Normalización y análisis exploratorio de datos inmobiliarios web

Del Río, Juan Pablo (CONICET/LINTA-CIC/FaHCE-UNLP) geodelry@gmail.com
Dioguardi, Felipe (LIFIA-UNLP) fdioguardi@lifia.info.unlp.edu.ar
May, Marcos (LINTA-CIC/FaHCE-UNLP) marcosmay@gmail.com
Torres, Diego (LIFIA-UNLP) dtorres@lifia.info.unlp.edu.ar

Resumen

Esta ponencia se enmarca en el Proyecto "Observatorio de valores del suelo e instrumentos de financiamiento del desarrollo urbano" (Programa Impactar - MinCyT). Con el propósito de generar líneas de base para el cálculo de plusvalías urbanas, el objetivo del proyecto es generar información sobre el comportamiento del mercado inmobiliario urbano en la Provincia de Buenos Aires. Para ello se analizan los anuncios de venta de inmuebles en los principales portales inmobiliarios, obtenidos a través de técnicas de web scraping con Python.

El documento presenta una identificación de los problemas que poseen estas fuentes de información, así como el proceso de curaduría y normalización de la base de datos generada a partir del scrapeo de los portales inmobiliarios y del análisis exploratorio de las principales variables obtenidas. Para ello se utiliza como entorno de trabajo el lenguaje de programación R y la herramienta de Qgis. Por último, se exponen las estrategias metodológicas en análisis para la puesta en valor de dichas matrices de datos.

Introducción

El acceso a la información del mercado inmobiliario y de suelo urbano es una tema crítico, dado que la opacidad es un rasgo característico de estos mercados imperfectos (Morales, 2005; Baer, 2013; Smolka y Mullahy, 2010). Sin embargo, la visibilidad que la oferta inmobiliaria ha adquirido en la web (Hernández, 2019) reconfiguran profundamente las posibilidades de sistematizar información sectorial. Esto supone una oportunidad, ya que contar con información pública de calidad se torna estratégico, tanto para conocer la dinámica de funcionamiento de estos mercados como para promover la regulación desde el sector público (Reese, 2006; Smolka y Mullahy, 2010).

En este marco, el monitoreo periódico de los valores inmobiliarios es una cuestión de agenda para los gobiernos locales y provinciales. En especial cuando se registra (Carranza, 2022 *et*

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

al.) que la incapacidad institucional de actualizar la valuaciones impacta en la magnitud y composición de los ingresos públicos, en la equidad del trato a los contribuyentes, en la condiciones de acceso al hábitat, en la capacidad de financiar la infraestructura pública y el desarrollo urbano.

En esta línea, la ponencia forma parte del Proyecto "Observatorio de valores del suelo e instrumentos de financiamiento del desarrollo urbano" (Desafío Nro 16 del Programa Impa.CT.AR - MinCyT). El mismo se enfoca en el problema de la carencia estructural de datos públicos respecto a los valores de mercado de los inmuebles urbanos y la desactualización que presentan las valuaciones fiscales. El objetivo general del proyecto es fortalecer las capacidades estatales de la provincia de Buenos Aires en materia de generación de información pública y cuantificación de la valorización inmobiliaria, con el propósito de contribuir al financiamiento urbano mediante instrumentos de recuperación de plusvalías urbanas[1] y disponibilizar datos abiertos en la materia.

En la actualidad la mayor parte del flujo de oferta inmobiliaria circula por plataformas web (como por ejemplo: Properati, Inmobusqueda, Zonaprop, MercadoLibre Inmuebles, Argenprop, etc.) especializadas en la difusión de avisos clasificados georreferenciados y con caducidad programada, sean estos gratuitos o pagos. En tal sentido, el análisis de las condiciones de publicación de ofertas, así como el volumen de anuncios y cobertura espacial de los avisos forman parte de las tareas realizadas en el proyecto, para priorizar la selección de las plataformas inmobiliarias.

En esta línea y con el propósito de generar información de base sobre el comportamiento del mercado inmobiliario urbano en los distritos bonaerenses de la Región Metropolitana de Buenos Aires[2] (en adelante RMBA), el presente trabajo analiza los anuncios de venta de inmuebles en los principales portales inmobiliarios con cobertura en el área de estudio.

[1] Las plusvalías urbanas son incrementos que acumula el valor de los inmuebles por causa externas a las acciones de los propietarios, es decir se vincula con la valorización que surge del esfuerzo comunitario, la inversión pública y las decisiones regulatorias vinculadas a los usos del suelo (Smolka y Amborski, 2003; Cuenya *et al*. 2009).

[2] La región metropolitana incluye los siguiente partidos de la provincia de Buenos Aires: Almirante Brown, Avellaneda, Berazategui, Berisso, Brandsen, Campana, Cañuelas, Ensenada, Escobar, Esteban Echeverría, Exaltación de la Cruz, Ezeiza, Florencio Varela, General Las Heras, General Rodríguez, General San Martín, Hurlingham, Ituzaingó, José C. Paz, La Matanza, La Plata, Lanús, Luján, Lomas de Zamora, Malvinas Argentinas, Marcos Paz, Merlo, Moreno, Morón, Pilar, Presidente Perón, Quilmes, San Fernando, San Isidro, San Miguel, San Vicente, Tigre, Tres de Febrero, Vicente López y Zárate.

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

El documento da cuenta de cómo los avisos fueron obtenidos a través de técnicas de *web scraping* con Python y presenta una identificación de los problemas que poseen estas fuentes de información, así como el proceso de curaduría y normalización de la base de datos generada a partir del scrapeo de los portales inmobiliarios y del análisis exploratorio de las principales variables obtenidas. Para ello se utiliza como entorno de trabajo el lenguaje de programación R y la herramienta de Qgis. Por último, se exponen las estrategias metodológicas en análisis para la puesta en valor de dichas matrices de datos.

Consideraciones metodológicas

Para analizar la información de los avisos inmobiliarios, se debe contar con una base de datos que sea representativa del panorama del mercado inmobiliario en la región. Esto hace que sea necesario pensar en técnicas de automatización que permitan recolectar grandes cantidades de datos en cortos períodos de tiempo, para evitar que el conocimiento se vuelva obsoleto.

El *web scraping*, también conocido como *web extraction* o *harvesting*, es una técnica para extraer datos de la World Wide Web (WWW) y guardarlos en un sistema de archivos o en una base de datos para su posterior recuperación o análisis. Por lo general, los datos de la web se extraen utilizando el protocolo HTTP o a través de un navegador web. Esto puede ser conseguido manualmente por un usuario o automáticamente por un bot o *web crawler* (Bo Zhao, 2017). En esa línea, un *web crawler* o *spider* es un programa creado para la descarga masiva de páginas web (Olston y Najork, 2010). Teniendo estas definiciones en cuenta y las especificidades y características de cada plataforma, crear una araña para la recolección de datos en cada una de ellas resultó la solución más natural al problema planteado. De esta forma, se construiría un programa que se comporte primero como un *crawler*, descubriendo links de interés a partir de un conjunto pequeño de enlaces URL; y luego como un *scraper*, aplicando las técnicas de *web scraping* en cada una de las páginas descubiertas, extrayendo la información perteneciente a cada una.

