Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Facultad de Ingeniería

Detección de villas y asentamientos informales en el partido de La Matanza mediante teledetección y sistemas de información geográfica

Tesis presentada para optar al título de
Magister en Explotación de Datos y Descubrimiento del Conocimiento

Lic. Federico Baylé

Director: Dr. Rafael Grimson

Buenos Aires, 2016

DETECCIÓN DE VILLAS Y ASENTAMIENTOS INFORMALES EN EL PARTIDO DE LA MATANZA MEDIANTE TELEDETECCIÓN Y SISTEMAS DE INFORMACIÓN GEOGRÁFICA

Realizar un relevamiento de campo requiere contar con recursos logísticos para poder hacerlo de manera exhaustiva. La creciente disponibilidad de datos abiertos, imágenes satelitales de alta resolución y software libre para procesarlos abre la puerta a poder hacerlo de manera escalable a partir del análisis de esas fuentes de información. En el presente trabajo se hizo un ejercicio de ese tipo para detectar villas y asentamientos en el Partido de La Matanza considerando el relevamiento realizado por la ONG Techo en 2013. El objetivo es proponer una metodología que reduzca el área del territorio a relevar, teniendo en cuenta la periodicidad y actualización de los conjuntos de datos. Se utilizaron datos censales, viales y naturales georreferenciados, imágenes satelitales y algoritmos de aprendizaje automático. Los resultados muestran que usando la metodología propuesta con todas las fuentes de datos mencionadas se logra reducir el territorio a un 16 % ($51 km^2$), mientras que considerando solamente imágenes se reduce a 30 % ($96 km^2$).

Palabras claves: SIG, Aprendizaje automático, Análisis de imágenes basado en objetos, Villas y asentamientos informales.

AGRADECIMIENTOS

En primer lugar quiero agradecer a mi familia, a mi novia Agustina y a mis amigos de siempre por el apoyo constante para terminar el trabajo.

En segundo lugar quiero agradecer a Rafael Grimson por haber aceptado dirigir la tesis. Durante este año adquirí mucho conocimiento a partir de sus consejos.

También a mis compañeros de trabajo en Properati[1] por su ayuda para resolver algunos problemas que se me fueron planteando a lo largo de la tesis. En particular le quería agradecer a Martín por haberme facilitado hardware para hacer el procesamiento.

Por último le quería agradecer a Ana, los profesores y las autoridades de la maestría. Tanto por todo el conocimiento brindado como por la buena predisposición ante cada inquietud de mi parte.

Muchas gracias a todos!

[1] `http://www.properati.com/` y `http://www.properati.com.ar/data`

Índice general

1. INTRODUCCIÓN

Para llevar a cabo un relevamiento preciso de asentamientos informales se requiere hacer una cobertura exhaustiva del territorio. Esta tarea requiere contar con recursos logísticos y materiales que tienen un costo asociado, lo que atenta contra la periodicidad y el alcance. Por ejemplo, la ONG Techo (también conocida como Un Techo para mi País, su anterior denominación) viene realizando desde 2011 y cada dos años un relevamiento en sólo siete grandes ciudades[1], basado en su estructura de voluntariado y generación de recursos. Otro ejemplo es el Observatorio del Conurbano (dependiente de la Universidad de General Sarmiento) que relevó en 2006 los partidos que forman el Gran Buenos Aires.

Una aproximación a la identificación de sitios potencialmente críticos puede realizarse a partir del análisis de datos demográficos. El problema en este caso está en la periodicidad de los relevamientos, como el censo poblacional argentino que se realiza aproximadamente cada 10 años. Como consecuencia de esto, con el correr del período intercensal los datos se tornan menos representativos de la realidad.

Para ilustrar esta situación se presenta en la Tabla 1.1 la evolución de la población residente en villas y la variación relativa porcentual intercensal para la Ciudad de Buenos Aires considerando los censos de los años 1962, 1980, 1991, 2001 y 2010. Esta información fue publicada en los "Resultados provisionales del Censo Nacional 2010", documento elaborado por la Dirección General de Estadística y Censos (Ministerio de Hacienda GCBA) sobre la base de datos censales y relevamientos del Instituto Municipal de la Vivienda. Como puede verse allí, la variación relativa intercensal es elevada período a período.

Tab. 1.1: Evolución de la población residente en villas y asentamientos y variación relativa porcentual. Ciudad Autónoma de Buenos Aires.

Año	**Población**	**Variación**
1962	42.462	-
1980	34.064	-20 %
1991	52.608	54 %
2001	107.422	104 %
2010	170.054	58 %

Fuente: "Censo 2010. Situación y caracterización de los asentamientos precarios en la Ciudad de Buenos Aires". Dirección General de Estadística y Censos 2015, Gobierno de la Ciudad de Buenos Aires.

Una posible solución a este tipo de situaciones viene dada por el enfoque de "small area estimation", el cual consiste en combinar datos censales con encuestas con un menor nivel de desagregación pero más actualizadas. Este método viene siendo promovido por el Banco Mundial durante los últimos 20 años, teniendo como principales ejes las mediciones basadas en hogares y en comunidades. La principal utilidad de estos métodos es poder medir la pobreza basada en ingresos o consumo, logrando estimar con mayor desagregación las mediciones arrojadas por encuestas de mayor periodicidad que un censo. Puede verse

[1] Ciudad de Buenos Aires, Buenos Aires, Córdoba, Gran Rosario, Salta (4 ciudades), Alto Valle de Río Negro y Neuquén y departamento capital de Misiones.

en Henninger y Snel (2002) una introducción a este tema con un resumen de aplicaciones en diversos países como el caso de Brasil, Ecuador, Guatemala y Vietnam entre otros.

Otro posible enfoque para remediar el problema consiste en utilizar imágenes satelitales y de radar y analizarlas con métodos de teledetección y procesamiento de imágenes. Uno de los puntos fuertes de este enfoque radica en la disponibilidad de imágenes con una periodicidad mayor que los relevamientos de campo[2]. Existen varios conjuntos de imágenes de libre disponibilidad, como los de la serie Landsat de la NASA.

Tomando estos dos posibles enfoques para poder identificar villas y asentamientos informales, se puede hacer un análisis conjunto para sacar provecho de las cualidades de cada metodología. Por ejemplo, se pueden determinar territorios potenciales a través de información que surge de relevamientos de campo, para luego refinar la detección de este tipo de sitios analizando imágenes satelitales. Puede verse en Sliuzas, Mboup y de Sherbinin (2008) los diferentes trabajos presentados durante el Global Slum Mapping Workshop (CIESIN, Universidad de Columbia) acerca de este tipo de estudios.

El objetivo de este trabajo es definir una metodología para la detección de potenciales villas y asentamientos que permita reducir el territorio a relevar. Para esto se utilizaron datos censales, viales y naturales georreferenciados, imágenes satelitales, acotando el alcance territorial al Partido de La Matanza (provincia de Buenos Aires) durante el año 2013. La elección de ese momento corresponde con la fecha del relevamiento por parte de Techo (en julio de ese año). Se analizó el Partido de La Matanza por la heterogeneidad de su territorio, puesto que se encuentran tanto zonas urbanas como rurales, lo que se traduce en diferentes tipos de villas y asentamientos.

Teniendo en cuenta el objetivo mencionado, se estudió el impacto de utilizar solamente imágenes satélitales en lugar de todos los conjuntos de datos. La idea es analizar el uso de fuentes de información actualizadas periódicamente.

Todos los experimentos fueron realizados utilizando software libre. Para lo relacionado con teledetección se utilizó Orfeo ToolBox 5.0.0. El procesamiento de datos georreferenciados fue llevado a cabo con QGIS 2.8.1. Para el procesamiento de datos censales, geográficos y manipulación de imágenes se utilizó el lenguaje Python 2.7.

La metodología consta de una etapa de preprocesamiento de datos que luego se utilizan para el desarrollo de un modelo de detección, comparando los resultados obtenidos con el relevamiento de Techo. Se utilizó el coeficiente κ (kappa) de Cohen (Cohen, 1960) como medida de comparación entre los posibles modelos de detección.

[2] Por ejemplo el satélite Landsat 8 tiene un período de revisita de 16 días.

2. MATERIALES Y MÉTODOS

2.1. Partido de La Matanza

El territorio de análisis corresponde al Partido de La Matanza, provincia de Buenos Aires. El partido, según el censo de 2010, tiene una población de 1.772.130 habitantes (41,17 % de variación intercensal) con una densidad de 5,441 hab/m^2. La población urbana es de 1.753.291 habitantes(99 % del total). Tiene una superficie de $326km^2$, siendo su cabecera la localidad de San Justo. Este partido es el más extenso del conurbano bonaerense y el más poblado de la provincia de Buenos Aires. Es el segundo municipio más poblado del país, el primer lugar corresponde a la Ciudad de Buenos Aires.

El Partido de La Matanza limita al noreste con la Ciudad Autónoma de Buenos Aires, al sudoeste con los Partidos de Cañuelas, Marcos Paz y Ezeiza, al sudeste con los partidos de Lomas de Zamora y Esteban Echeverría al oeste con Marcos Paz, y al noroeste con Merlo, Morón y Tres de Febrero. Está dividido en 15 localidades.

Fig. 2.1: Partido de La Matanza y localidades limítrofes.

Fuente: Elaboración propia en base a OpenStreetMap y Dirección Provincial de Estadística de la Provincia de Buenos Aires

Analizando los asentamientos correspondientes a La Matanza, se puede observar en la Tabla 2.1 la cantidad de habitantes del partido y el porcentaje que vive en una villa o asentamiento. Allí se ve que la tendencia es creciente tanto para el partido en análisis como para la Ciudad de Buenos Aires (CABA) y el conurbano bonaerense.

Tab. 2.1: Población total, población en villas y asentamientos en La Matanza, Ciudad de Buenos y Conurbano Bonaerense 1981 - 2010.

Distrito	1991		2001		2010	
	Población Total	Población V y A	Población Total	Población V y A	Población Total	Población V y A
La Matanza	1.121.298	2 %	1.255.288	6 %	1.338.386	10 %
CABA	2.965.403	2 %	2.776.138	4 %	2.890.151	4 %
Conurbano Bonaerense	7.969.324	5 %	8.684.437	7 %	9.257.707	10 %

Fuente: elaboración propia en base a Cravino, Del Río y Duarte (2009) y Dirección General de Estadística y Censos, Gobierno de la Ciudad de Buenos Aires.

2.2. Relevamientos de villas y asentamientos considerados

Para el desarrollo y calibración del modelo de detección se consideró como base el relevamiento de villas y asentamientos realizados por la ONG Techo en julio de 2013. Se expondrán a continuación las definiciones y criterios de demarcación de este último trabajo y del realizado por la UNGS, para explicitar la manera en la que se llevan a cabo este tipo de tareas.

Para el relevamiento de UNGS, se considera villa a aquellas ocupaciones irregulares de tierra urbana vacante que producen tramas urbanas irregulares. Se caracterizan por los diferentes grados de precariedad de las viviendas, con una alta densidad poblacional. Para el caso de los asentamientos, Cravino et al. [12] explican que estos se distinguen por la ubicación de sus trazados sobre tierra privada (usualmente terrenos que pertenecían a basurales o zonas inundables) pero que no resultaban atractivos por sus dueños para la explotación económica. Estos terrenos se encuentran loteados y correctamente definidos.

Según los autores, una de las distinciones entre los pobladores de villas y de asentamientos radica en la percepción de la explotación de la tierra de sus habitantes. Para el caso de los asentamientos, los residentes tienen pensado establecer allí su vivienda, logrando en la posteridad la regularización de su situación, por lo que las viviendas tienden a ser construcciones firmes.

Considerando el relevamiento de Techo, esta ONG modificó la definición operativa respecto del relevamiento realizado por ellos en 2011. En el del año 2013 buscaron que la definición fuera consistente con la situación y los diferentes matices que se presentan a lo largo de todo el territorio nacional[1]. Para Techo se define asentamiento informal como un conjunto de al menos ocho familias agrupadas o contiguas, en donde más de la mitad de la población no cuenta con título de propiedad del suelo, ni acceso regular a, como mínimo, dos de los servicios básicos: red de agua corriente, red de energía eléctrica con medidor domiciliario y/o red cloacal. Puede verse con más detalle esta caracterización en el informe metodológico publicado por dicha organización en 2013.

[1] Por ejemplo en las zonas áridas del país, los desagües pluviales no son considerados un servicio esencial, situación que no se comparte en el resto del territorio.

2.3. Datos censales

Se utilizaron los microdatos correspondientes al Censo Nacional de Población, Hogares y Viviendas en la Argentina realizado en 2010. Se encuentran disponibles en el sitio web del Instituto Nacional de Estadística y Censos (INDEC). Como se menciona en la metodología del relevamiento, este fue un censo de hecho, lo que refiere a que fueron censadas las personas que se encontraban presentes a la hora cero del día 27 de octubre de 2010, fueran residentes habituales o no. Para llevar a cabo la recolección de datos se utilizó el procedimiento de entrevista directa por parte del censista en cada vivienda.

Las unidades de empadronamiento para el Censo 2010 son la Población (personas), los Hogares, las Viviendas Particulares y las Viviendas Colectivas. Cabe destacar que se dividió el país en provincias y a estas en departamentos (o partidos en la Provincia de Buenos Aires). Los departamentos se dividieron en fracciones, radios y segmentos censales (en orden decreciente de tamaño, cada uno está incluido en el anterior). Para más información se sugiere ver la metodología citada. En la Figura 2.2 se ve la división en radios censales del Partido de La Matanza[2].

Siguiendo las definiciones que provee INDEC, las fracciones y los radios censales se definen por una determinada cantidad de unidades de viviendas a relevar dentro de un espacio territorial con límites geográficos. Se considera segmento censal a la subdivisión interna del radio que comprende un espacio territorial y una cantidad de unidades que se definen para cada operativo. Representa el área de trabajo de cada censista en particular. Estas definiciones fueron extraídas a partir de documentación publicada por la Provincia de Santa Fe[3].

Los radios censales se pueden clasificar en urbano, rural o rural mixto. Los primeros son aquellos con población agrupada, formados por manzanas y/o sectores pertenecientes a una localidad. Para el caso de los rurales, en estos la población se encuentra dispersa y las viviendas se distribuyen en campo abierto de manera diseminada. Los rurales mixtos son aquellos cuya población se encuentra dispersa en campo abierto pero agrupada en pequeños poblados o en bordes amanzanados de localidades.

Cabe destacar que un radio censal contiene en promedio trescientas viviendas, mientras que una fracción censal (siguiente nivel de agregación) contiene un promedio de cinco mil. Para bordes de localidades, el radio urbano puede bajar a doscientas viviendas y en localidades aisladas a cien aproximadamente. Las divisiones pueden sufrir particiones entre un censo u otro, debido al crecimiento en la cantidad de viviendas, lo que dificulta el siguimiento censo a censo. Otro inconveniente radica en que el INDEC menciona que la información a nivel de fracción o radio censal (provista por las provincias) puede no ser consistente con los límites interprovinciales.

Al momento de realizar este trabajo solo se encontraban publicadas las preguntas asociadas al cuestionario básico del censo.

[2] La totalidad de los datos censales fueron obtenidos de diversas consultas a lo largo de este trabajo, procesando los datos con Redatam+SP. La cartografía se encuentra disponible en `http://www.indec.gov.ar/codgeo.asp`.

[3] Se puede acceder en `https://www.santafe.gov.ar/index.php/web/content/view/full/185198`

2.4. Fuentes de datos georreferenciados consideradas

2.4.1. Datos georreferenciados

2.4.1.1. Sistemas de información geográfica

Los datos georreferenciados son aquellos referidos a una posición con respecto a un sistema de coordenadas terrestres. Un sistema de información geográfica (SIG) es un conjunto de software, hardware y datos diseñado para la captura, almacenamiento y análisis de información georreferenciada.

Un SIG funciona como una base de datos con información geográfica (datos alfanuméricos) que se encuentra asociada por un identificador común a los objetos geográficos. De esta forma, señalando un objeto se conocen sus atributos e, inversamente, preguntando por un registro de la base de datos se puede saber su localización en la cartografía.

Por ejemplo, para el conjunto de datos de radios censales, los atributos pueden ser la población y la distribución según rangos etáreos asociados a un determinado código de radio censal. A partir de ese código se puede acceder a las localización del radio.

Los SIG pueden trabajar con distintos tipos de datos georreferenciados, considerando los modelos de representación ráster y vectorial.

El modelo ráster se utiliza para fenómenos que ocurren de manera continua en el espacio. Lo divide en celdas regulares (píxeles) donde cada una de ellas tiene asociado un único valor. Ejemplos de este son las imágenes satelitales y las fotografías aéreas.

El modelo vectorial almacena los datos como una serie de pares de coordenadas (X, Y) representadas a partir de primitivas geométricas (puntos, líneas y polígonos). Es útil cuando se quiere representar datos con límites definidos como calles o parcelas. Ejemplos de este son los polígonos utilizados para representar los radios censales.

Es importante elegir el modelo apropiado según el tipo de análisis que se quiera realizar. Por ejemplo, el uso de ráster puede producir resultados mucho más rapido que el vectorial. Sin embargo puede tener dificultades para representar adecuadamente atributos lineales dependiendo de la resolución de las celdas.

2.4.1.2. Sistemas de coordenadas

Los datos georreferenciados son representados siguiendo un modelo de la forma de la Tierra. Mientras un elipsoide ofrece una aproximación para ese modelo, un datum define la posición del elipsoide relativa al centro de nuestro planeta y el origen y la orientación de las líneas de latitud y longitud. De este modo, proporciona un marco de referencia para medir las ubicaciones en la superficie. Un datum se genera encima del elipsoide seleccionado y puede incorporar variaciones locales en la elevación.

Un datum centrado en la tierra o geocéntrico utiliza el centro de masa de la tierra como origen. En los últimos años, los datos de los satélites han proporcionado nuevas mediciones para definir el elipsoide que mejor se ajusta a la Tierra, que relaciona las coordenadas con el centro de masa de nuestro planeta. El último datum desarrollado, ampliamente utilizado, es WGS 1984 (conocido también como WGS84).

Los sistemas de coordenadas son designaciones arbitrarias para datos espaciales. Su finalidad es la de proporcionar una base común de comunicación sobre un lugar determinado o área de la superficie de la Tierra. Uno de los aspectos más importantes al trabajar con sistemas de coordenadas consiste es saber cuál es la proyección. Existen dos tipos de sistemas de coordenadas: geográficos y proyectados.

Un sistema de coordenadas geográficas utiliza una superficie esférica tridimensional para definir ubicaciones en la Tierra. Incluye una unidad angular de medida, un meridiano base y un datum. Los valores de latitud y longitud hacen referencia a un punto. La longitud y la latitud son ángulos medidos desde el centro de la Tierra hasta un punto de la superficie de la misma. Los ángulos se suelen medir en grados (o en grados centesimales).

Un sistema de coordenadas proyectadas se define sobre una superficie plana de dos dimensiones. A diferencia de un sistema de coordenadas geográficas, este posee longitudes, ángulos y áreas constantes en las dos dimensiones. Se caracteriza a partir de un sistema de coordenadas geográficas basado en una esfera o un elipsoide.

En un sistema de coordenadas proyectadas, las ubicaciones se identifican mediante las coordenadas (x, y) en una cuadrícula, con el origen en el centro de la cuadrícula. Cada posición tiene dos valores de referencia respecto a esa ubicación central. Uno especifica su posición horizontal y el otro su posición vertical.

Los datos georreferenciados que se utilizaron en este trabajo utilizan la proyección universal transversa de Mercator (conocido comunmente como UTM). El sistema UTM divide la Tierra en sesenta zonas, donde cada una abarca meridianos de seis grados de longitud, utilizando una proyección Mercator normal pero secante a un meridiano en lugar de tangente al Ecuador. La zona que mejor se adapta al territorio en análisis es la 21 S. De este modo el datum considerado será el WGS84 con la proyección UTM zona 21 S.

2.4.2. Radios censales

El conjunto de datos de polígonos de radios censales correspondientes al Censo Nacional de 2010 se encuentran publicados en el sitio web del INDEC y puede ser descargado de manera libre. También puede ser descargado desde el sitio web de la Dirección Provincial de Estadística de la Provincia de Buenos Aires.

Los polígonos utilizan el sistema de coordenadas de referencia POSGAR 94 / Argentina 5, elaborado por el Instituto Geográfico Militar de nuestro país. Los conjuntos de datos de ejes, manzanas y usos del suelo tienen el mismo sistema de referencia que el de los radios censales. El Partido de La Matanza posee 1.302 radios censales para el censo mencionado.

2.4.3. Ejes

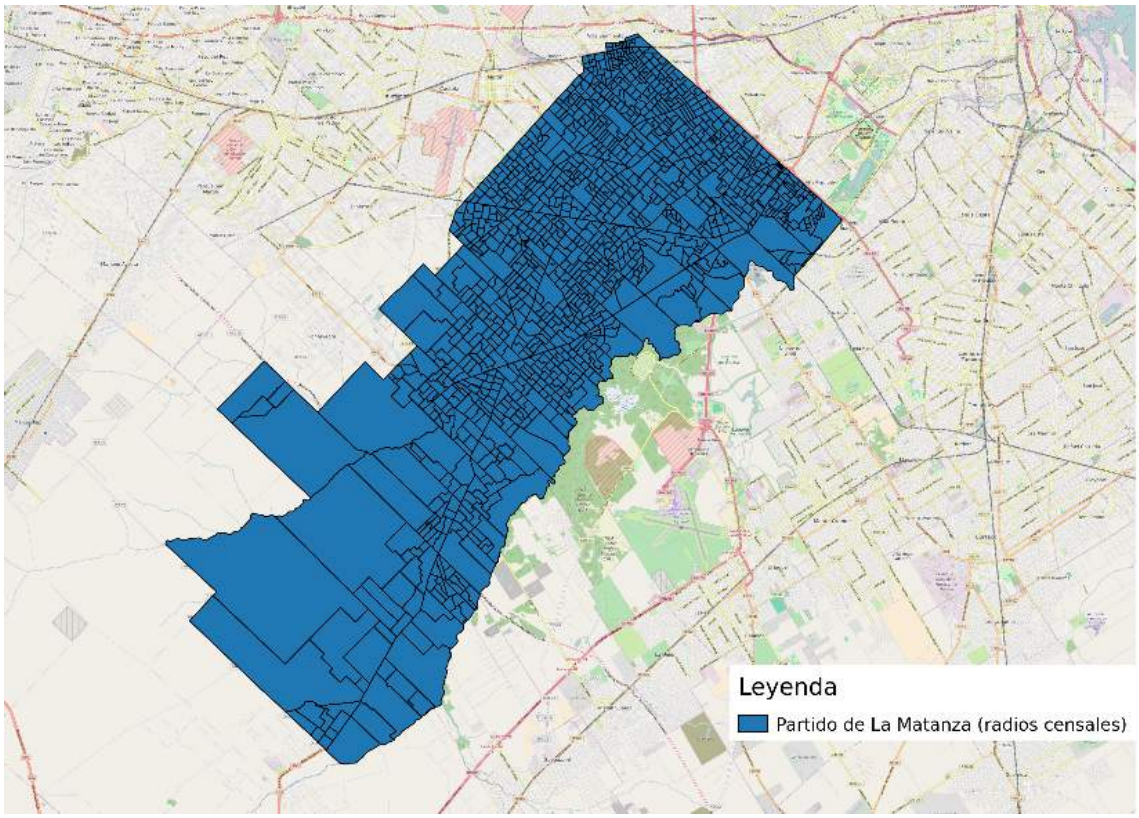
Este conjunto de datos es provisto por la Dirección Provincial de Estadística de la Provincia de Buenos Aires que lo publica para cada partido integrante de la provincia[4]. Presenta 36.059 ejes. Los atributos son: ancho de calle (en metros), nombre de calle, altura inicial izquierda, altura final izquierda, altura inicial derecha y altura final derecha.

Los distintos tipos de ejes que presenta son: curso de agua, líneas costeras, arroyos, calles, avenidas y vías de tren. La distribución puede verse en la Tabla 2.2.

En la Figura 2.3 se muestra cada eje graduado según su ancho. El eje de color oscuro que atraviesa el partido se denomina Avenida Brigadier Juan Manuel de Rosas (Ruta Nacional 3). Puede verse que a medida que se avanza en dirección sur, los ejes que parten desde dicha avenida son mas cortos, lo que se corresponde con localidades con un componente rural cada vez mayor, como el caso de Gonzalez Catán o Virrey del Pino.

[4] Estos pueden ser descargados de `http://www.ec.gba.gov.ar/estadistica/censo2010/cartografia.html`. A partir de noviembre de 2016 también se encuentran disponibles en el portal de datos abiertos de la Provincia de Buenos Aires (`http://datos.gba.gob.ar/`).

Fig. 2.2: Radios Censales correspondientes al Partido de La Matanza, Censo 2010.



Fuente: elaboración propia en base a datos del INDEC. Censo Nacional de Población, Hogares y Viviendas 2010 y Cartografía y códigos geográficos del Sistema Estadístico Nacional. Buenos Aires: INDEC, 2016.

2.4.4. Manzanas

El conjunto de datos de polígonos de manzanas es provisto por el mismo organismo que el de ejes y también se encuentra en el portal de datos abiertos mencionado. Está formado por 14.946 manzanas, presentando los atributos: código de provincia, código de departamento, código de localidad, fracción censal, radio censal y código de manzana.

Cabe destacar que al presentar el número de radio censal, este conjunto de datos puede asociarse a los datos censales, por lo que se puede saber qué manzanas integran cada radio.

2.4.5. Usos del suelo

Este conjunto de datos fue construido a partir de la información que se presenta en el Plan de Desarrollo Productivo 2020 presentado por la Secretaría de Hacienda del Partido de La Matanza durante el año 2014[5]. Partiendo del conjunto de datos de polígonos de manzanas como base, se asoció a cada uno de estos un uso del suelo en particular a partir de los publicados en ese informe.

Como se menciona en ese documento, la idea de este plan es superar las limitaciones técnicas para la implementación de la estrategia de ordenamiento y promoción industrial,

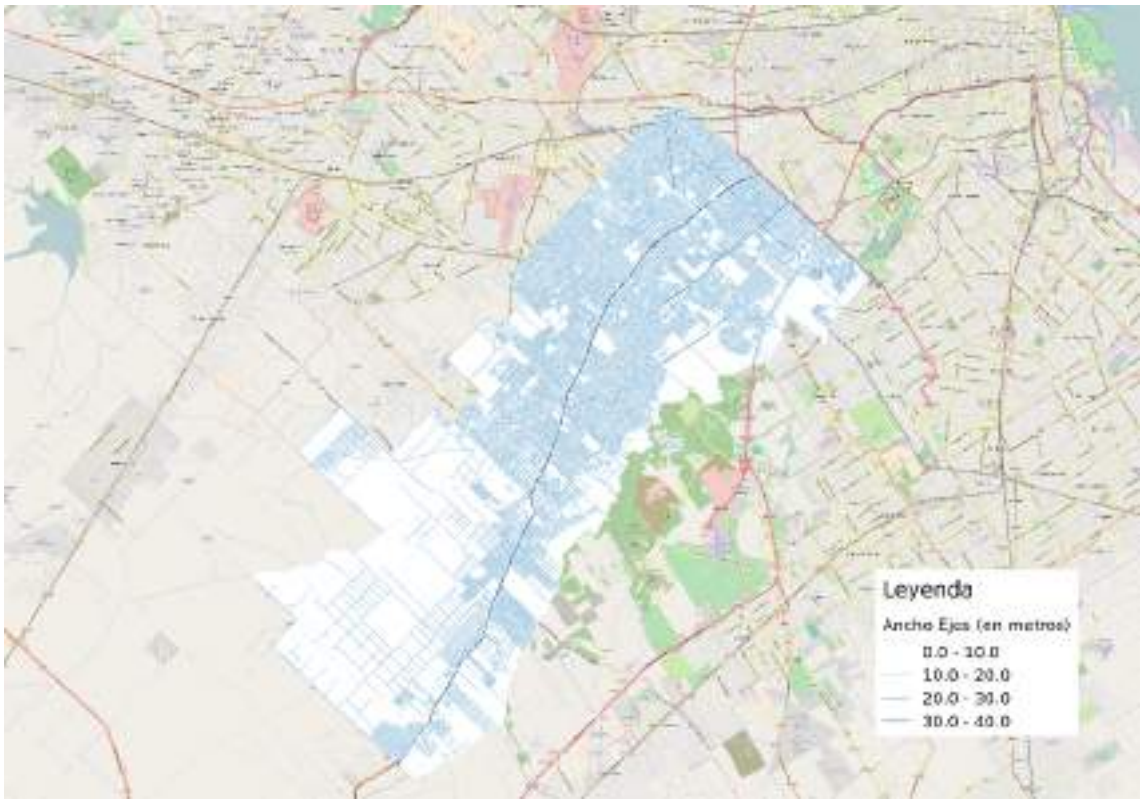
[5] `http://produccion.lamatanza.gov.ar/assets/files/PLAN%20DE%20DESARROLLO%20PRODUCTIVO%20MATANZA%202020%20-%20Anexo%20I.pdf`

Tab. 2.2: Cantidad y Porcentaje de Ejes según el ancho, Partido de La Matanza.

Ancho	Cantidad de Ejes	Porcentaje
hasta 5m	1.840	5 %
entre 5m y 10m	528	1 %
entre 10m y 15m	32.717	91 %
entre 15m y 25m	542	2 %
entre 25m y 40m	433	1 %

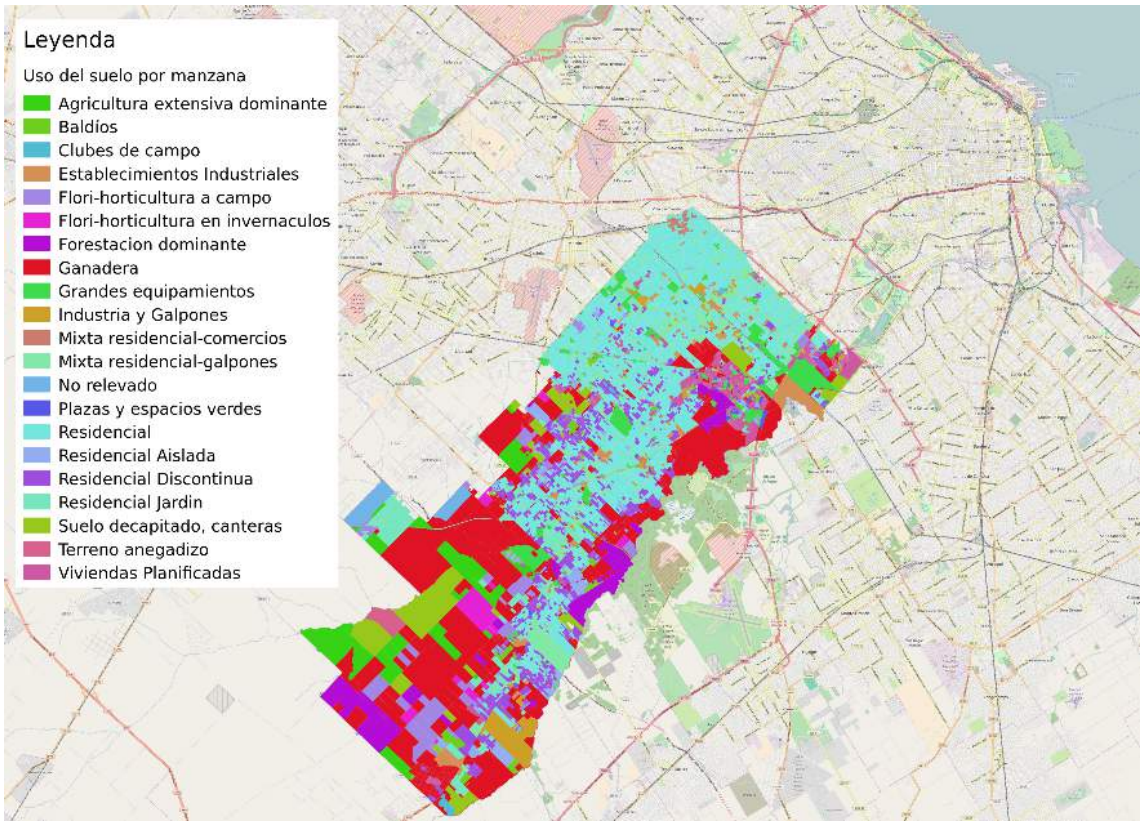
Fuente: elaboración propia en base a Dirección Provincial de Estadísticas de la Provincia de Buenos Aires.

