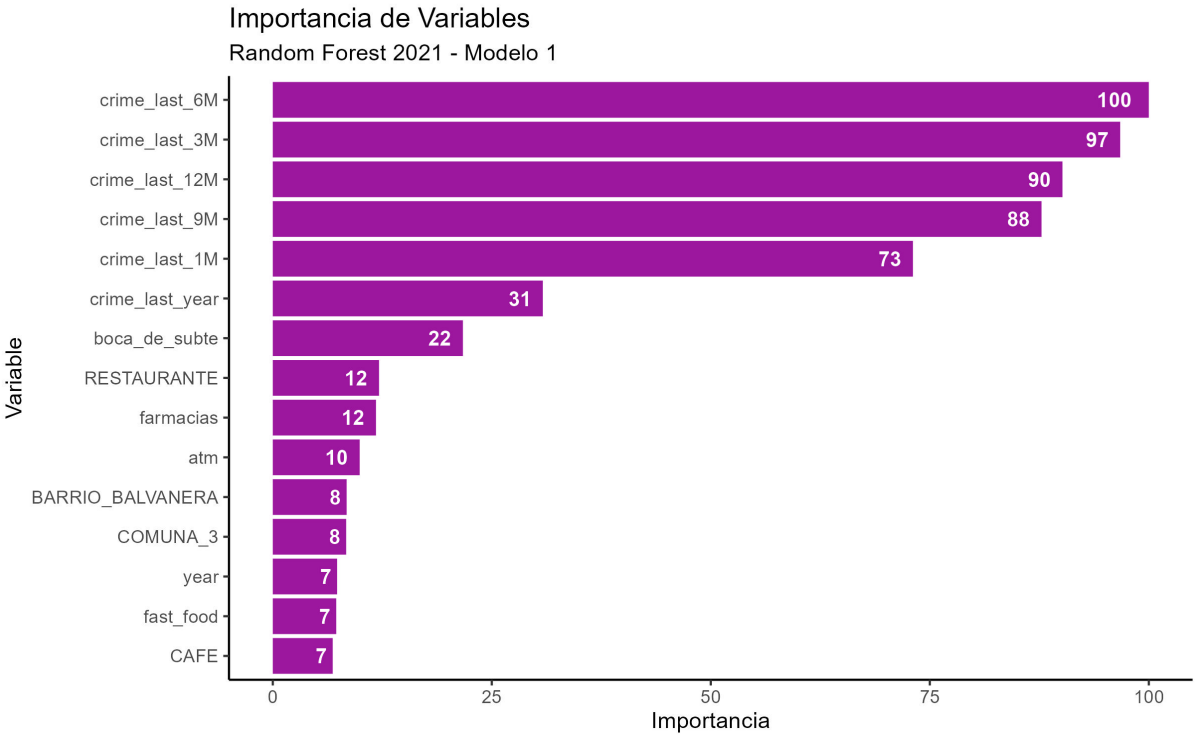
Por su parte, en *XGBoost* para todos los modelos, el peso de la primera variable es muy superior a la segunda. En cambio, para *Random Forest* no hay tanta diferencia entre la primera y las siguientes.

[14] Los modelos 2 y 3 al haber sido entrenados con los mismos datos, presentan las mismas variables.

Figura 50. Importancia de variables - Modelo 1

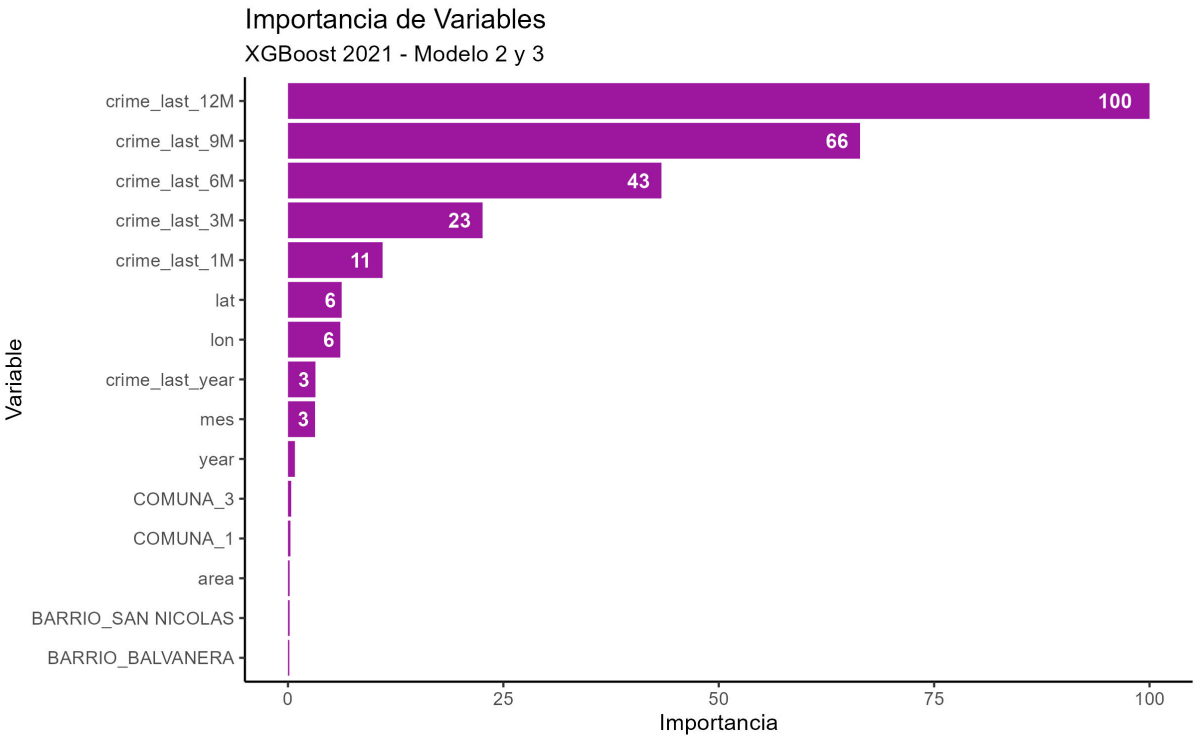


(a) XGBoost

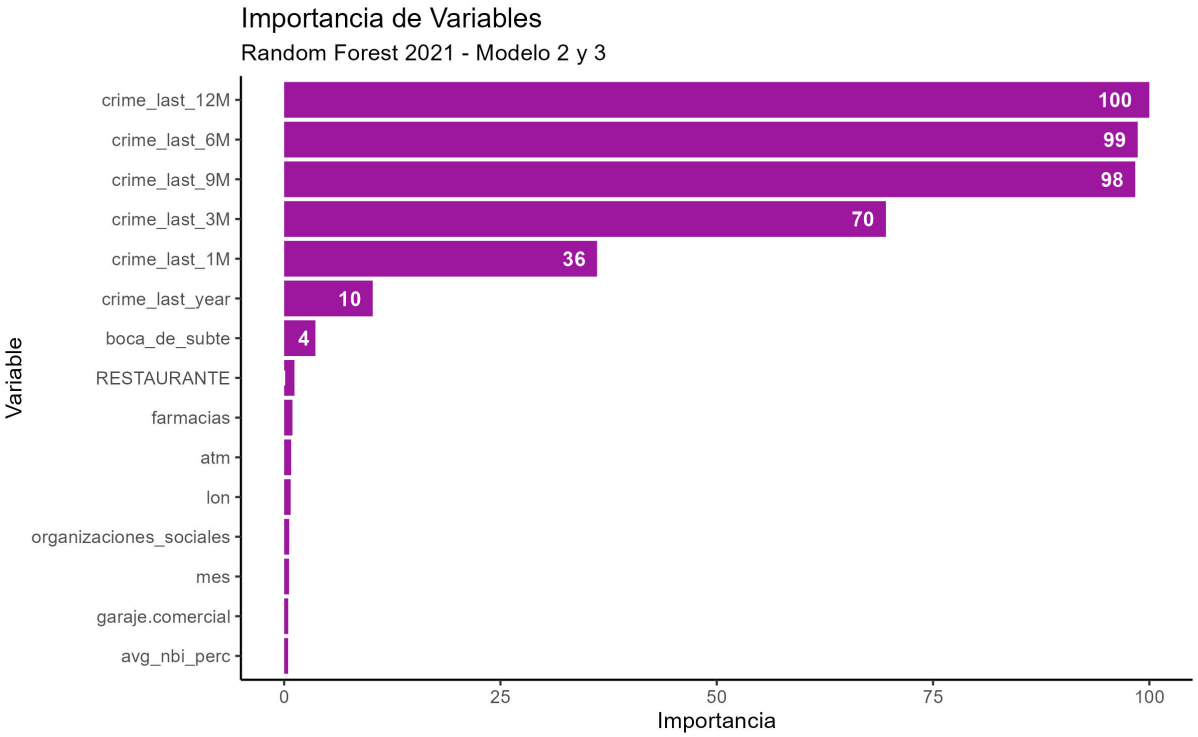


(b) Random Forest

Figura 51. Importancia de variables - Modelo 2 & 3



(a) XGBoost



(b) Random Forest

6 Discusión

6.1 Aplicaciones

El principal objetivo de este trabajo fue demostrar que es posible, a partir de modelos de *machine learning* y un enfoque espacio - temporal basado en una grilla, generar una predicción acertada de los delitos que se darán el mes siguiente en CABA. La correcta utilización de esta herramienta construida con datos completamente abiertos podría permitir alocar de manera más precisa a los patrulleros y al personal de las fuerzas de seguridad de la Ciudad Autónoma de Buenos Aires. De esta manera se lograría aumentar la eficiencia de los recursos disponibles para ubicarlos en donde realmente se necesitan.

Asimismo, cabe destacar que todo el análisis fue realizado con datos enteramente públicos. Por lo que si se pudiera incorporar información privada de la Policía de la Ciudad, posiblemente se podría perfeccionar aún más la *performance* de los modelos. Por ejemplo, se podría sumar la ubicación de las llamadas al 911 que recibe la policía para hurtos y robos. O las zonas con recorridos programados de patrulleros. Incluso, como se menciona al inicio, si las fuerza policiales experimentadas pudieran aportar su *know-how*, también se podrían lograr modelos con una mayor precisión.

Por otro lado, estos resultados también se podrían combinar con algún modelo que permita tomar como input la cantidad de recursos disponibles en cada mes y que, a partir de ello, los asigne de la forma más eficiente en relación al pronóstico de delitos, optimizando los recorridos, barrios y comunas en los que debería encontrarse la fuerza policial para evitar un delito.

6.2 Limitaciones y trabajo futuro

En futuros trabajos relacionados con este se podrían probar más algoritmos de regresión. Incluso, como se menciona al principio, podría ser interesante utilizar modelos más complejos como redes neuronales por ejemplo, para estimar los delitos. Así como también explorar más la cantidad de hiperparámetros para cada modelo y algoritmo, buscando reducir aún más el error. Por otro lado, se podría probar un modelo binario para predecir si en cada celda de la grilla la cantidad de delitos será Baja o Alta (se debería definir un valor a partir del cual ya se podría considerar alta la cantidad). Incluso, se podría generar un modelo binario para predecir las celdas con 0 y las mayores a 0 para luego, en las que el modelo predijo un valor mayor, estimar un modelo de regresión. De esta manera podría evitarse que el modelo, al encontrar muchas celdas con 0 delitos, prediga de manera errónea que no se dará un crimen por el desbalance de clases.