En general, todas las arañas funcionan de la misma forma (Schrenk, 2012):

1. Descargan la página objetivo inicial, conocida como *URL semilla*.
2. Buscan en esa página nuevos links a otras páginas de interés.
3. Si se determina que se llegó al nivel de profundidad deseado, finaliza; en caso contrario, descarga las nuevas páginas y vuelve al paso 2.

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

Diseñar arañas para la recolección de avisos inmobiliarios implica replicar este comportamiento para cada sitio específico. Entonces, un *crawler* para una plataforma determinada deberá (1) acceder a la primera página del listado de avisos inmobiliarios, (2) descargarla para conseguir su información, junto con los links a los avisos visibles en esa página del listado, (3) obtener la información respectiva a cada uno de los avisos, y (4) acceder a la siguiente página del listado para volver al paso 2.

Para implementar el *scraper* se utilizó el lenguaje Python, por su amplio catálogo de librerías de fácil uso y acceso, que permiten elaborar programas extensibles y funcionales en todo tipo de ambientes. El código de la herramienta se desarrolló sobre el framework Scrapy[3], un marco de trabajo que ofrece una estructura programática popular y distribuida para la producción de *crawlers* y *scrapers*. A propósito de este esquema, se elaboraron 3 arañas preparadas para recorrer cada una de las plataformas seleccionadas.

El primer paso en la confección de los *crawlers* es limitar el universo de anuncios publicados por cada plataforma tanto por su alcance geográfico (en este caso la RMBA) como por el tipo de operación (compra-venta de inmuebles). Cada uno de los *crawlers* producidos tiene un conjunto de URL semilla pertenecientes al mismo dominio, pero que dirigen a un listado distinto dependiendo del filtro aplicado. Esto hace que quitar un partido del conjunto de búsqueda sea tan fácil como remover el URL correspondiente de la lista de URL semilla de cada *spider*. Asimismo, realizar una búsqueda sobre un listado filtrado será posible haciendo el mismo movimiento, siempre y cuando las características del filtro estén reflejadas en el URL.

A la hora de recuperar los avisos inmobiliarios, fue necesario hacer uso de diversas técnicas para evitar el rechazo de los servidores. En primer lugar, se designó un tiempo de espera ajustado a cada sitio web, que determina la cantidad mínima de segundos que deben pasar entre descargas de clasificados de las plataformas. De esta manera se evita sobrecargar los sitios consultados y eludir algunos de los protocolos de protección contra bots que pudieran implementar.

Dentro de las tecnologías aplicadas para extraer la información de cada oferta inmobiliaria se destacan los lenguajes XPath y JavaScript. El primero es un lenguaje que se compone de expresiones, que permite representar la ruta de un atributo de interés en el código HTML de

[3] https://scrapy.org/

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

una página. La mayoría de los valores rescatados de cada aviso fueron extraídos a través de la identificación de su ruta XPath y posterior interpretación mediante la librería lxml[4]. En cuanto a Javascript, este conocido lenguaje de programación es el más empleado por los distintos sitios web para cargar contenido dinámicamente, lo que hace que no siempre se encuentre presente en la estructura HTML. Para acceder a esta información se localizaron las etiquetas <script> que almacenan el código JavaScript pertinente, para así evaluarlo con la librería Js2Py[5], simulando lo que hubiera sucedido al acceder desde un navegador web corriente. También se pudo aprovechar que uno de los sitios proporciona una API pública que permite obtener las ofertas inmobiliarias en formato JSON, facilitando la extracción de sus datos.

Variables	
address	latitude
advertiser_id	listing_antiquity
advertiser_name	listing_id
antiquity	longitude
bath_amnt	maintenance_fee
bed_amnt	maintenance_fee_currency
covered_surface	neighborhood
covered_surface_unit	price
currency	property_type
date_extracted	province
description	reconstructed_total_surface
	reconstructed_total_surface_unit
district	room_amnt
features	title
garage_amnt	toilete_amnt
is_finished_property	total_surface
is_new_property	total_surface_unit
is_studio_apartment	uncovered_surface
land_surface	uncovered_surface_unit
land_surface_unit	url

Fig. 1 Variables de la base inmo scrap

La herramienta desarrollada trae consigo resultados prometedores. Su ejecución en un entorno paralelo dispara la acción de todas las arañas, permitiendo recolectar en las tres plataformas más de 600 mil avisos y 38 variables en menos de una semana en la RMBA, aumentando el tamaño del corpus de datos en más de un 1.000% en relación a los esfuerzos manuales previos.

De los 623.845 avisos relevados, la Plataforma 1 aportó el 38% de los registros, la plataforma 2 el 32% y la plataforma 3 el 30% restante. Asimismo, se destaca que el 41% de los anuncios corresponde a casas, el 28% a departamentos, el 16% a terrenos y el 6% a Ph[6]. En otras palabras, el destino residencial representa el 77% del flujo de oferta del área de estudio y en conjunto con la oferta de tierra alcanza el 93%.

[4] https://lxml.de/
[5] https://github.com/PiotrDabkowski/Js2Py
[6] La categoría de "Ph" es un recurso utilizado por las plataformas para segmentar los departamentos en planta baja o departamentos en lote subdividido por régimen de propiedad horizontal. Aunque los departamentos en altura jurídicamente también constituyen unidades funcionales asociadas al régimen de propiedad horizontal, la categoría nativa de "Ph" es utilizada para tipificar un producto inmobiliario que se diferencia de los edificios de vivienda multifamiliar en altura.