Fig. 2.3: Ejes correspondientes al Partido de La Matanza, Censo 2010.



Fuente: elaboración propia en base a Dirección Provincial de Estadísticas de la Provincia de Buenos Aires.

Fig. 2.4: Usos del Suelo correspondientes al Partido de La Matanza, por polígono de manzana.



Fuente: elaboración propia en base a Ministerio de Hacienda del Partido de La Matanza (2014).

dada la falta de datos actualizados. Una vez llevado a cabo el relevamiento, la idea es que sirva de rector para ejecutar la planificación y ordenamiento de determinadas zonas del municipio, con énfasis en las zonas industriales. El objetivo final del documento es mejorar la zonificación de esas áreas para la aplicación del régimen de promoción industrial vigente.

En la Figura 2.4 se presenta el mapa creado a partir de los datos del informe. Puede verse como las manzanas residenciales se encuentran en mayor proporción a medidad que se acerca la Ciudad Autónoma de Buenos Aires, mientras que se vuelven más rurales a medida que se avanza hacia el sur. En la Tabla 2.3 se muestran aquellos usos que mayor cantidad de manzanas tienen asociadas.

2.5. Imágenes satelitales

2.5.1. Teledetección

Siguiendo a Carnegie y Lauer [6], "el término teledetección se refiere a la detección e identificación de objetos mediante el uso de cámaras aéreas o de otros dispositivos de detección que se encuentran a una apreciable distancia de los elementos objeto de la investigación.". Otros autores como Chuvieco [9] definen a esta disciplina como "la observación remota de la superficies de la Tierra desde sensores aerotransportados, plataformas espaciales, fotografía aérea y globos aerostáticos". El énfasis será puesto en los sensores ópticos,

Tab. 2.3: Cantidad y Porcentaje de Manzanas por Uso del Suelo, Partido de La Matanza, 2014

Uso del Suelo	Cantidad de Manzanas	Porcentaje
Residencial	7.023	47 %
Residencial Discontinua	1.769	12 %
Mixta Residencial-Galpones	1.035	8 %
Ganadera	1.148	8 %
Viviendas Planificadas	727	5 %
Residencial Aislada	483	3 %
Grandes Equipamientos	371	2 %
Residencial Jardín	297	2 %
Industria y Galpones	251	2 %
Baldíos	209	1 %
Mixta Residencial-Comercios	175	1 %
Residencial Discontinua	150	1 %
Otros (12 categorías más)	1.115	7 %

Fuente: elaboración propia en base a Ministerio de Hacienda del Partido de La Matanza (2014).

sin dejar de mencionar que existen otro tipo como los térmicos, microondas (tanto activas como el caso de los radares, como pasivas). En la Figura 2.5 se muestra un esquema de interacción a modo de ejemplo.

Fig. 2.5: Interacción fundamental en la teledetección.

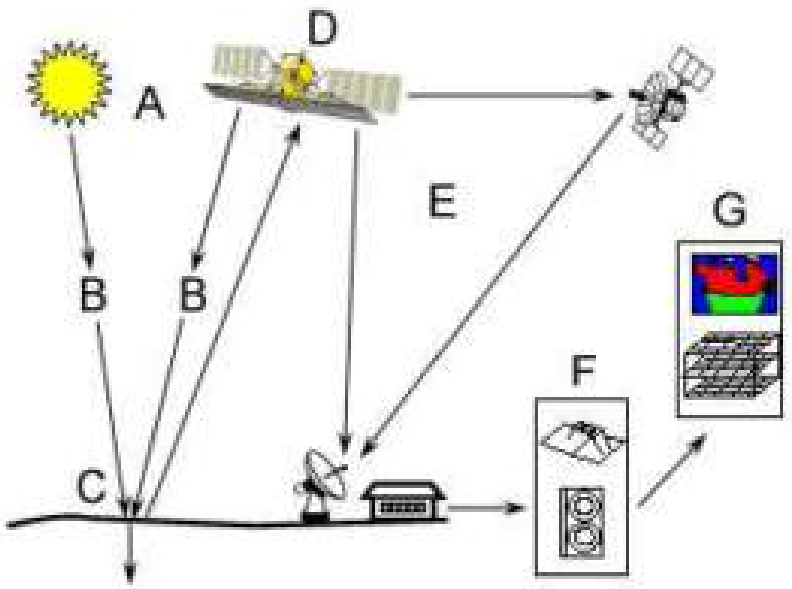
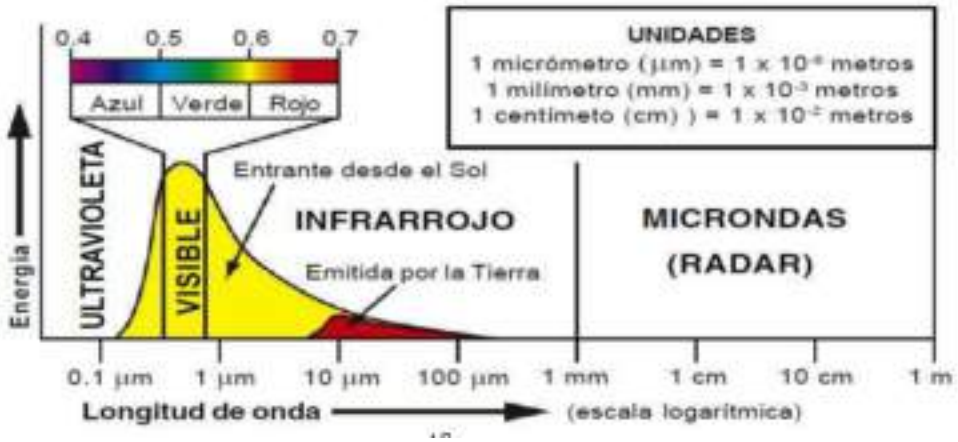


Fig. 2.6: Variación en longitud de onda de la radiación electromagnética.



Fuente: Miño (2012)

2.5.1.1. Radiación electromagnética

La radiación electromagnética es una forma de energía que se propaga mediante ondas que se desplazan por el espacio a la velocidad de la luz, transportando cantidades discretas de energía. Cuando un haz de luz incide en un material, una parte de esa energía (dependiendo de la longitud de onda) es absorbida por el material y otra parte es reflejada. En la Figura 2.6 se muestran los distintos tipos de longitud de onda de la radiación electromagnética.

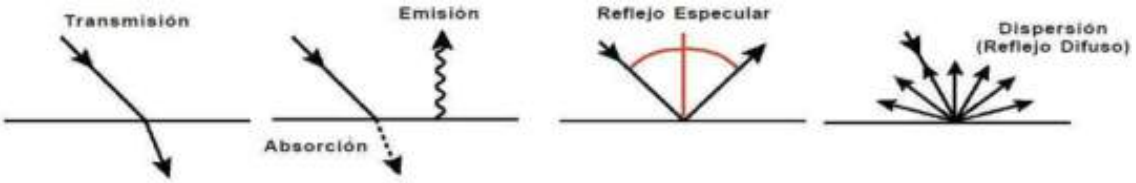
De este modo se puede definir reflectancia como la cantidad de energía que es reflejada por un objeto luego de que esta incide sobre él. El resto de la energía puede ser transmitida o absorbida (absorción) por el objeto.

Para una determinada superficie la reflectancia varía en función de la longitud de onda. La primera y más importante fuente es el Sol que ilumina los objetos que se encuentran en la superficie terrestre. Mediante reflexión parte de esa energía incide en los objetos y luego se refleja hasta llegar al sensor. Una fuente de tipo artificial puede ser la radiación emitida por los radares, midiendo la cantidad de energía que es retrodispersada hacia el mismo.

Las interacciones con la superficie terrestre se pueden dar a través del proceso de transmisión, absorción, emisión, reflexión o dispersión. Cabe destacar que las interacciones dependen fuertemente de los materiales y pueden cambiar la dirección, intensidad y polarización de la radiación. Por ejemplo, no es lo mismo la interacción con un techo de metal que con uno de tejas. En la Figura 2.6 de presentan estas interacciones.

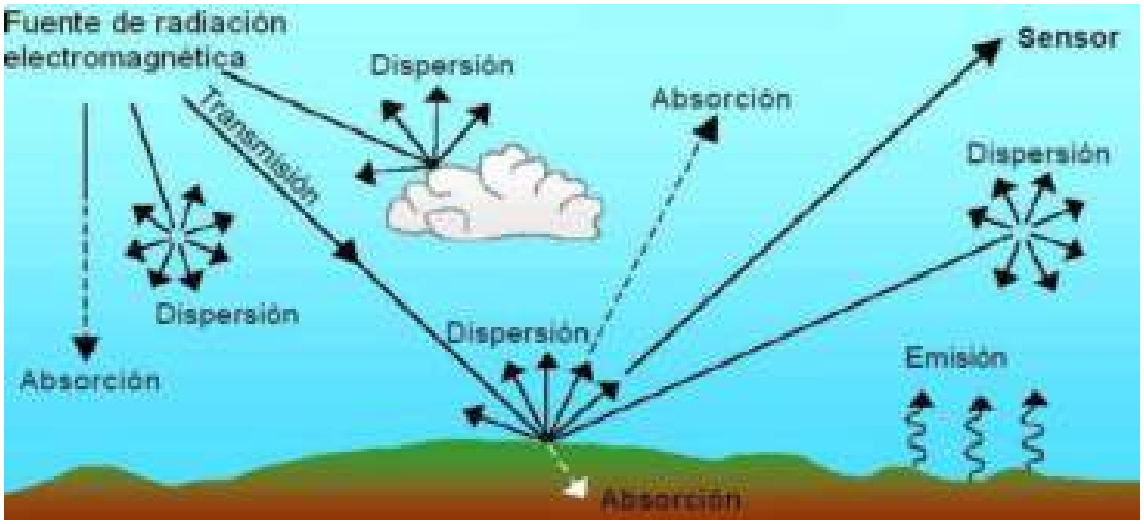
La radiación proveniente de la cubierta terrestre tiene que atravesar la atmósfera para llegar al sensor, viéndose afectada por los fenómenos de dispersión, absorción y emisión. Esto ocurre por la interacción con moléculas de gas, vapor de agua y aerosoles. El primero de estos afecta a las ondas de longitudes más cortas ocasionando un cambio de dirección de la radiación respecto de la que llevaba originalmente. La absorción consiste en la transformación energética en determinadas bandas del espectro cuando la radiación atraviesa el medio. Por último, la emisión está asociada a que todo cuerpo caliente introduce modificaciones en la radiación originalmente propagada entre la cubierta terrestre y el

Fig. 2.7: Interacciones de la radiación electromagnética con la superficie terrestre



Fuente: Smith R. B. (2006)

Fig. 2.8: Interacciones de la radiación electromagnética con la atmósfera y superficie terrestre.



Fuente: Smith R. B. (2006)

sensor.

El sensor remoto genera una señal eléctrica (analógica) proporcional a la intensidad de la señal electromagnética que recibe y la digitaliza en números enteros conocidos como números digitales (ND). Esto es, transforma una señal continua en un valor discreto. De este modo se forma una imagen de tipo ráster con diferentes valores numéricos que representan niveles de luminosidad. Se sugiere ver [9] para profundizar.

2.5.1.2. Tipos de imágenes generadas por los sensores ópticos

Debido a restricciones físicas en el diseño del sensor, resulta dificultoso lograr alta resolución espacial y espectral al mismo tiempo. Por esta razón, la mayoría de los satélites de alta resolución contienen dos tipos de sensores que generan dos tipo de imágenes diferentes:

- imágenes multiespectrales: compuestas de tres a ocho bandas presentando usualmente el canal azul, verde, rojo e infrarojo. Normalmente tienen una resolución que va de $2,8$m a 2m.
- imágenes pancromáticas: imagen en escala de grises adquirida por un sensor que cubre la parte más amplia del espectro de luz. Esto permite aumentar el flujo óptico y

por lo tanto reducir el tamaño del píxel. La resolución de este tipo de imágenes es por lo general aproximadamente cuatro veces mayor que la de la imágen multiespectral.

2.5.2. Imágenes utilizadas

Las imágenes utilizadas en este trabajo fueron tomadas por el satélite WorldView-2 (WV2), que fue lanzado en octubre de 2009. Es el primero de tipo comercial de alta resolución con 8 bandas multiespectrales. Estas definen el número y ancho de las bandas en el espectro electromagnético al que es sensible el sensor. El mismo opera a una altitud de 770km, con un período de revisita de $1,1$ días. Las imágenes en el canal pancromático tienen una resolución de 46cm mientras que las de las bandas multiespectrales tienen 1.85m.

En la Tabla 2.4 se presentan las principales características[6].

Tab. 2.4: Especificaciones Técnicas Satélite WorldView-2

Especificaciones WorldView-2	
Lanzamiento	8/10/2009, Base Aérea Vanderberg, California.
Orbita	770 km.
Duración Misión	10 - 12 años.
Bandas	- Pancromática: 450 - 800 nm - 8 Multiespectrales - Coastal: 400 - 450 nm - Azul: 450 - 510 nm - Verde: 510 - 580 nm - Amarillo: 585 - 625 nm - Rojo: 630 - 690 nm - Rojo edge: 705 - 745 nm - Infrarojo 1: 770 - 895 nm - Infrarojo 2: 770 - 895 nm
Período de Revisita	1,1 días.
Capacidad	1 millón km^2 por día.

Fuente: elaboración propia en base a especificaciones técnicas de Digital Globe.

2.6. Aprendizaje automático supervisado y no supervisado

De acuerdo a Mitchell [24] un programa aprende de su experiencia E respecto a una tarea T y una métrica P si el desempeño en la tarea T medido por P mejora con la experiencia E. El objetivo principal de todo proceso de aprendizaje es utilizar la evidencia conocida para poder crear un modelo y poder dar una respuesta a nuevas situaciones no conocidas.

Los problemas posibles de ser abordados pueden definirse como aprendizaje supervisado, no supervisado y por refuerzos.

- Aprendizaje supervisado: se le presenta al programa una serie de ejemplos (entradas) con su resultado esperado (salidas). Se pretende que infiera una correspondencia

[6] Para más información puede consultarse `https://www.digitalglobe.com/about/our-constellation`

entre las entradas y las salidas para poder luego aplicarse a nuevos ejemplos. Una posible subdivisión dentro de este tipo de aprendizaje se da según el tipo de dato que sea salida. Si es una categoría nos encontramos ante un problema de clasificación mientras que si la salida es un número real estamos frente a un problema de regresión. Un ejemplo asociado a esta tesis es la clasificación de un segmento como villa o no a partir de sus atributos.

- Aprendizaje no supervisado: se le presenta al programa un conjunto de entradas sin su salida correspondiente, dejando que encuentre una estructura o patrón común en los datos. Una posible aplicación separaría un grupo de segmentos de acuerdo a una estructura común de pertenencia.

- Aprendizaje por refuerzos: en este caso el programa interactúa con su entorno retroalimentándose de la respuesta a sus acciones, aprendiendo mediante prueba y error. Un ejemplo de aplicaciones de este tipo de aprendizaje se da en los programas que aprenden a jugar contra un oponente.

2.6.1. Algoritmos de aprendizaje no supervisado utilizados

2.6.1.1. Análisis de componentes principales

Siguiendo a Hastie [18], los componentes principales son una secuencia de proyecciones del conjunto de datos, no correlacionados entre si y ordenados decrecientemente de acuerdo a su varianza. Dado un conjunto de datos en $\mathbb{R}^p$, los componentes principales proveen una secuencia de las mejores aproximaciones lineales para todo rango $q \leq p$. Por ello, el número de componentes principales es menor o igual a la cantidad de atributos del conjunto de datos, por lo que se puede afirmar que permite reducir la dimensionalidad de los datos. Esto se debe a que necesitamos q atributos en lugar de p para caracterizar una observación.

El análisis de componentes principales construye una transformación lineal que elige un nuevo sistema de coordenadas para el conjunto original de datos en el cual la varianza de mayor tamaño del conjunto de datos es capturada en el primer eje (llamado el Primer Componente Principal), la segunda varianza más grande es el segundo eje, y así sucesivamente. Para reducir la dimensionalidad de un grupo de datos, retiene aquellas características del conjunto de datos que contribuyen más a su varianza.

Si representamos nuestro conjunto de datos de p variables $X_1, X_2, ..., X_p$ con n individuos, la matriz de datos resulta:

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{np} \end{bmatrix}$$

Sean $X = [X_1, X_2, ..., X_p]$ y $S = var(X)$ su matriz de covarianzas. Puesto que $S \geq 0$ y es simétrica, S puede descomponerse en $S = T\Lambda T'$ con $T = [t_1, ..., t_p]$ y $\Lambda = diag(\lambda_1, ..., \lambda p)$, siendo $\lambda_1 > ... > \lambda_p$.

Luego, las componentes principales de X son las nuevas variables $Y_j = Xt_j, j = 1, ..., p$.

Para cada j, la nueva variable Y_j se construye a partir del j-ésimo autovector de S. De este modo, el individuo i-ésimo de X se representa como $y_i = x_i'T = (x_i't_1, ..., x_i't_p)$.

Se sugiere ver Hastie, Tibshirani y Friedman (2010) para una aproximación más formal al tema.

2.6.2. Algoritmos de aprendizaje supervisado utilizados

En la siguiente sección se realizará una breve introducción a los algoritmos con los que se experimentó: Random Forests, XGBoost, Máquinas de Vectores de Soporte (conocido como SVM por sus siglas en inglés) y mezcla de distribuciones gaussianas (conocido como GMM). Se sugiere ver [18] para profundizar en cada caso.

2.6.2.1. Métodos básados en árboles de decisión

Antes de introducir el algoritmo Random Forests, se comentará brevemente el método de árboles de decisión para luego extenderlo a este.

El método de árboles de decisión representa el conocimiento a partir particiones de las variables consideradas. Cada nodo interno del árbol se ramifica en cada categoría posible del atributo asociado (las variables continuas son particionadas en categorías para tal fin). Esta metodología puede representar cualquier conjunción y disyunción con la facilidad de poder ser reescritas como reglas a partir de la concatenación de las sentencias. Cuanto más profundo el árbol, más complejas serán las reglas que generalizó el modelo.

Entre las ventajas de esta metodología se destaca la facilidad de interpretación del modelo resultante. Entre las desventajas se puede mencionar la inestabilidad ante pequeñas variaciones en los datos y la presencia de clases desbalanceadas (como sucede en esta tesis). Otra de las desventajas que presentan es la necesidad de ser implementados basándose en heurísticas puesto que el problema de encontrar el árbol de decisión óptimo es NP-completo. Como ejemplo de estas implementaciones puede verse el algoritmo C4.5 de Quinlan (1993).

Para poder solucionar el problema de la inestabilidad se suele clasificar a partir del resultado de la ejecución de varios árboles, lo que se conoce como ensamble de árboles de decisión. Las técnicas comúnmente utilizadas para trabajar con ensambles son bagging y boosting.

- Bagging (Bootstrap aggregating, Breinman [4]): se ajustan muchos árboles profundos a versiones muestreadas aleatoriamente con reemplazo (muestreo bootstrap) del conjunto de datos de entrenamiento. Se clasifica por voto mayoritario entre esos árboles. Se busca reducir la varianza de las predicciones generando datos adicionales a partir del muestreo con reemplazo.

- Boosting (Freund y Schapire [15]): se ajustan muchos árboles poco profundos a subconjuntos de los datos de entrenamiento. Se clasifica a partir de alguna regla como voto mayoritario. A diferencia del bagging, la creación de los subconjuntos es aleatoria pero la probabilidad de selección de cada instancia no es uniforme. Considerando todas las instancias, tienen mayor probabilidad de ser nuevamente seleccionadas aquellas clasificadas erróneamente por el modelo. De este modo, los nuevos árboles son creados para predecir las instancias clasificadas erróneamente en pasos anteriores.

2.6.2.2. Random Forests

Este algoritmo ensambla árboles de decisión, cuya cantidad es configurada al inicio, utilizando bagging.

Para cada árbol se selecciona de manera bootstrap un subconjunto de los datos de entrenamiento y se deja el resto de las instancias para evaluar el error. Para cada nodo del árbol se selecciona una muestra aleatoria (el tamaño es un parámetro del algoritmo) de los atributos del conjunto de datos para realizar la mejor partición posible de acuerdo a algún criterio predefinido (como puede ser entropía). De este modo se obtiene un árbol de decisión.

Ante una instancia a predecir, esta es clasificada por todos los árboles que se hayan computado en el ensamble. La regla de clasificación es voto mayoritario. El error de evaluación converge asintóticamente a un límite a medida que el número de estimadores va creciendo, se sugiere ver Breiman [5] para una demostración de esta propiedad.

Entre las ventajas que presenta esta metodología pueden mencionarse la estabilidad de las predicciones y la posibilidad de estimar la importancia de cada variable en la clasificación. Dentro de las desventajas se encuentra la posibilidad de generalizar casos particulares dada la cantidad y profundidad de los árboles que intervienen (lo que se suele denominar sobreajuste del modelo). Otra desventaja es la dificultad de interpretación de las reglas de clasificación.

Respecto del cómputo de importancia de variables, el procedimiento consiste en calcular cuánto desciende la medida de impureza (sea Gini o entropía) del árbol si se considera determinado atributo para un nodo. Llevado al ensamble de árboles, se promedia el descenso de impureza de cada atributo y a partir de esto se determina la importancia de cada variable.

Se le critica al cálculo de importancia de variables el desempeño ante la presencia de atributos altamente correlacionados. El problema es que estos se consideran como predictores individualmente, de modo que cuando uno es utilizado, los demás pierden su efectividad en reducir la impureza. Esto se debe a que su aporte es similar al del atributo considerado anteriormente y por lo tanto se diluye su importancia. Cuanto más correlacionados es peor. En la sección de modelado se muestra este problema en los atributos considerados en esta tesis.

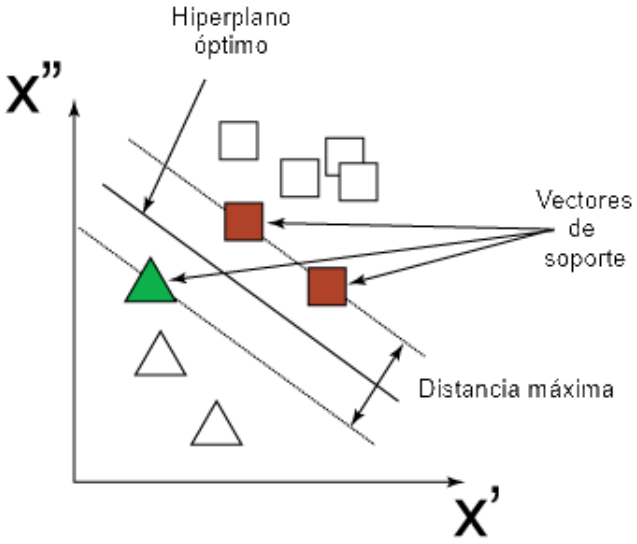
Para realizar los experimentos se utilizó la implementación de este algoritmo en Python 2.7 mediante la librería scikit-learn 0.17.

2.6.2.3. XGBoost

El algoritmo XGBoost (por *extreme gradient boosting*) es una implementación del método *gradient boosting decision tree*. Está basada en aprovechar los recuros de cómputo con los que se cuenta, permitiendo paralelizar y distribuir la creación de árboles de decisión optimizando el uso de memoria. La principal diferencia de este método con el de Random Forests es que en este último los árboles que se ensamblan son independientes mientras que en XGBoost dependen de los errores de los modelos anteriores (es decir, uno usa bagging y el otro boosting para el ensamble).

Tanto este algoritmo como el de Random Forests devuelven la importancia de cada atributo en el resultado final. En el caso de XGBoost, la calcula sumando cuantas veces cada atributo parte un nodo en cada árbol del modelo. La diferencia con el de Random Forests está en los atributos que se encuentran correlacionados: suponiendo que se tienen dos atributos A y B perfectamente correlacionados, como en Random Forests los árboles se crean de manera independiente, un porcentaje de estos va a contener a A mientras que el porcentaje restante contendrá a B. De este modo, la información que contienen estos atributos (que es la misma por estar perfectamente correlacionados) será divida en estos

Fig. 2.9: Ejemplo de representación en dos dimensiones de un hiperplano óptimo separando datos y vectores de soporte



dos. Esto en XGBoost no sucede puesto que cuando el algoritmo aprende la influencia de una variable en el resultado, no se vuelve a enfocar en ella. Por esta razón, la importancia aquí estará toda en A o en B pero no divida entre ambas como en el caso de Random Forests.

La implementación utilizada fue la de Python 2.7 mediante la librería xgboost 0.6. Se sugiere ver Chen [8] para mayor detalle.

2.6.2.4. Máquinas de Vectores de Soporte

Las Máquinas de Vectores de Soporte (comunmente conocidas como SVM por sus siglas en inglés) fueron introducidas en los años 90 por Vapnik y su equipo de colaboradores pensadas originalmente para resolver problemas binarios. Pertenecen al grupo de los clasificadores lineales dado que infieren separadores lineales o hiperplanos, ya sea en el espacio original de las instancias de entrada (para el caso que sean separables o cuasi-separables) o en un espacio transformado (si no lo son).

La idea es seleccionar un hiperplano de separación que equidiste de las instancias más cercanas de cada clase para obtener lo que se denomina *margen máximo* a cada lado del hiperplano. Las instancias que caen justo en la frontera de estos márgenes reciben el nombre de vectores de soporte. Desde un punto de vista algorítmico el problema de optimización del margen constituye un problema de optimización cuadrático con restricciones lineales que puede ser resuelto con técnicas estándar de programación cuadrática.

Para facilitar la resolución del problema de optimización se utilizan funciones denominadas kernel. Estas mapean la información a un espacio de mayor dimensión de manera que resulte más fácil el aprendizaje lineal (por ejemplo, ayudando a realizar ciertos cálculos más rápido). En esta tesis se experimentó con los siguientes tipos de funciones kernel:

- Polinomial: $(\gamma\langle x, x'\rangle + r)^d$, con d como grado del polinomio y r como la constante asociada a x^0.

- Lineal: $\langle x, x' \rangle$
- Radial: $\exp(-\gamma|x - x'|^2)$, con $\gamma > 0$.

Al igual que con Random Forests, se utilizó la implementación en Python 2.7 mediante la librería scikit-learn 0.17, la cual implementa este algoritmo partiendo de las librerías LIBSVM y LIBLINEAR.

2.6.2.5. Mezcla de distribuciones gaussianas

Los modelos de mezcla de distribuciones suelen ser utilizados para realizar inferencias acerca de propiedades de subpoblaciones. Parten de las observaciones de la población en general sin identificar las subpoblaciones que la conforman. Esto puede relacionarse con las técnicas de aprendizaje no supervisado, como los modelos de clustering. Se asume que las instancias del conjunto de datos son generadas por una mezcla de finitas distribuciones gaussianas con parámetros desconocidos.

Se utilizó la misma implementación de Python y scikit-learn que en los anteriores modelos. Esta utiliza el algoritmo de maximización de la esperanza (conocido como EM) para realizar la mezcla. La cantidad de componentes de esa mezcla debe definirse de antemano. Luego del entrenamiento, se clasifican nuevas observaciones asignándolas a alguno de los componentes.

2.7. Evaluación de algoritmos

Una vez seleccionados los modelos con los que se va a experimentar, resulta necesario evaluar el rendimiento de cada uno para elegir el que mejor funcione según los criterios de medición elegidos. Se supone en todo momento que nuestro conjunto de datos representa una muestra aleatoria de una distribución de probabilidades desconocida, asumiendo que esta muestra es representativa de algún modo de la población.

2.7.1. Conjuntos de datos de entrenamiento y validación

En primer lugar se separa el conjunto de datos en dos partes: una de entrenamiento y otra de validación. El objetivo es evitar introducir un sesgo optimista dado que, en caso contrario, se estaría evaluando sobre el mismo conjunto de datos sobre el cual el modelo intentó aprender. De este modo, bastaría con memorizar en lugar de generalizar para tener un mejor rendimiento. Un modelo con esas características presentaría peores resultados ante datos nuevos.

Para separar el conjunto de datos se puede proceder de manera aleatoria, asumiendo que todos nuestros datos provienen de la misma distribución de probabilidades (respecto a cada clase). Por ejemplo, se puede partir en 70 % para entrenamiento y 30 % para validación. Sin embargo, esto puede traer problemas si nuestro conjunto de datos no presenta clases balanceadas. En esta caso podría darse que el conjunto de datos de validación no contenga ninguna observación de la clase minoritaria, distorsionando las clases originales. Se sugiere realizar un muestreo estratificado para preservar el desbalanceo tanto en entrenamiento como en validación.

Para reducir la variabilidad de entrenar y evaluar siempre en la misma partición, surge la idea de iterar en este procedimiento sobre distintas particiones del conjunto de datos. Para calcular el resultado final puede considerarse el promedio de la evaluación en cada

ronda. Esta técnica se denomina validación cruzada y permite reducir esa variabilidad. En este trabajo se utilizó validación cruzada en k iteraciones. Este método divide el conjunto de datos en k subconjuntos del mismo tamaño, utilizando $k-1$ como entrenamiento y el restante como validación. Se obtienen k estimaciones, donde el resultado final será el promedio de las mismas.

En este trabajo se utilizó validación cruzada con 5 iteraciones, con muestreo estratificado. Esto se debe a la presencia de clases desbalanceadas.

2.7.2. Métricas utilizadas

En esta sección se describen las métricas que se utilizaron en esta tesis. A la hora de evaluar modelos se utilizaron las siguientes: precisión, coeficiente κ de Cohen, matriz de confusión y área bajo la curva ROC. Se procede a describir brevemente cada una.

2.7.2.1. Precisión

Esta métrica se calcula como la fracción de observaciones clasificadas correctamente sobre el total de observaciones. Más formalmente, si $\hat{y}$ es el valor predicho para la i-ésima observación y y_i es su correspondiente clase, entonces la fracción de predicciones correctas sobre el total de observaciones se define como:

$$\texttt{precision}(y, \hat{y}) = \frac{1}{n_{\text{muestra}}} \sum_{i=0}^{n_{\text{muestra}}-1} 1(\hat{y}_i = y_i)$$

donde $1(x)$ es la función indicador. Cabe destacar que esta métrica no será de mucha utilidad para el contexto del problema que se quiere resolver aquí, debido al desbalance de clases. Por ejemplo, si el 90 % de las observaciones tiene la misma clase, basta con plantear un clasificador que asigne esa clase a la totalidad de observaciones para poder tener una precisión del 90 %. Esto no resulta informativo de la capacidad de generalización del modelo.

