UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA CIVIL

CONSTRUCCIÓN Y ANÁLISIS DE UNA BASE DE DATOS MASIVOS Y DESAGREGADOS DE UNA CIUDAD

TESIS PARA OPTAR AL GRADO DE MAGÍSTER EN CIENCIAS DE LA INGENIERÍA, MENCIÓN TRANSPORTE

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERA CIVIL

CAMILA INÉS LIRA REVUELTA

PROFESOR GUÍA:
FRANCISCO MARTÍNEZ CONCHA

INTEGRANTES DE LA COMISIÓN:
PEDRO DONOSO SIERRA
MARCELA MUNIZAGA MUÑOZ

SANTIAGO DE CHILE
2024

RESUMEN DE LA TESIS PARA OPTAR AL
GRADO DE MAGÍSTER EN CIENCIAS DE LA
INGENIERÍA, MENCIÓN TRANSPORTE Y
MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERA CIVIL
POR: CAMILA INÉS LIRA REVUELTA
FECHA: 2024
PROF. GUÍA: FRANCISCO MARTÍNEZ CONCHA

CONSTRUCCIÓN Y ANÁLISIS DE UNA BASE DE DATOS MASIVOS Y DESAGREGADOS DE UNA CIUDAD

El rápido crecimiento urbano y la complejidad de las dinámicas de los sistemas urbanos resaltan la importancia de comprender cómo las personas eligen sus localizaciones y actividades dentro de las ciudades. Este entendimiento es crucial para mejorar la calidad de vida urbana y desarrollar una planificación de la ciudad informada y eficiente, lo cual se puede abordar con modelos de uso del suelo. Sin embargo, las ventajas de estos enfrentan desafíos significativos debido al alto costo y complejidad de la recolección de datos que se usan para estimar y aplicar los modelos. En ciudades como Santiago, a pesar de la creciente disponibilidad de datos y recursos computacionales, no existe una base de datos integrada que permita caracterizar y estimar modelos con precisión sobre el comportamiento de los agentes urbanos.

Esta tesis aborda este desafío presentando un procedimiento para construir y analizar una base de datos desagregada de uso de suelo, utilizando datos pasivos disponibles y actualizados periódicamente. Mediante la integración de múltiples fuentes y la utilización de datos censales, se busca mejorar la capacidad de predicción y reducir costos del uso de los modelos urbanos. La tesis se enfoca en avanzar en la integración de datos de los agentes residenciales, evaluando comparativamente métodos de imputación de variables y de desagregación espacial a nivel de manzana de los agentes en la ciudad.

Los resultados revelan que el diseño del proceso de integración de datos es eficiente y se logra un enriquecimiento significativo en el Censo 2017 con la imputación de ingresos residenciales de la ciudad utilizando un método de *Machine Learning*. Además, se logra una mayor desagregación espacial de la localización de los agentes a nivel de manzana censal ajustada a los datos disponibles, mediante métodos de optimización con variables enteras y restricciones lineales de igualdad. Estos avances no solo mejoran la precisión y aplicabilidad de las bases de datos requeridas para las predicciones urbanas, sino que también establecen una base empírica rica y actualizable a muy bajo costo para futuras investigaciones y aplicaciones en planificación urbana. La metodología desarrollada podría extenderse a otras ciudades que cuenten con bases de datos similares, como lo son todas las ciudades de Chile que cuentan con una encuesta tipo Casen, contribuyendo al entendimiento y modelación del Uso de Suelo urbano.

Agradecimientos

Antes de comenzar este extenso trabajo que marca el fin de un largo camino, quiero expresar mi gratitud a cada persona que ha sido un pilar fundamental para llegar a este punto. Mi más sincero agradecimiento a todos por su compañía, apoyo y cariño entregado en este tiempo.

Quiero comenzar agradeciendo a mi familia, desde mis abuelos, tíos/as, primos/as, quienes siempre han sido muy importantes para mí. Atesoro inmensamente su amor y apoyo incondicional. A mi abuelo, el Pipi, quien siempre, orgulloso, se interesó en saber cómo me iba en la universidad y en qué estaba. A mi abuela, la Mimi, que cada fin de semana nos recibió en su casa con amor y cuidados que solo una abuela puede ofrecer. También un especial agradecimiento a la Javi, que en el momento de más dudas y miedos logró emocionarme con su apoyo y cariño.

A mis padres, quienes con sus sacrificios constantes para educarnos desde pequeños hicieron esto posible. Han estado a mi lado en cada momento de estrés y duda, ofreciéndome su apoyo y compañía. Me han entregado todo lo que he necesitado y más. Quiero recordarles que sin ustedes no podría haber llegado a donde estoy, y les debo toda una vida.

A mis hermanos, Tomás, Sebastián, Macarena y María Jesús, gracias por ser una compañía toda la vida y en este proceso. Un especial agradecimiento a la Jesu que revisó cada punto y coma de esta tesis, menos los agradecimientos, No puedo dejar de mencionar al Kaiser, la Mamucha y la Lili, las compañías más lindas y desinteresadas que se pueden tener. Me acompañaron en la pandemia, y los extrañaré un montón; al menos la Lili se encargará de seguir llenando mis polerones de pelos.

A mis compañeras y amigas de transporte, Pauli y Fran, gracias por ser la mejor compañía que podría haber tenido en una carrera tan pequeña donde puedes pensar que estarás sola, pero con ustedes nunca fue así. Anhelo seguir viéndolas a lo largo de nuestras carreras, cumpliendo sueños y metas, y verlas ser tan felices como nos hemos imaginado este último tiempo. A mis amigos de primer año, Brook, Cata, Joaco, Pablo, Pauli, Nacho y los Martines, que hicieron esta etapa mucho más sencilla y pasadera. Y al grupo de Pichidangui, con quienes compartí almuerzos, ramos e incontables horas de estudios en plan común, especialmente a la Cata Acevedo, a quien agradezco su apoyo como compañera de universidad en civil, tutoría y como amiga.

Un agradecimiento especial a mi profesor guía, Francisco Martínez, por su constante dedicación y valiosa orientación en este trabajo de una manera que jamás hubiese imaginado. A la comisión, Pedro Donoso, por su compromiso con este trabajo, desafiándome y aportando nuevas ideas. Y a Marcela Munizaga, también por ser mi primera profesora en transporte, quien despertó mi pasión la carrera y me motivó a seguir este camino.

Al Rafa, mi compañero de vida que encontré en este camino, una persona que me inspira en lo que hace, que me ha acompañado y escuchado hablar hasta del problema más absurdo que me pudo haber pasado en estos últimos 5 años y un poquito más. Muchas gracias por estar a mi lado siempre.

Este viaje ha sido largo y desafiante, pero estoy profundamente agradecida del él y todo lo que he logrado. Gracias a todas estas personas y muchas más que lamentablemente no puedo mencionar, sino no terminaría nunca y ya quiero entregar la tesis. Finalmente, aunque pueda sonar egocéntrico, me gustaría agradecerme a mí misma. Tras superar un poco el síndrome del impostor, debo mencionar que gracias al esfuerzo y dedicación que le entregué a la universidad estos últimos siete años, puedo decir que, con este trabajo, finalmente seré ingeniera (y con magíster).

Tabla de Contenido

Índice de Tablas

Índice de Figuras

Capítulo 1
Introducción

El estudio de la configuración y la evolución de las ciudades, en particular de la economía urbana y la dinámica de los sistemas urbanos, de los modelos de localización, la accesibilidad y la interacción entre el transporte y el uso del suelo, conforma un área de estudio de creciente relevancia (ISCI, n.d.), debido a la influencia diaria de las ciudades en la vida de las personas que se acentúa en el contexto actual de urbanización acelerada. Según las estimaciones de urbanización mundial de las Naciones Unidas (UN, 2018), se proyecta que el 68% de la población mundial vivirá en las ciudades para el año 2050. Chile, por su parte, ya ha superado el 87% de urbanización de su población (BCN, n.d.). Por lo tanto, resulta esencial comprender cómo las personas eligen sus ubicaciones y actividades dentro de la ciudad para así contribuir en la calidad de vida de estas y al desarrollo eficiente de la ciudad a través de una planificación informada.

En este contexto, se ha desarrollado el Modelo de Uso de Suelo de Santiago, conocido como MUSSA, el cual sirvió de base para el desarrollo de CUBE Land (Virtuosity. 2023) que ha sido aplicado en distintas ciudades del mundo, como Minneapolis, Boston, París, Berlín, entre otras, y se utiliza en la docencia en varias universidades del mundo. CUBE Land es una herramienta que no solo permite realizar predicciones o simulaciones del mercado inmobiliario urbano, sino también analizar una gran cantidad y variedad de información que caracteriza a este mercado. Además, permite estudiar los efectos de la aplicación de variadas políticas de localización y transporte (ISCI, s.f.).

A pesar de sus ventajas de CUBE Land y su continuo desarrollo y aplicación, los modelos de uso del suelo, enfrentan limitaciones significativas debido al costo de su calibración. En el caso de MUSSA en la actualidad está calibrado con datos de los años 90´s solo para la ciudad de Santiago y no ha sido actualizado. La complejidad de la recolección de datos y de los métodos econométricos para calibrar los modelos con estos datos, genera dificultades para disponer de modelos actualizados de ciudades o de grupos de ciudades.

Por otra parte, cada vez es más común contar con bases de datos granulares de actualización periódica tanto de uso de suelo como de transporte en distintas ciudades del mundo, tales como el CENSO (INE, 2018), Catastro de Bienes Raíces (SII, s.f.) y ADATRAP (DTPM, s.f.). Estas bases proporcionan datos censales sobre la población y bienes inmuebles, así como, en el caso ADATRAP, información de movilidad detallada en el transporte público de Santiago, gracias a los sistemas de pago digital y la tecnología GPS en los buses. En los últimos años, los investigadores han recurrido a enfoques avanzados de fusión de *Big Data* urbana basados en *Machine Learning* y *Deep Learning* para abordar problemas de predicción y cálculo de la calidad del aire urbano, predicciones de flujo de multitudes urbanas y predicción de la calidad del agua urbana en las ciudades inteligentes (Khan et al., 2021). A

pesar de estos avances, en el caso de Santiago todavía no existe una base integrada de datos granulares que permita caracterizar o estimar el comportamiento de los agentes de la ciudad a un gran nivel de desagregación espacial que recoja apropiadamente las no linealidades de los sistemas urbanos y permita generar y evaluar políticas específicas.

Esta tesis busca avanzar en esta línea mediante el desarrollo de un procedimiento para construir y analizar una base de datos desagregada de uso de suelo y transporte basada en mega bases de datos con un alto nivel de granularidad, como el Censo de Población y Vivienda. Este procedimiento permite la actualización periódica de los datos, eliminando así el costo de recolección de datos y reduciendo también el costo de calibración del modelo MUSSA, además, incorpora el uso de inteligencia artificial mediante la comparación de métodos de *Machine Learning*. Para lograr tal base integrada, en esta tesis se define un procedimiento general del proyecto de esta base de datos de uso de suelo y transporte y avanza en el ámbito de la localización de agentes, lo cual tras ser complementado con datos de transporte permitirá la estimación futura actualizada, granular y periódica del modelo de uso de suelo. Además, permitirá disponer de una base de datos desagregada, actualizable y enriquecida con la aplicación de métodos de inteligencia artificial. Esta herramienta no solo servirá para los modelos de uso de suelo, sino que también puede ser utilizada para diferentes ámbitos y estudios que requieran información precisa de la ciudad.

Objetivos

Objetivo General

El objetivo general de la tesis es diseñar un procedimiento de procesamiento de datos para construir y analizar una mega base de datos, con integración de varias fuentes, de carácter censal, alta granulometría geográfica y actualizable periódicamente, generando datos desagregados espacialmente y censales. Este esfuerzo busca mejorar significativamente la precisión y la aplicabilidad de las predicciones en contextos urbanos. Además, se avanzará en la construcción de la base de datos desarrollando e implementando métodos para enriquecer el Censo de Población con información de otras bases de datos.

Objetivos Específicos

1. Identificar y recopilar las diversas bases de datos relevantes para la modelación del uso de suelo.

2. Desarrollar un proceso de integración de datos que permita la combinación eficiente de información relacionada de agentes, rentas, bienes inmuebles y transporte.

3. Desarrollar un método de imputación e imputar el ingreso de los hogares en el conjunto de datos censales, evaluando su efectividad y eficiencia de diversas técnicas.

4. Evaluar y comparar métodos de desagregación espacial, asignando las viviendas censales desde un nivel espacial de zona censal a manzana censal para una mayor granularidad en la modelación.

5. Evaluar la validez y utilidad de los procesos implementados para representar socioeconómicamente a los agentes de la ciudad a un nivel detallado.

Así, las preguntas de investigación de esta tesis son:

- ¿Qué metodología es adecuada para la integración de datos eficiente que permita construir y actualizar una mega-base de datos consistente en cuanto a geo-referencia y tempo-referencia del dato, integrando diferentes bases de datos de uso de suelo y transporte de la ciudad de Santiago, manteniendo un nivel desagregado?
- ¿Qué método es el mejor para procesar la base original de datos censales de los agentes, para imputar datos faltantes y asignar espacialmente?

Esta tesis se compone de seis capítulos. En el primer capítulo, se introduce el tema a tratar y se delimitan los objetivos y alcances de la investigación. El capítulo 2 se dedica a la revisión bibliográfica, marco conceptual y estado del arte relevante en el área. Posteriormente, en el capítulo 3, se describen las fuentes de datos urbanos y de transporte relevantes y disponibles en la actualidad para la ciudad de Santiago, y se diseña un diagrama de integración de bases de datos urbanos y la metodología para el procesamiento de datos de los agentes de la ciudad. El capítulo 4 se enfoca en detallar el proceso de aplicación de la metodología en el estudio de caso de Santiago y los resultados obtenidos. Luego, en el capítulo 5 se realiza un análisis de los resultados. Finalmente, el capítulo 6 presenta las conclusiones de la investigación y posibles trabajos y líneas de investigación futuras.

Capítulo 2
Recopilación de Antecedentes

Introducción

Este trabajo se enmarca en el contexto de generar una base de datos detallados de la ciudad que podría ser utilizada para la calibración de los modelos de uso de suelo, como en el caso aplicado a Santiago de Chile, es decir, el modelo de Uso de Suelo de Santiago (MUSSA). En este sentido, se debe tener en cuenta que el enfoque con el cual se va a trabajar tanto la recolección como tratamiento de los datos urbanos estará enfocado en obtener la información necesaria para calibrar este modelo. Lo cual no obsta que esta base y los procesos se puedan utilizar para diferentes objetivos que requieran este tipo de datos y/o procedimientos.

En este capítulo se tiene como objetivo proporcionar una revisión de la literatura y los conceptos fundamentales relacionados con el desarrollo del trabajo de tesis. Esto es crucial para comprender la base teórica y conceptual que sustenta este trabajo. De esta manera, la revisión bibliográfica proporcionará las bases para la formulación y desarrollo del enfoque de investigación.

Específicamente, se presentan conceptos fundamentales relacionados con el Modelo de Uso de Suelo de Santiago (MUSSA), con énfasis en la distinción entre agentes residenciales y no residenciales, la importancia de la localización en zonas y la oferta de transporte, así como la accesibilidad. Luego, se lleva a cabo una revisión bibliográfica sobre las bases de datos urbanos, su relevancia y las existentes en otras partes del mundo. Finalmente, se realiza una revisión bibliográfica y del estado del arte sobre el procesamiento de bases de datos en este contexto, incluyendo la fusión e integración de bases de datos, el uso de enfoques de *Machine Learning* y *Deep Learning*, como Random Forest, k-means y métodos para estimar variables faltantes y para la desagregación de datos, como la imputación de ingresos, creación de población sintética y maximización de la entropía, entre otros.

Conceptos clave

En esta sección, se exploran conceptos clave fundamentales para comprender este trabajo de investigación y el modelo MUSSA presentados en el libro "Microeconomic Modeling in Urban Science" de Martínez (2018). Se aborda la noción de agentes residenciales y no residenciales, así como conceptos relacionados con la localización en zonas, los movimientos en las redes y la accesibilidad en la ciudad.

1. Agentes residenciales y no residenciales

En el contexto del modelo MUSSA, los agentes son aquellos que toman decisiones sobre su localización, consumo o producción. Los agentes pueden ser personas, empresas o

instituciones y se dividen principalmente en dos categorías: agentes residenciales y agentes no residenciales. Los agentes residenciales, en su mayoría individuos y/u hogares, toman decisiones sobre dónde vivir y cómo utilizar su tiempo en la ciudad. En contraste, los agentes no residenciales, como las empresas o instituciones, están enfocados en la producción, el comercio y la prestación de servicios.

Los agentes residenciales eligen sus ubicaciones de residencia y las actividades que realizan en la ciudad. Los hogares, formados por grupos de individuos, comparten su localización y un presupuesto común. Por otro lado, los agentes no residenciales determinan dónde establecer sus operaciones económicas para maximizar sus beneficios reduciendo sus costos y maximizando sus ventas, lo que puede incluir la elección de ubicaciones estratégicas para garantizar un buen acceso a sus servicios o productos. El supuesto fundamental es que todos estos agentes se rigen por la racionalidad, buscando decisiones que optimicen su bienestar o beneficio, mientras compiten por recursos y mercados y se adaptan a las cambiantes condiciones del mercado.

2. Localización en Zonas y Oferta de Transporte

La localización en el Modelo MUSSA se divide en tres categorías principales: residencial, comercial e industrial, siendo estas últimas dos, parte de las localizaciones no residenciales. Las decisiones de localización incluyen la elección de un individuo sobre dónde vivir, la elección de una empresa comercial o de un servicio sobre la ubicación de su negocio y la elección de una empresa industrial sobre la ubicación de una planta de producción en las diferentes zonas disponibles en la ciudad.

La localización residencial, por ejemplo, depende de factores como la proximidad al trabajo, la calidad del entorno y las opciones de transporte disponibles. Las decisiones de localización comercial se centran en la maximización de la visibilidad y el acceso a los clientes, mientras que las decisiones de localización industrial buscan la eficiencia en la producción y la logística.

La oferta de transporte, compuesta por nodos y arcos, interconectan las zonas y destinos de los agentes en la ciudad. Esta es esencial para facilitar los movimientos de los agentes entre sus localizaciones y actividades en la ciudad. Esta oferta permite que las personas y las mercancías se desplacen eficientemente a través de la ciudad utilizando diversos medios de transporte.

3. Accesibilidad

La accesibilidad es un concepto crítico en el Modelo MUSSA, ya que influye en las decisiones de localización y movimiento de los agentes en la ciudad. Se define como la facilidad (beneficio) de visitar distintas actividades desde una localización y se compone de

tres factores principales: la distancia entre los agentes y las actividades que se desean realizar, el costo asociado al transporte requerido y la disponibilidad de opciones de transporte adecuadas.

La distancia desempeña un rol esencial, ya que cuanto más lejos se encuentren los agentes de sus destinos, mayor será el tiempo y el costo asociado a los desplazamientos. El costo del transporte también es un factor determinante, ya que afecta directamente la viabilidad económica de viajar a ciertos lugares. Además, la disponibilidad de opciones de transporte adecuadas, como sistemas de transporte público o carreteras eficientes, influye en la movilidad de los agentes.

Paralelamente, la atractividad se define como el beneficio de ser visitado sin realizar un viaje. Este concepto es considerado por los agentes y se ve influenciado tanto por la disponibilidad de opciones de transporte adecuadas que faciliten llegar a una localización como por las economías de aglomeración, donde a mayores actividades disponibles, mayores son los beneficios percibidos. En conjunto, los conceptos de accesibilidad y atractividad configuran lo que se denomina el acceso de una localización.

Modelos de Uso de Suelo

Los modelos de uso de suelo urbanos intentan comprender el comportamiento de la ciudad considerando individuos que maximizan su utilidad para elegir su localización y actividades en la ciudad. Estos individuos valoran y eligen entre las opciones de ubicaciones con tal de tener el máximo beneficio de pertenecer a la ciudad (Martínez, 2018). No obstante, también consideran que, para poder realizar diferentes actividades dentro de la ciudad, deben viajar dentro de esta. Por lo tanto, el uso del suelo y el transporte están intrínsecamente conectados (Duranton & Puga, 2015).

En este sentido, la accesibilidad de cada zona de la ciudad influye en los precios del suelo y la localización de las personas. Sin embargo, cabe destacar que a su vez la accesibilidad se verá afectada por las localizaciones de las viviendas y empresas en la ciudad, debido a que las actividades de estas afectan directamente en la demanda de transporte y, por ejemplo, una mayor densidad de empresas genera una mayor atractividad. Por lo tanto, el problema de uso de suelo es un problema de equilibrio iterativo, donde múltiples factores interactúan de manera compleja (Duranton & Puga, 2015). Para abordar y describir el comportamiento de este sistema de alta complejidad se requieren de herramientas matemáticas sofisticadas, además de datos ricos y detallados (Martínez, 2018).

Dado lo anterior, se requieren datos de la localización residencial y no residencial observada, de la oferta de bienes inmuebles residencial y no residencial disponibles al momento de la observación, los precios de los bienes inmuebles residenciales y no residenciales e indicadores de accesibilidad en las zonas de la ciudad. A continuación, se presentan las bases

de datos urbanos más comunes en la literatura utilizados en Estados Unidos, Londres y Australia. Estos ejemplos ayudarán a comprender la variedad y tipo de datos que se suelen tener disponibles y que son utilizados para abordar este tipo de problemas.

Una de las bases de datos urbanos más conocidas en EE. UU. es la encuesta American Community Survey (ACS) del U.S. Census Bureau. Es realizado mensualmente cada año y enviado a una muestra de direcciones (aproximadamente 3.5 millones) en los 50 estados (U.S. Census Bureau, s. f.-b). La ACS proporciona datos detallados sobre temas no incluidos en el Censo 2020, como educación, empleo, acceso a Internet y transporte, proporcionando información actualizada cada año. Este detalle complementa los datos del Censo que se realiza cada diez años y que cuenta a todas las personas que viven en los 50 estados. Además, la base de datos Longitudinal Employer-Household Dynamics (LEHD) del Centro de Estudios Económicos del U.S. Census Bureau, proporciona información sobre la relación entre el lugar de trabajo y la residencia, llenando vacíos de datos críticos para autoridades estatales y locales (U.S. Census Bureau, s. f.-a).

En Inglaterra, el Office for National Statistics proporciona datos extensos a través de su Censo (Office for National Statistics, s. f.). Este incluye variables demográficas, educativas, de grupo étnico, salud, discapacidad, vivienda, migración internacional, mercado laboral, viaje al trabajo, entre otros. También están disponibles los datos del Land Registry, que proporcionan información detallada sobre precios y transacciones inmobiliarias, con conjuntos de datos públicos como el Price Paid Data, Transaction Data, y el UK House Price Index, actualizados mensualmente y disponibles desde 1995, que rastrean las ventas de propiedades residenciales en Inglaterra y Gales (GOV.UK, s. f.).

Finalmente, en Australia, el Australian Bureau of Statistics ofrece datos censales detallados, incluyendo información sobre vivienda, empleo y transporte (Australian Bureau of Statistics, s. f.). Adicionalmente, el Australian Urban Research Infrastructure Network (AURIN) proporciona acceso a una variedad de bases de datos urbanas, como empleo, transporte, infraestructura, ingresos, migración, propiedades, parques, entre otros. Esta se trata de una red colaborativa nacional que involucra a investigadores y proveedores de datos líderes en los sectores académico, gubernamental y privado, proporcionando un banco de trabajo en línea con acceso a miles de conjuntos de datos multidisciplinarios (AURIN, s. f.).

Procesamiento de Bases de Datos

El tema central de la tesis consiste en realizar una fusión e integración de diferentes bases de datos y enriquecer los datos censales. Por lo tanto, se debe realizar una revisión bibliográfica de estos conceptos y de cómo se ha abordado este tema en términos de clasificaciones y metodologías específicamente para datos urbanos y espaciales. En las siguientes secciones se presentan los principales artículos que describen lo mencionado.

Fusión e integración de datos

La fusión de datos consiste en combinar diversas fuentes para obtener información mejorada, es decir, información menos costosa, de mayor calidad o más relevante (Castanedo, 2013). Los investigadores han aceptado la definición propuesta por los Joint Directors of Laboratories (JDL). Según JDL, la fusión de datos es "un proceso de múltiples niveles que se ocupa de la asociación, correlación, combinación de datos e información de múltiples fuentes para lograr una posición refinada, identificar estimaciones y evaluaciones completas y oportunas de situaciones, amenazas y su importancia". (White, 1991).

Con relación a la fusión de datos, se han propuesto diferentes clasificaciones que se presentan en Castanedo (2013) y Meng et al. (2020). A partir de estas, para la estructura general de esta tesis, se considera la propuesta por Durrant-Whyte (1988). En la Figura 1 se presenta la clasificación basada en las relaciones entre las fuentes de datos de Durrant-Whyte (1988), donde las fusiones complementarias se refieren a la unión de información diferente de diferentes fuentes (Sources) para formar una base más completa, mientras que la fusión redundante se refiere a la fusión de una misma información de diferentes fuentes que pueden fusionarse para incrementar la confianza. Finalmente, las fusiones cooperativas consisten en combinar información lo que permite generar información más compleja que la original.

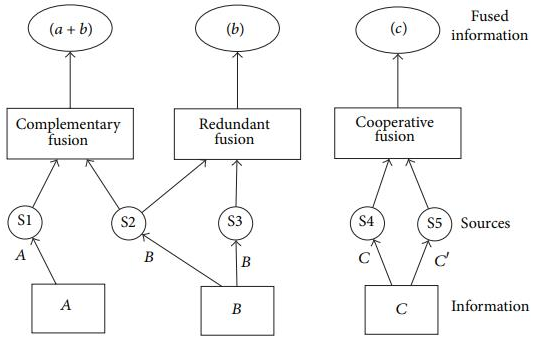


Figura 1: Clasificación Durrant-White (Castanedo, 2013)

Enriquecimiento de Datos Censales

Sobre el enriquecimiento del Censo, en primer lugar, es importante mencionar la importancia de trabajar con datos censales. Estos ofrecen una cobertura completa de la población objetivo, eliminando posibles sesgos por selección de muestras. La representatividad geográfica y socioeconómica de estos datos proporciona una visión más precisa y diversa de la realidad urbana, evitando distorsiones presentes en muestras limitadas. Los Censos y los datos espacialmente detallados son cruciales para una modelación más precisa y correcta del uso del suelo urbano.

Leslie Kish (1965) discute las ventajas y desventajas de los Censos y las muestras, señalando que los Censos son la única forma de obtener una estimación exacta de la población total, mientras que las muestras siempre introducen algún error en la estimación de la población total. El tamaño de la muestra y el diseño de la muestra determinan la magnitud del error que se introduce. Estudios recientes, como el de Held et al. (2018), destacan cómo los datos del Censo son cruciales para una estimación espacial precisa en los modelos de suelo urbano, ya que proporcionan una muestra estadísticamente representativa.

Sin embargo, los datos censales presentan desafíos, porque carecen de variables como ingresos o atributos relacionados con la movilidad (Held et al, 2018), periodicidad baja y georreferenciación a un nivel muy agregado por temas de privacidad. Los Censos también son costosos, laboriosos, especialmente en poblaciones grandes o dispersas, mientras que las muestras son más económicas y eficientes en la recolección de datos (Kish L, 1965). Kish argumenta que Censos y muestras son complementarios, no sustitutivos: los Censos son útiles para obtener estimaciones exactas de la población total, mientras que las muestras son útiles para obtener estimaciones rápidas y económicas, específicamente de subpoblaciones específicas.

En los últimos años, los investigadores han recurrido a enfoques de fusión de *big data* urbana basados en *deep learning* para abordar estos problemas. Por ejemplo, el artículo de Kovacs-Györi et al. (2020) analiza el papel del análisis geoespacial en la promoción de la habitabilidad urbana en el contexto del *big data* y el aprendizaje automático. Los autores argumentan que estas tecnologías emergentes ofrecen una oportunidad sin precedentes para comprender las complejas dinámicas de las ciudades y desarrollar soluciones enfocadas a mejorar la calidad de vida de sus habitantes. El uso de aprendizaje automático se ve potenciado debido a los desafíos que implica trabajar con *big data* utilizando técnicas tradicionales. Por esto, en este trabajo se utilizarán diferentes métodos de *Machine Learning* para abordar estos desafíos de variables faltantes y otros métodos estadísticos y geoespaciales. A continuación, se detallan los principales métodos de *Machine Learning* utilizados, destacando Random Forest (Breiman, 2001) y K-Means (Shindler et al., 2011).

Random Forest

Random Forest es un algoritmo de aprendizaje automático supervisado que se utiliza para clasificación y regresión. Fue desarrollado por Leo Breiman y Adele Cutler en 2001. Este algoritmo es una extensión del algoritmo Decision Tree. Un árbol de decisión es un modelo de aprendizaje automático que predice los valores de una variable objetivo en función de un conjunto de variables predictoras. El árbol se construye dividiendo el rango de la variable a predecir en grupos conformados por cortes de las variables predictoras. La variable predictora que mejor divide el conjunto de datos se elige como la variable de decisión. El proceso se repite hasta que todos los datos se hayan asignado a un subconjunto. Random Forest construye un conjunto de árboles de decisión independientes. Cada árbol se construye

utilizando un subconjunto aleatorio de las variables predictoras. Esto ayuda a reducir el riesgo de sobreajuste, que es un problema que puede ocurrir recurrentemente con los árboles de decisión (Breiman, 2001).

Funcionamiento del algoritmo Random Forest

1. Crear un bootstrapped dataset: Random Forest comienza tomando una muestra con reemplazo del conjunto de datos original. Si el número de casos en el conjunto de entrenamiento es N, se toman al azar N casos de los datos originales. Esta muestra se convierte en el conjunto de entrenamiento para cada árbol.

2. Crear un árbol de decisión con el bootstrapped dataset: Si hay M variables de entrada, se especifica un número m << M aleatorio de variables. En cada nodo del árbol, se seleccionan al azar m variables de las M y se utiliza la mejor división en estas m variables para dividir el nodo. El valor de m se mantiene constante durante el crecimiento del bosque.

3. Repetir paso 1 y 2: Se repiten los pasos anteriores varias veces, lo que entrega un conjunto de árboles de decisión.

4. Clasificar nuevos datos: Se clasifica cada dato en todos los árboles. Cada árbol proporciona una clasificación y "vota" por esa clase. La clasificación que obtiene más votos o el promedio entre todos los árboles se selecciona como la salida del bosque, a este proceso se le llama bagging.

5. Estimar la precisión: Los datos que no entran en el dataset (out-of-bag dataset) se utilizan para obtener una clasificación y estimar el error de Out-of-bag.

Sobreajuste

El sobreajuste es un problema que puede ocurrir con los modelos de aprendizaje automático. Ocurre cuando el modelo se ajusta demasiado a los datos de entrenamiento y no es capaz de generalizar a nuevos datos. Random Forest es menos propenso al sobreajuste que los árboles de decisión individuales. Esto se debe a que Random Forest utiliza un conjunto de árboles de decisión independientes. Cada árbol de decisión tiene un riesgo menor de sobreajuste, lo que reduce el riesgo de que el conjunto de árboles de decisión se sobreajuste. Por otro lado, Random Forest utiliza Cross Validation, un proceso donde utiliza un porcentaje de los datos para entrenar el modelo y otro para testear, esto sirve para obtener información sobre cómo las diferentes configuraciones de parámetros afectan al equilibrio entre sobreajuste y subajuste. (Pedregosa et al., 2011)

K-means

El algoritmo k-means fue desarrollado por MacQueen en 1967 y es un algoritmo de aprendizaje automático no supervisado que se utiliza para agrupar datos sin necesidad de conocer las categorías predefinidas. El algoritmo iterativamente divide las observaciones en k grupos (clusters) de manera que las observaciones dentro de cada grupo sean lo más similares posibles entre sí y lo más diferentes posible de las observaciones de otros grupos (MacQueen J., 1967).

Funcionamiento del algoritmo k-means

1. Se selecciona un número k de centroides/grupos que se desea crear.
2. Se seleccionan aleatoriamente k puntos como centroides iniciales.
3. Se asigna cada dato al grupo cuyo centroide sea el más cercano.
4. Se recalcula la posición de los centroides.
5. Se repiten los pasos 3 y 4 hasta que los centroides no cambien o hasta que se alcance un límite predefinido de iteraciones.

A continuación, se realiza una revisión bibliográfica sobre los desafíos presentados en los datos censales y cómo han sido abordados en la literatura, como el caso de tener variables faltantes o requerir una mayor granularidad espacial de los datos.

Métodos para estimar variables faltantes

El Censo de Población y Vivienda de Chile realizado el año 2017, es el último disponible, el cual fue un Censo abreviado debido a los problemas existentes con el Censo del año 2012. Por esta razón, el Censo del año 2017 no recopiló toda la información, como el ingreso del hogar. Es importante destacar que, en general, los censos alrededor del mundo tampoco suelen incluir información del ingreso de los hogares. Sin embargo, el ingreso del hogar ha sido fundamental para el modelamiento de la localización residencial, ya que permite segmentar los diferentes tipos de agentes residenciales. Por lo tanto, en este trabajo se realiza una búsqueda de diferentes métodos que se han utilizado para la caracterización de los agentes y clasificación de estos en grupos de agentes similares, así como para el modelamiento de los ingresos de los hogares.