Asimismo, teniendo en cuenta que los delitos registrados subestiman la cifra real, podría ser de mucha ayuda incorporar un índice de victimización[15] para obtener una

[15] "El objetivo del Índice de Victimización (IVI) es cuantificar la tasa de victimización de Argentina. Siguiendo estándares internacionales, se define como tasa de victimización al porcentaje de hogares cuyos miembros convivientes sufrieron al menos un delito en los últimos 12 meses, hayan sido denunciados o no a una autoridad competente."[LICIP, 2023]

cifra más acertada de la verdadera cantidad de delitos, más allá de si los mismos fueron o no registrados como una denuncia. Por ejemplo, se podría incorporar la encuesta de victimización del Laboratorio de Investigaciones sobre Crimen, Instituciones y Políticas (LICIP) de la Universidad Torcuato Di Tella [LICIP, 2023]. Dicha encuesta se realiza de manera mensual para distintas regiones del país, y una de estas es la Ciudad Autónoma de Buenos Aires. Por lo que, en caso de poder obtener la información desagregada de dicha región para los tipos de delitos particulares tratados en esta tesis, se podría utilizar para complementar los delitos registrados en las comisarias de CABA, ya que la misma tiene en cuenta también aquellos delitos que por el motivo que sea no están registrados. En esta misma línea se puede consultar la encuesta de victimización de CABA en el Informe Complementario del Mapa del Delito [Ministerio de Justicia y Seguridad, 2022]. De esta forma, al poder observar la proporción de encuestados que reporta haber denunciado y los que no, se podría generar una variable con la cantidad faltante en los registros que se tienen.

En línea con incorporar variantes, consideraría incluir la cantidad de delitos que se encuentran en las celdas vecinas para los distintos períodos [Lin et al., 2018]. Lo mismo se podría hacer con los puntos de interés, ya que si algún POI posee poder gran predictivo podría influir en la cantidad de delitos de una celda contigua. Por otro lado, también se podrían incorporar las variables para las que se posea información de todos los años de manera anual. Ejemplo de esto último son los datos socioeconómicos, para las cuales solo se ha utilizado el año 2020. Así como también se podrían experimentar distintas combinaciones para la asociación de los grupos de variables.

Además, podría ser relevante considerar variables macroeconómicas, como el índice de inflación, de pobreza, el tipo de cambio y coeficiente de Gini, debido a su potencial impacto en la incidencia de los delitos, especialmente en los casos de robos y hurtos.

Por otra parte, con el objetivo de generar modelos con una menor ventana de tiempo, se podrían estimar los delitos para una semana en lugar de utilizar un mes completo como variable objetivo. Siguiendo en esta línea, incluso se podría dividir el día en 4 franjas horarias, ya que se posee el horario del delito, para generar un modelo que estime los crímenes para cada una [Felson and Poulsen, 2003]. De esta manera, no sería necesario asignar fuerzas policiales para todo el día, sino que podrían rotarse por turnos en relación a la estimación de robos y hurtos. También se podría utilizar este mismo enfoque para estimar otros tipos de delitos, como los homicidios y lesiones que en este trabajo no fueron incorporados. Para esto sería importante tener en cuenta las diferencias que existen en el móvil de estos crímenes, ya que a diferencia de los robos y hurtos, podrían no estar relacionados directamente con características geográficas sino con el contexto [Instituto de Investigaciones del Poder Judicial de la Nación, 2020].

Asimismo, se podría incorporar teoría de juegos para modelar la interacción entre las predicciones de los modelos explicados en esta tesis y la asignación policial designada a partir de esto. Teniendo en cuenta que al incorporar un elemento nuevo, la locación de patrulleros policiales basado en la predicción de los modelos de los crímenes futuros, resulta necesario modelar la interacción entre los distintos "jugadores"para tener un mejor

entendimiento de cómo se van a comportar los mismos. Al entender y modelar estas interacciones entre delincuentes y fuerzas policiales, se puede proporcionar a las autoridades una herramienta valiosa para mejorar la efectividad de sus operaciones y la seguridad pública en general. El uso de la teoría de juegos en este contexto permite a los investigadores y responsables de hacer cumplir la ley analizar y anticipar cómo los delincuentes podrían reaccionar ante ciertas acciones policiales, y viceversa. Esto ayuda a desarrollar enfoques más sofisticados y efectivos para combatir el crimen y reducir su incidencia [Espejo et al., 2016]. De la misma forma, se puede pensar al crimen en las ciudades como una economía abierta en donde se buscará encontrar el equilibrio general a partir de asignar en diferentes ubicaciones a las fuerzas policiales [Galiani et al., 2018].

Para finalizar, es relevante considerar enfoques que se basan en la interpretación de modelos de aprendizaje automático en la predicción de crímenes [Zhang et al., 2022], como SHAP (SHapley Additive exPlanations) [Scott Lundberg, 2018]. A partir de esto, se podrá entender qué características dentro de cada celda de la grilla contribuyen a que la misma tenga como resultado una mayor cantidad de sustracciones predicha (robos y hurtos). Al utilizar algún enfoque de este estilo, se podrá profundizar en cuestiones importantes relacionadas con la estigmatización de ciertos barrios de CABA, ya que como se ha analizado, los modelos pueden estimar una cantidad de delitos superior en algunas áreas en comparación con otras. Este enfoque resulta valioso, ya que la relación entre desigualdad y delitos es un factor que ha sido probada y estudiada a lo largo del tiempo. Se ha demostrado como factores socioeconómicos pueden interactuar con otros determinantes para influir en la aparición y persistencia del delito en la sociedad. [Schargrodsky and Freira, 2022]

6.3 Conclusión

En función de los resultados obtenidos y de los análisis generados, es posible afirmar que algoritmos de *machine learning* basados en grilla pueden estimar la cantidad de delitos que se darán en cada celda para un mes en el futuro. Además, podemos concluir que no solo las variables de los delitos pasados tienen un impacto en la performance de los mismos, sino que también las variables de entorno, climáticas y socioeconómicas generan mejoras en los modelos.

En el caso de los resultados, si se hubieran mirado solo las métricas de error, la conclusión habría sido sin dudas que los modelos de *machine learning* tenían mejor poder predictivo. Sin embargo, luego, al haber observado las distribuciones de los valores predichos y la comparación de algoritmos de valores observados versus los predichos, la conclusión habría sido que el modelo *Naïve* posee la mejor *performance*. Por último, teniendo en cuenta que es un modelo que busca predecir geo espacialmente, al observar los mapas de las celdas de la grilla, barrios y comunas se puede concluir que nuevamente los modelos de aprendizaje automático son los que más se destacan.

Por otro lado, fue importante haber analizado los tres modelos, ya que a pesar de que todos tuvieron un RMSLE similar en evaluación, el Modelo 3, entrenado con datos hasta noviembre de 2019 y evaluado en diciembre 2021, se destacó en promedio de error para

el análisis hecho a nivel comunas y barrios. Indicando que el impacto de la baja en la cantidad de delitos por la pandemia, parecería tener un efecto en los resultados.

Referencias

[Appiolaza, 2010] Appiolaza, M. (2010). Prevención social de la violencia y el delito en la argentina. *PiPP (Plataforma de información para Políticas Públicas)*.

[Argentina.gob.ar, 2020] Argentina.gob.ar (2020). Ciudad Autónoma de Buenos Aires. `https://www.argentina.gob.ar/caba`.

[Bogomolov et al., 2014] Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., and Pentland, A. (2014). Once upon a crime: Towards crime prediction from demographics and mobile data.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

[Buenos Aires Ciudad, 2023] Buenos Aires Ciudad (Acceso en 2023). Ciudad de Buenos Aires. `https://www.buenosaires.gob.ar/laciudad/ciudad`.

[Chainey et al., 2008] Chainey, S., Tompson, L., and Uhlig, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. *Security journal*, 21(1):4–28.

[Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

[Cichosz, 2020] Cichosz, P. (2020). Urban crime risk prediction using point of interest data. *ISPRS International Journal of Geo-Information*, 9(7).

[CRAN, 2021] CRAN (2021). st_make_grid sf - simple features for r. `https://search.r-project.org/CRAN/refmans/sf/html/st_make_grid.html`.

[Cravino, 2016] Cravino, M. C. (2016). Desigualdad urbana, inseguridad y vida cotidiana en asentamientos informales del area metropolitana de buenos aires. *Etnografías Contemporáneas*, 2 (3):56–83.

[Dalla Via Monti, 2020] Dalla Via Monti, J. (2020). A machine learning approach for prediction of hospital bed availability. Master's thesis, Master in Management+Analytics Thesis, Universidad Torcuato Di Tella.

[DarkSky, 2023] DarkSky (Acceso en 2023). Weather API. `https://darksky.net`.

[Dash et al., 2018] Dash, S. K., Safro, I., and Srinivasamurthy, R. S. (2018). Spatio-temporal prediction of crimes using network analytic approach. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1912–1917.

[Espejo et al., 2016] Espejo, G., L'Huillier, G., and Weber, R. (2016). A game-theoretical approach for policing decision support. *European Journal of Applied Mathematics*, 27(3):338–356.

[Felson and Poulsen, 2003] Felson, M. and Poulsen, E. (2003). Simple indicators of crime by time of day. *International Journal of Forecasting*, 19(4):595–601.

[Galiani et al., 2018] Galiani, S., Lopez Cruz, I., and Torrens, G. (2018). Stirring up a hornets' nest: Geographic distribution of crime. *Journal of Economic Behavior Organization*, 152:17–35.

[Geofabrik GmbH & OSM Contributors, 2018] Geofabrik GmbH & OSM Contributors (2018). Download OpenStreetMap data for this region: Argentina. `https://download.geofabrik.de/south-america/argentina.html/`.

[Gobierno de la Ciudad Autónoma de Buenos Aires, 2012] Gobierno de la Ciudad Autónoma de Buenos Aires (2012). BA Data. `https://data.buenosaires.gob.ar/`.

[Instituto de Investigaciones del Poder Judicial de la Nación, 2020] Instituto de Investigaciones del Poder Judicial de la Nación (2020). Informe sobre Homicidios CABA. Consejo de la Magistratura Poder Judicial de la Nación.

[James et al., 2021] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer US.

[Kuhn, 2008] Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software, Articles*, 28(5):1–26.

[Kuhn and Johnson, 2013] Kuhn, M. and Johnson, K. (2013). *Applied predictive modeling*. Springer, New York, NY.

[LICIP, 2023] LICIP (2023). Índice de Victimización Mayo 2023. *Laboratorio de Investigaciones sobre Crimen, Instituciones y Políticas (LICIP), Universidad Torcuato Di Tella*, page 2.

[Lin et al., 2017] Lin, Y.-L., Chen, T.-Y., and Yu, L. (2017). Using machine learning to assist crime prevention. *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 1029–1030.

[Lin et al., 2018] Lin, Y.-L., Yen, M.-F., and Yu, L.-C. (2018). Grid-based crime prediction using geographical features. *ISPRS International Journal of Geo-Information*, 7(8).

[Ministerio de Justicia y Seguridad, 2022] Ministerio de Justicia y Seguridad (2022). Informe de Estadística Criminal 2021. pages 201–326.

[Montane, 2020] Montane, M. (2020). *Ciencia de Datos para curiosos - Capítulo 4: Datos espaciales en R*.

[OpenStreetMap Contributors, 2017] OpenStreetMap Contributors (2017). Planet dump retrieved from https://planet.osm.org . `https://www.openstreetmap.org`.

[Pebesma, 2018] Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439–446.

[Psyllidis et al., 2022] Psyllidis, A., Gao, S., Hu, Y., Kim, E.-K., McKenzie, G., Purves, R., Yuan, M., and Andris, C. (2022). Points of interest (poi): a commentary on the state of the art, challenges, and prospects for the future. *Computational Urban Science*, 2.

[Ratcliffe, 2004] Ratcliffe, J. H. (2004). Geocoding crime and a first estimate of a minimum acceptable hit rate. *International Journal of Geographical Information Science*, 18(1):61–72.

[Schargrodsky and Freira, 2022] Schargrodsky, E. and Freira, L. (2022). Inequality and crime in latin america and the caribbean: New data for an old question. *Economia LACEA*.

[Scott Lundberg, 2018] Scott Lundberg (2018). SHAP Latest Documentation . `https://shap.readthedocs.io/en/latest/index.html`.

[Van Dijk, 1990] Van Dijk, J. (1990). Crime prevention policy: current state and prospects. *G. Kaiser y HJ Albrecht, Crime and Criminal Policy in Europe: Criminological Research Report*, 43:205–220.

[Yuki et al., 2019] Yuki, J., Sakib, M. M., Zamal, Z., Habibullah, K., and Das, A. (2019). Predicting crime using time and location data. pages 124–128.

