Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL INTEGRADOR

Incidencia de las condiciones socio-económicas en las solicitudes ciudadanas

Análisis de la población según los radios censales de la Ciudad Autónoma de Buenos Aires

AUTOR: FLORENCIA DIAZ

MENTOR: BLANCA ROSA VITALE

12/2020

Resumen

El Gobierno de la Ciudad Autónoma de Buenos Aires, en una forma de interacción con la sociedad, pone a disposición el Sistema Único de Atención Ciudadana por el cual miles de personas realizan, cada año, distintos tipos de solicitudes a las autoridades. Las mismas abarcan dieciséis temáticas y entre las más solicitadas se encuentran las referidas al alumbrado, al arbolado, al estado de calles y veredas y a la limpieza. Una gran cantidad de estos pedidos podrían ser anticipados a partir de un buen entendimiento de las necesidades de los ciudadanos. El objetivo de este estudio es analizar las características socio-económicas de la población de las que se prevé cierta relación con determinados tipos de reclamos que realizan los residentes de la Ciudad Autónoma de Buenos Aires. A partir de la búsqueda de patrones que permitan caracterizar a la población en distintos segmentos, se podrá lograr un mejor conocimiento de la dinámica social con relación a los reclamos que realizan los ciudadanos. Este entendimiento se traduce en una mejor planificación y en la utilización eficiente de los recursos por parte de los representantes y en niveles de satisfacción más elevados por parte de los ciudadanos.

Palabras clave: condiciones socio-económicas, solicitudes ciudadanas, radios censales, población, Ciudad Autónoma de Buenos Aires.

Índice

1. Introducción

Miles de datos de diversos orígenes y naturaleza son los que almacena el Gobierno como consecuencia de las diferentes interacciones que realiza con los individuos. Cada año, el Gobierno de la Ciudad Autónoma de Buenos Aires recibe ciento de miles de peticiones por parte de los ciudadanos a través de su sistema de gestión de solicitudes: el Sistema Único de Atención Ciudadana (SUACI). Muchos de estos reclamos son resueltos, algunos en tiempos más oportunos que otros, pero es probable que varios de ellos habrían podido ser evitados si el gobierno se hubiese anticipado a las necesidades de los residentes.

El sistema vigente desde 2011 recibe un gran caudal de datos que, enriquecidos con otras fuentes externas y abiertas (como las provenientes del Portal de Datos Abiertos de la misma jurisdicción), permiten realizar un análisis de la sociedad que al gobierno le facilitaría anticiparse a los requerimientos de la misma de forma más precisa, atenta y proactiva. Un mejor entendimiento de las distintas necesidades de los ciudadanos conlleva a una planificación más óptima de los recursos además de mantener a los ciudadanos conformes. ¿Cómo puede caracterizarse a la población para entender y manejar sus solicitudes ciudadanas de forma conveniente?

La posibilidad de contar con diversidad de datos en un volumen considerable, facilita los medios para llevar adelante tareas de analítica avanzada de datos. Técnicas de aprendizaje automático como clustering permiten realizar tareas de segmentación que facilitan el estudio de patrones y es de suma utilidad para encontrar características en común, y a la vez diferentes, entre los grupos resultantes. Esta técnica puede ser fácilmente complementada con otras, ya que la segmentación es muchas veces utilizada como punto de partida. La interacción de diversas ramas como la programación y la estadística promueven y facilitan estas prácticas al permitir la manipulación de datos a la vez que se estudia su comportamiento ofreciendo un ambiente propicio para el análisis.

El objetivo de este estudio es analizar las características socio-económicas de la población de las que se prevé cierta relación con determinados tipos de reclamos que realizan los residentes de la Ciudad Autónoma de Buenos Aires. Este objetivo se desagregue en otros de menor alcance y en una serie de tareas que se irán comentado a lo largo del trabajo en

función de poner a prueba a la hipótesis bajo estudio. Las mismas serán comentadas en detalle y acompañadas de gráficos y/o tablas de encontrarse pertinente.

El trabajo consta de cuatro apartados. El primero comienza con una descripción formal del Gobierno de la Ciudad de Buenos Aires para terminar en una caracterización sobre la puesta en marcha de políticas de Gobierno Abierto en la Ciudad. La exposición de este paradigma es de vital importancia ya que este trabajo está desarrollado a partir del consumo de parte de sus resultados, que son los Datos Abiertos. Se comentan las ley y decretos sancionados que favorecen las políticas con este propósito y se describe cómo es su implementación en la actualidad.

En el apartado siguiente, se detallan las cuestiones metodológicas. Primero, se comenta sobre la recolección de los datos y el software utilizado, para luego, continuar con la explicación de cada una de las tareas desarrolladas. El objetivo de este apartado es la elaboración de la tabla input a la que se le implementará el algoritmo de aprendizaje automático seleccionado que es el clustering. Este apartado se centra en cuestiones de tratamiento de los datos como son la limpieza, normalización y transformación de los datos de las fuentes de diversos orígenes.

En el tercer apartado, se comenta el proceso de implementación del clustering. El mismo abarca la toma de la decisión sobre cuántos grupos es óptimo generar y las métricas que se utilizan para decidir esta definición. Una vez obtenidos los diferentes grupos se exponen los resultados del mismo y se caracteriza a cada uno de ellos en virtud de las características socioeconómicos y reclamos realizados por medio del Sistemas de Atención Único Ciudadano considerados para este fin.

Para finalizar, se hace una breve recapitulación de lo elaborado. Se exponen las conclusiones alcanzadas y una serie de reflexiones que surgieron como consecuencia del desarrollo del trabajo. A su vez, se plantean futuros temas de investigación que profundizarían el presente trabajo, así como desarrollos complementarios que se podrían realizar a partir de los resultados obtenidos.

2. Gestión de datos en contextos organizacionales

A lo largo de este apartado se estudia al Gobierno de la Ciudad de Buenos Aires con el fin de conocer su estructura y entender cómo afrontan el manejo de datos resultantes de la administración pública y de la interacción con los ciudadanos, según la política de Datos Abiertos. El estudio comienza con una descripción formal del Gobierno, en la primera sección, para proseguir en la siguiente con una caracterización sobre cómo aborda una de las temáticas en auge de la política: el Gobierno Abierto. Allí, se comentan la ley y decretos sancionados para su favorecimiento y cómo es su implementación en el entorno actual. Luego, en una tercera sección, se ahonda en cómo es la puesta en marcha de este paradigma mediante los Datos Abiertos. Se pone especial énfasis en la importancia de la apertura de datos y en el mantenimiento del portal que alimentan. Para concluir el estudio, en la última sección se exponen las problemáticas sobre la gestión de datos en el gobierno sobre diferentes aspectos: el tecnológico y el humano.

2.1 Descripción de la organización

Desde el año 1996, luego de declarada su autonomía, la forma de gobierno adoptada por el Poder Ejecutivo de la Ciudad Autónoma de Buenos Aires es la Jefatura de Gobierno de la Ciudad de Buenos Aires. Esta forma de gobierno es ejercida por el Jefe de Gobierno cuya duración en el cargo es de cuatro años y con la posibilidad de reelección consecutiva por un solo período. La representación es definida mediante elecciones democráticas por medio del por voto popular. Junto al Jefe de Gobierno también es elegido el Vicejefe de Gobierno, siendo ambos compañeros de fórmula. En caso de ausencia temporal o de forma definitiva por acefalía, la función del Vicejefe es la de asumir el mando del Poder Ejecutivo.

El Jefe de Gobierno puede designar a sus ministros y secretarios según la Constitución de la Ciudad de Buenos Aires. De él van a depender el Jefe de Gabinete de Ministros y los nueve ministros a cargo de las áreas de: Hacienda y Finanzas, Justicia y Seguridad, Salud, Educación, Desarrollo Económico y Producción, Cultura, Desarrollo Humano y Hábitat, Espacio Público e Higiene Urbana y el Ministerio de Gobierno. Además, de su figura dependen seis Secretarías: Secretaría General y Relaciones Internacionales, Secretaría Legal y Técnica, Secretaría de Medio, Secretaría de Asuntos Estratégicos,

Secretaría de Ambiente y la Secretaría de Comunicación, Contenidos y Participación Ciudadana.

En la Ciudad existe una intención de descentralización de las tareas. En septiembre de 2005, se reglamentó la Ley Orgánica de las Comunas por Ley Nº 1777. La misma formula en su cuerpo que en el ámbito de la Ciudad Autónoma de Buenos Aires la democracia tenga carácter participativo. Bajo este lineamiento y con el objetivo de crear unidades administrativas más pequeñas con gobiernos autónomos que puedan atender de forma más cercana la problemática vecinal, se propuso la descentralización del poder a través de la creación de las comunas y de la sanción de un presupuesto participativo. Este esquema fue implementado en 2007 respondiendo a un proceso de fortalecimiento institucional.

2.2 La apertura de datos como política de gobierno

El actual Jefe de Gobierno asumió al inicio de su primera gestión los denominados "Compromisos de Gobierno", una serie de objetivos específicos y medibles. La iniciativa, coordinada por la Secretaría General y Relaciones Internacionales, está basada en buenas prácticas internacionales y combina elementos de la administración por resultados y de gobierno abierto. Se trata de un método de gestión que promueve el uso de datos y evidencia sustentada para la toma de decisiones y el seguimiento de las mismas. Buenos Aires es una de las ciudades pioneras en América Latina en implementarla.

Los "Compromisos de Gobierno" tienen dos objetivos. Por un lado, permiten profundizar la cultura de rendición cuentas hacia los vecinos de la Ciudad, comunicando y publicando en forma periódica avances y problemáticas encontradas. El objetivo fundamental de la transparencia política es el de establecer y mantener una relación de confianza entre la ciudadanía y los poderes públicos. Por otro lado, alinean y orientan a las diferentes áreas de gobierno tras una visión común con objetivos claros y medibles, buscando conseguir una organización más eficaz e integrada de la gestión.