En primer lugar, se revisó la metodología de la Asociación de Investigadores de Mercado (AIM). La AIM presenta la clasificación de los Grupos Socioeconómicos (GSE) basada de tres variables: ingreso per cápita, nivel de estudios y ocupación del jefe de hogar, utilizando

el *Índice Socioeconómico Ingreso-Educación-Ocupación* (ISE-YEO), presentado a continuación en la Ecuación 1.

$$ISE - YEO = \frac{Y_{TOTH}}{n^{0,7}} \cdot E_{PSH} \cdot O_{PSH} \qquad (1)$$

donde, el primer término Y_{TOTH} corresponde al ingreso per cápita equivalente según el tamaño del hogar ajustado según un factor de economía de escala del hogar ($n^{0,7}$). Luego, el segundo término E_{PSH} corresponde a un factor según el nivel de educación formal alcanzada por el principal sostenedor del hogar. Finalmente, el tercer término O_{PSH} corresponde a un factor según la ocupación del principal sostenedor del hogar. La metodología se encuentra en el documento "Clasificación grupos socioeconómicos y manual de aplicación" (AIM, 2019) y se presenta un resumen de su metodología en el Anexo .

En segundo lugar, se investigó sobre el *Índice Socio Material Territorial* (ISMT), desarrollado por el Observatorio de Ciudades UC. Este índice utiliza las variables escolaridad, hacinamiento, allegamiento y materialidad de la vivienda para su cálculo. En su página web (IDE Observatorio de Ciudades UC, 2019) y en el Anexo se presenta un resumen de la metodología utilizada. Este índice se ve muy útil ya que presenta información por quintiles. Las categorizaciones del ISMT son alto, medio, bajo, AB, C1, C2, C3, D y E.

Por último, se considera la opción de calibrar un modelo de predicción de ingreso del Censo usando la información de ingreso de la encuesta Casen a partir de *Machine Learning, Deep Learning* o regresión lineal. Esto implica imputar ingresos al Censo 2017 utilizando la información disponible desde la encuesta Casen. Estudios como el de Heldt et al. (2018) afirman que la imputación es una opción viable para abordar la falta de una variable de elección importante en grandes fuentes de datos como los Censos. A continuación, se describe el método, nivel de agregación, variables explicativas utilizadas y principales conclusiones de las imputaciones realizadas revisadas en la literatura en las Tabla 1, Tabla 2 y Tabla 3.

- Heldt et al. (2018)

Tabla 1: Resumen imputación ingresos Heldt et al., 2018

Método	Modelo de regresión ordenada con los datos del *Zensus* y *Mikrozensus* de Alemania 2011
Nivel de agregación	Hogares: Se define al jefe de hogar como la persona que se supone que decide dónde localizarse, que en este modelo es la persona empleada de mayor edad si uno o más miembros están trabajando, o la persona de mayor edad si ningún miembro del hogar está trabajando
Variables explicativas	Thresholds: 1, 2, 3, Tipo de trabajo del Representante del Hogar (Household representative), Tipo de trabajo del segundo Representante del Hogar (Household representative), Si el hogar es de una sola persona, Si el

	Representante del Hogar tiene o no antecedentes de inmigración, Número de miembros por grupos de edad: [<18, 18-30, 31-50, 51-64, >64], entre otros.
Conclusiones de autor	El modelo de imputación con regresión ordenada es una herramienta eficaz para abordar la falta de datos de ingresos en los Censos. El modelo permite estimar el ingreso de los hogares con precisión, lo que es esencial para la estimación de modelos de localización residencial.

- Tarozzi and Deaton, (2009)

Tabla 2: Resumen imputación ingresos Tarozzi y Deaton, 2009

Método	Usan un extracto del Censo México 2000, limitando el análisis a la sección rural de tres de los estados mexicanos más grandes: Chiapas, Oaxaca y Veracruz. Cada estado se subdivide en un gran número de municipios. Generan un pseudo-censo con esos datos reemplazando cada observación en el extracto con réplicas idénticas en número idéntico al peso (entero) proporcionado en el conjunto de datos, no poseen identificadores para las zonas. Comparan modelos econométricos para imputar ingresos faltantes.
Nivel de agregación	Hogares y municipios
Variables explicativas	Jefe de hogar sabe leer y escribir, es una mujer, pertenece a grupo indígena, trabaja en agricultura/pesca/silvicultura/minería y está trabajando, Edad del jefe de hogar, Habla sólo lengua indígena, Habla tanto lengua indígena como español, Número de miembros del hogar de 0 a 12 años, mayores de 65 años, Número de miembros masculinos de 13 a 65 años, Número de miembros femeninos de 13 a 65 años. Acceso a la electricidad, Posee refrigerador, Posee TV, Posee radio, Número de habitaciones, Acceso a baño dentro de la vivienda, El principal combustible para cocinar es la madera, La vivienda tiene suelo de tierra, El material primario de la vivienda es ladrillo o piedra, El material primario del techo es mampostería, concreto o teja.
Conclusiones de autores	Los métodos de imputación de datos censales pueden producir estimaciones de la pobreza y la desigualdad que son consistentes con las estimaciones de las encuestas, pero la precisión de las estimaciones depende del método utilizado y de las características de los datos.

- Yee et al. (2023)

Tabla 3: Resumen imputación de ingresos Yee et al., 2023

Método	Modelo híbrido K-means y regresión lineal múltiple (MLR) en datos de la encuesta de ingresos de los hogares (HIS) 2012 de Malasia
Nivel de agregación	Hogares
Variables explicativas	Estrato, edad, género, ocupación y estado civil del jefe de hogar, tamaño del hogar y región.

Conclusiones de autores	Los resultados del estudio muestran que el modelo híbrido supera al modelo de regresión lineal múltiple solo, en la estimación del ingreso familiar. Esto se debe a que el uso de K-means para agrupar a los hogares antes de aplicar la regresión lineal múltiple mejora la precisión de la estimación.

Estos trabajos, al igual que las metodologías revisadas anteriormente, reafirman que para la categorización socioeconómica de los hogares, variables como la educación, el trabajo del jefe de hogar y otras variables de la materialidad de la vivienda son variables relevantes a la hora de realizar la imputación de ingresos.

Métodos de desagregación de datos

En la literatura sobre desagregación espacial de hogares, se destacan métodos como la Población Sintética y el Procedimiento de Ajuste Proporcional Iterativo (IPF) y otros métodos análogos. La metodología de población sintética, según la explicación de Müller y Axhausen (2010), implica la generación de conjuntos de datos para simular poblaciones en microsimulaciones, creando individuos y hogares ficticios que reflejan las características estadísticas de una población real. En paralelo, el método de Ajuste Proporcional Iterativo (IPF) (Lomax & Norman, 2016) ha sido ampliamente empleado como una forma de generar una distribución de valores que cumpla con un conjunto de restricciones de totales mediante la maximización de la entropía. Este método ha sido usado para la creación de poblaciones sintéticas y microsimulaciones. En estudios de microdatos, Birkin y Clarke, (1988) utilizaron IPF para estimar características de residentes en áreas geográficas pequeñas, mientras que Rees (1994) lo usó para actualizar la estructura por edad y sexo de poblaciones en áreas reducidas.

Además, se han explorado métodos análogos como el prorrateo y la proporción, tal como describen Rees, Norman y Brown (2004), los cuales garantizan la coherencia de los datos entre diferentes escalas geográficas. Otros enfoques, como los algoritmos de Hill Climbing utilizados por Kurban et al. (2011) para la creación de tabulaciones cruzadas de hogares cuando solo se dispone de distribuciones univariantes, o el método optimización combinatoria empleado por Ryan, Maoh y Kanaroglou (2009), que construye poblaciones sintéticas intercambiando individuos hasta que coinciden con una distribución observada, también han sido objeto de estudio.

Resumen

El capítulo de revisión bibliográfica sienta las bases teóricas y conceptuales fundamentales para el desarrollo del trabajo de tesis centrado en la integración y el enriquecimiento de una base de datos. A lo largo del capítulo se exploran conceptos clave y se revisa la literatura relacionada con el modelo, las bases de datos urbanas y el tratamiento/procesamiento de datos en el contexto de uso de suelo.

La experiencia internacional muestra que la fusión e integración de datos, junto con la imputación de ingresos, constituyen elementos cruciales en la construcción de la base de datos para el modelo. Métodos de *Machine Learning*, especialmente Random Forest y K-Means, se identifican como herramientas valiosas para el procesamiento de datos urbanos. Así mismo, la imputación de ingresos en Censos se presenta como un desafío abordable mediante diversos enfoques para modelos de ubicación residencial. En cuanto a la desagregación espacial, se espera abordar esta cuestión en los siguientes apartados del trabajo, profundizando en estrategias específicas para optimizar la representación geográfica de los datos y mejorar la precisión de la estimación del modelo.

Capítulo 3
Bases de datos y Metodología de procesamiento

En este capítulo, se presentan y describen cada una de las bases de datos y métodos necesarios para comprender y llevar a cabo este trabajo de tesis. En primer lugar, es relevante identificar qué fuentes de datos pueden ser útiles para la modelación de la ciudad, considerando fuentes, de preferencia públicas, de datos periódicos. A continuación, se detalla la metodología de integración de las bases de datos identificadas. Posteriormente, se describe en detalle la información que nos entregan las bases que se trabajarán específicamente en este trabajo de tesis. Finalmente, se presentan en detalle los modelos necesarios para llevar a cabo esta fusión.

El trabajo de esta tesis se llevará a cabo siguiendo un enfoque metodológico cuantitativo, seleccionado principalmente debido a la naturaleza y magnitud de los datos utilizados para describir y analizar el uso de suelo urbano. Este enfoque, centrado en el manejo de grandes volúmenes de datos numéricos y estadísticos, es fundamental para obtener una representación precisa y detallada de las dinámicas urbanas. A través del análisis cuantitativo, se buscará transformar extensos conjuntos de datos en información significativa y aplicable para los modelos de uso de suelo y transporte. Este proceso es esencial para construir y analizar eficientemente la base de datos, y para responder a las preguntas de investigación y a los objetivos planteados. En contraste, esta tesis se aleja del enfoque metodológico cualitativo, que prioriza la interpretación subjetiva y descriptiva.

Identificación de Fuentes de datos

Para la aplicación del Modelo de Uso de Suelo de Santiago (MUSSA), se debe tener en cuenta su interacción con otros modelos y los modelos que este mismo necesita. Así, a continuación, se presenta una discusión de las interacciones entre modelos, las bases de datos necesarias para el modelo y el trabajo realizado en estas para buscar una calibración del modelo MUSSA.

Interacción entre el Modelo de uso de suelo y el Modelo de Transporte

En primer lugar, se debe comprender que el modelo de uso de suelo utiliza medidas de acceso espacial, típicamente de los tiempos de viaje de las personas, los cuales se obtienen de la aplicación de un modelo de transporte. A su vez, el modelo de transporte depende de las localizaciones, las cuales provienen del modelo de uso de suelo. Así, se puede simular el desarrollo de una ciudad a partir de la interacción de estos. En el caso de la ciudad de Santiago, el modelo de transporte ESTRAUS podría interactuar con el modelo MUSSA, produciendo cada uno resultados utilizables por el otro modelo.

El modelo de uso de suelo consta de un total de 6 sub-modelos. Modelos de Precios, de Oferta y de Localización, cada uno de estos tanto para agentes residenciales como no residenciales. Estos sub-modelos se encuentran relacionados entre sí y además, existen interacciones adicionales por la presencia de atributos endógenos, dependientes de localización y oferta. Por tanto, para el modelo se requieren datos de información de los bienes inmuebles en la ciudad, de sus usos y precios. En particular, para los usos residenciales, es crucial conocer las características socioeconómicas de los agentes que los utilizan y las características del transporte y entorno de cada zona de la ciudad.

Identificación de las potenciales fuentes de datos

La obtención de datos desempeña un papel fundamental en este trabajo de tesis, ya que es esencial para la creación de una base de datos de uso de suelo y transporte completa, con la mejor información disponible. En este sentido, se ha llevado a cabo una identificación de las fuentes de información más importantes disponibles para todas las ciudades de Chile, las cuales se presentan a continuación.

Respecto al modelo de localización residencial, existe información del Censo de Población y Vivienda, así como la encuesta Casen. Estas fuentes proporcionan una base sólida para la localización residencial de la población, de los hogares y de sus características socioeconómicas. En el caso del modelo de localización no residencial, se encuentran datos valiosos de este en el Catastro de Bienes Raíces del Sistema de Impuestos Internos (SII). Esta fuente proporciona las localizaciones principales de los diferentes destinos económicos (usos) de los bienes inmuebles no residenciales.

Luego, para los modelos de oferta, tanto residenciales como no residenciales, nuevamente el Catastro de Bienes Raíces del SII es una fuente crucial, proporcionando información detallada de los bienes inmuebles disponibles en el país. Esta incluye datos como el rol y líneas de construcción de las unidades de oferta residencial y no residencial, así como otras características importantes, como, el formato (casa o depto.) y las superficies de terreno y de construcción.

Posteriormente, al abordar los modelos de precios, en el contexto de los precios residenciales, existen fuentes de datos como las avaluaciones fiscales de viviendas disponibles en el Catastro de Bienes Raíces del SII. También se dispone de información de transacciones proporcionada por el mismo organismo, que incluye datos como el rol de las unidades, junto con el precio y la fecha de transacción. Cabe resaltar que estas dos fuentes de datos son igualmente pertinentes y aplicables para el modelo de precios no residenciales, ofreciendo una base amplia y detallada para el análisis de precios en distintos tipos de bienes inmuebles.

Por otro lado, se reconoce la importancia de disponer de información detallada sobre la accesibilidad y características del entorno de cada zona. Para ello, se pueden utilizar fuentes de datos de transporte, como los tiempos de viaje en transporte público y cantidad de paraderos proporcionados por ADATRAP. Asimismo, los tiempos de viaje en transporte privado obtenidos por compañías telefónicas. Las generaciones de viajes entre zonas, conteos vehiculares y mediciones de tasa de ocupación son proporcionadas por las Encuestas Origen-Destino (EOD). Por último, la información de imágenes y mapas ayuda a comprender el entorno, incluyendo variables de detección de objetos como áreas verdes, superficies de construcción, entre otros aspectos relevantes.

En resumen, para el proyecto se cuenta con ocho fuentes de datos claves: el Censo de población y vivienda, las encuestas Casen, el Catastro de Bienes Raíces del SII, las transacciones registradas en el SII, datos de imágenes, las Encuestas Origen-Destino, los datos de ADATRAP y datos de compañías telefónicas. Cada una de estas fuentes son esenciales para llevar a cabo el análisis de manera integral y precisa. Por lo tanto, el análisis detallado de los datos disponibles de las fuentes de datos más importantes, incluyendo sus limitaciones y nivel de desagregación geográfico, se detallan más adelante en el estudio.

Diseño de la Base de Datos

En el contexto de este estudio, a partir de las bases de datos presentadas anteriormente se ha identificado una estrategia metodológica eficiente para fusionar los datos de estas diversas fuentes y construir una base de microdatos integrados, mostrada en la Figura 2. A continuación, se explicará a detalle cada una de las sub-bases necesarias para obtener la base de datos integrada de uso de suelo y transporte. Antes de abordar esto, es fundamental destacar que el área de estudio se ha definido a partir de la cartografía del Censo, enfocándose exclusivamente en el ámbito urbano, específicamente en las manzanas censales urbanas. Esta elección se debe a que la información disponible en las bases de datos, como el catastro y datos de transporte, se centran en áreas urbanas. Además, la investigación sobre una ciudad se centra en esta área de interés. No obstante, es relevante mencionar que existe la posibilidad de expandir el análisis en un futuro, considerando el Catastro de Bienes Inmuebles rural y la planimetría rural proporcionada por el INE.

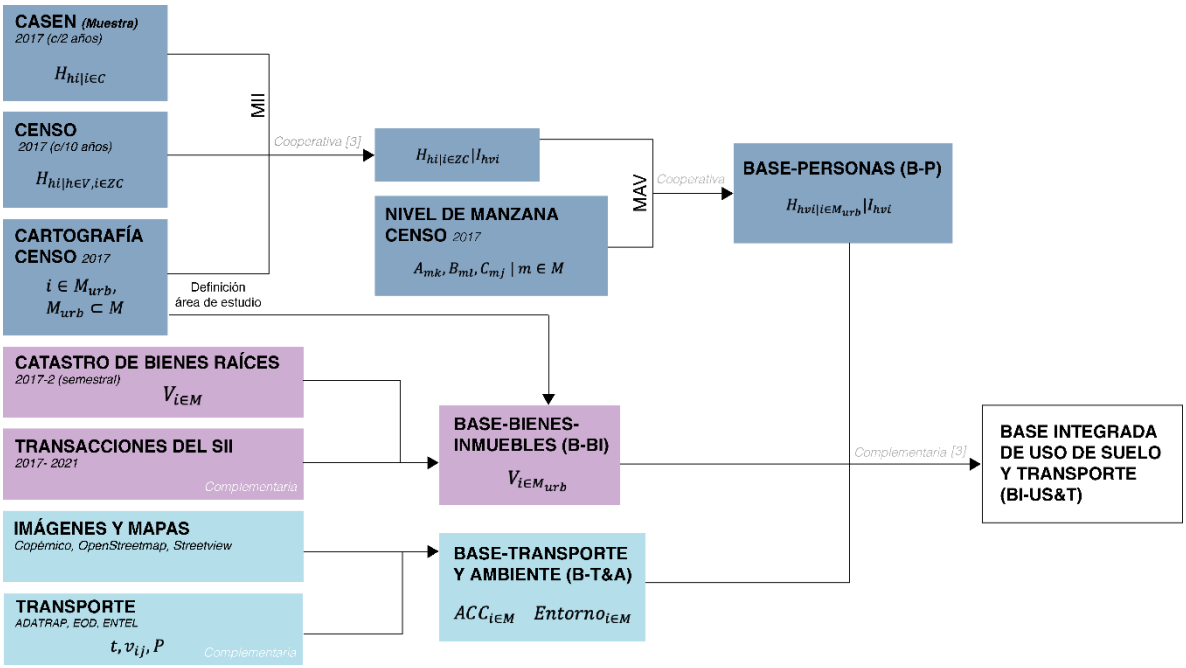


Figura 2: Diagrama de fusión de datos

Base de Personas (B-P)

Se busca construir en primer lugar una base censal, es decir, con información de todos los agentes residenciales de la ciudad, desagregada espacialmente a nivel de manzana censal, con el objetivo de conocer granularmente las decisiones de localización de los agentes con características socioeconómicas diferentes. Dentro de estas características, los ingresos son relevantes para diferenciarlos, como se mencionó en el capítulo anterior. Por esto, la metodología a seguir es enriquecer el Censo con información de otras fuentes, para obtener una base a nivel de manzanas y con nuevas variables no disponibles en el Censo, como lo son los ingresos.

La base de datos del Censo proporciona información desagregada a nivel de personas para los agentes residenciales y, geográficamente, está disponible a nivel de zona censal, que es más agregada que a nivel de manzana. Más adelante se detallarán las dimensiones de estas variables. La encuesta Casen posee variables comunes con el Censo además del ingreso de cada persona, pero solo para una muestra de la población y se encuentra georreferenciada a nivel de comuna. Se propone utilizar la Casen para enriquecer el Censo, imputando datos de ingreso al Censo mediante un Modelo de Imputación de Ingresos (MII) a nivel de personas o hogares, cuya calidad de predicción será validada estadísticamente.

Posteriormente, se propone aplicar un Modelo de Asignación de Viviendas (MAV), utilizando la base de datos del Censo a nivel de manzana y los resultados obtenidos del MII, lo que permitirá tener esta información a un nivel de detalle mayor, específicamente a nivel

de manzana censal. La base de datos del Censo a nivel de manzanas posee información de distribuciones marginales a nivel de manzana censal de diferentes atributos de personas, hogares y viviendas. La nueva base de datos con los procesos MII y MAV se denominará BASE-PERSONAS (B-P). Más adelante en este capítulo, se detallará a profundidad cada uno de los procesos y bases de datos requeridos para la construcción de esta base de datos.

Base de Bienes Inmuebles (B-BI)

Esta base busca describir las características de los bienes inmuebles y su entorno a nivel de manzana censal, para saber la oferta de bienes inmuebles disponibles en la ciudad, sus usos y precios. Para esto, el Catastro de Bienes Raíces proporciona el rol y líneas de construcción de las unidades de oferta residencial y no residencial. El rol denota un bien inmueble, un espacio que posee un único dueño y las líneas de construcción se diferencian por estructuras de construcción separadas. También incluye el destino económico, avaluación fiscal y otras características, como la superficie de terreno y de construcción. Mediante un procesamiento de webscrapping fue posible obtener las coordenadas geográficas de los bienes inmuebles, lo cual permite georreferenciarlos por manzana censal urbana.

Otra base de datos del SII es la de Transacciones, esta conforma una parte adicional del Catastro de Bienes Raíces que por tanto contiene el mismo rol de las unidades, pero incluye el precio (en UF y/o pesos chilenos) y la fecha de transacción desde el año 2017 a 2021 para los bienes inmuebles vendidos e inscritos en el Conservador de Bienes Raíces durante ese período. En esta segunda base, se debe ajustar el precio de las transacciones a UF, para que sean comparables debido a la diferencia de temporalidad de estas. Finalmente, la integración de estas dos bases se realiza a nivel de rol de los bienes inmuebles. Los desafíos de esta base incluyen que no se tiene transacciones de todos los bienes inmuebles y las avaluaciones fiscales no son representativas de las transacciones que realmente ocurren. La integración geo-referenciada de estas bases generará la BASE-BIENES-INMUEBLES (B-BI), con datos a nivel de cada predio conteniendo el arreglo de variables de las diversas fuentes.

Base de Transporte y Ambiente (B-T&A)

La base de transporte y ambiente busca describir la accesibilidad y el entorno de las localizaciones para comprender las características del barrio y cómo estas afectan en las localizaciones de los agentes. Para esto, existen datos de diferentes imágenes y mapas que se pueden analizar para integrar variables de detección de objetos, como áreas verdes, superficie de construcción, altura de edificios, entre otros. La disponibilidad de imágenes satelitales con cada vez más resolución o de las calles en Google StreetView, junto con las herramientas computacionales que existen en la actualidad para procesar estas, como redes neuronales y reconocimiento automático, permiten realizar procesamientos de imágenes que permitiría obtener información del entorno de los bienes inmuebles y las localizaciones.

Las bases de datos de transporte son esenciales en el análisis, ya que proporcionan información sobre los tiempos de viaje en la red de transporte público (en el caso de ADATRAP) y privado (en el caso de la base de datos de telefonía Entel). La integración de estas bases permitirá asignar a la B-BI y B-P atributos de accesibilidad y atractividad de cada registro. ADATRAP posee información de diferentes variables de transporte público, como viajes OD , tiempo de viaje, modo de viaje, propósito de viaje, paraderos georreferenciados, frecuencia de servicios y número de servicios. Con esta información se pueden construir diferentes indicadores de accesibilidad como la cantidad de paraderos por manzana censal o la distancia al paradero más cercano, además del tiempo de viaje promedio desde la zona o para llegar a la zona, entre otros que se deben investigar.

Las empresas telefónicas a través de las torres telecomunicaciones tienen información de la localización geoespacial de los usuarios, empresas como Entel han trabajado estos datos en busca de construir matrices de movilidad con sus propósitos y modo de viaje. La dificultad de esto radica en la privacidad de estos datos, a diferencia de todos los otros trabajados hasta el momento, ya que una sola empresa de teléfonos tiene información de solo una parte de la población y esto podría ser información sesgada. Otras fuentes como la Encuestas Origen destino pueden ser útiles para determinar los modos de transporte y matrices de viajes, a pesar de ello estas tienen problemas de periodicidad. Estas fuentes de transporte y entorno geo-referenciadas a nivel de manzana censal generarán la BASE-TRANSPORTE Y AMBIENTE (B-T&A).

Base Integrada de Uso de Suelo y Transporte (BI-US&T)

Para comprender el proceso de integración de datos en este proyecto descrito anteriormente y obtener la base de datos integrada de uso de suelo y transporte, se ha presentado el diagrama de fusión de datos en la Figura 2. El proceso de enriquecimiento del Censo de Población y Vivienda, la base de datos principal, se realiza a través de diversos procesos e inclusión de datos adicionales, lo cual se explica en referencia a la Figura 3. En esta figura p denota personas, h hogares, v viviendas, mc manzanas censales y zc zonas censales.

Las variables originales del Censo representan información de las personas (z_p), con 39 variables para 6.825.907 personas, hogares (z_h) con 14 variables para 2.157.337 hogares, viviendas (z_v) con 20 variables para 2.281.760 viviendas y la georreferenciación de estas últimas a nivel de zona censal (ZC_v) abarcando 1.864 zonas censales urbanas. En un primer paso, el modelo de imputación de ingresos (MII) incorpora la variable de ingresos a las personas y/o hogares (I_p, I_h). Posteriormente, el modelo de asignación de viviendas (MAV) aumenta el nivel de precisión de la variable de georreferenciación de las viviendas, ahora a un nivel de manzana censal (MC_v), cubriendo 48.809 manzanas censales urbanas. Así, se logra georreferenciar las variables de personas, hogares y viviendas a un nivel considerablemente más granular.

Siguiendo este proceso, el modelo de imputación-asignación contribuye con variables adicionales a la vivienda, tales como los metros cuadrados (m_v^2) y los precios (p_v), en concordancia con la información del Catastro de Bienes Raíces del Sistema de Impuestos Internos. La información de imágenes satelitales o StreetView, introduce variables del entorno de las manzanas censales ($Entorno_{mc}$), como las áreas verdes, características de edificios como la altura, características de las veredas, entre otros aspectos. Finalmente, los datos de transporte agregan medidas de accesibilidad (ACC_{mc}), a cada manzana censal.

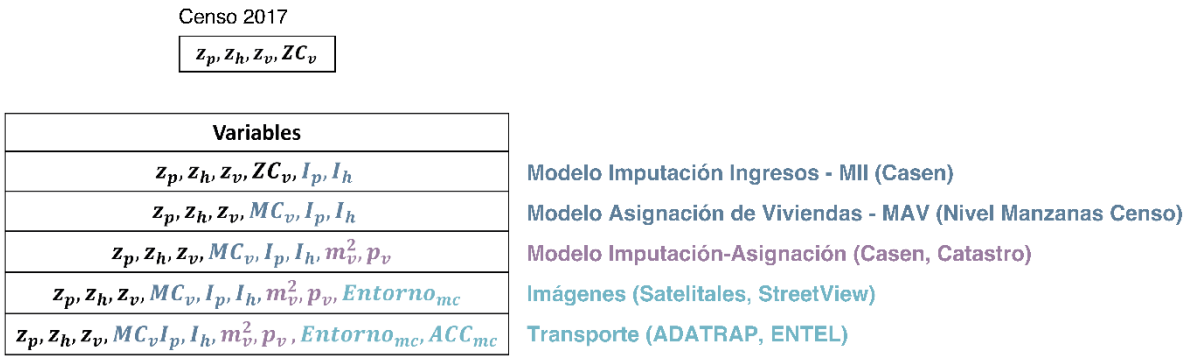


Figura 3: Diagrama de Enriquecimiento del Censo

Dimensionalidad de una base de datos desagregada de uso de suelo y transporte

Para el desarrollo eficiente de una base de datos integrada que abarque aspectos de uso de suelo y transporte, que sea censal y desagregada espacialmente, es crucial estimar su dimensionalidad potencial. Esta estimación no solo facilita la comprensión del volumen de datos a manejar, sino que también orienta las decisiones relativas al almacenamiento y procesamiento de datos.

En el contexto de esta investigación, las dimensiones consideradas se dividen en tres categorías. En primer lugar, a nivel espacial para el área urbana de la ciudad de Santiago, de un total de 985 km cuadrados aproximadamente, se desagrega a nivel de manzana censal, con un total de 48.809 manzanas (MC). Para estas manzanas, se estima un máximo de 20 variables relacionadas con el entorno, el transporte y/o accesibilidad (V), considerando aspectos como paraderos, viajes, tiempos, frecuencia, modo, propósito. En segundo lugar, la dimensión a nivel de bienes inmuebles incluye los bienes inmuebles no residenciales del Catastro de Bienes Raíces y las viviendas del Censo, con un total aproximado de 3.300.000 bienes inmuebles (BI). Para estos, se contemplan 30 atributos (A), incluyendo el destino económico, superficie de terreno, superficie construida, precios y características de las viviendas. Finalmente, en la dimensión de agentes residenciales, se cuenta con información para cada una de las 6.825.907 personas de la ciudad (P) y aproximadamente 55 características socioeconómicas de ellas y del hogar al que pertenecen (C) proporcionadas por el Censo.

Considerando que para un computador de 64 bits cada dato usa 8 bytes, la Tabla 4 presenta la magnitud de cada una de estas bases (PxC, BIxA y MCxV) y el tamaño en gigabytes de cada una de estas, además, de la base final. La magnitud de la base de datos tiene directas implicancias en la capacidad de almacenamiento y rendimiento de consultas; por ello, el requerimiento se calcula multiplicando la magnitud por los 8 bytes que requiere cada dato. Es importante destacar que este cálculo representa el requerimiento para almacenar la base de datos integrada, pero también se debe considerar el almacenamiento y procesamiento de las bases de datos originales, para las cuales se puede aplicar un cálculo de estimación similar.

Tabla 4: Dimensionalidad de la potencial base de datos

	Cantidad	Variables	Magnitud ($\cdot 10^7$)	Tamaño de la Base de Datos (GigaBytes)
Personas (B-P)	6.800.000	55	37,4	2,99
Bienes Inmuebles (B-BI)	3.300.000	30	9,90	0,790
Manzanas Censales (B-T&A)	48.809	20	0,0976	0,0781
B-US&T (total)	-	-	-	3,86

Descripción del contenido de las bases de datos a trabajar

En este apartado se describe la información básica a conocer de las bases de datos a trabajar, Censo de Población y Vivienda 2017 y encuesta Casen. Es importante recordar que en este trabajo de investigación se aplicará y trabajarán los datos de la Base-Personas, es decir, los primeros dos modelos de enriquecimiento del Censo. Esta limitación responde a los alcances definidos previamente en la introducción.

1. Censo de Población y Vivienda 2017

A continuación, se presenta un extracto de la información relevante del Censo para el procesamiento de los datos en esta investigación. Esta información es entregada por el Instituto Nacional de Estadísticas (INE) en el documento del Manual de Usuario de la base de datos del Censo de Población y Vivienda 2017 (INE, 2018) y se resume en:

> El objetivo principal del Censo es producir información sociodemográfica básica que actualice los datos sobre el tamaño de la población, su composición y distribución territorial, así como de los hogares y las viviendas existentes en el país, que permita proveer información para las estimaciones y proyecciones de población y para la conformación de un marco muestral base para las encuestas de hogares.

Para cada una de estas unidades, el Censo buscó recopilar información que permitiera cumplir con los objetivos que se definieron.

a) Personas: corresponden a todos los individuos comprendidos en el Censo.
b) Hogares: corresponden a la manera de organización de las personas dentro de las viviendas particulares, las que corresponden a una o más personas que, unidas o no por un vínculo de parentesco, alojaron la noche del 18 al 19 de abril en una misma vivienda o parte de ella y se benefician de un mismo presupuesto para alimentación.
c) Viviendas: corresponden a los lugares de alojamiento, estructuralmente separados e independientes, en los que pueden residir las personas. Estas pueden ser viviendas particulares o viviendas colectivas (hospitales, conventos, internados, cuarteles, establecimientos correccionales, hoteles, pensiones y residenciales, entre otros).

El cuestionario principal fue el cuestionario **de viviendas particulares** que constó de 21 preguntas contenidas en tres secciones para abordar las unidades de análisis especificadas.

División geográfica del territorio nacional

El Censo 2017 entrega información relevante para la toma de decisiones a nivel país y para áreas geográficas más pequeñas. Para ello, el territorio nacional se divide a partir de la división político-administrativa del país (en adelante DPA), de carácter legal, y la división censal, que es de ámbito operativo y permite obtener una desagregación a nivel de microdato.

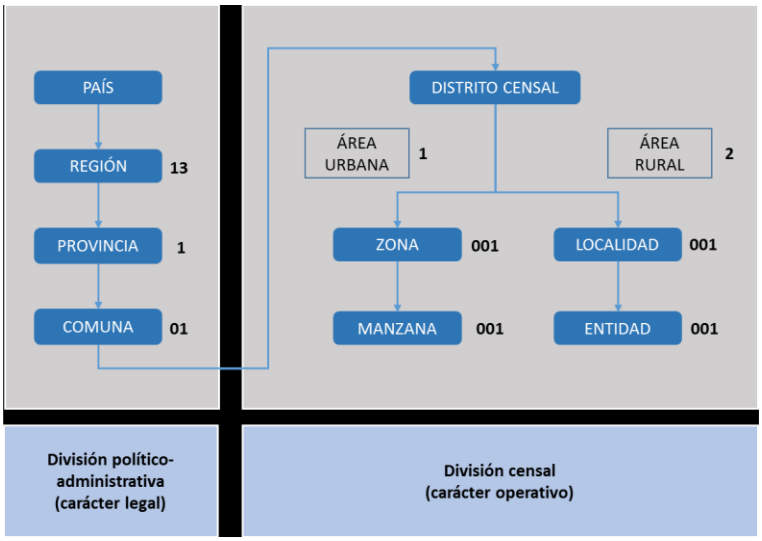


Figura 4: Diagrama territorial (INE, 2018)

1. **División político-administrativa:** para el cumplimiento de los objetivos de gobierno y administración, el país se divide en unidades territoriales menores llamadas regiones, las que se dividen, a su vez, en provincias y éstas, finalmente, en comunas. De acuerdo a la actual DPA, el país está conformado por 16 regiones, 56 provincias y 346 comunas.
2. **División censal:** para efectos operativos, las comunas se subdividen en unidades territoriales menores que permiten organizar de mejor forma el levantamiento del Censo. Los límites censales los define el INE y se encuentran enmarcados en los límites de la DPA. Por otro lado, su trazado no reviste un carácter legal.