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

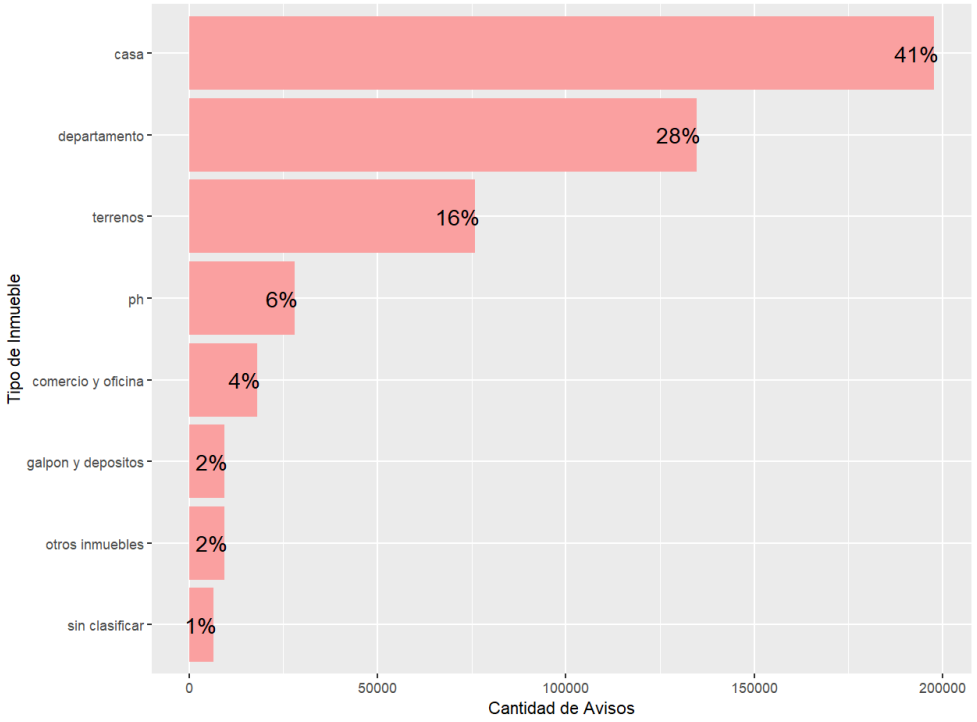

Fig. 2 .Cantidad de avisos según tipo de inmueble

La base inmo scrap, principales problemas encontrados

En este apartado se documentan las principales dificultades encontradas a nivel de la matriz de datos que se elaboró con el desarrollo del programa *inmo scrap,* así como también producto de las características de las fuentes de información consultadas.

En línea con la literatura especializada (Morales, 2005; Jaramillo, 2009; Abramo, 2011; Pérez Torres, 2015), la localización es una de las variables críticas para el análisis del mercado inmobiliario, ya que además de los atributos intrínsecos de los inmuebles la externalidades juegan un papel clave en la formación de los precios. De allí que la detección de los **problemas de calidad de ubicación** de la oferta fue una de las primeras cuestiones identificadas, registrándose grosso modo cuatro dificultades:

-Errores de carga de los datos por parte de los oferentes. Aspecto que se refleja rápidamente en la ubicación georreferenciada de oferta asociada a los partidos consultados de la RMBA por fuera de los límites jurisdiccionales de esta entidad territorial[7] o bien en la ubicación de ofertas perteneciente a un partido de la RMBA en otro partido de la región.

[7] A modo de ejemplo, se encontraron ofertas asociadas a partidos de la RMBA en el interior de la provincia de Buenos Aires, en otras provincias de la Argentina o en el exterior. Las cuales en conjunto representan el 6% del universo de datos scrapeados para los partidos del área de estudio.

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

-Limitaciones de la geocodificación automática de direcciones. Otro de los aspectos detectados se vincula con las restricciones que presentan los distintos servicios de *web mapping* que utilizan las plataformas inmobiliarias para automatizar la ubicación a partir de los parámetros de dirección (calle, altura, localidad o partido, provincia, país). Cuando *google maps* (o herramientas equivalentes) no cuentan con las alturas cargadas en una calle (o para la totalidad de los segmentos de un eje de calle), ubica el anuncio en posiciones equidistante de la calle o segmento de calle aunque no haya sido posible encontrar altura, sin que esto se notifique al oferente.

-Enmascaramiento de las ubicaciones por parte de las plataformas. Por razones comerciales o de seguridad, algunas plataformas ofrecen la posibilidad de distorsionar la ubicación del inmueble con el propósito de no publicar la ubicación exacta. Si bien, la ubicación se realiza en un radio de influencia aproximado a escala de la cuadra o manzana, cabe señalar que a efectos de ciertos análisis esto obliga a complementar la información con consulta directa al oferente.

-Entornos territoriales que restringen la precisión de la ubicación. En determinados contextos urbanos, suburbanos o periurbanos donde la estructura vial se encuentra menos estructurada, frecuentemente se registran calles sin nombres que dificultan la ubicación de los anuncios. Algo semejante sucede con sectores con rápido crecimiento o de expansión urbana. También, la oferta en áreas rurales sin referencias, o en rutas con kilometraje progresivo, suele ser difícil de validar al momento de controlar la ubicación. Los anuncios dentro de conjunto inmobiliarios mayores, como ser: barrios cerrados, countries, parques industriales, también presenta restricciones a efectos de contar referencias para validar la ubicación en su interior.

En relación a esta familia de dificultades la estrategia metodológica desplegada fue realizar un control manual de los datos georreferenciados en las plataformas, distinguiendo tres categorías de ubicación: aceptable (con tolerancia de una cuadra, 100-150 metros), rectificable (anuncios ubicados a más de una cuadra, pero con información complementaria que permite mejorar la ubicación) y sin posibilidad de validación (se trata de anuncios que no brindan información suficiente para controlar la ubicación o se encuentran en un contexto territorial que inhabilita la verificación y requerirían consultar directamente a los oferentes para ajustar el dato). En esta línea de trabajo, los controles realizados sobre un subuniverso de 41.529 anuncios presentan variaciones por municipio, pero en promedio arrojaron que el 76% de ubicaciones son aceptables, pudiéndose rectificar mediante trabajo manual un 19% de las

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

observaciones, perdiéndose el 5% restante por no contarse con información suficiente para validar los datos.

En paralelo al problema de ubicación, se identificó una dificultad concurrente asociada al conjunto de entidades territoriales o **topónimos utilizados por las plataformas** para clasificar la oferta, pudiendo esta dificultad objetivarse en problemas de ubicación real del aviso o no[8]. Es decir la forma en la cual las plataformas tipifican geográficamente los avisos es heterogénea y sólo la plataforma 2 contempla la jurisdicción oficial de partido. A su vez, debe tenerse en cuenta que los topónimos o nomencladores de ciudad o barrio varían entre las plataformas. En la figura 3, se observan las variables de ubicación con las cuales las plataformas organizan los avisos, no siendo obligatorio el completamiento de todas las variables.

Plataforma 1	Plataforma 2	Plataforma 3
Provincia	Provincia	Provincia
	Partido	
Ciudad		Ciudad
	Localidad	
Barrio	Barrio	
Subzona	Subarrio	
Calle	Calle	Calle
Altura	Altura	Altura

Fig. 3. Variables de ubicación de aviso según plataforma inmobiliaria

Todos estos aspectos dificultan la normalización de la matriz de datos por distrito o partido, en tanto jurisdicción oficial. Frente a este obstáculo la estrategia metodológica adoptada fue, una vez georreferenciados los avisos por latitud y longitud, realizar la intersección espacial de las observaciones con el límite jurisdiciconal de los partidos y asignar el nombre oficial. De este modo, fue posible contar con una primera segmentación territorial de la matriz de datos con el objetivo de analizar la jerarquía de toponímia emergente, las imprecisiones que poseen los anuncios o las estrategias de marketing a la que responden cuando se distorsionan los nombres de los lugares.

[8] A modo de ejemplo, si en apariencia la oferta de un inmueble se encuentra georreferenciada dentro del perímetro del límite de partido de Brandsen, la dirección y barrio al que hace referencia el anuncio corresponde a calle y barrio sito en Brandsen, no se objetiva un problema de ubicación real aunque la carga de la localidad o partido cargado sea La Plata. En otras palabras, en este caso podría verificarse un problema de ubicación real cuando dentro del límite del partido de Brandsen se encuentra un aviso cuya descripción en términos de calle, barrio o partido corresponde a La Plata, pero las coordenadas de aviso se encuentran dentro de Brandsen.

