Universidad de Chile

Facultad de Economía y Negocios **ů** Facultad de Ciencias Físicas y Matemáticas

Magíster en Analítica de Negocios

Datos públicos como catalizador de la industria:

el caso de los datos catastrales del Servicio de Impuestos Internos en Chile

Borrador Final

Alumno: Cristian Luis Hernández Milla

RUT: 13.885.872-1

Profesor: Richard Weber

Fecha: Abril de 2026

Índice

Resumen Ejecutivo

El Servicio de Impuestos Internos de Chile (SII) administra 9,5 millones de predios con avalúos fiscales, superficies, destinos y geometrías poligonales. Aunque formalmente públicos (Ley 20.285), estos datos resultan inutilizables para el análisis masivo debido a tres barreras técnicas: un archivo de ancho fijo sin delimitadores, una API unitaria con bloqueo de IPs y un servicio WMS que entrega geometrías solo como imágenes PNG. Esta tesis implementa un pipeline que supera las tres barreras mediante parseo de formatos heredados, extracción distribuida con 30 túneles WireGuard y vectorización raster con GDAL, produciendo un dataset georreferenciado con geometría vectorial (≈30 cm) para 5,67 millones de predios (60,3%) en 343 comunas de Chile. Los resultados confirman que las barreras son estrictamente técnicas y superables, y que el costo de la inaccesibilidad supera los 2 millones de USD anuales (véase sección 2.2.4).

Palabras clave: datos abiertos, datos catastrales, web scraping, vectorización raster, ingeniería de datos geoespaciales, SII Chile.

Abstract

Chile's Internal Revenue Service (SII) maintains 9.5 million property records with fiscal appraisals, areas, zoning classifications, and cadastral geometries. Although formally public (Law 20,285), these data are unusable for large-scale analysis due to three technical barriers: a fixed-width file with no delimiters, an API accepting only single-property queries with IP blocking, and a WMS delivering geometries solely as PNG images. This thesis implements a pipeline that overcomes all three barriers through legacy format parsing, distributed extraction via 30 WireGuard tunnels, and raster vectorization using GDAL, producing a georeferenced dataset with vector geometry (≈30 cm) for 5.67 million properties (60.3%) across 343 Chilean municipalities. Results confirm that access barriers are strictly technical and surmountable, and that the cost of inaccessibility exceeds USD 2 million annually (see Section 2.2.4).

Keywords: open data, cadastral data, web scraping, raster vectorization, geospatial data engineering, SII Chile.

Capítulo 1

Introducción

La información que el Estado recopila en el ejercicio de sus funciones (catastros fiscales, registros de propiedad y estadísticas tributarias) tiene un doble carácter: es un insumo administrativo para el gobierno y, al mismo tiempo, un bien económico de alto valor para el conjunto de la sociedad (World Bank, 2021). Cuando esa información fluye libremente, reduce la asimetría entre actores y habilita decisiones más eficientes en mercados de todo tipo: inmobiliario, crediticio, logístico, de inversión pública. Cuando permanece atrapada en formatos opacos o accesible solo a quienes tienen capacidad técnica para extraerla, produce el efecto contrario: concentra la capacidad analítica en los actores más grandes y deja al resto operando con información de segunda mano (Open Knowledge Foundation, 2023).

La evidencia internacional es consistente. El informe *Data for Better Lives* del Banco Mundial documenta que los países que han avanzado en apertura de datos públicos muestran ganancias de productividad en sectores que dependen de información territorial: construcción, logística, planificación urbana y servicios financieros (World Bank, 2021). La Open Knowledge Foundation estima que el valor generado por la reutilización de datos públicos en economías desarrolladas equivale a varios puntos porcentuales del PIB (Open Knowledge Foundation, 2023). En América Latina, CEPAL advierte que la ausencia de políticas activas de apertura de datos perpetúa asimetrías que frenan la innovación y concentran el poder de mercado en los actores con mayor acceso a información (Nasser and Concha, 2013; CEPAL, 2022; Janssen et al., 2012).

El mecanismo económico es directo. Cuando el dato es accesible sin restricciones técnicas, los actores de menor tamaño (PyMEs, investigadores, municipios y emprendedores) pueden competir en igualdad analítica con los grandes conglomerados.

Esta nivelación tiene efectos sistémicos: mercados con mayor simetría de información son más eficientes, más transparentes y más resistentes a la corrupción (Broxterman and Zhou, 2023). Por el contrario, cuando el acceso al dato requiere capacidades técnicas sofisticadas o contratos con proveedores privados, se reproduce una forma de exclusión que no es neutral: beneficia estructuralmente a quienes ya tienen más.

Chile tiene legislación de acceso a la información pública desde 2008. La Ley 20.285 de Transparencia (Biblioteca del Congreso Nacional de Chile, 2008) consagra el derecho de cualquier persona a solicitar y recibir información de los órganos del Estado. Los datos catastrales del SII son información pública por naturaleza: no contienen datos personales protegidos, sino registros administrativos del patrimonio inmobiliario del país. Y sin embargo, la publicación formal no equivale a acceso real cuando los datos se entregan en formatos que requieren conocimientos de programación para ser leídos, cuando la API no acepta consultas masivas, y cuando las geometrías del territorio están encerradas en imágenes PNG (Cetl et al., 2023).

1.1 El dato catastral del SII: un activo público inutilizable

El Servicio de Impuestos Internos de Chile (SII) administra el catastro fiscal de todos los predios urbanos y rurales del país: aproximadamente 9,5 millones de registros que incluyen avalúos fiscales actualizados semestralmente, valor comercial de suelo por metro cuadrado, superficie de terreno y construcción, destino predial, coordenadas geográficas y la geometría vectorial de cada lote. Es, en términos de cobertura y actualización, uno de los conjuntos de datos territoriales más completos de América Latina.

Este dato es formalmente público. La Ley 20.285 de Acceso a la Información (Biblioteca del Congreso Nacional de Chile, 2008) garantiza el derecho a consultarlo. El SII lo publica en su sitio web. Y sin embargo, en la práctica, es inutilizable para el análisis masivo.