   Esta división contempla:

   a) **Distrito censal:** es una división censal de carácter operativa definida por el INE. Se define como la parte en que se divide el territorio comunal y que se constituye en la unidad básica mayor para las operaciones de terreno censales y de muestras estadísticas. La suma de los distritos censales en el país es de 2.771, los cuales pueden ser urbanos, rurales y mixtos. Para la división de los distritos, el criterio general es el número de viviendas en el área urbana y la superficie en la parte rural.
   b) **Área geográfica (urbana/rural):** corresponde a la división dentro de los distritos entre áreas urbanas y rurales que se expresa territorialmente a través del Límite Urbano Censal (LUC).
   Específicamente se entiende como entidad urbana un asentamiento humano con continuidad y concentración de construcciones con población mayor a 2.000 habitantes, o entre 1.001 y 2.000 habitantes donde menos del 50% de la población que declara haber trabajado se dedica a actividades primarias.
   c) **Zona censal:** corresponde a la división del distrito censal urbano y área urbana de los distritos censales mixtos, formada por un conglomerado de manzanas, cuya finalidad es facilitar la organización, control y levantamiento del Censo. Para el caso de la Región Metropolitana se tienen 1.864 zonas censales urbanas.
   d) **Localidad:** corresponde a un área geográfica con nombre propio de conocimiento generalizado. Para efectos de la base de datos de un Censo, corresponde a la división del distrito censal rural y las áreas rurales de los distritos censales mixtos.

Información a nivel de manzana/entidad

Considerando que hay usuarios que requieren disponer de la información a niveles geográficos menores que zona/localidad, el INE pone a disposición una base de datos de manzana/entidad con datos agregados que tiene las siguientes características:

1. La base de datos a nivel de manzana/entidad presenta cantidades totales en dicha área geográfica para datos de personas y viviendas.
2. Por temas de indeterminación, no se presentan las características de sexo, edad, pueblos indígenas y migración en las manzanas y entidades con tres personas o menos en total, con el fin de resguardar la confidencialidad de los datos. Estos datos fueron marcados con un asterisco.
3. Considera una variable llave (key) que permite incorporar esta información como una capa en la cartografía censal. Esa variable es ID_ MANZENT, que se presenta como se describe en la Figura 5. Para el caso de la Región Metropolitana se tienen 48.809 manzanas urbanas.

Figura 5: ejemplo llave ID_MANZENT

Bases de datos

La Base de Datos del Censo 2017 se encuentra disponible en dos formatos: REDATAM y CSV. Entre ambos no hay diferencias en los datos publicados, pero sí en relación con las características y la lógica de funcionamiento de cada programa.

La base de datos se estructura a partir de tres niveles de información, correspondientes a las unidades de análisis del Censo 2017: vivienda, hogar y persona. Para cada uno de ellos hay variables asociadas de acuerdo con los contenidos del cuestionario censal.

Se tienen 20 variables de viviendas, 14 variables de hogares y 39 variables de personas. Para el caso de la región metropolitana se tienen 2.281.760 viviendas, 2.157.337 hogares y 6.825.907 personas en las zonas censales urbanas. A continuación, se presenta el detalle para cada comuna en la Tabla 5.

Tabla 5: Zonas Censales, viviendas y personas a nivel de comuna Censo

Nombre de la comuna	Código comuna	Nro. ZC	Nro. Viv	Nro. Per	Nombre de la comuna	Código comuna	Nro. ZC	Nro. Viv	Nro. Per
ALHUÉ	**13502**	1	1.083	2.706	**VITACURA**	**13132**	26	31.753	85.300
MARÍA PINTO	**13504**	6	2.257	6.656	**LA GRANJA**	**13111**	30	34.343	116.312
PIRQUE	**13202**	4	3.488	11.514	**COLINA**	**13301**	36	35.226	117.839
CALERA DE TANGO	**13403**	9	3.497	11.488	**INDEPENDENCIA**	**13108**	26	36.610	100.059
SAN JOSÉ DE MAIPO	**13203**	5	3.998	11.121	**CONCHALÍ**	**13104**	39	37.721	126.800
TILTIL	**13303**	5	4.336	13.057	**CERRO NAVIA**	**13103**	36	37.963	132.401

CURACAVÍ	**13503**	8	6.671	20.276	**QUINTA NORMAL**	**13126**	29	38.914	109.784
ISLA DE MAIPO	**13603**	8	8.563	26.910	**SAN MIGUEL**	**13130**	24	42.892	107.828
EL MONTE	**13602**	9	9.760	29.998	**MACUL**	**13118**	33	42.998	116.249
PAINE	**13404**	16	14.427	46.352	**RENCA**	**13128**	44	43.128	146.987
TALAGANTE	**13601**	15	18.367	59.097	**EL BOSQUE**	**13105**	39	47.912	162.415
PADRE HURTADO	**13604**	11	18.378	55.561	**LA PINTANA**	**13112**	48	49.655	176.105
SAN RAMÓN	**13131**	22	23.743	82.602	**RECOLETA**	**13127**	45	50.136	157.569
CERRILLOS	**13102**	21	24.523	80.710	**ESTACIÓN CENTRAL**	**13106**	43	52.413	140.809
PEÑAFLOR	**13605**	21	26.327	82.959	**QUILICURA**	**13125**	45	62.228	209.676
LAMPA	**13302**	20	26.339	80.683	**PUDAHUEL**	**13124**	62	67.463	222.754
LO ESPEJO	**13116**	25	26.744	98.651	**PEÑALOLÉN**	**13122**	56	70.322	241.394
BUIN	**13402**	21	26.753	82.338	**PROVIDENCIA**	**13123**	51	70.926	141.986
MELIPILLA	**13501**	26	28.725	84.286	**SAN BERNARDO**	**13401**	68	88.422	295.550
HUECHURABA	**13107**	21	28.905	98.500	**ÑUÑOA**	**13120**	61	92.179	208.048
LO BARNECHEA	**13115**	22	29.001	103.092	**LAS CONDES**	**13114**	82	117.878	294.480
LO PRADO	**13117**	31	29.417	95.901	**LA FLORIDA**	**13110**	107	119.864	366.376
LA REINA	**13113**	28	29.771	92.678	**MAIPÚ**	**13119**	129	160.126	517.393
PEDRO AGUIRRE CERDA	**13121**	28	29.874	101.035	**PUENTE ALTO**	**13201**	142	171.077	566.561
SAN JOAQUÍN	**13129**	28	30.054	94.325	**SANTIAGO**	**13101**	129	193.215	402.847
LA CISTERNA	**13109**	23	31.395	89.889	**Total**	-	1864	2.281.760	6.825.907

2. Encuesta Casen

A continuación, se presenta un extracto de la información principal de la encuesta Casen desarrollada a nivel nacional por el Ministerio de Desarrollo Social y presentada en el documento Libro de Códigos Base de Datos (Ministerio de Desarrollo Social, 2018) que se resume en:

> Uno de los objetivos fundamentales de la Encuesta Casen es conocer periódicamente la situación de los hogares y de la población, especialmente de aquélla en situación de pobreza y de aquellos grupos definidos como prioritarios por la política social (infancia, juventud, adultos mayores, mujeres, pueblos indígenas, personas en situación de discapacidad, nacidos fuera de chile, entre otros), principalmente con relación a aspectos demográficos, de educación, salud, vivienda, trabajo e ingresos. En particular, estimar la magnitud de la pobreza y la distribución del ingreso; identificar carencias y demandas de la población en las áreas señaladas; y evaluar las distintas brechas que separan a los diferentes segmentos sociales y ámbitos territoriales.
>
> La Encuesta Casen tiene como objeto de estudio los hogares que habitan las viviendas particulares que se ubican en el territorio nacional, exceptuando algunas zonas muy alejadas o de difícil acceso (ADA), así como las personas que forman parte de esos hogares. Al interior de cada vivienda seleccionada, se intenta entrevistar a todos los hogares y recoger información de todas las personas que sean miembros del mismo.

Además del contexto nacional, la Encuesta Casen 2017 considera como dominios de estudio las regiones y las áreas geográficas urbano y rural (según marcos muestrales del Instituto Nacional de Estadísticas). La siguiente ficha resume los antecedentes técnicos de la encuesta Casen 2017:

Tamaños de Unidades de Análisis, Casen 2017	**Muestral**
Número de viviendas	68.466 viviendas
Número de hogares	70.948 hogares
Número de personas	216.439 personas
Número de variables	806 variables

Para el caso de la Región Metropolitana se tienen 12.772 viviendas, 13.530 hogares y 42.601 personas. A continuación, se presenta el detalle para cada comuna en la Tabla 6.

Tabla 6: Viviendas y personas a nivel de comuna Encuesta Casen

Nombre de comuna	Código comuna	Nro. Viv	Nro. Per	Nombre de comuna	Código comuna	Nro. Viv	Nro. Per
San Pedro	**13505**	29	91	**San Joaquín**	**13129**	180	595
Pirque	**13202**	58	187	**Cerro Navia**	**13103**	188	686
San José de Maipo	**13203**	77	243	**La Reina**	**13113**	189	605
Curacaví	**13503**	80	255	**Conchalí**	**13104**	193	728
Alhué	**13502**	82	256	**Macul**	**13118**	194	645
Padre Hurtado	**13604**	84	304	**Renca**	**13128**	202	726
Calera de Tango	**13403**	85	352	**La Granja**	**13111**	213	803
El Monte	**13602**	85	307	**Lampa**	**13302**	214	818
María Pinto	**13504**	86	307	**Estación Central**	**13106**	226	796
Isla de Maipo	**13603**	94	335	**Quilicura**	**13125**	229	845
Cerrillos	**13102**	96	354	**Lo Barnechea**	**13115**	243	999
Huechuraba	**13107**	106	429	**Colina**	**13301**	245	856
Talagante	**13601**	106	341	**Recoleta**	**13127**	254	884
Buin	**13402**	108	392	**El Bosque**	**13105**	258	978
Tiltil	**13303**	114	407	**La Pintana**	**13112**	258	988
Paine	**13404**	118	420	**Pudahuel**	**13124**	259	935
Peñaflor	**13605**	130	469	**Vitacura**	**13132**	272	892
Lo Espejo	**13116**	131	594	**Peñalolén**	**13122**	399	1.458
San Ramón	**13131**	131	485	**San Bernardo**	**13401**	429	1.569
Independencia	**13108**	138	469	**Ñuñoa**	**13120**	598	1.669
Quinta Normal	**13126**	148	563	**La Florida**	**13110**	672	2.420
Pedro Aguirre Cerda	**13121**	149	561	**Puente Alto**	**13201**	677	2.317
La Cisterna	**13109**	154	546	**Las Condes**	**13114**	735	2.171
Lo Prado	**13117**	158	543	**Maipú**	**13119**	771	2.571
Melipilla	**13501**	158	507	**Santiago**	**13101**	829	2.093
San Miguel	**13130**	171	531	**Providencia**	**13123**	969	2.306
				Total	-	12.272	42.601

Metodología del procesamiento de las bases de datos

A continuación, se presenta detalladamente la base teórica y metodológica de los modelos utilizados para el desarrollo de la tesis. En primer lugar, el Modelo de Imputación de Ingresos (MII) y, en segundo lugar, el Modelo de Asignación de Viviendas (MAV), ambos ya mencionados anteriormente.

Modelo de Imputación de Ingresos (MII)

El Modelo de imputación de Ingresos (MII) tiene como objetivo estimar el ingreso de un hogar h en la vivienda v, ubicada en la zona censal i de la comuna c, para cada clúster del conjunto de comunas Ω_c. El modelo genera estos clústeres mediante el método K-means, agrupando comunas con distribuciones de ingresos similares debido a la cantidad de datos disponibles. El método de imputación se realiza mediante la utilización de variables comunes entre el Censo de Población y Viviendas y la encuesta Casen. Se compara la eficiencia de ocho modelos de *Machine Learning,* así como dos métodos de imputación, cuyos detalles se describen a continuación. Cabe destacar que este método podría aplicarse a cualquier variable que se desee imputar desde una base que comparta variables que se puedan usar como pivote.

Para esta comparación, se evalúan dos métodos de imputación distintos: uno indirecto y uno directo a los hogares. Con esto se busca identificar cual es el mejor método de imputación de ingresos, el indirecto, que imputa a nivel de personas y luego agrega para obtener el ingreso del hogar, o bien imputando directamente a nivel de hogares. En el caso indirecto, se emplea una imputación individual, según lo descrito en la Ecuación 2, calibrando el modelo de imputación de ingreso con los datos de la encuesta Casen para cada persona p de la comuna c que pertenezca al clúster de comunas Ω_c. Se consideran variables tanto de las personas (x_p) como de la vivienda (x_v) presentes tanto en la encuesta Casen como en el Censo. En el segundo método, se implementa una imputación directa al hogar, presentada en la Ecuación 3. Este también se calibra con los datos de la encuesta Casen, pero para cada hogar h perteneciente a la comuna c que pertenezca al clúster de comunas Ω_c. En este caso, se tienen en cuenta variables del hogar (x_h) y variables de la vivienda (x_v). Además, ambos métodos incorporan una variable *dummy* para cada comuna (1_c), donde su valor el 1 si el hogar pertenece a la comuna c y 0 en el caso contrario.

$$I_{pvc} = \sum_{c \in \Omega_c} \theta_{\Omega_c} \cdot 1_c + \sum_{l} \theta_{\Omega_c l} \cdot x_{pl} + \sum_{k} \theta_{\Omega_c k} \cdot x_{vk} \rightarrow I_{pvc} = f_{\Omega_c}\left(1_{c \in \Omega_c},\, x_p,\, x_v | \theta\right) \qquad (2)$$

$$I_{hvc} = \sum_{c \in \Omega_c} \theta_{\Omega_c} \cdot 1_c + \sum_{l} \theta_{\Omega_c l} \cdot x_{hl} + \sum_{k} \theta_{\Omega_c k} \cdot x_{vk} \rightarrow I_{hvc} = f_{\Omega_c}\left(1_{c \in \Omega_c}, x_h, x_v \middle| \theta\right) \qquad (3)$$

Aunque las Ecuaciones 2 y 3 se presentan de manera lineal para explicar fácilmente la diferencia de cada caso, en la aplicación práctica se usa una función general para estimar los

parámetros θ. La obtención de estos parámetros de cada método se realizará comparando diferentes modelos de estimación de *Machine Learning*: *Linear Regressión (without regularization, with Ridge regularization, with Lasso regularization) (LR), Gradient Boosting Regression (GBR), Random Forest (RF), Neural Network (NN), Suport Vector Regression (SVR)*. La elección de cada método y modelo utilizado se sustenta en criterios de eficiencia, calidad y otros aspectos relevantes. Se utilizan indicadores de bondad de ajuste como *Root Mean Square Error* (RMSE) para respaldar y justificar dicha elección.

La siguiente etapa consiste en aplicar el modelo con los mejores resultados en la estimación y predicción del ingreso, pero esta vez con los datos del Censo. Esto se realiza para cada persona p o hogar h en la zona censal i de la comuna c que pertenezca al clúster Ω_c, utilizando las funciones estimadas en las Ecuaciones 2 y 3, pero ahora aplicadas con las variables del Censo como se describe en las Ecuaciones 4, 5 y 6. Aquí, 1_c es la variable dummy para cada comuna, z_p son las variables de las personas, z_h son las variables del hogar y z_v son las variables de la vivienda del Censo, comunes con la encuesta Casen. De esta manera, se obtiene e imputa el ingreso de los hogares presentes en la base de datos del Censo, ya sea mediante el método indirecto (Ecuaciones 4 y 5), para el cual se agregan los ingresos de cada persona p perteneciente al hogar h (Ecuación 5) para tener el ingreso del hogar, o con el método de imputación directa al hogar, como se muestra en la Ecuación 6. La determinación del método a utilizar realiza considerando los indicadores de bondad de ajuste.

$$I_{pvi} = f_{\Omega_c}\left(1_{c\in\Omega_c},\, z_p,\, z_v|\theta\right), \quad \forall i \in c \in \Omega_c \qquad (4)$$

$$I_{hvi} = \sum_{p} I_{pvi}\,, \; \forall p \in h \qquad (5)$$

$$I_{hvi} = f_{\Omega_c}\left(1_{c\,\in\,\Omega_c}, z_h, z_v\middle|\theta\right), \qquad \forall i \in c \in \Omega_c \qquad (6)$$

Modelo de Asignación de Viviendas (MAV)

Este es un modelo de desagregación espacial, que tiene como objetivo asignar las personas, hogares y viviendas a las que estos pertenecen, desde un nivel de zona censal en las manzanas censales dentro de esta. Esto permite obtener la información de las personas a un nivel de desagregación más detallado, pasando de 1.864 zonas censales a las 48.160 manzanas censales urbanas en la región metropolitana. Esta asignación se realiza mediante la información del propio Censo, usando datos de viviendas y las personas, que está a nivel de zona censal, y los datos adicionales de distribuciones marginales a nivel de manzana del Censo.

Se conoce el valor de estos atributos para cada una de las personas, hogares y viviendas del Censo, denotados como:

a_{hk} = valor del atributo k del hogar h
b_{pl} = valor del atributo l de la persona p
c_{vj} = valor del atributo j de la vivienda v

Además, se cuenta con el total para cada atributo a nivel de manzana que representan cantidades totales de diferentes características de personas, hogares y viviendas. Estos datos son comúnmente conocidos como las distribuciones marginales, que en este caso denotaremos:

A_{mk} = suma del atributo k sobre todos los hogares localizados en la manzana m
B_{ml} = suma del atributo l sobre todas las personas localizadas en la manzana m
C_{mj} = suma del atributo j sobre todas las viviendas localizadas en la manzana m

También se tiene información sobre cómo están relacionadas las viviendas con los hogares y las personas del Censo, representadas por:

$$\delta_{ph} = \begin{cases} 1 \text{ si la persona } p \text{ pertenece al hogar } h \\ 0 \text{ si no} \end{cases}$$
$$\delta_{hv} = \begin{cases} 1 \text{ el hogar } h \text{ pertenece a la vivienda } v \\ 0 \text{ si no} \end{cases}$$

El modelo, como se mencionó anteriormente, busca localizar cada vivienda v de una zona censal en una manzana m de esa zona, de manera que se cumplan las distribuciones marginales de los atributos en cada manzana. La variable incógnita se define como y_{vm}, una variable dummy que tiene un valor igual a 1 si la vivienda v se localiza en la manzana m y 0 si no. Así, las Ecuaciones 7, 8 y 9 representan mediante restricciones el cumplimiento de las distribuciones marginales para viviendas, hogares y personas respectivamente. La Ecuación 10 asegura que una vivienda se asigna una sola vez, es decir, se puede localizar en una sola manzana.

$$C_{mj} = \sum_{v=1}^{V} c_{vj} y_{vm} \ \forall\, m, j \qquad (7)$$

$$A_{mk} = \sum_{v=1}^{V} \sum_{h=1}^{H} a_{hk} \delta_{hv} y_{vm} = \sum_{v=1}^{V} d_{vk} y_{vm} \ \forall\, m, k \qquad (8)$$

$$B_{ml} = \sum_{v=1}^{V} \sum_{h=1}^{H} \sum_{p=1}^{P} b_{pl} \delta_{ph} \delta_{hv} y_{vm} = \sum_{v=1}^{V} e_{vl} y_{vm} \ \forall\, m, l \qquad (9)$$

$$\sum_{m=1}^{M} y_{vm} = 1 \ \forall\, v \qquad (10)$$

Definiendo d_{vk} y e_{vl} según las Ecuaciones 11 y 12, se pueden reducir las Ecuaciones 7, 8 y 9 a la expresión genérica presentada en la Ecuación 13, donde $F = \{C_{mj}, A_{mk}, B_{ml}\}$, $f = \{c_{vj}, d_{vk}, e_{vl}\}$ y $r = \{j, k, l\}$.

$$d_{vk} = \sum_{h=1}^{H} a_{hk}\delta_{hv} \qquad (11)$$

$$e_{vl} = \sum_{h=1}^{H} \delta_{hv} \sum_{p=1}^{P} b_{pl}\delta_{ph} \qquad (12)$$

$$F_{mr} = \sum_{v=1}^{V} f_{vr}y_{vm} \;\; \forall\, m, r \qquad (13)$$

Para abordar el desafío de identificar el vector solución $y = y_{vm}$, se pueden formular diferentes problemas de optimización. Los problemas presentados a continuación son los que se prueban y comparan en este trabajo de investigación para asignar las viviendas.

Maximización de la entropía

El problema de la maximización de la entropía busca la distribución de probabilidades p_i donde no hay preferencia o predicción sesgada. La entropía es máxima cuando todas las probabilidades son iguales, porque esto representa el mayor grado de incertidumbre, así la maximización de la entropía sería $p_i = y_{vm}$ probabilidad de que la vivienda v esté en la manzana m. Sujeto a las condiciones mencionadas en las Ecuaciones 10 y 13. Así se tiene el problema primal presentado a continuación en las Ecuaciones 14, 15 y 16.

$$\max_{y} -\sum_{v,m} y_{vm} \ln y_{vm} \qquad (14)$$

$$F_{mr} = \sum_{v=1}^{V} f_{vr}y_{vm} \;\; \forall\, m, r \qquad (15)$$

$$\sum_{m=1}^{M} y_{vm} = 1 \;\forall v \qquad (16)$$

Por ejemplo, consideremos una zona censal compuesta por 2 manzanas (m) y 3 viviendas (v), enfocándonos únicamente en el atributo 'MUJERES' (r), la distribución marginal de este atributo representa la cantidad total de mujeres en cada manzana (F_{mr}). Supongamos que en la manzana 1 hay 10 mujeres ($F_{1,MUJERES} = 10$) y en la manzana 2 hay 4 mujeres ($F_{2,MUJERES} = 4$). Si en la vivienda 1 hay 4 mujeres ($f_{1,MUJERES} = 4$), en la vivienda 2 hay

3 mujeres ($f_{2,MUJERES} = 3$) y en la vivienda 3 hay 7 mujeres ($f_{3,MUJERES} = 7$) y se desea saber que viviendas pertenecen a cada manzana, entonces la entropía busca maximizar la expresión $-\sum_{v,m} y_{vm} \ln y_{vm}$ sujeta a que el total de mujeres de cada manzana sea igual a la suma de las mujeres de las viviendas que se asignan en la manzana ($F_{mr} = \sum_{v=1}^{V} f_{vr} y_{vm}$). Esto se traduce en que para la manzana 1: $F_{1,MUJERES} = f_{1,MUJERES}\, y_{11} + f_{2,MUJERES}\, y_{21} + f_{3,MUJERES}\, y_{31}$, es decir, la suma de las mujeres de las viviendas asignadas a la manzana debe ser 10 ($10 = 4y_{11} + 3y_{21} + 7y_{31}$) y para la manzana 2 análogamente debe ser 4 ($4 = 4y_{12} + 3y_{22} + 7y_{32}$). Se debe encontrar el vector y_{vm} que maximice la entropía cumpliendo estas dos restricciones.

Luego, a partir de la demostración del Anexo se obtiene el problema dual de la maximización de la entropía, presentado en la Ecuación 17, para su aplicación computacional más eficiente de la cual se hablará más adelante. Este se presenta a continuación, donde $y_{vm} = \frac{\exp(\sum_r \beta_{mr} f_{vr})}{\sum_{m'} \exp(\sum_r \beta_{m'r} f_{vr})}$ $\forall v, m$ y β_{mr} son los multiplicadores de Lagrange de las restricciones.

$$\underset{\beta}{\text{Max}} - D(\beta) = \underset{\beta}{\text{Max}} \sum_r \beta_{mr} F_{mr} - \sum_v \ln \sum_m \exp\left(\sum_r \beta_{mr}\, f_{vr}\right) \qquad (17)$$

Norma 1 ($||\ ||_1$)

El segundo problema de optimización que se prueba es la minimización de la diferencia entre la distribución marginal (F_{mr}) y los atributos asignados ($\sum_{v=1}^{V} f_{vr} y_{vm}$). El objetivo es reducir el valor absoluto de la diferencia entre las distribuciones conocidas y las asignadas para toda manzana m y atributo r, representada por $|F_{mr} - \sum_{v=1}^{V} f_{vr} y_{vm}|$. Para linealizar el valor absoluto en la aplicación de la optimización en Python, se introduce α_{mr}, que se define como mayor o igual a $F_{mr} - \sum_{v=1}^{V} f_{vr} y_{vm}$ y su negativo $-\alpha_{mr}$ menor o igual, como se detalla en la Ecuación 19. Al minimizar la suma de α_{mr} en la función objetivo (Ecuación 18), la restricción se activa, lo que resulta en la minimización del valor positivo de la diferencia, es decir, minimizar el valor absoluto de la diferencia en la función objetivo. Este problema es más flexible que el anterior, ya que permite que exista esta diferencia entre las distribuciones marginales y las distribuciones asignadas, es decir, aunque lo queremos minimizar al no ser una restricción que se cumpla la igualdad $F_{mr} = \sum_{v=1}^{V} f_{vr} y_{vm}$ permite una mayor flexibilidad en la solución. A continuación, se presenta como se formula este problema de optimización para su aplicación.

$$\underset{y}{Min} \sum_{m,r} \alpha_{mr} \qquad (18)$$

$$-\alpha_{mr} \leq F_{mr} - \sum_{v=1}^{V} f_{vr} y_{vm} \leq \alpha_{mr}\ \forall m, r \qquad (19)$$

$$\sum_{m=1}^{M} y_{vm} = 1 \; \forall v \qquad (20)$$

$$y_{vm} = 0{,}1 \; \forall v, m \qquad (21)$$

$$\alpha_{mr} \geq 0 \; \forall m, r \qquad (22)$$

Norma 2 ($||\,||_2$)

Este problema, tiene como función objetivo la minimización de la diferencia de las distribuciones al cuadrado, generalmente conocido como la minimización de los errores al cuadrado. Este en comparación con norma 1 buscará que los errores sean más pequeños para cada atributo dado que penaliza al cuadrado a las diferencias que se alejan más de la igualdad. A continuación, se presenta como se formula este problema de optimización en las Ecuaciones 23, 24 y 25.

$$\underset{y}{\text{Min}} \sum_{m,r} ||F_{mr} - \sum_{v=1}^{V} f_{vr} y_{vm}\,||^2 \qquad (23)$$

$$\sum_{m=1}^{M} y_{vm} = 1 \; \forall v \qquad (24)$$

$$y_{vm} = 0{,}1 \; \forall v, m \qquad (25)$$

Norma 1 y Norma 2 con restricciones ($||\,||_{1r}, ||\,||_{2r}$)

Los problemas de optimización de norma 1 y norma 2 con restricciones, resuelven los mismos problemas descritos anteriormente, pero con restricciones que buscan disminuir la flexibilidad que presentan estos. Para esto, se debe imponer que ciertos atributos, en este caso $F' = \{C_{mj}, A_{mk}\}$, $f' = \{c_{vj}, d_{vk}\}$ y $r' = \{j, k\}$ que son los atributos de hogares y viviendas cumplan en la asignación la igualdad de las distribuciones marginales. Así los problemas de optimización de la norma 1 y norma 2 con restricciones se presentan respectivamente a continuación.

Norma 1 con restricciones:

$$\underset{y,\alpha}{\text{Min}} \sum_{m,l} \alpha_{ml} \qquad (26)$$

$$-\alpha_{ml} \leq B_{ml} - \sum_{v=1}^{V} e_{vl} y_{vm} \leq \alpha_{ml} \; \forall m, l \qquad (27)$$

$$F'_{mr'} = \sum_{v=1}^{V} f'_{vr'} y_{vm} \ \forall m, r' \qquad (28)$$

$$\sum_{m=1}^{M} y_{vm} = 1 \ for \ all \ v \qquad (29)$$

$$y_{vm} = 0{,}1 \ \forall v, m \qquad (30)$$

$$\alpha_{ml} \geq 0 \ \forall m, l \qquad (31)$$

Norma 2 con restricciones:

$$\underset{y}{Min} \sum_{m,l} ||B_{ml} - \sum_{v=1}^{V} e_{vl} y_{vm}||^2 \qquad (32)$$

$$F'_{mr'} = \sum_{v=1}^{V} f'_{vr'} y_{vm} \ \forall m, r' \qquad (33)$$

$$\sum_{m=1}^{M} y_{vm} = 1 \ \forall v \qquad (34)$$

$$y_{vm} = 0, \ 1 \ \forall v, m \qquad (35)$$

Norma infinito ($||\ ||_\infty$)

Finalmente, se tiene el problema de la minimización de la norma infinito, la cual busca minimizar la distancia máxima de F_{mr} y $\sum_{v=1}^{V} f_{vr} y_{vm}$, para lo cual se tiene la misma formulación que en el caso de la norma 1, pero se debe definir α mayor o igual a todos los α_{mr}. Al minimizar este en la función objetivo, se tendrá α_{mr} igual al mayor α_{mr} de todas las manzanas y atributos. A continuación, se presenta la formulación de este último problema de optimización.

$$\underset{y,\alpha}{Min} \ \alpha \qquad (36)$$

$$-\alpha_{mr} \leq F_{mr} - \sum_{v=1}^{V} f_{vr} y_{vm} \leq \alpha_{mr} \ \forall m, r \qquad (37)$$

$$\sum_{m=1}^{M} y_{vm} = 1 \ \forall v \qquad (38)$$

$$y_{vm} = 0, \ 1 \ \forall v, m \qquad (39)$$

$$\alpha_{mr} \geq 0 \ \forall m, r \qquad (40)$$

$$\alpha \geq \alpha_{mr} \ \forall m, r \qquad (41)$$

Resumen

En este capítulo se han establecido las bases fundamentales para el desarrollo metodológico de una base de datos integrada que abarca aspectos cruciales de uso de suelo y transporte en el área urbana de Santiago. A través de una metodología detallada y rigurosa, se han delineado los pasos necesarios para la imputación de ingresos y asignación de viviendas, asegurando así la creación de una base de personas más completa y detallada.

Inicialmente se identificaron las fuentes de datos relevantes para la modelación del uso de suelo y transporte y se diseñó una estructura de integración de datos y enriquecimiento del Censo que permite delinear la estructura del proyecto. Se destacó la importancia de la dimensionalidad de la base de datos y cómo las distintas dimensiones, desde las variables espaciales a nivel de manzanas censales hasta las características socioeconómicas de los individuos, contribuyen a crear un conjunto de datos robustos y detallados.

Se describen a detalle las bases de datos del Censo de Población y Vivienda y la encuesta Casen, que se trabajarán en el procesamiento de datos, que se detallan a continuación de estas. Se aborda la metodología del procesamiento de las bases de datos, centrándose en dos modelos claves: el Modelo de Imputación de Ingresos (MII) y el Modelos de Asignación de Viviendas (MAV). Estos modelos son esenciales para enriquecer y refinar los datos del Censo, permitiendo una mayor precisión y relevancia en el análisis del uso de suelo y transporte. La aplicación de técnicas avanzadas de *Machine Learning* en el modelo MII y la meticulosa asignación de datos en el MAV ilustran el compromiso del estudio con la precisión analítica.

En resumen, este capítulo ha sentado una sólida base metodológica y técnica para la tesis. A la vez estableció un camino claro para la integración y análisis de datos complejos, procesos de que se presentarán en los siguientes dos capítulos, asegurando que la base de datos resultante será una herramienta valiosa para comprender y modelar las dinámicas de uso del suelo y transporte en Santiago. Este trabajo no solo proporciona una base para análisis actuales, sino que también establece una estructura adaptable para futuras investigaciones y actualizaciones de la base de datos.

Capítulo 4
Generación de la base de datos

Este capítulo detalla el proceso de generación de la base de datos de personas (B-P) a nivel de manzana censal, comenzando con la implementación del modelo de imputación de ingresos (MII). La siguiente etapa aborda el desarrollo del modelo de asignación de viviendas (MAV). A continuación, se presenta un desglose paso a paso de la generación de esta base de datos.

Modelo de Imputación de Ingresos

Esta primera tarea consiste en desarrollar un modelo que permita obtener una base de datos censal de los hogares de la ciudad, incluyendo sus ingresos. Para esto, se debe identificar las variables comunes entre el Censo y la Casen, e igualar sus respuestas, es decir, clasificar las respuestas en los mismos grupos para garantizar su correcta comparación en la imputación. Posteriormente, a través del método de clustering k-means, se generan clústeres de comunas con distribuciones de ingresos similares, según los ingresos entregados en la encuesta Casen. Se verifica que las variables explicativas para la imputación no tengan un nivel de correlación muy alto que deteriore el modelo y se comparan los distintos modelos de *Machine Learning* y tipos de imputación (a nivel de personas o de hogares). Finalmente, se selecciona según indicadores de bondad de ajuste el enfoque más efectivo para la imputación y se imputan los ingresos. En los siguientes apartados se detallan cada uno de estos pasos.