[Zambrano, 2021] Zambrano, R. (2021). Un enfoque espaciotemporal para la predicción de delitos en la ciudad de buenos aires. *Revista de Investigación en Modelos Matemáticos Aplicados a la Gestión y la Economía - Año 7*, II.

[Zhang et al., 2022] Zhang, X., Liu, L., Lan, M., Song, G., Xiao, L., and Chen, J. (2022). Interpretable machine learning models for crime prediction. *Computers, Environment and Urban Systems*, 94:101789.

[Zheng and Casari, 2018] Zheng, A. and Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media, Inc., Sebastopol, CA, first edition.

Apéndice A

Puntos de Interés

BA Data

Listado de puntos de interés:

- Bar
- Banco
- Basílica
- Biblioteca
- Boca de subte
- Café
- Cajero automático
- Cartel luminoso
- Centro de salud de acción comunitaria
- Centro de salud privado
- Centro médico barrial
- Club
- Comisaria
- Consulado
- Cuartel/Destacamento de bomberos
- Embajada
- Establecimiento educativo
- Estación de ferrocarril
- Estación de metro bus
- Estación de servicio
- Estadio
- Farmacia
- Garaje comercial

- Gimnasio
- Hospital
- Hotel de alta categoría
- Hotel económico
- Iglesia
- Librería
- Local bailable
- Organización social
- Parada de bus turístico
- Parada de taxi
- Parroquia
- Pista de skate
- Premetro
- Restaurante
- Sala de teatro
- Universidad

OpenStreetMap

Listado de puntos de interés:

- Almacén/Supermercado pequeño
- Arte (estatuas, murales)
- Atracciones
- Cámaras de vigilancia
- Centro de deportes
- Local de comidas rápidas
- Local de indumentaria
- Monumentos
- Museo

- Peluquería
- Shopping
- Super e Hipermercado

Regiones de Interés

OpenStreetMap

Listado de regiones de interés:

- Atracciones
- Campo de golf
- Cancha de deportes
- Cementerio
- Centros deportivos
- Cine
- Estadio
- Memorial
- Zoológico

Variables

Listado de variables:

- area
- artwork
- atm
- attraction
- avg_depto_dolarm2
- avg_nbi_perc
- avg_terreno_dolarm2
- banco
- bar

- barriopopular
- BARRIO_AGRONOMIA
- BARRIO_ALMAGRO
- BARRIO_BALVANERA
- BARRIO_BARRACAS
- BARRIO_BELGRANO
- BARRIO_BOCA
- BARRIO_BOEDO
- BARRIO_CABALLITO
- BARRIO_CHACARITA
- BARRIO_COGHLAN
- BARRIO_COLEGIALES
- BARRIO_CONSTITUCION
- BARRIO_FLORES
- BARRIO_FLORESTA
- BARRIO_LINIERS
- BARRIO_MATADEROS
- BARRIO_MONSERRAT
- BARRIO_MONTE CASTRO
- BARRIO_NUEVA POMPEYA
- BARRIO_NUÑEZ
- BARRIO_PALERMO
- BARRIO_PARQUE AVELLANEDA
- BARRIO_PARQUE CHACABUCO
- BARRIO_PARQUE CHAS
- BARRIO_PARQUE PATRICIOS
- BARRIO_PATERNAL

- BARRIO_PUERTO MADERO
- BARRIO_RECOLETA
- BARRIO_RETIRO
- BARRIO_SAAVEDRA
- BARRIO_SAN CRISTOBAL
- BARRIO_SAN NICOLAS
- BARRIO_SAN TELMO
- BARRIO_VELEZ SARSFIELD
- BARRIO_VERSALLES
- BARRIO_VILLA CRESPO
- BARRIO_VILLA DEL PARQUE
- BARRIO_VILLA DEVOTO
- BARRIO_VILLA GRAL MITRE
- BARRIO_VILLA LUGANO
- BARRIO_VILLA LURO
- BARRIO_VILLA ORTUZAR
- BARRIO_VILLA PUEYRREDON
- BARRIO_VILLA REAL
- BARRIO_VILLA RIACHUELO
- BARRIO_VILLA SANTA RITA
- BARRIO_VILLA SOLDATI
- BARRIO_VILLA URQUIZA
- basilica
- biblioteca
- boca_de_subte
- boliche
- cafe

- camera_surveillance
- cartel_luminoso
- centro_de_salud_comunitario
- centro_medico_barrial
- centro_de_salud_privado
- cinema
- clothes
- club
- comisaria
- COMUNA_1
- COMUNA_10
- COMUNA_11
- COMUNA_12
- COMUNA_13
- COMUNA_14
- COMUNA_15
- COMUNA_2
- COMUNA_3
- COMUNA_4
- COMUNA_5
- COMUNA_6
- COMUNA_7
- COMUNA_8
- COMUNA_9
- consulado
- convenience
- crime_last_12M

- crime_last_1M
- crime_last_3M
- crime_last_6M
- crime_last_9M
- crime_last_year
- cuartel_de_bomberos
- distancia_autopista
- distancia_avenida
- distancia_barrio_popular
- distancia_bomberos
- distancia_comisaria
- distancia_hospital
- distancia_tunel
- embajada
- espacio_verde
- establecimiento_educativo
- estacion_de_servicio
- estaciones_ferrocarril
- estadios
- farmacias
- fast_food
- garaje_comercial
- gimnasios
- golf_course
- graveyard
- hairdresser
- hospital

- hotelȧlta
- hotel_baja
- iglesia
- lat
- libreria
- lon
- mall
- memorial
- mes
- monument
- museum
- organizaciones_sociales
- paradaḃustṫuristico
- paradaṁetrobus
- paradaṫaxi
- parroquia
- pistaḋeṡkate
- pitch
- precipitaciones_last_12M
- precipitaciones_last_1M
- precipitaciones_last_3M
- precipitaciones_last_6M
- precipitaciones_last_9M
- precipitaciones_last_year
- premetro
- restaurante
- sports_centre

- stadium
- supermarket
- teatro
- temperatura_last_12M
- temperatura_last_1M
- temperatura_last_3M
- temperatura_last_6M
- temperatura_last_9M
- temperatura_last_year
- universidad
- viento_last_12M
- viento_last_1M
- viento_last_3M
- viento_last_6M
- viento_last_9M
- viento_last_year
- year
- zoo

Apéndice B

Construcción de una grilla: paso a paso

En primer lugar, se debió cargar el archivo *.shp* correspondiente a la superficie de la Ciudad Autónoma de Buenos Aires. En este caso, el mismo fue obtenido a partir del portal de BA Data. Luego, es necesario definir el tamaño que tendrá cada cuadrado de la grilla. En este caso, dado que la unidad del área de los datos de CABA está en m^2, los lados de los cuadrados de la grilla se deben definir en metros. Una vez definida la cantidad de m^2 de cada celda de la grilla, se debe generar un cuadro delimitador o *bounding box* (usualmente llamados solamente *bbox*). Esta es un área definida por dos longitudes y dos latitudes, como se mencionó anteriormente. En este caso, el *bbox* comienza en la parte inferior izquierda (suroeste), donde la línea vertical y la línea horizontal son tangentes al límite de la ciudad izquierdo e inferior, respectivamente. Continuando hasta llegar a la esquina superior derecha (noreste), que no necesariamente debe ser tangente a los límites de la ciudad, ya que se debe completar el tamaño del último cuadrado. A partir de este cuadro delimitador, se generó la grilla de forma tal que lo cubra por completo. Luego, se le asignó a cada cuadrado un ID que permitirá identificarlo. Como último paso, se intersecó la superficie de la ciudad, y se eliminaron aquellas celdas de la cuadrícula que no se encuentran en el territorio de CABA. Es decir, aquellas que no contienen territorio en ellas. Si bien la mayoría de las celdas de la grilla son cuadrados, si la celda se encuentra en el límite de la ciudad, se definió la misma como la intersección del cuadrado y el interior de la ciudad. Por lo tanto, puede haber celdas que sean polígonos e incluso algunas que estén desconectadas del resto, dado que está el río en el medio.