El problema no es legal. Es técnico. Los datos se publican en tres formatos que, de manera individual y combinada, impiden su uso analítico directo: un archivo de texto plano de ancho fijo sin cabeceras ni delimitadores (codificación `latin-1`, 3 GB), cuya estructura es característica de los sistemas de procesamiento por lotes de entornos mainframe; una API que solo acepta consultas predio a predio con bloqueo automático de IPs que consultan en volumen; y un servicio de mapas que

entrega las geometrías prediales únicamente como imágenes PNG, sin formato vectorial descargable. El resultado es un dato visualmente accesible pero inutilizable para el análisis automatizado: cualquiera puede *ver* un predio en el visor del SII; nadie puede *analizar* todos los predios de Chile sin superar barreras técnicas significativas (Janssen et al., 2012; Open Knowledge Foundation, 2023).

1.2 Contribución de esta AFE

Esta AFE aborda el problema en dos dimensiones complementarias. Por un lado, un desafío técnico de ingeniería de datos geoespaciales que requiere métodos avanzados de web crawling, parseo de formatos heredados y vectorización de cartografía raster para construir el dataset nacional. Por otro lado, un desafío de accesibilidad que demanda el desarrollo de una interfaz que permita a actores con distintos niveles de conocimiento técnico consultar y descargar los datos sin escribir una línea de código.

La solución concreta es doble: un pipeline de extracción y procesamiento que supera las tres barreras técnicas del SII, produciendo un dataset georreferenciado con geometría vectorial para 5,67 millones de predios (60,3% del total nacional) en 343 comunas de Chile; y la plataforma **Catastral.cl**, que distribuye ese dataset públicamente a través de una interfaz web sin requerir conocimientos de programación. Ambos productos no existían previamente en el ecosistema de datos públicos de Chile.

Para efectos de evaluación académica, conviene distinguir de manera explícita qué parte de este trabajo constituye una contribución original y qué parte corresponde a la integración competente de tecnologías existentes.

Las **contribuciones originales** de esta AFE son cuatro. Tres son de carácter técnico-metodológico:

1. **El descubrimiento del comportamiento escala-dependiente del WMS del SII.** El servidor cartográfico del SII solo renderiza los polígonos prediales cuando la consulta se realiza por debajo de un umbral de escala determinado, y los identifica con el valor de píxel DN=182 en el estilo vigente. Este hecho no estaba documentado en ninguna fuente pública y fue establecido empíricamente en este trabajo (Capítulo 6). Es el hallazgo que hace posible la vectorización masiva del catastro: sin él, el servicio WMS devuelve imágenes sin predios y el método simplemente no funciona.

2. **El pipeline de dos pasadas (zoom 19 + zoom 17) con re-descarga focalizada.** Es un método diseñado en este trabajo para un problema específico que ninguna herramienta existente resuelve: la cobertura geométrica incompleta que deja una única pasada de descarga, causada por el comportamiento escala-dependiente anterior. La combinación de una segunda pasada a resolución distinta, la re-descarga focalizada mediante *supercells* y el join espacial multi-pasada con tolerancia geométrica constituye un aporte metodológico replicable para la vectorización de cualquier WMS con comportamiento similar.

3. **El filtrado por *compactness* para clasificar geometrías catastrales.** El índice de compacidad es una métrica conocida en morfometría espacial, pero su aplicación como criterio de clasificación para separar polígonos prediales válidos de artefactos de renderizado en cartografía catastral vectorizada es una aplicación novedosa desarrollada en esta tesis.

La cuarta contribución es de carácter conceptual, y es la tesis central del trabajo: **la caracterización de las barreras de acceso al dato catastral chileno como un problema de ingeniería, no legal ni político**. Esta caracterización se formuló como hipótesis falseable (Capítulo 3) y quedó confirmada empíricamente por la construcción del dataset completo usando exclusivamente fuentes públicas y métodos estándar. Su implicancia es directa para la política pública: la inaccesibilidad del dato no es una decisión normativa que haya que disputar, sino una deuda técnica que el propio Estado puede saldar a bajo costo.

A estas contribuciones se suma una abstracción metodológica de alcance más acotado, el concepto de `rol_base` para vincular el rol predial tributario con el lote físico (Capítulo 6), que resuelve la integración entre la dimensión tabular y la dimensión espacial del catastro.

El resto del sistema corresponde a **integración de tecnologías existentes**: la polygonización raster se apoya en GDAL, la extracción distribuida en WireGuard y espacios de nombres de red de Linux, el almacenamiento y la consulta espacial en PostgreSQL/PostGIS, y la plataforma de distribución en FastAPI, React y almacenamiento de objetos S3. En estos componentes el mérito del trabajo es de diseño, dimensionamiento e implementación de sistemas, no de descubrimiento, y así debe leerse.

El resto del documento se organiza como sigue: el Capítulo 2 describe el problema y sus tres barreras con detalle; el Capítulo 3 formula la hipótesis y los objetivos;

el Capítulo 4 desarrolla el marco teórico; el Capítulo 5 presenta la metodología; el Capítulo 6 detalla el desarrollo e implementación; el Capítulo 7 reporta los resultados y casos de uso; y el Capítulo 8 cierra con la discusión, limitaciones y trabajo futuro.

Capítulo 2

Planteamiento y Formulación del Problema

2.1 El mercado que opera a ciegas

La inaccessibilidad del dato catastral genera asimetría de información en múltiples mercados. El más expuesto es el inmobiliario.

Esta inaccesibilidad ha generado un mercado secundario de intermediarios que comercializan el dato catastral sin agregar inteligencia significativa. Diversos operadores privados y corredores de datos territoriales basan su modelo de negocio en superar las barreras técnicas de extracción del SII y revender la información estructurada a precios que, según cotizaciones observadas durante 2024–2025, oscilan entre los 2.000 y los 6.000 USD por dataset comunal (el costo agregado de este mercado se analiza en la sección 2.2.4). El valor que ofrecen no radica en análisis avanzados ni en inteligencia territorial, sino en el simple hecho de haber logrado acceder y estructurar un dato que es formalmente público.