2.7.2.2. Coeficiente kappa de Cohen

Esta medida tiene en cuenta el acuerdo entre dos observadores (en este caso el relevamiento de Techo en 2013 y las clasificaciones del modelo de detección) en sus correspondientes clasificaciones. Constan de n elementos en c categorías mutuamente excluyentes (en este caso c es igual a 2, puesto que se clasifica de manera positiva o negativa). Cabe destacar que el coeficiente resulta conservador respecto de considerar la precisión, ya que penaliza el hecho de que los dos observadores clasifiquen igual de manera azarosa (como se mencionó en el ejemplo de la precisión del 90 %). Su fórmula es:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

Donde $Pr(a)$ es el acuerdo observado relativo entre los observadores y $Pr(e)$ es la probabilidad hipotética de acuerdo al azar, utilizando los datos observados para calcular esta última. El coeficiente es un número real entre -1 y 1.

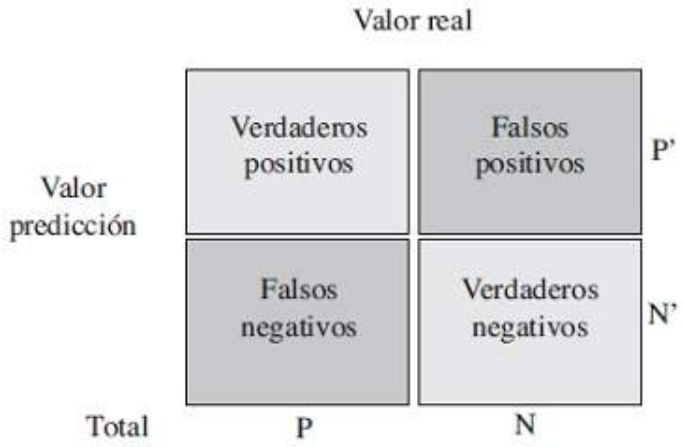
Se sugiere ver Olofsson et al. [27] para encontrar críticas al uso de esta métrica. Uno de los cuestionamientos es que está altamente correlacionada con la precisión global. De esta manera, a pesar de ser más conservador, termina siendo redundante informarlo puesto

que no modifica las conclusiones. Más allá de esto, sigue siendo el indicador comúnmente utilizado en la literatura de teledetección.

2.7.2.3. Matriz de confusión

La matriz de confusión puede definirse como aquella matriz cuya entrada i, j es el número de observaciones que pertenecen a la clase i pero cuya predicción a partir de un modelo es j.

Fig. 2.10: Matriz de confusión genérica



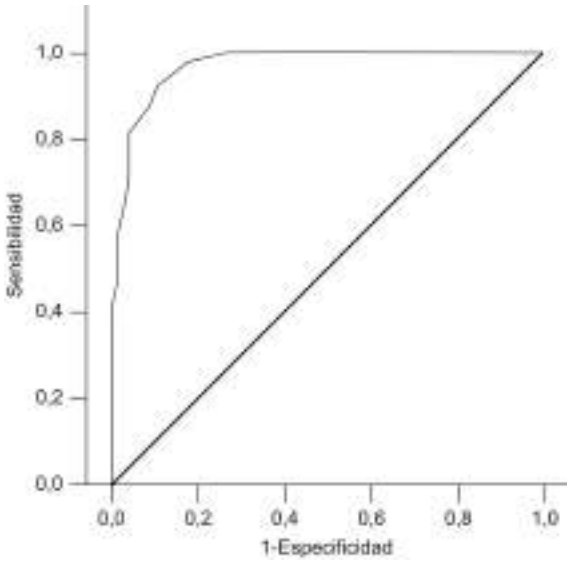
2.7.2.4. Área bajo la curva ROC

Antes de describir esta métrica se definen algunos conceptos necesarios para facilitar la explicación.

- Sensibilidad $= \frac{\text{Verdaderos positivos (VP)}}{\text{Positivos (P)}}$
- Especificidad $= \frac{\text{Verdaderos negativos (VN)}}{\text{Negativos (N)}}$
- Cociente de falsos positivos $= \frac{\text{Falsos positivos (FP)}}{\text{Negativos (N)}} = 1 - \text{Especificidad}$

De este modo, la curva ROC es la representación de la sensibilidad (en el eje vertical) frente al cociente de falsos positivos (en el eje horizontal). Se tiene en cuenta el umbral de discriminación a partir del cual se define si una observación pertenece a una clase u otra. El espacio definido a partir de esos ejes se denomina espacio ROC. Como se ve en la figura 2.11, un clasificador perfecto se encontraría en el punto $(0, 1)$ mientras que una clasificador al azar sería un punto sobre la diagonal, también llamada línea de no-discriminación.

Fig. 2.11: Representación de espacio ROC



Para el caso de los clasificadores que se utilizaron en este tesis, las implementaciones de estos devuelven una probabilidad de pertenencia a cada clase para cada observación. De este modo, fijando diferentes umbrales se podría armar un clasificador para cada umbral que puede ser representado con un punto en el espacio ROC. Se define curva ROC a la curva que une todos esos puntos, tal como se muestra en la figura 2.11.

El área bajo la curva ROC justamente se calcula a partir de la curva mencionada en el párrafo anterior y es utilizada usualmente para comparar modelos en la literatura de aprendizaje supervisado. Se la suele denominar AUC por sus siglas en inglés.

En resumen, se calibraron los modelos a partir de utilizar validación cruzada utilizando el coeficiente κ. También se calculó la precisión y el área bajo la curva ROC para ver la diferencia entre estos indicadores comúnmente usados.

2.8. Software utilizado

Para lo relacionado con teledetección (procesamiento y extracción de atributos a partir de las imágenes satelitales) se utilizó el software Orfeo ToolBox 5.0.0, que es una librería de C ++ desarrollada por el CNES (agencia espacial francesa). Es de tipo open-source, distribuido bajo la licencia CeCILL-v2. En sus comienzos este software se utilizaba para acompañar y promocionar el uso de imágenes captadas por los satélites Pléiades. El objetivo de Orfeo Toolbox es procesar imágenes de gran volumen mediante los algoritmos que forman parte del estado del arte de la teledetección. Se sugiere ver Inglada, Christophe (2009) para más detalles de implementación.

Para las tareas de procesamiento de datos georeferenciados se utilizaró el software QGIS 2.8.1 “Wien”. Este también es open-source, distribuido bajo la licencia GNU (General Public License) Version 2 o superior. QGIS está desarrollado usando Qt toolkit[7] y el lenguaje C++. El proyecto comenzó en mayo de 2002 como un visualizador de datos, pero fue evolucionando soportando cada vez más formatos, permitiendo que se le agreguen

[7] `http://qt.io`

extensiones que le den potencial de procesamiento. Cabe destacar que QGIS posee un complemento para incorporar los algoritmos de Orfeo Toolbox dentro de su plataforma.

Para el preprocesamiento de datos censales, geográficos y manipulación de imágenes procesadas se utilizó el lenguaje Python 2.7. Este permite armar esquemas de procesamiento incorporando tareas de QGIS y Orfeo ToolBox. Python es un lenguaje de alto nivel, propósito general, interpretado y dinámico, que soporta multiples paradigmas como imperativo, funcional u orientado a objetos.

3. PROCESAMIENTO Y GENERACIÓN DE ATRIBUTOS

En este capítulo se detallarán los pasos llevados a cabo para obtener los atributos con los que se experimentó finalmente.

Se presentarán las tecnicas utilizadas para calcular los atributos asociados a ejes, datos censales e imágenes satelitales. Para cada caso se mencionan los algoritmos y las decisiones tomadas según el tipo de variable.

3.1. Preprocesamiento

3.1.1. Datos censales

La menor unidad de relevamiento publicada actualmente para el Censo Nacional de Población, Hogares y Viviendas realizado por el INDEC en 2010 es el radio censal. Como se comentó anteriormente, cada una de estas unidades se encuentra dentro de una fracción censal y estas a su vez dentro de departamentos que conforman provincias. Cada radio censal tiene un código de 9 dígitos, donde los primeros dos refieren a la provincia (para el caso de Buenos Aires es 06), los siguientes tres al departamento (para el caso del Partido de La Matanza es 427), los siguientes dos a la fracción dentro de ese departamento y los últimos dos al radio dentro de esa fracción.

Para cada pregunta del cuestionario básico del censo, se calculó el porcentaje de casos para cada respuesta respecto del total de unidades censadas dentro del radio correspondiente. Por ejemplo, para la pregunta acerca de si el hogar tiene al menos un indicador de necesidad básica insatisfecha, se calculó para cada radio censal la proporción de hogares cuya respuesta es afirmativa y la proporcion cuya respuesta es negativa. De este modo, en el radio censal 064271701 el 88,93 % de los hogares no tienen ningún indicador de necesidad básica instatisfecha, mientras que 11 % sí. Cabe destacar que a partir de tener información desagregada al mayor nivel posible, se puede agrupar a fracción censal y departamento. Para el caso de la Ciudad de Buenos Aires un departamento es una comuna mientras que para la provincia de Buenos Aires es un partido, como el caso de La Matanza. Se consideraron para el procesamiento todas las preguntas del cuestionario básico[1].

Finalmente se obtuvo un conjunto de datos donde cada fila es un radio censal y cada atributo es un ítem de cada pregunta del cuestionario. Dado el objetivo de este trabajo, sólo se consideraron los radios censales correspondientes al Partido de La Matanza.

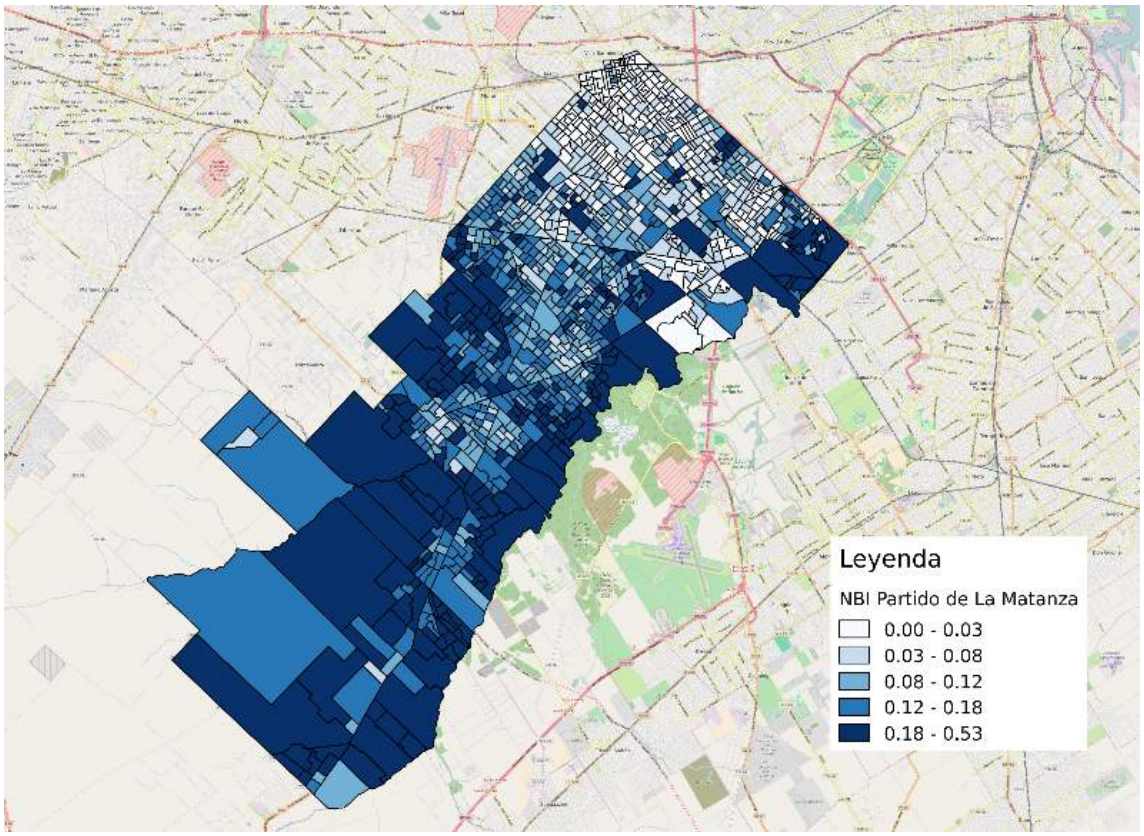
Se presenta, a modo de ejemplo, en la figura 3.1 un mapa con el indicador “alguna necesidad básica insatisfecha” para el Partido de La Matanza.

3.1.2. Imágenes satelitales

Las imágenes utilizadas para esta tesis fueron tomadas por el satélite el 22 de noviembre de 2013, siendo nula la cobertura nubosa. Resulta importante destacar que las imágenes fueron cedidas de manera gratuita por parte de Digital Globe Foundation[2] mediante la

[1] Disponible en `http://www.indec.gob.ar/ftp/cuadros/poblacion/cuestionario_basico_2010.pdf`.
[2] `http://www.digitalglobefoundation.org/`

Fig. 3.1: Variable censal Alguna Necesidad Básica Insatisfecha a nivel radio censal, Partido de La Matanza.



Fuente: elaboración propia en base a datos del INDEC. Censo Nacional de Población, Hogares y Viviendas 2010 y Cartografía y códigos geográficos del Sistema Estadístico Nacional. Buenos Aires: INDEC, 2016.

aplicación a su programa de otorgamiento de imágenes. Dicha fundación es una organización sin fines de lucro creada en 2007 con el objetivo de proveer de imágenes satelitales a estudiantes y profesores dedicados a la investigación en Geografía y Teledetección.

Se presenta en la Figura 3.2 una imagen pancromática y en 3.3 una multiespectral, asociadas de la localidad de Isidro Casanova, dentro del Partido de la Matanza. Puede verse la mayor resolución de la pancromática respecto de la otra imagen.

Como se mencionó anteriormente, el ráster asociado a cada imagen esta compuesto de números enteros denominados números digitales. La mayor parte de los métodos de análisis de imágenes tienen como objetivo establecer diferencias entre tipos de cubiertas y no tanto en caracterizarlas de manera absoluta. Por lo tanto, en primer lugar se convertirán los ND a parámetros físicos para poder hacer una medición absoluta a partir de estos valores. Otras ventajas de este procedimiento es que permite realizar con mayor naturalidad los cocientes entre bandas (será de interés a la hora de calcular algunos indices radiómetricos).

Tal como se puede ver en la documentación publicada por Digital Globe para el sensor WorldView-2[3], también resulta necesario corregir atmosféricamente las imágenes dado que la energía captada por el sensor sufre una serie de interacciones con la atmósfera antes

[3] Disponible en `http://global.digitalglobe.com/`.

Tab. 3.1: Valores de Gain y Offset para el sensor WorldView-2.

Banda	Gain	Offset
Pancromática	1,041	3,157
Coastal	0,889	11.558
Azul	1,004	9,809
Verde	1,063	4,455
Amarillo	1,049	3,408
Rojo	1,042	1,752
Rojo edge	0,953	2,671
Infrarojo 1	0,955	2,558
Infrarojo 2	0,941	2,444

Fuente: Kuester et al. 2015

de llegar al mismo. Esto provoca que la radiancia registrada por el sensor no represente fielmente la radiancia emitida y reflejada por las coberturas. Luego de llevar a cabo las correcciones, se puede proceder a extraer información cuantitativa de las mismas. Por otro lado, también se necesita verificar que la imagen este correctamente georreferenciada, lo que se denomina corrección geométrica.

3.1.2.1. Conversión de ND a reflectancia

Tal como se mencionó anteriormente, esta conversión es realizada para transformar los números digitales en valores de reflectancia de superficie.

La reflectividad es la relación entre la energía reflejada y la incidente, por lo tanto varía entre 0 y 1 donde 0 corresponde a superficie perfectamente absorbente y 1 a superficie perfectamente reflectora.

El proceso se inicia a partir de los ND grabados por el sensor, los cuales son traducciones digitales de la radiancia espectral detectada por el sensor. Para convertirlos nuevamente a valores de radiancia se calcula:

$$L_{sat} = ND * Gain + Offset$$

Las unidades en que se mide la radiancia en el sensor (L_{sat}) son $\frac{watts}{m^2 * sr * \mu m}$, es decir, la cantidad de energía ($watts$) por unidad de superficie en cada banda espectral considerando la distribución angular de la radiación (sr). Se considera que los valores de radiancia que llegan al satélite guardan una relación lineal con los valores de ND y por lo tanto $Gain$ y $Offset$ son la pendiente y la ordenada al origen de la recta de regresión que relaciona ambas magnitudes. Se transcribe en la Tabla 3.1 esa información a la fecha de realización de este trabajo, extraída de Kuester et.al (2015). El Offset se calcula a partir de la radiancia mínima ($Offset = L_{min}$) y el Gain a partir de dicha medida y la radiancia máxima ($Gain = \frac{L_{max} - L_{min}}{255}$) registradas por el sensor.

3.1.2.2. Corrección en el tope de atmósfera (TOA)

Este cálculo indica la relación existente entre la energía incidente y la reflejada, convirtiendo la radiancia de brillo en reflectancia de superficie. Se corrige parcialmente el efecto

producido por la atmósfera, considerando solamente la dispersión Rayleigh. No posee unidades al ser un cociente de irradiancias. La fórmula es:

$$\rho_{TOA} = \frac{\pi * L_{sat} * d^2}{E_0 * cos(\theta_z)}$$

Donde:

- L_{sat} = radiancia espectral.
- d = distancia Tierra-Sol en unidades astronómicas. Se calcula como: $d = 1 - 0,0167 * cos(2 * \pi * (\text{día juliano} - 3)/365)$.
- E_0 = irradiancia solar espectral a tope de atmósfera.
- θ_z = ángulo solar en el zenit (en grados).

Estas correcciones fueron realizadas utilizando el software Orfeo Toolbox a través de la aplicación *OpticalCalibration*.

3.1.2.3. Corrección geométrica

El objetivo de esta corrección es remover la distorción geométrica para que los valores de las coordenadas de las imágenes se correspondan con las coordenadas terrestres. A partir de esto se puede relacionar la información obtenida de las imágenes con datos de otras fuentes georreferenciadas.

Existen dos tipos de errores geométricos, los sistemáticos (predecibles) y los no sistemáticos (aleatorios). Los primeros son errores internos introducidos por el mismo sistema de sensores remotos o en combinación con características de la Tierra tales como rotación o curvatura. Los no sistemáticos están asociados a fenómenos introducidos que varían en tiempo y espacio relacionados con movimientos de la plataforma y relieve del suelo entre otros. Estos últimos son los más difíciles de corregir.

Las imágenes provistas por Digital Globe se encuentran georreferenciadas utilizando el datum WGS84 y el sistema de coordenadas UTM 21 S. Se verificó a partir de puntos de control (considerados en la cartografía oficial del Partido de La Matanza) la correcta georreferenciación.

3.1.2.4. Fusión de imágenes pancromáticas y multiespectrales

El proceso de fusión de imágenes pancromáticas y multiespectrales (denominado en la literatura como *pan-sharpening*) consiste en la combinación de píxeles de color de baja resolución con píxeles de alta resolución pancromáticos para generar una imagen color de alta resolución.

El objetivo de este proceso es mejorar la calidad global de la imagen resultante mediante la combinación de la información espectral de la imagen multiespectral con la información espacial de la imágen pancromática, aprovechando su caracter complementario. De este modo se genera una sola imagen de alta resolución en color. Se sugiere ver Padwick et al. (2010) para más detalles de este procedimiento para el sensor WorldView-2.

Para llevar a cabo este procedimiento se utilizó la aplicación *BundleToPerfectSensor* disponible en Orfeo Toolbox.

3.2. Segmentación de imágenes satelitales

Desde principios de la década del 70, se resolvían los problemas de clasificación en imágenes satelitales a partir de los píxeles. Para cada uno se estudiaban las propiedades espectrales sin tener en cuenta la información del contexto tanto espacial como espectral, para luego intentar clasificarlo utilizando métodos estadísticos. Luego de clasificarlos pueden formarse mosaicos de parcelas uniformes a partir de píxeles con la misma clase asignada. Sin embargo, suele ocurrir que dada la variabilidad asociada a considerar los píxeles de manera aislada, dicha uniformidad no sea tal. Esto se denomina en la literatura como efecto "sal y pimienta", donde píxeles que pertenecen a la misma clase son clasificados de manera distinta.

Como se mencionó en el párrafo anterior, solo se analiza el color (reflectancia) de cada píxel y la textura de estos pero no la forma ni la vecindad ni la ubicación de los mismos. Una de las críticas que se le hace a esta metodología pasa por no explotar dicha información, sin capturar las interacciones entre píxeles (esto se hacen más evidente en el caso de imágenes de alta resolución). En este tipo de imágenes, por ejemplo, puede suceder que el techo de una casa este compuesto por varios píxeles, los cuales en el contexto del problema que se quiere analizar deberían tener la misma clasificación. Otra crítica que se encuentra frecuentemente en la literatura pasa por comparar los métodos de clasificación con la tarea humana de interpretar una imagen, donde el observador considera la forma y las relaciones espaciales para clasificar las regiones de interés. Para analizar en mayor detalle las críticas a la clasificación por píxel puede verse Blaschke [3].

Como respuesta a esto se desarrollan técnicas para extraer objetos que agrupen píxeles contiguos (llamadas técnicas de segmentación). Este nuevo paradigma se denomina "análisis de imágenes basado en objetos" (mencionado en la literatura como OBIA por sus siglas en inglés). De este modo, el objeto más pequeño que puede considerarse es un píxel, para luego ir agrupándolos de manera jerárquica formando regiones contiguas a partir de uno o más criterios de homogeneidad en una o más dimensiones. Una de las ventajas que se pueden mencionar es que al usar objetos en lugar de píxeles se pueden calcular en estos estadísticos (como el promedio, desvío y mediana entre otros) para los valores espectrales y de textura, los cuales resumen la información que para el caso anterior se obtenía de considerar los píxeles como una unidad. Estas medidas pueden no ser informativas en el caso de que las regiones formadas no sean homogéneas.

Siguiendo a Blaschke [1], a partir de 2003 comienza a darse un incremento en la cantidad de trabajos empíricos utilizando este tipo de técnicas publicados en revistas con referato. Estos artículos mostraban mejores resultados aplicando el paradigma basado en objetos por sobre el basado en píxeles. Se sugiere ver esa publicación para encontrar ejemplos de trabajos que muestran esa mejora. También se mencionan [1] las posibilidades que presenta este tipo de análisis de interactuar con otro tipo de datos georreferenciados, permitiendo, por ejemplo, poder calcular la distancia entre un segmento y un punto de interés. Esto último se menciona en la literatura como GEOBIA y fue utilizado en este trabajo. Para más referencias sobre GEOBIA se sugiere ver Blaschke et al. [2].

En el siguientes apartado se explicará el procedimiento empleado en esta tesis para llevar a cabo la segmentación de las imágenes utilizando el algoritmo *mean-shift*.

3.2.1. Técnicas empleadas para segmentar las imágenes satelitales

Las técnicas de segmentación de imágenes no se utilizan solamente en el campo de la teledetección. También han sido utilizadas en visión por computadora, análisis de imágenes médicas o reconocimiento de rostros, entre otros. Estas surgen en la década del 80 con usos principalmente en la industria de visión por computadora, comenzando a ser utilizada en el ámbito de teledetección a finales de la década del 90.

Los algoritmos de segmentación de imágenes pueden ser divididos en tres categorías:

- basados en histograma: se calcula el histograma del color o intensidad a partir de la totalidad de los píxeles de la imagen. Se realiza el agrupamiento analizando máximos y mínimos a partir del gráfico. Este proceso puede volverse recursivo dentro de cada grupo.
- basados en detección de bordes: se busca encontrar los bordes entre las regiones para determinar los segmentos, aprovechando el fuerte ajuste en la intensidad en los límites de las regiones.
- basados en regiones: pueden dividirse entre crecimiento de regiones, fusión y división (que a su vez pueden combinarse). A partir de celdas semilla distribuidas a lo largo de la imagen, se van construyendo las regiones de acuerdo al método seleccionado. Metodologías similares a los algoritmos de agrupamiento utilizados en análisis no supervisado.

En esta tesis se utilizó un algoritmo basado en regiones denominado *mean-shift*. Siguiendo a Comaniciu y Meer [11], *mean-shift* es una técnica no-paramétrica iterativa que considera al conjunto de datos de entrada como una muestra de la función de densidad que genera esos datos. Intuitivamente, si existen agrupamientos, estos corresponderán al modo (o máximo local) de la función de densidad de probabilidad (la cual es estimada mediante funciones tipo kernel). Para cada punto, el método define una ventana en su entorno y calcula la media. Luego corre el centro de dicha ventana hacia la media, repitiendo el procedimiento hasta la convergencia (utilizando la técnica de gradiente descendente). En cada iteración se corre la ventana a una región cada vez más densa. Los puntos asociados al mismo punto estacionario pertenecerán al mismo grupo. Siendo t el número de iteración, para cada punto x_i, se procede de la siguiente manera:

1. se crea una ventana en el entorno de cada punto x_i.
2. se computa la media dentro de dicha ventana para cada punto, $m(x_i^t)$.
3. se mueve la ventana según $m(x_i^t)$ siguiendo la regla: $x_i^{t+1} = x_i^t + m(x_i^t)$.
4. repetir hasta convergencia.

La función $m(x)$ se denomina *mean-shift*. Está basada en la función de kernel elegida. Siendo $N(x_i)$ la vecindad de puntos dentro de un entorno de x_i y K el kernel utilizado, se define:

$$m(x_i) = \frac{\sum_{x_j \in N(x_i)} K(x_j - x_i) x_j}{\sum_{x_j \in N(x_i)} K(x_j - x_i)}$$

A partir de estos pasos el algoritmo determina el número de agrupamientos. Si bien es una técnica no paramétrica, requiere que se calibre el parámetro asociado a la ventana

móvil. Este influye en la cantidad de agrupamientos y en la convergencia. Cuanto menor sea la ventana que se espera obtener mayor cantidad de grupos. A medida que el conjunto de datos presente mayor cantidad de dimensiones, se podrían encontrar más rápido los máximos locales afectando la convergencia. Se sugiere ver el mencionado artículo para una descripción formal del algoritmo.

3.2.2. Implementación utilizada para segmentar las imágenes

Para procesar las imágenes satelitales se utilizó la implementación de este algoritmo en Orfeo Toolbox 5.0.0 a partir del procedimiento se segmentación *Large-Scale Mean-Shift (LSMS)*. Está basado en 4 procesos encadenados que a partir de una imagen devuelve los polígonos correspondientes a dicha imagen segmentada. Informa la media y la varianza de la radiometría de cada banda para cada polígono. Siguiendo a Michel et al. [23], se describe cada proceso:

1. se usa el algoritmo mean-shift para suavizar las imágenes, utilizando la aplicación $MeanShiftSmoothing$. En este paso se debe calibrar el parámetro asociado a la ventana del algoritmo.

2. segmentación: a partir de la imagen suavizada, se agrupan los píxeles adyacentes de acuerdo a los grupos encontrados por el algoritmo mean-shift en el paso anterior. Se utiliza la aplicación $LSMSSegmentation$. Cabe destacar que este proceso puede arrojar un gran número de segmentos.

3. fusión de segmentos pequeños (paso opcional): se puede definir un tamaño (en cantidad de píxeles) de segmento mínimo para acotar el número de regiones obtenidas.

4. vectorización: una vez segmentada la imagen, se vectoriza, convertiendo cada segmento en un polígono con la misma proyección que la imagen original. Para cada polígono se informa el promedio, varianza y la cantidad de píxeles del segmento asociado.

Se probaron varias configuraciones de parámetros para luego comprobar, inspeccionando en las imágenes, el ajuste de los segmentos a distintos tipos de objetos. Se terminó optando por tamaño mínimo de segmento de 30 píxeles.

3.3. Atributos considerados

Esta es una de las etapas más importantes de la tesis. Más allá de los algoritmos que se consideren, la elección de las variables resulta fundamental para lograr buenos resultados.

Una vez segmentadas las imágenes se procedió a calcular atributos asociados a los segmentos, para luego pasar a la etapa de clasificación. A la hora de elegir los atributos a considerar se utilizaron las variables que usualmente son calculadas en la literatura de clasificación de usos del suelo orientadas a detectar asentamientos precarios y villas. En particular se consideraron Kohli et al. [21] y Banzhaf et al. [25], donde el último aplica conceptos del anterior.

Para poder crear los atributos que se mencionan en esos artículos se utilizaron todos los conjuntos de datos que fueron caracterizados anteriormente. En esta sección se describirá el enfoque de estos autores para luego mostrar cómo se calculó cada uno de ellos.

En el primer artículo se busca identificar villas a partir de proponer una ontología basada en indicadores asociados a variables espectrales de las imágenes (para detectar características de las construcciones). También se consideran variables relacionadas al entorno de cada lugar (como caminos de acceso y locación). Los autores plantean tres tipos de niveles espaciales:

- Nivel entorno
  - Ubicación: cercanía a autopistas, caminos asfaltados y cursos de agua. Asentamientos y villas se encontrarían cerca de zonas inundables, granjas y al lado de autopistas.
  - Características de vecindad: barrios diferenciados según el nivel socioeconómico. Se encontrarían cerca de lugares con nivel socioeconómico bajo y de zonas industriales o donde se ofrezca empleo.
- Nivel asentamiento
  - Forma: forma de las manzanas, trazado de calles. Asentamientos y villas están asociados a un irregular trazado de manzanas, calles y pasillos internos.
  - Textura: densidad de techos y vegetación. Villas y asentamientos asociados a mayor densidad de techos en cada manzana y escasez de vegetación y espacios abiertos.
- Nivel vivienda
  - Construcción: material de construcción predominante, material de los techos. Villas y asentamientos suelen estar vinculados con techos pequeños y metálicos, ocupando entre 10 metros cuadrados y 40 metros cuadrados.

3.3.1. Atributos calculados a partir del conjunto de datos de ejes

Partiendo de los segmentos calculados en apartados anteriores, en esta sección se hablará del cálculo de atributos a partir de la información que aporta el conjunto de datos de ejes. Como se menciona en la sección anterior, dentro de las variables a nivel entorno se destaca la infraestructura vial en un entorno de este tipo de lugares. Se destaca allí que la cercanía de vías de tren, autopistas o grandes caminos es una característica común en las villas y asentamientos de otras ciudades en otros países. Para el caso de las vías de tren, los asentamientos más cercanos a ellas se encuentran sobre terrenos fiscales que fueron ocupándose de manera irregular, lo que conlleva a la falta de acceso a servicios básicos.

Como se menciona en el relevamiento realizado por Techo en 2013 en la provincia de Buenos Aires, el 12,6 % de los barrios informales cuenta con todas o casi todas sus calles asfaltadas, el 14 % algunas, el 19,2 % sólo la calle principal y el 54 % ninguna. Las implicancias de esto pasan por ser vulnerables en temporadas de lluvia generando inundaciones, además de complicar el acceso de vehículos de emergencia y de transporte público. Si bien en el conjunto de datos de ejes no se cuenta con información acerca si la calle está asfaltada, el hecho de contar con el ancho puede ser una aproximación.