Apéndice C

Figura 52. Cantidad de delitos por barrio 2021

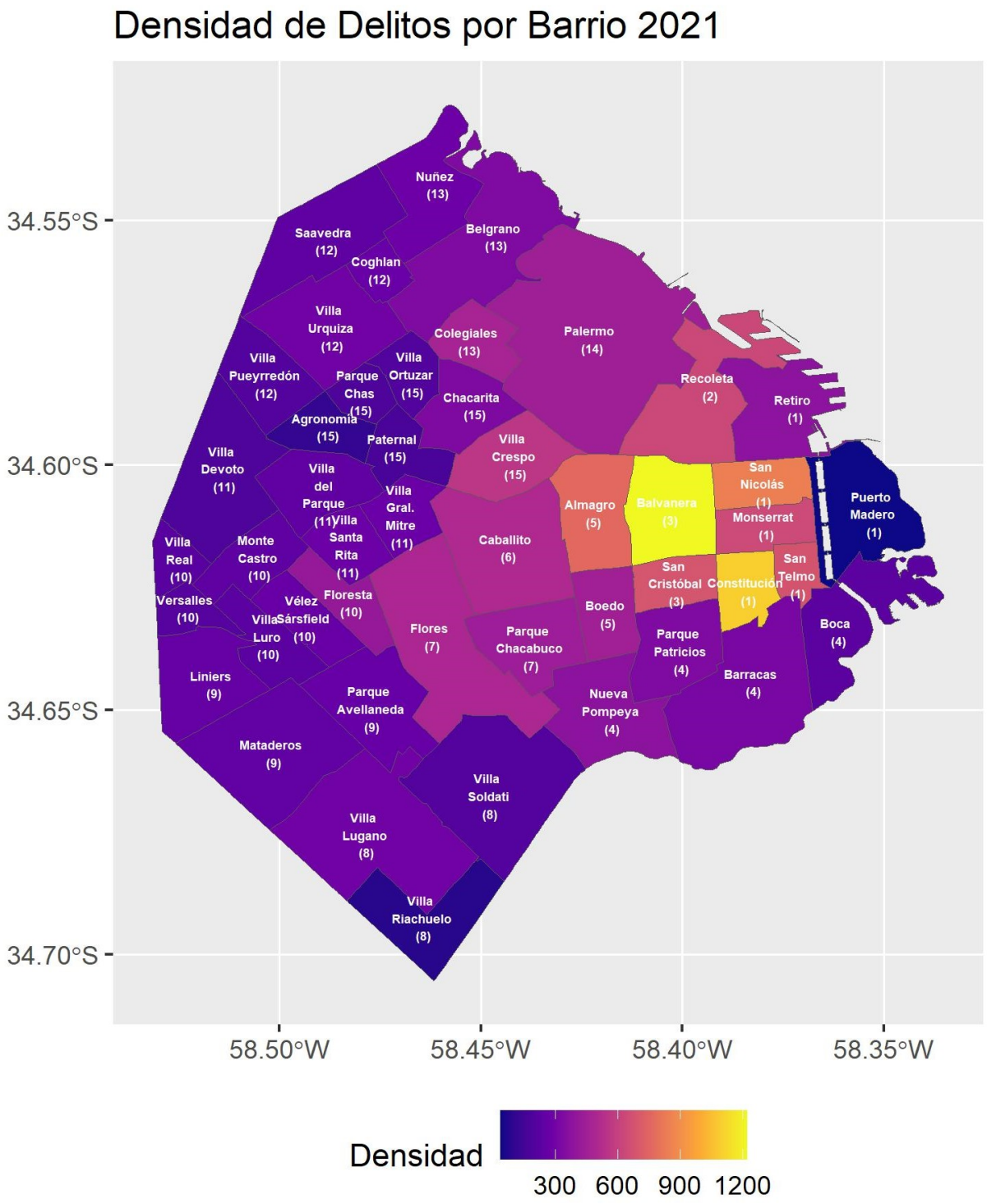


Figura 53. Precio promedio de los terrenos por m^2 en dólares por Barrio

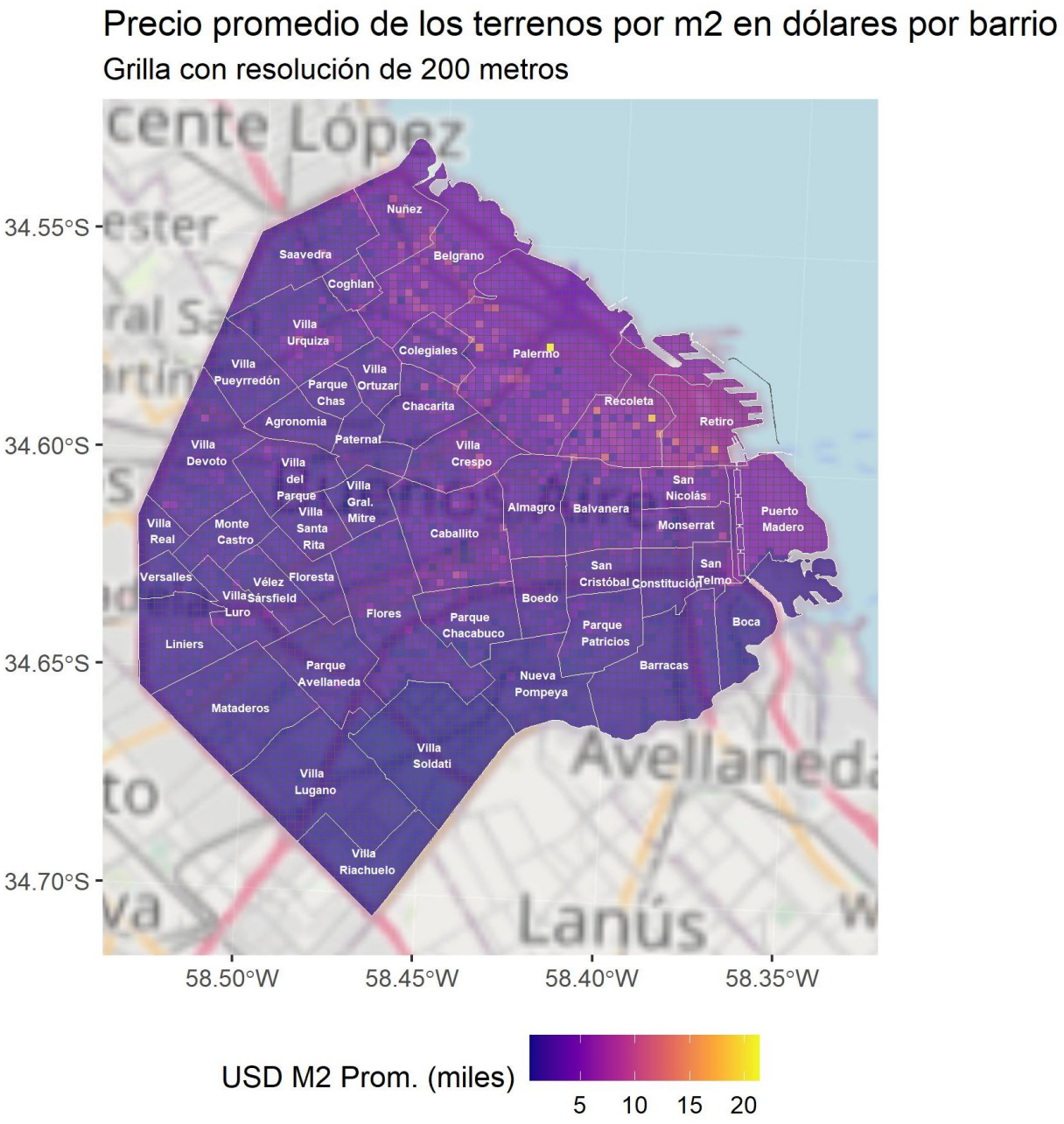


Figura 54. Precio promedio de los departamentos por m^2 en dólares por Barrio