El mercado inmobiliario comercial en Chile (oficinas, bodegas, locales, industria) opera sobre transacciones de alto valor donde cada decisión de inversión puede involucrar decenas de millones de dólares. En mercados maduros como Estados Unidos o el Reino Unido, estas decisiones se apoyan en plataformas de inteligencia territorial que integran datos catastrales, transacciones registradas, permisos de edificación y variables socioeconómicas en tiempo real (Broxterman and Zhou, 2023). En Chile, ese ecosistema no existe para la mayoría de los actores.

Esta asimetría tiene dos efectos medibles. Primero, concentra la capacidad analítica

en los actores con departamentos técnicos propios: grandes fondos de inversión, consultoras multinacionales, o el Estado mismo. El resto del mercado (PyMEs inmobiliarias, desarrolladores regionales, periodistas de datos, investigadores universitarios y municipios) opera con información de segunda mano o sin información (Janssen et al., 2012). Segundo, encarece la toma de decisiones: cuando los datos no están disponibles en forma estructurada, las decisiones se fundamentan en la experiencia tácita del consultor individual, lo que introduce sesgos no auditables y eleva el costo de due diligence (Redman, 2008).

Un ejemplo ilustrativo es el de la prospección de ubicaciones para estaciones de servicio. La identificación de un sitio óptimo requiere cruzar superficie de terreno disponible, uso de suelo permitido, flujo vehicular y contexto de precios en la zona. Con el dato catastral accesible, este análisis puede automatizarse completamente sobre los 9,5 millones de predios del país en 30 minutos, como demuestra la plataforma **combustible.tremen.tech** desarrollada como caso de uso aplicado de esta tesis. Sin ese dato, el proceso requiere scouts en terreno durante 6 a 12 semanas por región.

La brecha entre disponibilidad formal y acceso efectivo tiene consecuencias documentadas por actores del ecosistema. Profesionales del sector geoespacial y de planificación urbana en Chile han descrito sus experiencias al intentar acceder a los datos catastrales del SII: emprendedores de tecnología inmobiliaria reportan la imposibilidad de construir geocoders prediales porque el SII no entrega la información fuente y sus servicios WMS no son estables; investigadores en planificación urbana señalan haber recurrido al Consejo de la Transparencia para obtener los polígonos catastrales, sin resultado; y consultores geoespaciales constatan que la información existe, pero su formato de distribución la hace inutilizable en flujos de trabajo SIG.

2.2 Las tres barreras de acceso

El análisis de los mecanismos por los cuales el dato catastral permanece inaccesible permite identificar tres barreras estructurales, de naturaleza estrictamente técnica. El dato no está protegido por ley; está bloqueado por su forma de distribución.

2.2.1 Barrera 1: el archivo de ancho fijo

El SII publica semestralmente el **Rol Semestral de Contribuciones de Bienes Raíces**, un archivo con el registro fiscal de todos los predios de Chile. El archivo es técnicamente accesible: cualquier contribuyente con RUT y clave del

SII puede descargarlo desde la sección "Detalle Catastral y Rol de Cobro" del portal institucional (https://www4.sii.cl/sismunInternet6/?caller=DETALLE_CAT_Y_ROL_COBRO). La barrera no es legal ni de acceso, sino de formato.

El archivo, denominado `BRTMPNACROL_NAC_2025_2` en su versión más reciente, es un texto plano sin cabeceras, con campos de longitud fija definidos por posición de carácter (Servicio de Impuestos Internos, 2025b). Para extraer el avalúo total de un predio se debe leer exactamente desde la posición 82 hasta la 96 de cada línea. No existe delimitador de columnas, no existe cabecera descriptiva, y la codificación es `latin-1`. Este formato, diseñado para sistemas mainframe de los años ochenta, no puede ser abierto con ninguna herramienta de análisis convencional (Excel, Google Sheets, Power BI) sin un proceso previo de parseo específico. El ciudadano que descarga el archivo de 3 GB no puede leer su contenido sin escribir código (Redman, 2008).

2.2.2 Barrera 2: la API sin capacidad de consulta masiva

El SII expone la API `getPredioNacional` que devuelve los datos completos de un predio dado su código de comuna, manzana y número de predio: coordenadas, avalúos, destino, superficie construida, área homogénea y valor comercial por metro cuadrado. Esta API está diseñada para consultas unitarias. No existe un endpoint que permita obtener todos los predios de una comuna ni aplicar filtros sobre el catastro completo. Para construir el dataset nacional completo, con 9,5 millones de predios, sería necesario realizar 9,5 millones de llamadas individuales. Adicionalmente, el servidor detecta y bloquea IPs que realizan consultas en volumen, haciendo prácticamente imposible la extracción masiva desde una conexión convencional.

2.2.3 Barrera 3: datos espaciales atrapados en imágenes

El visor cartográfico del SII (https://www4.sii.cl/mapasui) muestra los polígonos de cada predio en un mapa interactivo. Esta geometría es extraordinariamente valiosa: permite calcular áreas reales, identificar colindancias, construir grillas de valor de suelo y ejecutar consultas espaciales. Sin embargo, el SII solo expone esta geometría a través de un servicio WMS que entrega imágenes PNG renderizadas (Open Geospatial Consortium, 2006). El servicio no ofrece los polígonos como datos vectoriales descargables. El SII declara en su documentación que no cuenta con la información de los polígonos en formato vectorial, afirmación que contradice lo que se ve en el visor, donde los predios aparecen claramente delimitados como polígonos

independientes.

2.2.4 El costo económico de la inaccesibilidad

La barrera técnica que impide el acceso al dato catastral tiene un costo económico estimable, que puede observarse directamente en los precios que hoy cobra el mercado por sortearla. Se identifican tres modelos de comercialización vigentes en Chile: (i) la venta de datasets catastrales estructurados por comuna, en un rango de 2.000 a 6.000 USD por comuna; (ii) la venta del servicio de extracción (*scraping*) del catastro, que el propio autor comercializó a 5.000 USD por 60 comunas (≈83 USD por comuna); y (iii) el acceso por consulta unitaria bajo modelo SaaS, como el de operadores de datos de movilidad, a razón de aproximadamente 1 UF por cada 50 consultas. Los tres modelos monetizan lo mismo: la superación de la barrera técnica de formato. Considerando que existen 346 comunas en Chile y que múltiples actores del mercado inmobiliario, financiero, académico y municipal requieren estos datos, el costo agregado de la inaccesibilidad se traduce en transferencias significativas desde los usuarios hacia intermediarios cuyo principal valor agregado es superar esas barreras.