En el Gobierno de la Ciudad, la política de apertura de datos se inició en 2012 con la firma del Decreto 156/2012 el cual impulsó la creación de su portal "data.buenosaires.gob.ar". La plataforma de datos públicos de la Ciudad se originó con el

objetivo de facilitar la búsqueda y acceso de los datos abiertos producidos por la Ciudad, promoviendo, de esta forma la transparencia, participación y colaboración con la ciudadanía. En 2013, se le dio continuidad a la política de apertura y se sancionó el Decreto 478/2013 que dio un nuevo impulso a esta política al establecer la apertura por defecto: desde entonces, todos los datos producidos, almacenados y/o recolectados en medios digitales por los órganos de la Administración Centralizada y Descentralizada, así como Entidades Autárquicas del Gobierno de la Ciudad, deben ser publicados en el portal anteriormente mencionado.

Adicionalmente, a fines de 2015, la Ciudad firmó la Carta Internacional de Datos Abiertos a través de la cual se compromete a seguir e implementar los lineamientos que establecen que los datos públicos deben ser: abiertos por defecto, oportunos y exhaustivos, accesibles y utilizables, comparables e interoperables, orientados a mejorar la gobernanza y la participación ciudadana y para el desarrollo incluyente y la innovación. La sanción de la nueva Ley de Acceso a la Información Pública (Ley Nº104) implicó también un avance en materia de apertura ya que estableció el "formato abierto" como criterio fundamental.

Asimismo, el Gobierno de la Ciudad Autónoma de Buenos Aires lleva adelante la denominada “Agenda de Transparencia e Innovación Institucional”. La misma trata de iniciativas orientadas a promover un gobierno abierto y mejorar la calidad de las instituciones públicas. Mediante esta agenda, se fomentan políticas e iniciativas que puedan incorporar estándares y herramientas que fortalezcan los procesos de apertura, datos abiertos, calidad institucional e innovación, como resultado del trabajo de forma conjunta, por ejemplo, entre la sociedad civil y organismos internacionales. La plataforma de datos abiertos de la Ciudad, “data.buenosaires.gob.ar”, es donde se encuentran centralizados los activos de información de todas las áreas de gobierno. En este sentido, funciona como un medio único que permite la interacción entre los ciudadanos y el gobierno. El portal de datos se soporta en Andino (una instancia desarrollada por el Gobierno Nacional) basado en CKAN, una de las plataformas más populares para este tipo de iniciativas. El código es abierto y cuenta con una API pública que permite un método alternativo por el cual acceder a los datos por parte de terceros.

2.3 La gestión de los Datos Abiertos

El Gobierno Abierto aparece como una nueva forma de vinculación entre los ciudadanos y la administración pública. Miles de datos que surgen como consecuencia de las actividades allí desarrolladas están al alcance de la mano. La filosofía de este paradigma busca facilitar los mecanismos de participación de los ciudadanos y promover una cualidad fundamental en los gobiernos: la transparencia. La concepción del término "Gobierno Abierto" surgió a fines de los 70 en Inglaterra. Se hablaba de la existencia de un gobierno secreto y opaco a los ojos de los ciudadanos. En aquel entonces, todavía era lejana su puesta en práctica tal como se conoce actualmente. Tal como plantea Álvaro Ramírez-Aluja (2011), hoy en día hay un consenso en los pilares que sientan las bases para llevar adelante un Gobierno Abierto: transparencia, participación y colaboración. Por transparencia se entiende a un gobierno que rinde cuentas a la ciudadanía, proporciona información sobre lo que está realizando y sus planes de actuación. Por participación a un gobierno que impulsa herramientas que aumentan la participación de las personas en la elaboración de políticas públicas y promueve la responsabilidad cívica. Por último, la colaboración*:* un gobierno que no solo involucra a la ciudadanía, sino también a asociaciones, empresas y demás agentes públicos y privados para el intercambio de conocimientos, necesidades, problemas y soluciones, es un gobierno que se abre, también, al resto de las Administraciones.

Una de las formas en que se materializa el Gobierno Abierto es a través de los portales de Datos Abiertos. Se tratan de medios que, buscando la simpleza en la interacción y la abundancia en la información, posibilitan un mejor conocimiento del funcionamiento del gobierno y facilitan la colaboración civil. Como señala Alejandra Lagunes (2015), los datos no solo son un elemento primordial de la transparencia sino, que son facilitadores de co-creación entre gobierno y sociedad para incrementar la productividad y promover la innovación. Los datos bajo este origen son catalogados por Kolanovic y Krishnamachari (2017) como datos alternativos producto de los procesos de negocio de entidades públicas, que posibilitan la obtención de datos que favorecen el descubrimiento de nueva información no contenida en fuentes tradicionales.

En el marco de la estrategia de Gobierno Abierto, la iniciativa del Portal de Datos Abiertos de la Ciudad "BA Data" (anteriormente mencionado como "data.buenosaires.gob.ar") pone a los datos resultantes de la gestión pública a disposición de todos los vecinos a fin de favorecer la rendición de cuentas, fomentar la transparencia y la innovación como valores de gestión y promover el desarrollo económico. Sobre el último punto, se prevé que a través del libre acceso a los datos es posible desarrollar nuevos conocimientos e ideas innovadoras para generar beneficios sociales y valor económico en el desarrollo de las ciudades. La estrategia de Gobierno Abierto está a cargo de la Dirección General de Calidad Institucional y Gobierno Abierto, que forma parte de la Subsecretaría de Gestión Estratégica y Calidad Institucional, dependiente de la Secretaría General y Relaciones Internacionales.

La unidad mínima disponible en BA Data son los llamados "recursos" que pueden encontrarse en formatos varios según el tipo de dato que contengan: hay desde csv hasta archivos tipo shape que representan información geográfica. Los recursos están agrupados por conjuntos de datos denominados "datasets" y estos son clasificados en categorías que facilitan su localización y acceso por parte del usuario. Los datasets cuentan con un recurso especial conocido como la guía de datos. Estas guías tienen la finalidad de funcionar como un diccionario de datos que permite la interpretación de los mismos y las variables contenidas en los recursos. En este portal, además, se disponibilizan APIs desarrolladas por otras áreas, tales como la API Unificada de Transporte o los servicios web de la Unidad de Sistemas de Información Geográfica, que representan otra forma de consumir los datos.

La apertura de datos es una estrategia que involucra la participación activa de todas las áreas de Gobierno. Periódicamente, las bases con orígenes en los distintos órganos son recolectadas por el equipo encargado de llevar adelante las tareas de mantenimiento del Portal de Datos Abiertos. Estas tareas van desde la recolección del dato, la transformación y normalización del mismo, hasta su publicación o actualización en el portal. Para llevar adelante estas tareas, el equipo cuenta con la Guía de Apertura de Datos de la Ciudad de Buenos Aires. Se trata de una guía para la publicación de datos en formatos abiertos del Gobierno de la Ciudad, la cual define los estándares de calidad a ser utilizados al momento de realizar el tratamiento de los recursos y datasets publicados en el sitio.

Por otra parte, el equipo encargado de BA Data debe garantizar las cualidades de los datos para que sean considerados abiertos. La iniciativa Open Government Data[1] definió ocho principios que promueven estas cualidades: completitud (debe estar disponible y no tener restricciones para la apertura), primario (estar publicado con el mayor nivel posible de detalle), oportuno (estar disponible tan rápido como sea necesario), accesible (estar disponible para el mayor espectro de ciudadanos), procesable (estar razonablemente estructurado), no discriminatorio (estar disponible para todos), no propietario (estar disponible en un formato tal que ninguna entidad tenga un control exclusivo) y libre de licencias (no estar sujetos a ningún derecho de autor, parante, marca registrada o regulación de secreto comercial).

Al día de hoy, se encuentran publicados 391 dataset que involucran a 31 organismo de gobierno diferentes. Los mismo se encuentran distribuidos en 12 categorías que cubren las distintas temáticas posibles: Administración Pública, COVID-19, Cultura y Turismo, Desarrollo Humano, Economía y Finanzas, Educación, Género, Medioambiente, Movilidad Salud, Seguridad y Urbanismo y Territorio. Asimismo, se pueden acceder a trabajos realizados por ciudadanos con los datos disponibles en el portal.

2.4 Problemáticas de la organización sobre la gestión de los datos

Los gobiernos se enfrentan ante la necesidad de generar las capacidades adecuadas en sus empleados para poder aprovechar las nuevas tecnologías (Big Data e Inteligencia Artificial) y fortalecer así la toma de decisiones basadas en evidencia. Buscan poder desarrollar investigaciones aplicando técnicas de aprendizaje automático y visualización de datos para distintos ámbitos de aplicación. Como se comentó en el apartado anterior, el gobierno tiene la capacidad de, en cierta forma, centralizar datos relevantes de distintas áreas y garantizar la calidad de los mismos gracias al trabajo que lleva adelante el equipo encargado de implementar las técnicas que exhibe la la Guía de Apertura de Datos de la Ciudad de Buenos Aires.

[1] Open Government Data (OGD) es una filosofía que promueve la transparencia, la rendición de cuentas y la creación de valor al hacer que los datos gubernamentales estén disponibles para todos.

Los beneficios del Gobierno Abierto no son solo para los ciudadanos. La apertura de datos permite encontrar información complementaria en otras áreas de gobierno para diseñar políticas más focalizadas en los problemas y necesidades de los ciudadanos. Entre más organismos trabajen con catálogos estandarizados, más oportuna será la información disponible para planificar, implementar y evaluar las políticas públicas. La posibilidad de complementar estos datos con otros de diversos orígenes como otros organismos públicos (por ejemplo, el INDEC) o con entes privados que trabajen también en función de este paradigma, hace que los horizontes de creación de conocimiento se acrecienten.