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

Una tercera dificultad se vincula con los **precios erróneos** o las estrategias de los anunciantes de evitar publicar el precio, el cual resulta un campo obligatorio en todas las plataformas. En este sentido, se detectaron patrones recurrentes en el completamiento de precios considerados datos no válidos, como por ejemplo: "999999" o "111111". Para ello en programa *inmo scrap* incorporó una nueva variable (*price control*) orientada a la detección rápida de estos valores atípicos.

En paralelo, el análisis exploratorio de los valores por metro cuadrado para el mismo tipo de inmueble en contextos próximos, permitió observar problemas de completamiento de un dígito en el precio o en la relación precio/moneda. Es decir se registraron precios atípicos que cuando se les agrega un 0, o se traduce la moneda de pesos a dólares (o viceversa), inmediatamente se asemejan a los valores vecinos. Para este problema no se automatizó una solución, pero (una vez determinado el subuniverso de ofertas con ubicaciones aceptables) se registró la necesidad de aplicar técnicas de detección *outliers* espaciales, .

Un cuarto problema se vincula con la carga y estructuración de las **variables de superficie** en cierto tipo de inmuebles y plataformas. En el caso de las casas, la plataforma 1 y 3 presentan una configuración confusa de superficie: total y cubierta; no siendo claro a qué corresponde la superficie total para los oferentes. A partir de la interpretación de los avisos, se observó que éstos mayormente cargan en la variable superficie total la superficie del terreno.

Tipo de inmueble	Plataforma 1	Plataforma 2	Plataforma 3
Casa	Superficie total Superficie cubierta Frente Fondo	 Superficie cubierta Superficie descubierta Superficie terreno Frente Fondo	Superficie total Superficie cubierta
Departamento	Superficie total Superficie cubierta	Superficie total Superficie cubierta Superficie descubierta	Superficie total Superficie cubierta
Ph	Superficie total Superficie cubierta Frente Fondo	Superficie total Superficie cubierta Superficie descubierta	Superficie total Superficie cubierta
Terreno	Superficie total Superficie cubierta Frente Fondo	Superficie total Frente Fondo	Superficie total Superficie cubierta

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

Fig. 4. Variables asociadas a la superficie de casa, ph, departamento y terrenos

En este sentido, para comparar los datos de la plataforma 1 y 3 con la 2, para el tipo de inmueble casa se reconstruyó la superficie de terreno a partir de homologar la superficie del terreno a la superficie total cuando ésta era mayor a la superficie cubierta. Asimismo, cuando la plataforma disponía de las variables frente y fondo del terreno (y ésta había sido completada por los oferentes), se recurrió a las mismas para reconstruir la superficie del terreno. No obstante, resta analizar casos menos frecuentes donde la superficie cubierta podría ser mayor a la superficie del terreno, por ser terrenos construidos en dos plantas u otras situaciones particulares.

Otro de los problemas asociados a la plataforma 2, se vincula con la no obligatoriedad del completamiento de la superficie total en el caso de los departamentos y Ph. Para subsanar esta ausencia se reconstruyó la superficie total a partir de la suma de la superficie cubierta y descubierta. En sentido inverso en las plataformas 1 y 3, para reconstruir la superficie descubierta de los departamentos y Ph, se restó la superficie cubierta a la superficie total.

De esta forma, se avanzó en un primer nivel de normalización de las tres variables críticas vinculadas al mercado analizados: ubicación, precio y superficie.

Resultados preliminares en base al análisis de terrenos

Al realizar el análisis de la distribución del peso relativo del flujo de oferta de terrenos[9] por partido, se pone en evidencia que los distritos con mayor protagonismo son aquellos de la segunda o tercera corona metropolitana donde se está desarrollando un intenso proceso de expansión urbana, pero que no sólo se explica por el tamaño de los distritos en términos de superficie o población, sino que además es posible distinguir que la subdivisión del suelo se realiza principalmente bajo la modalidad de urbanizaciones cerradas orientadas a los sectores de mayor ingresos. De allí el protagonismo que adquieren partidos como Escobar, Pilar, Tigre, Ezeiza, San Vicente y Berazategui en la explicación del flujo de oferta global de terreno.

[9] Se consideraron para los análisis que se muestran a continuación terrenos de tamaño urbano de 150 a 2.500 m2 por ser normalmente la superficie mínima y máxima de la subdivisión de parcelas urbanas.

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

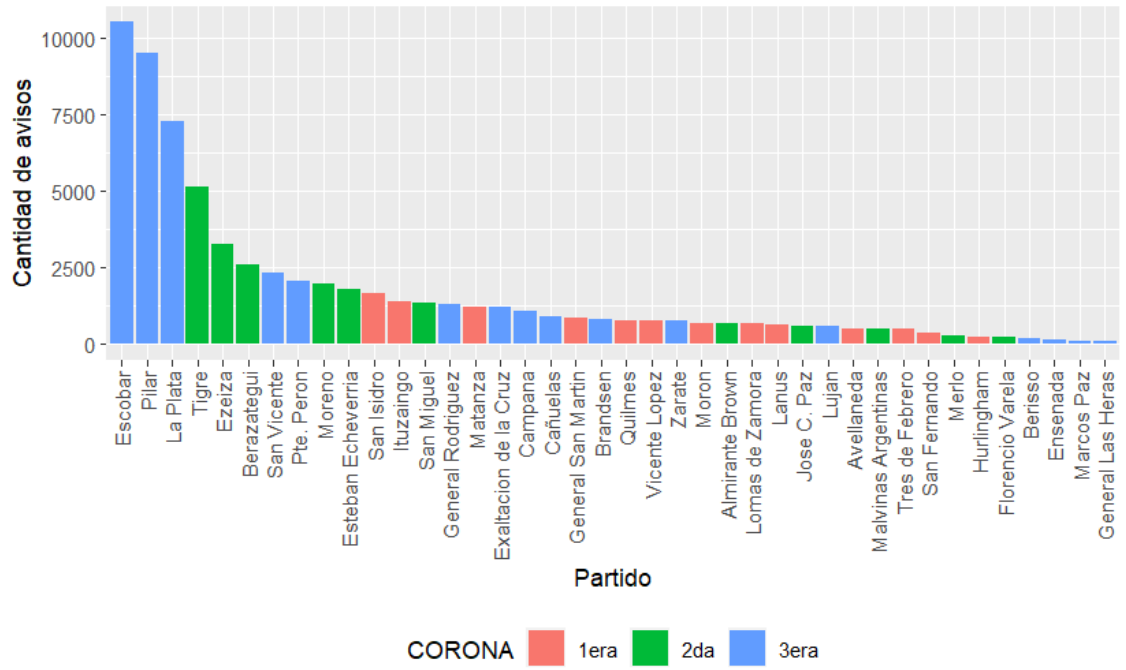

Fig. 5. Cantidad de anuncios por partido.