Para dimensionar este costo se propone una estimación de orden de magnitud basada en tres supuestos explícitos: (i) el número de actores que adquieren datos catastrales anualmente, (ii) el precio promedio por dataset comunal, y (iii) el número promedio de comunas adquiridas por actor. La Tabla 2.1 presenta un análisis de sensibilidad que varía estos parámetros dentro de rangos plausibles.

Tabla 2.1: Análisis de sensibilidad del costo anual de la inaccesibilidad catastral

Escenario	**Actores**	**Precio/comuna**	**Comunas/actor**	**Costo anual (USD)**
Conservador	200	USD 2.000	1	400.000
Base	500	USD 4.000	1	2.000.000
Moderado	500	USD 4.000	3	6.000.000
Alto	800	USD 5.000	3	12.000.000

El escenario base (500 actores, USD 4.000 por comuna, 1 comuna por actor) arroja un costo agregado de aproximadamente 2 millones de USD anuales (véase sección 2.2.4). El supuesto de 500 actores se basa en el número de inmobiliarias, consultoras ambientales, municipios con presupuesto para datos y centros de investigación que operan en el mercado inmobiliario chileno. El precio promedio de USD 4.000 corresponde al punto medio del rango observado. Cabe señalar que esta estimación no incluye los costos indirectos de decisiones subóptimas tomadas por actores que operan sin información territorial adecuada (costos que la teoría de la asimetría de información

predice como significativos (Akerlof, 1970; Stiglitz, 2000)) ni el valor de los nuevos modelos de negocio que la disponibilidad del dato habilitaría. A nivel internacional, el European Data Portal estimó en 2020 que el mercado europeo de datos abiertos alcanzó los 184 mil millones de euros, lo que sugiere que el valor potencial de liberar datos catastrales excede el ahorro directo en adquisición (European Data Portal, 2020).

2.3 Por qué este problema no se ha resuelto antes

La persistencia del problema se explica por la confluencia de tres desafíos técnicos que deben resolverse de forma simultánea: el parseo de formatos heredados de ancho fijo, la extracción masiva distribuida frente a un servidor con detección de bots, y la vectorización de cartografía raster. Cada uno de estos problemas tiene soluciones conocidas en la literatura; sin embargo, su combinación sobre un dataset de 9,5 millones de registros presenta una complejidad operativa que no ha sido abordada previamente en el contexto chileno.

Adicionalmente, una vez construido el dataset, su distribución genera un segundo problema: los formatos analíticos (GeoJSON, GeoPackage, Parquet) son útiles para técnicos pero inaccesibles para la mayoría de los actores que podrían beneficiarse de los datos. La sola existencia del dataset no resuelve el problema de acceso si no existe una interfaz que lo haga consultable sin conocimientos de programación.

Capítulo 3

Hipótesis y Objetivos

3.1 Hipótesis

El análisis presentado en los capítulos anteriores conduce a una hipótesis central que orienta este trabajo:

> *Las barreras que impiden el acceso masivo a los datos catastrales del SII son de naturaleza exclusivamente técnica y de formato, sin componente legal ni político, y son superables con métodos estándar de ingeniería de datos disponibles públicamente.*

Esta formulación es falseable bajo al menos tres condiciones: (i) que el SII impusiera restricciones legales explícitas al acceso automatizado, lo que convertiría la barrera en legal y no técnica; (ii) que las barreras técnicas requirieran acceso privilegiado no disponible públicamente (por ejemplo, credenciales internas o servidores no expuestos a internet), lo que las haría insuperables con métodos estándar; o (iii) que los métodos de ingeniería de datos disponibles fueran insuficientes para producir un dataset con cobertura y precisión aceptables para análisis territorial.

La hipótesis tiene una implicancia directa: si se confirma, la brecha de información que hoy favorece a los actores con mayor capacidad técnica no es un resultado inevitable del mercado, sino una consecuencia de la ausencia de herramientas apropiadas (World Bank, 2021; Stiglitz, 2000). La teoría de la asimetría de información (Akerlof, 1970) predice que la reducción de esta brecha debería mejorar la eficiencia del mercado inmobiliario chileno.

3.2 Objetivos

Objetivo general: Proponer y estimar el impacto económico que la apertura efectiva de datos públicos técnicamente opacos puede generar: la reducción de la asimetría de información en mercados que dependen de inteligencia territorial, la nivelación del acceso analítico entre actores de distinto tamaño y la habilitación de nuevos emprendimientos y modelos de análisis que hoy no existen por ausencia del dato. Para ello, esta AFE toma el caso del catastro predial del SII (formalmente público pero prácticamente inaccesible) y construye la infraestructura técnica y la plataforma de distribución necesarias para que cualquier actor, independientemente de su capacidad técnica o presupuesto, pueda consultar, analizar y descargar la información territorial de los 9,5 millones de predios de Chile.

Los **objetivos específicos** son:

1. Desarrollar un pipeline de ingeniería de datos que supere las tres barreras de acceso identificadas: parseo del archivo de ancho fijo, extracción masiva distribuida vía API con múltiples identidades de red, y vectorización de la cartografía raster del SII.

2. Construir un dataset predial nacional con cobertura de 343 comunas de Chile, integrando atributos tabulares (avalúos, superficies, destinos prediales, series históricas 2018–2025) y geometría vectorial (≈30 cm de precisión) para cada uno de los 9,5 millones de predios.

3. Diseñar e implementar la plataforma **Catastral.cl**, que materializa el objetivo de acceso: descarga gratuita de las series históricas de avalúo (16 semestres, 2018–2025) sin registro ni pago, y descarga de datos enriquecidos con geometría vectorial en formatos analíticos (CSV y Parquet) para las 343 comunas disponibles mediante un modelo de pago social sin barrera monetaria.