Los atributos que se calcularon a partir de los ejes están basados en la distancia de cada segmento a cada tipo de eje más cercano de acuerdo al ancho de estos últimos. Es decir, se calculó la distancia (en kilómetros) al eje más cercano de hasta 5 metros de ancho,

hasta el más cercano de hasta 10 metros de ancho y así para cada de unos de los tipos de ejes mencionados anteriormente. A continuación se describe el procedimiento.

3.3.1.1. Cálculo de distancia mínima a cada tipo de eje

Para computar la distancia mínima de cada segmento a cada tipo de eje se utilizó la librería *Fiona 1.6.4* del lenguaje Python. Esta librería permite leer y escribir datos georreferenciados e integrarlos con otras librerías de dicho lenguaje. Se sugiere ver el sitio oficial de la misma para más información[4].

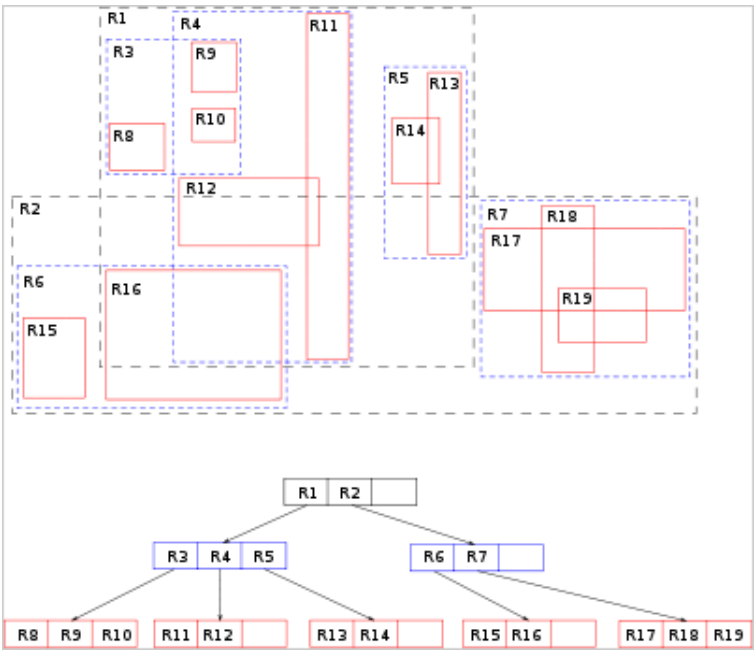
Para llevar a cabo el cálculo mencionado para un segmento y un tipo de eje en particular, un enfoque posible sería calcular la distancia de dicho segmento a todos los ejes de esa categoría y tomar la distancia mínima. El problema de esto es el tiempo de procesamiento que se necesitaría para realizarlo, puesto que si se tienen k segmentos y n ejes dentro del radio de 1 kilómetro, se necesitarán realizar $k * n$ cálculos. Dada la cantidad de segmentos y de ejes para cada tipo, este cálculo demandaría demasiado tiempo de procesamiento.

Para mejorar el tiempo de cómputo se utilizó un radio de 1 kilómetro en un entorno de cada segmento. De este modo, si este tiene un tipo de eje más cercano fuera de dicho radio, se consideró que tiene distancia mayor a uno sin realizar el cálculo para los ejes fuera del radio. Esta decisión reduce la cantidad de cálculos a realizar.

Para implementar esto con la librería *Fiona* se utilizó la estructura de datos R-tree (R refiere a rectángulo). Esta estructura es comúmente usada en bases de datos espaciales para indexar las coordenadas tanto de puntos como de polígonos, favoreciendo la velocidad de cálculo de puntos más cercanos. Su funcionamiento está basado en agrupar los objetos cercanos en el mínimo rectángulo que los envuelva, agrupándolos en forma de árbol hasta llegar a contener todos los objetos. De esta manera se aprovecha la estructura para evitar leer todos los ejes para ver cuáles se encuentran dentro del radio de 1 kilómetro de cada segmento. Se utilizó la implementación en Python de R-tree en la librería *Rtree 0.7.0*[5]. Se sugiere ver Guttman (1984) para más detalle y comparaciones con otros tipos de estructuras.

[4] `https://github.com/Toblerity/Fiona`
[5] `http://toblerity.org/rtree/`

Fig. 3.4: Ejemplo de aplicación de la estructura R-tree para rectángulos en dos dimensiones.



Fuente: R-trees: A dynamic index structure for spatial searching (A. Gutman, 1984).

A partir de lo comentado anteriormente, puede darse que un segmento tenga su correspondiente eje asociado a más de 1 kilómetro de distancia, por lo que queda fuera del radio considerado. Este segmento no va a tener un valor de distancia asociado para ese tipo de eje. Para los que no ocurre esto, hay una distancia calculada entre 0 y 1 kilómetro. Puede ocurrir entonces que dentro de una misma variable se podrían mezclar valores de variable continuos (distancias) con categóricos (aquellos con distancia mayor a 1).

Si bien esto podría no ser un problema, se decidió discretizar estas variables considerando las categorías entre 0 y 0,5 kilómetros (se los denomina categoría 0), entre 0,5 y 1 kilómetro (se los denomina categoría 1) y más de 1 kilómetro (se los denomina categoría 2). Por lo tanto, para cada tipo de eje se tendrán a lo sumo tres categorías distintas.

3.3.2. Atributos calculados a partir de variables censales

Como se muestra en la enumeración de los indicadores en [21], a nivel entorno se consideran las características socioeconómicas del lugar donde se sitúa el asentamiento o villa. Dado que no se cuenta con datos demográficos acerca de ingresos y gastos, se calculó un índice que sirva como aproximación al bienestar material de los hogares e individuos. A partir del conjunto de datos de radios censales y de las preguntas del cuestionario básico, se buscó que el índice refleje las desigualdades socioeconómicas en término de bienes y acceso a servicios. La idea detrás del indicador es poder realizar la detección de villas y asentamientos a partir de aquellos radios censales asociados a los estratos socioeconómicos más vulnerables, para poder reducir el territorio a analizar.

Revisando la literatura acerca de este tipo de indicadores, Vyas y Kuamaranayake (2006) listan los puntos débiles de esta metodología. Allí se comenta que este tipo de mediciones representan mejor la riqueza material de largo plazo que de corto, debido a

la dificultad para captar cambios de riqueza transitorios. Otra cuestión que dificulta el uso radica en la imposibilidad de captar la calidad de los bienes que posee una unidad censal. Por ejemplo, dos hogares pueden tener televisor y ser estos de muy diferente calidad, pero para este análisis cuenta que ambos coinciden en tener ese bien. Similar a este último argumento, contar con agua corriente puede ser un elemento separador en ámbitos urbanos pero no así en zonas rurales, por lo que no todas las variables tienen la misma significatividad en todos los subgrupos que pueden considerarse.

Lo que se buscó es ponderar la contribución de cada uno de los atributos considerados para determinar el grado de desigualdad asociado a la unidad censal en análisis. Existen diferentes metodologías para realizar este tipo de cálculos, como el caso de las regresiones multivariadas, análisis de factores o análisis de componentes principales. Este último método fue el empleado en esta tesis, adaptando el análisis propuesto por el Banco Mundial para las encuestas de demografía y salud (EDS, denominadas DHS en idioma inglés) en países en desarrollo.

Puede verse Falkingham, Namazie (2002), Montgomery et al. (2000), Filmer and Pritchet (2001) y Gasparini, Cicowiez y Sosa Escudero (2013) para estudiar en mayor detalle las herramientas analíticas asociadas a desigualdad y pobreza. En el siguiente capítulo se expondrán los resultados de este ańalisis reflejado en la selección de radios censales a utilizar para la posterior clasificación.

3.3.2.1. Selección de variables para índice socioeconómico

Para construir el índice socioeconómico se consideraron aquellas variables frecuentemente utilizadas en la literatura relacionada a este tópico que estén disponibles en el cuestionario básico del censo 2010. Hay que tener en cuenta que las más útiles son aquellas cuya distribución es más heterogénea entre los radios censales. Estas variables son:

- Calidad constructiva de la vivienda: insuficiente
- Calidad de los materiales: calidad II[6]
- Al menos un indicador NBI: hogares con NBI
- Condición de Actividad: Desocupado
- Nivel educativo que cursa o cursó: primario
- Nivel educativo que cursa o cursó: polimodal
- Tiene heladera: no
- Tiene computadora: no
- Tiene teléfono celular: no
- Tiene teléfono de línea: no
- Tiene baño / letrina: no
- Tipo de vivienda particular: casilla[7]

Se analizó la correlación entre las variables consideradas, para explorar cuáles son las más relacionadas y bajo que tipo de asociación (negativa o positiva). Se presentan en la Tabla 3.2 los tres pares de variables más correlacionadas de manera negativa y de manera positiva.

[6] La vivienda presenta materiales resistentes y sólidos tanto en el piso como en el techo. Y techos sin cielorraso o bien materiales de menor calidad en pisos.

[7] Vivienda con salida directa al exterior, construida originalmente para que habiten personas (sus habitantes no pasan por pasillos o corredores de uso común). Habitualmente está construida con materiales de baja calidad o de desecho y se considera propia de áreas urbanas.

Tab. 3.2: Variables censales consideradas con mayor correlación positiva y negativa

Variable 1	Variable 2	Correlación
Tiene Computadora	Tiene telefóno línea	0,92
Tiene computadora	Educ. polimodal	0,87
Tiene teléfono línea	Heladera	0,82
Tiene heladera	NBI	-0,76
Tiene computadora	NBI	-0,80
Tiene teléfono línea	NBI	-0,83

Fuente: elaboración propia en base a datos del INDEC. Censo Nacional de Población, Hogares y Viviendas 2010. Buenos Aires: INDEC, 2016.

Se presenta en la Figura 3.5 la matriz de correlaciones asociada a esas variables. Allí las variables más correlacionadas de manera positiva están asociadas a celdas rojas, mientras que para el caso negativo corresponden las celdas más azules. A partir de analizar la misma puede verse como las variables "Tiene teléfono línea", "Tiene computadora" y "Tiene heladera" se encuentran correlacionadas con atributos asociados a nivel educativo y a necesidades básicas insatisfechas. Cabe destacar que esto impacta en el análisis de componentes principales, puesto que al agregar variables que están altamente correlacionadas puede producir que se sobrestime la varianza capturada por los ejes. De este modo, se mantuvo la variable "Tiene heladera", cuya posesión está más relacionada en el largo plazo con el bienestar material que las dos variables restantes (se quitaron).

En el siguiente capítulo se expondrá el cómputo de este índice utilizando la metodología de componentes principales a partir de las variables seleccionas en este apartado.

3.3.3. Atributos calculados a partir de imágenes satelitales

Siguiendo los artículos mencionados al comienzo de este capítulo, se calcularon, a partir de las imágenes satelitales, las variables asociadas a nivel asentamiento (densidad de techos y vegetación) y a construcción de la vivienda (material de los techos y superficie que ocupa la misma).

Los índices de vegetación son combinaciones de las bandas espectrales cuya función es realzar la vegetación en función de su respuesta espectral y atenuar los detalles de otros elementos como el suelo y la iluminación, entre otros. Estos se calculan a partir de distintas bandas espectrales obteniendo una nueva imagen en la que se destacan los píxeles relacionados con coberturas vegetales. El más utilizado es el Índice de Vegetación de Diferencia Normalizada (conocido como NDVI por sus siglas en inglés) que se define a partir de un cociente entre valores del espectro infrarrojo cercano (NIR) y la región espectral roja. El índice varía entre -1 y 1, cuanto más cercano a 1 mayor relación con la presencia de cobertura vegetal.

$$NDVI = \frac{NIR - Rojo}{NIR + Rojo}$$

La densidad y la dispersión de los techos de las viviendas fue considerada a partir de 6 atributos de textura de Haralick para las bandas azul y roja, tal como se aplica en la literatura mencionada. Estos son energía, entropía, correlación, inercia (a veces llamado

contraste), diferencia de momentos inversa y correlación de Haralick. Para computar dichas métricas se utilizó el método de la matriz GLCM (siglas de Grey Level Co-occurrence Matrix) propuesto por Haralick. Para ver detalles de la implementación se sugiere la documentación de Orfeo ToolBox 5.0.0.

Tanto el índice de vegetación como las medidas de textura fueron calculados en la imagen sin segmentar. Luego, para cada segmento, se computaron la media, mediana, varianza, mínimo, máximo y desvío estándar para esos indicadores.

Para cada una de las bandas de las imágenes se calcularon los mismos estadísticos descriptivos, buscando capturar el color, la forma y la orientación de los techos. En particular, las banda roja y verde son útiles para discriminar entre agua y vegetación mientras que la azul para el caso del suelo, la vegetación y los caminos.

Por último, para capturar la influencia del tamaño de las viviendas y de los objetos se calculó el área medida en píxeles de cada segmento.

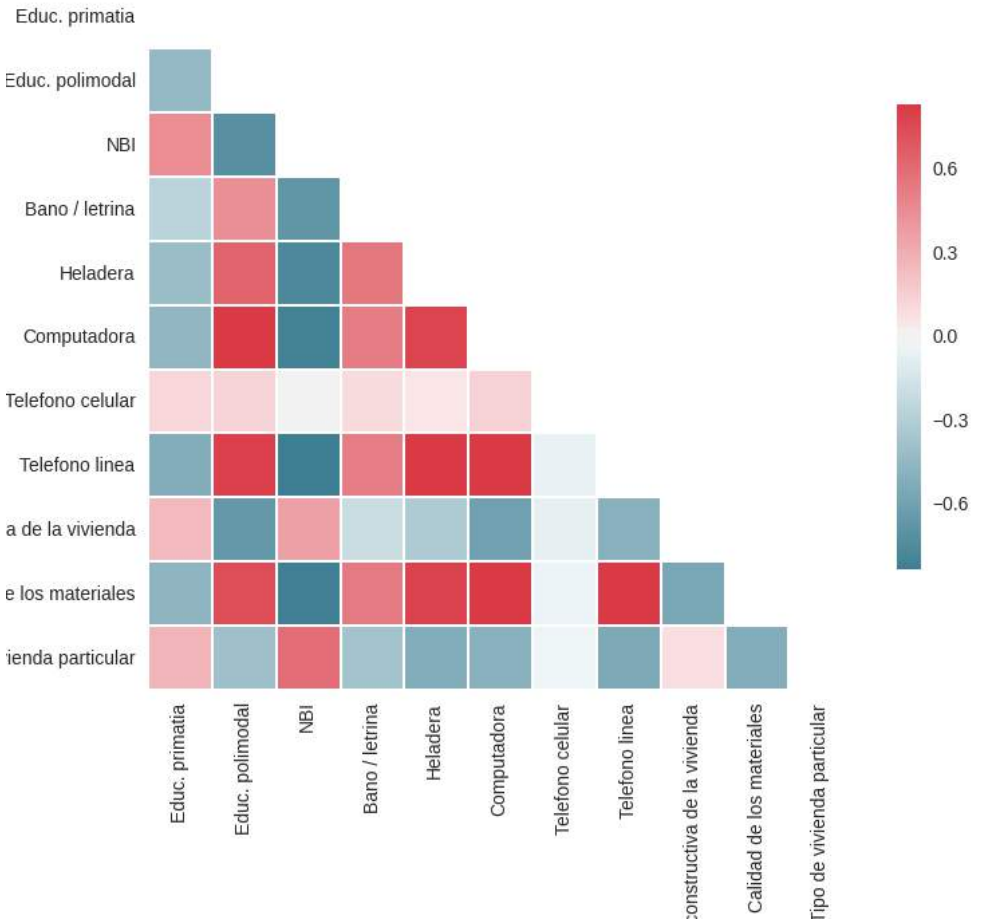
Fig. 3.2: Imagen Pancromática de la localidad de Isidro Casanova (resolución de 48cm), noviembre de 2013.

Fuente: Imágenes Cedidas por Digital Globe Foundation `http://www.digitalglobefoundation.org/`

Fig. 3.3: Imagen Multiespectral de la localidad de Isidro Casanova (resolución de 1,85 m.), noviembre de 2013.

Fuente: Imágenes Cedidas por Digital Globe Foundation `http://www.digitalglobefoundation.org/`

Fig. 3.5: Matriz de correlaciones correspondiente a las variables censales consideradas.



Fuente: elaboración propia en base a datos del INDEC. Censo Nacional de Población, Hogares y Viviendas 2010. Buenos Aires: INDEC, 2016.

4. MODELADO

En este apartado se aplicarán diferentes técnicas estadísticas y de aprendizaje automático para detectar las villas y asentamientos con los conjuntos de datos que se fueron mencionando a lo largo de la tesis.

En primer lugar se mostrarán los resultados de la clasificación utilizando imágenes, datos censales y de ejes. Luego se llevará a cabo la misma tarea pero utilizando solamente las imágenes. La intención de esto último es analizar el impacto de considerar solamente datos que estén actualizados al momento de llevar a cabo la detección.

4.1. División regional Partido de La Matanza

Para trabajar con territorios más homogéneos se dividió el partido de La Matanza en cuatro distritos. Esto se debe a la marcada diferencia en indicadores urbanos que hay a lo largo del partido (se encuentran áreas predominantemente rurales como asi también residenciales e industriales). Basado en varios proyectos de división de La Matanza que se han ido planteando [1] se muestra a continuación la división territorial que se utilizó en esta tesis.

- La Matanza: formado por Ramos Mejía, San Justo y Villa Luzuriaga.
  - Cantidad de segmentos: 1.683.399 (0.5 % villas/asentamientos).
- Los Tapiales: Lomas del Mirador, La Tablada, Ciudad Evita, Aldo Bonzi y Villa Madero.
  - Cantidad de segmentos: 1.853.200 (5 % villas/asentamientos).
- Gregorio de Laferrere: Isidro Casanova, Rafael Castillo y Laferrere.
  - Cantidad de segmentos: 2.944.163 (3 % villas/asentamientos).
- Juan Manuel de Rosas: González Catán, Virrey del Pino y 20 de Junio.
  - Cantidad de segmentos: 4.721.667 (8 % villas/asentamientos).

4.2. Cálculo de índice socioeconómico

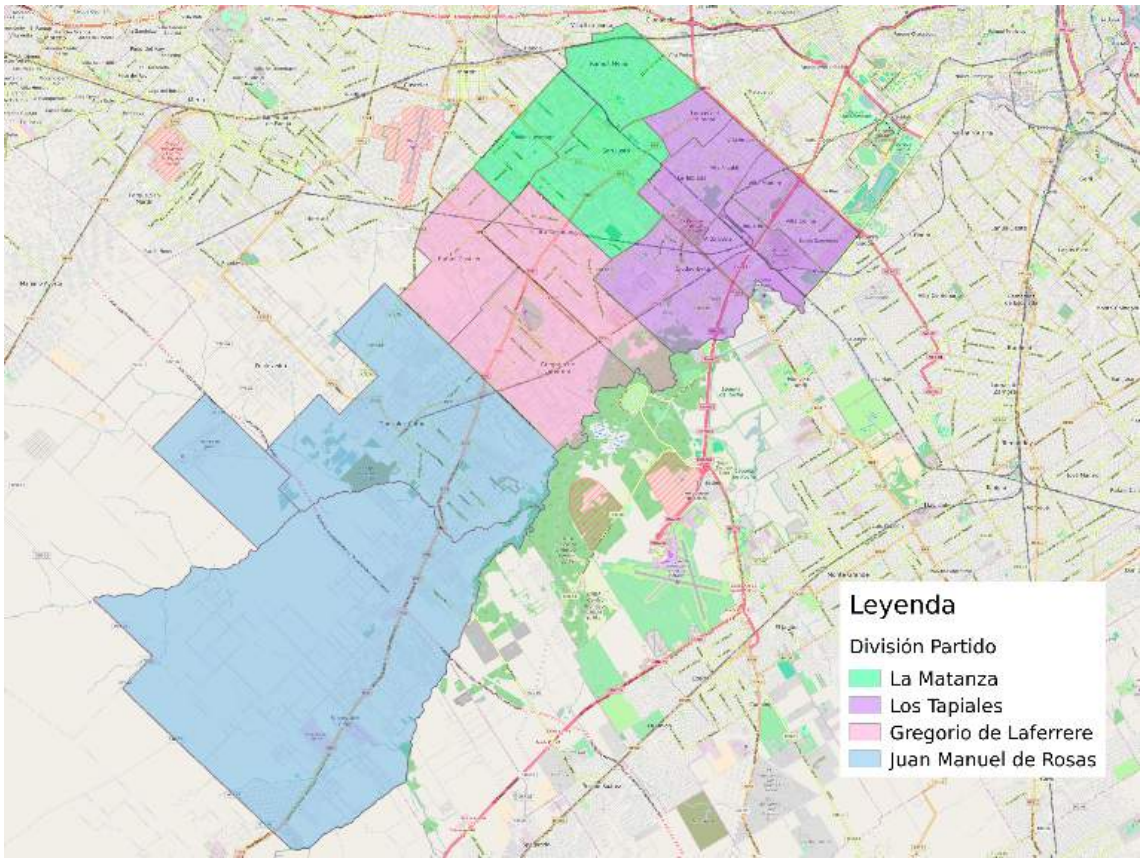
En esta sección se desarrolla el procedimiento utilizado para calcular un índice que permitió reducir el terreno a analizar. El indicador está basado en la aplicación del análisis de componentes principales presentado en la Sección 2.6.1.1.

4.2.1. Aplicación del análisis de componentes principales

Una vez calculados los componentes principales para las variables seleccionadas en la Sección 3.3.2.1, se detallan los resultados y sus implicancias.

[1] A modo de ejemplo se sugiere ver el proyecto D- 31/16-17 del Diputado Provincial Marcelo Díaz

Fig. 4.1: División propuesta para el Partido de La Matanza



Fuente: Elaboración propia en base a OpenStreetMap.

Existen varios criterios a la hora de elegir la cantidad de componentes a considerar. Por ejemplo, seleccionar aquellos cuyo autovalor asociado sea mayor a 1 o analizar el gráfico de porcentaje de varianza explicada acumulada de acuerdo a la cantidad de componentes.

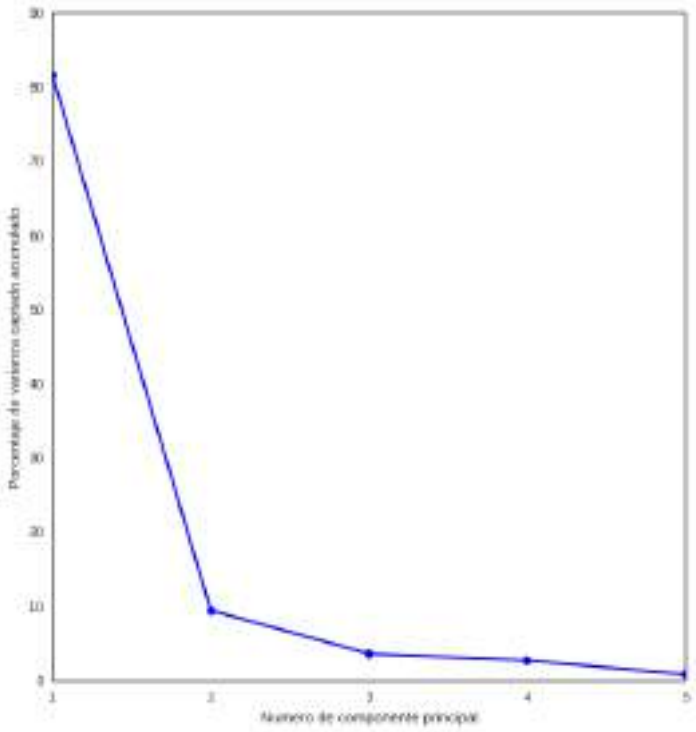
Como puede verse en la Figura 4.2, la primer componente principal captura el 81 % de la varianza, mientras que la segunda y las subsiguientes presentan valores menores a 10 %, de modo que alcanza con utilizar solo la primera. Las ponderaciones asociadas se muestran en la Tabla 4.1.

Analizando los resultados, las variables que más impacto tienen son “Calidad de los materiales”, “NBI” y “Calidad constructiva de la vivienda”. Ponderaciones altas generalmente contribuyen a valores más altos del índice. Por ejemplo, cuando el porcentaje de hogares con NBI sea mayor en el radio censal correspondiente, mayor puntaje tendrá el índice. Siguiendo este razonamiento, cabe esperar que si se divide el índice en quintiles, el bienestar material va a ser decreciente, lo que tiene sentido dadas las variables consideradas.

4.2.2. Clasificación en grupos socioeconómicos

Luego de aplicar la primer componente principal al conjunto de datos, obtenemos cada variable transformada. Esto arroja un valor a modo de puntaje para cada radio censal.

Fig. 4.2: Porcentaje de varianza captado acumulado según el número de componente principal.



Fuente: elaboración propia en base a datos del INDEC. Censo Nacional de Población, Hogares y Viviendas 2010. Buenos Aires: INDEC, 2016.

A partir de estos puntajes se dividirá en 5 grupos (quintiles) para obtener el grupo más rico, el más pobre y a los 3 grupos intermedios restantes. En la Tabla 4.2 se muestran los promedios asociados a cada variable según el grupo de pertenencia. En la Figura 4.3 se encuentra el mapa asociado a esta clasificación.

4.2.3. Clasificación radios censales que contienen villas y asentamientos

En este apartado se cuantificará el impacto de reducir el total de radios censales a analizar a partir de los resultados que arroja el índice socioeconómico elaborado en la sección anterior. En la Tabla 4.3 se muestra el porcentaje de área de villas y asentamientos que captura cada quintil del índice. Con esa información se puede evaluar la pérdida asociada en favor de la reducción del área para analizar. Como se ve en esa tabla, considerando solamente los quintiles 4 y 5 se captura aproximadamente el 94 % del área de las villas y asentamientos del partido con solo analizar el 48 % del territorio. Esto permite reducir el análisis considerablemente. En las Figuras 4.4 y 4.5 se muestra esta información a través de mapas.

Del análisis de estos mapas puede verse que la mayor cantidad de radios censales capturados están al sur del partido, dejando solamente un solo asentamiento de lado. Esta característica se va perdiendo conforme se avanza hacia el norte, lo que da la pauta de

Tab. 4.1: Promedio, desvión estándar y ponderación en 1era componente principal según variable.

Variable	Promedio	Desvío Estándar	Ponderación
Condición de Actividad	0,51	0,05	-0,13
Educación Primaria	0,04	0,01	0,02
Educación Polimodal	0,05	0,03	-0,11
NBI	0,11	0,09	0,35
Baño \Letrina	0,97	0,03	-0,07
Calidad constructiva	0,22	0,09	0,25
Calidad de los materiales	0,51	0,19	-0,87
Tipo de vivienda particular	0,01	0,02	0.05
Heladera	0,96	0,04	-0,16

Fuente: elaboración propia en base a datos del INDEC. Censo Nacional de Población, Hogares y Viviendas 2010. Buenos Aires: INDEC, 2016.

Tab. 4.2: Promedio de variables consideradas de acuerdo a pertenencia a grupo socioeconómico según análisis de componentes principales.

Variable	1er quintil	2do quintil	3er quintil	4to quintil	5to quintil
Condición de actividad	0,55	0,53	0,50	0,48	0,47
Educación primaria	0,04	0,41	0,42	0,46	0,51
Eduacación polimodal	0,09	0,06	0,04	0,03	0,01
NBI	0,02	0,06	0,1	0,15	0,23
Calidad constructiva	0,11	0,20	0,26	0,27	0,29
Calidad de los materiales	0,77	0,64	0,53	0,40	0,22
Tipo de vivienda particular	0	0,01	0,02	0,02	0,04
Baño \letrina	0.99	0.98	0.98	0.97	0.95
Heladera	0.99	0.98	0.96	0.94	0.89

Fuente: elaboración propia en base a datos del INDEC. Censo Nacional de Población, Hogares y Viviendas 2010. Buenos Aires: INDEC, 2016.

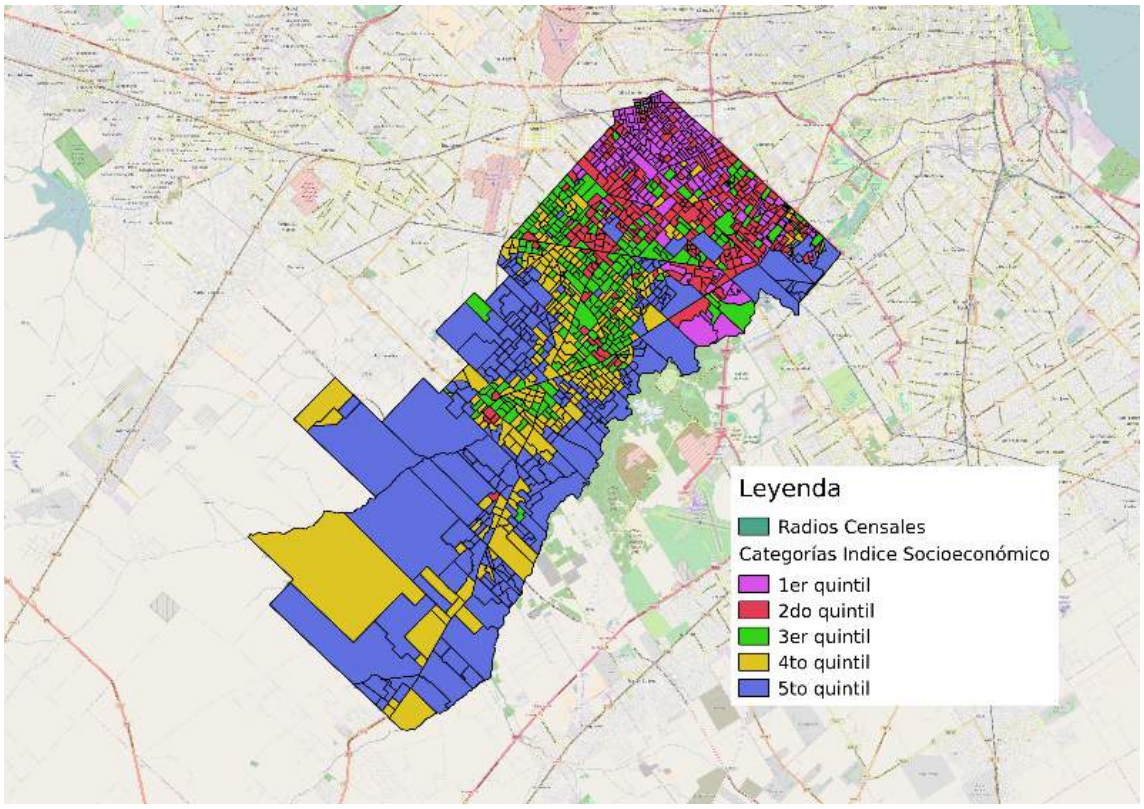
que funciona mejor para zonas con menor densidad de población. Una de las razones de este comportamiento radica en que, para el caso del Partido de La Matanza, las villas y asentamientos que se encuentran en territorios más densamente poblados suelen tener un tamaño menor respecto de los demás. Por lo tanto, el porcentaje de área respecto al radio censal de pertencencia es lo suficientemente bajo como para que las variables asociadas a esos radios reflejen adecuadamente la situación de los territorios más vulnerables.

Este análisis se utilizó una vez calibrados los modelos de clasificación para cada división del territorio: se asignó de antemano como no villa todo aquel segmento que este por fuera de los radios censales seleccionados.