La frecuencia de distribución de los precios por partido también aporta un resultado preliminar interesante. Los partidos de primera corona suelen tener rango intercuartílico más extendido, con precios relativos más altos pero sin elevadas concentraciones de precios en rangos específicos tan definidos. Puede observarse que Vicente López, San Isidro, San Fernando, Tres de Febrero y San Martín en la zona norte, presentan un patrón equivalente a Morón y La Matanza en zona oeste o Avellaneda, Lanús, Lomas de Zamora y Quilmes en zona sur (ver figura 6).

En la tercer corona los partidos asociados a la ciudades más pequeñas, distantes y aún no conurbadas -como por ejemplo: San Vicente, Exaltación de la Cruz, Brandsen, Cañuelas, Luján, Zárate, Campana, Marcos Paz- presentan valores más bajos, rango más acotados y concentraciones de precios más definidas. En cambio, en los restantes partidos de la tercera y segunda corona metropolitana tiende a mostrar distribuciones mucho más asimétricas y extendidas que informa de una mayor heterogeneidad, vinculándose la tendencia a la dispersión de los precios elevados a efecto de la densificación de las áreas centrales o a la dinámica de diferenciación que las cabeceras históricas conurbadas adquieren, como subcentralidades en entornos de rápido crecimiento. Este tipo de perfiles puede observarse en el caso de Escobar, Pilar, La Plata, Moreno, Presidente Perón, Esteban Echeverría, Almirante Brown, Florencio Varela, Berisso y Ensenada.

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

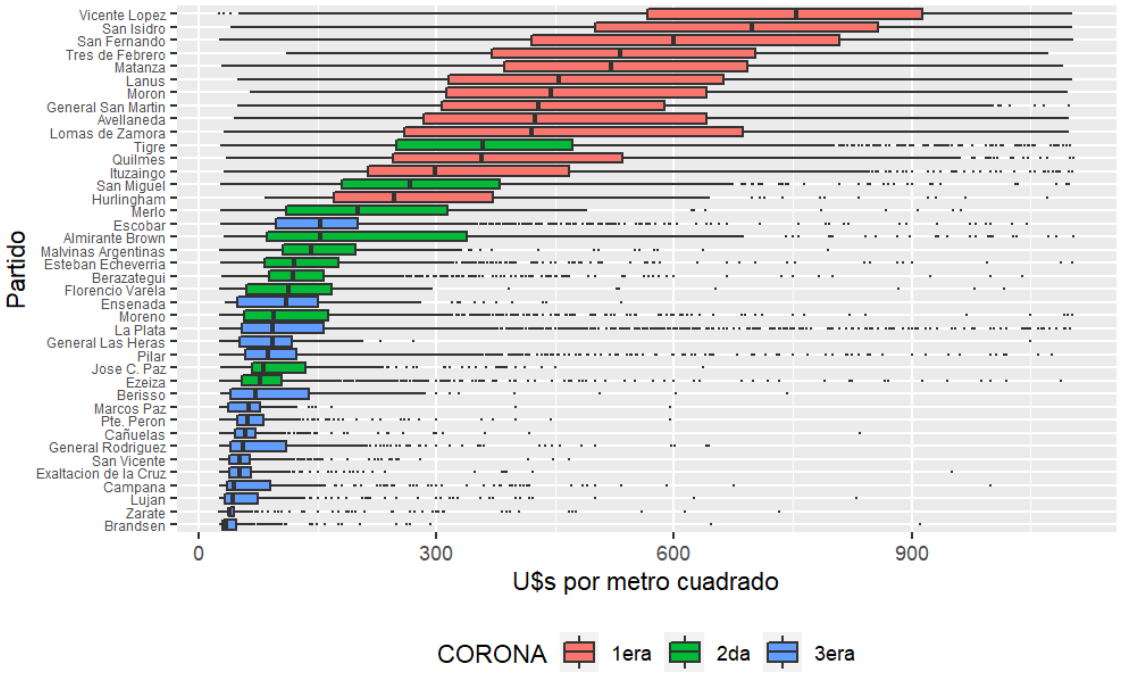

Fig. 6. Distribución de precios de terrenos por partido

De modo complementario, el mapa de valores de los terrenos ayuda a interpretar la figura 6. En el mapa puede verse tres grandes factores que organizan la distribución espacial de los precios en los distritos bonaerenses de la RMBA: la distancia a la Capital Federal que configura el centro histórico en torno al cual se estructuró el crecimiento de los distritos metropolitanos de jurisdicción bonaerense; los corredores radiocéntricos vinculados al poblamiento primero vinculado a las estaciones ferroviarias y luego al papel de las vialidades primarias, alrededor de los cuales crecen las zonas suburbanas de menor precios; en tercer término, las subcentralidades de los partidos aún no conurbados que tienden a tener valores mayores que los de su entorno inmediato.