Capítulo 4

Marco Teórico y Revisión de la Literatura

El concepto de *open data* o datos abiertos designa conjuntos de datos que cualquier persona puede usar, modificar y redistribuir libremente, sin restricciones de propiedad intelectual, patentes o mecanismos de control adicionales (Open Knowledge Foundation, 2023). En el ámbito gubernamental, los datos abiertos constituyen un mecanismo de transparencia y rendición de cuentas: al poner a disposición del público información sobre el territorio, la economía y los servicios del Estado, se habilita la fiscalización ciudadana y se reduce la asimetría de información entre el gobierno y los administrados (Janssen et al., 2012; Zuiderwijk and Janssen, 2014).

Sin embargo, la experiencia internacional muestra una brecha persistente entre el dato *técnicamente público* y el dato *prácticamente accesible*. Un dato es técnicamente público cuando no existe restricción legal para acceder a él; es prácticamente accesible cuando puede ser efectivamente utilizado sin barreras técnicas significativas (Attard et al., 2015). Esta distinción es central para entender el problema que este trabajo aborda.

4.1 Asimetría de información en mercados inmobiliarios

El problema de la inaccesibilidad de datos catastrales se enmarca en la teoría económica de la asimetría de información. Akerlof (1970) demostró que cuando los participantes de un mercado disponen de información desigual sobre la calidad

de los bienes transados, se producen ineficiencias que pueden derivar en la selección adversa y el colapso del mercado. Stiglitz (2000) extendió este análisis al mostrar que las asimetrías informacionales son ubicuas en mercados de activos y que su reducción genera ganancias de eficiencia medibles.

El mercado inmobiliario es particularmente susceptible a estas asimetrías. DiPasquale and Wheaton (1996) establecieron que la heterogeneidad de los activos inmobiliarios, combinada con la naturaleza localizada de la información catastral, produce *search costs* elevados: los compradores y vendedores incurren en costos significativos para obtener información sobre precios, características y ubicación de las propiedades disponibles. Broxterman and Zhou (2023) documentaron que estas fricciones informacionales persisten incluso en mercados con alta penetración tecnológica, y que la disponibilidad de datos de valor de suelo a escala granular permite a los actores ajustar precios, identificar tendencias de densificación y reducir el riesgo de inversión mediante análisis comparativos objetivos.

En este contexto, los registros catastrales públicos funcionan como un mecanismo institucional de reducción de asimetrías: al proveer información verificable sobre la propiedad, el valor fiscal y las características de cada predio, reducen los costos de transacción y mejoran la eficiencia del mercado (Deininger and Feder, 2009; Williamson and Kerekes, 2011). Pero este mecanismo solo opera cuando los datos son efectivamente accesibles. Cuando el formato de publicación impone barreras técnicas que limitan el acceso a un grupo reducido de actores, la asimetría no se elimina sino que se desplaza: ya no es entre gobierno y ciudadano, sino entre quienes pueden procesar los datos y quienes no.

4.2 Datos abiertos gubernamentales: de la publicidad formal a la accesibilidad efectiva

En Chile, la Ley de Transparencia (Biblioteca del Congreso Nacional de Chile, 2008) garantiza el derecho de acceso a la información de los organismos del Estado. El SII, en cumplimiento de esta normativa, publica los datos catastrales en su sitio web. Formalmente, los datos son públicos. Pero el formato en que se publican (un archivo de texto plano de ancho fijo sin cabeceras, de 3 GB de tamaño y con codificación `latin-1`) hace que su uso efectivo requiera conocimientos de programación que solo posee una fracción mínima de la población.

La literatura sobre datos abiertos gubernamentales ha documentado extensamente

esta brecha entre publicación y accesibilidad. Zuiderwijk and Janssen (2014) propusieron un marco de evaluación de políticas de datos abiertos que distingue entre la disponibilidad formal del dato y su usabilidad efectiva, identificando que muchos gobiernos satisfacen el requisito legal de publicación sin garantizar que los datos sean procesables por máquinas o comprensibles por usuarios no técnicos. Attard et al. (2015), en una revisión sistemática de 55 iniciativas de datos abiertos gubernamentales, encontraron que las barreras técnicas de formato son tan restrictivas como las barreras legales para el acceso efectivo.

Según CEPAL (2022), la brecha de habilidades digitales avanzadas en América Latina agrava esta situación: solo una fracción pequeña de la población posee las competencias necesarias para procesar formatos de baja usabilidad. En la práctica, la publicidad formal no equivale a accesibilidad real (Janssen et al., 2012).

4.3 El valor económico del dato catastral abierto

Cuando los datos catastrales son genuinamente accesibles, el impacto económico es significativo. World Bank (2021) estima que la apertura efectiva de datos gubernamentales puede generar valor equivalente al 1–3% del PIB en economías desarrolladas, principalmente a través de la reducción de costos de transacción en mercados de activos, la mejora en la asignación de recursos y la habilitación de nuevos modelos de negocio. A escala global, McKinsey Global Institute (2013) estimó el potencial de los datos abiertos en 3 a 5 billones de USD anuales. Deininger and Feder (2009) demostraron que la calidad de los registros catastrales tiene efectos directos sobre la inversión, el acceso al crédito y la recaudación fiscal, especialmente en economías en desarrollo donde la informalidad de la tenencia de tierra es alta.

La experiencia internacional ofrece ejemplos concretos. En el Reino Unido, la apertura del registro catastral (*Land Registry*) permitió el surgimiento de plataformas de inteligencia inmobiliaria que generan cientos de millones de libras anuales en valor económico (World Bank, 2021). En la Unión Europea, el European Data Portal estimó que el mercado de datos abiertos alcanzó los 184 mil millones de euros en 2020 (European Data Portal, 2020). En América Latina, Colombia liberó su catastro multipropósito en 2020, lo que permitió a municipios rurales acceder por primera vez a información territorial estructurada para la planificación de servicios públicos (CEPAL, 2022).

4.4 Fundamentos legales del scraping sobre datos públicos

La construcción del dataset que esta AFE produce requiere extracción automatizada de datos desde los servidores del SII. Este método, conocido como *web scraping* o *API harvesting*, descansa sobre un marco legal que es necesario precisar.