4.3. Elección de modelos utilizando un territorio de muestra

Dado el tamaño del conjunto de datos resultante y la necesidad de probar diferentes configuraciones de parámetros para cada modelo, se utilizó una porción del territorio para

Fig. 4.3: Quintiles del índice socioeconómico asociados a cada radio censal del Partido de La Matanza



Fuente: elaboración propia en base a datos del INDEC (Censo Nacional de Población, Hogares y Viviendas 2010).

poder calibrarlos. A partir de este resultado, se entrenaron los mejores en cada una de las divisiones territoriales que se definieron. La porción comprende parte de las localidades de Tapiales, Aldo Bonzi, Ciudad Evita, Isidro Casanova y Gregorio de Laferrere (que corresponden a las subdivisiones de Los Tapiales y Gregorio de Laferrere respectivamente).

Fue elegida para que contenga no solamente segmentos de áreas urbanas sino también de áreas rurales para capturar la diversidad que presenta el partido. Presenta:

- Cantidad de segmentos: 411.702
- Cantidad de segmentos villas / asentamientos: 82.545 (20 %)

En la Tabla 4.4 se presentan los resultados de las experimentaciones con cada uno de los modelos. La idea es calibrar los parámetros de cada algoritmo maximizado el coeficiente κ. Luego se desarrollará para cada modelo los pasos que se siguieron para llevar a cabo la calibración.

Como se ve allí, los modelos que mejor κ presentan son XGBoost y Random Forests, aunque los otros dos los siguen muy de cerca. Esto muestra que el poder de detección viene dado en gran parte por los atributos considerados y en menor medida por los algoritmos elegidos.

Tab. 4.3: Porcentaje de área de Partido de La Matanza y porcentaje de villas y asentamientos según Techo (2013), simples y acumulados, capturado por cada quintil de índice socioeconómico.

Quintil	Área quintil (%)	Área VyA (%)	Área quintil acum.(%)	Área VyA acum.(%)
5	48,38 %	78,81 %	48,38 %	78,81 %
4	22,70 %	15,17 %	71,08 %	93,98 %
3	10,97 %	4,11 %	82,05 %	98,09 %
2	9,38 %	1,18 %	91,14	99,28 %
1	8,57 %	0,72 %	100 %	100 %

Fuente: elaboración propia en base a datos del INDEC (Censo Nacional de Población, Hogares y Viviendas 2010) y Techo 2013.

Tab. 4.4: Medidas de precisión, área bajo la curva ROC y coeficiente κ según el algoritmo

Modelo	Precisión	AUC	κ
Random Forests	93 %	90 %	0.79
XGBoost	94 %	91 %	0.81
SVM	95 %	92 %	0.81
GMM	91 %	90 %	0.80

Analizando las variables más importantes para esos dos modelos, en Random Forests tienen mayor importancia las variables binarias asociadas a la distancia a cursos de agua y distintos anchos de calle mientras que para XGBoost tienen mayor importancia aquellas relacionadas con atributos de las imágenes (como el caso del NDVI). De este modo puede verificarse la distinción en el cómputo de importancia de variables según cada algoritmo. En el caso de XGBoost, al considerar el atributo asociado a la distancia de un curso de agua, luego deja de tener en cuenta las demás variables de distancia debido a que estas están altamente correlacionadas entre sí. En vista de los resultados para ese algoritmo, se evidencia la presencia reiterada de medidas asociadas al NDVI y a entropía, contraste y correlación. Estas superaron a las relacionadas con los estadísticos descriptivos de las ocho bandas de la imagen, validando las conclusiones de Kohli, Kerle y Sliuzas (2012) respecto de su utilidad.

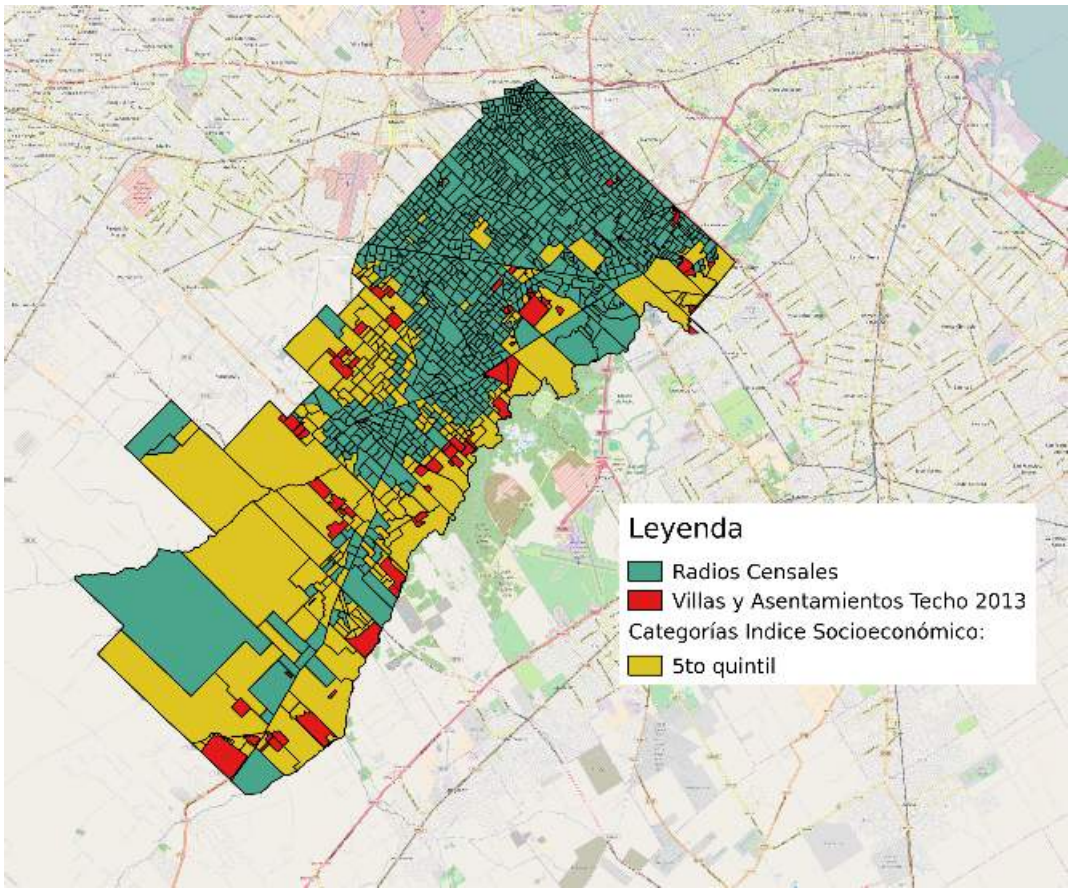
4.3.1. Experimentación con Random Forests

La implementación de este algoritmo en la librería scikit-learn permite configurar varios parámetros[2], se experimentó con los siguientes:

- Máxima profundidad de los árboles.
- Mínima cantidad de observaciones para dividir un nodo interno.
- Criterio de medición de calidad de una partición: medida de la impureza de una partición. Se experimentó con la medida de impureza de Gini y con ganancia de información (también llamada entropía).

[2] Se sugiere ver la documentación de scikit-learn para ver la totalidad de estos y sus posibles configuraciones

Fig. 4.4: Polígonos radios censales asociados a 5to quintil de índice socioeconómico y villas y asentamientos según Techo (2013)



Fuente: elaboración propia en base a datos del INDEC (Censo Nacional de Población, Hogares y Viviendas 2010) y Techo 2013.

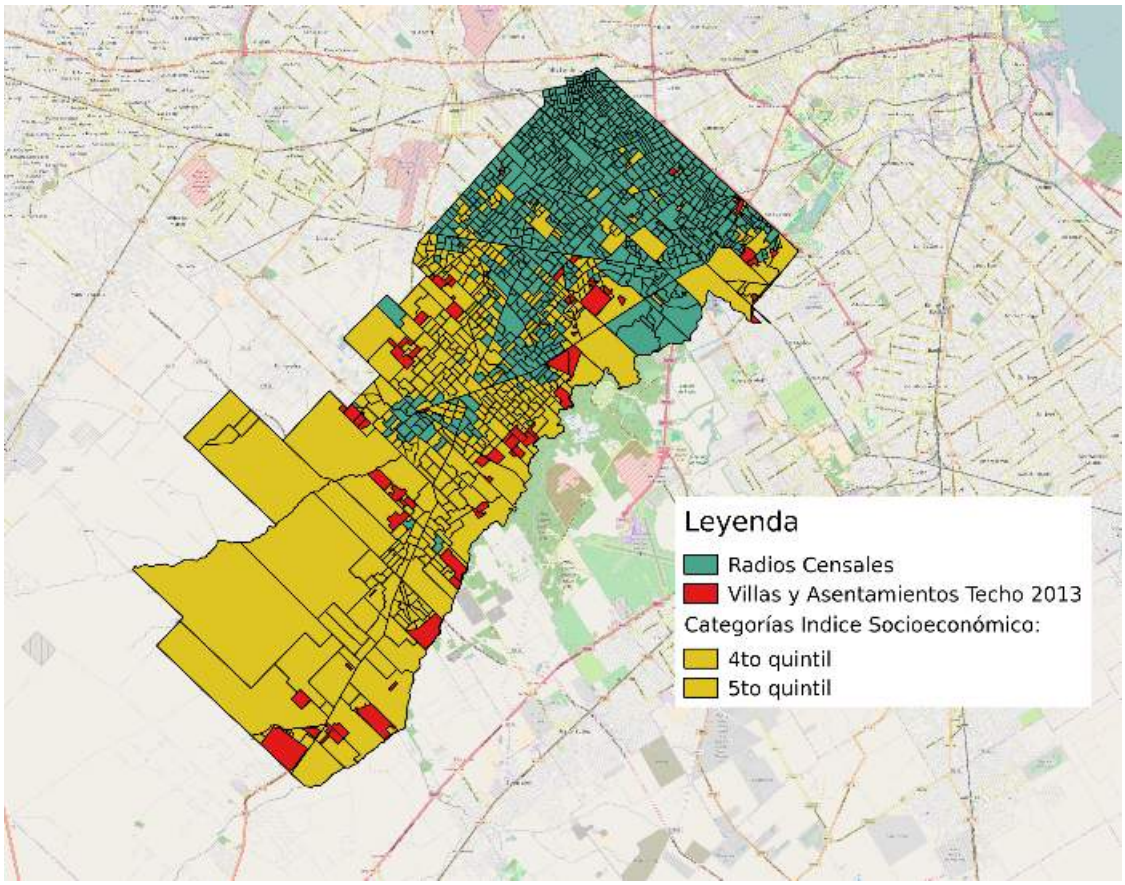
Se buscó llegar a una configuración de parámetros tales que el modelo no sobreajuste los datos de entrenamiento, logrando el mayor coeficiente κ posible. Se presenta en las Figuras 4.9 y 4.10 gráficos con pruebas para los dos primeros.

Puede verse en el gráfico de máxima profundidad como el modelo sobreajustó a medida que aumenta la cantidad máxima permitida. Esto se evidencia en el incremento del coeficiente κ para el conjunto de datos de entrenamiento comparado con la estabilidad del mismo para el conjunto de prueba. Puede explicarse por la mayor complejidad que adquiere el modelo a medida que los árboles van tomando mayor profundidad, por lo que empieza a haber menos sesgo pero más varianza en las clasificaciones.

Para el caso de la cantidad mínima de observaciones para dividir un nodo interno, cuantas más observaciones se le pidan al algoritmo para poder partir un nodo, menos profundidad adquiere el árbol. Por ello, a medida que esa cantidad aumente, menor complejidad tendrá el modelo.

Respecto de utilizar la medida de impureza de Gini o ganancia de información, los resultados en ambos casos fueron similares, presentando un κ de 0.85 para entrenamiento y 0.80 para prueba.

Fig. 4.5: Polígonos radios censales asociados a 4to y 5to quintil de índice socioeconómico y villas y asentamientos según Techo (2013)



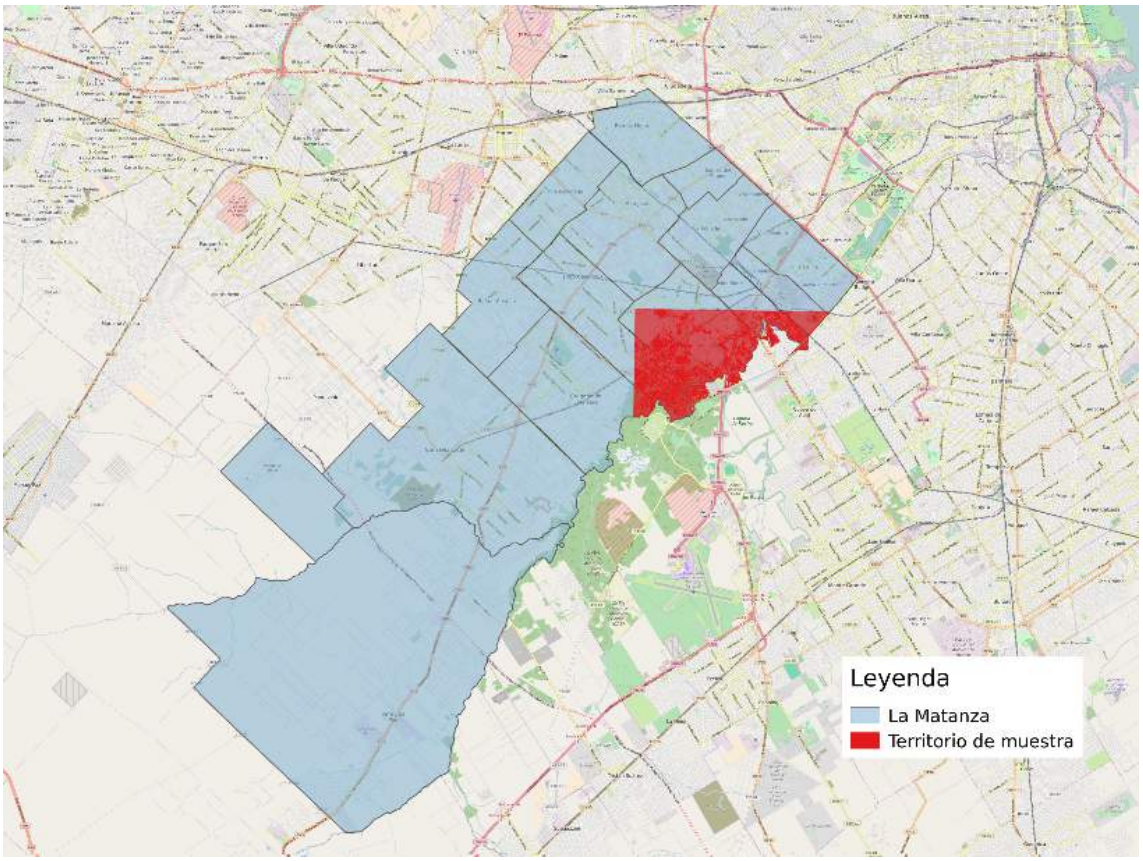
Fuente: elaboración propia en base a datos del INDEC (Censo Nacional de Población, Hogares y Viviendas 2010) y Techo 2013.

A partir de lo comentado anteriormente, se decidió utilizar una máxima profundidad entre 10 y 15 y una mínima cantidad de observaciones por nodo entre 5 y 20. El modelo seleccionado será aquel que obtenga un mejor coeficiente κ a partir de medir utilizando la técnica de validación cruzada. En la Tabla 4.5 se presenta esa información.

Tab. 4.5: Coeficientes *kappa* para modelos entrenados con validación cruzada

Profundidad	División	*kappa*
10	5	0.80
10	10	0.80
10	15	0.80
10	20	0.79
15	5	0.81
15	10	0.81
15	15	0.81
15	20	0.81

Fig. 4.6: Territorio de muestra considerado para el entrenamiento de los modelos.



Fuente: Elaboración propia.

En la misma se ve una gran paridad entre todos los modelos, por lo que se utilizó el de profundidad 15 y el de mínima cantidad para división de 20.

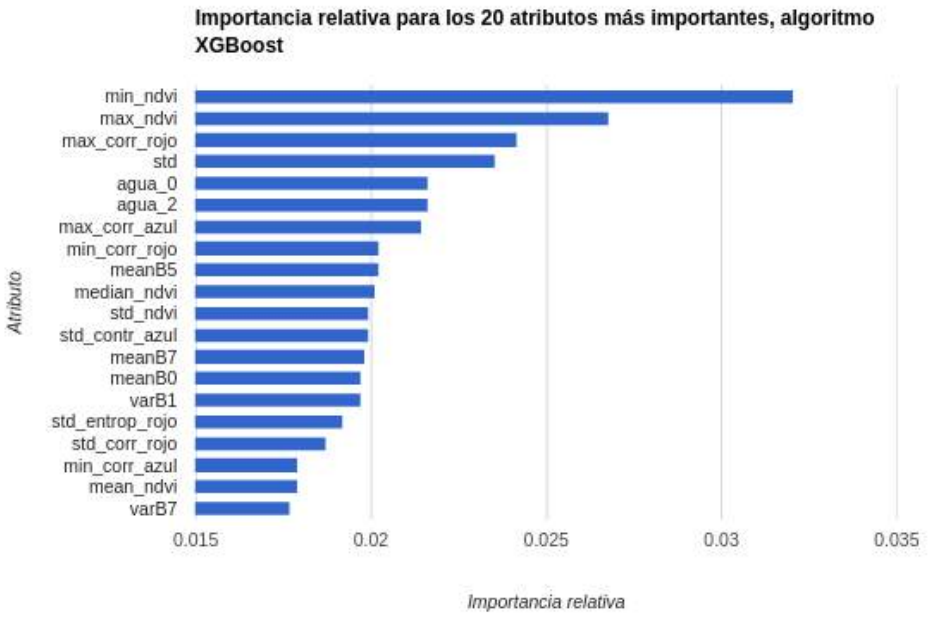
4.3.2. Experimentación con XGBoost

Para el caso del algoritmo XGBoost se experimentó con los siguientes parámetros:

- Máxima profundidad de los árboles.
- Gamma: un nodo de un árbol será dividido si la partición resultante provoca una reducción en la función de pérdida. El parámetro gamma especifica la mínima pérdida aceptada para poder realizar la división. Cuanto más grande, más conservador el algoritmo.
- Porcentaje de atributos: porcentaje de los atributos que se seleccionarán aleatoriamente para desarrollar cada árbol.

Para el caso de la máxima profundidad de los árboles, se vió un comportamiento similar al de Random Forests. Al aumentar la profundidad máxima de cada árbol este sobreajustó cada vez más los datos de entrenamiento en perjuicio del conjunto de prueba.

Fig. 4.7: Importancia relativa de los primeros 20 atributos más importantes a partir del algoritmo XGBoost



Fuente: Elaboración Propia.

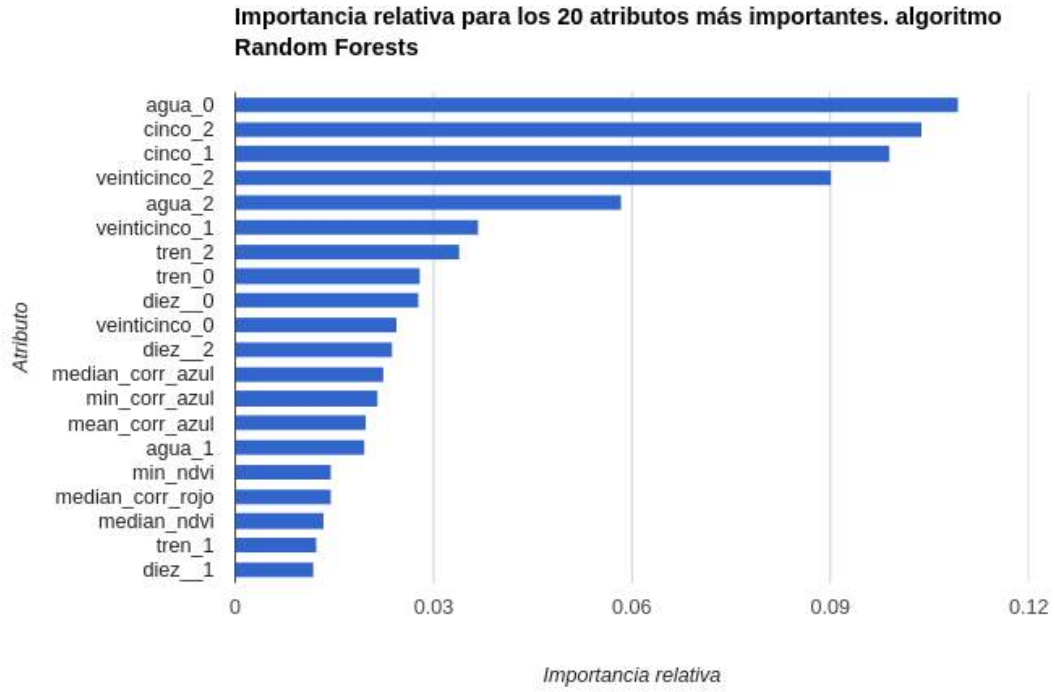
En la figura 4.11 se muestran los resultados. A partir del análisis del mismo se optó por utilizar una máxima profundidad de 4 y 5 para evaluar utilizando validación cruzada.

Respecto del parámetro gamma, se encontró que los valores 0.1 y 0.2 muestran un rendimiento levemente superior en la etapa de entrenamiento (en el conjunto de datos de prueba es pequeña la mejoría).

Por último, para el porcentaje de atributos para cada árbol, se experimentó con valores entre 60 % y 90 %. Allí el κ de entrenamiento y prueba presentó variaciones del orden del tercer decimal (alrededor de 0.84 para entrenamiento y 0.80 para prueba). De este modo se calibró en validación cruzada utilizando 70 %, que arrojó 0.84 para entrenamiento y 0.81 para prueba.

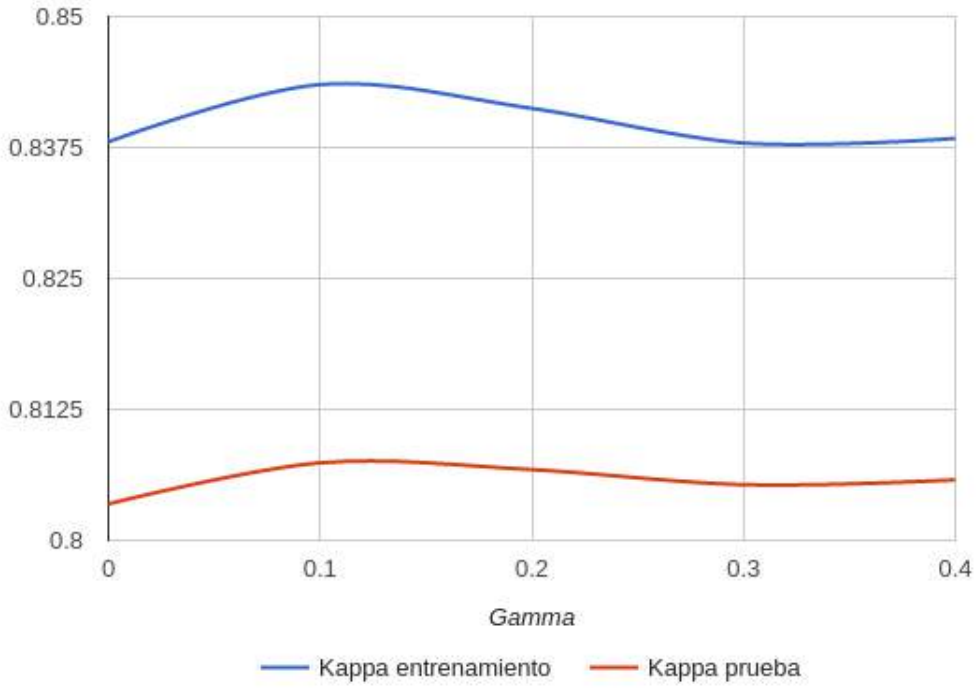
Como se ve en los resultados, no se presentan grandes variaciones a medida que se modifican los parámetros. Esto respalda lo mencionado acerca de la importancia del conjunto de datos por sobre los algoritmos considerados.

Fig. 4.8: Importancia relativa de los primeros 20 atributos más importantes a partir del algoritmo Random Forests



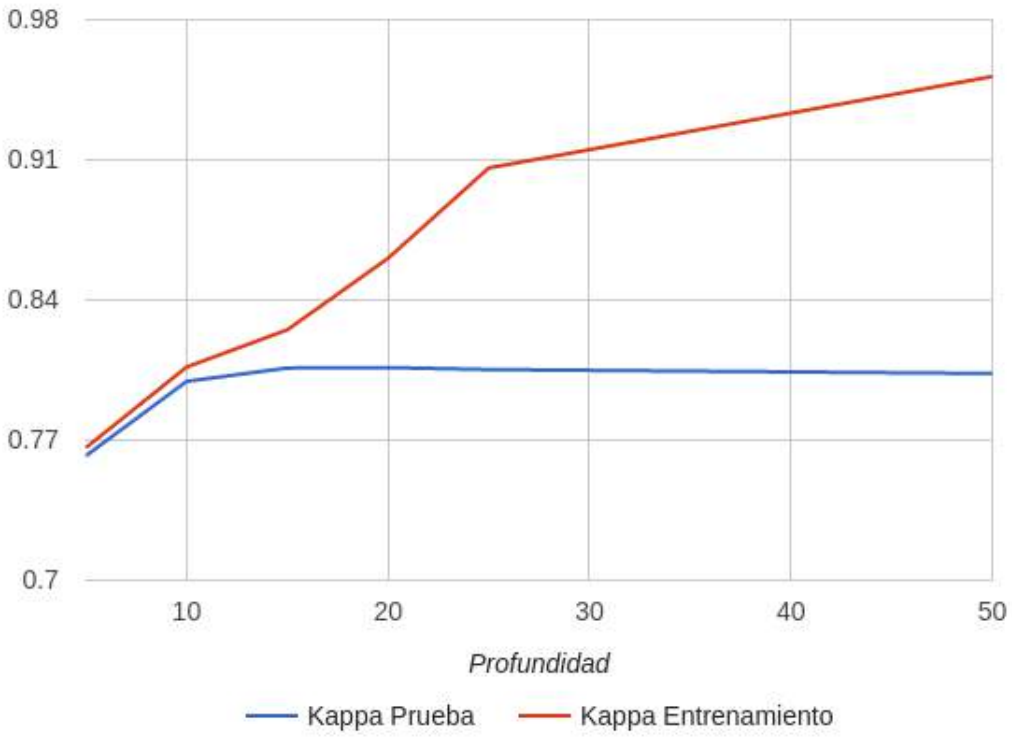
Fuente: Elaboración Propia.

Fig. 4.12: Coeficiente κ según el parámetro gamma para el conjunto de datos de entrenamiento y validación



Fuente: Elaboración Propia.

Fig. 4.9: Coeficiente κ según la máxima profundidad del árbol para el conjunto de datos de entrenamiento y validación



Fuente: Elaboración Propia.

Se muestran en la Tabla 4.6 los valores obtenidos de evaluar con la metodología de validación cruzada. Se ve un mejor desempeño utilizando árboles de 5 de profundidad, donde no tiene incidencia en este caso la elección de gamma. De este modo se utilizó máxima profundidad 5, gamma 0.1 y 70 % de los atributos para desarrollar cada árbol.

Tab. 4.6: Coeficientes *kappa* para modelos entrenados con validación cruzada

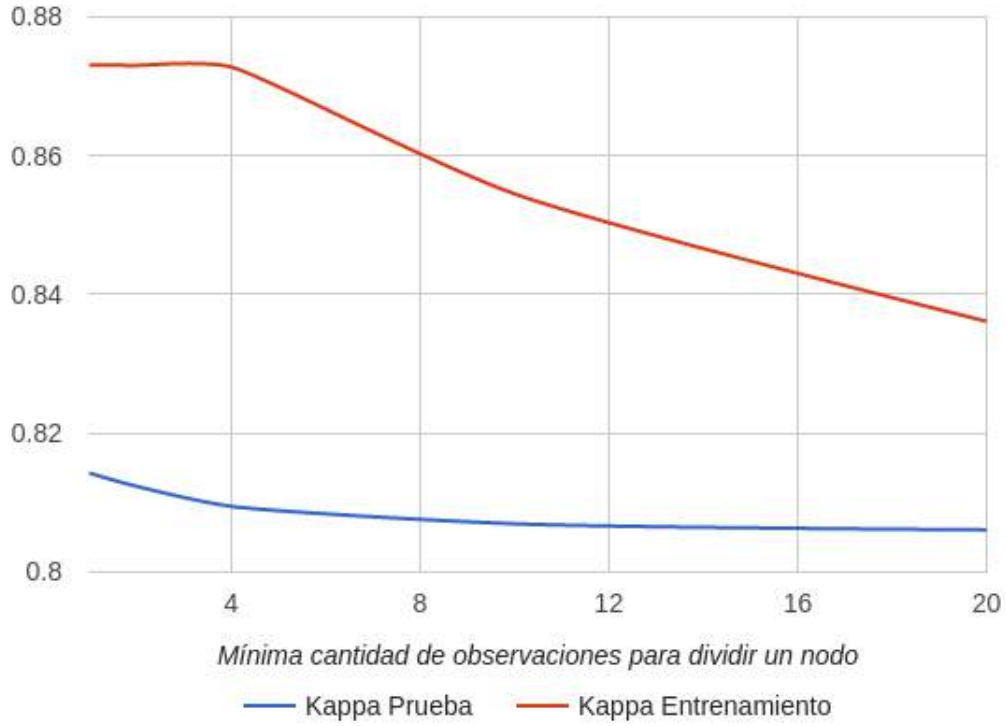
Profundidad	División	*kappa*
4	0.1	0.78
4	0.2	0.78
5	0.1	0.81
5	0.2	0.81

4.3.3. Experimentación con Máquinas de Vectores de Soporte (SVM)

Para el algoritmo SVM se experimentó modificando los siguientes parámetros que ofrece la implementación utilizada:

- Funciones kernel: polinómica, radial y lineal

Fig. 4.10: Coeficiente κ según la mínima cantidad de observaciones para dividir un nodo interno para el conjunto de datos de entrenamiento y validación



Fuente: Elaboración Propia.

- Parámetro C: tiene en cuenta los errores de clasificación y la complejidad de la superficie de decisión del modelo. Cuanto más alto, mayor tasa de clasificación correcta de observaciones de entrenamiento. Cuanto más bajo, más simple la superficie de decisión.