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

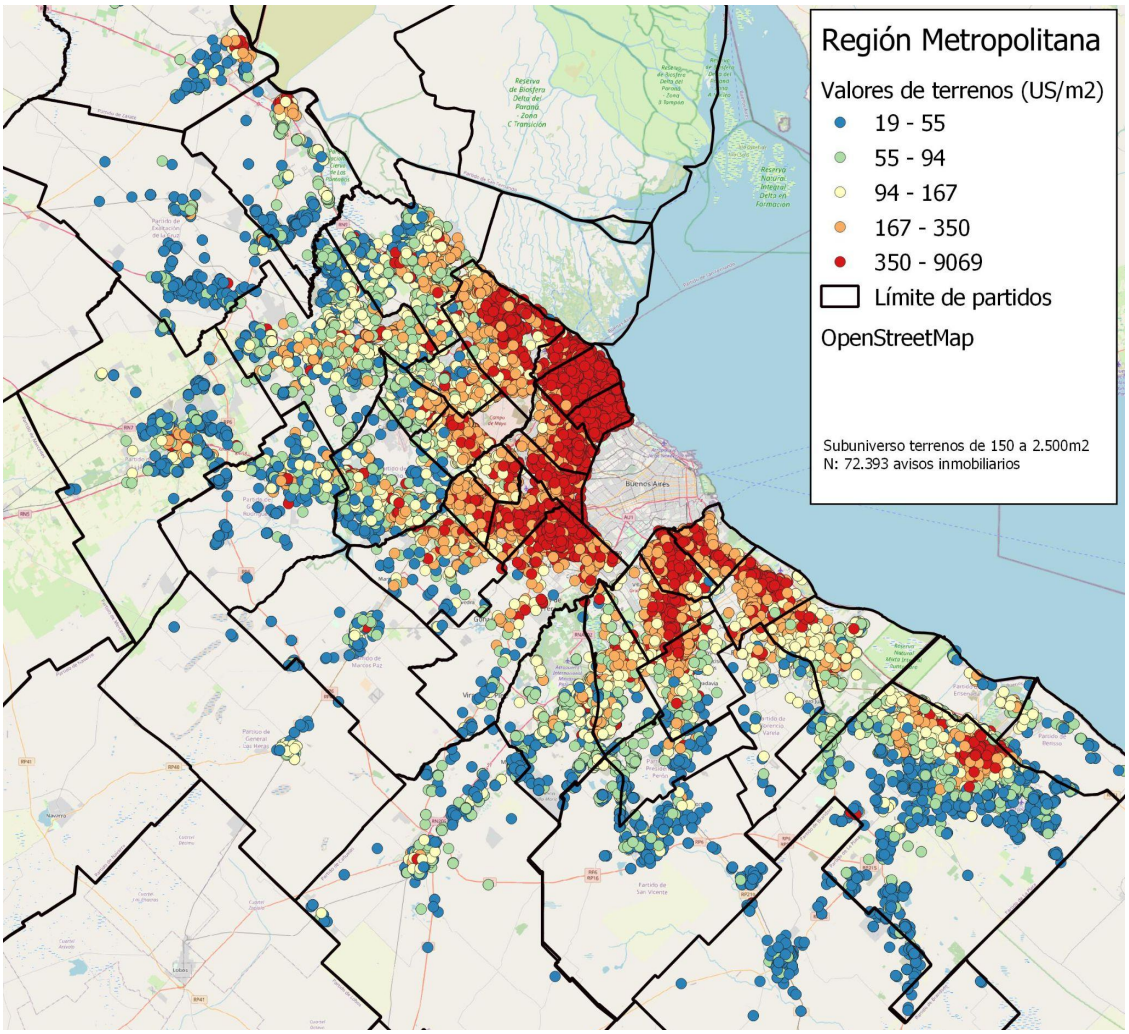

Fig. 7. Precios de terrenos de 150 a 2500 m2. Cartografía base: OSM.

Por otra parte, a efectos de explorar en qué grado las fuentes de información consultadas ofrecen datos suficientes con el propósito de utilizarlos como insumos para el cálculo plusvalías urbanas, una cuestión clave fue evaluar la densidad y cobertura espacial de la matriz de datos de terrenos. Para ello, mediante la técnica *DBSCAN* se reconocieron cluster espaciales en base grupos de oferta de mínimo 5 avisos y un distanciamiento máximo 500 metros de distancia. Este análisis se realizó para el conjunto de área urbana y complementaria según la zonificación vigente de usos del suelo de la provincia de Buenos Aires (decreto-ley 8.912) a partir de la consulta a Urbasig[10] (DPOUT). Para facilitar la interpretación, los resultados fueron integrados en una grilla de 17.230 celdas de 500 por 500 metros. Las zonas donde la densidad de oferta cumplió con el criterio establecido pueden verse a continuación en color verde, mientras que las zonas donde el flujo de oferta no llegó a configurar un cluster

[10] https://urbasig.gob.gba.gob.ar/urbasig/

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

espacial se observan en rojo. Para las restantes áreas urbanas y complementarias en gris no se relevó oferta (ver figura 8).

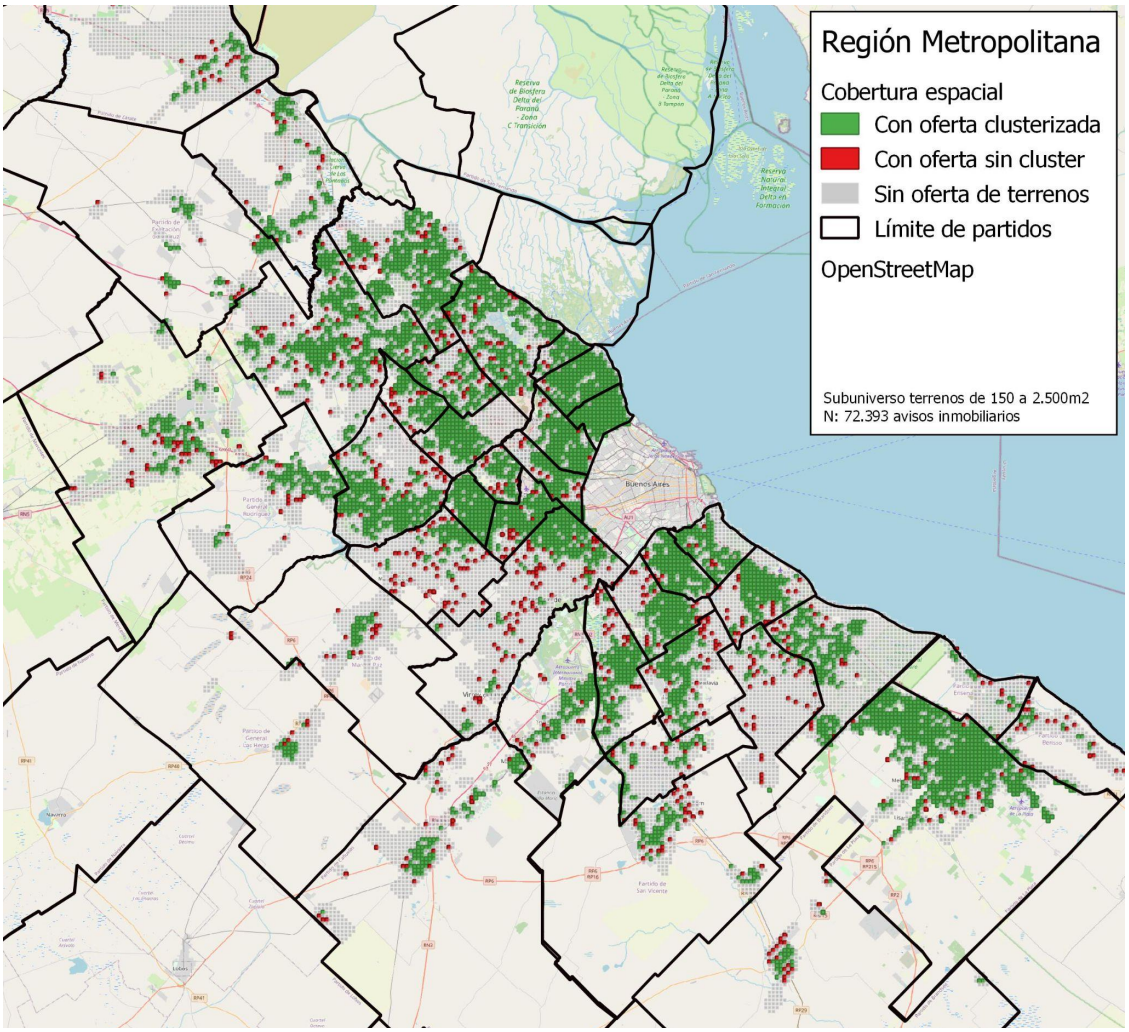

Fig. 8. Cobertura de oferta de terrenos en áreas urbanas y complementarias. Cartografía base: OSM.