La posibilidad de contar con esta diversidad y volumen de datos, facilita los medios para llevar adelante análisis más complejos y ricos en información. Técnicas de minerías de datos como clustering o series temporales, permiten darles entidad y personalidad a los distintos grupos resultantes de las características similares y disímiles existentes, a la vez que se pueden estudiar patrones y proyectar sobre el accionar futuro por parte de la población, respectivamente. Otras técnicas como los sistemas de recomendación, podrían mejorar la experiencia de usuario personalizando webs o aplicaciones de forma tal de ajustarse a los gustos o necesidades de los ciudadanos. Herramientas como la programación, en diversos lenguajes, promueven y facilitan estas prácticas al permitir la manipulación de datos en sus formas más amplias y versátiles como el ofrecer, también, un ambiente propicio para la experimentación. Esta ramificación del análisis de datos convencional no solo modernizará al Estado y lo impulsará en su búsqueda de toma de decisiones basadas en evidencia, sino que también se traducirá en un Gobierno más atento y cercano a la ciudadanía

Como plantea David Salgado (2016), otro de los desafíos vinculados con el procesamiento de datos es la posibilidad de contar con un marco tecnológico adecuado con la infraestructura y recursos necesarios. La adopción de fuentes Big Data y/o implementación de algoritmo de aprendizaje automático trae consigo la modernización de las oficinas públicas, además de la necesidad de disponer de empleados con perfiles profesionales más heterogéneo y flexibles. La conformación de estos equipos interdisciplinarios con visiones conjuntas más completas que permitan la interpretación de la realidad desde diferentes ópticas no le es ajeno al Gobierno. Las inversiones en tecnología comprenden la necesidad de contar con espacios de trabajo que respondan a los requisitos

de los equipos de forma tal que puedan escalar o responder en tiempo a las exigencias de los algoritmos más costosos, tanto computacional como económicamente hablando. Hoy en día, los servicios en nube permiten lidiar con varias de estas problemáticas, pero en determinadas cuestiones podrían entrar en conflicto con las normativas regulatorias en asuntos de seguridad y privacidad dispuestas por tratarse de datos de los ciudadanos resultantes de la interacción con la administración pública.

Por último, dejando de lado cuestiones tecnológicas, aparece el reto de la comunicabilidad. Tal como afirmó Walter Sosa Escudero (2017) en la conferencia realizada sobre la implementación de big data en las estadísticas públicas en el marco de la firma del acuerdo para la cooperación en temáticas de innovación estadística del INDEC con el Central Bureau of Statistics (CBS) de los Países Bajos, los algoritmos tienen un objetivo que no es necesariamente el de explicar, sino el de predecir correctamente y se diseñan con ese objetivo. Pero, cuando se trata de políticas públicas de gobierno, el poder brindar una explicación de cómo se llegó a los resultados toma especial relevancia. Es por esto que se tiene que tener en cuenta al momento de llevar adelante un proyecto de datos, que las técnicas elegidas pueden ser no necesariamente las mejores dese un punto de vista técnico y analítico, pero pueden tener características comunicacionales deseables. El servicio que ofrece la política pública muchas veces no es necesariamente un servicio algorítmico y computacional, sino que debe fundirse, también, de algún tipo de acuerdo que es entre social, comunicacional y, fundamentalmente, político.

3. Descripción metodológica

En este apartado, se mencionan las distintas fuentes de datos consideradas, el proceso de recolección de sus datos y el tratamiento de cada una para poder vincular diversos orígenes en una única tabla input para el modelo de aprendizaje automático no supervisado. Los procesos desarrollados comprenden la asignación del código de radio censal a cada registro, la limpieza de datos y la generación del indicador anteriormente mencionado sobre el valor del m2 de cada radio. Para finalizar, se mencionan los criterios finales de selección de registros y se detallan las variables creadas derivadas de las originales.

3.1 Recopilación de la información

Para el desarrollo del trabajo se consideraron diversas fuentes de datos. La principal, por ser la unidad de estudio geoestadístico, son los radios censales de la Ciudad. A cada radio se lo enriquece con datos estadísticos poblacionales del censo 2010 (en la sección 3.5 se especifican las variables seleccionadas) y con datos del Portal de Datos de la jurisdicción bajo análisis, tal como las solicitudes ciudadanas (2019) y del área de seguridad, los delitos (2019). Otra variable de interés a considerar referente a las condiciones socio-económicas es el valor del m2 representativo de cada radio censal de la Ciudad. Para ello, se crea un símil indicador del precio del m2 para cada caso a partir de datos entre 2018 y 2019 del mercado inmobiliario con origen en el portal "Properati". Estos datos responden a la política de apertura de datos de la empresa y se encuentra disponible para su consumo vía API. La construcción del indicador se detalla en la sección 3.4.

La fecha de publicación del dataset de SUACI es del 8 de enero de 2019 y de los datos de delitos la fecha es del 5 de marzo de 2020. Para las fuentes de radios censales, censo 2010 y Properati no se posee información. En cuanto a la fecha de actualización, al momento de acceder a los datos, la fecha notificada para SUACI es del 7 de mayo de 2020 y para delitos el 29 de junio de 2020. Para el resto de las fuentes, al igual que el punto anterior, no se posee información. Para lo que es fecha de relevamiento, ninguna de las fuentes especifica fecha precisa.

Como puede observarse en el universo de datos contemplado, una parte considerable provienen del Portal de Datos Abiertos. Por un lado, se encuentra SUACI que son los datos correspondientes a las solicitudes ciudadanas realizadas en la Ciudad Autónoma de Buenos Aires, Argentina. El responsable de esta fuente es la Dirección General de Atención y Cercanía Ciudadana de la Secretaría de Atención y Gestión Ciudadana que forma parte de la Jefatura de Gabinete de Ministros en su jerarquía mayor. Por otro lado, se encuentra la base de los delitos que ocurrieron en la Ciudad Autónoma de Buenos Aires cuyo responsable es la Policía de la Ciudad del Ministerio de Justicia y Seguridad.

En contraposición a lo mencionado anteriormente, hay una parte de las bases utilizadas que no forman parte de un portal de datos gubernamental. Siguiendo con los datos oficiales, quedan por mencionar los datos abiertos de radios censales y el censo 2010. El primero, se trata de la base geográfica de los radios censales de la Ciudad Autónoma de Buenos Aires (CABA) elaborada por el Instituto Nacional de Estadística y Censos (INDEC) para el Censo Nacional de Población, Hogares y Viviendas 2010 (CNPHyV 2010) en base a información provista por las Direcciones Provinciales de Estadística (DPE). El responsable es el área de Cartografía y códigos geográficos del Sistema Estadístico Nacional que forma parte de Unidades Geoestadísticas. Los datos correspondientes se obtienen por medio de la página oficial del organismo. El segundo, se trata de las estadísticas poblacionales correspondientes al censo oficial de 2010 cuyo responsable también es el INDEC. En este caso, los datos se obtienen por medio del sistema REDATAM. El censo se realiza a nivel país, pero para el estudio en cuestión se limitan los datos a CABA.

En cuanto a Properati, se trata de una iniciativa de datos abiertos pero privada. Esta empresa disponibiliza avisos de propiedades publicadas en su portal bajo responsabilidad de “Properati Data”. Si bien el alcance de los datos excede a publicaciones en Argentina, para este estudio se limita a las operaciones en la Capital Federal. Los datos de Properati Data fueron extraídos con la API puesta a disposición por parte de la empresa mediante BigQuery.

Para el análisis, preprocesamiento de datos y e implementación del modelo de clustering se utilizó el software R en su versión 3.6.3 mediante la interfaz RStudio 1.2.1335. Para el análisis y procesamiento geográfico de los datos se usó el software de información geográfica QGIS en su versión 3.14.1 con GRASS 7.8.3. A continuación, se procede a comentar el análisis y tratamiento de los datos realizado con las herramientas mencionadas para generar la tabla input del modelo.

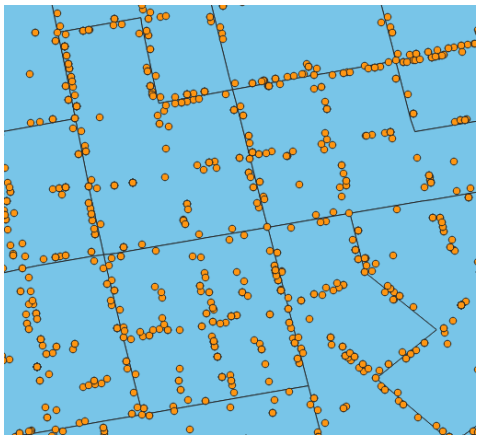
3.2 Asignación del código de radio censal a los datos

Originalmente, la base de radios censales cuenta con un total de 3.555 registros debido a que dos radios se encontraban fragmentados, pero cada parte mantenía el mismo código de radio. Se procedió a la reunificación de las unidades con QGIS llegando a la cantidad total de 3.553 radios en la Ciudad Autónoma de Buenos Aires. Una vez corregidos los radios censales, tanto a las propiedades de Properati como a las solicitudes de SUACI y a los delitos, se les asignó el código de radio censal correspondiente a partir de la geolocalización de los mismos mediante su latitud y longitud. El proceso también fue realizado con el software QGIS.

Mapa de los 3.553 radios censales de la Ciudad

Fuente: elaboración propia con QGIS.

Ejemplo de geolocalización de solicitudes de SUACI



Fuente: elaboración propia con QGIS.

3.3 Limpieza de datos

Un valor atípico (outlier) es una observación que es numéricamente distante del resto de los datos. A lo largo de esta sección se tomó como referencia el rango intercuartílico (IQR) para la detección de valores atípicos. El IQR Se define como la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1), es decir: RQ = Q3 - Q1. A su vez, se diferencian entre valores atípicos leves y extremos. Los valores atípicos leves son aquellos que se encuentra a una distancia equivalente a 1.5 veces el IQR tanto respecto del Q1 como del Q3. En cambio, los extremos son aquellos cuya distancia es de 3 veces el IQR. Para ciertas

fuentes utilizadas en este trabajo se aplicó la detección de outliers extremos cuando se quiso preservar cierto grado de dispersión de datos que se considere oportuna.

a) Censo

La base del censo se filtró para que considere únicamente a los datos de los 3.553 radios censales de la Ciudad. No se realizó ningún otro tratamiento sobre el dataset además del mencionado en el apartado 3.2. En el apartado 3.5 se identifican las variables consideradas de esta fuente de datos y de las demás.

b) SUACI

El dataset completo de SUACI 2019 cuenta con 989.629 registros. Primero, se seleccionaron aquellos contactos con radio censal asignado y bajo el tipo de prestación "Solicitudes". Luego, se dejaron en consideración los que tuvieran la subcategoría no nula y se seleccionaron aquellos cuya subcategoría del reclamo se encontrase dentro de los 10 tipos de solicitudes más frecuentes. Por último, se agruparon la cantidad de solicitudes por radio censal y se detectaron outliers extremos (IQR * 3) en cada variable a las que se le imputó el valor del límite superior correspondiente. El proceso finalizó con 208.366 registros.