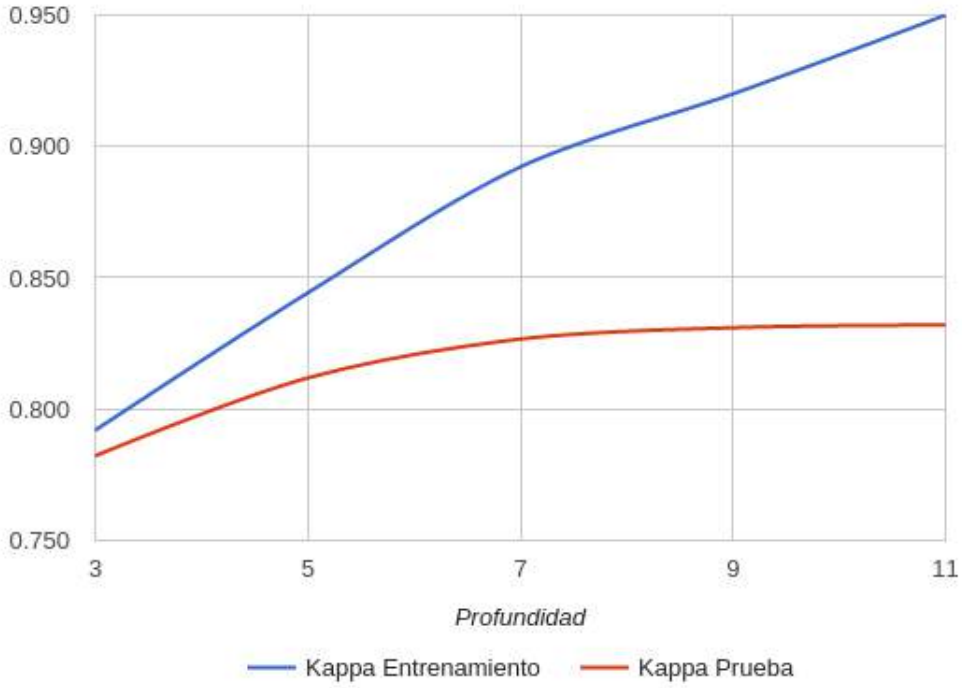
De la Figura 4.13 surge que el mejor rendimiento se dió para las funciones radial y polinómica, aunque la primera parece mucho más estable. De este modo se evualuaron ambas funciones utilizando validación cruzada con C tomando los valores 1, 10 y 100.

Se ve en los resultados de la validación cruzada una paridad para el caso de $C = 100$. En los demás valores prevalece el kernel radial, por lo que fue elegido junto con $C = 100$.

Tab. 4.7: Coeficientes *kappa* para modelos entrenados con validación cruzada

Kernel	C	*kappa*
polinómico	1	0.75
polinómico	10	0.77
polinómico	100	0.80
radial	1	0.78
radial	10	0.79
radial	100	0.80

Fig. 4.11: Coeficiente κ según la máxima profundidad del árbol para el conjunto de datos de entrenamiento y validación



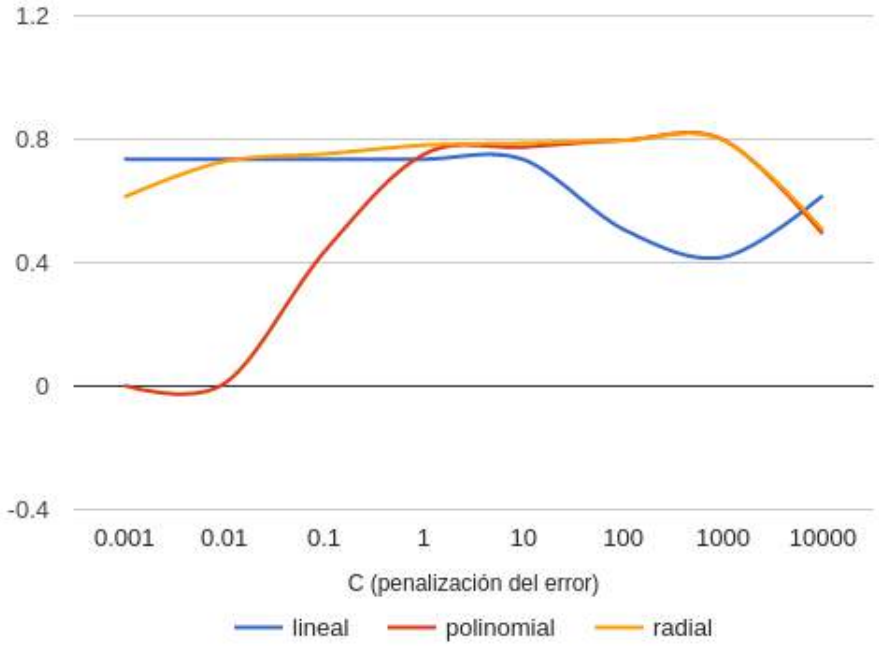
Fuente: Elaboración Propia.

4.3.4. Experimentación con mezcla de distribuciones gaussianas (GMM)

En este caso se experimentó con la cantidad de componentes que se mezclaron, utilizando matriz de covarianzas diagonal. La implementación devuelve un componente para cada observación, por lo que hay que transformar este resultado para obtener una clase binaria y adaptarlo al problema. Se propone, para cada componente, que si el número de instancias con clase villa/asentamiento es mayor a la proporción de estas que hay en el conjunto de datos, entonces todos las observaciones que lo integran serán categorizadas como villa/asentamiento.

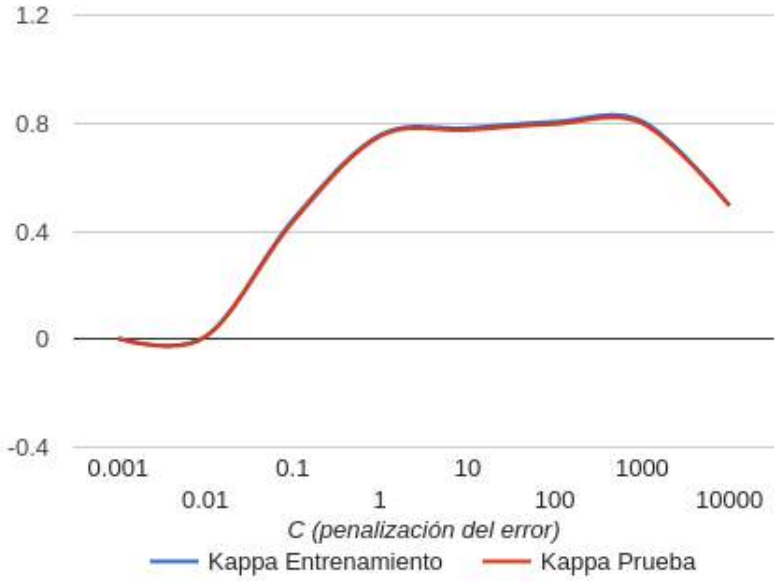
Se experimentó con la cantidad de componentes entre 20 y 100. Se observó que el coeficiente κ para el conjunto de entrenamiento se encuentra estable a medida que crece la cantidad de componentes. De este modo, se seleccionó el modelo con 80 componenentes para calibrar en las subdivisiones territoriales.

Fig. 4.13: Coeficiente κ según valor de C para el conjunto de datos de prueba según el tipo de función kernel



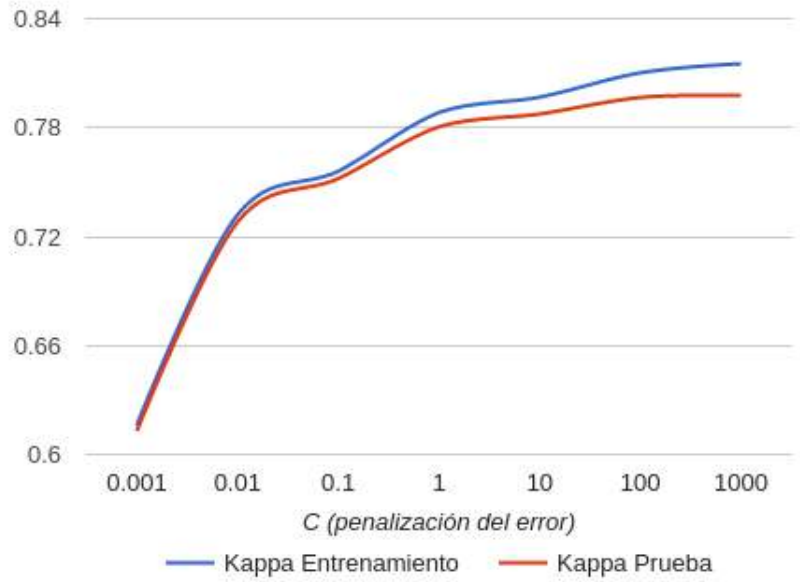
Fuente: Elaboración Propia.

Fig. 4.14: Coeficiente κ según valor de C para el conjunto de datos de entrenamiento y prueba para la función kernel polinómico



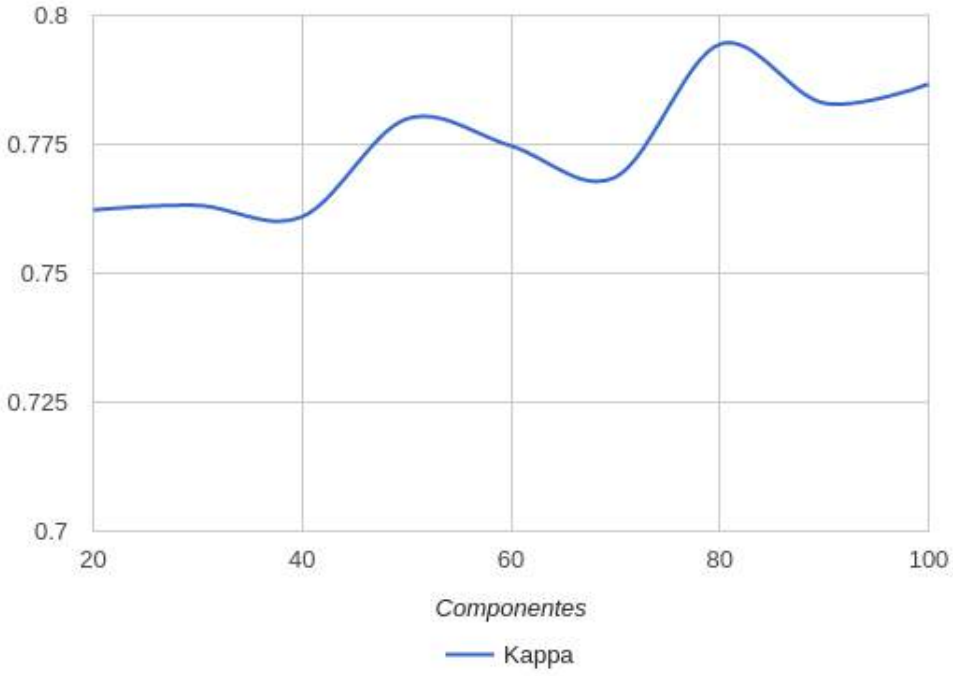
Fuente: Elaboración Propia.

Fig. 4.15: Coeficiente κ según valor de C para el conjunto de datos de entrenamiento y prueba para la función kernel radial



Fuente: Elaboración Propia.

Fig. 4.16: Coeficiente κ según cantidad de componentes para el conjunto de datos de prueba



Fuente: Elaboración Propia.

5. RESULTADOS CONSIDERANDO IMÁGENES SATELITALES Y DATOS GEORREFENCIADOS

En este capítulo se presentarán los resultados de utilizar la configuración de los modelos calibrados con la imagen de muestra 4.6 aplicados a cada subdivisión propuesta de La Matanza. Se informará, para cada región, el coeficiente κ, la precisión y el área bajo la curva ROC. También se mostrará en un mapa el resultado de la clasificación utilizando la matriz de confusión.

Cabe destacar que los resultados contemplan la aplicación del índice socioeconómico, de modo que los segmentos correspondientes a radios censales que hayan sido excluidos por dicho indicador son clasificados de manera negativa.

Los mapas que se presentan en cada caso (también los del próximo capítulo) están hechos en base a los resultados de la experimentación con el algoritmo XGBoost, puesto que es el que mejor resultados presentó.

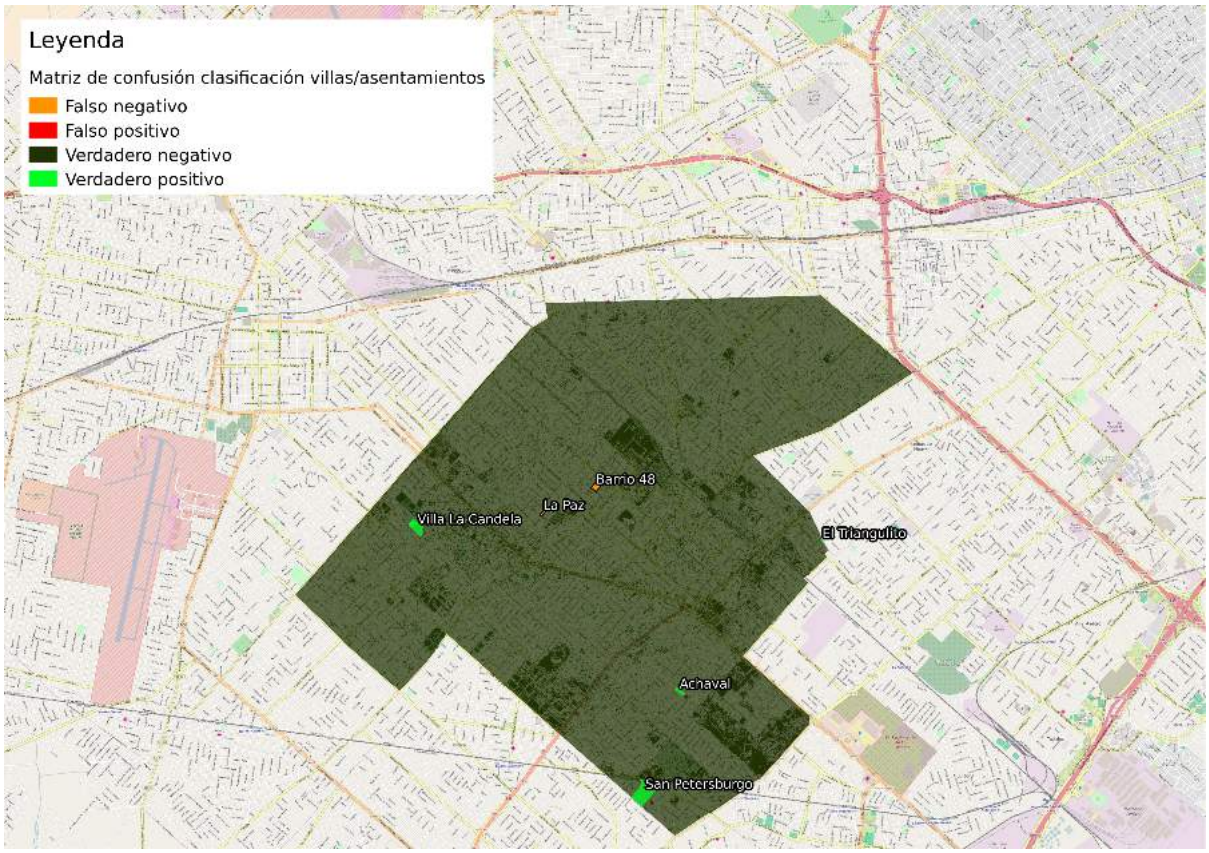
5.1. Región La Matanza

Esta región presenta la particularidad de que la villa Barrio 48 y el asentamiento La Paz se encuentran en radios censales excluídos por el índice socioeconómico, de modo que los segmentos correspondientes son falsos negativos. En la Tabla 5.1 se presentan los resultados de la clasificación. En la Figura 5.1 se muestran los resultados en términos de los indicadores de la matriz de confusión (verdaderos positivos, verdaderos negativos, falsos positivos, falsos negativos). Como se ve en esta, se detectan correctamente los segmentos asociados a los asentamientos San Petersburgo, Achaval y Villa La Candela.

Tab. 5.1: Resultados clasificación para la región La Matanza

Modelo	Precisión	Área bajo curva ROC	κ
Random Forests	85.00 %	79 %	0.70
XGBoost	90.00 %	83 %	0.71
SVM	83.00 %	83 %	0.73
GMM	86.00 %	79 %	0.66

Fig. 5.1: Indicadores matriz de confusión para clasificación en región La Matanza



Fuente: Elaboración Propia.

5.2. Región Los Tapiales

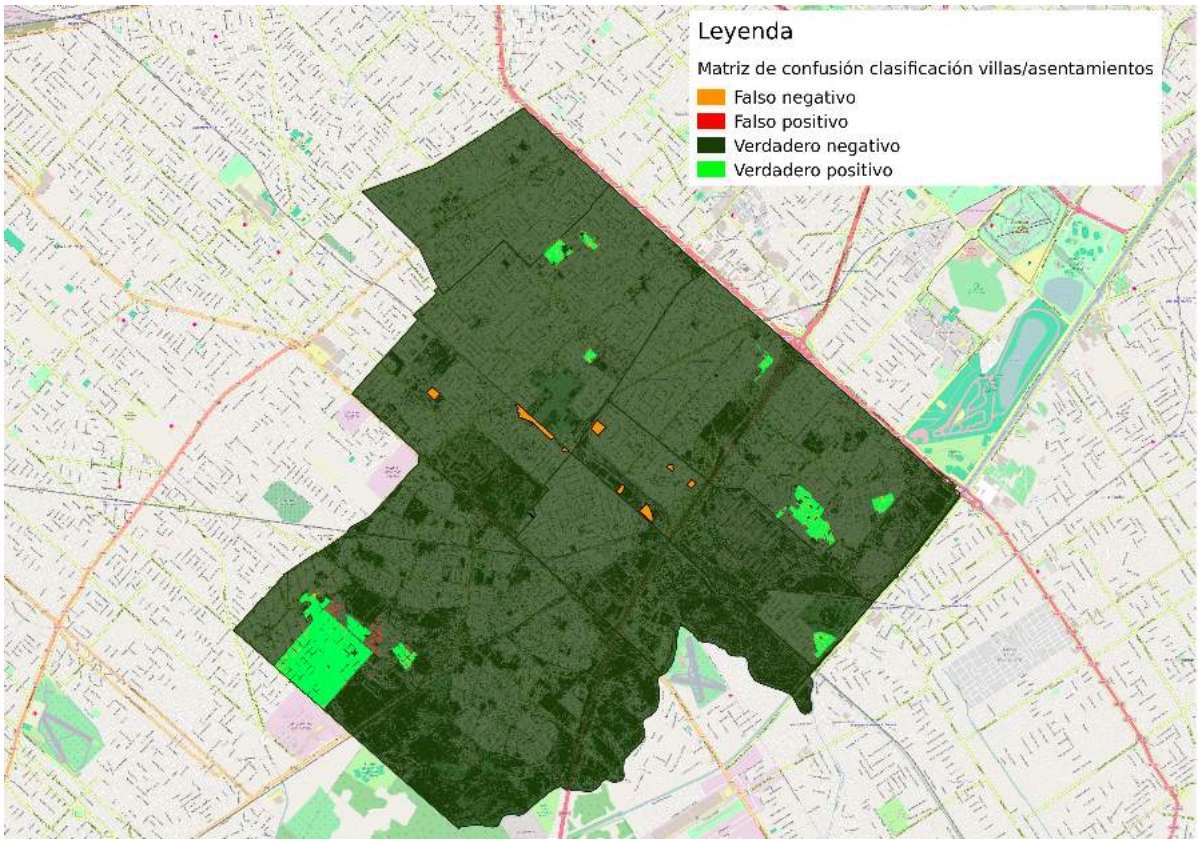
En la Tabla 5.2 se muestran los resultados de la clasificación en esta región. En la Figura 5.2 se muestra el mapa donde se representa la matriz de confusión.

Tab. 5.2: Resultados clasificación para la región Los Tapiales

Modelo	Precisión	AUC	κ
Random Forests	90 %	89 %	0.80
XGBoost	92 %	90 %	0.82
SVM	93 %	92 %	0.83
GMM	87 %	88 %	0.70

Se presenta a continuación, en la Figura 5.3, el caso del asentamiento 22 de Enero. Se clasificó correctamente gran parte de este asentamiento, incluyendo también los tres que están a su alrededor (Tierra y Libertad, El Gauchito Gil y 21 de marzo). En medio de estos se presentan falsos positivos, lo que puede explicarse por su ubicación. Un argumento de ese estilo puede utilizarse para los falsos negativos en los bordes del asentamiento 22 de Enero.

Fig. 5.2: Indicadores matriz de confusión para clasificación en región Los Tapiales



Fuente: Elaboración Propia.

Cabe destacar que se presentan falsos negativos para los asentamientos Ayacucho, Esteban de Luca, Villa Tuyutí, Los Vagones y Alberti.

5.3. Región Gregorio de Laferrere

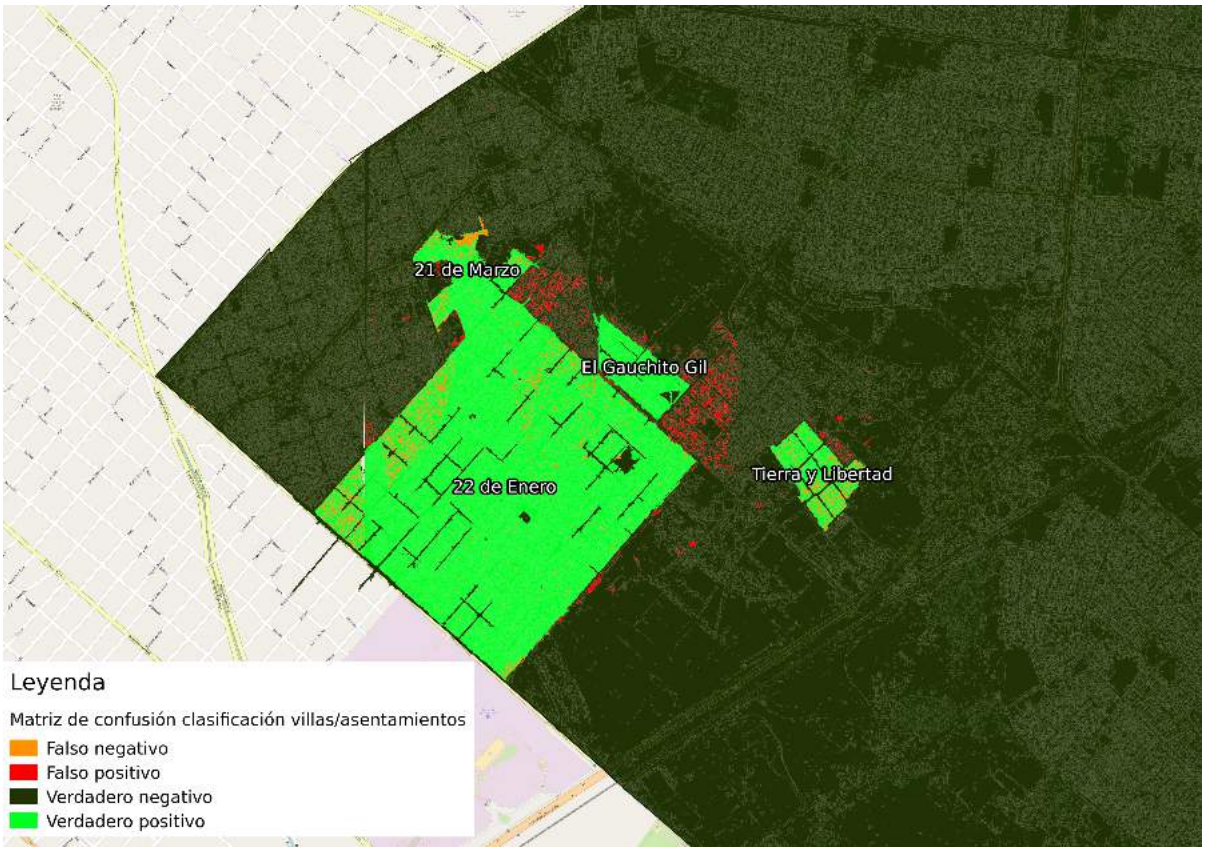
En la Tabla 5.3 se muestran los resultados de la clasificación en esta región.

Tab. 5.3: Resultados clasificación para la región Gregorio de Laferrere

Modelo	Precisión	AUC	κ
Random Forests	88 %	84 %	0.70
XGBoost	88 %	91 %	0.72
SVM	86 %	85 %	0.77
GMM	87 %	91 %	0.72

Se observa en la Figura 5.4 que se detectan correctamente los asentamientos, pero hay una elevada tasa de falsos positvos entre La Palangana, Barrio Luján, Primero de Mayo, José Luis Cabezas y Madre Teresa de Calcuta (en la Figura 5.5 se hace un acercamiento). Analizando las imágenes sin procesar, se identifica allí que los segmentos clasificados

Fig. 5.3: Indicadores matriz de confusión para clasificación en región Los Tapiales para los asentamientos 22 de Enero, 21 de Marzo, El Gauchito Gil y Tierra y Libertad



Fuente: Elaboración Propia.

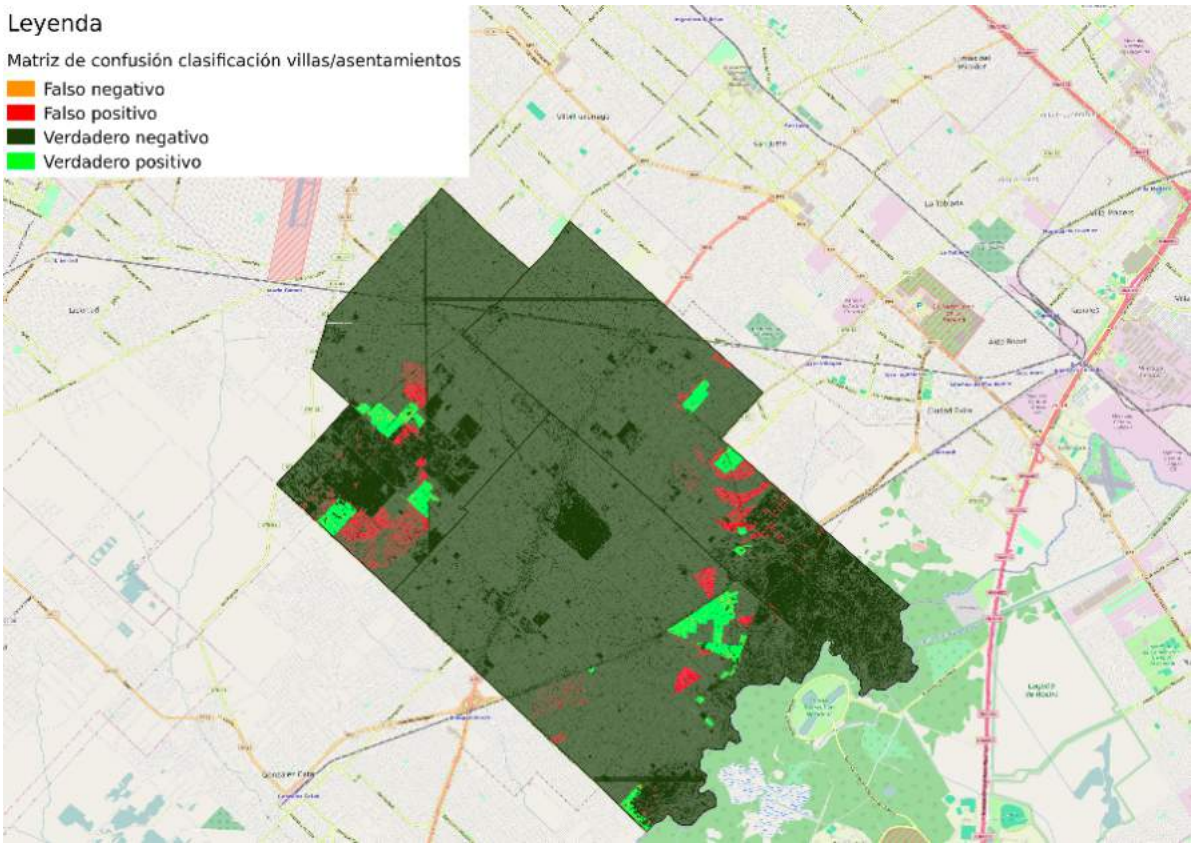
erróneamente están asociados a la cercanía al arroyo Don Mario y la calle Antartida Argentina (corren en paralelo). Cabe destacar que los atributos asociados a cursos de agua se destacaban entre los más importantes cuando se experimentó con la imagen de muestra.

Al igual que para el caso anterior, se observan falsos positivos en torno a las villas/asentamientos, lo que puede interpretarse como consecuencia del peso de las variables relacionadas a los ejes.

5.4. Región Juan Manuel de Rosas

En la Tabla 5.4 se muestran los resultados de la clasificación en esta región.

Fig. 5.4: Indicadores matriz de confusión para clasificación en región Gregorio de Laferrere



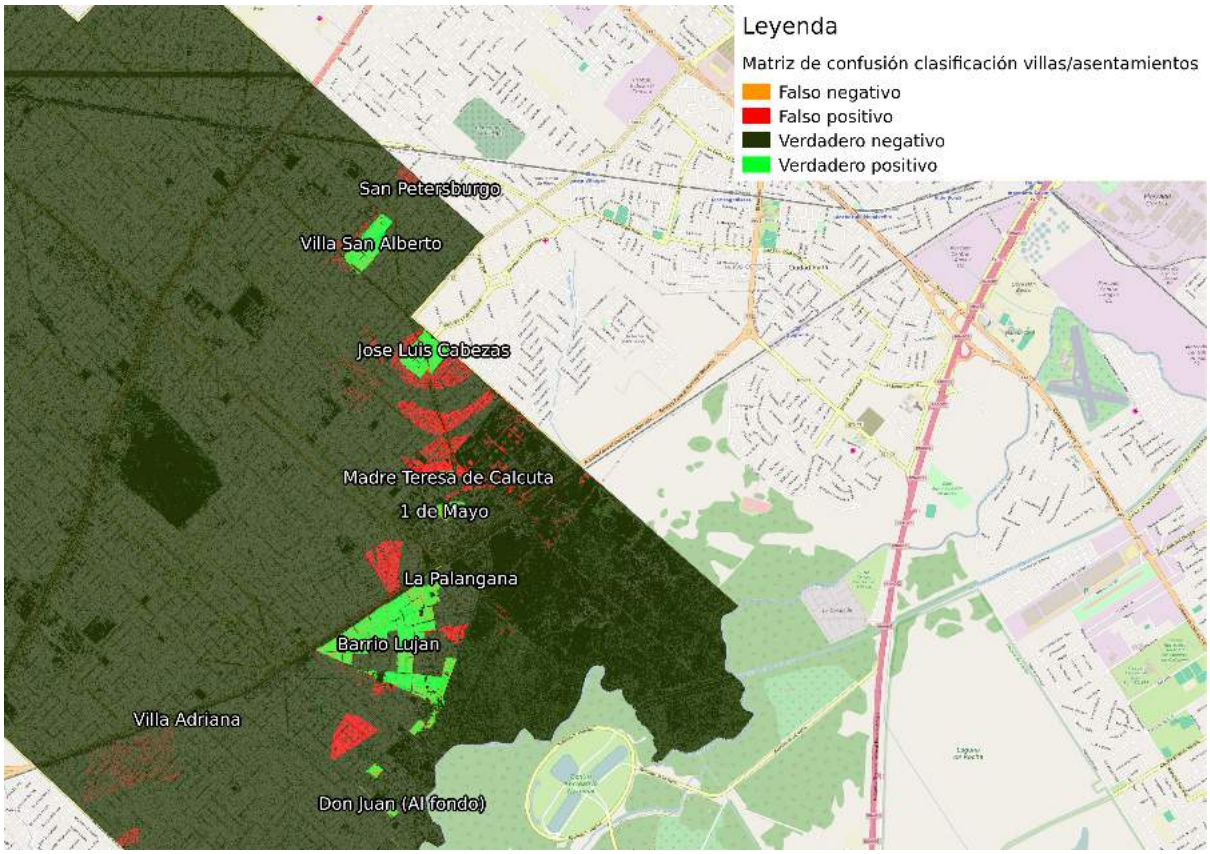
Fuente: Elaboración Propia.