En Chile, no existe legislación específica que prohíba el acceso automatizado a sitios web de organismos públicos, siempre que dicho acceso no genere daño al servicio ni implique apropiación de datos protegidos por derechos de autor. Los datos catastrales del SII son de dominio público por mandato de la Ley 20.285 (Biblioteca del Congreso Nacional de Chile, 2008). A nivel internacional, el caso *hiQ Labs v. LinkedIn* (2021) en la Corte Suprema de Estados Unidos sentó un precedente relevante al confirmar que el scraping de datos públicamente accesibles no constituye necesariamente una violación del *Computer Fraud and Abuse Act*: el tribunal razonó que si los datos son accesibles sin autenticación para cualquier usuario, la extracción automatizada de los mismos no puede equipararse a un acceso no autorizado. En la Unión Europea, la Directiva 2019/1024 sobre datos abiertos (*Open Data Directive*) va más lejos: establece explícitamente que los datos gubernamentales *deben* ser accesibles en formatos legibles por máquina, lo que convierte la extracción automatizada no solo en legítima sino en el mecanismo previsto por el legislador para el uso efectivo de los datos públicos (Attard et al., 2015).

El principio rector que guía la práctica responsable de scraping sobre servicios públicos es la no interferencia: el proceso de extracción debe diseñarse para consumir la menor cantidad posible de recursos del servidor, respetando las limitaciones de tasa de consultas (*rate limits*) y distribuyendo la carga a lo largo del tiempo. Esta AFE sigue este principio mediante la distribución de consultas entre 30 túneles WireGuard con rotación dinámica, lo que mantiene la tasa de consultas por IP dentro de umbrales que no afectan la disponibilidad del servicio (Cetl et al., 2023).

Capítulo 5

Metodología

El presente capítulo describe la metodología empleada para abordar el problema descrito en el Capítulo 1. Se organiza en seis secciones: los principios de ingeniería de datos masiva aplicada a fuentes públicas, el estándar WMS y las limitaciones de la cartografía raster para el análisis espacial, los algoritmos de vectorización de imágenes catastrales, los métodos de integración espacial entre fuentes heterogéneas, los criterios de validación del dataset resultante, y la metodología de refinamiento iterativo para maximizar la cobertura geométrica en predios urbanos.

5.1 Ingeniería de datos masiva sobre fuentes públicas

La extracción sistemática de datos desde fuentes públicas en internet es una disciplina consolidada en la ciencia de datos, conocida genéricamente como *web scraping* o, cuando se trata de APIs, como *API harvesting*. Los fundamentos legales que habilitan esta práctica sobre fuentes gubernamentales se discuten en el Capítulo 4; esta sección describe los aspectos técnicos de su implementación.

5.1.1 El problema del bloqueo por volumen

Los servidores web modernos implementan mecanismos de detección de comportamiento automatizado, principalmente mediante el análisis de la frecuencia y el origen de las solicitudes HTTP. Cuando un único cliente realiza miles de consultas en un período corto, el servidor puede interpretar este comportamiento como un

ataque de denegación de servicio y bloquear la dirección IP del solicitante (Donenfeld, 2017).

Para la extracción a escala nacional de datos prediales, con 9,5 millones de consultas necesarias, este mecanismo constituye una barrera técnica determinante. La solución implementada en este trabajo es la distribución de la carga entre múltiples identidades de red mediante el uso de túneles VPN con rotación dinámica, lo que permite mantener una tasa de consultas por IP dentro de umbrales aceptables mientras se maximiza el throughput agregado del sistema.

5.1.2 Infraestructura de extracción distribuida

El concepto de *network namespace* en sistemas Linux permite crear entornos de red completamente aislados dentro de un mismo servidor físico (Linux man-pages project, 2024). Cada namespace tiene su propia interfaz de red, tabla de rutas y dirección IP, lo que permite ejecutar múltiples procesos de scraping en paralelo, cada uno saliendo por una IP distinta hacia el servidor objetivo.

La combinación de namespaces con túneles WireGuard (Donenfeld, 2017) (protocolo VPN moderno de alta performance) permite construir una infraestructura de extracción distribuida donde cada túnel actúa como un cliente independiente desde la perspectiva del servidor del SII. La detección de bloqueo y la rotación automática de relays permite recuperar un túnel bloqueado en segundos sin interrumpir los demás procesos en ejecución.

5.2 Servicios WMS y cartografía catastral

5.2.1 El estándar OGC Web Map Service

El *Web Map Service* (WMS) es un estándar del Open Geospatial Consortium (OGC) que define una interfaz HTTP para solicitar imágenes de mapas georreferenciadas desde un servidor cartográfico (Open Geospatial Consortium, 2006). Una solicitud WMS especifica una capa de datos, un sistema de referencia de coordenadas (CRS), un bounding box geográfico y unas dimensiones en píxeles; el servidor responde con una imagen renderizada que representa los datos de la capa en esa región y resolución.

El SII implementa un servidor WMS basado en GeoServer (GeoServer Project Steer-

ing Committee, 2024) sobre la URL:

```
https://www4.sii.cl/mapasui/services/ui/wmsProxyService/call
```

Las capas catastrales siguen la convención de nombres `sii:BR_CART_{NOMBRE_COMUNA}_WMS` con el estilo `PREDIOS_WMS_V0`. El servicio entrega imágenes PNG de 256Œ256 píxeles donde cada tipo de elemento del mapa tiene un color específico y consistente.

5.2.2 Las limitaciones del raster para el análisis espacial

La diferencia fundamental entre un dato *raster* y un dato *vectorial* es la naturaleza de la representación geométrica. Un raster es una grilla regular de píxeles, donde cada píxel almacena un valor (color, temperatura, elevación). Un dato vectorial es una colección de geometrías (puntos, líneas, polígonos) definidas por coordenadas exactas, con atributos asociados a cada geometría (GDAL/OGR contributors, 2024).

Para el análisis espacial catastral, la representación vectorial es indispensable. Solo con polígonos vectoriales es posible calcular áreas exactas, identificar colindancias, realizar intersecciones geométricas, construir grillas de valor de suelo y ejecutar consultas espaciales del tipo "todos los predios con destino comercial dentro de un radio de 500 metros de este punto". El raster, al ser una imagen, no permite ninguna de estas operaciones de forma directa (Gillies et al., 2024b).