Identificación de Variables y Clustering

Para la imputación de ingresos en el Censo 2017 se requiere de un análisis detallado de las bases de datos de la encuesta Casen y del Censo, llevado a cabo con herramientas de software y lenguaje de programación en Python. De esta forma se identifican las variables comunes que podrían ser explicativas para una imputación de ingresos y se verifica si la encuesta es una fuente representativa que nos daría una buena imputación. Se denominan x y z estas variables comunes entre la encuesta Casen y el Censo respectivamente, las cuales se presentan en la Tabla 7, donde se incluye la pregunta realizada y la abreviación correspondiente a cada variable en la base de datos. Estas variables presentan rangos de respuestas diferentes en las bases de datos, lo que requiere una igualación en las respuestas de ambas bases de datos, detalle que se presenta en el Anexo .

Tabla 7: Variables comunes Casen y Censo

Variables de las personas (x_p y z_p)	
¿Qué relación de parentesco tienen con el jefe/a de hogar?	jh
¿Cuál es el nivel más alto alcanzado o el nivel educacional actual?	niveducc
Durante la semana pasada, ¿Trabajó o no trabajó?	trab

¿A qué se dedica o qué hace el negocio, empresa o institución donde usted trabaja?	tipotrabc
Variables del hogar (x_h y z_h)	
¿Cuántos hogares hay en esta vivienda?	hog
Cantidad de personas por grupos de edad.	edadm, edadj, edadaj, edada, edadam
Porcentaje de hombres en el hogar.	prophombres
Edad, sexo, nivel educacional y tipo de trabajo del jefe de hogar.	edadjh, sexojh, niveducjh, tipotrabjh
Variables de la vivienda (x_v y z_v):	
¿Cuál es el tipo de vivienda que ocupa el entrevistado?	tipoviv
¿Cuál es el material de construcción principal en las paredes exteriores?	muros
¿Cuál es el material de construcción principal en el piso?	piso
¿Cuál es el material de construcción principal en la cubierta del techo?	techo
El agua que usa esta vivienda proviene principalmente de:	agua
¿Cuántas piezas de esta vivienda se usan exclusivamente como dormitorio?	dormitorios

La desagregación espacial de los datos de la encuesta Casen a nivel comunal se detalla en la Tabla 6 del capítulo anterior. Dado que estos datos constituyen una muestra, algunas comunas presentan datos insuficientes para un modelo de imputación individual, lo que hace necesario un procedimiento previo. Para abordar este desafío, se crean clústeres, denominados Ω_c, que consisten en conjuntos de comunas. Estos clústeres se utilizarán para calibrar un modelo de imputación de ingresos para estas mismas comunas, utilizando la metodología k-means.

Este método se basa en las variables claves: el ingreso promedio de los hogares y la varianza del ingreso de cada comuna. Como resultado de este análisis de clúster, se obtienen 4 grupos distintos gracias a la aplicación del método del codo (*elbow method*), que determina la cantidad óptima de grupos minimizando la suma de los errores cuadrados hasta que no se pueda lograr una mejora sustancial. En la Figura 6 se presenta la agrupación realizada con k-means, donde cada punto representa una comuna de la región metropolitana según su ingreso promedio y varianza. Cada color representa un clúster y los centroides de estos están marcados con estrellas negras.

En la Figura 7 se presentan los clústeres obtenidos, junto con su el límite superior de los ingresos promedio, cantidad de viviendas y personas, así como la distribución espacial de estos clústeres en la región metropolitana. Cabe destacar que solo se utilizan datos del sector urbano (delineado en rojo), razón por la cual la comuna de San Pedro no se clasifica en ninguno de los clústeres. Dichos clústeres agrupan comunas con características similares en términos de ingresos y su varianza, clasificándolas en categorías denominadas comunas de ingresos bajos, medios, altos y muy altos. Cada uno de estos clústeres proporciona los datos necesarios, de cantidad de viviendas y personas, para llevar a cabo un modelo de imputación.

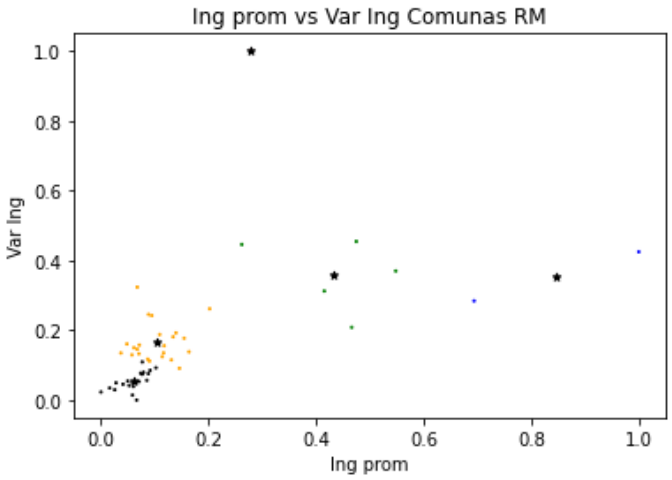


Figura 6: Gráfico de clústeres de comunas k-means

	Clúster	Corte superior ing. Promedio (Millones de $)	Nro. viviendas	Nro. personas
0	Ingresos bajos	1,08	3.298	11.891
1	Ingresos medios	1,60	5.363	18.118
2	Ingresos altos	3,05	2.434	7.145
3	Ingresos muy altos	5,06	1.007	3.063

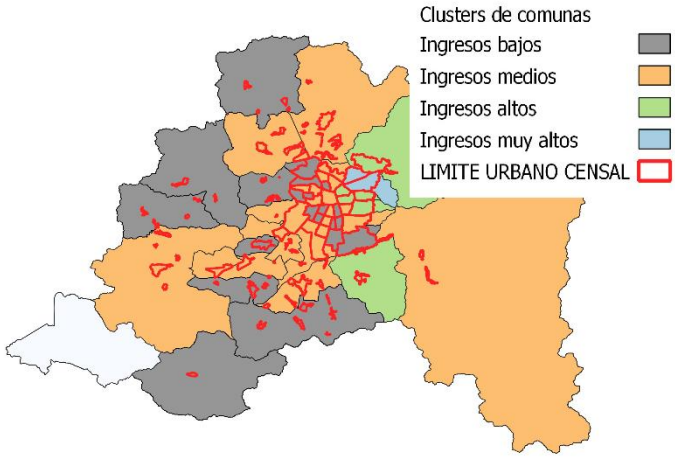


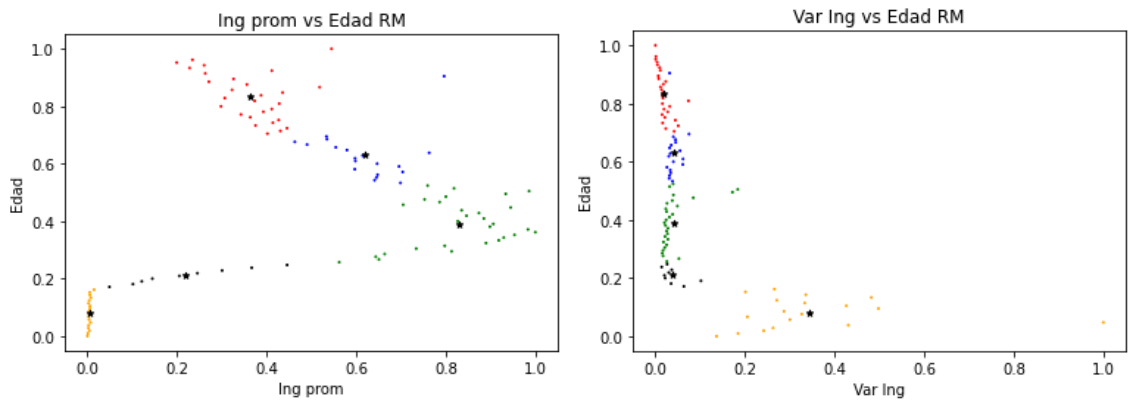
Figura 7: Clústeres de comunas

Para las variables explicativas de los modelos de ingreso con respuestas granulares o con múltiples opciones, que complican el proceso de modelación, también se crean grupos de respuestas utilizando k-means. Un ejemplo de esto es la variable edad. Inicialmente, esta se considera como una variable continua de 0 a 105 años, pero mediante el análisis de clúster, se identifican 5 grupos de edades en función de cómo estos se relacionan con sus ingresos. Esta distinción se basa en la comprensión de que diferencias de edades pequeñas, como entre 40 y 41 años, no tienen un impacto significativo en los ingresos a un nivel tan detallado.

En la Figura 8 se presentan dos gráficos: el de la izquierda muestra de manera normalizada el ingreso promedio de las personas frente a la edad, y el de la derecha muestra la varianza de los ingresos en relación con la edad. Estas serán las tres variables clave utilizadas para el análisis de clúster, aplicando k-means y el método del codo, con las que se tienen los 5 grupos presentados.

Para segmentación de los grupos de edades obtenida, los menores de edad (0 a 17 años) muestran ingresos mínimos y una gran varianza, reflejando la limitada capacidad de generación de ingresos que se da en esta etapa de la vida, aunque algunos comienzan a trabajar o a tener otras fuentes de ingresos. Los jóvenes (18 a 26 años) experimentan un aumento en sus ingresos con menor varianza, marcando el inicio de la vida laboral. El grupo denominado adulto joven (27 a 55 años) alcanza los ingresos máximos, reflejando el peak de sus carreras laborales. Los adultos (56 a 73 años) y adultos mayores (74 a 105 años) ven disminuir progresivamente sus ingresos, lo que indica el fin de sus carreras laborales y la transición hacia la jubilación.

Se aplica un procedimiento similar al nivel educacional, agrupando los datos en categorías de educación baja, media y alta. En cuanto a la variable del tipo de trabajo, se identifican 5 categorías distintas. Finalmente, en relación al porcentaje de hombres en el hogar, se identifican 3 categorías: hogares con mayor proporción de hombres, hogares con proporciones similares y hogares con mayor proporción de mujeres. Estos detalles se presentan en el Anexo .



1	0-17	Menores de edad
2	18-26	Jóvenes
3	27-55	Adulto joven
4	56-73	Adulto
5	74-105	Adulto mayor

Figura 8: Clústeres de edad

Tras un análisis inicial, se reduce el número de variables explicativas de los ingresos. En un inicio se cuenta con 18 variables comunes entre el Censo y la Casen, pero solo se conservan aquellas pertinentes al modelo. Se excluye el origen del agua potable, al no aportar información relevante en áreas urbanas; un 97,97% de las viviendas encuestadas en la Casen tiene como origen el suministro de la red pública, por lo que la variable no varía lo suficiente como para aportar en la diferenciación de ingresos. Esta variable podría ser útil al extender el estudio a áreas rurales para diferenciar este tipo de viviendas. En lo que respecta al

parentesco, aunque no está directamente relacionado con los ingresos de las personas, se utiliza para construir las variables relacionadas con el jefe de hogar.

Se aplica también un análisis de correlaciones entre las variables, identificando que las variables 'trab' y 'tipotrab' presentan una alta correlación, con valores que oscilan entre 0.84 y 0.82 en los distintos clústeres de comunas. En este caso, se opta por mantener únicamente la variable que proporciona información más significativa, es decir, 'tipotrab', ya que esta variable además de decirnos si una persona trabaja o no, nos dice en qué tipo de empresa la persona trabaja. Además, se verifica que las distribuciones de las variables explicativas en ambas bases de datos son similares, lo que respalda la decisión de utilizar la encuesta Casen como fuente de datos para imputar los ingresos en el Censo. Los detalles sobre las distribuciones de estas variables se presentan en el Anexo .

Comparación de modelos y métodos de imputación

Con todos los elementos previamente mencionados, se obtienen los *inputs* necesarios para desarrollar y comparar diferentes modelos de imputación. En total, se ponen a prueba siete modelos: Regresión Lineal - *Linear Regressión (without regularization, with Ridge regularization, with Lasso regularization) (LR),* Regresión de Aumento de Gradiente - *Gradient Boosting Regression (GBR),* Bosque Aleatorio - *Random Forest (RF*), Redes Nuronales - *Neural Network (NN),* Regresión de Vectores de Soporte - *Support Vector Regression (SVR).* Los resultados de estos modelos se presentan en las Tabla 8 y Tabla 9, donde se comparan los indicadores: raíz del error cuadrático medio (RMSE), coeficiente de determinación (R2) y tiempo de ejecución (TIME). Estos indicadores se presentan para cada clúster de comunas, modelo y método de imputación. El método indirecto se presenta en la Tabla 8, mientras que el directo se muestra en la Tabla 9. Cabe resaltar que los resultados entre ambas tablas no son comparables, ya que se aplican a diferentes conjuntos de variables y cantidades de datos, tal como se explicó en el capítulo de la metodología.

En todos los casos, se observa que el modelo RF ofrece las mejores predicciones de ingresos según los indicadores de bondad de ajuste RMSE y R2. Un RMSE más bajo significa un mejor ajuste del modelo, y un R2 cercano a 1 sugiere que el modelo explica de mejor manera la variabilidad de los datos en relación con la media. En resumen, estos indicadores son cruciales para evaluar la calidad de un modelo, ya que RMSE mide la magnitud de los errores y el R2 indica que tan bien el modelo ajusta a la variabilidad de los datos. Es importante destacar que, aunque el modelo RF muestra resultados superiores en términos de ajuste, presenta los tiempos de ejecución más prolongados. Por otro lado, el modelo SVR obtiene los resultados menos favorables en términos de indicadores de ajuste.

Tabla 8: Comparación modelos de imputación a nivel de personas

	ESTIMACIÓN DE MODELOS DE INGRESO A NIVEL DE PERSONAS											
	CLÚSTER DE COMUNAS DE INGRESOS BAJOS			CLÚSTER DE COMUNAS DE INGRESOS MEDIOS			CLÚSTER DE COMUNAS DE INGRESOS ALTOS			CLÚSTER DE COMUNAS DE INGRESOS MUY ALTOS		
MODELO	RSME ($\cdot 10^5$)	R2	TIME (seg)	RSME ($\cdot 10^5$)	R2	TIME (seg)	RSME ($\cdot 10^5$)	R2	TIME (seg)	RSME ($\cdot 10^5$)	R2	TIME (seg)
LINEAR REGRESSION (LR)	2.31	0.437	0.947	3.83	0.341	1.62	14.8	0.205	0.907	20.8	0.229	0.668
LR WITH RIDGE REGULARIZATION	2.31	0.437	1.59	3.83	0.341	3.01	14.8	0.205	1.54	20.8	0.229	1.01
LR WITH LASSO REGULARIZATION	2.31	0.437	0.930	3.83	0.341	1.51	14.8	0.205	0.708	20.8	0.229	0.734
GRADIENT BOOSTING	2.16	0.507	28.6	3.58	0.424	96.5	15.0	0.175	23.5	18.8	0.370	9.98
SUPPORT VECTOR REGRESSION	3.26	-0.118	129	4.40	0.129	537	17.0	-0.051	73.0	24.7	-0.093	10.2
NEURAL NETWORK	2.31	0.437	70.8	3.83	0.341	256	14.8	0.205	66.1	19.3	0.332	33.5
RANDOM FOREST	0.960	0.903	449	1.60	0.884	977	7.40	0.800	136	11.6	0.759	35.0

Tabla 9: Comparación modelos de imputación a nivel de hogares

	ESTIMACIÓN DE MODELOS DE INGRESOS A NIVEL DE HOGARES											
	CLÚSTER DE COMUNAS DE INGRESOS BAJOS			CLÚSTER DE COMUNAS DE INGRESOS MEDIOS			CLÚSTER DE COMUNAS DE INGRESOS ALTOS			CLÚSTER DE COMUNAS DE INGRESOS MUY ALTOS		
MODELO	RSME ($\cdot 10^5$)	R2	TIME (seg)	RSME ($\cdot 10^5$)	R2	TIME (seg)	RSME ($\cdot 10^5$)	R2	TIME (seg)	RSME ($\cdot 10^5$)	R2	TIME (seg)
LINEAR REGRESSION (LR)	5.34	0.235	0.732	8.55	0.178	1.07	26.9	0.195	0.658	38.7	0.202	0.766
LR WITH RIDGE REGULARIZATION	5.34	0.235	1.17	8.55	0.178	1.54	26.9	0.194	1.00	38.7	0.201	0.845
LR WITH LASSO REGULARIZATION	5.34	0.235	0.592	8.55	0.178	0.748	26.9	0.195	0.563	38.7	0.202	0.523
GRADIENT BOOSTING	4.95	0.343	48.3	7.96	0.288	20.5	24.1	0.353	41.3	32.5	0.437	4.20
SUPPORT VECTOR REGRESSION	6.24	-0.043	12.9	9.70	-0.057	56.9	31.0	-0.067	7.42	44.5	-0.053	1.47
NEURAL NETWORK	5.34	0.235	24.2	8.55	0.178	38.1	26.9	0.195	22.9	38.7	0.202	17.8
RANDOM FOREST	2.05	0.888	169	3.33	0.875	305	13.0	0.813	62.2	17.7	0.834	21.3

Es importante mencionar que, aunque el modelo Random Forest (RF) presenta tiempos de ejecución más largos en comparación con los otros modelos, su superioridad en términos de ajuste, lo que se ve reflejado en mejores indicadores de RMSE y R2 compensa este aspecto. Para evidenciar esto se presenta la Figura 9, que ilustra gráficamente las predicciones de los diferentes modelos frente al ingreso de referencia (valor real) para el clúster de comunas con ingresos bajos. Las predicciones realizadas con el modelo RF son significativamente más precisas y acertadas, especialmente en comparación con los otros modelos que no logran predecir valores mayores a un millón de pesos, mientras que en los datos reales existen

valores que superan los 5 millones de pesos. A pesar de que RF tiende a subestimar los valores en este rango, sigue siendo mucho más preciso en sus predicciones que los demás modelos evaluados. En consecuencia, dados los resultados obtenidos en el análisis comparativo, se determina que el modelo Random Forest es el más adecuado para la tarea de imputación de ingresos en el Censo.

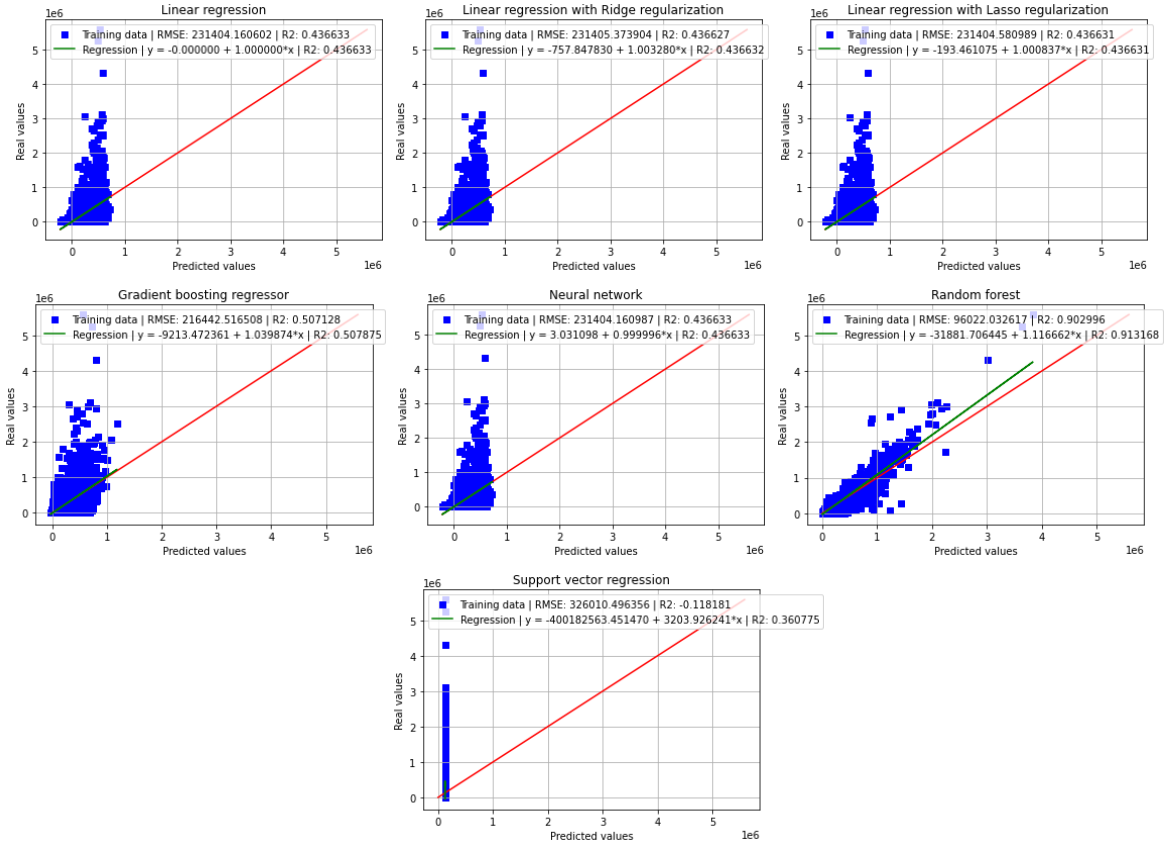


Figura 9: Gráficos de predicción MII caso clúster de comunas de ingresos bajos

Un aspecto metodológico crucial es determinar si es más preciso estimar los ingresos a las personas, para luego eventualmente agregar a un nivel de hogar, o hacerlo directamente a nivel de hogar. Al aplicar estos dos enfoques, los resultados obtenidos, mostrados en la Tabla 10, ofrecen una comparación entre ambos métodos a nivel de cada clúster de comunas y de manera general ponderando cada uno de los clústeres según la cantidad de hogares presentes. Cabe destacar que, aunque las imputaciones a nivel individual y a nivel de hogar no son directamente comparables, una comparación efectiva es factible usando los indicadores RMSE y R2 de las predicciones a nivel de hogar en ambos casos.

Los resultados generales, evaluados en términos de RMSE y R2, indican que el enfoque imputación a nivel de personas seguida de su agregación al hogar es más precisa. Esta conclusión se mantiene consistente al comparar los resultados en cada clúster. En la mayoría

de los casos la imputación a nivel de personas es más precisa, con la excepción del indicador R2 del clúster de comunas de ingresos altos.

Tabla 10: Comparación imputación a nivel de personas y hogares

	PERSONA		AGREGACIÓN		HOGAR		
Clúster de comunas de	RMSE ($\cdot 10^5$)	R2	RMSE ($\cdot 10^5$)	R2	RMSE ($\cdot 10^5$)	R2	Hogares
Ingresos bajos	0.960	0.903	1.81	0.912	2.05	0.888	2867
Ingresos medios	1.60	0.884	3.02	0.893	3.33	0.875	5579
Ingresos altos	7.40	0.800	12.8	0.806	13.0	0.813	2434
Ingresos muy altos	11.6	0.759	14.5	0.882	17.7	0.834	1007
Total	5.71	0.879	6.21	0.862			

Además de los indicadores presentados en la Tabla 10, la Figura 10 ilustra claramente los resultados de las imputaciones realizadas en el Censo en comparación con la distribución observada en la encuesta Casen. En el gráfico de la izquierda, se muestra la distribución de los ingresos de los hogares según la encuesta Casen, el gráfico central presenta la distribución resultante de las imputaciones indirectas agregadas a nivel de hogar, mientras que el gráfico de la derecha muestra la distribución resultante de las imputaciones directas a los hogares. Es notorio que la distribución de la variable imputada mediante el método indirecto, a nivel de personas, es más similar a la observada en la Casen. Este enfoque proporciona una mayor riqueza de información, al ofrecer datos detallados para cada individuo, además de la información agregada a nivel de hogar. En consecuencia, se concluye que la imputación indirecta de los ingresos al hogar representa la opción óptima para la imputación de los ingresos. Es importante destacar que este método incluye una consideración especial para las personas menores de edad que no trabajan, las cuales generalmente no tienen ingresos. Esta suposición se valida con los datos de la encuesta Casen, que indican que el 82% de las personas de este grupo no poseen ingresos y en promedio el total de este grupo poseería un ingreso de $5.000 aproximadamente.

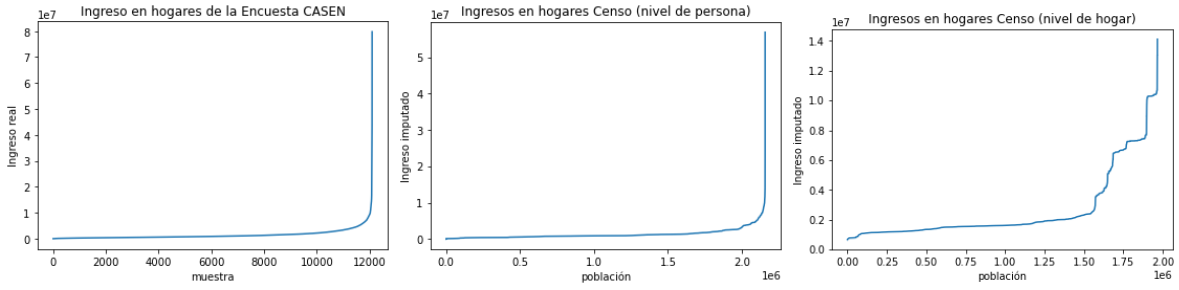


Figura 10: Gráficos de comparación de distribución de ingresos imputados

Modelo de Asignación de viviendas

El modelo de asignación de viviendas (MAV) tiene como objetivo alcanzar un nivel de granularidad espacial superior al actual, desagregando la información sobre personas, hogares y viviendas a un nivel más detallado. Este modelo utiliza la información censal de las personas de la ciudad de Santiago, obtenida del Censo, y la complementa con los datos de distribuciones marginales a nivel de manzana proporcionados por el INE, tal como se describió en el capítulo anterior. Se espera obtener la mejor asignación posible de las personas de la ciudad, con sus respectivos hogares y viviendas, a las manzanas censales pertenecientes a cada zona censal.

Para lograr esto, se identifican y corroboran los atributos comunes entre las informaciones censales disponibles mencionadas anteriormente y se preparan los datos para abordar los problemas de optimización descritos en la metodología. Se comparan diferentes puntos de inicialización y tipos de problemas de optimización, seleccionando el enfoque más adecuado según indicadores de bondad de ajuste para la asignación. La implementación de este proceso se realiza utilizando el lenguaje de programación Python. En las siguientes secciones, se explican en detalle cada uno de estos pasos necesarios para la implementación del modelo.

Identificación de Atributos y preparación de los datos

Al igual que en el modelo de imputación de ingresos, para el modelo de asignación de viviendas (MAV) se tienen atributos comunes entre las dos fuentes de datos: el Censo detallado a nivel de hogar, personas y vivienda, con los atributos denominados a_{hk}, b_{pl} y c_{vj} respectivamente, y de las distribuciones marginales a nivel de manzana, denotadas como A_{mk}, B_{ml} y C_{mj}. Primero se identifican los atributos comunes entre las fuentes de datos a nivel de personas, hogares y viviendas, que se listan en la Tabla 11. Luego, se verifica que la base de datos a nivel de zonas censales urbanas, de personas, hogares y viviendas detallada del Censo, coincida con las restricciones marginales de la base de datos a nivel de manzana. Esta verificación es fundamental para garantizar la consistencia del modelo, especialmente porque si no hay concordancia, los problemas con restricciones de igualdad $F_{mr} = \sum_{v=1}^{V} f_{vr} y_{vm}$ resultan ser infactibles. Como resultado de este proceso, se tiene que la información de distribuciones marginales coincide con la información detallada del Censo para 1.863 zonas censales, involucrando un total de 6.780.655 personas, 2.273.488 viviendas y 2.138.740 hogares, que deben asignarse a 48.160 manzanas censales urbanas.

Tabla 11: Atributos comunes Censo detallado y distribuciones marginal

Atributos de las personas[1]	
Cantidad de mujeres	MUJERES
Cantidad de hombres	HOMBRES
Cantidad de personas entre edades 0 a 5	EDAD_0A5
Cantidad de personas entre edades 6 a 14	EDAD_6A14
Cantidad de personas entre edades 15 a 64	EDAD_15A64
Cantidad de personas entre edades 65 y más	EDAD_65YMAS
Cantidad de personas que se consideran pertenecientes a un grupo indígena	PUEBLO
Cantidad de inmigrantes	INMIGRANTES
Atributos de los hogares y viviendas	
Cantidad de hogares	CANT_HOG
Cantidad de viviendas colectivas	VIV_COL
Cantidad de viviendas particulares ocupadas con moradores presentes	VPOMP
Cantidad de viviendas con materialidad aceptable	MATACEP
Cantidad de viviendas con materialidad irrecuperable	MATIRREC
Cantidad de viviendas con materialidad recuperable	MATREC
Cantidad de viviendas particulares del tipo: casa, departamento, tradicional indígena, pieza o conventillo, mediagua, otro.	P01_1, P01_1, P01_1, P01_1, P01_1, P01_1, P01_1
Cantidad de viviendas con muros tipo: hormigón armado, albañilería, tabique forrado, tabique sin forrar, adobe, materiales precarios.	P03A_1, P03A_2, P03A_3, P03A_4, P03A_5, P03A_6
Cantidad de viviendas con techo tipo: tejas, losa hormigón, planchas metálicas, fonolita, paja, materiales precarios, sin cubierta.	P03B_1, P03B_2, P03B_3, P03B_4, P03B_5, P03B_, P03B_7
Cantidad de viviendas con piso tipo: parquet o piso flotante, radier sin revestimiento, baldosa de cemento, capa de cemento sobre tierra, tierra.	P03C_1, P03C_2, P03C_3, P03C_4, P03C_5
Cantidad de viviendas con origen del agua en: red pública, pozo o noria, camión aljibe, rio o lago.	P05_1, P05_2, P05_3, P05_4

Para aplicar las formulaciones de los problemas de optimización presentados en la metodología, es imprescindible tener disponibles todos los atributos a nivel de vivienda, tal como se establece en las Ecuaciones 10 y 11 del capítulo anterior. Se trabaja con las bases de datos de personas y hogares para integrarlas en la base de viviendas. En la base de personas del Censo, cada individuo está asociado a una vivienda y hogar específico, lo que permite determinar la cantidad de personas en cada uno de los atributos de las personas y hogares que ocupan cada vivienda.

Optimización y comparación de los problemas de optimización

La implementación de los problemas de optimización se realiza en Python, utilizando la librería *Pyomo* y el *solver Gurobi*, detalles de los cuales se presenta un ejemplo en el Anexo

[1] No todas las manzanas tienen toda la información de las variables de personas debido a la confidencialidad de los datos.

. *Gurobi* es adecuado para los problemas de optimización lineales de variables enteras. Sin embargo, la maximización de la entropía, al no ser un problema lineal, requiere el *solver Ipopt*, especializado en problemas no lineales de variables continuas. Este modelo, siendo el más riguroso y estricto al incluir todas las distribuciones marginales como restricciones, se prueba y optimiza con el problema Dual, un enfoque más sencillo y rápido que entrega resultados equivalentes al problema primal. Cabe recalcar que los resultados numéricos de la implementación son dependientes del computador y sistema operativo que se utilicen.

La maximización de la entropía asigna probabilísticamente si una vivienda pertenece a una manzana censal, con resultados en el intervalo [0,1] para la variable a optimizar (y_{vm}). Al aplicar un criterio de asignación de mayor a menor probabilidad, es decir, ir asignando valor 1 o 0 a la variable, no siempre se cumplen todas las distribuciones marginales. Sin embargo, al ser el problema de optimización más estricto en las restricciones marginales, sus resultados sirven como un punto de inicialización de la variable a optimizar para los otros problemas de optimización. La Tabla 12 muestra una comparación de resultados utilizando la entropía como punto de inicialización y sin un punto inicial especifico. Denominaremos $D = \sum_{m,r}(F_{mr} - \sum_{v=!}^{V} f_{vr}y_{vm})$, que mide la distancia a una asignación ideal que cumpla con todos los marginales, es decir, $D = 0$. También se presenta la condición de término de la optimización, ya sea maxTimeLimit/maxIterationLimit cuando se alcanza el tiempo/iteración máxima de búsqueda de soluciones (250 segundos/100 iteraciones) u optimal en caso de encontrar una solución óptima al problema. Finalmente, se incluye el tiempo de ejecución en segundos.

Los resultados de la Tabla 12 indican que al usar entropía como punto de partida para la optimización puede disminuir el tiempo de ejecución y/o la distancia D en ciertos casos, e incluso cambiar la condición de termino de tiempo límite a óptimo. Sin embargo, en otros casos, como en la zona censal 13104031004, no se observan mejoras, lo que se puede deber a la calidad de la solución de entropía, que en caso de no ser buena no aporta mejoras en la búsqueda de una solución óptima. Por lo tanto, se recomienda utilizar la entropía como punto de partida solo cuando esta ofrezca buenas soluciones, si su tiempo de ejecución es pequeño o se tenga optimizada con anterioridad, como lo es en nuestro caso, por el contrario, puede no presentar mayores ventajas. Es crucial recalcar que el punto de partida no afecta la solución óptima alcanzada para un mismo problema de optimización, como en los casos 13403021003 y 13301061011, donde la asignación final es la misma independiente del punto de inicio de la búsqueda de la solución.