Tab. 5.4: Resultados clasificación para la región Juan Manuel de Rosas

Modelo	Precisión	AUC	κ
Random Forests	91 %	93 %	0.76
XGBoost	91 %	94 %	0.81
SVM	90 %	93 %	0.75
GMM	87 %	93 %	0.73

Esta región del partido es la que más sectores rurales presenta, en un claro contraste con Los Tapiales y La Matanza. Se observa en el conjunto de datos de villas que la mayoría de estas se encuentran a los costados de la ruta nacional 3, lo que coincide con la ubicación de los centros de mayor densidad poblacional de esta región. En la Figura 5.6 se ven los falsos positivos en la trancisión entre cada una, lo que da la pauta de la influencia de la variable distancia a dicha ruta (distancia a eje entre 25 y 40 metros de ancho). Se presenta en la Figura 5.7 un mapa con mayor foco en esa zona.

Fig. 5.5: Indicadores matriz de confusión para clasificación en asentamientos seleccionados en región Gregorio de Laferrere



Fuente: Elaboración Propia.

5.5. Potenciales villas y asentamientos para relevar

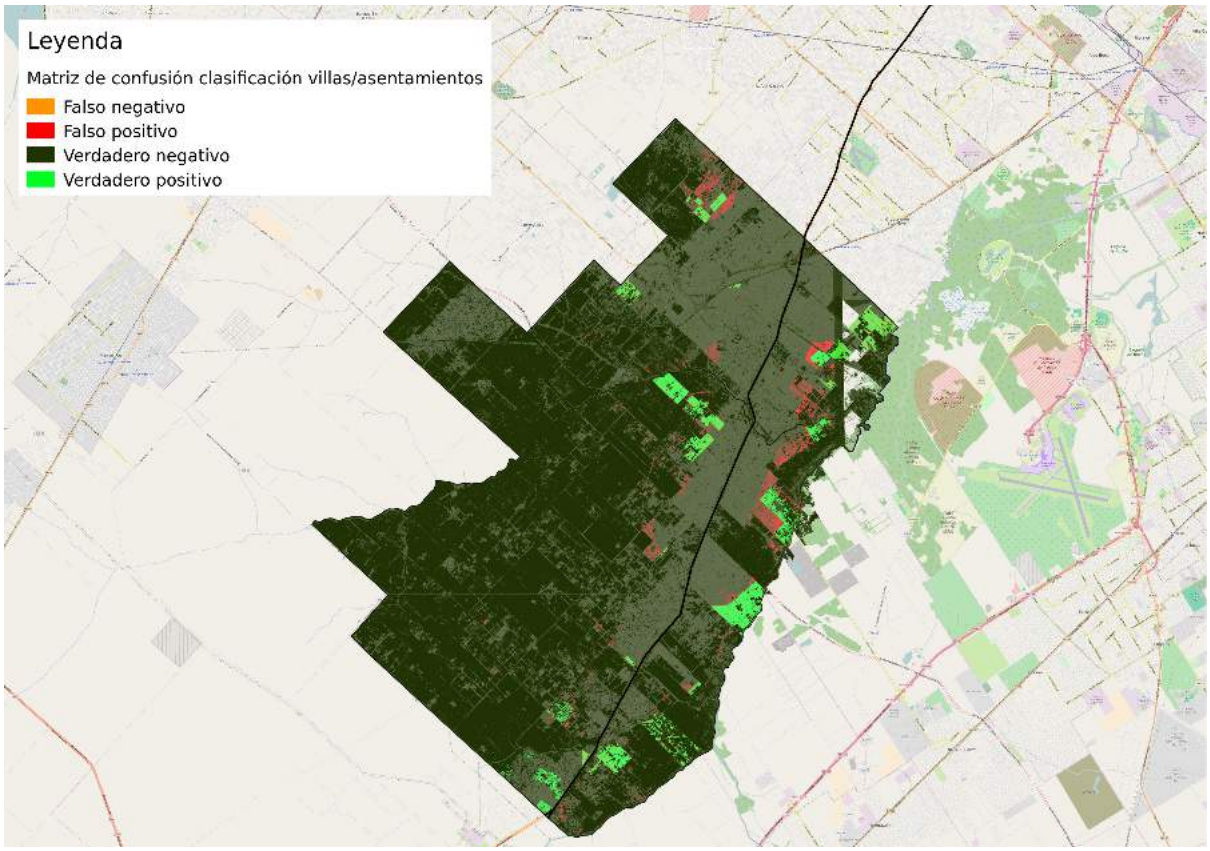
A partir de los resultados para cada región se cuantificó el área del municipio a relevar para detectar zonas potencialmente críticas. Para llevar a cabo esta tarea se utilizó el algoritmo XGBoost, cuya implementación informa la probabilidad de pertenecer a la clase positiva para cada instancia (en este caso segmentos). Como se mencionó durante el trabajo, el área del partido de La Matanza es de $326km^2$.

En las secciones anteriores el umbral de probabilidad para decidir la clase de un segmento fue $0,5$. A partir de la clasificación se puede construir la matriz de confusión que, entre otras métricas, presenta la cantidad de falsos postivos y de falsos negativos. En este caso, los primeros son aquellos segmentos clasificados erróneamente como villa/asentamiento, mientras que los falsos negativos son los que corresponden a esas zonas pero son categorizados de manera negativa por el modelo.

Para reducir el territorio a estudiar se requiere tener una baja cantidad de falsos negativos para que el relevamiento no pierda eficacia, controlando el aumento en la cantidad de falsos positivos que esa baja acarrea. Para analizar esto, se experimentó modificando el umbral de probabilidad a partir de los resultados de XGBoost.

En la Figura 5.8 se ve que la tasa de falsos positivos comienza en un 45 % para comen-

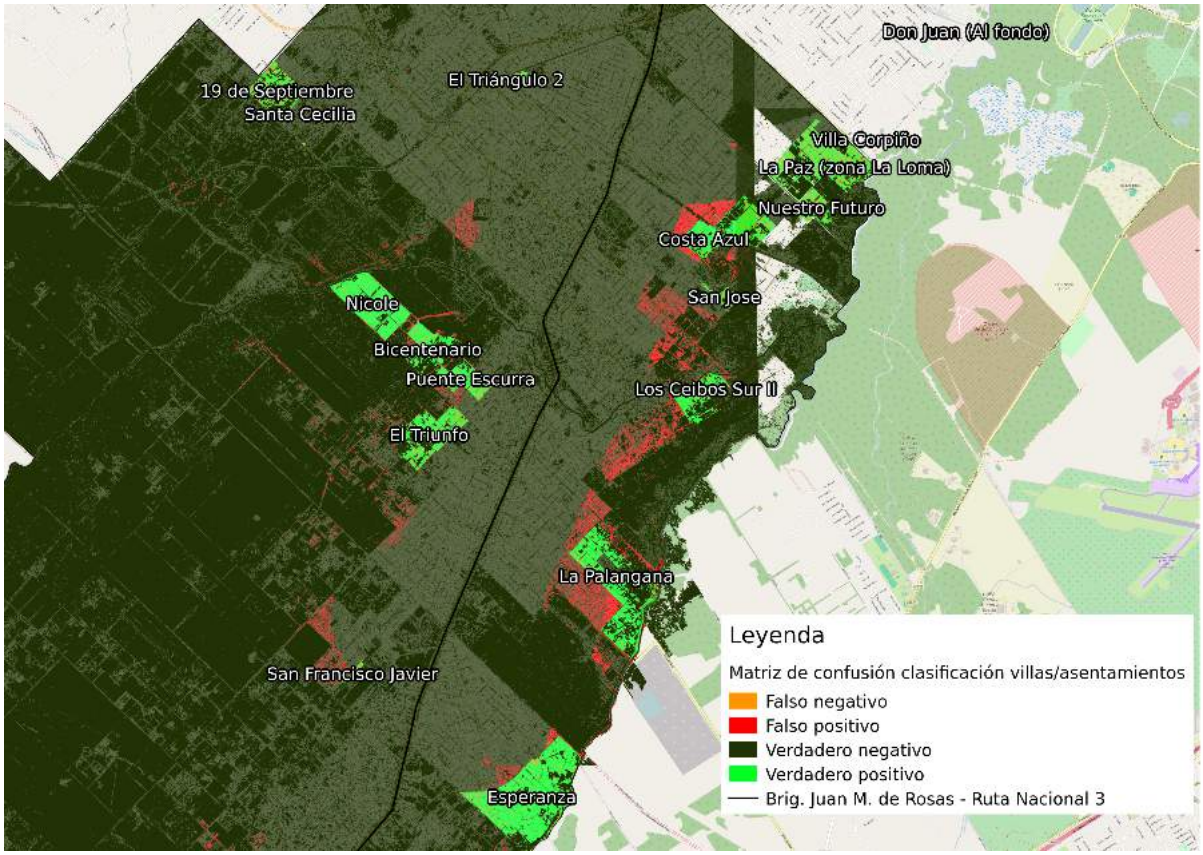
Fig. 5.6: Indicadores matriz de confusión para clasificación en región Juan Manuel de Rosas



Fuente: Elaboración Propia.

zar a descender a aproximadamente 15 % en el umbral $0,20$. Más allá de la elección de un único valor, seleccionar el umbral en el rango entre $0,4$ y $0,6$ cumple con los requisitos planteados anteriormente. En la Tabla 5.5 se presentan los resultados para diferentes valores de umbrales, donde el recorrido se calcula sumando los falsos positivos y los verdaderos positivos. En la Figura 5.9 se muestra el mapa del municipio con el territorio a relevar resaltado, utilizando un umbral de $0,40$. Cabe destacar que considerando esas zonas a relevar, quedan afuera las villas Barrio 48 y La Paz.

Fig. 5.7: Indicadores matriz de confusión para clasificación en región Juan Manuel de Rosas



Fuente: Elaboración Propia.

Fig. 5.8: Tasa de falsos negativos y falsos positivos para la clasificación utilizando XGBoost.

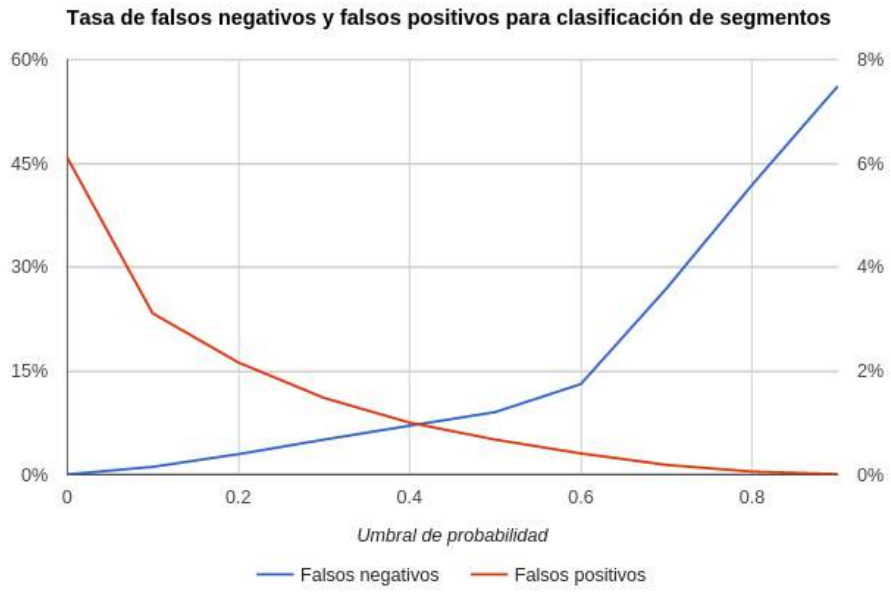
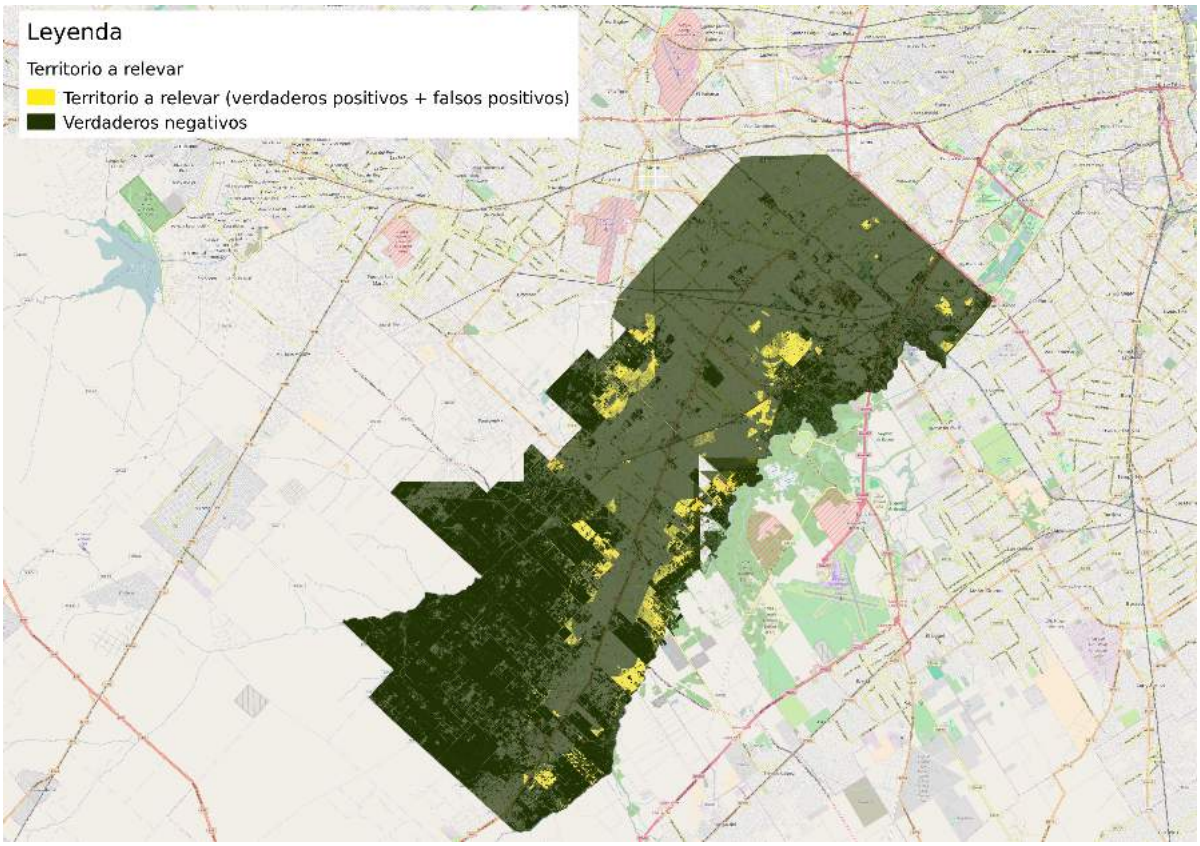


Fig. 5.9: Territorio a relevar utilizando el algoritmo XGBoost con umbral de probabildad 0,40.



Tab. 5.5: Tasa de falsos negativos, falsos positivos, porcentaje y kilómetros cuadrados del territorio a recorrer.

Umbral	**Falsos negativos**	**Falsos positivos**	**Recorrido**	**Recorrido (km2)**
0	0 %	46 %	55 %	179
0,1	0 %	23 %	32 %	105
0,2	0 %	16 %	25 %	81
0,3	1 %	11 %	19 %	63
0,4	1 %	8 %	16 %	51
0,5	1 %	5 %	13 %	42
0,6	2 %	3 %	10 %	34
0,7	4 %	1 %	7 %	22
0,8	6 %	0 %	4 %	12
0,9	7 %	0 %	2 %	5

6. RESULTADOS CONSIDERANDO SÓLO IMÁGENES SATELITALES

En este apartado se resolverá el mismo problema pero sin utilizar los datos censales ni los de ejes. La idea de considerar solamente las imágenes satelitales está basada en utilizar solo datos actualizados al momento del análisis.

Utilizando el mismo territorio de muestra que para el caso anterior, se calibran los algoritmos para luego aplicarlos a las regiones ficticias en las que se dividió el territorio.

El algoritmo SVM no tuvo que ser recalibrado, mientras que para el caso de Random Forests y XGBoost se utilizó una mayor profundidad de árboles para poder captar una complejidad mayor (para el primero se consideró una máxima profundidad de 20, mientras que par el segundo se aumentó a 7). Para el caso de GMM se utlizaron 120 componentes en lugar de 80.

6.1. Elección de modelos utilizando un territorio de muestra considerando solo imágenes

Utilizando el mismo procedimiento que para el capítulo anterior, se presenta en la Tabla 6.1 el resultado de la calibración de modelos. Se ve allí que el desempeño medido por el coeficiente κ desciende en aproximadamente veinte puntos porcentuales respecto de considerar todos los conjuntos de datos. Esta baja era esperable de acuerdo a lo que se fue mencionando acerca de la importancia de la distancia a cursos de agua y a otros tipos de eje.

Tab. 6.1: Medidas de precisión, área bajo la curva ROC y coeficiente κ según el algoritmo

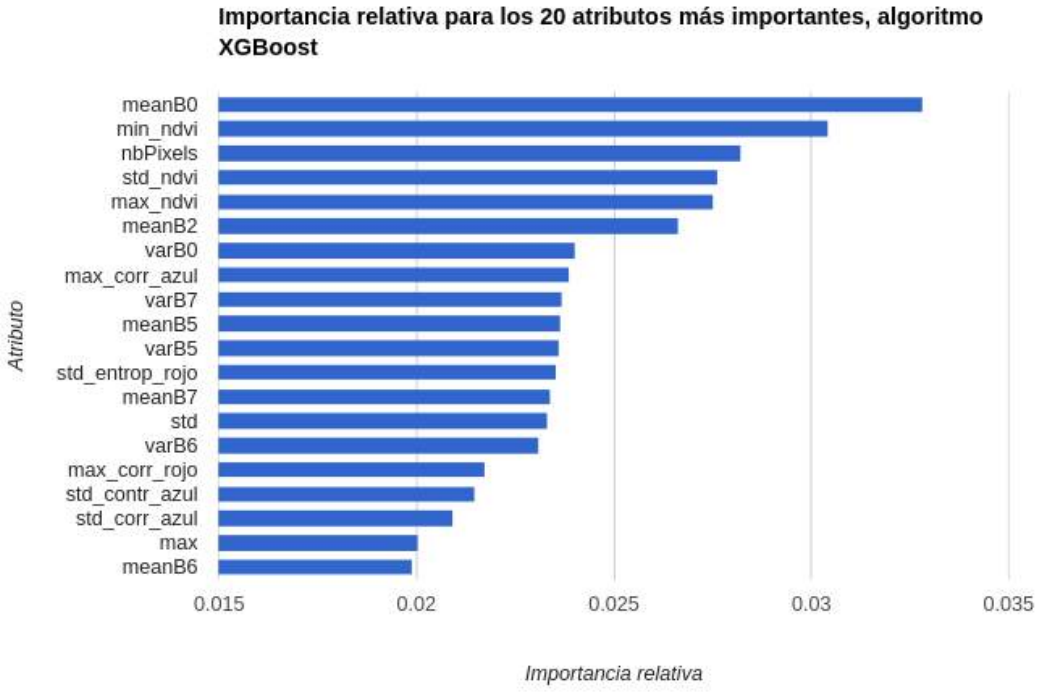
Modelo	Precisión	AUC	κ
Random Forests	86 %	68 %	0.58
XGBoost	84 %	70 %	0.60
SVM	84 %	69 %	0.61
GMM	82 %	70 %	0.60

Analizando la importancia de variables, para el algoritmo Random Forests se observa que los atributos relacionados con la banda azul aparecen en los primeros lugares. Esto también sucedía en el capítulo anterior pero precediendo a las relacionadas a ejes. También se encuentran las relacionadas al indicador NDVI . Un aspecto a destacar es que no aparecen numerosamente las variables asociadas a las bandas de las imágenes, lo que confirma que los atributos seleccionados a partir de la literatura resultaron importantes.

Por otro lado, siguiendo a Manghara y Odindi (2013), la importancia de los atributos asociados a la correlación en la banda roja viene dada por la capacidad de este indicador de separar los caminos de la vegetación. Dado que no se cuenta con información de ejes en este apartado, tiene sentido que estas variables cobren relevancia en estos experimentos. Cabe destacar que, como se mencionó anteriormente, los atributos asociados a textura están relacionados a la densidad en la cobertura de techos, vegetación y espacios abiertos.

Para el caso del contraste en la banda azul, este tiene importancia en la separación de los caminos de la vegetación escasa y a su vez de esta última respecto de las edificaciones. Para el caso del NDVI, el territorio de muestra presenta una creciente densidad de vegetación a medidad que se acerca al río Matanza, por lo que esta variable cobra relevancia aquí. Este argumento también resulta válido para el atributo asociado a la cantidad de pixeles del segmento, puesto que los segmentos más grandes están relacionados aquí con zonas de alta densidad de vegetación.

Fig. 6.1: Importancia relativa de los primeros 20 atributos más importantes a partir del algoritmo XGBoost

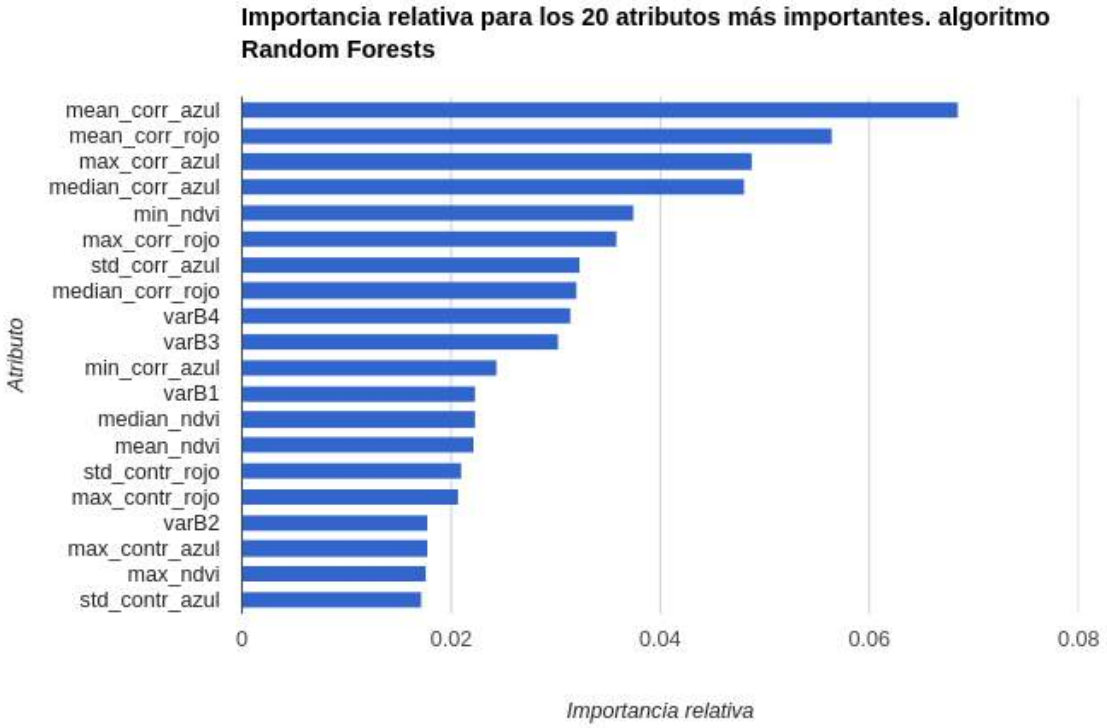


Fuente: Elaboración Propia.

6.2. Región La Matanza

Comparando la clasificación resultante con la anterior sección, en la Figura 6.3 se ven un conjunto de manzanas clasificadas erróneamente como villa/asentamiento cerca de la villa San Petersburgo. Estas manzanas corresponden a terrenos (baldíos en gran parte) atravesados por vías del ferrocarril Belgrano Sur, lo que podría causar este error de clasificación.

Fig. 6.2: Importancia relativa de los primeros 20 atributos más importantes a partir del algoritmo Random Forests



Fuente: Elaboración Propia.

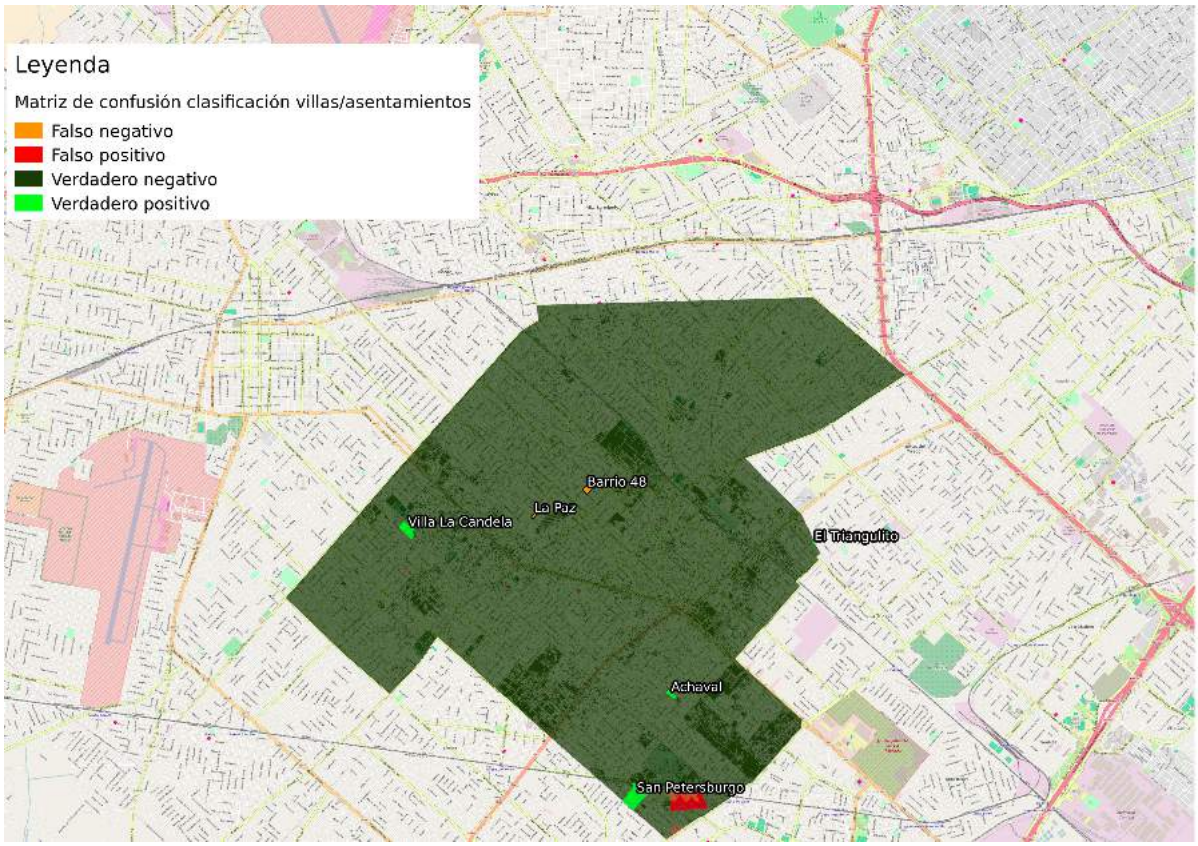
Tab. 6.2: Medidas de precisión, área bajo la curva ROC y coeficiente κ según el algoritmo

Modelo	Precisión	AUC	κ
Random Forests	80 %	80 %	0.69
XGBoost	78 %	78 %	0.67
SVM	81 %	76 %	0.66
GMM	77 %	74 %	0.63

6.3. Region Los Tapiales

Para el caso de la región de Los Tapiales, los algoritmos Random Forests, XGBoosts y Máquinas de Vectores de Soporte presentan el mejor rendimiento en término del coeficiente κ. El descenso en dicha métrica fue de 0.15 en general. En la Tabla 6.3 se presentan los resultados de la experimentación.

Fig. 6.3: Indicadores matriz de confusión para clasificación en región La Matanza utilizando atributos asociados a imágenes



Fuente: Elaboración Propia.

Tab. 6.3: Resultados clasificación para la región Los Tapiales utilizando atributos asociados a imágenes

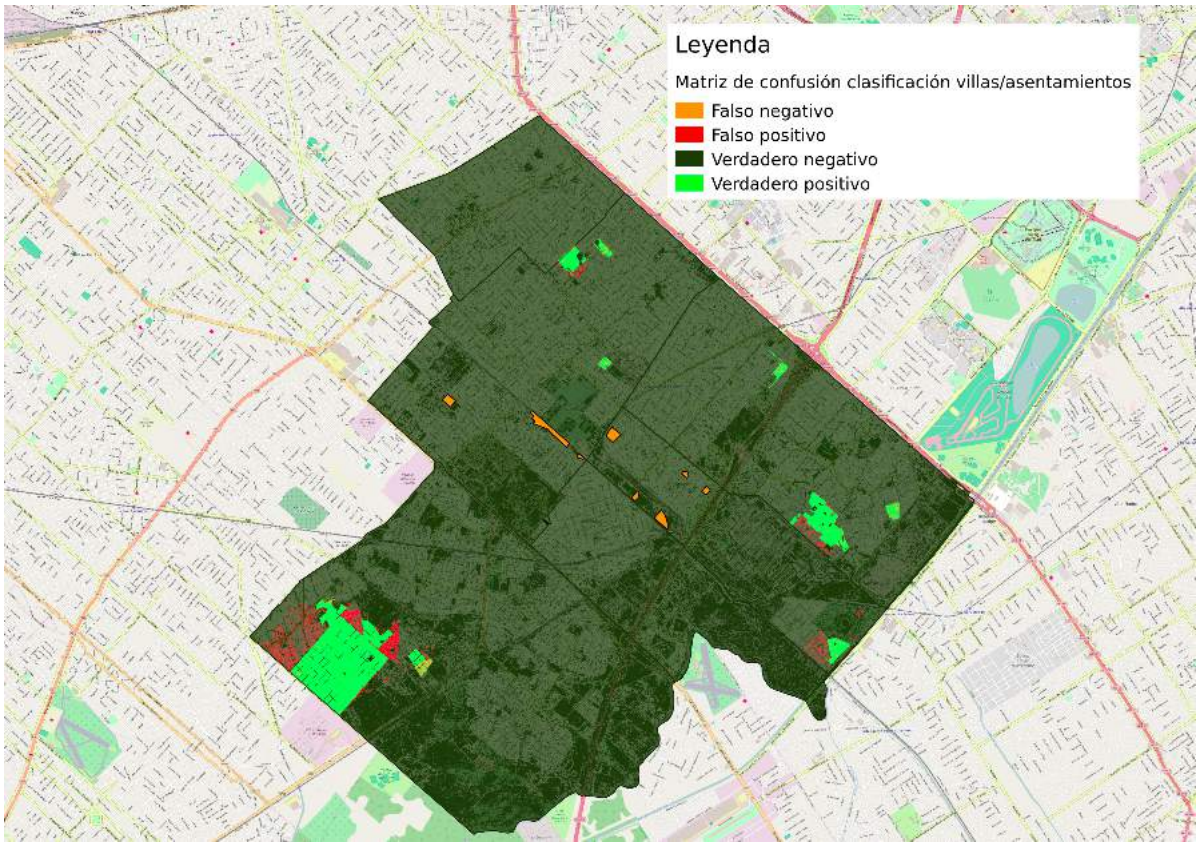
Modelo	Precisión	AUC	κ
Random Forests	87 %	87 %	0.67
XGBoost	85 %	87 %	0.61
SVM	83 %	87 %	0.68
GMM	83 %	85 %	0.63

Comparando con los resultados del capítulo anterior, en los bordes del asentamiento 22 de Enero se observa una mayor presencia de falsos positivos. Lo mismo sucede para los asentamientos 21 de Marzo, El Gauchito Gil y Tierra y Libertad.