En la tabla se presentan el top 10 de los tipos de solicitudes más frecuentes. Los mismos acumulan casi el 85% de los reclamos y se consideran suficientemente relevantes y representativos del total. Además, abarcan diferentes temáticas: estado de calles y veredas, limpieza, arbolado e iluminaria.

Top 10 de los tipos de solicitudes más frecuentes

	SUBCATEGORÍA	CANT. RECLAMOS	CANT. ACUMULADA	% TOTAL	% RELATIVO
1	REPARACIÓN DE VERDA	46.231	46.231	22%	19%
2	LIMPIEZA DE VÍA PÚBLICA	37.799	84.030	18%	34%
3	PODA DE ÁRBOL Y DESPEJE DE LUMINARIA	29.818	113.848	14%	46%
4	CESTO Y CONTENEDORES	23.431	137.279	11%	55%

5	REPERACIÓN DE LUMINARIA	19.698	156.977	9%	63%
6	VEHÍCULOS ABANDONADOS	15.636	172.613	8%	69%
7	PROBLEMAS CON INTERVENCIÓN DE ARBOLADO	10.530	183.143	5%	74%
8	REPARACIÓN DE BACHES	9.142	192.285	4%	77%
9	EXTRACCIÓN DE ÁRBOL	8.815	201.100	4%	81%
10	MAYOR ILUMINACIÓN EN CALLE	7.266	208.366	3%	84%

Fuente: elaboración propia con R.

c) Delitos

El dataset completo de Delitos 2019 cuenta con 117.661 registros. En primer lugar, se seleccionaron aquellos delitos con radio censal asignado y los tipificados como “Hurto (sin violencia)” y "Robo (con violencia)" por ser, notoriamente, los de mayor frecuencia. Luego, se filtraron los delitos que no tenían datos sobre la franja horaria en que ocurrió. Por último, se agrupó la cantidad de delitos por radio censal y se detectaron outliers extremos (IQR * 3) sobre el total de delitos a los que se les imputó el valor del límite superior. El proceso finalizó con 99.948 registros.

d) Properati

Al momento de realizar la consulta vía API se extrajeron las publicaciones realizadas entre enero de 2018 hasta diciembre de 2019 en Argentina, un total de 1.468.880 registros. Una vez que los datos fueron obtenidos, se realizaron los siguientes filtros:

- Se desechó el id de publicación y se detectaron duplicados sobre los campos restantes para su eliminación.
- Publicaciones de propiedades en Capital Federal.
- Publicaciones con radio censal asignado.
- Precio de la propiedad expresada en dólares y mayor que 0.
- Tipo de propiedad: casa, departamento o PH.
- Tipo de operación: venta.

- Número de habitaciones mayores a 0.
- Superficie cubierta total mayor a 20 mts2.

Por último, se detectaron y eliminaron los registros con valores atípicos aplicando la técnica del rango intercuartílico (1.5 * IQR) sobre el campo "relación" creado a partir de dividir el precio sobre la superficie total. El proceso finalizó con 169.939 registros.

3.4 Generación del indicador "nivel precio m2"

Una vez que la data de Properati fue homogenizada, se creó el indicador "nivel precio m2" a partir de la valorización de las propiedades resultantes con la generación de deciles a partir de la relación precio/superficie total. Los deciles 1 y 2 fueron categorizados como "bajo", 3 y 4 como "medio-bajo", 5 y 6 como "medio", 7 y 8 como "medio-alto", y 9 y 10 como "alto" haciendo referencia al nivel del precio del m2. En la tabla se presentan los rangos de precios por nivel del precio m2 resultante del desarrollo.

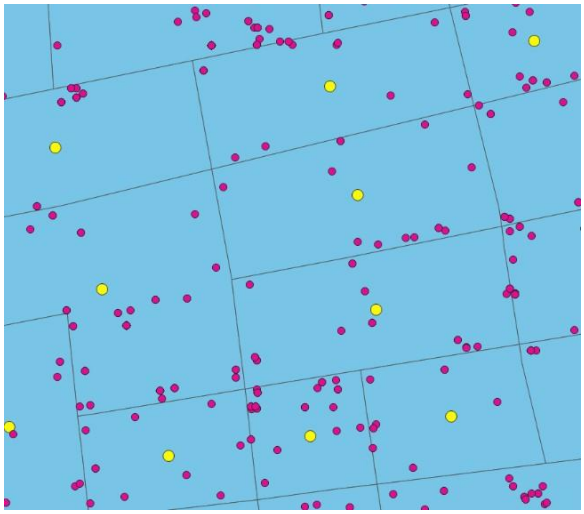
Rangos de precios por nivel del precio m2

NIVEL PRECIO M2	PRECIO M2 MIN	PRECIO M2 MAX
ALTO	3250	4824
MEDIO-ALTO	2764	3250
MEDIO	2400	2764
MEDIO-BAJO	1975	2400
BAJO	493	1975

Fuente: elaboración propia en R

Luego, se realizó un proceso iterativo en el que, para cada radio censal, se utilizó la técnica de Vecinos Más Cercanos (KNN) lo que permitió asignarle a cada radio el nivel del precio m2 correspondiente. Dado que un radio censal es un polígono, se detectaron los centroides de cada figura que fueron utilizados como punto de referencia del radio censal. En la imagen se representan los centroides de los radios en amarillo y las propiedades de Properati en rosa.

Representación de centroides y propiedades

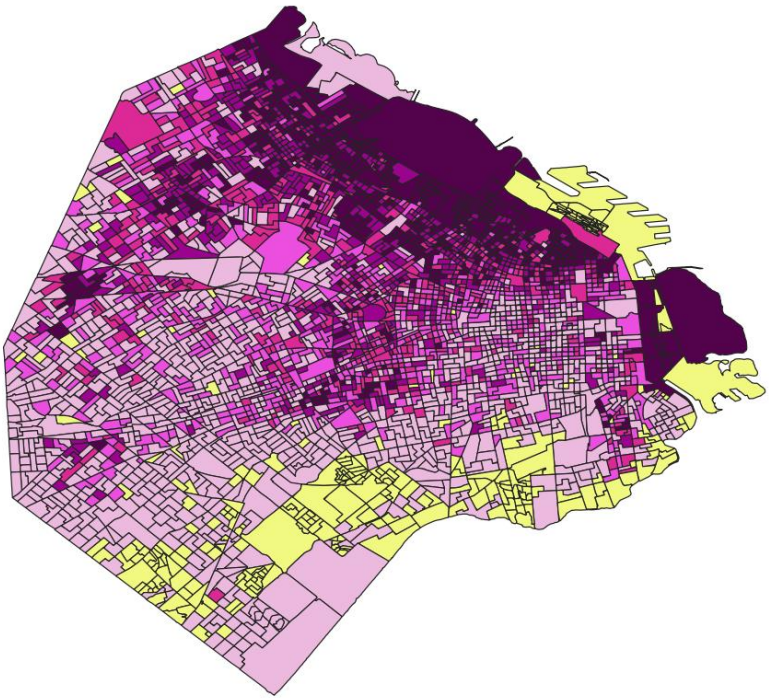


Fuente: elaboración propia en QGIS.

El paquete utilizado en R fue Caret. Este paquete proporciona una interfaz general para casi 150 algoritmos machine learning. También proporciona excelentes funciones para muestrear los datos (separar entre datos de entrenamiento y datos de testeo), preprocesamiento, evaluación del modelo, entre otros. El algoritmo que provee este paquete determina el número óptimo de "k" para cada caso, por lo que se aplicó a medida para cada radio censal utilizando la distancia euclidiana. Dado que no se encontraron propiedades en todos los radios censales, se tomaron en consideración aquellos radios en los que hay, por lo menos, siete propiedades publicadas. De esta manera, se les asignó a 3.255 radios censales el nivel precio m2 elaborado anteriormente, delimitando las publicaciones de Properati a cada radio censal.

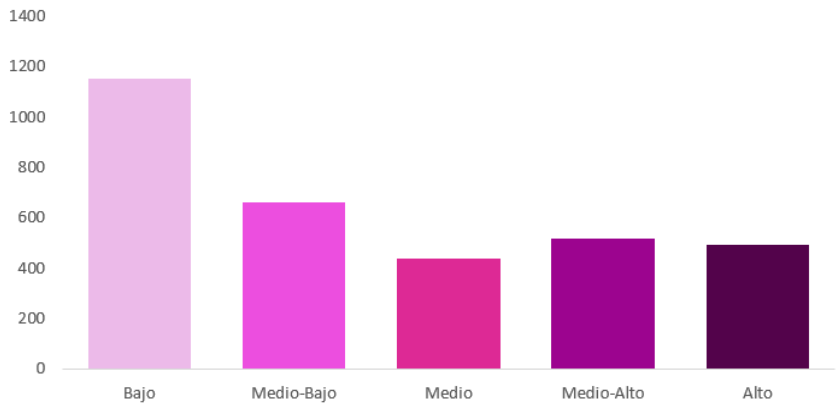
Para visualizar los resultados se generó el mapa resultante en QGIS. En la escala del violeta se representa el nivel del precio m2: colores más intensos para niveles más alto y viceversa para los más bajos. En amarillo se identifican aquellos radios censales que quedaron excluidos del proceso. Puede observarse que no se encontraron propiedades, en parte, en zonas portuarias y barrios vulnerables. Además, se percibe hay una predominancia de los niveles considerados como precio m2 bajo.

Mapa de los radios censales de la Ciudad según el nivel precio m2.



Fuente: elaboración propia en QGIS en base a datos del INDEC y Properati Data.

Cantidad de radios censales según nivel precio m2.



Fuente: elaboración propia en R en base a datos del INDEC y Properati Data.

3.5 Criterios de selección de registros y estructura final

En el apartado 3.4 se comentó el proceso de KNN y se enunció la condición de que el radio censal debía contar con un mínimo de siete propiedades publicadas. También, se filtraron aquellos radios censales en los que había una cantidad menor a diez solicitudes y menor a tres denuncias de delito. De esta manera se obtiene mayor robustez en el análisis al conservar únicamente radios censales con inquietudes. En cuanto a SUACI, la cantidad de solicitudes finales pasó de 208.366 a 207.529 y en Delitos de 99.948 a 99.899 registros.