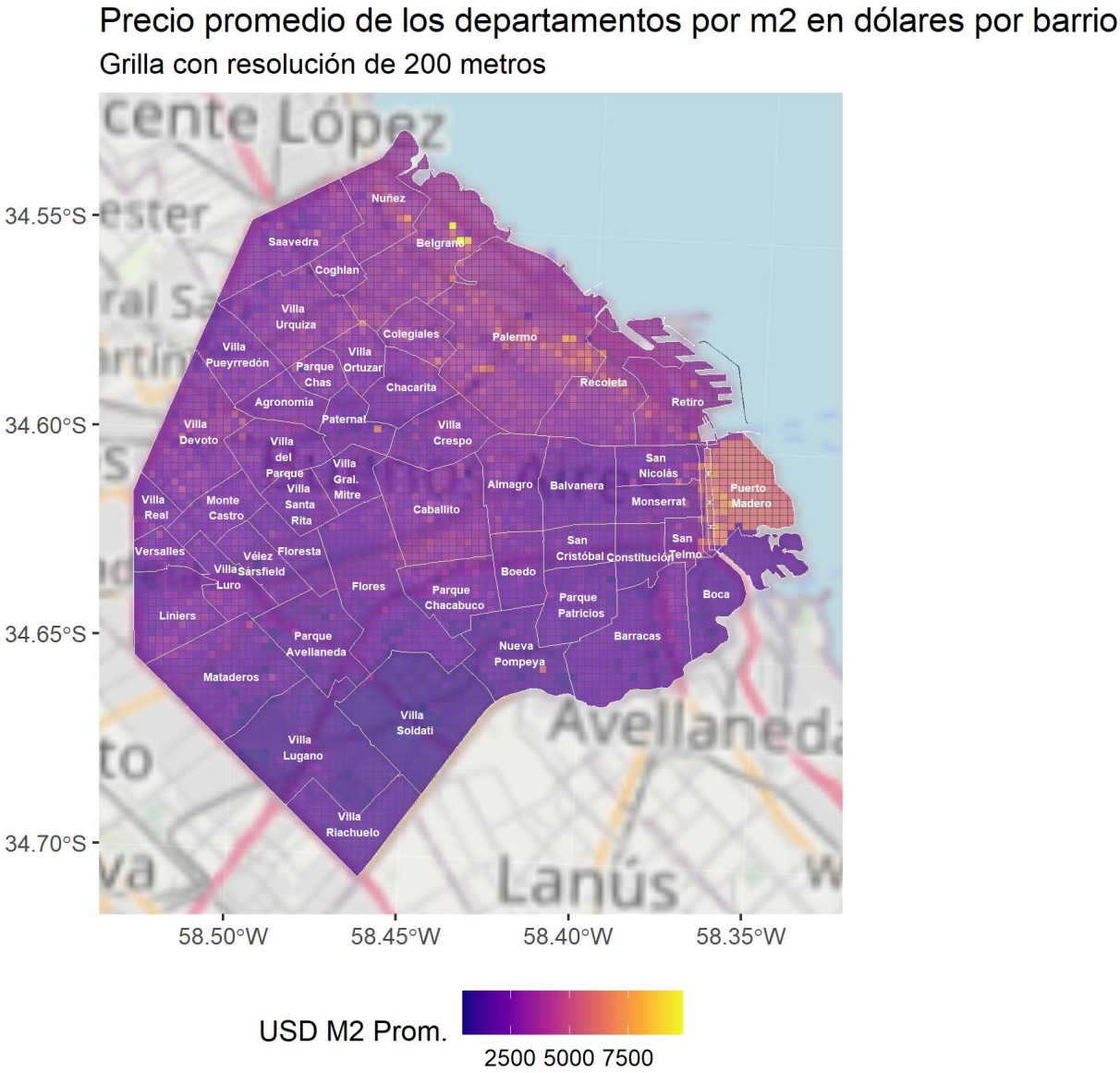


Figura 55. Porcentaje de Necesidades Básicas Insatisfechas por Barrio

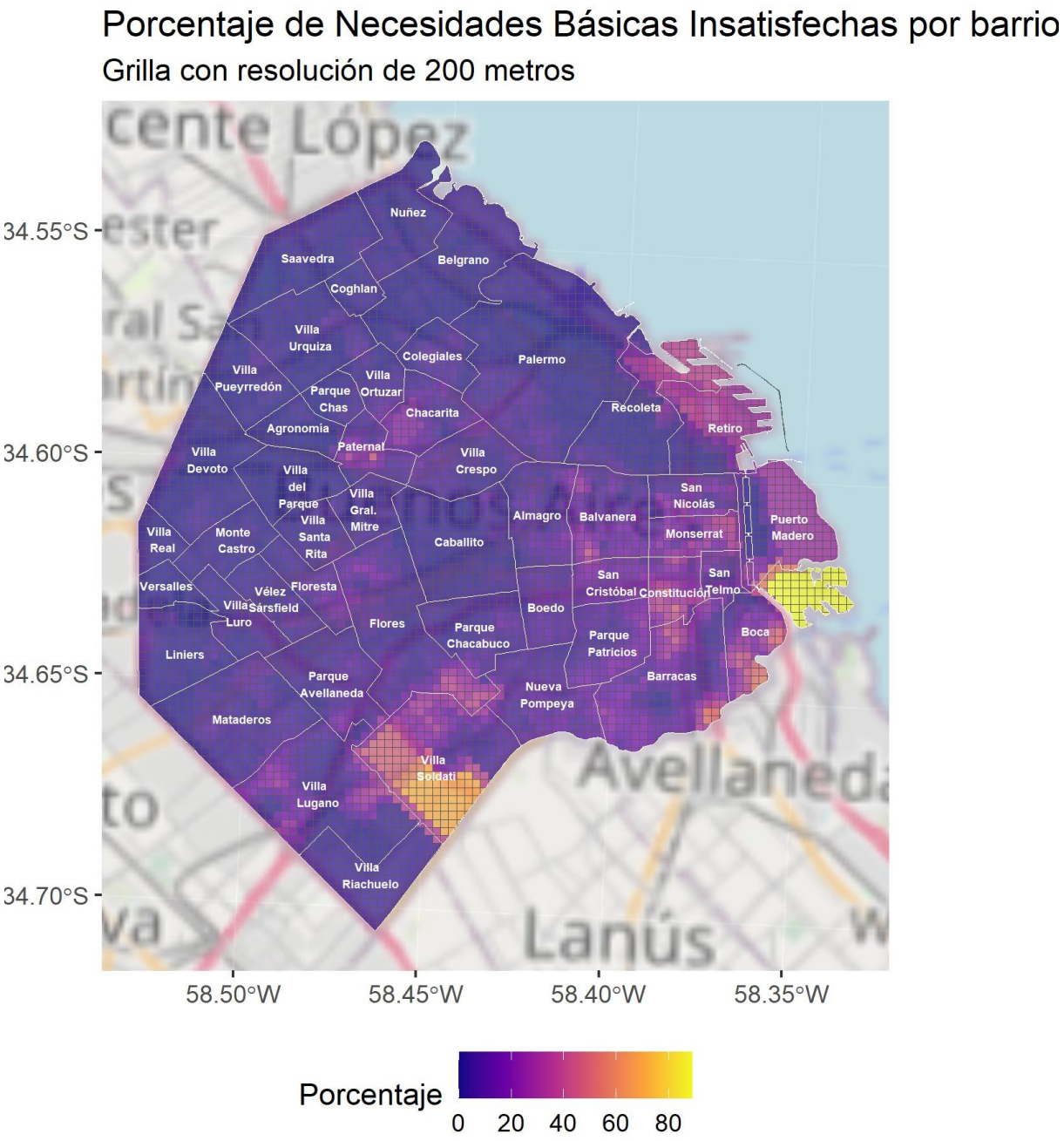


Figura 56. Correlaciones Espaciales

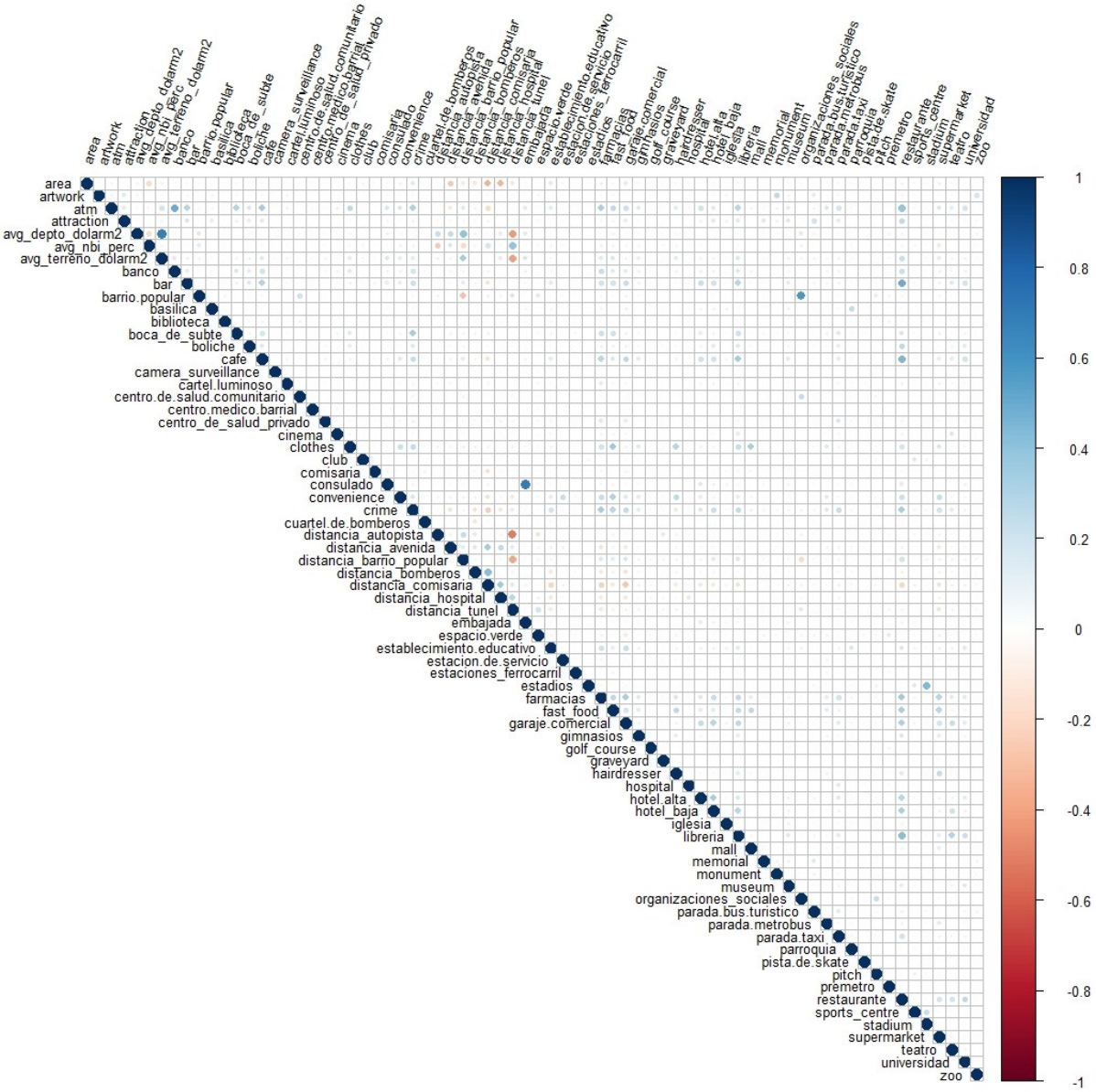
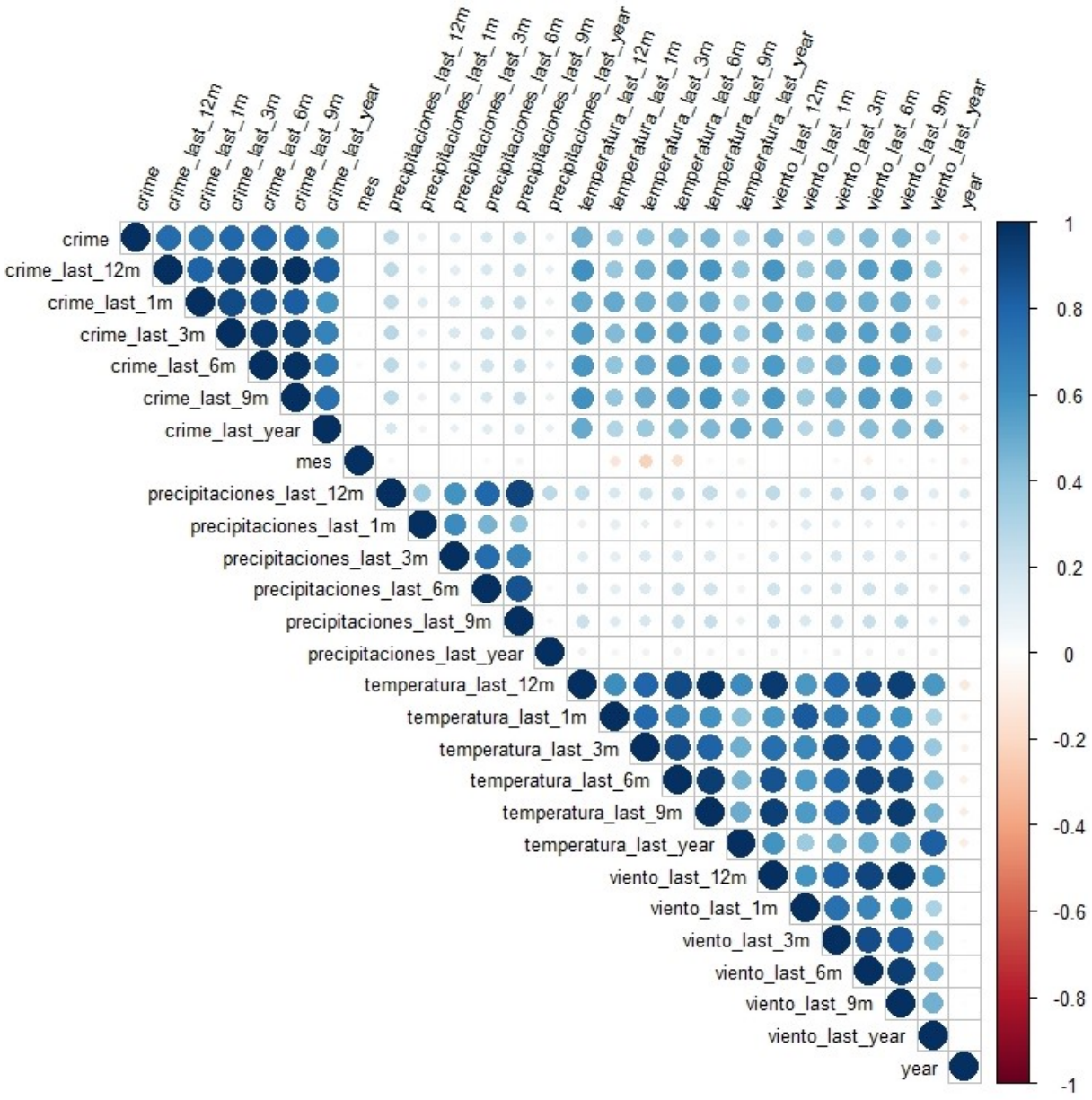