Tabla 12: Punto de inicialización con y sin entropía

		Entropía		Norma 2 sin punto inicial			Norma 2 con punto inicial entropía		
Zona Censal	Nro viv	D	Condición de término	D	Condición de término	Tiempo (s)	D	Condición de término	Tiempo (s)
13301041001	86	58	maxIterationLimit	76	maxTimeLimit	50,36	0	optimal	0,73
13128021002	801	836	maxIterationLimit	128	maxTimeLimit	252,46	101	maxTimeLimit	252,43
13104031004	694	1.828	maxIterationLimit	156	maxTimeLimit	252,78	156	maxTimeLimit	252,82
13403021003	33	20	optimal	0	optimal	5,39	0	optimal	0,18
13301061011	123	22	optimal	0	optimal	0,21	0	optimal	0,27

Para evaluar el desempeño de los problemas de optimización basados en las diferentes normas propuestas, se lleva a cabo un caso de estudio comparativo en zonas censales de la comuna de Providencia. La Tabla 13, presenta los resultados de esta comparación, incluyendo la distancia D, el tiempo de ejecución en segundos y la condición de término (Time o Optimal, dependiendo de si el proceso de optimización termina porque alcanza el límite de tiempo o una solución óptima). Se analizan tres casos distintos (clasificados según las variables manzanas, personas, hogares, viviendas y distribuciones marginales de la comuna de Providencia): un caso sencillo (CS) con 10 manzanas, 932 personas, 301 hogares, 373 viviendas y 408 distribuciones marginales que cumplir; un caso promedio de la comuna (CP) con 12 manzanas, 2733 personas, 1323 hogares, 1494 viviendas y 506 distribuciones marginales que cumplir; y el caso más difícil de la comuna (CD) con 42 manzanas, 4365 personas, 1544 hogares, 1752 viviendas y 1754 distribuciones marginales que cumplir. Se compararon diferentes problemas de optimización: norma 1 ($||\ ||_1$), norma 1 con restricciones ($||\ ||_{1r}$), norma 2 ($||\ ||_2$), norma 2 con restricciones ($||\ ||_{2r}$) y norma infinito ($||\ ||_\infty$), todos ellos inicializados con la solución de entropía.

En el caso sencillo (CS), todos los métodos logran una asignación óptima, cumpliendo con todas las distribuciones marginales. En el caso promedio (CP), la norma 2 con y sin restricciones no alcanza una solución óptima dentro del límite de tiempo, destacando que la norma 2 con restricciones no encuentra ninguna solución, quedándose con la solución inicial de entropía. En el caso más difícil (CD), solo las normas 1 e infinito logran mejorar la solución inicial proporcionada por la entropía.

Tabla 13: Comparación métodos de asignación, caso de estudio en Providencia

	ID_ZC	$\|\|\ \|\|_1$	$\|\|\ \|\|_{1r}$	$\|\|\ \|\|_2$	$\|\|\ \|\|_{2r}$	$\|\|\ \|\|_\infty$	Entropía
D	13123071003 (CS)	0	0	0	0	0	903
	13123031007 (CP)	0	0	25	4.358	0	4.358
	13123041003 (CD)	304	5.130	5.130	5.130	677	5.130
tiempo (s)	13123071003 (CS)	20	15	44	30	19	
	13123031007 (CP)	113	139	418	333	110	
	13123041003 (CD)	645	530	1.267	781	643	
Condición de término	13123071003 (CS)	Optimal	Optimal	Optimal	Optimal	Optimal	
	13123031007 (CP)	Optimal	Optimal	Time	Time	Optimal	
	13123041003 (CD)	Time	Time	Time	Time	Time	

Aunque inicialmente podría parecer que la norma 1 es la opción más favorable debido a su menor distancia D, es importante considerar que la norma infinito es más restrictiva, permitiendo como máximo una distancia de 2 en los marginales, es decir, tiene una mayor distancia total pero permite distancias menores que la norma 1. Por lo tanto, es importante determinar cuál método comete menos errores específicamente en la cantidad total asignada de personas, hogares y vivienda. Para esto, se utilizan los indicadores de error de primer orden presentados en las Ecuaciones 42 y 43. Estas ecuaciones calculan, respectivamente el error relativo y absoluto en la asignación de la variable i (personas, hogares o viviendas) para cada manzana censal m de la zona censal (Error relativo) y para la zona censal zc en su conjunto (Error absoluto). Aquí, x_{im} representa la cantidad total de i asignadas en la manzana m, y X_{im} indica la cantidad total de i que realmente pertenecen a la manzana m.

$$\frac{|x_{im} - X_{im}|}{X_{im}} \ \forall m, i \qquad (42)$$

$$\sum_{m=1}^{M} |x_{im} - X_{im}| \ \forall zc, i \qquad (43)$$

La Tabla 14 ofrece una comparación de estos indicadores de primer orden, para el error relativo se presenta el mínimo, promedio y máximo de las manzanas censales, revelando que la norma 1 resulta tener mejores indicadores de primer orden, a excepción del error relativo promedio y error absoluto en el caso de las personas. Estas excepciones son marginales y no compensan la mejora de en la asignación de hogares y viviendas que presenta la norma 1 por sobre la norma infinito. La diferencia porcentual entre ambas normas es pequeña y, en un contexto ideal se compararían los problemas de norma 1 e infinito en todas las zonas censales. Si la diferencia siguiera siendo pequeña, se podría recurrir al criterio de máxima entropía para decidir cuál es el problema más adecuado. Sin embargo, debido a los alcances de la tesis y limitaciones computacionales, este análisis exhaustivo no es posible. No obstante, el estudio de caso realizado en la comuna de Providencia justifica y respalda la elección de la

norma 1, ya que demuestra que la norma infinito no es significativamente mejor que la seleccionada.

Tabla 14: Indicadores de primer orden norma 1 v/s norma infinito

		Error relativo			Error absoluto
		mínimo	promedio	máximo	
Personas	$\|\|\ \|\|_1$	0,00%	1,77%	5,08%	60 (1,37%)
	$\|\|\ \|\|_\infty$	0,00%	1,65%	8,57%	50 (1,15%)
Hogares	$\|\|\ \|\|_1$	0,00%	0,90%	7,41%	12 (0,78%)
	$\|\|\ \|\|_\infty$	0,00%	3,16%	11,76%	34 (2,20%)
Viviendas	$\|\|\ \|\|_1$	0,00%	0,91%	5,56%	14 (0,80%)
	$\|\|\ \|\|_\infty$	0,00%	3,19%	10,53%	38 (2,17%)

Finalmente, se aplica el modelo de asignación en toda la región metropolitana utilizando el método de norma 1 con punto inicial de entropía. Este obtiene los errores de primer orden para las personas, hogares y viviendas mostrados en la Tabla 15, que se interpreta de la siguiente manera: por ejemplo, en el caso de la región metropolitana se asignan erróneamente un total de 85.716 personas (error absoluto), lo que equivale a un 1,26% de las personas asignadas en la región. A nivel de manzana censal, en promedio se asignan incorrectamente 0,43 hogares (error absoluto), o un 1,87% de los hogares de cada manzana (error relativo). En algunos casos extremos, el error relativo máximo en una manzana supera el 100%, lo que ocurre cuando se asignan más personas, hogares o viviendas de las que realmente existen en esa manzana censal. Estos aspectos se analizarán más a fondo en el siguiente capítulo de análisis de la base de datos resultante.

Tabla 15: Errores de primer orden para los resultados asignación en la RM

		Error relativo			Error absoluto		
	Nivel espacial	mínimo	promedio	máximo	mínimo	promedio	máximo
Error personas	RM	-	-	1,26%	-	-	85.716
	Comuna	0,06%	1,45%	12,11%	4	1.681	17.186
	Zona Censal	0,00%	0,88%	84,52%	0	46	9.138
	Manzana Censal	0,00%	2,50%	1.611%	0	1,79	1.338
Error hogares	RM	-	-	0,97%	-	-	20.756
	Comuna	0,00%	1,19%	12,01%	0	407	4.538
	Zona Censal	0,00%	0,69%	85,89%	0	11	2.588
	Manzana Censal	0,00%	1,87%	975%	0	0,43	416
Error viviendas	RM	-	-	0,73%	-	-	16.610
	Comuna	0,00%	0,91%	13,15%	0	326	4.996
	Zona Censal	0,00%	0,44%	87,02%	0	8,92	2.734
	Manzana Censal	0,00%	1,47%	875%	0	0,34	541

Resumen

Este capítulo ha sido fundamental para detallar el proceso meticuloso de la generación de la base de datos de personas (B-P). A través de dos modelos clave, el Modelo de Imputación de Ingresos (MII) y el Modelo de Asignación de Viviendas (MAV), se ha logrado una representación detallada y precisa de la distribución de la población y sus características socioeconómicas, incluyendo los ingresos en la ciudad.

El desarrollo del MII fue un paso crucial para obtener una base censal enriquecida con datos de ingresos. Se ha demostrado a partir de indicadores de bondad de ajuste que la imputación de ingresos a nivel de persona, seguida de una agregación al nivel de hogar, es un enfoque más preciso, proporcionando una distribución de ingresos que refleja la observada en la encuesta Casen y también enriquece el análisis con un nivel de detalle más profundo. La implementación del modelo Random Forest se ha destacado como el modelo más efectivo, equilibrando precisión y eficiencia a pesar de tener tiempos de ejecución más largos.

El MAV, por su parte, ha permitido asignar de manera precisa a personas, hogares y viviendas en manzanas censales específicas, logrando una granularidad espacial sin precedentes. Utilizando métodos de optimización avanzados y comparando diferentes enfoques, se ha determinado que el problema de optimización de norma 1, con un punto inicial basado en entropía, es el más adecuado para este propósito en este caso. A pesar de los desafíos presentados en la asignación de algunas zonas censales, este método ha demostrado ser eficiente y confiable en la asignación precisa de las unidades de análisis a nivel de manzana.

En conclusión, este capítulo ha sido crucial en la construcción de una base de datos integrada y desagregada de uso de suelo y transporte para la ciudad. Los modelos implementados han mejorado significativamente la precisión y relevancia de la base de datos de personas, superando limitaciones previas con la precisión en la georreferenciación de los datos y disponibilidad de información socioeconómica relevante, como lo es el ingreso. Este avance proporciona una herramienta valiosa para estudios futuros y la planificación urbana.

Capítulo 5
Análisis de la base de datos resultante

Este capítulo está dedicado al análisis de los resultados obtenidos de la base de datos resultante para personas (B-P) en el capítulo anterior. Inicialmente, se examinarán los resultados generales relativos a los ingresos imputados a la ciudad y su distribución a nivel de zona censal. Seguidamente, se aborda la asignación de viviendas a nivel de manzana censal, incluyendo un análisis de la distribución granular de los ingresos y la heterogeneidad de estos dentro de cada zona censal, comuna y clústeres de comunas. Además, se evalúan tanto los errores de primer orden como los errores específicos asociados a cada atributo en la asignación. Finalmente, se analiza una variable adicional de interés: la distribución del nivel de educación de las personas mayores a 18 años en las manzanas censales.

Resultados de la imputación de ingresos

En cuanto a la imputación de ingresos se destaca que el 99,99% de los ingresos imputados en Censo se encuentran dentro del rango de valores proporcionados por la Encuesta Casen. Por otro lado, solo un 0,041% de los ingresos registrados en la Casen no se representan en el rango imputado en el Censo. La Tabla 16 presenta los resultados generales de la imputación, que revelan similitudes significativas en las medias de ambos conjuntos de datos con un 99% de confianza. No obstante, es importante destacar que en el último decil se tiene una tendencia a subestimar los ingresos, tanto a nivel individual como de hogares, aunque esta diferencia es poco significativa en el contexto general. En conclusión, se estima que los resultados de la imputación, tanto de personas como para la agregación a hogares, tienen resultados satisfactorios y se ajustan dentro de los rangos esperados.

Tabla 16: Resultados imputación de ingresos

	Personas		Hogares	
	CASEN	CENSO	CASEN	CENSO
Moda del último decil	$ 1.500.000	$ 1.324.203	$ 3.500.000	$ 2.648.406
Media de todos los datos	$ 436.129	$ 432.182	$ 1.449.331	$ 1.400.872
Moda del primer decil	$ 0	$ 0	$ 200.000	$ 411.838
Mínimo de todos los datos	$ 0	$ 0	$ 0	$ 0
Cantidad de datos	40.217	6.753.655	12.102	2.083.564

Los resultados de la imputación también se visualizan gráficamente en el ámbito espacial a nivel de zona censal, agrupando los valores en 10 clases de ingresos separadas por puntos de ruptura. Esta agrupación sigue el método de clasificación de rupturas naturales de Jenks, que busca minimizar la varianza dentro de cada clase y maximizarla entre las clases (Jenks, 1967). La Figura 11 muestra esta distribución espacial de los ingresos promedio de las personas a nivel de zona censal. Es notorio que los ingresos promedio más altos se concentran

en zonas censales interiores de las comunas de Las Condes y Vitacura, información más detallada de lo que se tenía a partir de la encuesta Casen, pero que concuerda a niveles comunales con lo esperado basado en esta encuesta y la división de clústeres previamente establecida. En contraste, los ingresos promedio más bajos se encuentran en las zonas periféricas, especialmente en zonas censales de las comunas de La Pintana, Pudahuel y Puente Alto.

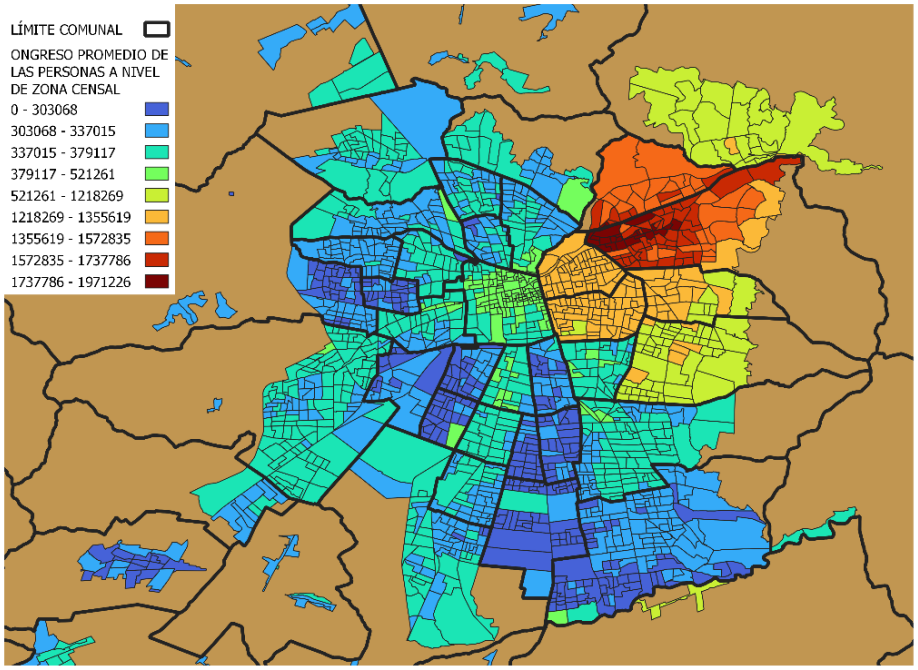


Figura 11: Ingreso promedio de las personas a nivel de zona censal

Por otro lado, en la Figura 12 se presenta la representación de los ingresos promedio por hogar en cada zona censal, revelando una distribución que difiere del caso de los ingresos individuales. En este caso, se observa que los ingresos más altos se concentran en la periferia del sector oriente, particularmente en las zonas censales periféricas de las comunas de Las Condes, Lo Barnechea y Vitacura. Por el contrario, los ingresos más bajos se encuentran principalmente en el centro de la ciudad, específicamente en la comuna de Santiago y sus alrededores.

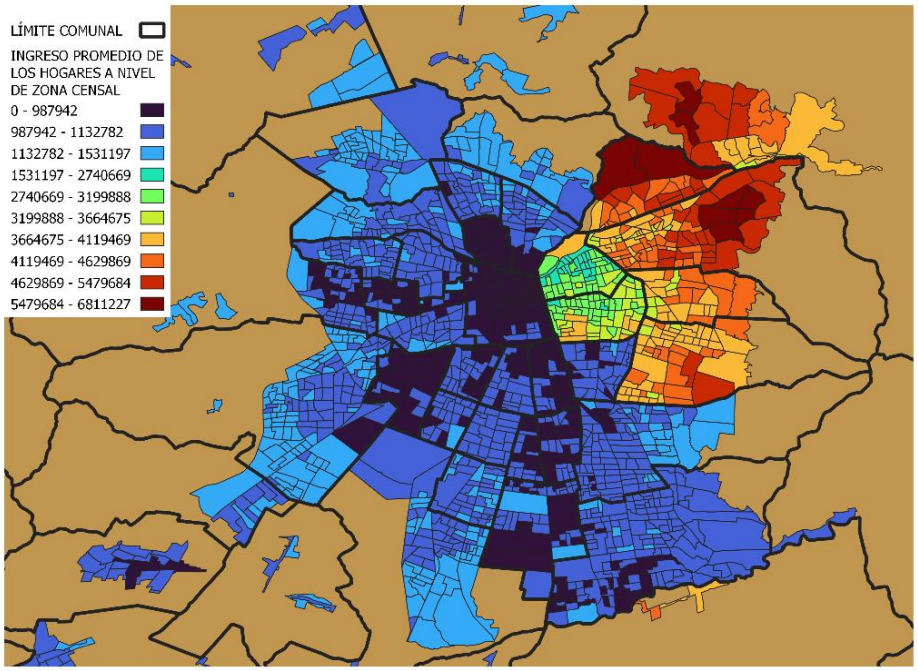


Figura 12: Ingreso promedio de los hogares a nivel de zona censal

Asignación de Viviendas

Tras la asignación de las viviendas a nivel de manzana en Santiago, es posible realizar análisis más detallados. A continuación, se analizarán los errores cometidos en la asignación, los ingresos de las personas y hogares espacialmente, examinando su heterogeneidad y, adicionalmente, se explorará otra variable que ahora puede ser analizada de manera granular gracias a la asignación de viviendas: el nivel de educación.

Errores en la asignación

En la asignación realizada existen zonas censales que alcanzan el óptimo, teniendo asignaciones sin errores en las distribuciones marginales, pero otras no. La Figura 13 muestra la condición de término de la optimización para cada una de las zonas censales de la región metropolitana. Se observa que el 73% de estas zonas finalizan el proceso por el límite máximo de tiempo establecido, lo que resalta la importancia de analizar detalladamente el error de asignación permitido en estos casos. Para abordar esto, las Figura 14, Figura 15 y Figura 16 exponen los errores de primer orden en la asignación de personas a nivel de manzana censal, zona censal y comuna respectivamente. La definición de estos errores se presentó en las Ecuaciones 42 y 43 del capítulo anterior. Errores que también se han evaluado para hogares y viviendas, y los resultados, que son similares, se pueden consultar en el Anexo . A pesar de que en la Tabla 15 del capítulo anterior se indicaba error máximo considerable a nivel de manzana censal, el análisis gráfico revela que estos errores se concentran

específicamente en algunas manzanas censales periféricas. Se observa una tendencia similar a nivel de zonas censales, teniéndose los errores elevados específicamente en casos de zonas censales donde el problema de asignación eran los más grandes y complejos de la ciudad. Y a nivel comunal se observan errores pequeños, donde los mayores se presentan en las comunas periféricas de la ciudad de Santiago.

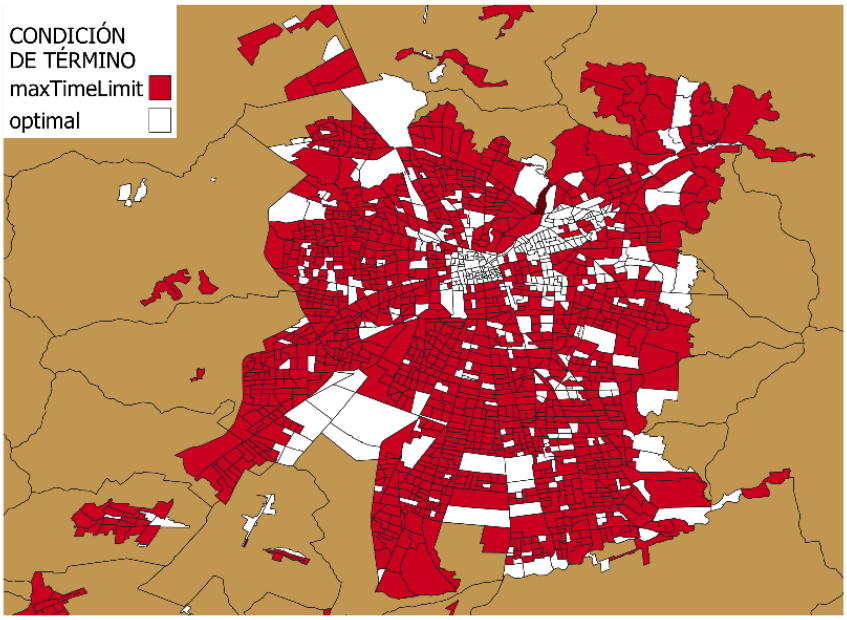


Figura 13: Condición de término de la asignación en las zonas censales

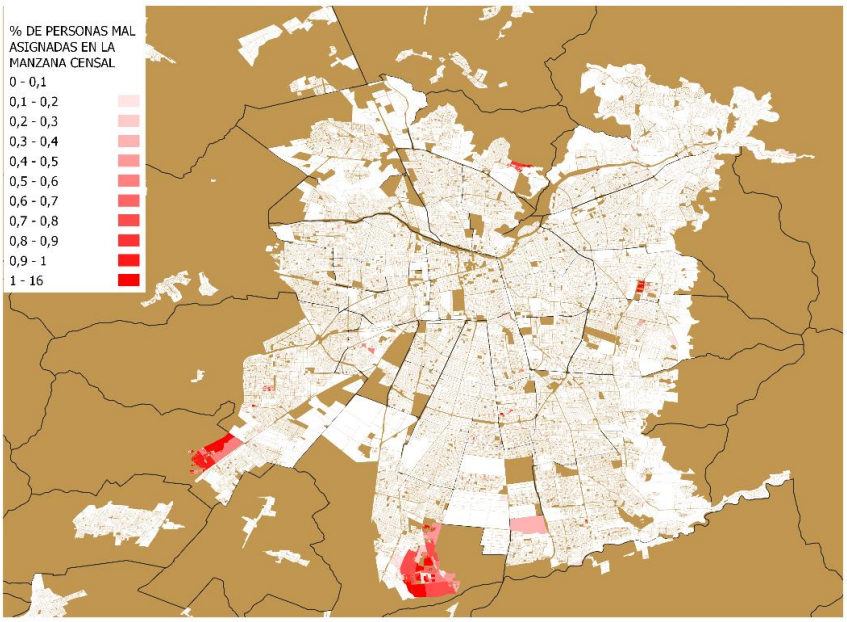


Figura 14: Porcentaje de personas mal asignadas a nivel de manzana censal

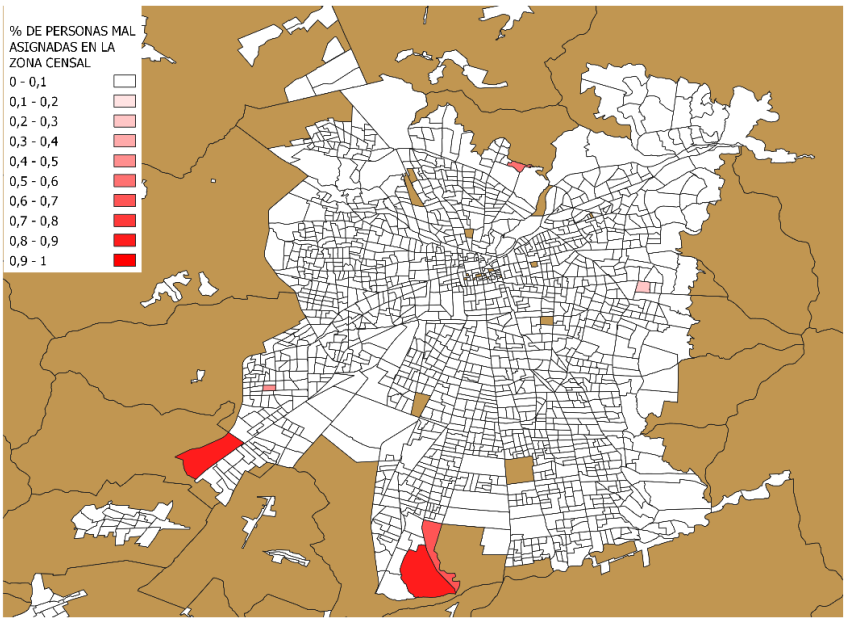


Figura 15: Porcentaje de personas mal asignadas a nivel de zona censal

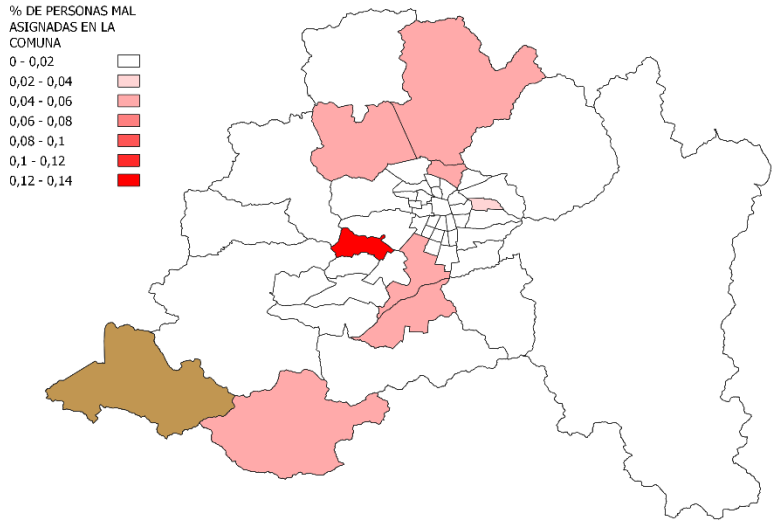


Figura 16: Porcentaje de personas mal asignadas a nivel de comuna

La Tabla 17 detalla los errores absolutos y relativos de asignación para cada atributo a nivel de la región metropolitana y zonas censales, estos también podrían obtenerse a nivel de manzana censal. El valor mínimo del error relativo no se reporta, ya que es cero para todos estos. Un hallazgo notable es que para el atributo de personas inmigrantes se presenta el mayor error relativo promedio y máximo, superando incluso el 100% en el caso máximo. Esto se debe a la falta de información sobre inmigrantes en numerosas manzanas por razones

de confidencialidad. No obstante, en promedio, el atributo de inmigrantes presenta un error de 3,33%, mientras que el error más bajo de 0,20% corresponde al atributo de viviendas tipo casa (P01_1). En cuanto al error absoluto, el mayor error se observa en la asignación de viviendas colectivas, con un 3,27% de estas mal asignadas, y el menor error en los atributos de viviendas tipo departamento (P01_2) y materialidad del techo de losa de hormigón (P03B_2) con un error de 0,13%. En general, aunque los errores máximos son elevados, representan casos puntuales, como se evidencia en los histogramas de error de los atributos del Anexo .

Tabla 17: Indicadores de error en la asignación de cada atributo

	Error absoluto en la RM		Error relativo a nivel de zona censal			Error absoluto en la RM		Error relativo a nivel de zona censal	
Atributo	Cantidad	% de error	Prome-dio	Máximo	Atributo	Cantidad	% de error	Prome-dio	Máximo
MUJERES	23.865	0,68%	0,49%	42,95%	P03A_1	3.762	0,56%	0,91%	77,78%
HOMBRES	23.545	0,72%	0,51%	43,17%	P03A_2	6.662	0,59%	0,43%	45,45%
EDAD_0A5	6.629,5	1,39%	1,10%	40,95%	P03A_3	2.855	1,70%	1,00%	85,19%
EDAD_6A14	8.744,5	1,20%	0,96%	37,74%	P03A_4	1.092	1,87%	0,81%	100,00%
EDAD_15A64	28.739	0,61%	0,43%	42,30%	P03A_5	260	0,72%	1,17%	100,00%
EDAD_65YMAS	7.203,5	1,03%	0,92%	53,56%	P03A_6	32	1,50%	1,00%	100,00%
PUEBLO	14.373	2,20%	1,76%	43,54%	P03B_1	5.697	1,05%	0,84%	71,58%
INMIGRANTES	8.076,5	2,69%	3,33%	182,14%	P03B_2	552	0,13%	0,92%	100,00%
CANT_HOG	10.378	0,49%	0,35%	42,94%	P03B_3	6.435	0,61%	0,43%	67,00%
VIV_COL	67	3,27%	2,82%	100,00%	P03B_4	79	1,96%	1,11%	100,00%
VPOMP	7.486	0,36%	0,22%	43,49%	P03B_5	1	0,57%	0,70%	100,00%
MATACEP	7.171	0,39%	0,27%	44,52%	P03B_6	64	1,69%	1,23%	100,00%
MATIRREC	122	1,97%	0,98%	80,28%	P03B_7	14	1,43%	1,15%	100,00%
MATREC	2.464	1,15%	0,66%	100,00%	P03C_1	7.322	0,38%	0,25%	44,15%
P01_1	7.289	0,48%	0,20%	43,62%	P03C_2	865	1,97%	1,13%	100,00%
P01_2	925	0,13%	0,68%	100,00%	P03C_3	1.225	1,58%	0,98%	83,78%
P01_3	1	0,34%	0,49%	100,00%	P03C_4	355	1,97%	1,15%	100,00%
P01_4	335	1,13%	1,25%	100,00%	P03C_5	44	2,60%	1,22%	100,00%
P01_5	365	2,26%	1,35%	100,00%	P05_1	7.421	0,36%	0,23%	43,43%
P01_6	3	1,57%	2,24%	100,00%	P05_2	69	1,40%	1,19%	100,00%
P01_7	106	1,28%	0,98%	100,00%	P05_3	41	1,65%	1,38%	100,00%
					P05_4	17	1,06%	1,90%	100,00%

Análisis espacial de los ingresos

Anteriormente se realizó un pequeño análisis de la distribución de los ingresos a nivel de zona censal, pero gracias a la asignación de viviendas a nivel de manzana en la región metropolitana, se logra apreciar una granularidad más detallada en atributos de las personas,

hogares y viviendas. En la Figura 17, se presentan los ingresos promedio de las personas a nivel de manzana censal, revelando variaciones significativas dentro de las zonas censales. Un ejemplo claro se aprecia a lo largo de los ejes como la Gran Avenida José Miguel Carrera en el sector sur, y en el sector poniente, Av. Los Pajaritos y Av. Alameda Libertador Bernardo O'Higgins, grandes avenidas de transporte (destacadas en rojo en la figura) también vinculadas a diferentes líneas de metro (línea 2, 1 y 5 respectivamente). Las manzanas situadas cerca a estos ejes presentan ingresos superiores en comparación con aquellas más alejadas. Este patrón resalta una correlación clara entre la proximidad a infraestructuras de transporte clave y los niveles de ingreso en las manzanas adyacentes.

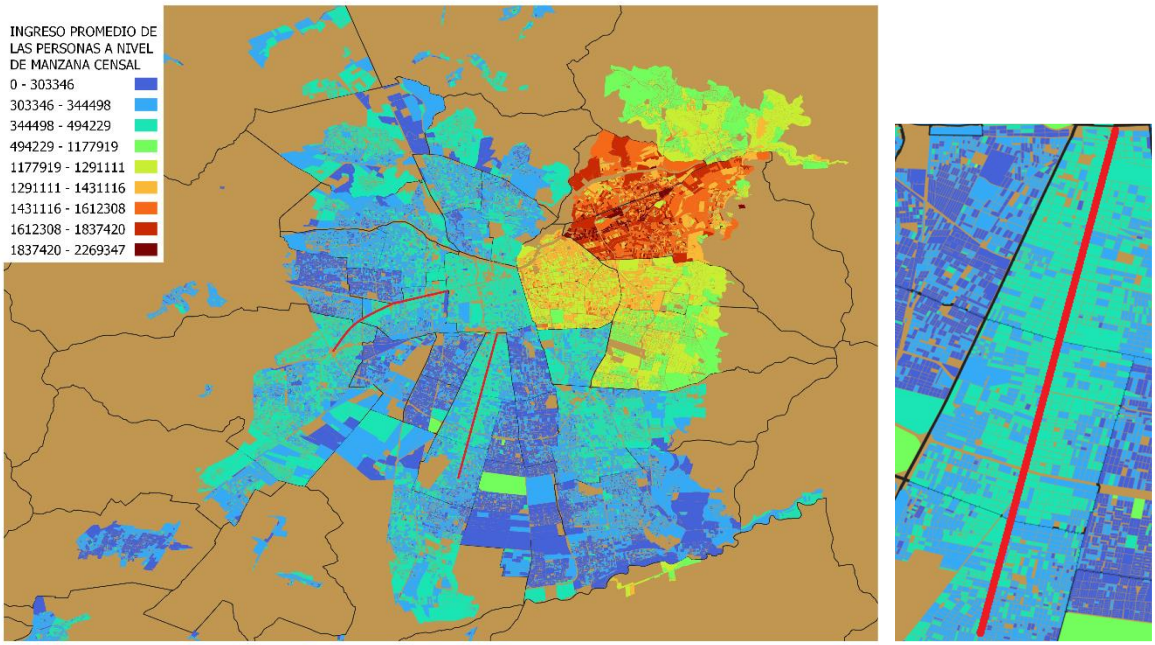


Figura 17: Ingreso promedio de las personas a nivel de manzana censal

Por otro lado, la Figura 18 ilustra los ingresos promedio de los hogares, siguiendo un patrón similar al observado en el análisis por zona censal (Figura 12), pero con más detalle. Los ingresos más altos se ubican en la periferia del sector oriente, mientras que los más bajos se encuentran en el centro de la ciudad. Esta diferencia en la distribución de los ingresos entre personas y hogares se puede explicar por la tendencia en las áreas periféricas a tener mayores tamaños de las viviendas y hogares con más personas, lo que conduce a un mayor ingreso agregado del hogar. En contraste, en el centro de la ciudad son comunes los departamentos pequeños, lo que implica hogares de tamaños más reducidos (menor cantidad de personas en el hogar) y, por ende, ingresos agregados menores por hogar.