Como última condición, el radio censal debía contar con información de las tres fuentes de datos consideradas: SUACI, Delitos y Properati. De esta manera, de los 3.553 radios censales se consideraron, finalmente, 3.169. Una vez que finalizado el preprocesamiento de los datos, se procedió a la unión de las salidas de las distintas fuentes según el código de radio censal. El dataset resultante finalizó con 31 variables (sin tener en cuenta el radio censal que es la unidad de análisis) y 3.169 observaciones. Por último, para realizar la técnica de clustering el dataset fue escalado: a cada variable se le restó su media y se lo dividió por la desviación típica de la misma. En la tabla a continuación, se presenta la estructura de la tabla input a la cual se le implementará el modelo de clustering.

Tabla input del modelo

NOMBRE DEL ATRIBUTO	ORIGEN	TIPO DE DATO	CREADA
radio_censal	Radios censales	Texto	NO
personas	Censo	Número entero	NO
persona_sexo_varon	Censo	Número decimal	NO
persona_sexo_mujer	Censo	Número decimal	NO
viviendas	Censo	Número entero	NO
hogar_al_menos_un_indicador_ nbi_hogares_con_nbi	Censo	Número decimal	NO
hogar_al_menos_un_indicador	Censo	Número decimal	NO

_nbi_hogares_sin_nbi			
vivienda_calidad_constructiva_ de_la_vivienda_satisfactoria	Censo	Número decimal	NO
vivienda_calidad_constructiva_ de_la_vivienda_basico	Censo	Número decimal	NO
vivienda_calidad_constructiva_de_la_vivienda_insuficiente	Censo	Número decimal	NO
vivienda_tipo_de_vivienda _particular_departamento	Censo	Número decimal	NO
vivienda_tipo_de_vivienda _particular_casa	Censo	Número decimal	NO
ocupacion_vivienda	Censo	Número decimal	SI
rango_etario_0_30	Censo	Número decimal	SI
rango_etario_31_60	Censo	Número decimal	SI
rango_etario_61_110	Censo	Número decimal	SI
persona_condicion_de_actividad _ocupado	Censo	Número decimal	NO
persona_condicion_de_actividad_desocupado	Censo	Número decimal	NO
persona_condicion_de_actividad_inactivo	Censo	Número decimal	NO
cestos_y_contenedores	SUACI	Número entero	NO
extraccion_de_arbol	SUACI	Número entero	NO
limpieza_de_via_publica	SUACI	Número entero	NO
mayor_iluminacion_en_calle	SUACI	Número entero	NO
poda_de_arbol_y_despeje _de_luminaria	SUACI	Número entero	NO
problema_con_intervenciones _de_arbolado	SUACI	Número entero	NO
reparacion_de_baches	SUACI	Número entero	NO

reparacion_de_luminaria	SUACI	Número entero	NO
reparacion_de_vereda	SUACI	Número entero	NO
vehiculos_abandonados	SUACI	Número entero	NO
cantidad_solicitudes	SUACI	Número entero	SI
cantidad_delitos	Delitos	Número entero	SI
nivel_precio_m2	Properati	Texto	SI

Fuente: elaboración propia en R.

La columna “creada” hace referencia a si es un dato elaborado o no. Los campos creados son:

- Ocupación vivienda: es el cociente entre viviendas habitadas y la cantidad de viviendas.
- Cantidad de solicitudes: es la sumatoria de la cantidad de solicitudes correspondientes a las subcategorías consideradas por radio censal.
- Cantidad de delitos: es la sumatoria de la cantidad de delitos correspondientes a las categorías consideradas por radio censal.

- Nivel precio m2 Num.: la elaboración se explica en el punto 2.3.4. Fue codificada numéricamente de la siguiente manera: Bajo = 1, Medio-Bajo = 2, Medio = 3, Medio-Alto = 4, Alto = 5.
- Rango etario 0-30: sumatoria de la proporción de edades entre 0 y 30.
- Rango etario 31-60: sumatoria de la proporción de edades entre 31 y 60.
- Rango etario 61-110: sumatoria de la proporción de edades entre 61 y 110

4. Implementación

En este apartado se introduce al método de clustering, que es la técnica de aprendizaje automático no supervisada implementada en este estudio. También, se comenta y explica el funcionamiento del algoritmo "k-means" que permite su implementación. Dado que el clustering requiere de conocer de forma previa el número de grupos a generar, se comparan distintas métricas con este objetivo: el WCSS y el silhouette. Luego, se aplica la técnica de PCA que permitirá representar gráficamente a los grupos resultantes en función de las dos primeras componentes principales. También se muestra la distribución de los radios censales en un mapa de la Ciudad y en una tabla. Por último, se caracterizan los clusters resultantes en función de los valores obtenidos en las variables consideradas.

4.1. Clustering

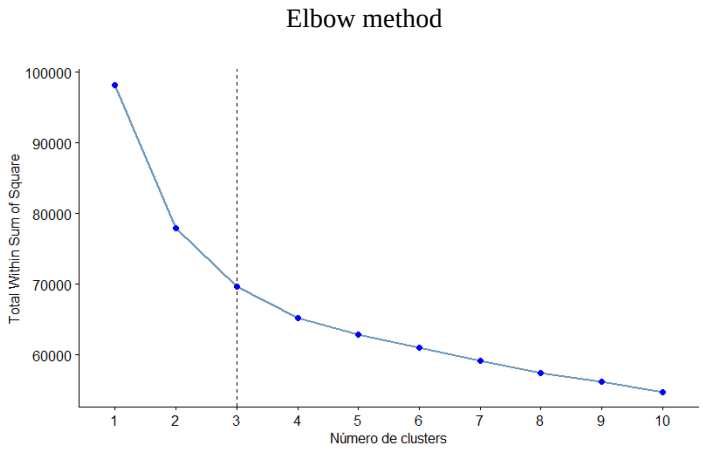
El clustering es una técnica que consiste en agrupar un conjunto de observaciones en subconjuntos de objetos llamados clusters. Cada cluster está formado por una colección de observaciones que son similares entre sí, pero que son distintas respecto a los objetos de otros grupos. Este agrupamiento se basa en la idea de distancia o similitud entre las observaciones. La obtención de dichos clusters va a depender del criterio elegido.

El método de k-medios (k-means) es uno de los algoritmos existentes para realizar la tarea de agrupamiento. Su funcionamiento se base en agrupar a los datos de entrada en un total de k conjuntos definidos por un centroide, cuya distancia con los puntos que pertenecen a cada uno de los datos es la menor posible. Este algoritmo minimiza las variaciones dentro del cluster y utiliza la distancia euclidiana al cuadrado. Para implementarlo se deben definir previamente la cantidad de clusters a generar.

K-means consta de tres pasos: una vez elegido el número de grupos se establecen k centroides en el espacio de los datos. Luego, cada observación es asignada a su centroide más cercano. Por último, se actualiza el centroide de cada grupo en función de la posición promedio de las observaciones asignadas. Se repiten los últimos pasos hasta que las observaciones no se mueven, se mueven por debajo de un umbral o se llega al límite de iteraciones.

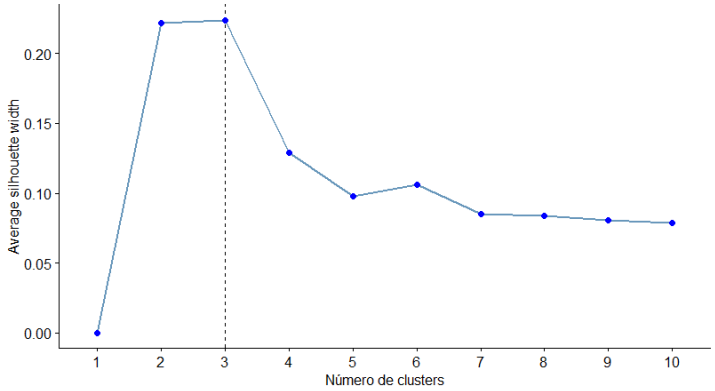
Para implementar el algoritmo de k-means se deben definir previamente la cantidad de clusters a generar. Para ello se pueden analizar diferentes métricas que ayudan en la toma de la decisión. En este estudio, se utilizaron las técnicas del "método del codo" (elbow method) y el silhouette para comprar las métricas de distintos valores simulados de k. El "método del codo" busca seleccionar la cantidad ideal de grupos a partir de la optimización de la suma de cuadrados dentro del cluster (WCSS). Cuando se grafican los resultados, se puede observar como a medida que aumenta el número de clusters, el valor de WCSS tiende a disminuir. Como número óptimo, se busca el "codo" a partir del cual ya no se producen variaciones importantes del valor WCSS al aumentar el número de k. El silhouette es un coeficiente que mide cuán estrechamente se relaciona cada dato con los datos dentro de su grupo y qué tan poco se relaciona con los datos del grupo vecino, es decir, el grupo cuya distancia promedio es más baja. Un valor del silhouette cercano a 1 implica que el elemento está en el grupo apropiado, mientras que valores cercanos a -1 implica que el elemento está en el grupo incorrecto. Por lo tanto, valores del coeficiente más altos para todo el agrupamiento son los deseados.

En este trabajo, se compararon los valores del WCSS, como resultado de aplicar el método del codo, y los valores del silhouette tanto gráficamente como en una tabla comparativa para distintos valores de k. Por último, se graficó la salida del k-means en un eje de coordenadas a partir de la detección de las dos primeras componentes principales. A continuación, se presentan los resultados obtenidos de aplicar las diferentes técnicas mencionadas para encontrar el número óptimo de clusters.



Fuente: elaboración propia en R.

Silhouette



Fuente: elaboración propia en R.

Comparación de valores para distintos k según métricas WCSS y Silhouette.

k	WCSS	SILHOUETTE
1	98208	0
2	77979.07	0.22145015
3	**69655.58**	**0.22366554**
4	65230.68	0.12870553
5	62854.08	0.09756518
6	60962.02	0.10598754
7	59228.76	0.08479101
8	57394.42	0.08351844
9	56155.47	0.08087463
10	54748.71	0.07849218

Fuente: elaboración propia en R.