Figura 57. Correlaciones Temporales



Mapas error promedio por barrio

Figura 58. Error promedio por barrio - Naïve 2021 Model 1

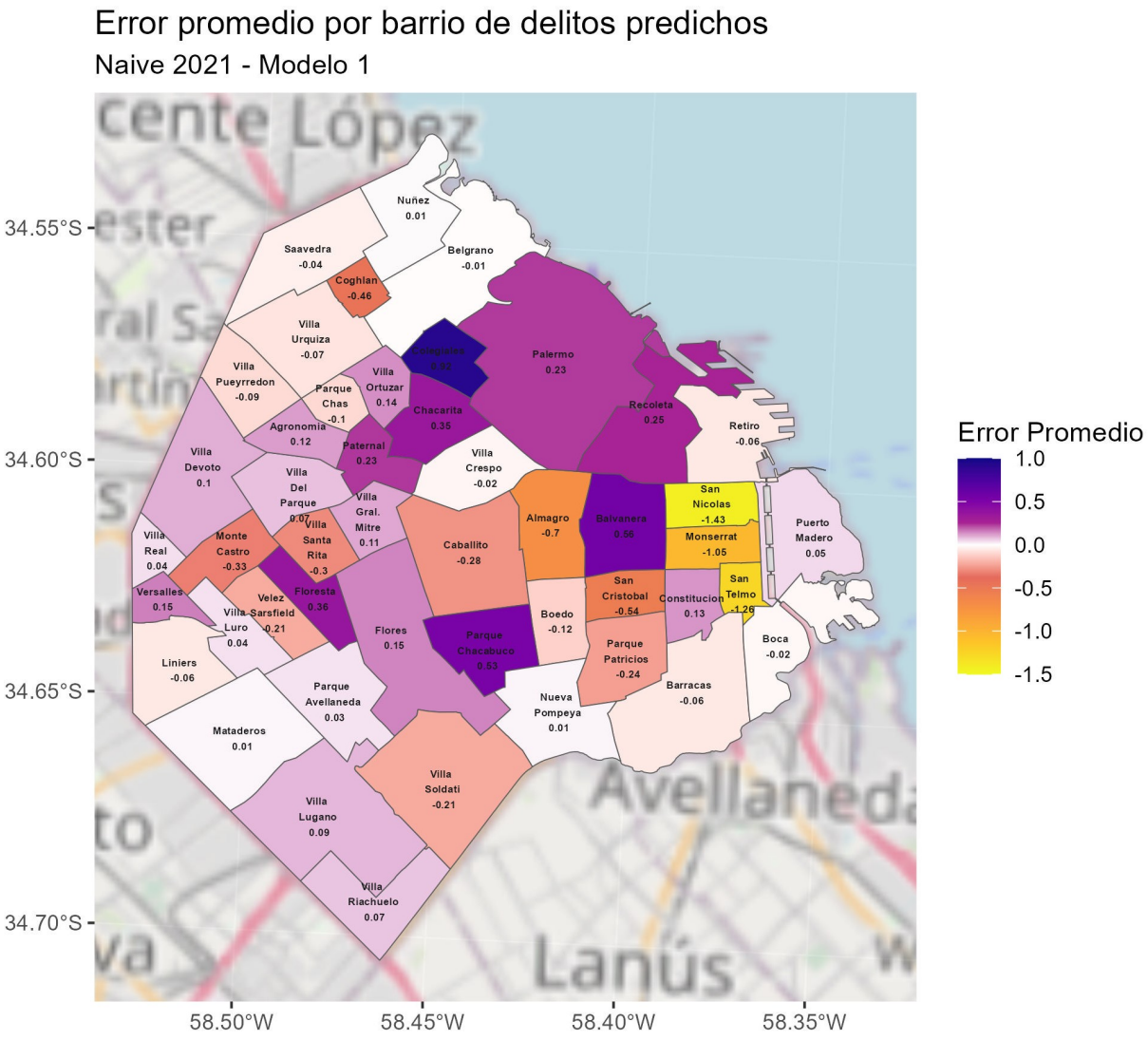


Figura 59. Error promedio por barrio - XGBoost 2021 Model 1

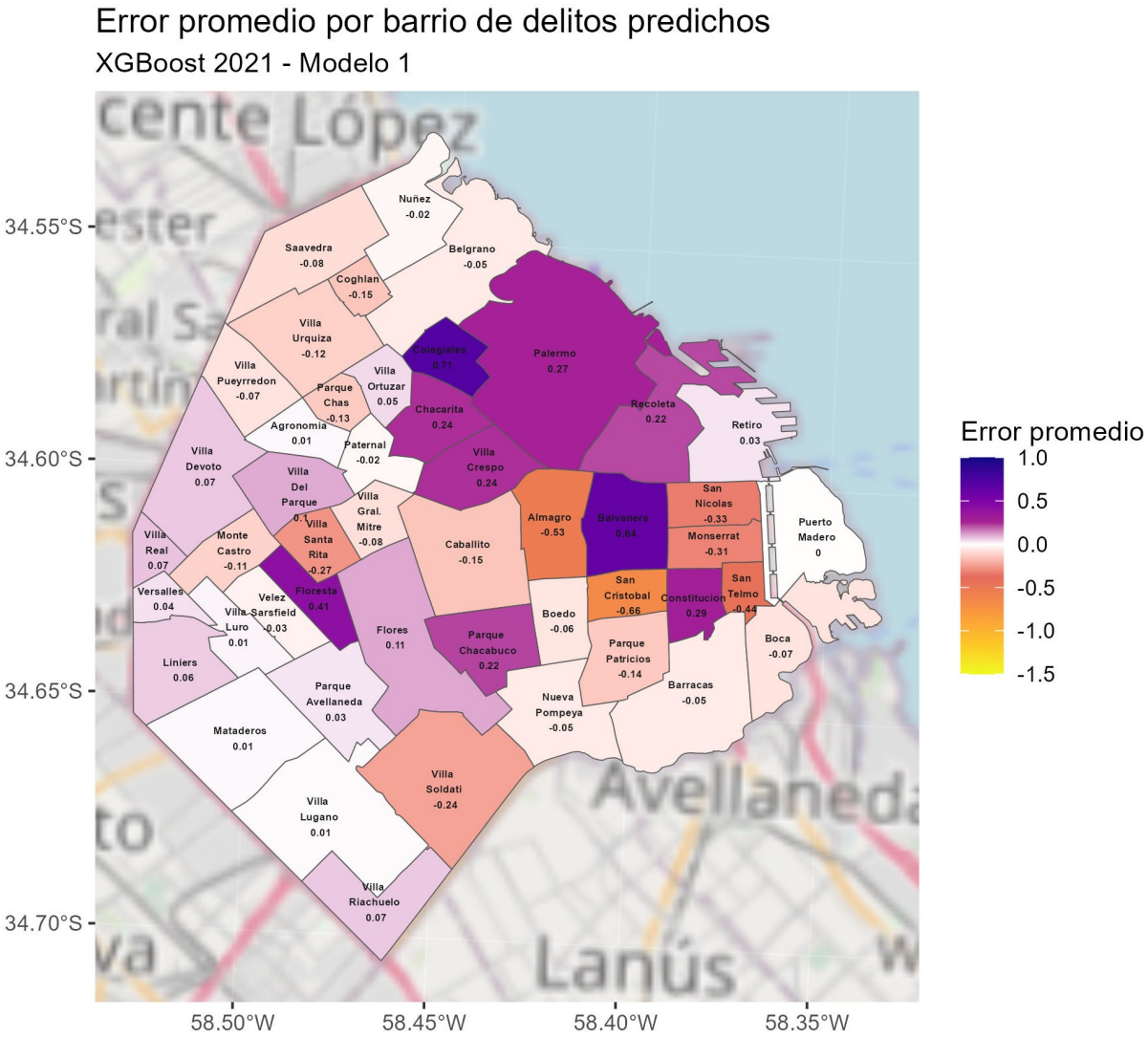


Figura 60. Error promedio por barrio - Random Forest 2021 Model 1

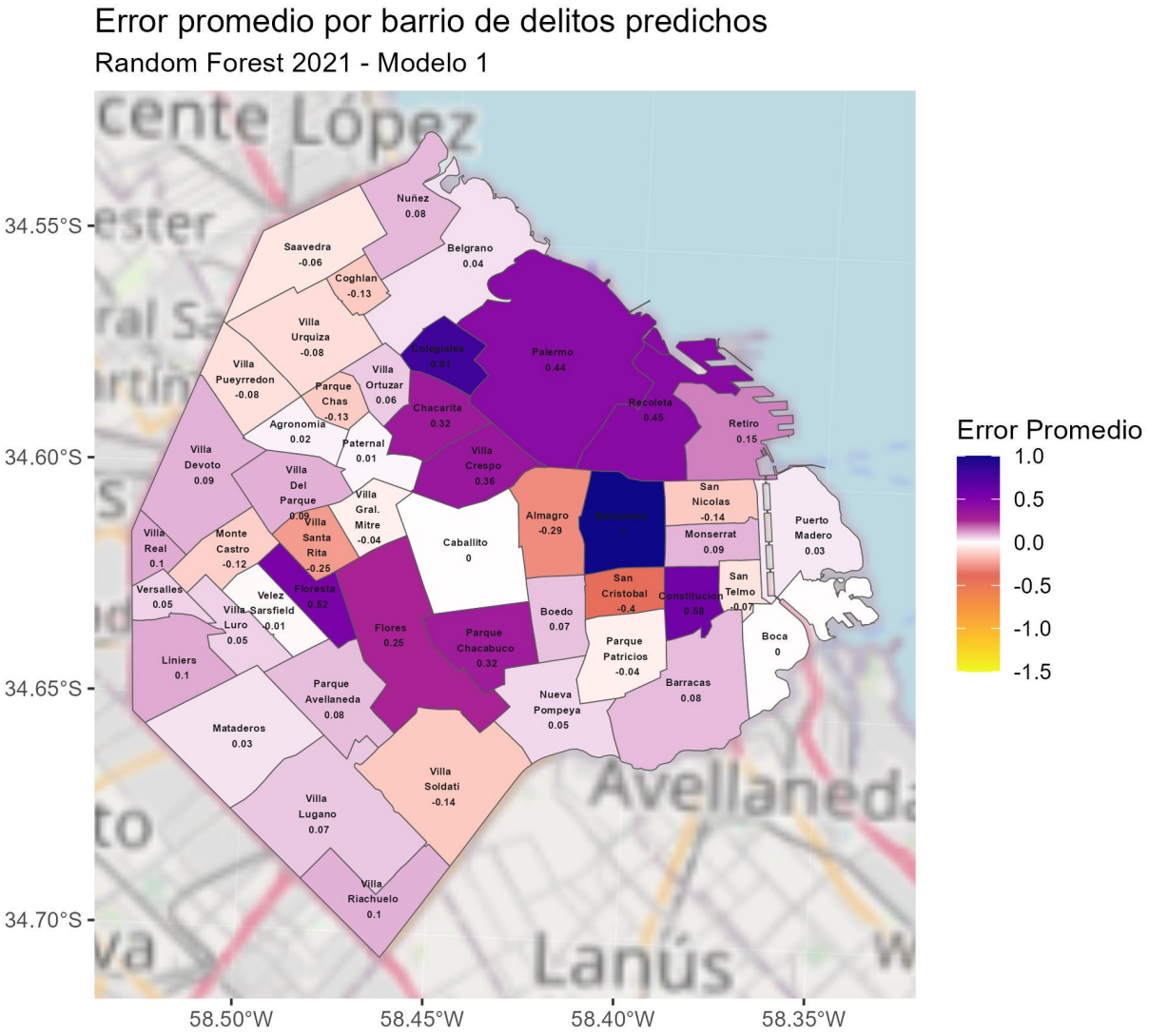


Figura 61. Error promedio por barrio - Naïve 2019 Model 2

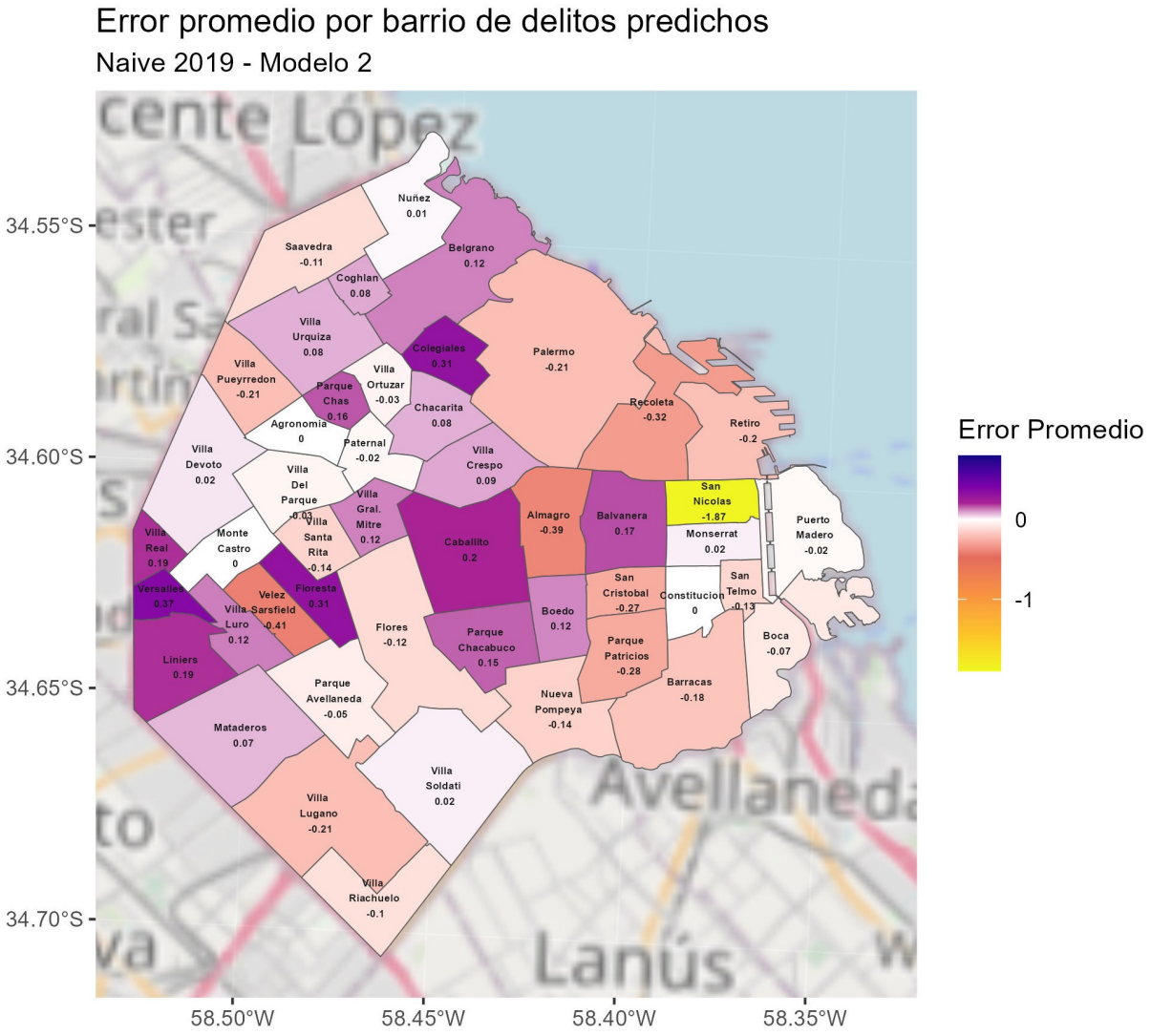


Figura 62. Error promedio por barrio - XGBoost 2019 Model 2

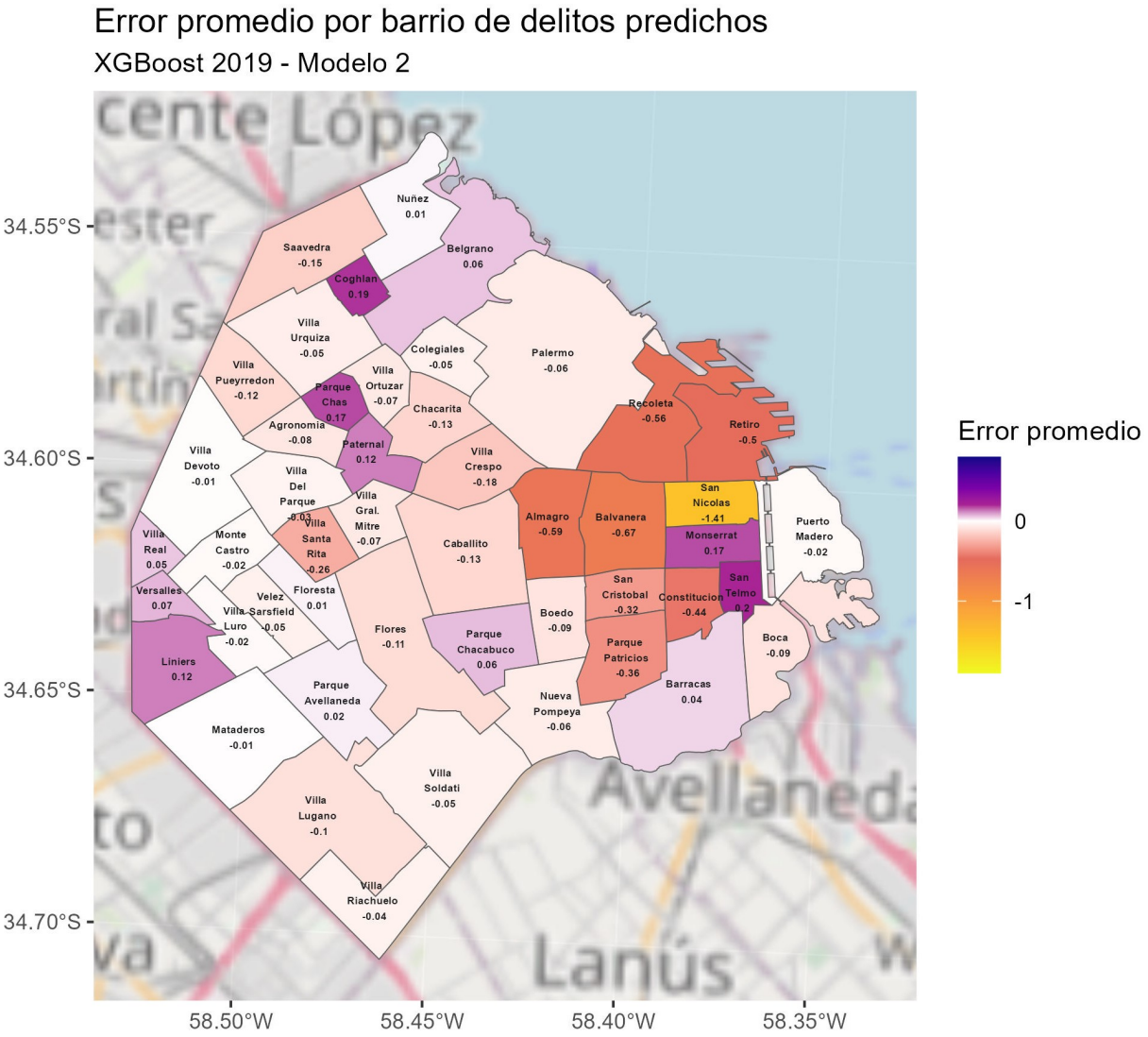


Figura 63. Error promedio por barrio - Random Forest 2019 Model 2

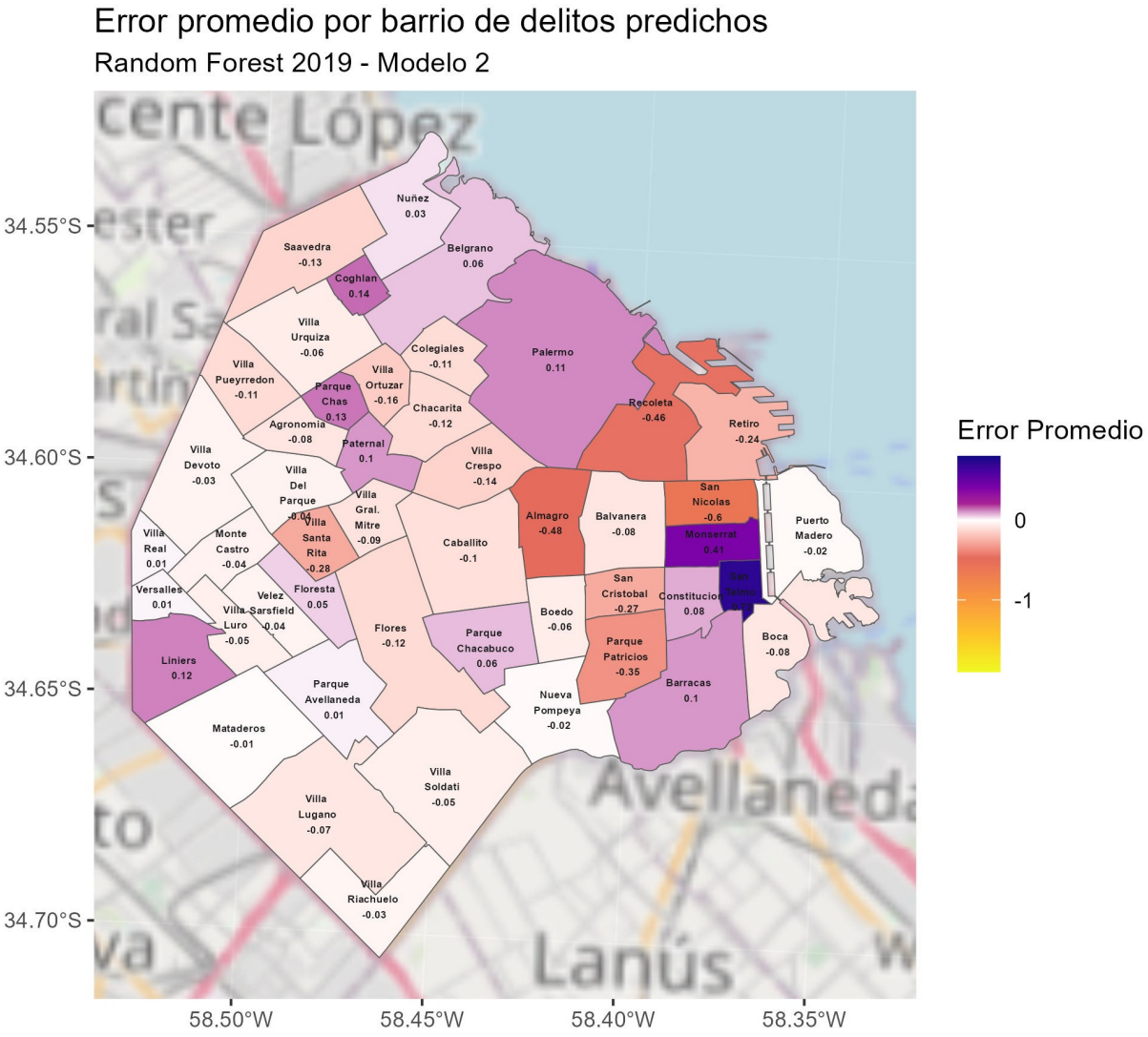


Figura 64. Error promedio por barrio - Naïve 2021 Model 3

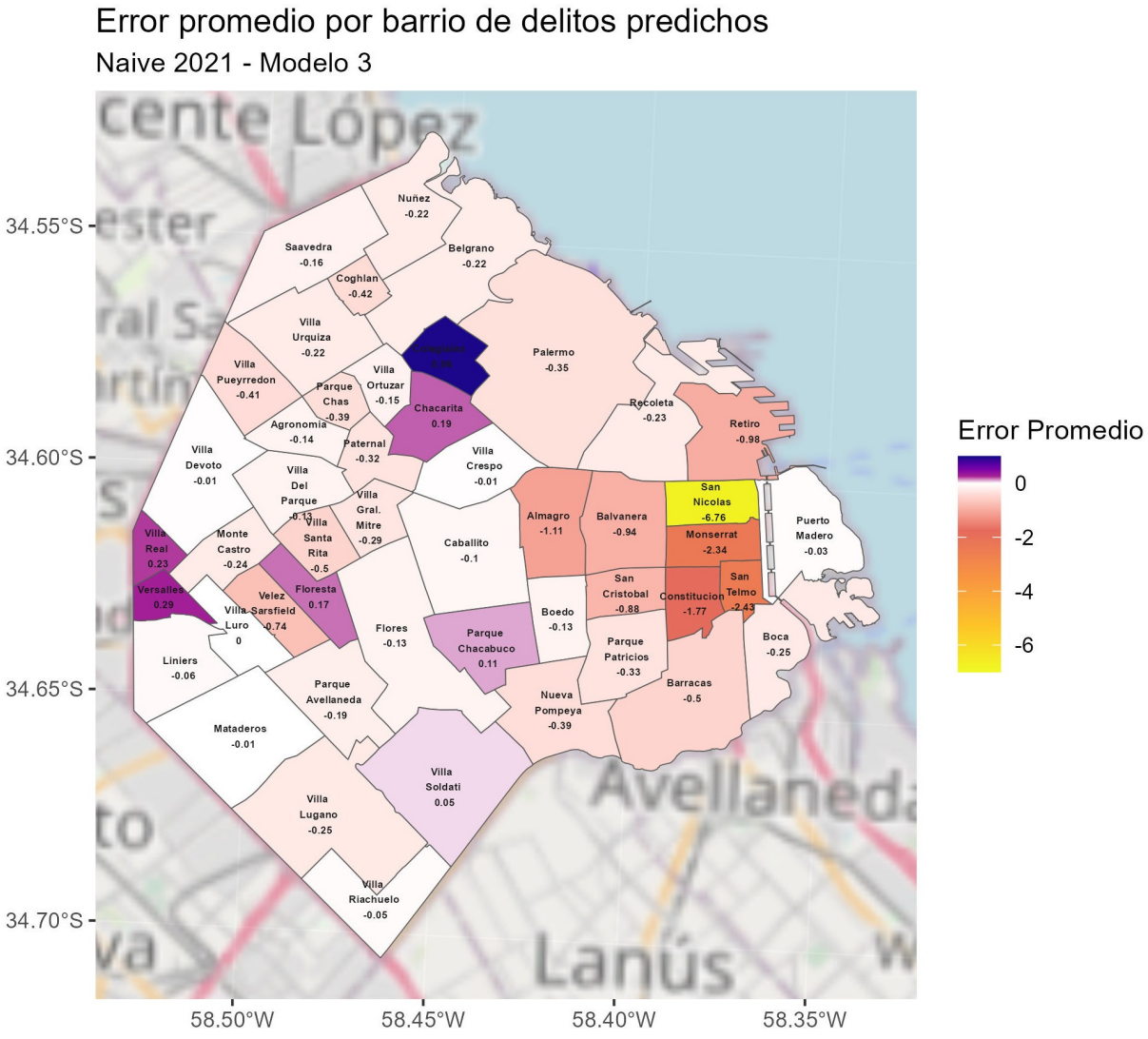


Figura 65. Error promedio por barrio - XGBoost 2021 Model 3

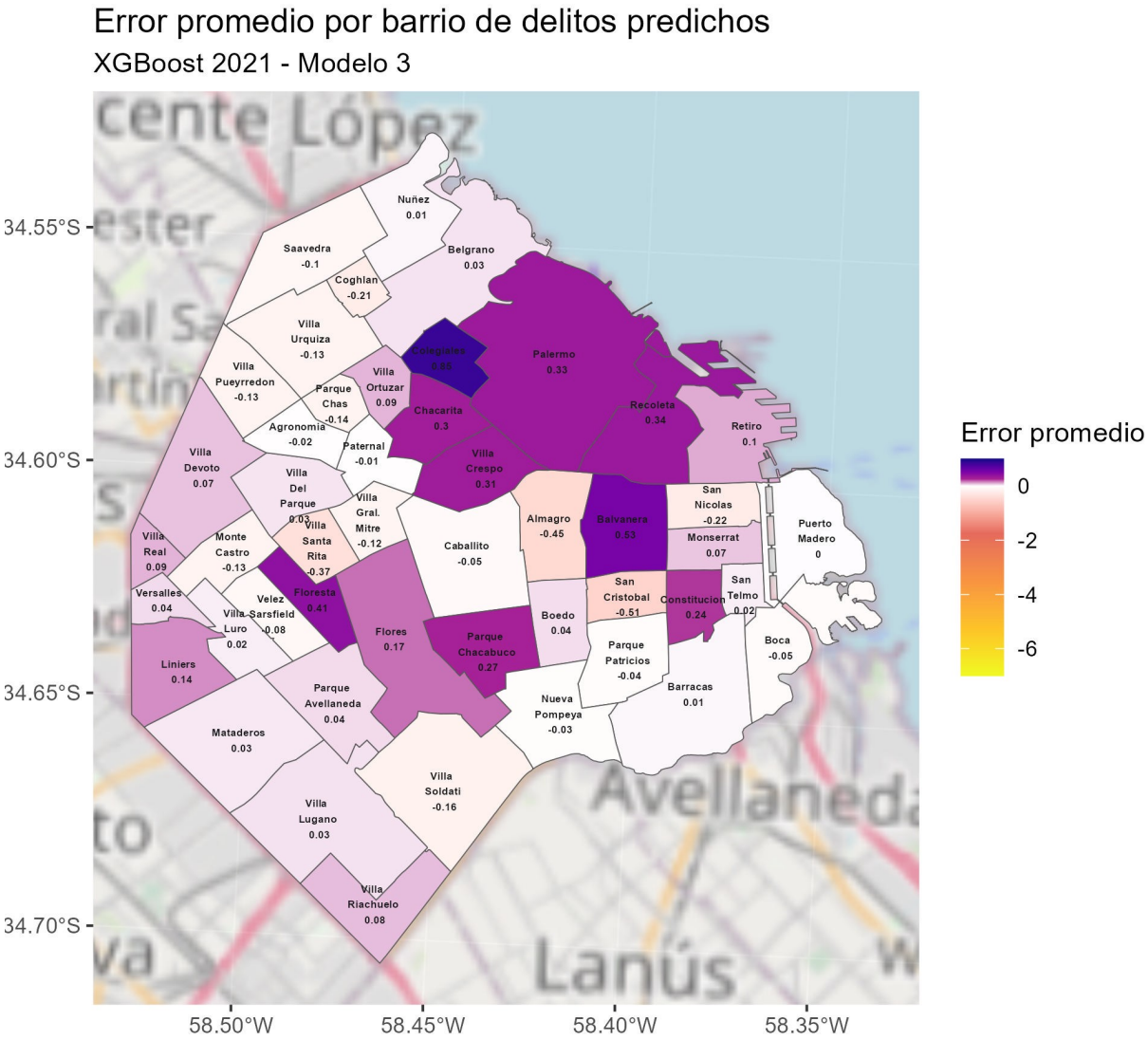
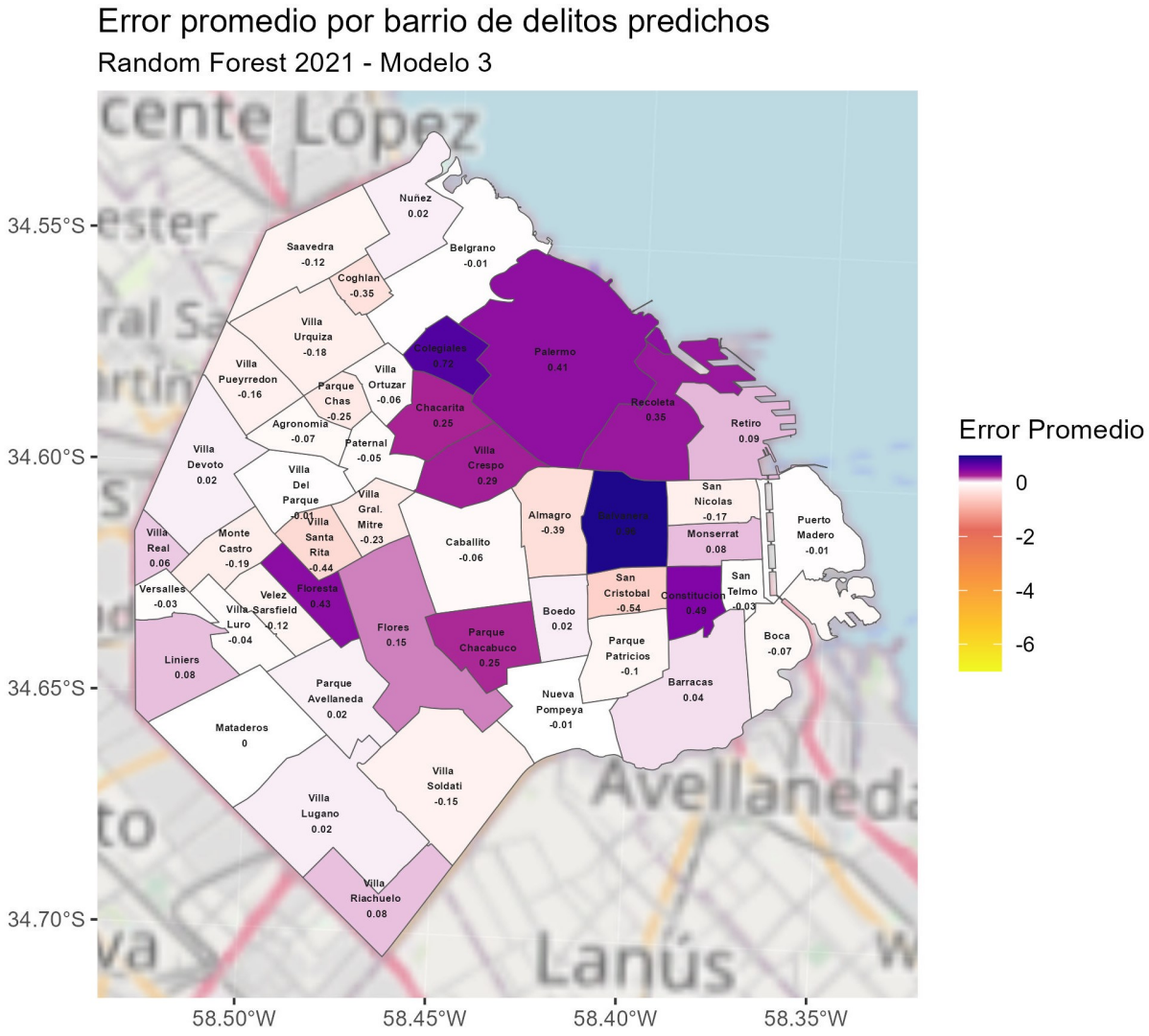


Figura 66. Error promedio por barrio - Random Forest 2021 Model 3



Apéndice D

Tabla 17. Ranking de puntos de interes

#	Tipo	Cant
1	Garaje comercial	2759
2	Restaurante	2134
3	Establecimiento educativo	1892
4	Farmacias	1376
5	Atm	1279
6	Organizaciones sociales	721
7	Librería	602
8	Café	389
9	Boca desubte	379
10	Parada de taxi	333
11	Club	301
12	Bar	300
13	Teatro	273
14	Hotel (bajo)	237
15	Estación de servicio	229
16	Parada de metrobus	213
17	Boliche	202
18	Hotel (alta)	188
19	Gimnasios	187
20	Parroquia	185
21	Universidad	128
22	Banco	122
23	Consulado	97
24	Embajada	91
25	Cartel luminoso	68
26	Centro de salud privado	65
27	Iglesia	50
28	Comisaria	49
29	Centro de salud comunitario	45
30	Parada bus turistico	37
31	Hospital	36
32	Centro médico barrial	31
33	Biblioteca	30
34	Cuartel de bomberos	30
35	Estadios	30