La Figura 19 presenta la cantidad promedio de personas por hogar a nivel de manzana censal, corroborando la inferencia mencionada anteriormente respecto a las distribuciones de los tamaños de hogar en la ciudad. Se observa que en el centro de la ciudad el tamaño del hogar

es menor, mientras que en la periferia es mayor, lo que se nota aún más en el sector oriente. Al combinar estos datos con los ingresos del hogar, se puede calcular el ingreso per cápita, el cual se presenta en el Anexo y muestra resultados similares al ingreso de las personas de la Figura 17, destacando la relevancia del tamaño del hogar en la determinación de la distribución del ingreso y como este se puede analizar.

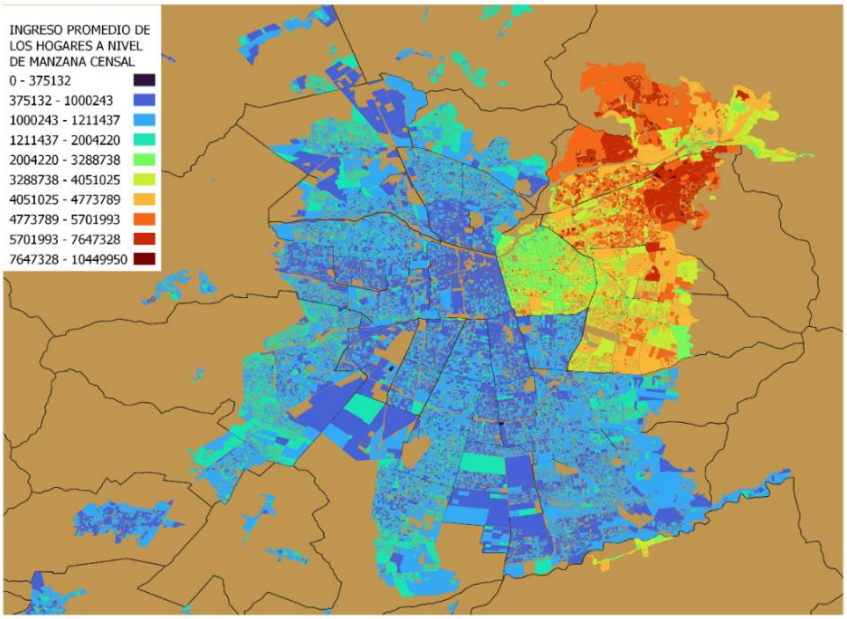


Figura 18: Ingreso promedio de los hogares a nivel de manzana censal

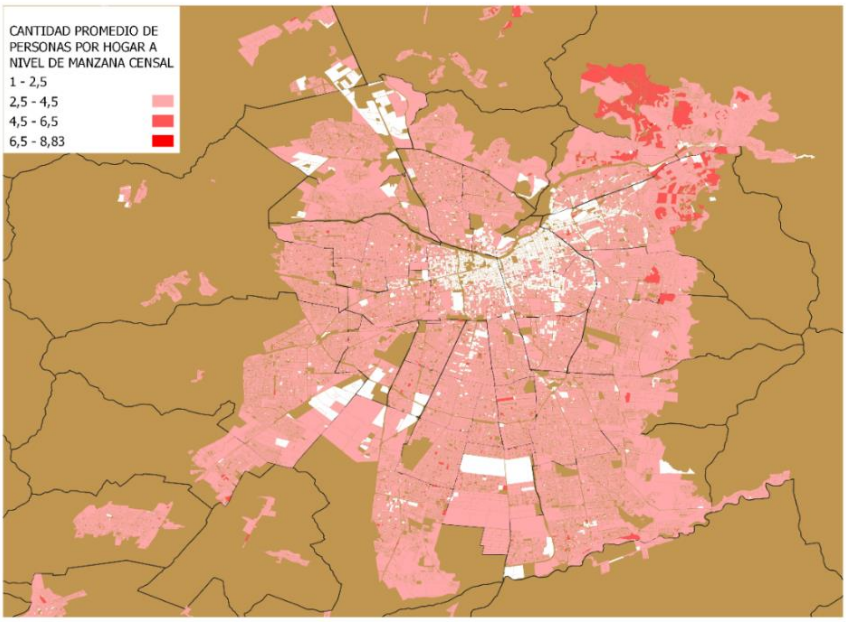


Figura 19: Cantidad de personas por hogar promedio a nivel de manzana censal

Heterogeneidad de los ingresos

Este segmento de análisis se centra en calcular diversos indicadores de heterogeneidad de ingresos en las manzanas dentro de cada zona censal. Se plantea la posibilidad de realizar, en un futuro, un análisis similar para cada manzana de la ciudad, permitiendo comparar la heterogeneidad interna dentro de ellas. En la Figura 20, se examina la desviación estándar de los ingresos de las manzanas de cada zona censal. Se destaca que las zonas censales del sector oriente tienen desviaciones estándar significativamente mayores en comparación con el resto de la ciudad, lo que sugiere una mayor variabilidad en los niveles de ingreso dentro de estas zonas censales y, posiblemente, una mayor diversidad socioeconómica.

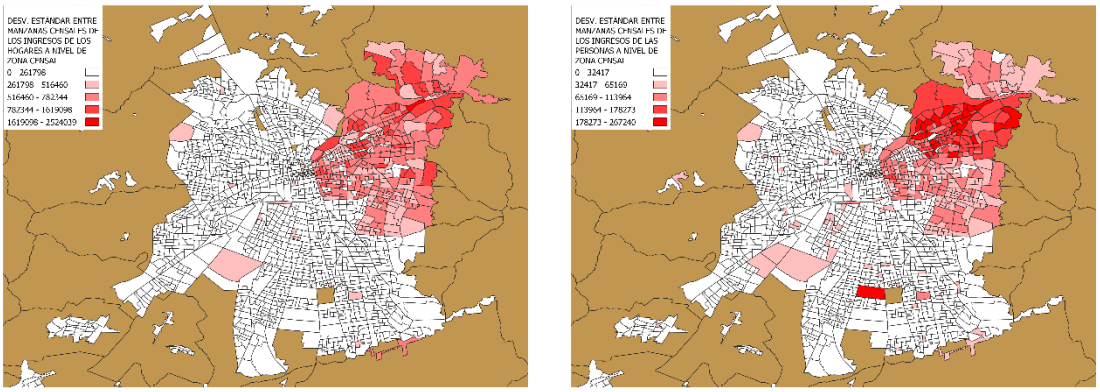


Figura 20: Desviación estándar de los ingresos promedio entre MC de cada ZC

Luego, la Figura 21 se enfoca en otro indicador de heterogeneidad de ingresos: el rango de ingresos dentro de cada zona censal. Este rango se define como la diferencia entre el ingreso promedio más alto y el más bajo de las manzanas pertenecientes a la misma zona censal. Los resultados de este indicador evidencian una vez más una mayor heterogeneidad en las zonas censales del sector oriente. Sin embargo, es importante considerar que estos indicadores pueden estar influenciados por el nivel promedio de los ingresos, ya que zonas con ingresos más altos suelen presentar rangos más altos.

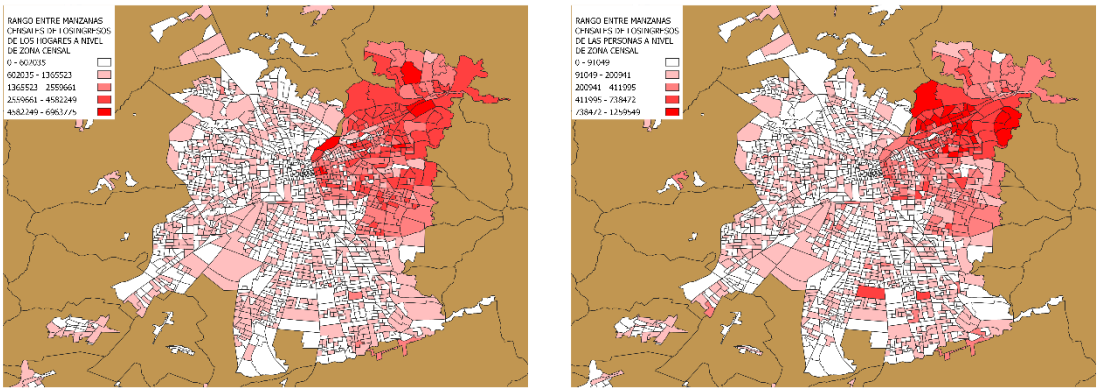


Figura 21: Rango de los ingresos promedio entre MC de cada ZC

Para contrarrestar la influencia del nivel de ingresos promedio en los indicadores de heterogeneidad, se introduce un tercer indicador: el coeficiente de variación para personas y hogares (CV_p, CV_h). Este coeficiente, calculado como la desviación estándar dividida por el promedio de ingresos de la zona censal, se ilustra en la Figura 22. Aunque las zonas censales del sector oriente siguen mostrando una heterogeneidad notable, ya no es evidente que sean las más heterogéneas de la ciudad.

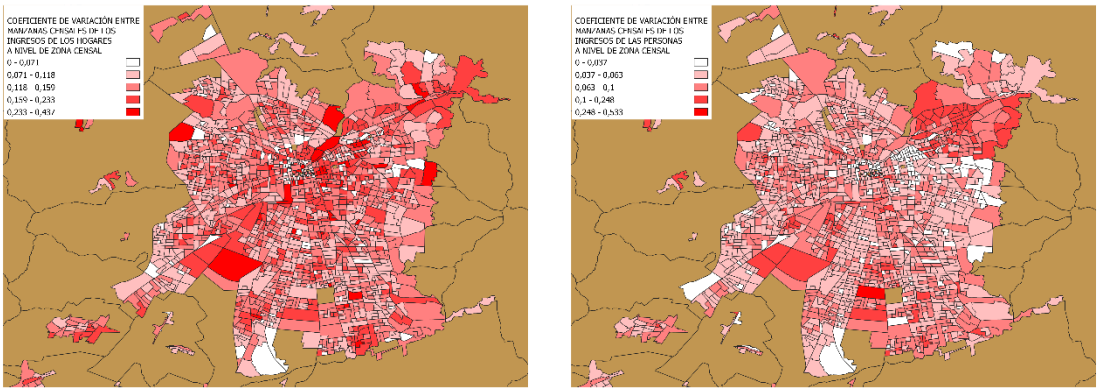


Figura 22: Coeficiente de variación de los ingresos promedio entre MC de cada ZC

Para poder analizar más a fondo este último indicador, se calculan los coeficientes de variación tanto a nivel comunal como a nivel de clústeres de comunas. En la Figura 23 el análisis comunal revela que algunas comunas son más heterogéneas que otras, destacando la comuna de Vitacura en particular. En el análisis por clústeres, se obtienen coeficientes de variación de personas (CV_p) de 9,1%, 4,50%, 5,61% y 7,00% y coeficientes de variación de hogares (CV_h) de 13,7%, 14,2%, 12,4 y 13,0% correspondientes a los clústeres de comunas de ingresos muy altos, altos, medios y bajos respectivamente. Estos resultados indican que, aunque existen diferencias de heterogeneidad entre zonas censales y comunas, la variabilidad entre clústeres a nivel de personas no es clara, ya que es mayor para las comunas de ingresos muy altos, pero baja para personas de comunas de ingresos altos. Sin embargo, a nivel de

hogares, se observa una heterogeneidad con relación directa a los clústeres de ingresos, teniendo que a mayores ingresos mayor es la heterogeneidad, aunque estas diferencias son menos pronunciadas que en el caso de las de personas.

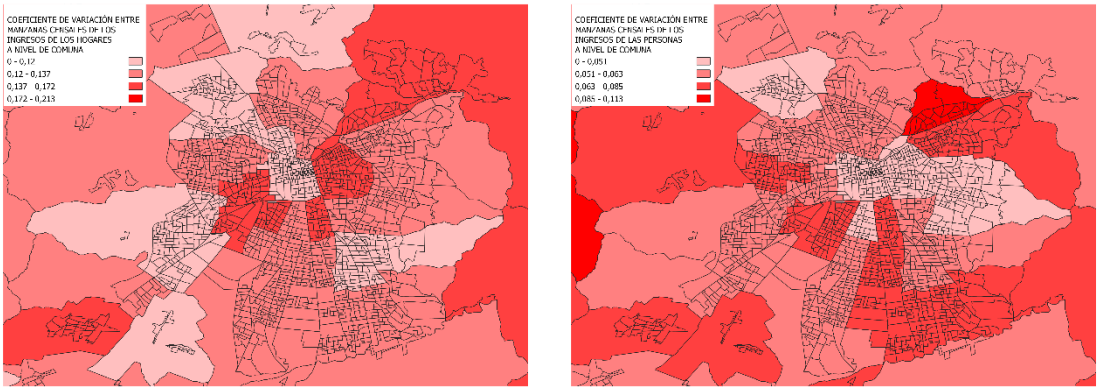


Figura 23: Coeficiente de variación de los ingresos a nivel de comuna

Finalmente, la Figura 24 presenta ejemplos de zonas censales representativas de los sectores con ingresos altos y bajos en la ciudad, específicamente en las comunas de La Reina y Puente Alto respectivamente. En estas zonas censales, la asignación alcanza el óptimo, sin errores en las distribuciones marginales de los atributos de personas, hogares y viviendas. En la zona de menores ingresos, las diferencias de ingresos entre las manzanas son menos pronunciadas, con un ingreso promedio mínimo de 890.816 y un máximo de 1.262.875, frente a un promedio general de la zona censal de 944.129. En cambio, en la zona censal de ingresos altos, las diferencias son mucho más marcadas en términos absolutos. El ingreso promedio en esta zona censal es de 3.698.631, con un mínimo de 2.925.464 y un máximo de 4.475.586 en sus manzanas, lo que revela una brecha significativa de aproximadamente un millón y medio.

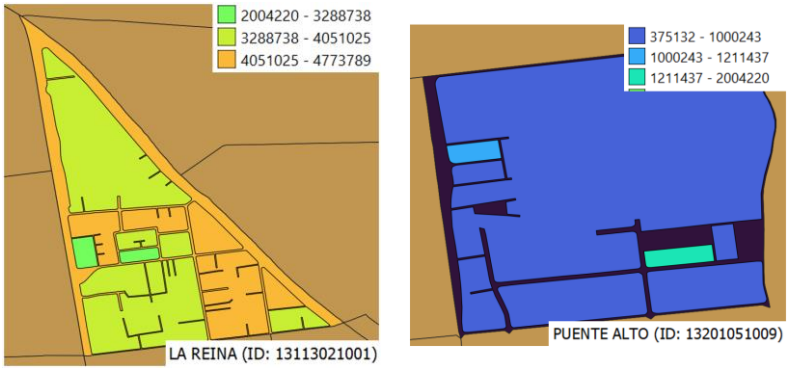


Figura 24: Ejemplos desagregación en zonas censales

Análisis del nivel de educación a nivel granular

Además del análisis de ingresos que hemos realizado a profundidad, la base de datos granular permite el estudio de otras variables relevantes a nivel de manzana, que antes de este proceso no se podían analizar, como el tipo de trabajo de las personas, de nivel de educación y, como se mencionó anteriormente, la cantidad de personas en los hogares (ver Figura 19). La Figura 25 se enfoca en el nivel de educación promedio de las personas mayores de edad a nivel de manzana censal. Para una comparación a nivel de zona censal se puede consultar el Anexo J. Según la clasificación establecida en el capítulo anterior, el 'nivel 1' de educación corresponde desde personas que no han asistido a ningún tipo de educación formal hasta personas con un máximo de educación básica completa. El 'nivel 2' de educación abarca desde educación escolar completa hasta educación técnica completa o profesional incompleta. Por último, el 'nivel 3' comprende desde una educación profesional completa hasta la realización de estudios de postgrado completos.

Los resultados mostrados en la Figura 25 revelan que, en promedio, las manzanas en el sector oriente, donde se encuentran los ingresos más altos, presentan niveles de educación superiores en las personas mayores a 18 años. Esto sugiere una correlación entre los niveles de ingreso y su educación muy marcado en la ciudad de Santiago. Esto destaca la importancia de considerar la educación junto con otros factores socioeconómicos al analizar la estructura y dinámica de las ciudades.

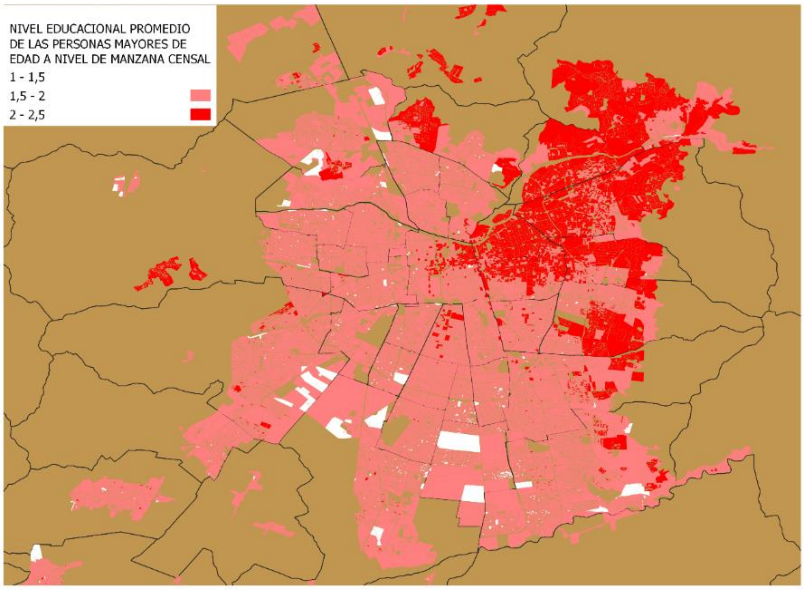


Figura 25: Nivel educacional promedio de las personas a nivel de manzana censal

Capítulo 6
Conclusiones y trabajos futuros

Esta investigación ha logrado avances significativos en la capacidad de análisis de uso de suelo y transporte urbano, particularmente en Santiago. A lo largo de esta investigación, se ha destacado la relevancia de este trabajo para su utilización en de los modelos de uso de suelo como MUSSA, para efectos de enfrentar desafíos considerables relacionados con el alto costo y la complejidad en la recolección de datos, así como los costos asociados a los métodos econométricos. En un contexto donde la tecnología y la disponibilidad de datos han avanzado significativamente, la existencia de una base de datos integrada y espacialmente detallada en Santiago, que permita caracterizar y estimar con precisión el comportamiento de los agentes urbanos y que pueda utilizarse en numerosas aplicaciones constituye un avance relevante.

El proceso de integración de múltiples fuentes y la utilización de datos censales han permitido construir métodos con alta precisión y aplicabilidad para construir y actualizar los datos en base a fuentes de datos disponibles y gratuitas. Un enfoque específico ha sido el avance en la dimensión de los agentes residenciales, evaluando comparativamente métodos de imputación de ingresos y desagregación de estos agentes en la ciudad. Los resultados obtenidos han revelado un diseño de proceso eficiente en la integración de datos y un enriquecimiento significativo en la imputación de ingresos residenciales utilizando el método Random Forest, el cual, tras comparar diferentes métodos, demostró ser consistentemente el mejor para este propósito. Además, se ha logrado una mayor precisión en la desagregación espacial de la localización de los agentes a nivel de manzana censal, luego de comparar diferentes problemas de optimización. Se encontró que optimizando un problema lineal de variables enteras con restricciones de distribuciones marginales y minimizando la norma 1 se obtienen los mejores resultados.

En consecuencia, esta tesis ha cumplido con su objetivo general de diseñar un procedimiento eficiente para la construcción y análisis de una base de datos integrada y desagregada de uso de suelo y transporte y avanzar en la construcción y análisis de la base de datos de personas (B-P). Los objetivos específicos, como la integración de datos, imputación de ingresos y evaluación de métodos de desagregación espacial, han sido alcanzados, contribuyendo a una comprensión más profunda y precisa de los patrones urbanos. Así, el aporte principal de esta tesis radica en la creación de un marco de datos urbanos integrado y detallado, actualizable a bajo costo, que permite una descripción detallada de la ciudad para diferentes estudios, como lo sería una mejor futura calibración y aplicación de modelos de uso de suelo como MUSSA. Esta integración y enriquecimiento de datos representan un avance significativo en la planificación urbana y en la modelación de la dinámica urbana, mejorando la precisión y aplicabilidad de las predicciones en contextos urbanos. Esto permite abordar el desafío de comprender el rápido crecimiento urbano y las dinámicas complejas de los sistemas urbanos,

para poder comprender la ciudad, mejorar la calidad de vida urbana y desarrollar una planificación de la ciudad informada y eficiente.

No obstante, como toda investigación, este estudio tiene sus limitaciones y por tanto trabajos futuros, sobre los cuales se puede seguir profundizando en la investigación. A continuación, se listan algunos de ellos reconocidos durante el desarrollo de este trabajo y/o al finalizarlo.

- Continuar con el proyecto de integración y enriquecimiento de esta base de datos granular, censal e integrada de datos de uso de suelo y transporte (BI-US&T). Aunque la tesis avanzó en el desarrollo de la Base de Personas (B-P), las otras bases, como la de Bienes Inmueble (B-BI) y la de Transporte y Ambiente (B-T&A), ofrecen un vasto campo para la investigación futura. El desarrollo de estos componentes, que se han descrito de manera preliminar en este trabajo, como la integración de precios, accesibilidades y el procesamiento avanzado de imágenes para la detección de objetos, son esenciales para lograr una descripción de la ciudad de una manera integral y la generación de modelos de uso de suelo y transporte más precisos y detallados.

- Esta integración de grandes cantidades de datos en esta base representa desafíos significativos de almacenamiento, procesamiento y comprensión de esta *big data*. Tradicionalmente, no hemos tenido disponible tal magnitud de información de la ciudad, y las metodologías de trabajo para manejarla están evolucionando constantemente. Es crucial entender estas metodologías para garantizar que la base final no solo sea detallada sino también eficiente. Así es como la identificación de las variables más relevantes y su análisis espacial es fundamental para el entendimiento de esta base debido a las grandes cantidades de variables. Se debe identificar aquellas que entregan información precisa, certera e importante para comprender el comportamiento de la ciudad y sus agentes.

- La metodología desarrollada en este estudio conlleva el desafío de extenderse a otras ciudades de Chile y del mundo, adaptándose a bases de datos disponibles que sean similares. Además, se puede extender esta metodología a las zonas rurales adyacentes a la ciudad, lo que permitiría explorar interacciones entre lo urbano y lo rural, proporcionando una visión integral de cómo los cambios en un área pueden influir en la otra.

En conclusión y como reflexión final, este trabajo de tesis avanzó en una dirección importante en cuanto al diseño de un sistema de datos integrados, basándose en datos disponibles de carácter desagregado y que se actualizan regularmente por agencias estatales u otras. Este estudio compara diferentes métodos para lograr un procesamiento eficiente de estos datos. Aunque aún quedan trabajos por hacer para poder modelar, comprender y aportar en la

mejora de la planificación de las ciudades, se destaca que los resultados obtenidos para esto, es decir, una base de datos censal y granular de los agentes residenciales de la ciudad con sus ingresos, es algo único, al menos en el caso de Santiago y Chile.

BIBLIOGRAFÍA

AIM (2019). *Clasificación Grupos Socioeconómicos y Manual de aplicación Chile*. https://aimchile.cl/wp-content/uploads/2022/03/Actualizacio%CC%81n-y-Manual-GSE-AIM-2019-1.pdf

AURIN (s. f.). *Home*. Recuperado 5 de diciembre de 2023, de https://aurin.org.au/

Australian Bureau of Statistics (s. f.). *Census*. Recuperado 5 de diciembre de 2023, de Australian Bureau of Statistics

BCN. (n.d.). *SIIT Estadísticas Territoriales: Población por zona Urbana-Rural, Censos 1992, 2002 y 2017*. Retrieved 14 November 2023, from https://www.bcn.cl/siit/estadisticasterritoriales/tema?id=95

Birkin, M., Clarke, M. (1988). Synthesis—A Synthetic Spatial Information System for Urban and Regional Analysis: Methods and Examples. *Environment and Planning A: Economy and Space*, *20*(12), 1645–1671. https://doi.org/10.1068/a201645

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Castanedo, F. (2013). A Review of Data Fusion Techniques. *The Scientific World Journal*, *2013*, 1–19. https://doi.org/10.1155/2013/704504

Donoso, P., & de Grange, L. (2010). A Microeconomic interpretation of the maximum entropy estimator of multinomial logit models and its equivalence to the maximum likelihood estimator. *Entropy*, *12*(10), 2077–2084. https://doi.org/10.3390/e12102077

DTPM. (s.f.). *Matrices de Viaje*. Recuperado 27 de enero de 2024, de https://www.dtpm.cl/index.php/documentos/matrices-de-viaje

Duranton, G., Puga, D. (2015). Urban Land Use. *Handbook of Regional and Urban Economics* (Vol. 5, pp. 467–560). Elsevier B.V. https://doi.org/10.1016/B978-0-444-59517-1.00008-8

Durrant-Whyte, H. F. (1988). Sensor Models and Multisensor Integration. *The International Journal of Robotics Research*, *7*(6), 97–113. https://doi.org/10.1177/027836498800700608

GOV.UK (s. f.). *HM Land Registry*. Recuperado 5 de diciembre de 2023, de https://www.gov.uk/government/organisations/land-registry

Heldt, B., Donoso, P., Bahamonde-Birke, F., Heinrichs, D. (2018). Estimating bid-auction models of residential location using census data with imputed household income. *Journal of Transport and Land Use*, *11*(1). https://doi.org/10.5198/jtlu.2018.1040

IDE Observatorio de Ciudades UC. (2019). *ISMT*. Recopilación de https://ideocuc-ocuc.hub.arcgis.com/datasets/97ae30fe071349e89d9d5ebd5dfa2aec_0/about

INE. (2018). *Manual de Usuario de la Base de Datos del Censo de Población Y Vivienda 2017*. www.ine.cl, Ministerio de Desarrollo Social. (2018). *Libro de Códigos Base De Datos*.

ISCI. (s.f.). *Líneas de Investigación*. Retrieved 21 May 2023, from https://isci.cl/nosotros/lineas-investigacion/

Jenks, G. F. (1967). The data model concept in statistical mapping. *International yearbook of cartography*, *7*, 186-190.

Khan, S., Nazir, S., García-Magariño, I., & Hussain, A. (2021). Deep learning-based urban big data fusion in smart cities: Towards traffic monitoring and flow-preserving fusion. *Computers & Electrical Engineering*, *89*, 106906. https://doi.org/10.1016/j.compeleceng.2020.106906

Kurban, H., Gallagher, R., Kurban, A., Persky, J. (2011). A Beginner's Guide To Creating Small-Area Cross-Tabulations. *Cityscape* (Vol. 13, 3).

Lomax, N., Norman, P. (2016). Estimating population attribute values in a table: "Get me started in" iterative proportional fitting. *Professional Geographer*, *68*(3), 451–461. https://doi.org/10.1080/00330124.2015.1099449

MacQueen J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, *1*, 281–297.

Martínez, F. (2018). *Microeconomic Modeling in Urban Science* (pp. 1–19). Elsevier. https://doi.org/10.1016/B978-0-12-815296-6.00001-9

Meng, T., Jing, X., Yan, Z., Pedrycz, W. (2020). A survey on machine learning for data fusion. *Information Fusion*, *57*, 115–129. https://doi.org/10.1016/J.INFFUS.2019.12.001

Müller, K., Axhausen, K. W. (2010). Population synthesis for microsimulation: State of the art. Working Paper. *Arbeitsberichte Verkehrs- und Raumplanung*, *638*. https://doi.org/10.3929/ethz-a-006127782

Office for National Statistics. (s. f.). *Census 2021*. Recuperado 5 de diciembre de 2023, de https://www.ons.gov.uk/census

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ..., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825–2830.

Rees, P. (1994). Estimating and projecting the populations of urban communities. *Environment and Planning A: Economy and Space*, *26*(11), 1671–1697. https://doi.org/10.1177/0308518X9402601101

Rees, P., Norman, P., Brown, D. (2004). A Framework for Progressively Improving Small Area Population Estimates. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *167*(1), 5–36. https://doi.org/10.1111/j.1467-985X.2004.00289.x

Ryan, J., Maoh, H., Kanaroglou, P. (2009). Population Synthesis: Comparing the Major Techniques Using a Small, Complete Population of Firms. *Geographical Analysis, 41*(2), 181–203. https://doi.org/10.1111/j.1538-4632.2009.00750.x

Shindler, M., Wong, A., Meyerson, A. (2011). Fast and Accurate k-means For Large Datasets. *Advances in neural information processing systems*.

Tarozzi, A., Deaton, A. (2009). Using census and survey data to estimate poverty and inequality for small areas. *Source: The Review of Economics and Statistics* (Vol. 91, 4). https://about.jstor.org/terms

UN. (2018). *68% of the world population projected to live in urban areas by 2050, says UN*. https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html

U.S. Census Bureau. (s. f.-a). *Longitudinal Employer-Household Dynamics*. Recuperado 5 de diciembre de 2023, de https://lehd.ces.census.gov

U.S. Census Bureau. (s. f.-b). *American Community Survey (ACS)*. Recuperado 5 de diciembre de 2023, de https://www.census.gov/programs-surveys/acs

Virtuosity. (2023, November 1). *CUBE Land: Transportation simulation software*. https://virtuosity.bentley.com/product/cube-land/

White, F. E. (1991). Data fusion lexicon. *Joint Directors of Labs*.

Yee, G. P., Rusiman, M. S., Ismail, S., Suparman, Hamzah, F. M., Shafi, M. A. (2023). K-means clustering analysis and multiple linear regression model on household income in Malaysia. *IAES International Journal of Artificial Intelligence*, *12*(2), 731–738. https://doi.org/10.11591/ijai.v12.i2.pp731-738

ANEXOS
Anexo A

Metodologías ISMT y GSE

En este anexo se presenta un resumen de las metodologías ISMT y GSE respectivamente.

Índice Socio Material Territorial (ISMT)

Índice de Calidad de la vivienda

1. Paredes exteriores de la vivienda

▪ ACEPTABLES: Hormigón, armado; albañilería, tabique forrado por ambas caras.
▪ RECUPERABLES: Tabique sin forro interior.
▪ IRRECUPERABLES: Materiales precarios o de desechos.

2. Techo

▪ ACEPTABLE: Tejas o tejuela, fibrocemento.
▪ RECUPERABLE: Fonolita; paja, coirón, totora o caña.
▪ IRRECUPERABLE: Materiales precarios o de desecho; sin cubierta en el techo.

3. Piso

▪ ACEPTABLE: Parquet, madera, piso flotante o similar; cerámico, flexit; alfombra o cubre piso.
▪ RECUPERABLE: Baldosa de cemento, radier, enchapado de cemento.
▪ IRRECUPERABLE: Piso de tierra.

Índice de Escolaridad del Jefe de Hogar

Categorías:

1. Sin instrucción: Sala cuna o jardín infantil, pre-kínder, kínder.
2. Primario: Educación básica, primaria o preparatoria (sistema antiguo).
3. Secundario: Científico-humanista, técnica profesional, humanidades (sistema antiguo), técnica comercial, industrial/normalista (sistema antiguo).
4. Profesional técnico: Técnico superior (0 a 3 años).
5. Profesional pregrado: Profesional (4 o más años).
6. Magister
7. Doctorado

Índice de Hacinamiento y Allegamiento

Hacinamiento:

Hacinamiento (persona/ dormitorio)	Categoría de hacinamiento
2,4 y menos	Sin hacinamiento
2,5 a 4,9	Hacinamiento medio
6 y más	Hacinamiento critico

Allegamiento:
Cantidad de hogares por vivienda

Ponderaciones

Escolaridad: 0.78
Calidad de la vivienda: 0.15
Hacinamiento: 0.045
Allegamiento: 0.025

GSE dependen del Ingreso per cápita

$$\frac{Ing}{n}$$

¿Economía de escala?

Elasticidad de equivalencia = 0.7 (EPF 2011,2012)

$$\frac{Ing}{n^{0.7}}$$

$si\ elasticidad = 0 \rightarrow ing$

$si\ elasticidad = 1 \rightarrow \frac{ing}{n}$

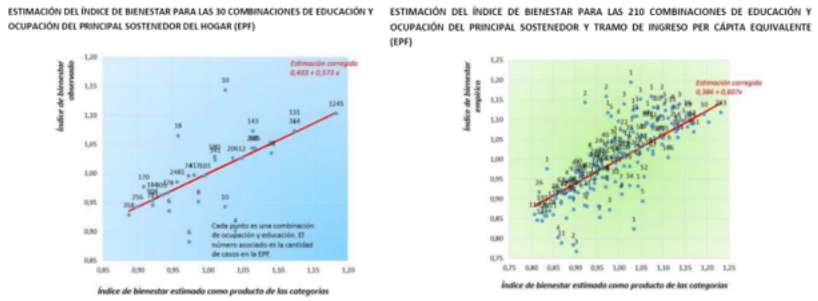
Ley de Engel (siglo XIX)

$\uparrow ing, \downarrow \%alimentación \rightarrow Bienestar$

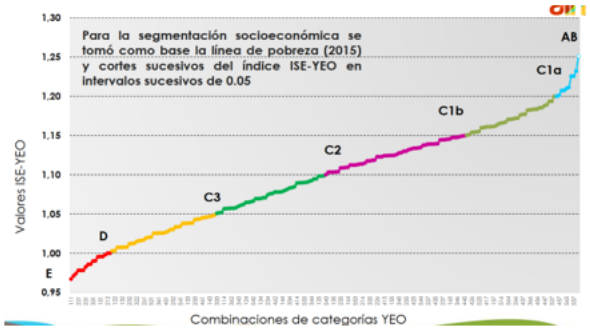
Se demuestra que $= ing\ pcápita$

$para \neq n$ tienen $= bienestar$

Bienestar para conectar: Ingresos, nivel de estudios y ocupación del Jefe de Hogar.