Una primera conclusión del análisis anterior se vincula con la heterogeneidad que presenta la cobertura espacial de la oferta por partido. Por ejemplo, mientras que partidos como Vicente López (86%), Ituzaingó (78%), Lanús (61%) o La Plata (60%) poseen una elevada proporción de cobertura espacial de la oferta de terrenos, La Matanza (17%), Berisso (14%), Merlo (12%) o Florencio Varela (6%) son distritos con mayores restricciones en términos de cobertura de oferta, con vacíos significativos o falta de densidad de observaciones. Si bien, estos resultados se encuentran influenciados tanto por el peso efectivo que cada plataformas inmobiliarias pueden adquirir en un partido, por la estructura parcelaria y la variabilidad jurídico-administrativa de la superficie asociada a las áreas urbanas o complementarias utilizadas para el cálculo; como hipótesis de trabajo surge que el principal condicionante que

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

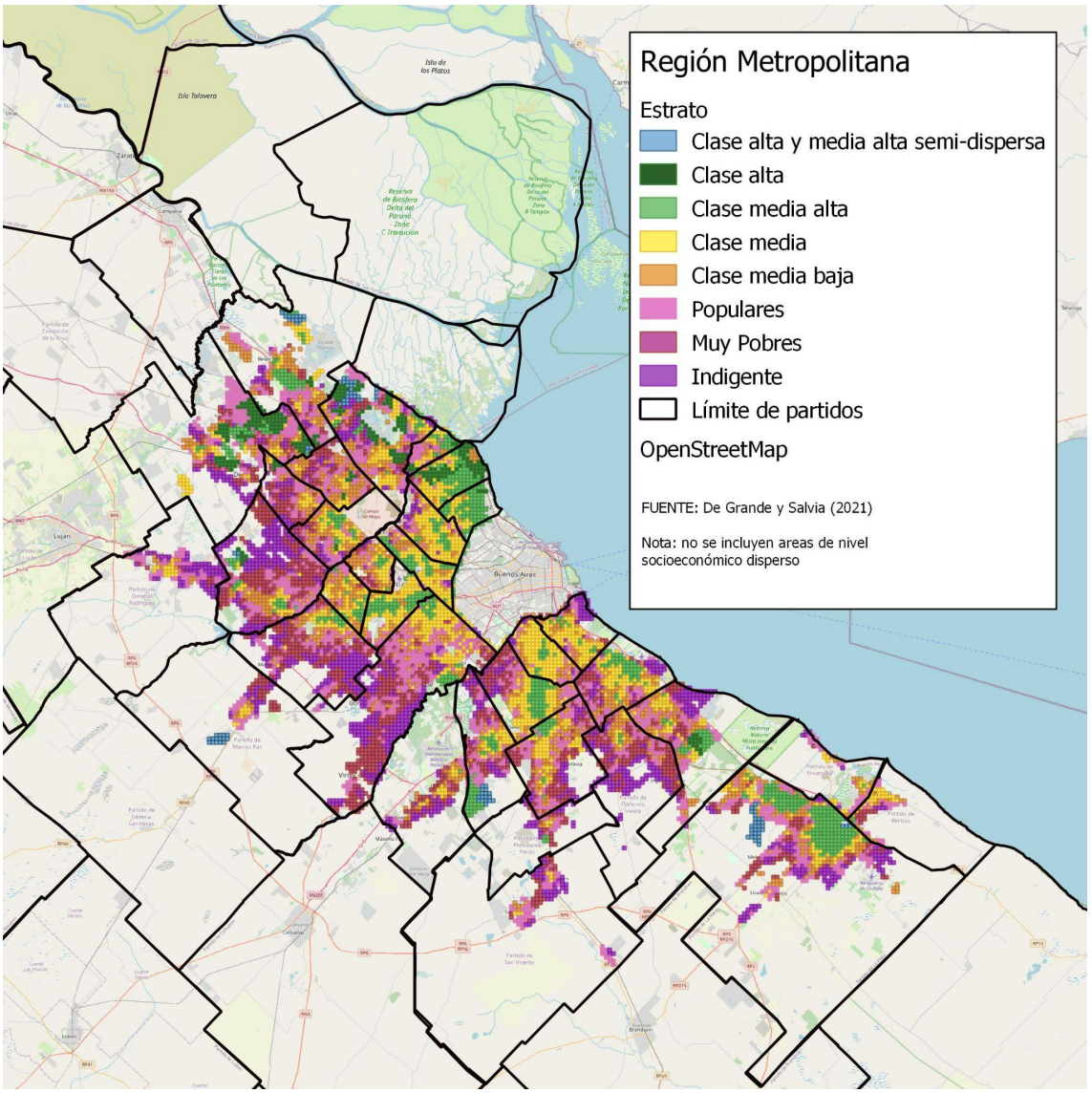

Fig. 9. Estratificación socioeconómica (De Grande y Salvia, 2019) integrada en el archivo de grilla en el cual se analizó la cobertura y densidad del flujo de oferta. Cartografía base: OSM.

restringe la densidad y cobertura del flujo de oferta guarda estrecha relación con el nivel socioeconómico de la población de los distintos contextos urbanos.

En tal sentido, un primer dato que valida esta interpretación se vincula con la asociación entre las áreas con oferta de terrenos (figura 8) y el mapa de estratificación socioeconómica (elaborado por De Grande y Salvia, 2019), el cual fue integrado en el mismo archivo de grilla en el que se analizó la presencia o ausencia de oferta (figura 9). En los espacios metropolitanos vinculados a sectores de clase media y alta el 70% del área registró la presencia de flujo de oferta (color verde o rojo en la figura 8) mientras que el porcentaje

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

desciende al 33% en el caso de los espacios con prevalencia de estratos populares. Esta diferencia del 37% entre ambas zonas permite dar cuenta de una asociación entre estrato socioeconómico y oferta de terrenos.

En relación a esto último, cabe realizar dos interpretaciones concurrentes: por un lado, la fuentes de información web utilizadas tienen menor sensibilidad en los contextos territoriales vinculados a los sectores de menores ingresos, siendo otros los canales por los que circula la información asociadas a la intermediación inmobiliaria (redes de parentesco o vecindad, micro inmobiliarias barriales, etc.); por otra parte, el mercado formal -al que mayormente se vinculan las fuentes de información consultadas- tiene en estos contextos menor intensidad.

Principales desafíos a futuro

En relación al trabajo en curso, el análisis exploratorio hasta aquí realizado plantea la necesidad de diseñar distintas estrategias metodológicas en relación a las siguientes líneas de trabajo: recuperación de variables no estructuradas en los avisos, cobertura y distribución geográfica, problema de doble contabilidad, homogeneización de valores, calidad de ubicación y outliers espaciales, evolución de los valores, entre otros.

La recuperación de las variables descriptivas de los anuncios e inmuebles, se vincula con la necesidad de sistematizar información valiosa que se encuentra en la descripción de la oferta. Es decir son variables que no se encuentran estructuradas en las plataformas o que estando estructuradas no fueron completadas por los oferentes, pero pueden detectarse en el texto de descripción general de la oferta. A modo de ejemplo: a efecto de describir el estatus jurídico de la venta (venta con escritura, venta de derecho de posesión y pre-venta) o el tipo de servicios urbanos del lugar donde se encuentra el inmueble (agua, cloaca, gas, pavimento, etc.), la descripción de los avisos suele aportar información útil para reconstruir estas variables. De allí se plantea la necesidad de aplicar técnicas de análisis del lenguaje natural para recuperar variables de interés.