5.2.3 El comportamiento escala-dependiente del WMS del SII

Durante el desarrollo de este trabajo se descubrió empíricamente un comportamiento no documentado del servidor WMS del SII: la representación de los bordes internos entre predios depende del tamaño del bounding box solicitado. Cuando el BBOX es pequeño (equivalente a tiles de zoom 19, aproximadamente 76 metros por lado), en ciertas zonas el servidor no renderiza las líneas de borde que separan predios contiguos, entregando bloques de color uniforme sin subdivisiones internas. Cuando el mismo territorio se solicita con un BBOX mayor (equivalente a zoom 17 o menor), el servidor sí renderiza todos los bordes.

Este comportamiento es consistente con la configuración de umbrales de escala (*scale denominators*) en GeoServer (GeoServer Project Steering Committee, 2024), que permiten al administrador definir a partir de qué nivel de zoom se renderizan capas

o estilos específicos. En la práctica, este comportamiento genera polígonos anormalmente grandes en la vectorización, que deben ser detectados y corregidos mediante un proceso de re-consulta multi-resolución.

5.3 Vectorización de datos raster catastrales

5.3.1 Polygonización raster con GDAL

La *polygonización raster* es el proceso de convertir una imagen raster en un conjunto de polígonos vectoriales, donde cada polígono representa una región contigua de píxeles con el mismo valor. GDAL (GDAL/OGR contributors, 2024) implementa este algoritmo en la herramienta `gdal_polygonize`, que es también el motor interno del comando equivalente en QGIS.

El algoritmo implementa *region labeling* con 4-conectividad: recorre la imagen píxel a píxel y agrupa en un mismo polígono todos los píxeles conectados horizontal o verticalmente que tienen el mismo valor digital (DN, *Digital Number*). Los píxeles con valores distintos actúan como separadores naturales entre regiones, generando fronteras entre polígonos.

En el mapa catastral del SII, el canal rojo (banda 1) tiene valor exactamente igual a 182 en el interior de todos los predios. Los bordes entre predios son líneas de 1 a 2 píxeles de ancho con valores inferiores a 182. Al polygonizar la banda 1, cada predio se convierte en un polígono independiente, separado de sus vecinos por las líneas de borde.

Formalmente, dado un raster R de dimensiones $W \times H$ píxeles, donde $R_{i,j}$ denota el valor del píxel en la columna i y fila j, la polygonización produce un conjunto de polígonos $\{P_k\}$ tal que cada polígono P_k cubre la región conexa maximal de píxeles con el mismo valor:

$$P_k = \{(i,j) \mid R_{i,j} = v_k \text{ y } (i,j) \text{ es 4-conexo con todos los demás píxeles de } P_k\} \tag{5.1}$$

5.3.2 El índice de compactness para clasificación de geometrías

Los polígonos extraídos de la polygonización contienen agujeros internos (*holes*) causados por el texto sobreimpreso en el mapa: números de rol predial, nombres de calles y otros elementos gráficos tienen colores distintos al relleno del predio y generan discontinuidades en la región 4-conectada.

Es necesario distinguir entre dos tipos de holes:

- **Holes textuales:** causados por números y letras sobreimpresos. Son formas relativamente compactas y de área pequeña.
- **Holes reales:** corresponden a pasajes peatonales, patios interiores o espacios públicos encerrados dentro del contorno de un predio. Son formas alargadas o de área considerable.

Para distinguirlos se utiliza el índice de compactness, también conocido como cociente isoperimétrico (Gillies et al., 2024b):

$$C = \frac{4\pi \cdot A}{P^2} \tag{5.2}$$

donde A es el área del hole y P es su perímetro. Este índice vale 1 para un círculo perfecto, aproximadamente 0,785 para un cuadrado, y valores cercanos a 0 para formas muy alargadas. Los holes textuales, al ser caracteres tipográficos, tienen compactness típicamente superior a 0,25. Los pasajes interiores, al ser formas estrechas y alargadas, tienen compactness típicamente inferior a 0,25.

Se aplica la siguiente regla de clasificación:

$$\text{tipo}(h) = \begin{cases} \text{textual (rellenar)} & \text{si } C(h) > 0{,}25 \text{ y } A(h) < 100 \text{ m}^2 \\ \text{real (conservar)} & \text{en caso contrario} \end{cases} \tag{5.3}$$

5.3.3 Georeferenciación y sistema de coordenadas

Los tiles WMS del SII se entregan en el sistema de proyección Web Mercator (EPSG:3857). La georeferenciación del GeoTIFF ensamblado se calcula a partir de los bounds exactos de la grilla de tiles descargados, utilizando la fórmula estándar del sistema de tiles XYZ (Open Geospatial Consortium, 2006):

$$\text{pixel\_size} = \frac{2\pi R_{\oplus}}{2^z \cdot 256} \tag{5.4}$$

donde $R_{\oplus} = 6.378.137$ m es el semieje mayor del elipsoide WGS84 en proyección Mercator esférica y z es el nivel de zoom. A zoom 19:

$$\text{pixel\_size}_{z=19} = \frac{2\pi \times 6.378.137}{2^{19} \times 256} \approx 0{,}2986 \text{ m/píxel} \tag{5.5}$$

Esta resolución de aproximadamente 30 cm por píxel garantiza que los bordes entre predios (de 1 a 2 píxeles de ancho) sean suficientemente visibles para la polygonización, y que la precisión geométrica resultante sea adecuada para análisis catastrales a escala de predio.

5.4 Integración espacial de fuentes heterogéneas

El producto final de este trabajo requiere combinar dos fuentes de naturaleza distinta: los datos tabulares extraídos de la API del SII (atributos de cada predio: avalúo, destino, superficie, coordenadas) y los polígonos vectoriales extraídos del WMS (geometría de cada lote). La integración de estas fuentes es un problema de *spatial join* con particularidades importantes (Gillies et al., 2024b).

5.4.1 El problema del rol predial y el lote físico

En el catastro chileno, la relación entre predios registrales y lotes físicos no es uno a uno. Existen dos configuraciones posibles:

Lote simple: un terreno físico corresponde a un único predio registral. El predio tiene coordenadas únicas y su polígono en el mapa corresponde directamente a su rol.