Según los resultados del elbow method, podría concluirse que es en k = 3 donde se produce el "codo" a partir del cual las variaciones del WCSS empiezan a disminuir con menos fuerza, aunque el quiebre es muy leve. Por el contrario, en el silhouette si es notorio que en k = 3 alcanza su máximo valor. Es por eso que se concluye que el número óptimo de clusters a implementar es de tres.

4.2. PCA

El Análisis de Componentes Principales (sus siglas en inglés: PCA) es un método de análisis multivariante que permite disminuir la dimensionalidad perdiendo la menor cantidad de información posible. Por sus características, el PCA también es utilizado como una herramienta de visualización potente ya que permite apreciar la distribución de los datos en función de nuevas variables – componentes principales – que contienen una cantidad significativa de información. Con el objetivo de visualización, por lo general se consideran las primeras dos componentes principales.

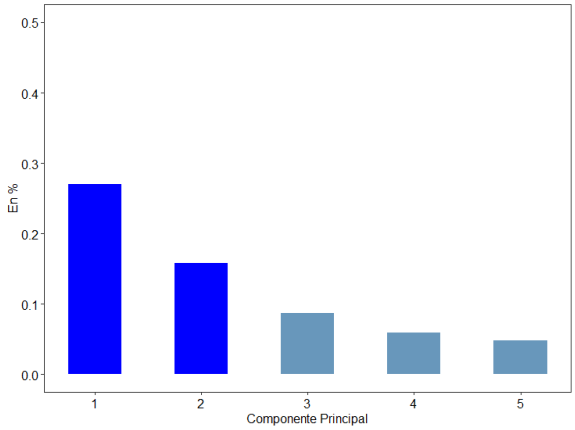
En nuestro caso de estudio, la utilidad dada a esta técnica es la de visualizar la salida de los clusters en un gráfico de coordenadas bidimensional. La misma se realizó sobre la matriz de correlaciones. En los resultados exhibidos a continuación producto de aplicar PCA sobre el dataset, se puede observar que las dos primeras componentes principales explican una cantidad significativa de información con respecto al resto de las componentes, teniendo una proporción de la varianza del 27% y 16% respectivamente. En su visión acumulada, las dos primeras componentes principales acumulan el 43% de la varianza. Luego, las proporciones de las variables individualmente quedan por debajo del 10%.

Resultados de las componentes principales obtenidas

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Standard deviation	2.8963	2.2125	1.64217	1.35843	1.21711	1.12930	1.0084	0.9944	0.93021
Proportion of Variance	0.2706	0.1579	0.08699	0.05953	0.04779	0.04114	0.0328	0.0319	0.02791
Cumulative Proportion	0.2706	0.4285	0.51550	0.57503	0.62281	0.66395	0.6967	0.7287	0.75656
	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	
Standard deviation	0.88267	0.86486	0.84487	0.82299	0.76875	0.73854	0.71578	0.67318	
Proportion of Variance	0.02513	0.02413	0.02303	0.02185	0.01906	0.01759	0.01653	0.01462	
Cumulative Proportion	0.78169	0.80582	0.82885	0.85070	0.86976	0.88736	0.90388	0.91850	
	PC18	PC19	PC20	PC21	PC22	PC23	PC24	PC25	
Standard deviation	0.65722	0.63307	0.61580	0.60078	0.58266	0.51848	0.49516	0.28470	
Proportion of Variance	0.01393	0.01293	0.01223	0.01164	0.01095	0.00867	0.00791	0.00261	
Cumulative Proportion	0.93244	0.94536	0.95760	0.96924	0.98019	0.98886	0.99677	0.99939	
	PC26	PC27	PC28	PC29	PC30	PC31			
Standard deviation	0.08990	0.08797	0.05594	0.006781	0.003686	0.0000000000000009923			
Proportion of Variance	0.00026	0.00025	0.00010	0.000000	0.000000	0.0000000000000000000			
Cumulative Proportion	0.99965	0.99990	1.00000	1.000000	1.000000	1.0000000000000000000			

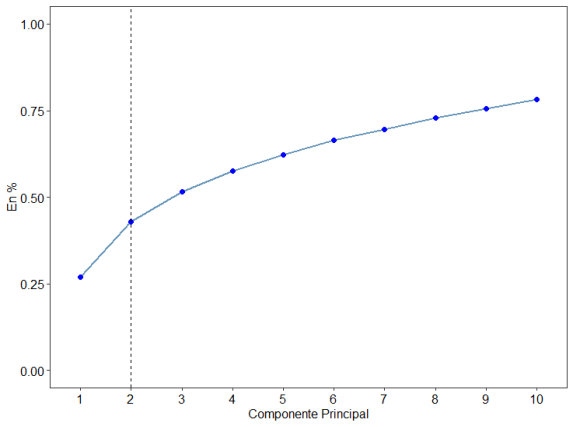
Fuente: elaboración propia en R.

Varianza explicada - Primeras 5 PC



Fuente: elaboración propia en R.

Varianza explicada acumulada – Top 10 PC



Fuente: elaboración propia en R.

4.3. Resultados

En este apartado se exponen los resultados del clustering. Una vez definido el número óptimo de clusters, se representan de forma gráfica aplicando la técnica de PCA y en un mapa de la Ciudad de Buenos Aires según radio censal. De esta forma puede verse fácilmente la distribución de los clusters desde diferentes visiones. Por último, se exponen las tablas con los resultados promedio para cada variable en cada cluster y se procede a caracterizarlos.

4.3.1 Clusters resultantes

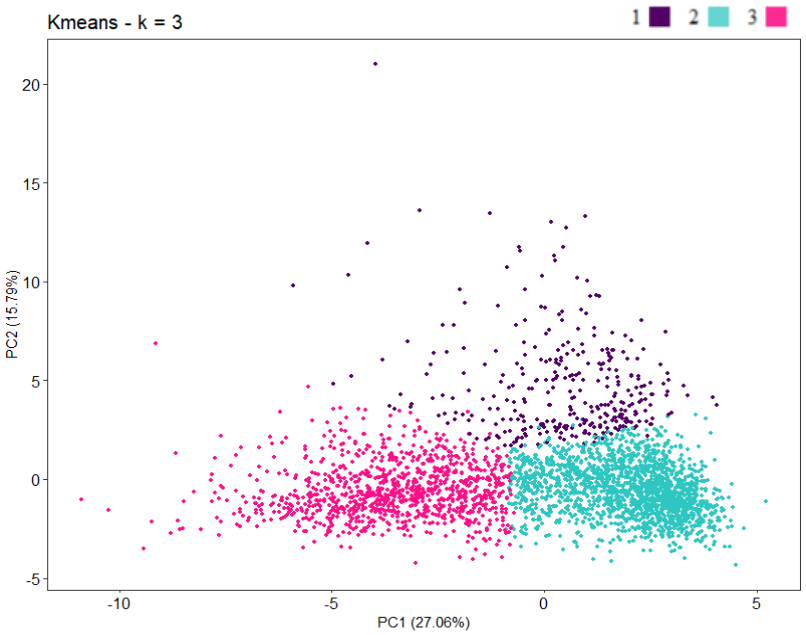
A partir de la detección de las dos primeras componentes principales, se graficó la salida del k-means. También se exponen en forma de tabla la distribución de los radios censales en los clusters resultantes y en un mapa de la Ciudad. El número de clusters implementado es de tres.

Cantidad de radios censales según cluster

CLUSTER	RADIOS CENSALES
1	334
2	1.803
3	1.032

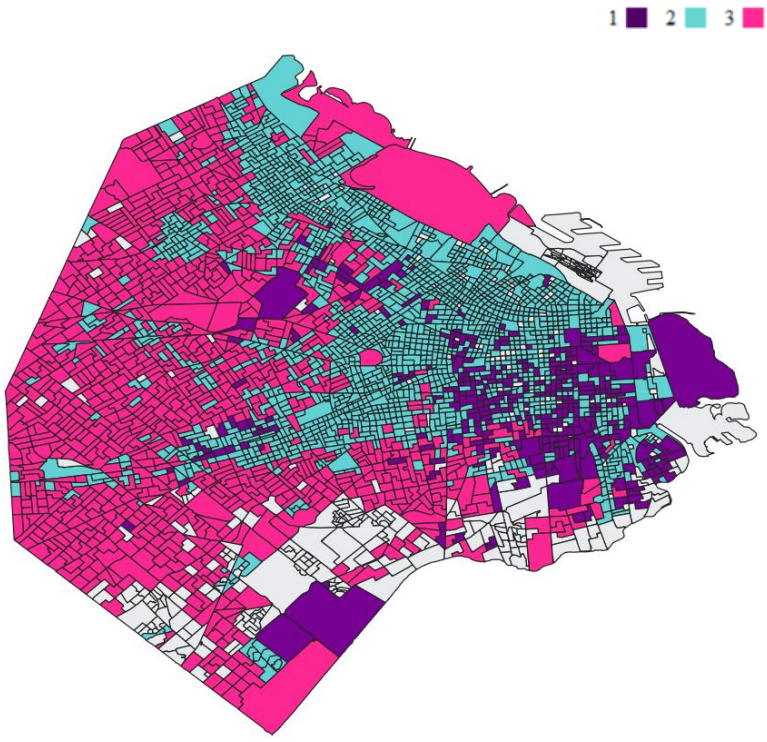
Fuente: elaboración propia en R.

Gráfico de clusters resultantes



Fuente: elaboración propia en R.

Mapa de los radios censales de la Ciudad según número de cluster



Fuente: elaboración propia en QGIS.

A partir de lo visualizado en el gráfico de coordenadas, puede observarse a simple vista que hay dos grupos concentrados – el grupo 2 más que el 3 – pero bien delimitados entre ellos y un tercer cluster bastante más disperso. En cuanto a la distribución de los radios centrales, el cluster 1 mencionado como disperso es el que concentra la menor cantidad, siendo de 334. Los cluster concentrados 2 y 3, tienen 1.803 y 1.032 radios cenales, respectivamente. A nivel radio censal en el mapa de la Ciudad, se observa una concentración del cluster 3 principalmente en la zona oeste, mientras que el cluster 2 se encuentra en el centro/norte y el cluster 1 en el sector este/sur.