Respecto a la distribución y cobertura espacial de la oferta se presentan múltiples desafíos. En aquellos sectores sin cobertura de terrenos pero con inmuebles edificados en venta se encuentra en estudio la aplicación de métodos aditivos (Jaramillo, 2009) o métodos de costo de reemplazo (Pérez Torres, 2015) para deducir el precio del suelo. Complementariamente, la

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

realización de trabajo de campo focalizado permitiría el acceso a otros circuitos de comercialización invisibles en la web. Por último, se evalúa la posibilidad de tramitar acceso a tasaciones oficiales, aunque la cobertura y densidad de observaciones suele ser escasa.

El problema de la doble contabilidad o entidades duplicadas se vincula con que -producto de los efectos de competencia- un mismo inmueble puede ser ofertado por más de una plataforma y por más de una inmobiliaria u ofertantes en una misma plataforma. Para la detección rápida de estos casos se prevé aplicar teoría de grafos por métodos computacionales (Dioguardi *et. al.* 2022) a la matriz de datos recabados, para lo cual se encuentra en elaboración un grafo de conocimiento a partir de los estándares ontológicos para la representación de inmuebles: RealEstate-Core (Hammar *et. al.* 2019) y SIOC (Berslin *et. al.* 2006).

La homogeneización de valores de los terrenos es una cuestión central con el objetivo de hacer comparables las observaciones, tanto para aplicar métodos de cálculo residual (Jaramillo, 2009; Pérez Torres, 2015) como para avanzar en el desarrollo de modelos de valuación automática masiva, siendo algunas de las variables centrales a considerar (IDECOR, 2021): superficie, medida de frente, ubicación en la cuadra (medial, esquina, interno, salida a dos calles), situación jurídica del inmueble (con o sin escritura, posesión o preventa) y tipo de valor relevado (oferta o venta-tasación). En tal sentido, una vez realizados los análisis de cobertura y control de ubicación resulta estratégico avanzar en la consolidación de una muestra para relevar la variables que permitan construir los coeficientes de ajuste para homogeneizar los valores.

En relación al control de la calidad de ubicación de las ofertas, dado la cantidad de datos y el tiempo que demanda el control manual, se presenta el desafío de identificar las áreas donde con mayor frecuencia las ubicaciones no son aceptables para trabajar en estrategias focalizadas en dichos sectores. Asimismo, en las áreas con mucha densidad de observaciones, se evalúa comenzar a trabajar con una lógica muestral, para solo en dicho subuniverso validar de modo manual la calidad de los datos. De modo complementario, se prevé trabajar con técnicas de detección de *outliers* espaciales (índice de Moran local).

Por último, otro de los desafíos del proyecto en curso se vincula con el monitoreo de la oferta en el área de estudio con una frecuencia cuatrimestral o semestral. En este sentido, el análisis de cómo evolucionan los valores presenta una doble consideración: por una parte, requiere

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

identificar intertemporalmente qué inmuebles continúan siendo ofertados y si los mismos presentan o no variación en los precios; por otra parte, implica detectar el allegamiento de nuevos inmuebles al mercado al interior del flujo de oferta.

Bibliografía y referencias

Abramo, P. (2011). *La producción de las ciudades latinoamericanas: mercado inmobiliario y estructura urbana*. Ecuador: OLACCHI.

Breslin, J., Decker, S., Harth, A., Bojars, U. (2006): SIOC: An approach to connect web-based communities. *IJWBC* (2), 133–142.

Baer, L. (2013). Principios de economía urbana y mercados de suelo. En D. Erba (Ed.), *Definición de políticas de suelo urbano en América Latina: teoría e práctica* (pp. 221-241). Lincoln Institute of Land Policy: Viçosa, MG.

Carranza, J.P.; Piumetto, M.A.; Lucca, C.M.; Sa Silva, E. (2022). Mass appraisal as affordable public policy: Open data and Machine learning for mapping urban land values. *Land Use Policy*, (119) 106-211.

Cuenya, B. (Coord.) (2009). *Recuperación de plusvalías urbanas. Aspectos conceptuales y gama de instrumentos*. Rosario: LILP y Municipalidad de Rosario.

De Grande, P. y Salvia, A. (2019). Estratificación y desigualdad social, 2010. [Cartografía: Principales 32 aglomerados urbanos del país]. Recuperado el 5 de junio, 2021, de https://mapa.poblaciones.org/map/7101

Dioguardi, F., Torres, D., Antonelli, L. y Del Río, J.P. (2022) Construcción de un grafo de conocimiento para un observatorio inmobiliario. XXVIII Congreso Argentino de Ciencias de la Computación. UNLaR. https://cacic2022.unlar.edu.ar/

DPOUT (s/f). Sistema de Información Geográfica de la Dirección Provincial de Ordenamiento Urbano y Territorial. Ministerio de Gobierno de la provincia de Buenos Aires. Recuperado el 23 de noviembre, 2021 https://urbasig.gob.gba.gob.ar/urbasig/

IDECOR (2021) Estudio del mercado de suelo urbano de la provincia de Córdoba 2020. Córdoba: Infraestructura de Datos Espaciales - Gobierno de la provincia de Córdoba.

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

Hammar, K., Wallin, E.O., Karlberg, P., Hälleberg, D. (2019): The RealEstateCore Ontology. In: Ghidini, C., Hartig, O., Maleshkova, M., Svátek, V., Cruz, I., Hogan, A., Song, J., Lefrançois, M., Gandon, F. (eds.) *The Semantic Web* – ISWC. Springer International Publishing.

Hernández, J. (2019). *Marketing Inmobiliario en la Era Digital.* Ed. Jorge Hernández. Disponible en : https://marketingdigitalinmobiliario.com/

Jaramillo, S. (2009) *Hacia una teoría de la renta del suelo.* Bogotá: Universidad de los Andes.

Morales Schechinger, C. (2005). Algunas reflexiones sobre el mercado de suelo urbano. *Curso de Especialización en Mercado y Políticas de Suelo*. Bogotá: LILP-Universidad Nacional de Colombia.

Olston, C., & Najork, M. (2010). Web crawling. *Foundations and Trends® in Information Retrieval*, 4(3), 175-246.

Pérez Torres, F. (2015) Economía Política y Métodos de Avalúo del suelo. *Equidad Desarrollo*, 24, 53-95.

Reese, E. (2006). La situación actual de la gestión urbana y la agenda de las ciudades en la Argentina. *Medion. Ambiente y Urbanización*, 65, (1), 3-21.

Schrenk, M. (2012). *Webbots, Spiders, and Screen Scrapers: A Guide to Developing Internet Agents with PHP/CURL.* No Starch Press.

Smolka, M. y Amborski, D. (2003) Recuperación de plusvalías para el desarrollo urbano: una comparación inter-americana. *Eure*, XXIX, (88), 55-77.

Smolka, M. O. y Mullahy, L. (Ed.) (2010). *Perspectivas Urbanas*. Cambridge, Massachusetts: LILP.

Zhao, B. (2017). Web Scraping. In: Schintler, L., McNeely, C. (eds) *Encyclopedia of Big Data*. Cham: Springer International Publishing.

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/