36	Estaciones de ferrocarril	29
37	Premetro	18
38	Basílica	15
39	Pista de skate	11

Tabla 18. Ranking precio promedio en dólares por m2 de terrenos en cada barrio de CABA - BA Data

#	Barrios	Prom.
1	Retiro	6449.08
2	Recoleta	5302.96
3	Puerto Madero	5085.54
4	Belgrano	4359.25
5	Palermo	4159.94
6	Nuñez	3877.99
7	San Nicolas	3658.70
8	Colegiales	3169.44
9	Caballito	3052.36
10	Villa Crespo	2829.05
11	Villa Urquiza	2725.25
12	Coghlan	2724.32
13	Almagro	2591.35
14	Monserrat	2531.43
15	Chacarita	2446.84
16	Villa Ortuzar	2364.41
17	Villa del Parque	2301.05
18	Balvanera	2299.53
19	Saavedra	2205.72
20	Boedo	2057.52
21	Villa Devoto	2028.53
22	Flores	2025.03
23	Parque Chas	1963.54
24	Villa Pueyrredon	1856.07

#	Barrios	Prom.
25	Agronomia	1821.91
26	Parque Chacabuco	1785.01
27	Villa Santa Rita	1784.89
28	Villa Luro	1763.82
29	Parque Patricios	1755.88
30	Villa Gral. Mite	1711.83
31	San Telmo	1708.03
32	Floresta	1665.63
33	San Cristobal	1590.73
34	Paternal	1559.20
35	Constitucion	1524.98
36	Monte Castro	1522.09
37	Velez Sarsfield	1494.20
38	Parque Avellaneda	1466.72
39	Liniers	1458.05
40	Barracas	1433.35
41	Nueva Pompeya	1399.18
42	Boca	1349.02
43	Versalles	1178.46
44	Villa Real	1172.03
45	Mataderos	1147.47
46	Villa Lugano	902.25
47	Villa Riachuelo	677.93
48	Villa Soldati	597.90

Tabla 19. Ranking precio promedio en dólares por m2 de departamentos en cada barrio de CABA - BA Data

#	Barrios	Prom.	#	Barrios	Prom.
1	Puerto Madero	6323.31	25	Boedo	2448.11
2	Palermo	3625.10	26	Villa Gral. Mite	2433.06
3	Belgrano	3541.08	27	Villa Luro	2427.18
4	Nuñez	3445.13	28	San Telmo	2423.59
5	Recoleta	3300.81	29	Monte Castro	2408.19
6	Retiro	3251.98	30	San Nicolas	2405.14
7	Colegiales	3180.00	31	Monserrat	2342.84
8	Villa Urquiza	3091.47	32	Villa Real	2314.51
9	Coghlan	2950.62	33	Mataderos	2288.36