Clasificación:



Línea de la pobreza (CASEN 2015) = 1

Anexo B

Demostración Problema Dual de la Entropía

Tenemos el problema primal de la siguiente forma:

$$\underset{y}{\text{Max}} - \sum_{v,m} y_{vm} \ln y_{vm}$$

$$F_{mr} = \sum_{v=1}^{V} f_{vr} y_{vm} \ \forall\, m, r$$

$$\sum_{m=1}^{M} y_{vm} = 1 \ \forall v$$

Para realizar la demostración del problema dual paso a paso nos guiamos por el enfoque de Donoso & de Grange (2010), quienes demuestran que el dual del problema de máxima entropía es el problema de maximización de la verosimilitud logarítmica para los modelos logit multinomiales. Por lo tanto, igualando el problema al caso del artículo se tiene:

$$\underset{y}{\text{Max}} - \sum_{v,m} y_{vm} \ln y_{vm}$$

$$F_{m'r} = \sum_{m=1}^{M} \sum_{v=1}^{V} \delta_m f_{vrm'} y_{vm'} \ \ for\ all\ r = \{j, k, l\}\ and\ m' \in M$$

$$donde\ \delta_m = 1\ si\ m = m', 0\ si\ no$$

$$\sum_{m=1}^{M} y_{vm} = 1\ for\ all\ v$$

Función Lagrangiana:

$$L = \sum_{v,m} y_{vm} \ln y_{vm} - \sum_{r} \beta_r \left(\sum_{v,m} \delta_m f_{vrm'} y_{vm'} - F_{m'r} \right) - \sum_{v} \theta_v \left(\sum_{m} y_{vm} - 1 \right)$$

Condiciones de KKT:

$$\frac{dL}{dy_{vm}} = \ln y_{vm} + 1 - \sum_{r} \beta_r\, \delta_m f_{vrm'} - \theta_v = 0\ (a)$$

$$\frac{dL}{d\beta_r} = \sum_{v,m} \delta_m f_{vrm'} y_{vm'} - F_{m'r} = 0\ (b)$$

$$\frac{dL}{d\theta_i} = \sum_{m} y_{vm} - 1 = 0\ (c)$$

Despejando (a) y reemplazando en (b) y (c):

$$y_{vm} = \exp\left(\sum_r \beta_r \, \delta_m f_{vrm'} + \theta_v - 1\right) \ (1)$$

$$\sum_{v=1}^{V} f_{vr} \exp\left(\sum_r \beta_r \, \delta_m f_{vrm'} + \theta_v - 1\right) = F_{m'r} \ (2)$$

$$\sum_m \exp\left(\sum_r \beta_r \, \delta_m f_{vrm'} + \theta_v - 1\right) = 1 \ (3)$$

Entonces de (1):

$$y_{vm} = \frac{\exp(\sum_r \beta_r \, \delta_m f_{vrm'}) \cdot \exp(\theta_v)}{\exp(1)}$$

Y de (3):

$$\frac{exp(\theta_v)}{\exp\,(1)} = \frac{1}{\sum_m \exp(\sum_r \beta_r \, \delta_m f_{vrm'})}$$

Entonces esto en lo anterior nos da:

$$y_{vm} = \frac{\exp(\sum_r \beta_r \, \delta_m f_{vrm'})}{\sum_m \exp(\sum_r \beta_r \, \delta_m f_{vrm'})} \ (*)$$

Dual:

$$\underset{y,\beta,\theta}{\text{Max}} \ L(y,\beta,\theta) = \underset{y,\beta,\theta}{\text{Max}} \sum_{v,m} y_{vm} \ln y_{vm} - \sum_r \beta_r \left(\sum_{v,m} \delta_m f_{vrm'} y_{vm'} - F_{m'r}\right) - \sum_v \theta_v \left(\sum_m y_{vm} - 1\right)$$

Reemplazando $(*)$ en el primer término:

$$L = \sum_{v,m} y_{vm} \ln \frac{\exp(\sum_r \beta_r \, \delta_m f_{vrm'})}{\sum_m \exp(\sum_r \beta_r \, \delta_m f_{vrm'})} - \sum_r \beta_r \left(\sum_{v,m} \delta_m f_{vrm'} y_{vm'} - F_{m'r}\right) - \sum_v \theta_v \left(\sum_m y_{vm} - 1\right)$$

Usando (c) en lo anterior:

$$L = \sum_{v,m} y_{vm} \ln \frac{\exp(\sum_r \beta_r \, \delta_m f_{vrm'})}{\sum_m \exp(\sum_r \beta_r \, \delta_m f_{vrm'})} - \sum_r \beta_r \left(\sum_{v,m} \delta_m f_{vrm'} y_{vm'} - F_{m'r}\right)$$

Usando propiedades del logaritmo natural:

$$L = \sum_{v,m} y_{vm} \sum_{r} \beta_r \, \delta_m f_{vrm'} - \sum_{v,m} y_{vm} \ln \sum_{m} \exp\left(\sum_{r} \beta_r \, \delta_m f_{vrm'}\right) - \sum_{r} \beta_r \sum_{v,m} \delta_m f_{vrm'} y_{vm'} + \sum_{r} \beta_r \; F_{m'r}$$

De la última restricción sabemos que $\sum_m y_{vm} = 1$, esto en el segundo término y reordenando el tercero:

$$\cancel{\sum_{v,m} y_{vm} \sum_{r} \beta_r \, \delta_m f_{vrm'}} - \sum_{v} \ln \sum_{m} \exp\left(\sum_{r} \beta_r \, \delta_m f_{vrm'}\right) - \cancel{\sum_{v,m} y_{vm'} \sum_{r} \beta_r \, \delta_m f_{vrm'}} + \sum_{r} \beta_r \; F_{m'r}$$

$$L = - \sum_{v} \ln \sum_{m} \exp\left(\sum_{r} \beta_r \, \delta_m f_{vrm'}\right) + \sum_{r} \beta_r F_{m'r}$$

Por lo tanto, el dual será:

$$D(\beta) = -\left(- \sum_{v} \ln \sum_{m} \exp\left(\sum_{r} \beta_r \, \delta_m f_{vrm'}\right) + \sum_{r} \beta_r F_{m'r}\right)$$

$$\max_{\beta} -D(\beta) = \max_{\beta} - \sum_{v} \ln \sum_{m} \exp\left(\sum_{r} \beta_r \, \delta_m f_{vrm'}\right) + \sum_{r} \beta_r F_{m'r}$$

$$\max_{\beta} -D(\beta) = \max_{\beta} \sum_{r} \beta_r F_{m'r} - \sum_{v} \ln \sum_{m} \exp\left(\sum_{r} \beta_r \, \delta_m f_{vrm'}\right)$$

Llevando esto a nuestra formulación inicial del problema:

$$\max_{\beta} -D(\beta) = \max_{\beta} \sum_{r} \beta_r F_{mr} - \sum_{v} \ln \sum_{m} \exp\left(\sum_{r} \beta_r \, f_{vrm}\right)$$

Donde $y_{vm} = \frac{\exp(\sum_r \beta_r f_{vrm})}{\sum_m \exp(\sum_r \beta_r f_{vrm})} \; \forall v, m$ y β_r son los multiplicadores de Lagrange de las restricciones.

Anexo C

Igualación respuestas variables comunes entre el Censo y la encuesta Casen

En este anexo se presentan todas las variables comunes entre el Censo de Población y Vivienda y la encuesta Casen, para personas, hogares y viviendas y sus respuestas en el Censo, la encuesta Casen y agrupadas de la manera en la que se utilizan en el modelo de imputación de ingresos (MII) respectivamente.

Variables de las personas (x_p y z_p):

1. Parentesco con el Jefe de Hogar:
    a. (Censo) ¿Qué relación de parentesco tienen con el jefe/a de hogar?
        1 Jefe/a de hogar
        2 Esposo/a o cónyuge
        3 Conviviente por unión civil
        4 Conviviente de hecho o pareja
        5 Hijo/a
        6 Hijo/a del cónyuge, conviviente o pareja
        7 Hermano/a
        8 Padre/madre
        9 Cuñado/a
        10 Suegro/a
        11 Yerno/nuera
        12 Nieto/a
        13 Abuelo/a
        14 Otro pariente
        15 No pariente
        16 Servicio doméstico puertas adentro
        17 Persona en vivienda colectiva
        18 Persona en tránsito
        19 Persona en operativo calle
    b. (Casen) ¿Qué relación tiene [NOMBRE] con el jefe(a) de este hogar?
        1 Jefe(a) de hogar
        2 Esposo(a) o pareja de distinto sexo
        3 Esposo(a) o pareja de igual sexo
        4 Hijo(a) de ambos
        5 Hijo(a) sólo del jefe(a)
        6 Hijo(a) sólo del esposo(a)/pareja
        7 Padre o Madre
        8 Suegro(a)
        9 Yerno o Nuera

10 Nieto(a)
11 Hermano(a)
12 Cuñado(a)
13 Otro familiar
14 No familiar
15 Servicio Doméstico P. Adentro

c. Respuestas agrupadas:
1 Jefe(a) de hogar
2 Esposo(a) o pareja
3 Hijo(a)
4 Hijo(a) solo del esposo(a)/pareja
5 Padre o Madre
6 Suegro(a)
7 Yerno o Nuera
8 Nieto(a)
9 Hermano(a)
10 Cuñado(a)
11 Otro familiar
12 No familiar
13 Servicio doméstico P. Adentro

2. Nivel de educación:
   a. (Censo) El curso anteriormente declarado, cuál de los siguientes niveles corresponde:
   1 Sala cuna o jardín infantil
   2 Prekínder
   3 Kínder
   4 Especial o diferencial
   5 Educación básica
   6 Primaria o preparatorio (sistema antiguo)
   7 Científico-humanista
   8 Técnica profesional
   9 Humanidades (sistema antiguo)
   10 Técnica comercial, industrial/normalista (sistema antiguo)
   11 Técnico superior (1-3 años)
   12 Profesional (4 o más años)
   13 Magíster
   14 Doctorado
   98 No aplica
   99 Missing
   b. (Casen) ¿Cuál es el nivel más alto alcanzado o el nivel educacional actual?

1 Nunca asistió
2 Sala cuna
3 Jardín Infantil (Medio menor y Medio mayor)
4 Prekinder/Kinder (Transición menor y Transición Mayor)
5 Educación Especial (Diferencial)
6 Primaria o Preparatoria (Sistema antiguo)
7 Educación Básica
8 Humanidades (Sistema Antiguo)
9 Educación Media Científico-Humanista
1 Técnica, Comercial, Industrial o Normalista (Sistema Antiguo
11 Educación Media Técnica Profesional
12 Técnico Nivel Superior Incompleto (Carreras 1 a 3 años)
13 Técnico Nivel Superior Completo (Carreras 1 a 3 años)
14 Profesional Incompleto (Carreras 4 o más años)
15 Profesional Completo (Carreras 4 o más años)
16 Postgrado Incompleto
17 Postgrado Completo
99 No sabe/no responde

c. Respuestas agrupadas:
1 Nunca asistió
2 Sala cuna, Jardín Infantil (Medio menor y Medio mayor)
3 Prekinder/Kinder (Transición menor y Transición Mayor)
4 Educación Especial (Diferencial)
5 Primaria o Preparatoria (Sistema antiguo)
6 Educación Básica
7 Humanidades (Sistema Antiguo)
8 Educación Media Científico-Humanista
9 Técnica, Comercial, Industrial o Normalista (Sistema Antiguo
10 Educación Media Técnica Profesional
11 Técnico Nivel Superior Incompleto (Carreras 1 a 3 años)
12 Técnico Nivel Superior Completo (Carreras 1 a 3 años)
13 Profesional Incompleto (Carreras 4 o más años)
14 Profesional Completo (Carreras 4 o más años)
15 Postgrado Incompleto
16 Postgrado Completo
20 No sabe/no responde, No aplica

3. Trabajó o no:
    a. (Censo) Durante la semana pasada, ¿Trabajó o no trabajó?
    1 Por un pago en dinero o especies
    2 Sin pago para un familiar

3 Tenía empleo pero estuvo de vacaciones, con licencia, en descanso laboral, etc.
4 Se encontraba buscando empleo
5 Estaba estudiando
6 Realiza quehaceres de su hogar
7 Es jubilado, pensionado o rentista
8 Otra situación
98 No aplica
99 Missing

b. (Casen) La semana pasada, ¿trabajó al menos una hora, sin considerar los quehaceres del hogar?
1 Sí
2 No [2]
9 No sabe/no responde

c. Respuestas agrupadas:
1 Si
2 No
3 Jubilado/a
4 No sabe/no responde, No aplica

4. Empresa en la que trabaja:
   a. (Censo) En ese trabajo ¿A qué se dedica esa empresa, institución o actividad por cuenta propia?
   'A' Agricultura, ganadería, silvicultura y pesca
   'B' Explotación de minas y canteras
   'C' Industrias manufactureras
   'D' Suministro de electricidad, gas, vapor y aire acondicionado
   'E' Suministro de agua, evacuación de aguas residuales, gestión de desechos y descontaminación
   'F' Construcción
   'G' Comercio al por mayor y al por menor, reparación de vehículos automotores y motocicletas
   'H' Transporte y almacenamiento
   'I' Actividades de alojamiento y de servicios de comidas
   'J' Información y comunicaciones
   'K' Actividades financieras y de seguros
   'L' Actividades inmobiliarias
   'M' Actividades profesionales, científicas y técnicas

[2] Complementada con la pregunta: ¿Cuál es la razón o razones por la(s) que no buscó trabajo o realizó alguna gestión para iniciar una actividad por cuenta propia en las últimas cuatro semanas? Respuesta 12: Jubilado(a), pensionado(a) o montepiado(a)

'N' Actividades de servicios administrativos y de apoyo
'O' Administración pública y defensa; planes de seguridad social de afiliación obligatoria
'P' Enseñanza
'Q' Actividades de atención de la salud humana y de asistencia social
'R' Actividades artísticas, de entretenimiento y recreativas
'S' Otras actividades de servicios
'T' Actividades de los hogares como empleadores; actividades no diferenciadas de los hogares como productores de bienes y servicios para uso propio
'U' Actividades de organizaciones y órganos extraterritoriales
'Z' Rama no declarada
98 No Aplica
99 Missing

b. (Casen) ¿A qué se dedica o qué hace el negocio, empresa o institución donde usted trabaja?
1 Agricultura, ganadería, caza y silvicultura
2 Pesca
3 Explotación de minas y canteras
4 Industrias manufactureras
5 Suministro de electricidad, gas y agua
6 Construcción
7 Comercio al por mayor y al por menor
8 Hoteles y restaurantes
9 Transporte, almacenamiento y comunicaciones
10 Intermediación financiera
11 Actividades inmobiliarias, empresariales y de alquiler
12 Administración pública y defensa
13 Enseñanza
14 Servicios sociales y de salud
15 Otras actividades de servicios comunitarios, sociales y p
16 Hogares privados con servicio doméstico
17 Organizaciones y órganos extraterritoriales
999 Sin dato

c. Respuestas agrupadas:
1 Agricultura, ganadería, caza, silvicultura y pesca
2 Explotación de minas y canteras
3 Industrias manufactureras
4 Suministro de electricidad, gas y agua
5 Construcción
6 Comercio al por mayor y al por menor

7 Hoteles y restaurantes
8 Transporte, almacenamiento y comunicaciones
9 Intermediación financiera
10 Actividades inmobiliarias, empresariales y de alquiler
11 Administración pública y defensa
12 Enseñanza
13 Servicios sociales y de salud
14 Otras actividades de servicios comunitarios, sociales y p
15 Hogares privados con servicio doméstico
16 Organizaciones y órganos extraterritoriales
17 Actividades profesionales, científicas y técnicas, de servicios administrativos y de apoyo
20 Sin dato, No aplica

Variables del hogar (x_h y z_h):

1. Cantidad de hogares en la vivienda:
    a. (Censo) Entonces, contando el de usted, ¿Cuántos grupos tienen gastos separados de alimentación?
    Variable continua
    b. (Casen) ¿Cuántos hogares hay en esta vivienda?
    Variable continua
    99 No sabe/no responde
    c. Respuestas agrupadas:
    0 0 hogares
    1 1 hogar
    2 2 hogares
    3 3 hogares
    4 4 hogares
    5 5 hogares
    6 6 o más hogares

Variables del hogar construidas:

2. Cantidad de personas por grupos de edad.
    a. (Censo) ¿Cuántos años cumplidos tiene?
    Variable continua
    b. (Casen) ¿Qué edad tiene [NOMBRE]?
    Variable continua
    c. Respuesta construida:
    Para cada grupo de edades definidos por el Clustering realizado con k-means se construye la variable como la suma de las personas que pertenecen a ese grupo de edad.

3. Porcentaje de hombres en el hogar.
    a. (Censo) ¿Cuál es su sexo?
        1 Hombre
        2 Mujer
    b. (Casen) ¿Es [NOMBRE] hombre o mujer?
        1 Hombre
        2 Mujer
    c. Respuesta construida:
        Se construye la variable como la cantidad de hombres del hogar dividido en la cantidad de personas del hogar.
4. Edad, sexo, nivel educacional y tipo de trabajo del jefe de hogar.
        Estas variables se construyen para cada hogar a partir de las variables anteriormente mencionadas (Edad, sexo, nivel educacional y tipo de trabajo) para la persona que corresponda al Jefe/a de hogar.

Variables de la vivienda (x_v y z_v):

1. Tipo de vivienda
    a. (Censo) Indique el tipo de vivienda
        1 Casa
        2 Departamento en edificio
        3 Vivienda tradicional indígena (ruka, pae pae u otras)
        4 Pieza en casa antigua o en conventillo
        5 Mediagua, mejora, rancho o choza
        6 Móvil (carpa, casa rodante o similar)
        7 Otro tipo de vivienda particular
        8 Vivienda colectiva
        9 Operativo personas en tránsito (no es vivienda)
        10 Operativo calle (no es vivienda)
        0 No Aplica
        11 Missing
    b. (Casen) ¿Cuál es el tipo de vivienda que ocupa el entrevistado?
        1 Casa aislada (no pareada)
        2 Casa pareada por un lado
        3 Casa pareada por ambos lados
        4 Departamento en edificio con ascensor
        5 Departamento en edificio sin ascensor
        6 Pieza en casa antigua o conventillo
        7 Mediagua, mejora o vivienda de emergencia
        8 Vivienda tradicional indígena
        9 Rancho o choza

10 Vivienda precaria de materiales reutilizados (latas, plástic
99 No sabe/no responde

c. Respuestas agrupadas:
1 Casa
2 Departamento
3 Pieza en casa antigua o conventillo
4 Mediagua, mejora, vivienda de emergencia, rancho o choza
5 Vivienda tradicional indígena
6 Otro tipo de vivienda particular

2. Material paredes:
   a. (Censo) ¿Cuál es el material de construcción principal en las paredes exteriores?
      1 Hormigon armado
      2 Albañilería: bloque de cemento, piedra o ladrillo
      3 Tabique forrado por ambas caras (madera o acero)
      4 Tabique sin forro interior (madera u otro)
      5 Adobe, barro, quincha, pirca u otro artesanal tradicional
      6 Materiales precarios (lata, carton, plastico, etc.)
      98 No Aplica
      99 Missing
   b. (Casen) ¿Cuál es el material que predomina en los muros exteriores de la vivienda?
      1 Hormigón armado
      2 Albañilería (bloque de cemento, piedra o ladrillo)
      3 Tabique forrado por ambas caras (madera, acero, lata u otro)
      4 Tabique sin forro interior (madera u otro)
      5 Adobe, barro, quincha, pirca u otro artesanal tradicional
      6 Materiales precarios o de desecho (cartón, latas, sacos, plá
      9 No sabe/no responde
   c. Respuestas agrupadas:
      1 Hormigón armado
      2 Albañilería (bloque de cemento, piedra o ladrillo)
      3 Tabique forrado por ambas caras (madera, acero, lata u otro)
      4 Tabique sin forro interior (madera u otro)
      5 Adobe, barro, quincha, pirca u otro artesanal tradicional
      6 Materiales precarios o de desecho (cartón, latas, sacos, etc.)
      9 No sabe/no responde, No aplica

3. Material piso:
   a. (Censo) ¿Cuál es el material de construcción principal en el piso?

1 Parquet, piso flotante, cerámico, madera, alfombra, flexit, cubrepiso u otro similar, sobre radier o vigas de madera
2 Radier sin revestimiento
3 Baldosa de cemento
4 Capa de cemento sobre tierra
5 Tierra
98 No Aplica
99 Missing

b. (Casen) ¿Cuál es el material que predomina en el piso de la vivienda?
1 Parquet, madera, piso flotante o similar
2 Cerámico, porcelanato, flexit o similar
3 Alfombra o cubrepiso
4 Baldosa de cemento
5 Radier
6 Tierra
9 No sabe/no responde

c. Respuestas agrupadas:
1 Parquet, madera, piso flotante, cerámico, porcelanato, flexit o similar, Alfombra o cubrepiso
2 Baldosa de cemento
3 Radier sin revestimiento, capa de cemento sobre tierra
4 Tierra
9 No sabe/no responde, No aplica

4. Material techo:
   a. (Censo) ¿Cuál es el material de construcción principal en la cubierta del techo?
   1 Tejas o tejuelas de arcilla, metálicas, de cemento, de madera, asfálticas o plásticas
   2 Losa hormigón
   3 Planchas metálicas de zinc, cobre, etc. o fibrocemento (tipo pizarreño)
   4 Fonolita o plancha de fieltro embreado
   5 Paja, coirón, totora o caña
   6 Materiales precarios (lata, cartón, plásticos, etc.)
   7 Sin cubierta sólida de techo
   98 No Aplica
   99 Missing
   b. (Casen) ¿Cuál es el material que predomina en el techo de la vivienda?
   1 Tejas o tejuela (arcilla, metálica, cemento, madera, asfálti
   2 Losa hormigón
   3 Planchas metálicas (zinc, cobre, etc.)

4 Plancha de fibrocemento (pizarreño)
5 Fonolita o plancha de fieltro embreado
6 Paja, coirón, totora o caña
7 Materiales precarios o de desecho.
8 Sin cubierta en el techo
9 No sabe/no responde

c. Respuestas agrupadas:
1 Tejas o tejuela (arcilla, metálica, cemento, madera, etc.)
2 Losa hormigón
3 Planchas metálicas (zinc, cobre, etc.) o fibrocemento (pizarreño)
4 Otros
9 No sabe/no responde, No aplica

5. Origen del agua:
   a. (Censo) El agua que usa esta vivienda proviene principalmente de:
   1 Red publica
   2 Pozo o noria
   3 Camión aljibe
   4 Rio, vertiente, estero, canal, lago, etc.
   98 No Aplica
   99 Missing
   b. (Casen) ¿De dónde proviene el agua de la vivienda?
   1 Red pública con medidor propio
   2 Red pública con medidor compartido
   3 Red pública sin medidor
   4 Pozo o noria
   5 Río, vertiente, lago o estero
   6 Camión aljibe
   7 Otra fuente. ¿Cuál?
   9 No sabe/no responde
   c. Respuestas agrupadas:
   1 Red pública
   2 Pozo o noria
   3 Camión aljibe
   4 Río, vertiente, lago, estero, canal, etc
   9 No sabe/no responde, No aplica

6. Cantidad de dormitorios:
   a. (Censo) ¿Cuántas piezas de esta vivienda se usan exclusivamente como dormitorio?
   0 0 piezas

1 1 pieza
2 2 piezas
3 3 piezas
4 4 piezas
5 5 piezas
6 6 o más piezas
98 No Aplica
99 Missing

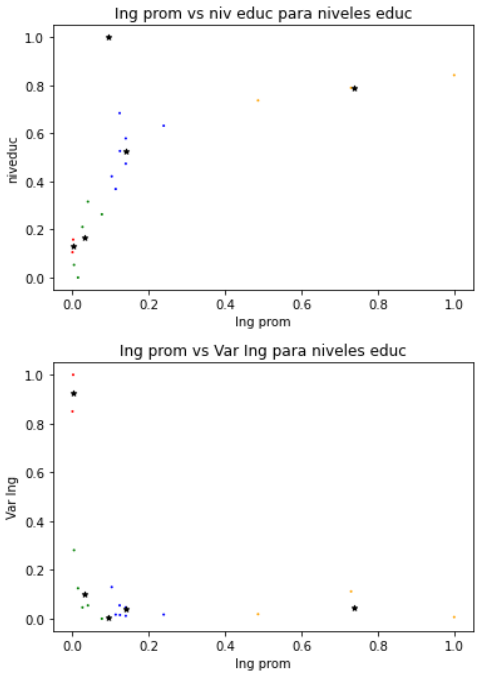
b. (Casen) ¿Cuántas piezas de cada tipo tiene la vivienda? a) Dormitorios (uso exclusivo para dormir)
Variable continua
99 No sabe/no responde

c. Respuestas agrupadas:
0 0 piezas
1 1 pieza
2 2 piezas
3 3 piezas
4 4 piezas
5 5 piezas
6 6 o más piezas
9 No sabe/no responde, No aplica

Anexo D

Análisis de Clúster de las variables: nivel de educación, tipo de trabajo y proporción de hombres.

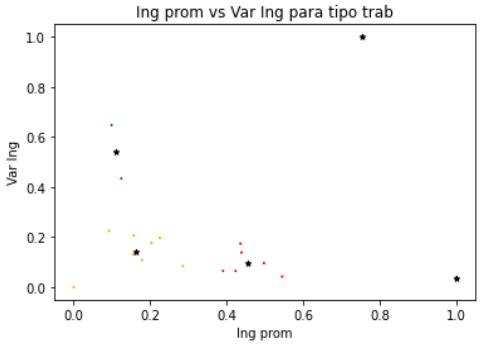
En este anexo se presentan los gráficos y tablas de análisis de clústeres de las variables educación, tipo de trabajo y proporción de hombres. En primer lugar, se presenta el análisis para el nivel de educación, donde se utiliza el método k-means donde se comparan el nivel de educación, los ingresos promedio de cada uno de estos niveles de educación y la varianza de estos ingresos. Así, se distinguen tres grupos: nivel de educación bajo (desde no poseer ninguna educación formal a educación básica), nivel de educación medio (desde educación media a educación técnica completa o profesional incompleta) y nivel de educación alto (desde educación profesional completa a postgrado completo). El análisis de clúster en el nivel de educación bajo diferencia prekinder/kinder y educación especial de este grupo debido a que poseen una mayor varianza, pero debido a las pequeñas diferencias que estos presentan con otros niveles como jardín infantil se mantienen en este grupo.



1	Nunca asistió	Nivel de educación bajo
	Sala cuna, Jardín Infantil (Medio menor y Medio mayor)	
	Prekinder/Kinder (Transición menor y Transición Mayor)	
	Educación Especial (Diferencial)	
	Primaria o Preparatoria (Sistema antiguo)	
	Educación Básica	
2	Humanidades (Sistema Antiguo)	Nivel de educación medio
	Educación Media Científico-Humanista	
	Técnica, Comercial, Industrial o Normalista (Sistema Antiguo)	
	Educación Media Técnica Profesional	
	Técnico Nivel Superior Incompleto (Carreras 1 a 3 años)	
	Técnico Nivel Superior Completo (Carreras 1 a 3 años)	
	Profesional Incompleto (Carreras 4 o más años)	
3	Profesional Completo (Carreras 4 o más años)	Nivel de educación alto
	Postgrado Incompleto	
	Postgrado Completo	
9	No sabe/no responde, No aplica	Sin datos

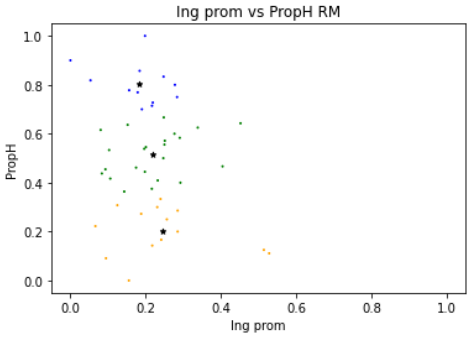
En el caso de la variable tipo de trabajo, el análisis de clúster identifica cinco grupos, uno de ingresos bajos con varianza baja, como los son trabajos en servicio domésticos, suministro de electricidad, gas y agua, entre otros. Un segundo grupo, también con ingresos bajos, pero con una varianza mayor al grupo anterior en trabajos de hoteles y restaurantes, transporte, almacenamiento y comunicaciones. Luego, se reconoce un grupo de trabajos donde se tienen ingresos medios con baja varianza, como los son la construcción, enseñanza, entre otras.

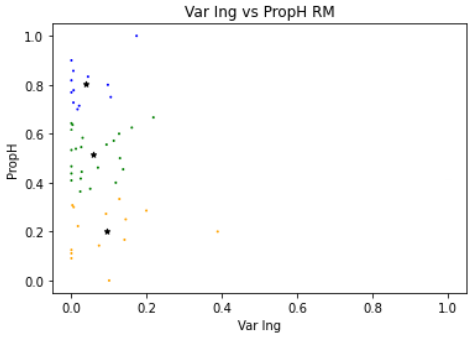
Finalmente, se encuentran dos grupos con un solo tipo de trabajo, ambos de ingresos altos, pero el primero, trabajo en industrias manufactureras, posee una alta varianza y el segundo, de actividades profesionales, científica y técnicas tiene una varianza baja.



1	Agricultura, ganadería, caza, silvicultura y pesca	Ingresos bajos, varianza baja
	Suministro de electricidad, gas y agua	
	Comercio al por mayor y al por menor	
	Intermediación financiera	
	Servicios sociales y de salud	
	Hogares privados con servicio doméstico	
	Organizaciones y órganos extraterritoriales	
2	Hoteles y restaurantes	Ingresos bajos, varianza media
	Transporte, almacenamiento y comunicaciones	
3	Explotación de minas y canteras	Ingresos medio, varianza baja
	Construcción	
	Actividades inmobiliarias, empresariales y de alquiler	
	Administración pública y defensa	
	Enseñanza	
	Otras actividades de servicios comunitarios, sociales y p	
4	Actividades profesionales, científicas y técnicas, de servicios administrativos y de apoyo	Ingresos altos, varianza baja
5	Industrias manufactureras	Ingresos altos, varianza alta
9	Sin dato, No aplica	Sin datos
99	no trabajó	Sin trabajo

Por último, el análisis de clúster de la proporción de hombres en el hogar, se obtienen 3 grupos, un grupo donde la mayoría son mujeres (0 a 33% de hombres en el hogar), otro donde la proporción es similar (33% a 66% de hombres en el hogar) y otro donde la mayoría son hombres (66% a 100% de hombres en el hogar). En este caso es difícil observar alguna diferencia entre los grupos, ya sea respecto al ingreso promedio o la varianza, pero el análisis nos ayuda a discretizar esta variable continua, sabiendo que el número óptimo de grupos es tres.





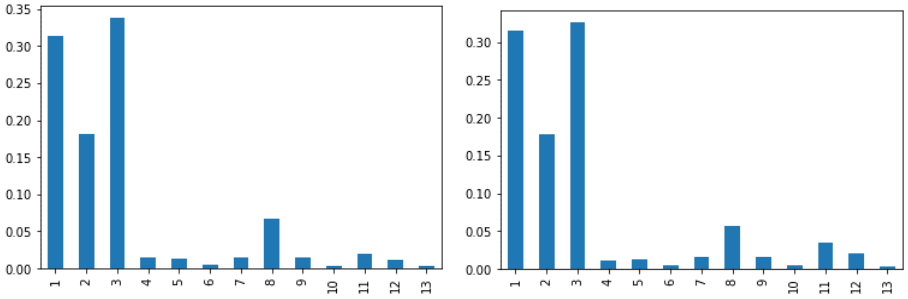
1	0 - 33%	Mayoría mujeres
2	33 - 66%	% balanceados
3	66 - 100%	Mayoría hombres

Anexo E

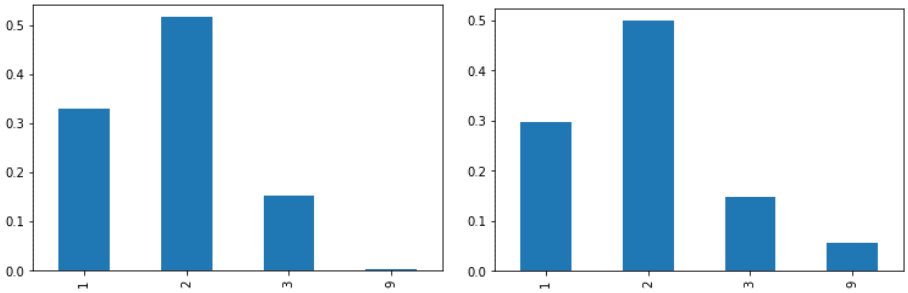
Comparación de las distribuciones de las variables de imputación en la encuesta Casen y el Censo de Población y Vivienda

En este anexo se presentan los gráficos comparativos entre las distribuciones de respuestas de las variables presentadas en el Anexo , que finalmente son utilizadas para la imputación. En el gráfico de la izquierda se presenta la distribución de las respuestas de la encuesta Casen y en el de la derecha del Censo de Población y Vivienda.