6.4. Región Gregorio de Laferrere

Para el caso de esta región se incrementó la tasa de falsos positivos en la zona analizada en el capítulo anterior (entre La Palangana, Barrio Luján, Primero de Mayo, José Luis Cabezas y Madre Teresa de Calcuta). Si bien no se cuenta con la variable asociada a la

Fig. 6.4: Indicadores matriz de confusión para clasificación en región Los Tapiales utilizando atributos asociados a imágenes



Fuente: Elaboración Propia.

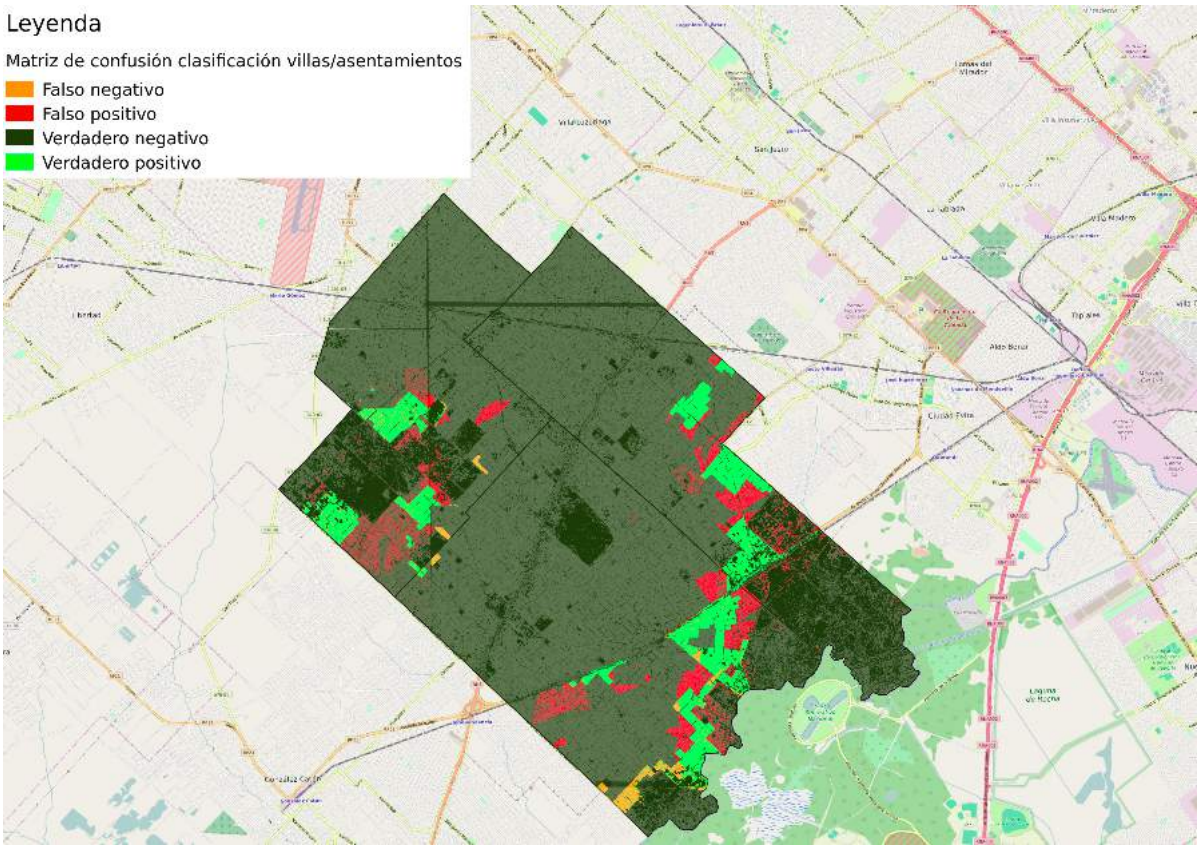
calle Antartida Argentina y al arroyo, toman mayor importancia las características que capturan la presencia de vegetación y la organización y brillo de los techos.

A medida que más al sur nos acercamos, más errores se cometen. Esto se debe a que comienza la transición entre los sectores netamente urbanos con los sectores rurales y terrenos cercanos al río Matanza.

Tab. 6.4: Resultados clasificación para la región Gregorio de Laferrere utilizando atributos asociados a imágenes

Modelo	Precisión	AUC	κ
Random Forests	79 %	84 %	0.61
XGBoost	86 %	85 %	0.64
SVM	82 %	84 %	0.60
GMM	80 %	79 %	0.55

Fig. 6.5: Indicadores matriz de confusión para clasificación en región Gregorio de Laferrere utilizando atributos asociados a imágenes



Fuente: Elaboración Propia.

6.5. Región Juan Manuel de Rosas

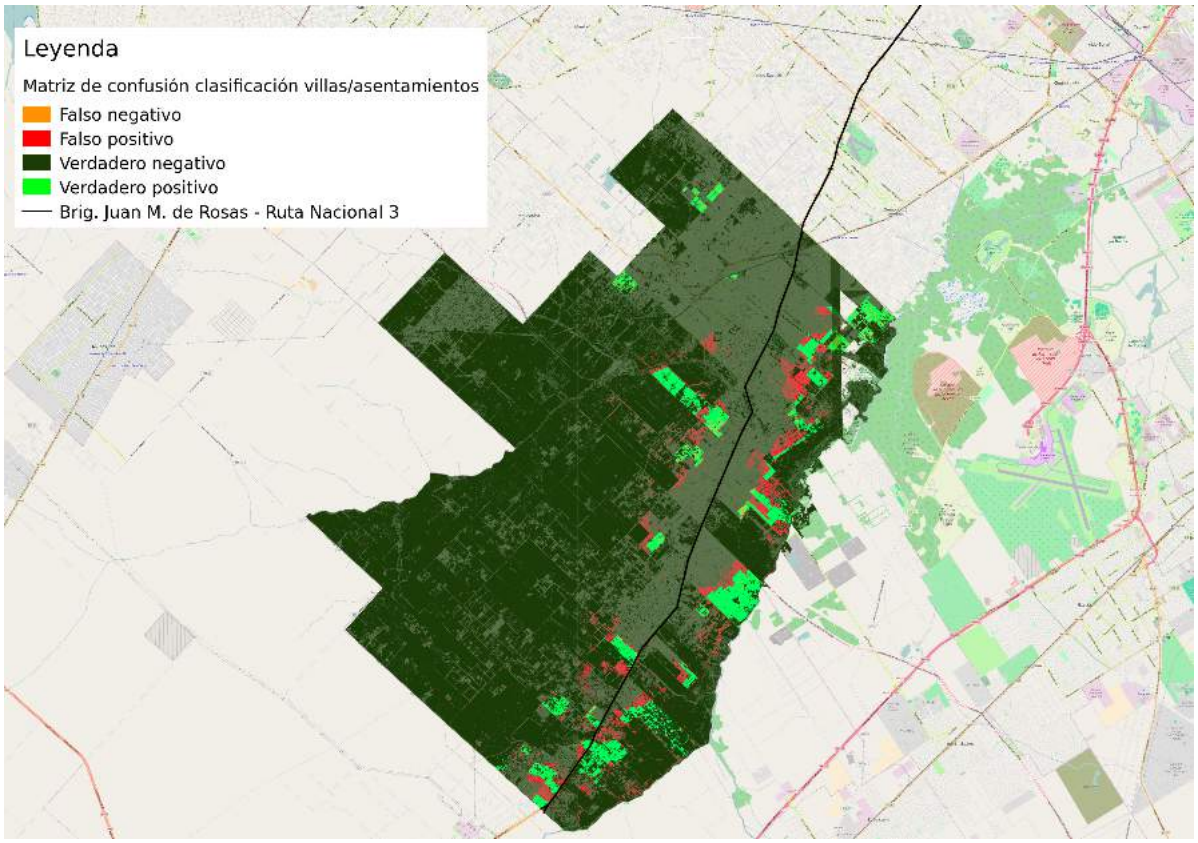
El descenso en el κ para este caso es similar al que se da en Los Tapiales. Se ve un peor desempeño en los asentamientos que se encuentran en el límite sur de esta región, mostrando una mayor cantidad de falsos positivos. El hecho de no contar con la variable asociada a la distancia a calles de 25 metros a 40 metros de ancho es una de las causas de la caída del κ.

Tab. 6.5: Resultados clasificación para la región Juan Manuel de Rosas utilizando atributos asociados a imágenes

Modelo	Precisión	AUC	κ
Random Forests	84 %	85 %	0.62
XGBoost	85 %	87 %	0.57
SVM	83 %	84 %	0.63
GMM	85 %	82 %	0.59

En la Figura 6.7 se muestran las probabilidades de detectar villas/asentamientos para

Fig. 6.6: Indicadores matriz de confusión para clasificación en región Juan Manuel de Rosas utilizando atributos asociados a imágenes



Fuente: Elaboración Propia.

esta subdivisión. Al no contar con el atributo asociado a la distancia a la Ruta Nacional 3, la probabilidad de clasificar un segmento positivamente se mantiene alta a lo largo del sector urbanizado de esta región. Los valores comienzan a descender a medida que se acerca la zona rural.

6.6. Potenciales villas y asentamientos para relevar

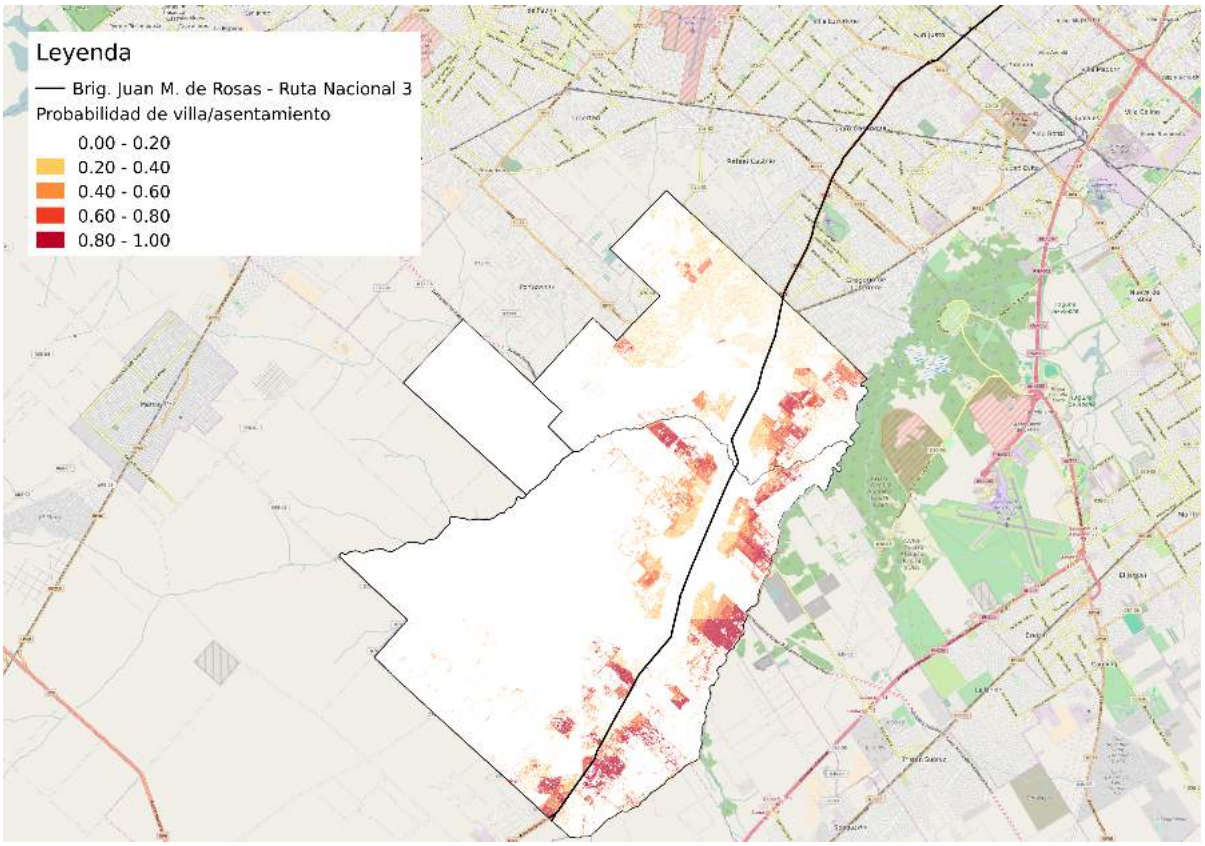
En esta sección se presenta el mismo análisis que en el capítulo anterior pero considerando solamente el usos de imágenes. En la Tabla 6.6 se presentan los resultados.

En este caso, nuevamente un umbral entre $0,40$ y $0,60$ cumpliría con los requisitos de falsos negativos y positivos. En la Figura 6.10 se presenta el mapa de territorios a relevar utilizando el valor $0,40$. El área a recorrer es mayor que en el capítulo anterior, pasando de $51km^2$ a $96km^2$ para ese umbral.

Considerando esas zonas a relevar, quedan afuera las villas Barrio 48 y La Paz. Esto coincide con los resultados del capítulo anterior.

En la Figura 6.9 se compara el territorio a estudiar según las fuentes de información consideradas. Se ve allí que el mejor desempeño se obtiene utilizando todos los conjuntos georreferenciados y las imágenes, seguido por utilizar solo los georreferenciados y en último

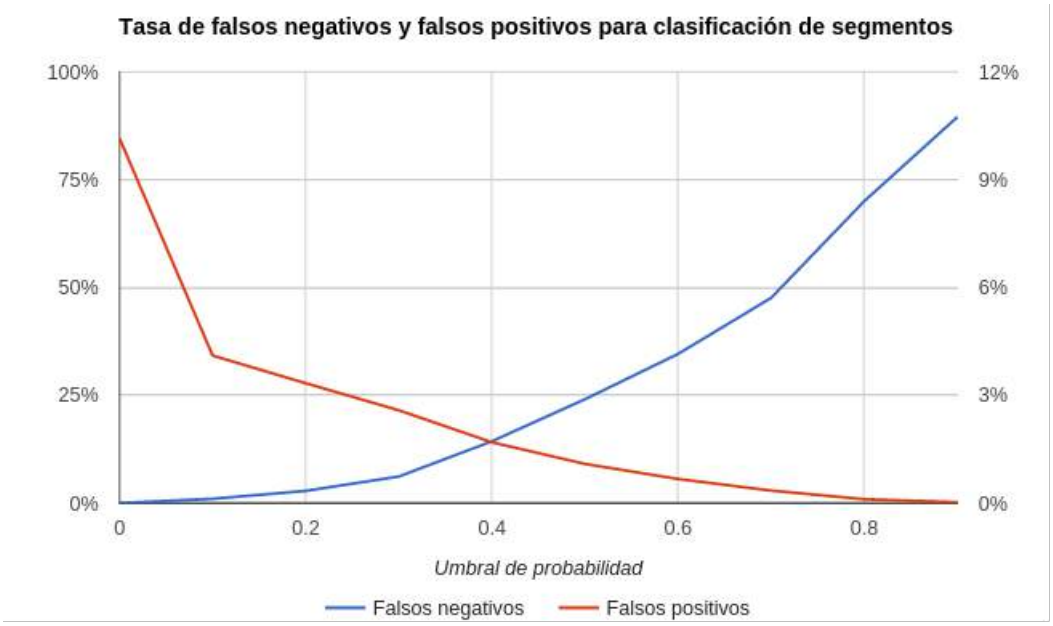
Fig. 6.7: Probabilidad de clase villa/asentamiento en región Juan Manuel de Rosas utilizando atributos asociados a imágenes



Fuente: Elaboración Propia.

lugar solo las imágenes. A partir de este análisis se verifica la mayor contribución de las variables basadas en los datos censales, viales y naturales respecto de las asociadas a las imágenes. De este modo, se puede concluir que las imágenes aportaron más a través de los segmentos formados que de los atributos calculados.

Fig. 6.8: Tasa de falsos negativos y falsos positivos para la clasificación utilizando XGBoost.



Tab. 6.6: Tasa de falsos negativos, falsos positivos, porcentaje y kilómetros cuadrados del territorio a recorrer.

Umbral	Falsos negativos	Falsos positivos	Recorrido	Recorrido (km2)
0	0 %	85 %	100 %	326
0,1	0 %	34 %	50 %	162
0,2	0 %	28 %	43 %	141
0,3	1 %	22 %	37 %	121
0,4	2 %	14 %	30 %	96
0,5	3 %	9 %	25 %	80
0,6	4 %	6 %	21 %	69
0,7	6 %	3 %	18 %	60
0,8	8 %	1 %	16 %	53
0,9	11 %	0 %	16 %	51

Fig. 6.9: Porcentaje de territorio de La Matanza a relevar según umbral de probabilidad y conjuntos de datos utilizado.

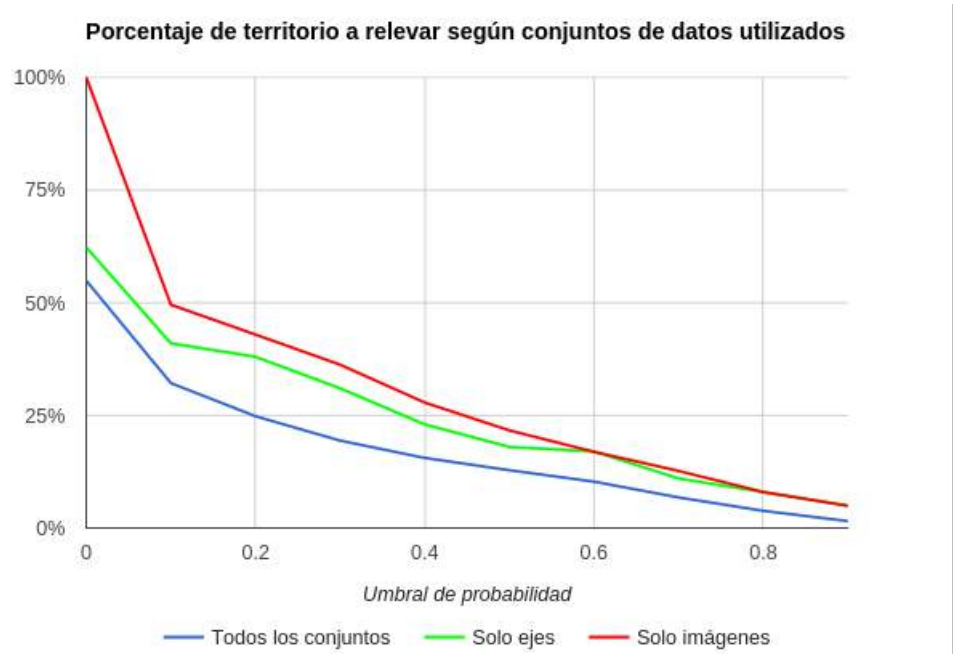
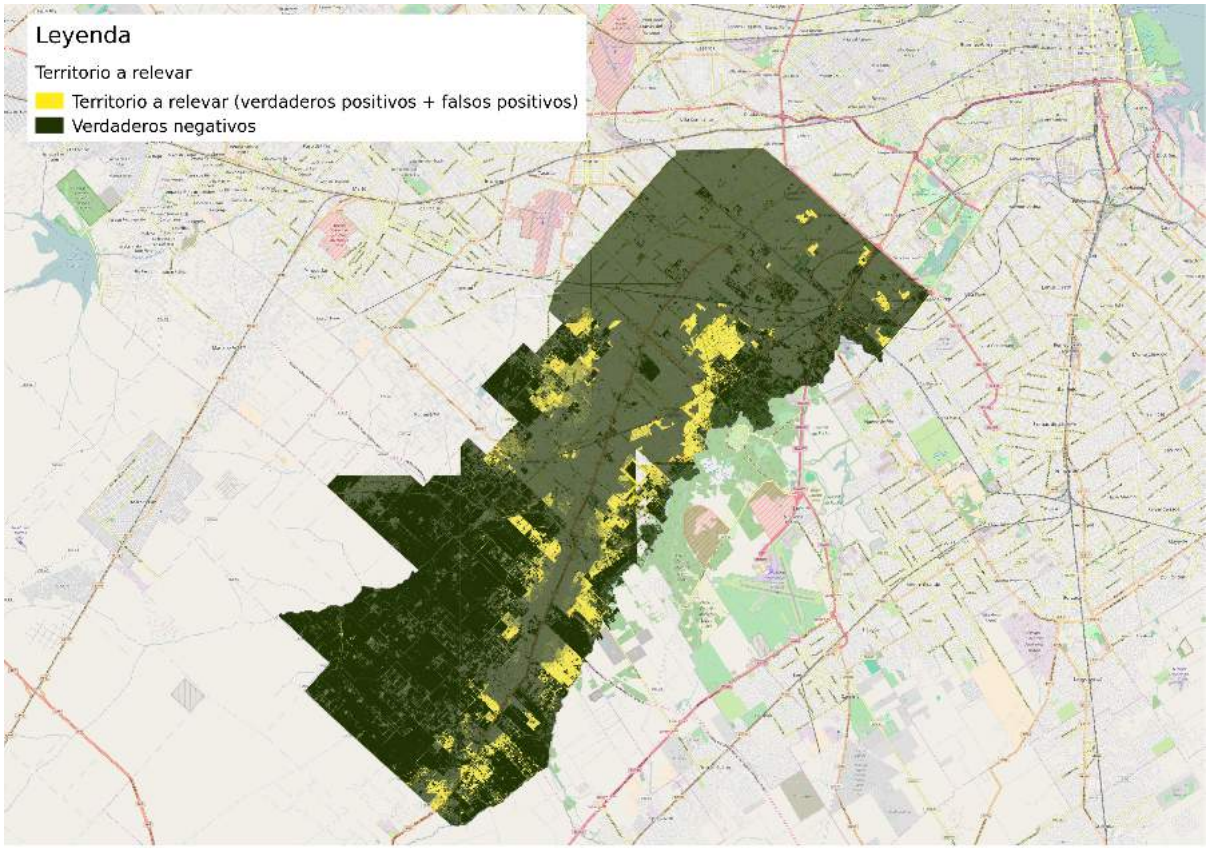


Fig. 6.10: Territorio a relevar utilizando el algoritmo XGBoost con umbral de probabildad 0,40.



7. DISCUSIÓN Y CONCLUSIONES

El objetivo de este trabajo es elaborar una metodología que identifique potenciales villas y asentamientos permitiendo reducir las zonas donde realizar trabajo de campo para relevarlas. El lugar elegido es el partido de La Matanza para el año 2013. A partir de este objetivo surge la idea de analizar el impacto de considerar conjuntos de datos con alta y con baja periodicidad de actualización. Para el caso de alta periodicidad se utilizaron imágenes satelitales y para el de baja, datos del censo de 2010, relevamiento de ejes y usos del suelo.

La metodología consta de una etapa de obtención y preprocesamiento de datos e imágenes, segmentación de la imágenes, cálculo de atributos y aplicación de un algoritmo de clasificación de segmentos previamente entrenado.

Considerando todos los conjuntos de datos, la metodología reduce el territorio a relevar a $51km^2$ (16 % del área total). La región de Los Tapiales es la que mejor desempeño tuvo, mientras que la de Gregorio de Laferrere fue la que más problemas presentó. Los cuatro algoritmos de clasificación utilizados mostraron resultados similares en términos de κ, siendo el más destacado XGBoost. La paridad de resultados muestra la relevancia de la elección de atributos por sobre los algoritmos. Los atributos más importantes en la clasificación cambiaban según el algoritmo que se considere. Los estadísticos calculados sobre el NDVI, cercanía a distinto tipo de calles, vías de tren y cursos de agua fueron más importantes. Este resultado valida las variables mencionadas en la literatura como relevantes para este tipo de problema. Además, se verificó en este punto las diferencia en el cálculo de la importancia de variables para Random Forests y XGBoost.

Considerando solamente imágenes satelitales, la metodología reduce a $96km^2$ (30 % del área total) el territorio a estudiar. Nuevamente la región de Los Tapiales fue la mejor clasificada, siendo el NDVI y los relacionados con la textura los atributos más importantes. En este caso se evidencia la necesidad de sumar más variables al análisis para poder mejorar el desempeño, lo que puede lograrse a partir de indagar más en la literatura de clasificación de usos del suelo utilizando imágenes satelitales.

Tanto considerando todas las fuentes de datos como utilizando solamente imágenes, queda fuera el 2 % del territorio de villas y asentamientos relevado por Techo. Parte de este corresponde a las villas Barrio 48 y La Paz. Esto es un punto débil de la metodología.

A continuación se plantean caminos a seguir con el objetivo de poder mejorar el rendimiento de la metodología, utilizando solamente datos con alta periodicidad de actualización.

El descenso en la representatividad de los datos a partir de su progresiva desactualización afecta principalmente a los datos censales, puesto que los relacionados a ejes pueden obtenerse de fuentes alternativas como el caso de Open Street Map. Esto abre la puerta a nuevas experimentaciones que permitan contar con datos actualizados sin perder la calidad y alcance de un relevamiento oficial.

Respecto de los algoritmos considerados, se podría experimentar con otro tipo de modelos como el caso de regresión logística o incluso probar con ensambles de los modelos que se estimaron en esta tesis. Por otro lado, se les está dando importancia a los algoritmos relacionados con redes neuronales profundas (como el caso de redes neuronales convulcionales), estos presentan un alto rendimiento en conjuntos de datos de imágenes. Como se

mencionó anteriormente, mas allá de los algoritmos, la calidad de los atributos es lo que verdaderamente mejora la clasificación.

Por último, un comentario relacionado con el costo de realizar el relevamiento. Si bien la metodología en ambos casos permite reducir el territorio a analizar (y por ende los costos logísticos asociados), las imágenes utilizadas son de alta resolución (menor a 1 metro) y tienen un costo monetario de adquisición.

Una alternativa que podría evaluarse es usar imágenes de dominio público, como el caso de las Landsat, provistas libremente por la NASA con una periodicidad alta. Si bien estas tienen una resolución del orden de 30 metros, la combinación con datos como los de Open Street Map podría tener un desempeño razonable con la característica de reducir el costo de adquisición de imágenes. La contracara sería un incremento en la tasa de falsos positivos que aumentaría el territorio a relevar. El uso de imágenes libres sumado a la disponibilidad de software libre (como el caso de Orfeo Toolbox y Python) permitiría reducir los costos de mantenimiento de una potencial herramienta para llevar a cabo políticas públicas.

Bibliografía

[1] T. Blaschke. Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(1):2–16, 2010.

[2] Thomas Blaschke, Geoffrey J. Hay, Maggi Kelly, Stefan Lang, Peter Hofmann, Elisabeth Addink, Raul Queiroz Feitosa, Freek van der Meer, Harald van der Werff, Frieke van Coillie, y Dirk Tiede. Geographic Object-Based Image Analysis - Towards a new paradigm. *ISPRS Journal of Photogrammetry and Remote Sensing*, 87:180–191, 2014.

[3] Thomas Blaschke y Josef Strobl. What's wrong with pixels? Some recent developments interfacing remote sensing and GIS. *GeoBIT/GIS*, 6(1):12–17, 2001.

[4] Leo Breiman. Bagging predictors. *Mach Learn*, 24(2):123–140, 1996.

[5] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[6] David M Carneggie y Donald T Lauer. Uses of multiband remote sensing in forest and range inventory. *Photogrammetria*, 21(4):115–141, 1966.

[7] Tianqi Chen. Introduction to boosted trees. 2014.

[8] Tianqi Chen y Carlos Guestrin. Xgboost: A scalable tree boosting system. *arXiv preprint arXiv:1603.02754*, 2016.

[9] Emilio Chuvieco. Fundamentos De Teledeteccion, 1995.

[10] J Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[11] D. Comaniciu y P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):1–37, 2002.

[12] María Cristina Cravino, Juan Pablo Del Río, y Juan Ignacio Duarte. Magnitud y crecimiento de las villas y asentamientos en el Área Metropolitana de Buenos Aires en los últimos 25 años. *Facultad de Arquitectura, Diseño y Urbanismo. Universidad de Buenos Aires*, páginas 1–17, 2001.

[13] Jane Falkingham y Ceema Namazie. Measuring health and poverty: a review of approaches to identifying the poor.

[14] Deon Filmer y Lant Pritchett. The effect of household wealth on educational attainment: evidence from 35 countries. *Population and development review*, 25(1):85–120, 1999.

[15] Yoav Freund y Robert E Schapire. Experiments with a New Boosting Algorithm, 1996.

[16] Leonardo Gasparini, Mariana Marchionni, y Walter Sosa Escudero. La distribucion del ingreso en la Argentina. 2001.

[17] Antonin Guttman. R-trees. In *Proceedings of the 1984 ACM international conference on Management of data*. Association for Computing Machinery ACM, 1984.

[18] Trevor Hastie, Robert Tibshirani, y Jerome Friedman. Unsupervised learning. In *The elements of statistical learning*, páginas 485–585. Springer, 2009.

[19] Norbert Henninger y Mathilde Snel. *Where are the poor? experiences with the development and Use of poverty Maps*. 2002.

[20] Jordi Inglada y Emmanuel Christophe. The orfeo toolbox remote sensing image processing software.

[21] Divyani Kohli, Richard Sliuzas, Norman Kerle, y Alfred Stein. An ontology of slums for image-based classification. *Computers, Environment and Urban Systems*, 36(2):154–163, 2012.

[22] Paidamwoyo Mhangara y John Odindi. Potential of texture-based classification in urban landscapes using multispectral aerial photos. *South African Journal of Science*, 109(3-4):1–8, 2013.

[23] Julien Michel, David Youssefi, y Manuel Grizonnet. Stable Mean-Shift Algorithm and Its Application to the Segmentation of Arbitrarily Large Remote Sensing Images. {*IEEE*} *Trans. Geosci. Remote Sensing*, 53(2):952–964, feb 2015.

[24] Thomas M Mitchell. Machine learning. *New York*, 1997.

[25] Maik Netzband, Ellen Banzhaf, René Höfer, y Katrin Hannemann. Identifying the poor in cities: how can remote sensing help to profile slums in fast growing cities and megacities. *IHDP Update*, 1:22–28, 2009.

[26] Miguel Ochoa, Alberto Dayer, Jared Levin, David Aaron, Dennis L Helder, Larry Leigh, Jeff Czapla-meyers, Nik Anderson, Brett Bader, Fabio Pacifici, William Baugh, Milan Karspeck, Nathan Longbotham, y Gregory Miecznik. Absolute Radiometric Calibration of the DigitalGlobe Fleet and updates on the new WorldView-3 Sensor Suite. 2015.

[27] Pontus Olofsson, Giles M Foody, Stephen V Stehman, y Curtis E Woodcock. Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation. *Remote Sensing of Environment*, 129:122–131, 2013.

[28] Bernhard Schölkopf y Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. 2002.

[29] Richard Sliuzas, Gora Mboup, y Alex de Sherbinin. Report of the expert group meeting on slum identification and mapping. *Report by CIESIN, UN-Habitat, ITC*, página 36, 2008.

[30] Robert B Smith. Introduction to Remote Sensing of Environment. 2006.

[31] S Vyas y L Kumaranayake. Constructing socio-economic status indices: how to use principal components analysis. *Health Policy and Planning*, 21(6):459–468, 2006.