10	Villa Ortuzar	2896.06	34	Paternal	2286.37
11	Saavedra	2827.65	35	Versalles	2242.06
12	Caballito	2822.75	36	Velez Sarsfield	2228.46
13	Villa Devoto	2820.45	37	Villa Santa Rita	2221.45
14	Parque Chas	2766.52	38	Balvanera	2182.15
15	Villa Pueyrredon	2728.91	39	San Cristobal	2124.76
16	Agronomia	2705.72	40	Boca	2078.32
17	Almagro	2697.30	41	Parque Patricios	2076.20
18	Villa Crespo	2676.92	42	Floresta	2071.60
19	Chacarita	2674.69	43	Parque Avellaneda	2041.91
20	Parque Chacabuco	2612.81	44	Constitucion	1982.75
21	Villa del Parque	2548.38	45	Nueva Pompeya	1867.75
22	Barracas	2515.31	46	Villa Riachuelo	1856.04
23	Liniers	2492.89	47	Villa Lugano	1506.71
24	Flores	2460.81	48	Villa Soldati	1097.04

Tabla 20. Ranking barrios de CABA con mayor porcentaje de hogares con necesidades basicas insatisfechas - BA Data

#	Barrios	Porcentaje
1	Boca	0.34
2	Villa Soldati	0.26
3	Constitucion	0.25
4	Monserrat	0.19
5	Retiro	0.17
6	Puerto Madero	0.15
7	Barracas	0.14
8	San Cristobal	0.14
9	Paternal	0.13
10	Balvanera	0.12
11	San Nicolas	0.12
12	Parque Patricios	0.11
13	Nueva Pompeya	0.10
14	San Telmo	0.10
15	Villa Lugano	0.09
16	Chacarita	0.09
17	Flores	0.09
18	Floresta	0.07
19	Boedo	0.07
20	Almagro	0.07
21	Recoleta	0.05
22	Parque Avellaneda	0.05
23	Villa Crespo	0.05
24	Colegiales	0.05
25	Villa Ortuzar	0.05
26	Parque Chacabuco	0.04
27	Mataderos	0.04
28	Villa Gral. Mite	0.04
29	Velez Sarsfield	0.03
30	Villa Santa Rita	0.03
31	Villa Riachuelo	0.03
32	Palermo	0.03
33	Villa Luro	0.03
34	Caballito	0.02
35	Liniers	0.02
36	Nuñez	0.02
37	Monte Castro	0.02
38	Villa Pueyrredon	0.02
39	Villa Urquiza	0.02
40	Agronomia	0.02
41	Saavedra	0.01
42	Coghlan	0.01
43	Parque Chas	0.01
44	Belgrano	0.01
45	Villa Devoto	0.01
46	Villa Real	0.01
47	Villa del Parque	0.01
48	Versalles	0.01