4.3.2 Tablas para análisis

Para complementar el análisis, se calcularon los valores promedio para cada variable según cluster. En la escala del naranja se colorearon las que caracterizan o diferencian a cada cluster con respecto a los demás. El color es más intenso para valores más altos y viceversa. Estas tablas facilitan la caracterización de los clusters al poder visualizar fácilmente los valores resultantes del agrupamiento.

a) Socio-económicas

cluster	personas	persona sexo varon	persona sexo mujer	viviendas	hogares con nbi
C1	740	47%	52%	370	23%
C2	782	45%	55%	452	3%
C3	847	47%	53%	358	3%

cluster	hogares sin nbi	calidad constr. vivienda satisfactoria	calidad constr. vivienda basico	calidad constr. vivienda insuficiente	vivienda particular departamento
C1	77%	75%	13%	11%	69%
C2	97%	91%	8%	1%	88%
C3	97%	87%	12%	1%	45%

cluster	Vivienda particular casa	condición actividad ocupado	condición actividad desocupado	condición actividad inactivo	ocupación vivienda
C1	16%	72%	4%	24%	76%
C2	11%	70%	3%	27%	74%
C3	54%	67%	3%	30%	82%

cluster	Rango etario 0 - 30	Rango etario 31 - 60	Rango etario 61 - 110
C1	43%	40%	17%
C2	37%	40%	23%
C3	39%	40%	21%

b) Solicitudes y Delitos

cluster	cestos y contenedores	extraccion de arbol	limpieza de via publica	mayor iluminacion en calle	Poda de arbol y despeje de luminaria
C1	9%	0%	19%	4%	11%
C2	9%	3%	19%	3%	13%
C3	13%	5%	17%	3%	15%

cluster	problema con intervenciones de arbolado	reparacion de baches	reparacion de luminaria	reparacion de vereda	vehiculos abandonados
C1	2%	6%	14%	24%	2%
C2	4%	5%	10%	25%	5%
C3	6%	4%	8%	20%	3%

cluster	cantidad solicitudes	cantidad delitos	ratio solicitudes	ratio delitos	cantidad radios censales
C1	13.826	13.676	41	41	334
C2	67.662	47.111	38	26	1.803
C3	114.991	29.508	111	29	1.032

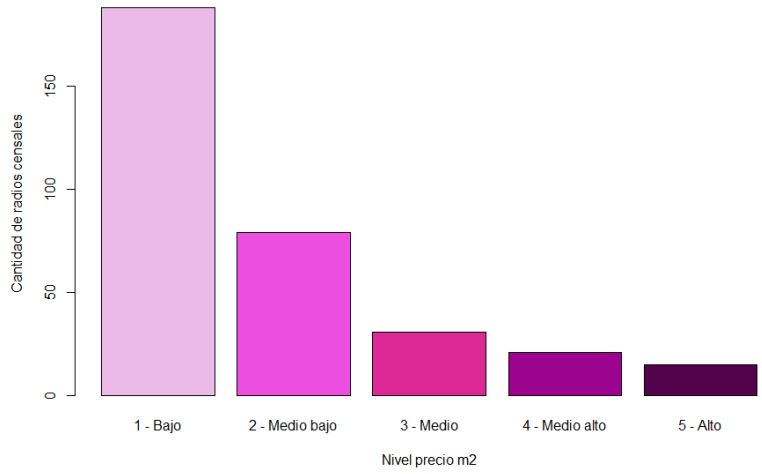
4.3.3 Caracterización de los clusters

Se procede a describir los distintos grupos producto del clustering a partir de las variables que constituyeron la tabla input del modelo.

a) Cluster 1

Con 334 radios censales agrupados es el cluster más pequeño. Este grupo es, en promedio, el cluster más joven destacándose por sobre los otros grupos en el rango etario de 0-30 y, a su vez, siendo el que menor proporción tiene en el rango 61-110. Una característica sobresaliente es que se trata del cluster donde hay mayor proporción de hogares con necesidades básicas insatisfechas, así como con calidad de la vivienda insuficiente. Parte de estas características puede verse reflejada en la distribución de los radios censales del grupo según su nivel precio m2.

Radios censales por nivel precio m2



Fuente: elaboración propia en R

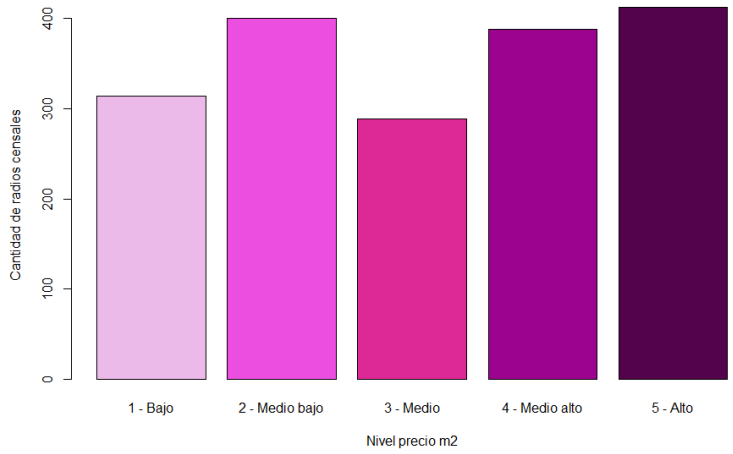
En cuanto a las solicitudes, este cluster tiene tendencia a realizar reclamos referidos a la iluminación. Como puede observarse, el 14% de los reclamos están orientados a la solicitud de reparación de luminaria (muy por encima de los otros clusters que tienen el 10% y 8% de sus solicitudes en esta subcategoría). Además, también es el cluster que mayor proporción tiene de solicitudes sobre mayor iluminación en calle, si bien en este caso la diferencia con los demás grupos es menor.

Es interesante notar que el cluster 1 es el que mayor ratio de denuncias de delito presenta. El ratio de denuncias hace referencia a la cantidad de delitos sobre la cantidad de radios censales del grupo. Es lógico pensar que es más probable que un delito ocurra en condiciones propicias como calles poco transitadas o poco iluminadas, por lo que se puede suponer que las solicitudes en mejoras de iluminación se ven condicionadas, en parte, por la elevada cantidad de denuncias de delito. Este punto de análisis excede el alcance del presente trabajo, pero queda abierto a futuras investigaciones complementarias que ahonden sobre esta suposición.

b) Cluster 2

Con 1.803 radios censales agrupados es el cluster más grande. El cluster 2, anticipando el resto del análisis, es el único cluster cuyo nivel predominante en el nivel precio m2 no es el "bajo". Por el contrario, se destacan los niveles medios y altos y casi el 90% de las propiedades son tipo departamentos siendo un rasgo claramente distintivo de esta clase. En línea con el gráfico, en este grupo solo un 3% de los hogares tiene alguna necesidad básica insatisfecha y el 91% de la calidad constructiva de la vivienda es satisfactoria, siendo la más alta entre los tres clusters.

Radios censales por nivel precio m2



Fuente: elaboración propia en R

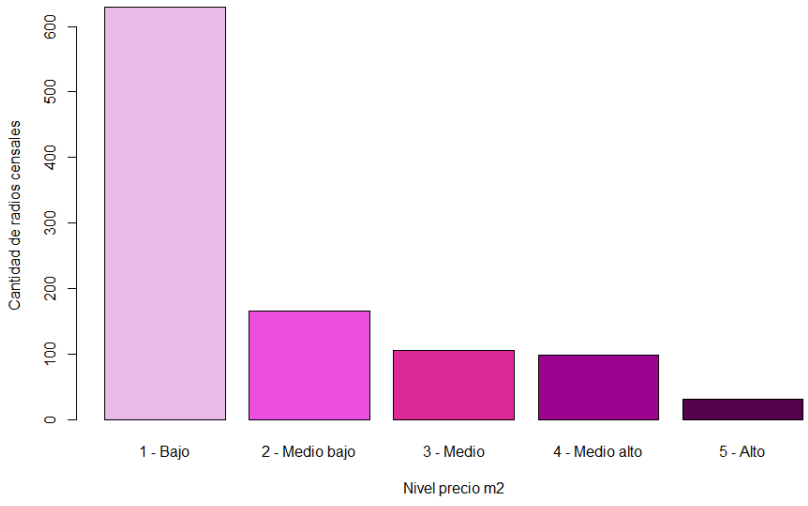
Este grupo tiene la mayor proporción de viviendas, pero, asimismo, la menor ocupación en promedio. Podría decirse que, en comparación, tiende a ser un cluster de

gente adulta siendo más fuerte en el rango 61-110 y menor en 0- 30. En cuanto a las solicitudes y delitos, tiene el ratio más bajo en ambos casos por lo que podría concluirse que es un cluster menos demandante que los demás. De caracterizarse en un tipo de reclamo sería en los orientados al estado de veredas y calles destacándose los pedidos de retiro de vehículos abandonados y los de reparación de la vereda.

c) Cluster 3

Con 1.032 radios censales agrupados es el cluster de tamaño mediano. En el cluster 3, la distribución de los radios censales con respecto al nivel precio m2 es bastante similar a la del cluster 1, pero disminuye el nivel medio-bajo y sube el medio-alto. A diferencia de los clusters anteriores, más del 50% del tipo de propiedad corresponde a casas mientras que en las otras agrupaciones este porcentaje no supera el 20%. La ocupación de las viviendas es la más alta a su vez que es el grupo con mayor cantidad de población promedio. Por otra parte, las necesidades básicas de los hogares están bien cubiertas y la calidad constructiva de la vivienda es buena. Aun así, es el cluster con mayor porcentaje de población inactiva. Se podría pensar que estos radios censales representan los típicos barrios tranquilos de construcciones bajas que no están dentro de las zonas con altos valores del m2.

Radios censales por nivel precio m2



Fuente: elaboración propia en R

El cluster 3 es claramente el que tiene la mayor tasa de reclamos (cantidad de reclamos sobre la cantidad de radios censales) siendo de 111 frente a los 38 y 42 de los otros grupos. Así como en el cluster 1 se concluyó que el mayor problema era con la iluminación y en el cluster 2 con el estado de calles y veredas, en este cluster los problemas se orientan por el lado del arbolado. Este grupo se destaca por sobre los demás en: extracción de árboles, poda de árbol y despeje de luminaria y problemas con intervención de arbolado. Sería interesante profundizar el análisis y ver si los radios censales de esta clase agrupan la mayor cantidad de espacios verdes, los espacios verdes de mayor superficie o bien si tiene plantaciones de un tipo particular de árbol que trae dificultades. Por último, también se destaca en lo que es reclamos sobre cestos y contenedores.