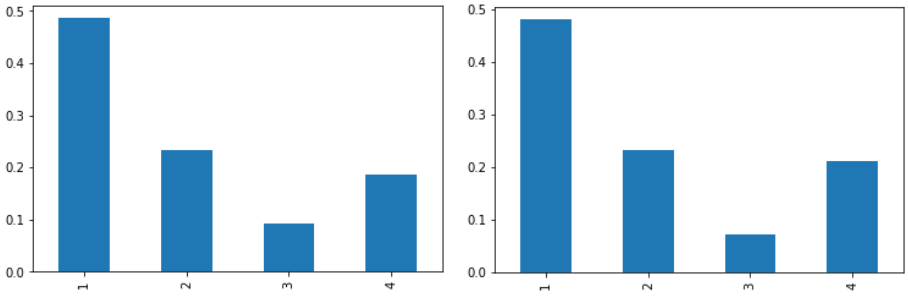
1) ¿Qué relación de parentesco tienen con el jefe/a de hogar?



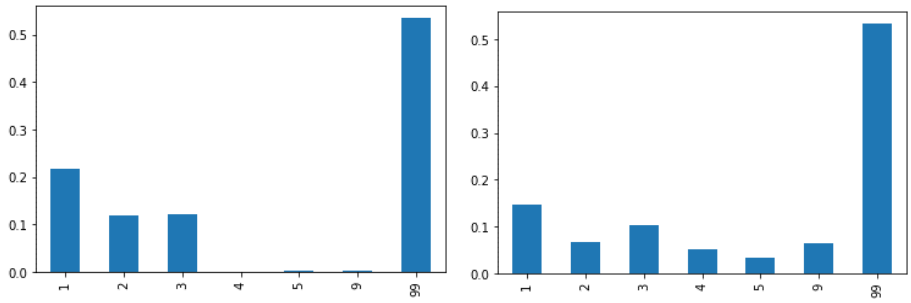
2) ¿Cuál es el nivel más alto alcanzado o el nivel educacional actual?



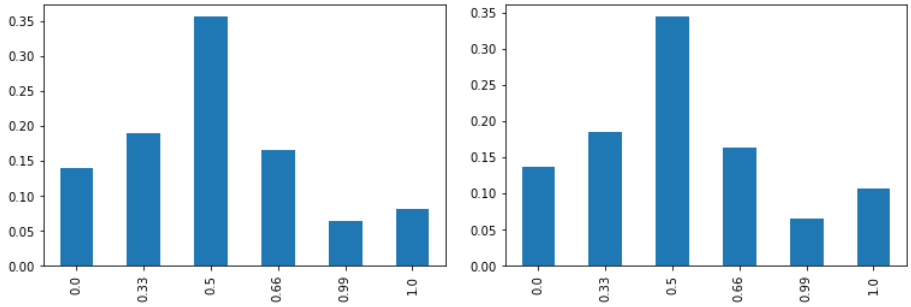
3) Durante la semana pasada, ¿Trabajó o no trabajó?



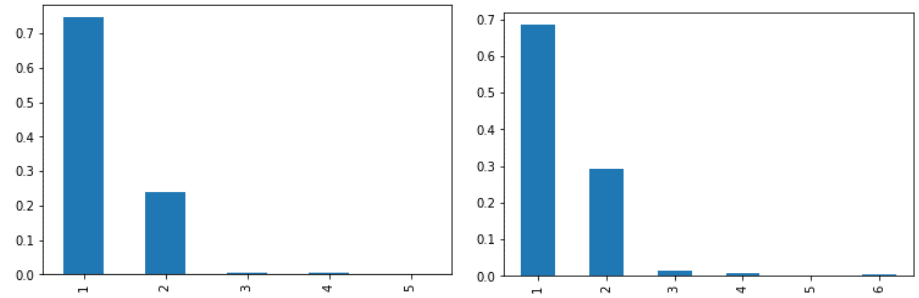
4) ¿A qué se dedica o qué hace el negocio, empresa o institución donde usted trabaja?



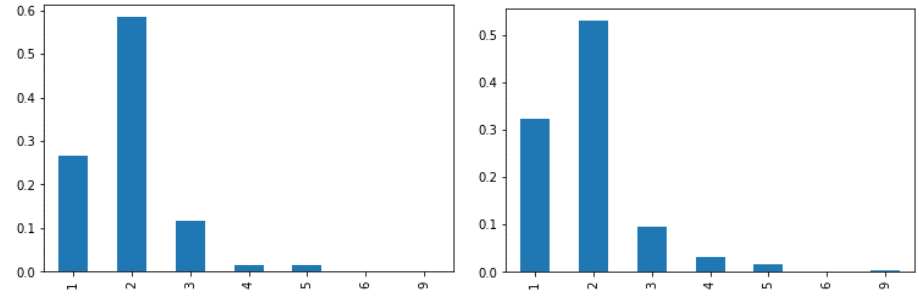
5) Porcentaje de hombres en el hogar.



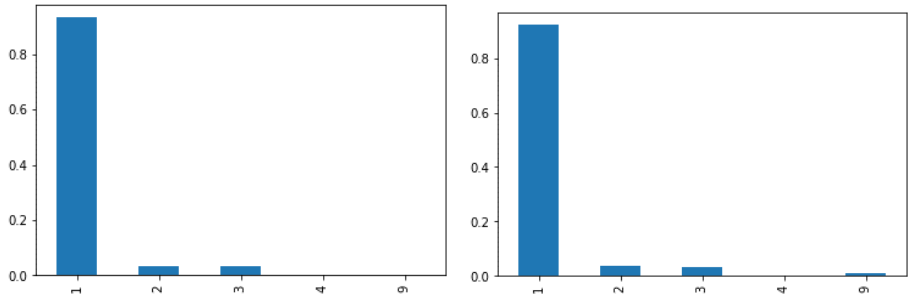
6) ¿Cuál es el tipo de vivienda que ocupa el entrevistado?



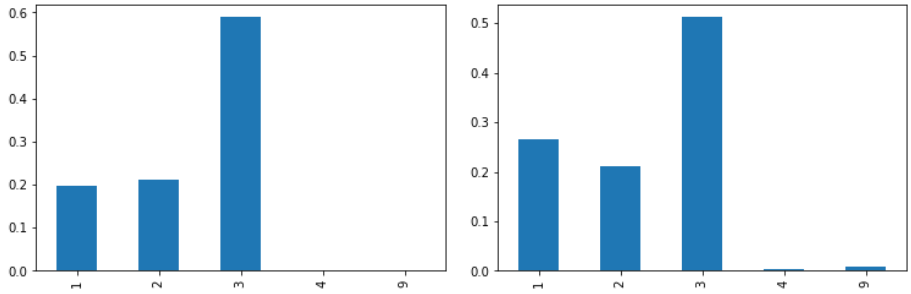
7) ¿Cuál es el material de construcción principal en las paredes exteriores?



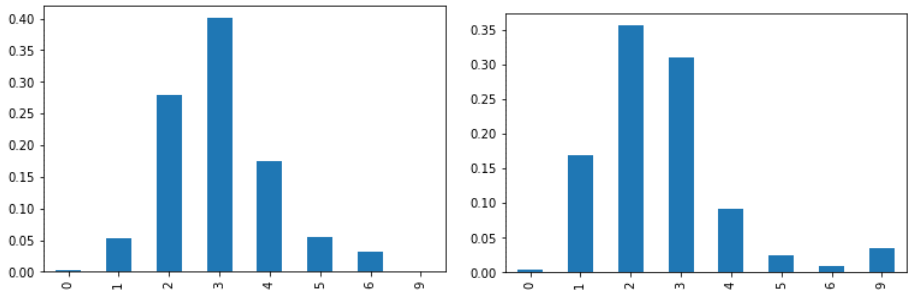
8) ¿Cuál es el material de construcción principal en el piso?



9) ¿Cuál es el material de construcción principal en la cubierta del techo?



10) ¿Cuántas piezas de esta vivienda se usan exclusivamente como dormitorio?



Anexo F

Implementación Pyomo problema de optimización norma 1

En este anexo se presenta la implementación del problema de optimización de la norma 1 utilizado en el modelo de asignación de viviendas (MAV) con el software Python, específicamente la librería Pyomo y el solver Gurobi.

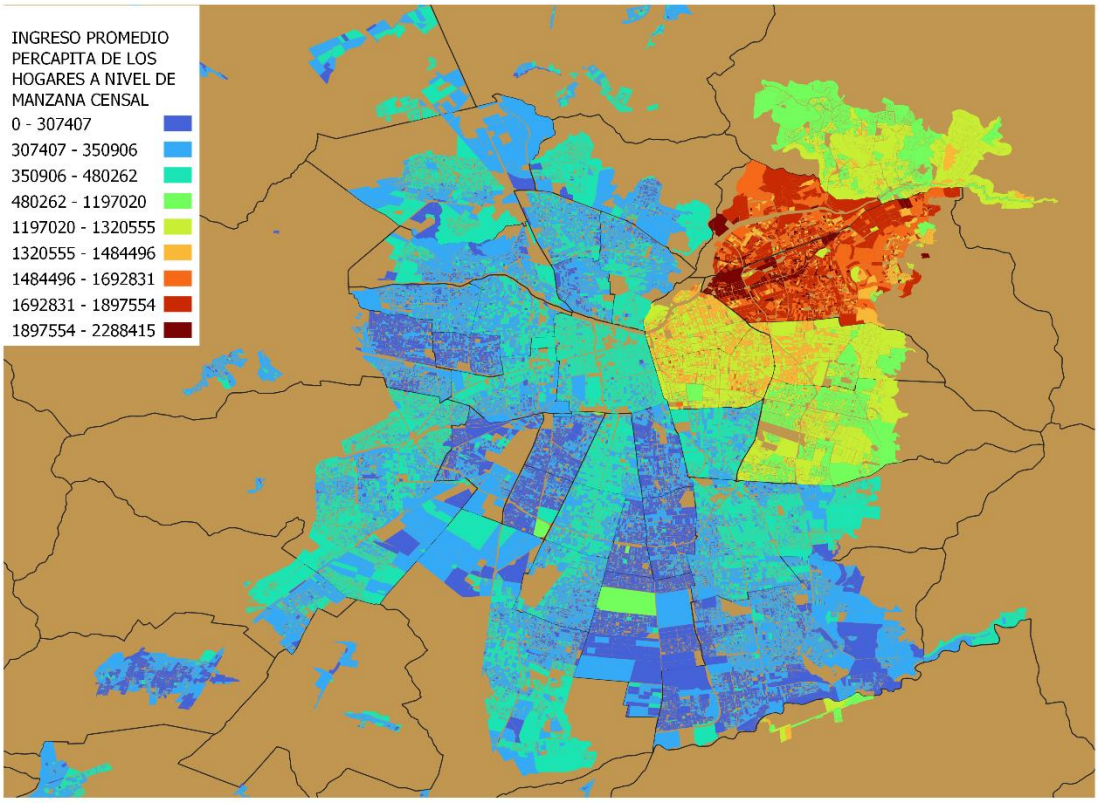
```
df_manzejzc[df_manzejzc['zc']==m] #distribuciones marginales de los atributos
df_parametros = df_parametroszc[df_parametroszc['zc']==m] #atributos de las viviendas
df_parametrosy = df_parametrosyzc[df_parametrosyzc['zc']==m] #df_parametros con el id de la vivienda

from pyomo.environ import *
from pyomo.environ import SolverFactory
import gurobipy as grb

model = ConcreteModel()

#definicion de los sets
model.V = Set(initialize=df_parametrosy['VIV'].tolist(), doc='lista de viviendas')
model.M = Set(initialize=df_manzej['ID_MANZENT'].tolist(), doc='lista de manzanas')
model.R = Set(initialize=df_parametros.columns.tolist(), doc='atributos viv, hog y pers')

#Parametros
dic = {}
for indice, fila in df_parametrosy.iterrows():
    columna_a = fila['VIV']
    for columna in df_parametrosy.columns:
        clave = (columna, columna_a)
        valor_diccionario = fila[columna]
        dic[clave] = valor_diccionario
dic_parametros = {clave: valor for clave, valor in dic.items() if 'VIV' not in clave}

model.r = Param(model.R, model.V, initialize=dic_parametros, doc='atributo r para la vivienda v')

dic = {}
for indice, fila in df_manzej.iterrows():
    columna_a = fila['ID_MANZENT']
    for columna in df_manzej.columns:
        clave = (columna, columna_a)
        valor_diccionario = fila[columna]
        dic[clave] = valor_diccionario
dic_parametros2 = {clave: valor for clave, valor in dic.items() if 'ID_MANZENT' not in clave}

# Eliminar las claves con valor '*'
claves_a_eliminar = [clave for clave, valor in dic_parametros2.items() if valor == '*']
for clave in claves_a_eliminar:
    dic_parametros2.pop(clave)
for clave, valor in dic_parametros2.items():
    dic_parametros2[clave] = int(valor)

model.F = Param(model.R, model.M, initialize=dic_parametros2, doc='suma del atributo r para la manzana m')
print('termina parametros:', time.time() - t0)

#iniciar y_vm con entropía
df_asiginic = df_asiginiczc[df_asiginiczc['zc']==m]

```

```
# Obtener la lista de valores únicos de 'VIV' y 'ID_MANZENT'
viventes = df_asiginic['VIV'].unique()
manzents = df_asiginic['ID_MANZENT'].unique()

# Crear un diccionario con todas las combinaciones posibles
initial_solution = {(viv, manz): 0 for viv in viventes for manz in manzents}
# Iterar sobre el DataFrame y actualizar el diccionario
for index, row in df_asiginic.iterrows():
    initial_solution [(row['VIV'], row['ID_MANZENT'])] = 1

#variables
model.y = Var(model.V, model.M, initialize= initial_solution,  within =Integers, bou
nds=(0.0, 1.0), doc='localizacion o no de la vivienda v en la manzana M')
#model.y = Var(model.V, model.M,  within =Integers, bounds=(0.0, 1.0),
doc='localizacion o no de la vivienda v en la manzana M')
model.a = Var(model.R, model.M, within=NonNegativeReals, bounds=(0.0, None), doc='Di
stancia de la manzana m en la variable r a la asignación')

#restricciones
print('termina variables:', time.time() - t0)
def supply_rule(model,V):
  return sum(model.y[V,M] for M in model.M) == 1
model.supply = Constraint(model.V, rule=supply_rule, doc='cada vivienda v debe estar
en una manzana')
def demand_rule(model, R, M):
    if hasattr(model, 'F') and (R, M) in model.F:
        return model.F[R, M] - sum(model.r[R, V] * model.y[V, M] for V in
model.V) <= model.a[R, M]
    else:
        return model.a[R, M] == 0
model.demand = Constraint(model.R, model.M, rule=demand_rule, doc='la cant total del
atributo r en la manz debe calzar a los atributos de las viviendas v en la manzana
m')
def demand_rule2(model,R,M):
   # for R in model.R:
   #    for M in model.M:
   if hasattr(model, 'F') and (R, M) in model.F:
        return -
model.a[R, M] <= model.F[R, M] - sum(model.r[R, V] * model.y[V, M] for V in model.V)
    else:
        return model.a[R, M] == 0
model.demand2 = Constraint(model.R, model.M, rule=demand_rule2, doc='la cant total
del atributo r en la manz debe calzar a los atributos de las viviendas v en la
manzana m')

#Función objetivo
def objective_rule(model):
    total_obj = 0
    for R in model.R:
        for M in model.M:
           # inner_sum = 0
           if hasattr(model, 'F') and (R, M) in model.F:
              #  inner_sum +=
              # for V in model.V:
               #    inner_sum += model.y[V, M] * model.r[R, V]
             #  total_obj += abs(model.F[R, M]  - inner_sum) #**2
              total_obj += model.a[R, M]
    return total_obj

model.objective = Objective(rule=objective_rule, sense=minimize, doc='Define
objective function')

       solver = SolverFactory("gurobi")
       solver.options['Presolve'] = 0
       solver.options['MIPGap'] = 0.1
       results = solver.solve(model, options = {'threads':12,'timelimit':200}, tee=T
rue)
       termination = results.solver.termination_condition
       tiempo = results.solver.time
```

Anexo G

Ingresos per cápita a nivel de manzana censal

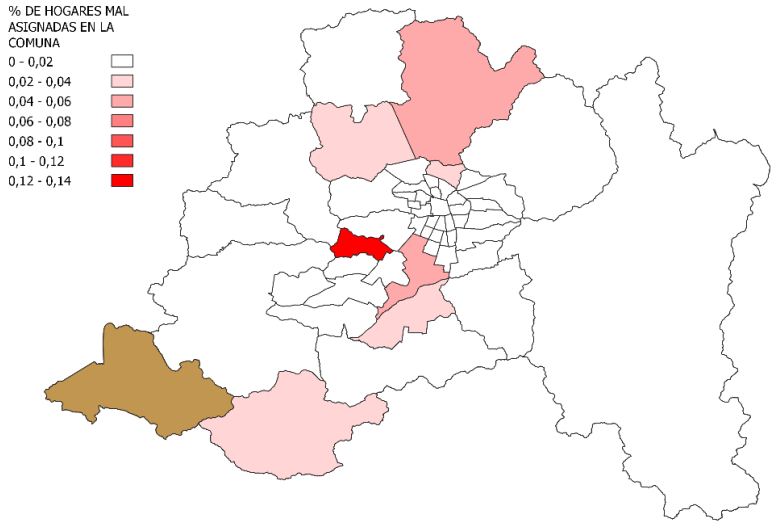
En este anexo se presenta el gráfico espacial de la distribución de los ingresos promedio per cápita de los hogares a nivel de manzana censal, es decir, el ingreso del hogar dividido en la cantidad de personas pertenecientes a este. Este gráfico muestra a un nivel granular las diferencias de los ingresos per cápita, los cuales presentan similitudes con el de los ingresos promedio de las personas (ver Figura 17), pero no son exactamente iguales.

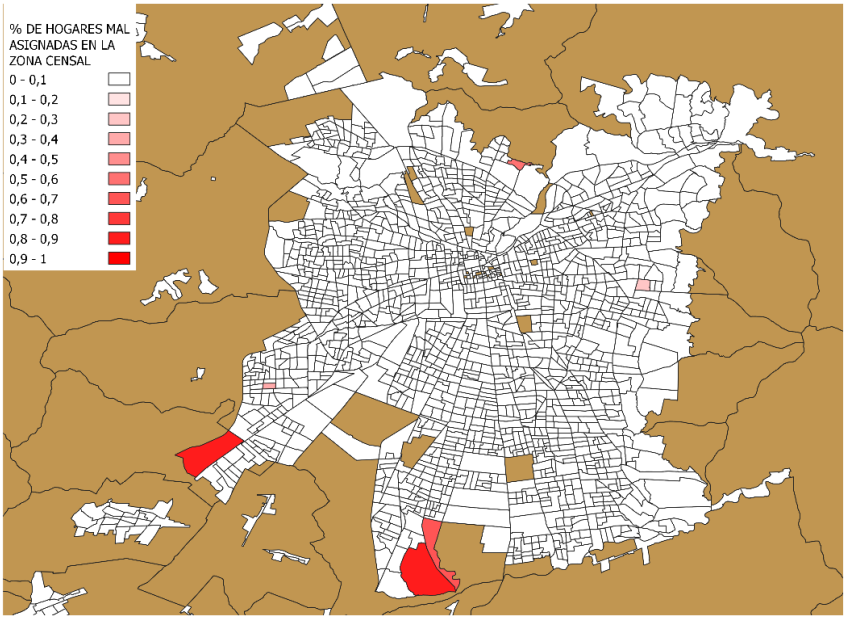


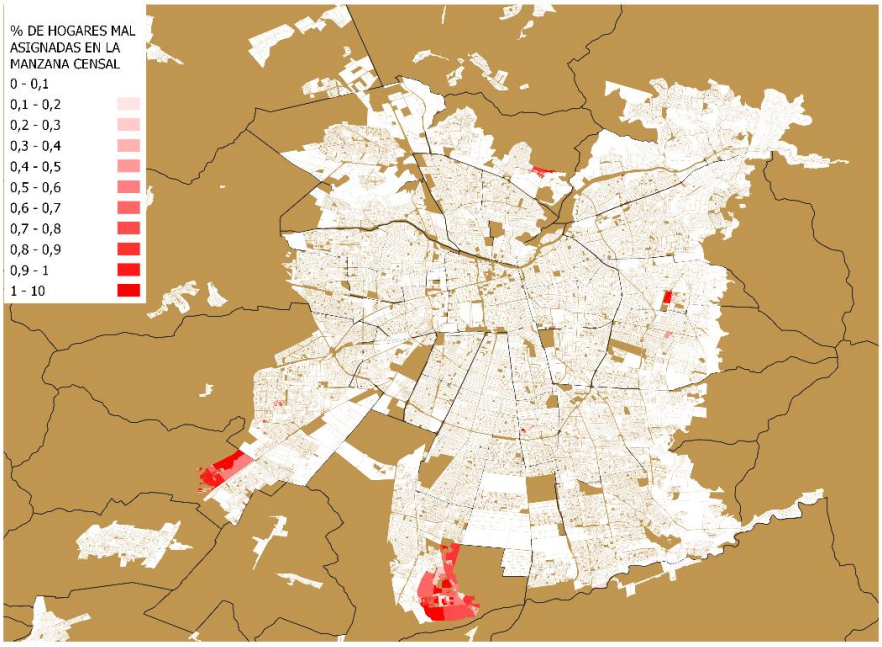
Anexo H

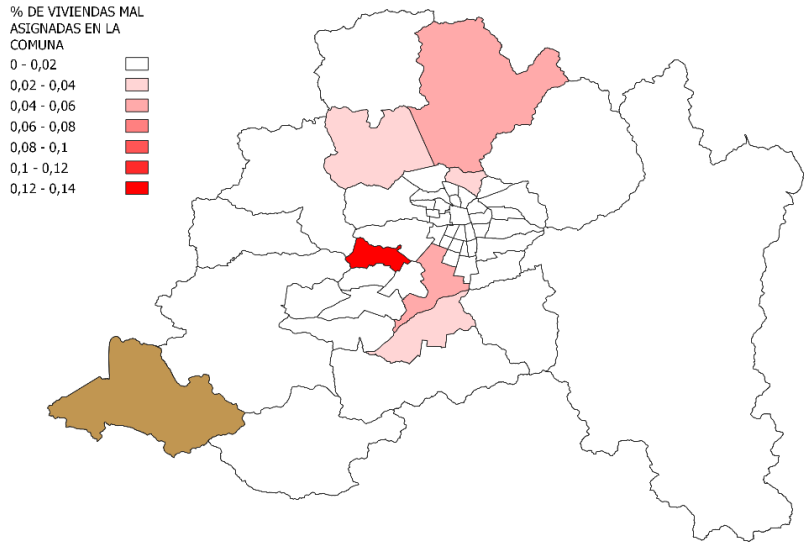
Error de asignación de los hogares y viviendas a nivel de comuna, zona y manzana censal

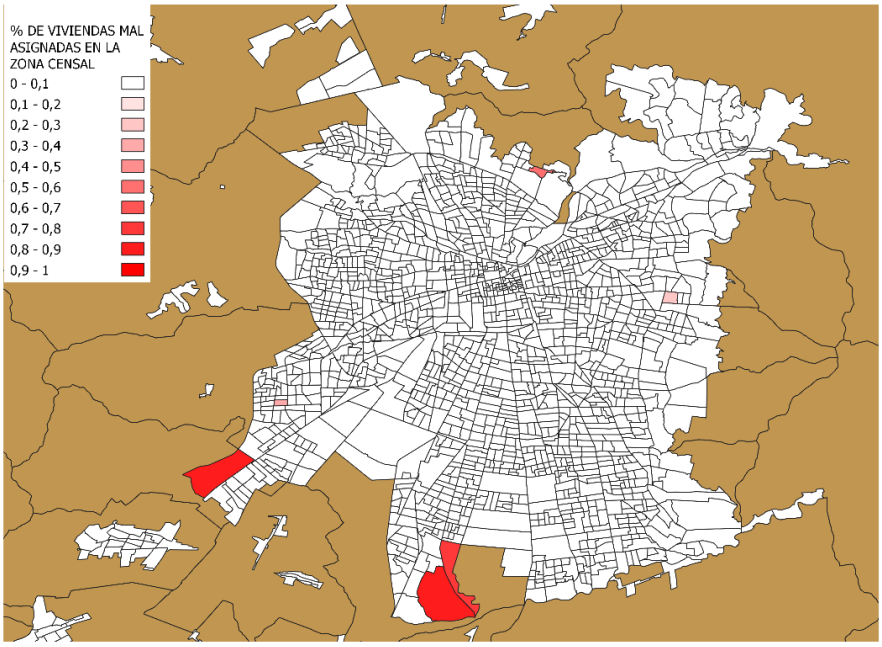
Este anexo presenta los gráficos espaciales de los errores de primer orden para hogares y viviendas a nivel de comuna, zona y manzana censal respectivamente. Estos muestran nuevamente que los errores máximos se dan en las mismas manzanas y zonas censales periféricas que corresponden a 3 casos de las 1.863 zonas urbanas desagregadas.

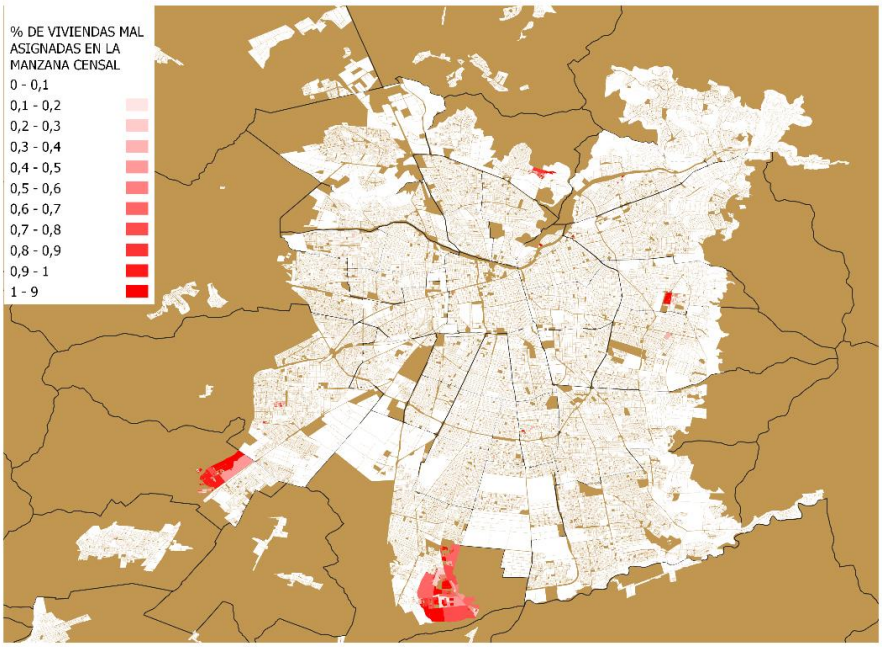





% DE HOGARES MAL ASIGNADAS EN LA MANZANA CENSAL
0 - 0,1
0,1 - 0,2
0,2 - 0,3
0,3 - 0,4
0,4 - 0,5
0,5 - 0,6
0,6 - 0,7
0,7 - 0,8
0,8 - 0,9
0,9 - 1
1 - 10


% DE VIVIENDAS MAL ASIGNADAS EN LA COMUNA
0 - 0,02
0,02 - 0,04
0,04 - 0,06
0,06 - 0,08
0,08 - 0,1
0,1 - 0,12
0,12 - 0,14

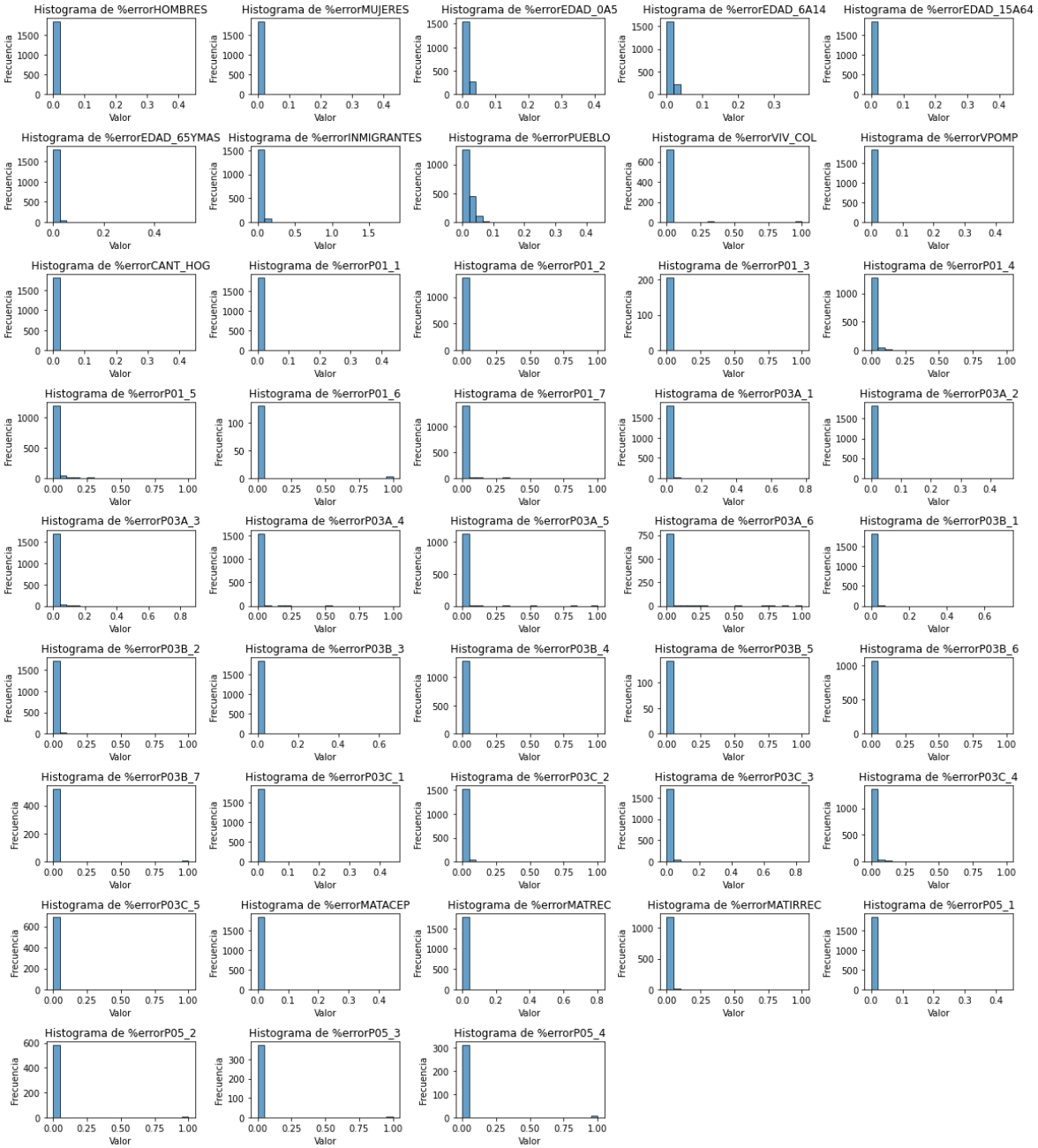

% DE VIVIENDAS MAL ASIGNADAS EN LA ZONA CENSAL
0 - 0,1
0,1 - 0,2
0,2 - 0,3
0,3 - 0,4
0,4 - 0,5
0,5 - 0,6
0,6 - 0,7
0,7 - 0,8
0,8 - 0,9
0,9 - 1


% DE VIVIENDAS MAL ASIGNADAS EN LA MANZANA CENSAL
0 - 0,1
0,1 - 0,2
0,2 - 0,3
0,3 - 0,4
0,4 - 0,5
0,5 - 0,6
0,6 - 0,7
0,7 - 0,8
0,8 - 0,9
0,9 - 1
1 - 9

Anexo I

Histogramas error de asignación de los atributos

En este anexo se presentan los histogramas de los errores de cada uno de los atributos. Es claro que en la mayoría de los casos los errores son mínimos y los errores grandes se dan en casos aislados.



Anexo J

Ejemplo comparación de la distribución de educación a nivel de zona censal versus manzana censal

Este anexo compara el nivel educacional de las personas mayores de edad en la comuna de Ñuñoa, contrarrestando los datos a nivel geográfico de zona censal con aquellos a nivel de manzana censal. Conforme a la clasificación previamente establecida, el 'nivel 1' de educación incluye a personas sin educación formal hasta aquellas con educación básica completa. El 'nivel 2' abarca desde quienes han completado la educación escolar hasta aquellos con educación técnica completa o profesional incompleta. Finalmente, el 'nivel 3' se refiere a personas con educación profesional completa hasta quienes han realizado estudios de postgrado.

Los gráficos espaciales muestran el promedio de nivel de educación de las personas mayores de edad de la zona o manzana censal según corresponda. Estos gráficos ilustran que la información a nivel de manzana censal es más detallada y revela diferencias dentro de las manzanas de una misma zona censal. Es interesante notar que incluso una zona censal incluye manzanas pertenecientes a las tres clasificaciones promedio de educación establecidas (1 - 1,5; 1,5 – 2; 2 – 2,5).