5. Conclusiones

A lo largo del trabajo se ha conocido a la organización mediante su descripción formal y su relación con el Gobierno Abierto. Como se ha comentado, del mismo se desprenden los Datos Abiertos que han servido para nutrir la base elaborada para realizar este estudio. Se ha expuesto la importancia de esta filosofía política tanto en cuestiones de transparencia como de generación de conocimiento o rédito económico y se describió cómo trabaja el equipo a cargo de mantener el Portal de Datos Abiertos de la Ciudad.

También, se comentó cómo los datos del censo fueron vinculados con datos sobre solicitudes de diferente naturaleza y de denuncias de delitos, junto con el indicador de nivel precio m2 elaborado a partir de datos abiertos de Properati. De esta forma, se complementó y extendió el alcance de las variables representativas de las condiciones socio-económicas utilizadas con origen en el censo 2010. Asimismo, se expuso la metodología aplicada, describiendo las distintas fuentes de datos utilizadas y su limpieza asociada para generar la tabla input que alimentó a la técnica de aprendizaje automático no supervisada utilizada, clustering, mediante el algoritmo k-means. Para su implementación, se compararon las métricas de WCSS y silhouette y se concluyó que el número óptimo de clusters era de tres por ser donde el silhouette alcanzó su mayor valor. El método del codo no presentó resultados determinantes.

Con los resultados provenientes de aplicar el algoritmo, se ha realizado una caracterización de los grupos que han sido resultado de la investigación a partir de las variables consideradas. El resultado del k-means se exhibió en un eje de coordenadas a partir de obtener las dos primeras componentes principales (acumulando un 43% de la varianza) mediante la aplicación de PCA al dataset para posibilitar su visualización y, también, en una mapa de la Ciudad. En resumen, y a modo de puntualizar lo relevante de los conglomerados resultantes, se obtuvo que:

- El cluster 1 es el grupo con mayores carencias estructurales, siendo los niveles por precio m2 bajos proporcionalmente relevantes. En cuanto a las solicitudes, se observa un predominio de los reclamos en luminaria y se planteó una posible relación con la elevada tasa de denuncias de delitos que presenta este grupo.
- El cluster 2 es el grupo con las mejores condiciones de la vivienda en cuanto a necesidades básicas satisfechas y calidad constructiva. Es el único cluster donde es fuerte la influencia de los niveles de m2 más altos. A su vez, tiene el ratio de delitos y solicitudes más bajo y la mayor cantidad de reclamos tienen origen en el estado de veredas y calles.
- El cluster 3 es el grupo con mayor cantidad de solicitudes y se lo determinó como el más demandante. Si bien el nivel del precio de m2 es predominantemente bajo la calidad de la vivienda es buena y las necesidades básicas están satisfechas, pero se resaltó que tiene el mayor porcentaje de población inactiva. En cuanto a las solicitudes, se destacan los reclamos sobre arbolado y se planteó una posible relación con los espacios verdes.

En conclusión, dentro del alcance de este trabajo, se puede advertir que existe cierta relación entre la naturaleza de las solicitudes realizadas y las características socio-económicas estudiadas de los radios censales. Por ende, existe evidencia suficiente para aceptar la hipótesis de la investigación. Por otra parte, se dejan abiertos nuevos interrogantes como puntos de partida de futuras investigaciones que sirvan como una segunda fase de análisis más exhaustiva del presente trabajo y que abren nuevas posibilidades de validación, o no, de la hipótesis planteada.

En línea con lo anterior, también podrían generarse estudios complementarios a partir de la implementación de otras técnicas de aprendiza automático como, por ejemplo, proyección de cantidad de solicitudes sobre algún tipo de reclamo considerado como relevante. En el cluster 3 son predominantes los reclamos sobre arbolado. En el caso de la poda, intuitivamente podría afirmarse que hay cierto componente estacionario relacionado con el rebrote. De esta forma, podría proyectarse la demanda y disponerse de más o menos recursos según la época del año. Esto indudablemente mejoraría la planificación y permitiría una mejor designación de los recursos del Estado.

De tal manera, y aludiendo a una de las problematizaciones planteadas, queda expuesto que no necesariamente se requieren de implementaciones complejas o costosas ni de contar con una infraestructura imponente. Si serán necesarios en entonos de Big Data, como pueden ser cuando se consumen datos de sensores o datos de la tarjeta SUBE. Aun así, es amplio el horizonte de posibles investigaciones que posibilitan los Datos Abiertos y que no requieren de entornos grandilocuentes, sino más bien de hacer interactuar datos de distinta naturaleza y origen, pero orientados hacia el mismo propósito.

Es amplio y diverso el catálogo de datos que ofrece el Portal de Datos Abiertos de la Ciudad de Buenos Aires. También, son cada vez más los organismos públicos y privados que entienden de los beneficios e importancia de poner los datos a disposición de la ciudadanía. Esta importancia no radica únicamente en cuestiones de transparencia, que es siempre fundamental sobre todo en el ámbito público, sino también como nueva forma de crear conocimiento y valor en conjunto con la sociedad.

6. Referencias bibliográficas

AFIT Data Science Lab R Programming Guide. (s.f.). Obtenido de https://afit-r.github.io/kmeans_clustering

Agenda de Transparencia e Innovación Institucional. (s.f.). Obtenido de https://www.buenosaires.gob.ar/agendadetransparencia

Aglú, E. (2017). *Universidad de Buenos Aires, Centro de Cooperativas y Economía.* Obtenido de http://www.economicas.uba.ar/wp-content/uploads/2017/09/Situacio%CC%81n-espacial-de-las-desigualdades-socioecono%CC%81micas-en-la-CABA.pdf

Bron, M. (2015). *Open Data: Miradas y Perspectivas de los Datos Abiertos.* La Rioja: Universidad Nacional de La Rioja.

Buenos Aires Ciudad. (s.f.). Organigrama. Recuperado el 27 de 11 de 2020, de https://www.buenosaires.gob.ar/organigrama/

Buenos Aires Data. (s.f.). *Arbolado público lineal*. Obtenido de https://data.buenosaires.gob.ar/dataset/arbolado-publico-lineal

Buenos Aires Data. (s.f.). *Delitos*. Obtenido de https://data.buenosaires.gob.ar/dataset/delitos

Buenos Aires Data. (s.f.). *Sistema Único de Atención Ciudadana*. Obtenido de https://data.buenosaires.gob.ar/dataset/sistema-unico-atencion-ciudadana

Calderón, C., & Lorenzo, S. (2010). *Open Government: Gobierno Abierto.* Alcalá la Real: Algón Editores.

Compromisos de Gobierno. (s.f.). Obtenido de https://www.buenosaires.gob.ar/jefedegobierno/secretariageneral/compromisos-de-gobierno

Estrategia de Apertura de Datos. (s.f.). Obtenido de https://datosgcba.github.io/guia-datos/politica-datos-abiertos/

Fachelli, S., Goicoechea, M., & López-Roldán, P. (Enero de 2015). Trazando el mapa social de Buenos Aires: Dos décadas de cambios en la ciudad. *Población de Buenos Aires, 12*(21), 7-39.

González, Á., Llinás Solano, H., & Tilano, J. (2008). Análisis multivariado aplicando componentes principales al caso de los desplazados. *Ingeniería y Desarrollo*, 121-132. Obtenido de https://www.redalyc.org/articulo.oa?id=85202310

INDEC Argentina. (2017). Encuentro sobre big data y estadísticas oficiales (10). Presentación de Walter Sosa Escudero [video]. YouTube. Obtenido de https://www.youtube.com/watch?v=fSxU9k8skZs

INDEC Argentina. (2017). Encuentro sobre big data y estadísticas oficiales (11). Presentación de Walter Sosa Escudero y Diego Bendersky. [video]. YouTube. Obtenido de https://www.youtube.com/watch?v=BzyIZGN6wH4&t=30

Insituto Nacional de Estadísticas y Censos. (s.f.). *Censo 2010*. Obtenido de https://www.indec.gob.ar/indec/web/Institucional-Indec-BasesDeDatos-6

Instituto Nacional de Estadísticas y Censos. (s.f.). *Radios censales*. Obtenido de https://www.indec.gob.ar/indec/web/Institucional-Indec-Codgeo

Kolanovic, M., & Krishnamachari, R. (2017). *Big Data and AI Strategies Machine Learning and Alternative Data Approach to Investing.* New York: JP Morgan.

López, L. S. (2005). *Red de Bibliotecas Virtuales de Ciencias Sociales de la red CLACSO.* Obtenido de http://biblioteca.clacso.edu.ar/ar/libros/becas/2005/demojov/lopez.pdf

Manzano, J., & Ezequiel Uriel, J. (2017). *Análisis multivariante aplicado con R* (2ª ed.). Madrid: Paraninfo.

McAfee, A., & Brynjolfsson, E. (1012). Big Data: The Management Revolution. *Harvard Business Review*.

Pla, I. L., & Bojórquez Pereznieto, J. A. (2015). *Gobierno Abierto y el valor social de la información pública.* (A. Hofmann, Ed.) Ciudad de México. Recuperado el 8 de junio de 2020, de https://biblio.juridicas.unam.mx/bjv/detalle-libro/4016-gobierno-abierto-y-el-valor-de-la-informacion-publica

Portales de datos abiertos. (8 de junio de 2020). Obtenido de https://www.argentina.gob.ar/aaip/accesoinformacion/datospublicos

Properati Data. (s.f.). *BigQuery*. Obtenido de properati-dw.public.ads

Ramírez-Alujas, Á. (2011). Gobierno Abierto y Modernización de la Gestión Pública. Tendencias Actuales y el (inevitable) camino que viene. Reflexiones Seminales. *Revista Enfoques*, 99-125.

Salgado, D. (2016). Big Data y la Estadística Oficial: retos. *Índice*, 14-17.

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2018). *Introduction to Data Mining* (Segunda ed.). Pearson.

Villoria, M., & Ramírez-Alujas, Á. (2011). La transparencia: marco conceptual. *Revista Democracia y Gobierno Local*, 10.