Edificio multi-unidad: un terreno físico alberga múltiples predios registrales (departamentos, oficinas, bodegas, estacionamientos), todos con la misma coordenada geográfica. El polígono en el mapa corresponde al terreno completo, no a cada unidad.

Esta distinción es fundamental para el proceso de join. Si se intenta emparejar cada predio registral con el polígono que contiene sus coordenadas, los edificios multi-unidad generan un problema de asignación: decenas o cientos de predios comparten

las mismas coordenadas y por lo tanto apuntarían al mismo polígono (GDAL/OGR contributors, 2024).

5.4.2 El concepto de rol base

La solución adoptada es el concepto de *rol base*: el identificador del predio que representa al terreno o edificio completo, independientemente de cuántas unidades registrales contenga. En el sistema catastral chileno, los predios de bienes comunes de edificios multi-unidad tienen números de predio en el rango 90.000-99.999, convención que permite identificarlos y utilizarlos como clave de enlace.

El proceso de asignación del rol base opera en dos pasos:

1. **Clasificación:** se agrupan los predios del CSV por coordenada geográfica (latitud y longitud redondeados a 6 decimales, equivalente a una precisión de aproximadamente 0,1 metros). Las coordenadas con un único predio corresponden a lotes simples; las coordenadas con múltiples predios corresponden a edificios multi-unidad.

2. **Identificación del predio 90xxx:** para cada grupo multi-unidad, se consulta la API `getPredioNacional` con números de predio candidatos en el rango 90.000-99.999 hasta identificar el predio de bienes comunes. Este predio se convierte en el `rol_base` del grupo.

El `rol_base` actúa entonces como la clave foránea que enlaza la tabla de atributos prediales con la capa de polígonos catastrales, permitiendo construir el dataset final con una fila por predio registral y la geometría correspondiente al lote físico al que pertenece (Gillies et al., 2024b; GDAL/OGR contributors, 2024).

5.4.3 Validación y control de calidad a escala nacional

La validación de un dataset geoespacial de cobertura nacional requiere métricas que operen a dos niveles: el nivel de predio (integridad de atributos y consistencia geométrica de cada fila) y el nivel de comuna (cobertura del join, proporción de predios con geometría asignada, ausencia de geometrías fuera de los bounds de Chile) (GDAL/OGR contributors, 2024).

Las métricas de cobertura del join son especialmente críticas porque revelan la fracción del universo predial que efectivamente pudo ser georreferenciada. Un predio sin

geometría asignada es un predio cuya ubicación en el espacio no pudo determinarse, lo que limita su utilidad para análisis territoriales. Los factores que pueden causar ausencia de geometría incluyen predios agrícolas sin coordenadas en la API, errores de georeferenciación en los tiles WMS, y predios cuyas coordenadas no caen dentro de ningún polígono vectorizado.

El proceso de QA implementado verifica, para cada una de las 343 comunas del dataset, la consistencia entre el número de predios en el CSV de Fase 0 y el número de filas en el producto final, la proporción de predios con geometría asignada, la validez topológica de los polígonos (ausencia de auto-intersecciones), y que todas las coordenadas se encuentren dentro de los bounds geográficos de Chile (longitud entre $-76°$ y $-65{,}5°$; latitud entre $-56°$ y $-17°$) (GDAL/OGR contributors, 2024).

5.5 Refinamiento iterativo de cobertura geométrica

Tras la consolidación del dataset (Fases 0–7), un porcentaje significativo de predios urbanos queda sin geometría asignada. El análisis del producto de Fase 6 revela que aproximadamente el 15,6% de los predios con ubicación urbana carecen de polígono, cifra que resulta inaceptable para un dataset que aspire a constituir una capa base de análisis territorial. Esta sección describe la metodología de refinamiento diseñada para recuperar la mayor cantidad posible de estos predios.

5.5.1 Diagnóstico del problema residual

La ausencia de geometría en predios urbanos tiene dos causas principales. La primera es el comportamiento escala-dependiente del WMS descrito en la sección 2.2.3: en ciertas zonas, los tiles de zoom 19 no renderan los bordes internos entre predios, produciendo bloques de color uniforme que se vectorizan como un único polígono fusionado. La segunda es una limitación del *spatial join* por punto-en-polígono: cuando la coordenada del predio cae en el borde de un polígono o a pocos centímetros de su límite, el join estricto falla.

El hallazgo clave que motiva el diseño del refinamiento es empírico: el 99% de los predios urbanos sin geometría tienen un polígono vectorizado a menos de 3 metros de su coordenada en el producto de la fase anterior. El problema no es la ausencia del polígono en el WMS, sino la precisión del emparejamiento.

5.5.2 Re-descarga focalizada con supercells

Para las zonas afectadas por el problema de escala del WMS, se implementa una re-descarga focalizada. En lugar de descargar tiles individuales de 256×256 píxeles (como en la Fase 1), se utilizan *supercells*: cada solicitud WMS cubre un área de 4×4 tiles regulares (1024×1024 píxeles), lo que equivale a un bounding box 16 veces mayor. Al solicitar un BBOX más amplio, el servidor renderiza correctamente los bordes internos que no aparecen en tiles individuales, resolviendo el problema de escala sin necesidad de recurrir a zoom 17.

Las supercells se calculan únicamente alrededor de los predios faltantes (con un buffer de ±2 tiles y alineamiento a la grilla), lo que limita la descarga a las zonas estrictamente necesarias.

5.5.3 Spatial join multi-pasada con tolerancia geométrica

El emparejamiento entre predios y polígonos se realiza en dos pasadas sucesivas:

1. **Pasada estricta (*point-in-polygon*):** la coordenada del predio debe caer estrictamente dentro de un polígono. Esta pasada resuelve los casos triviales donde la coordenada está claramente en el interior del lote.

2. **Pasada con tolerancia (≤3 m):** para los predios no emparejados en la primera pasada, se aplica un buffer de 3 metros alrededor de cada polígono y se repite el join. Este umbral cubre el *drift* típico entre la coordenada reportada por la API y el centroide real del polígono WMS, que empíricamente es de aproximadamente 0,2 metros pero puede alcanzar hasta 2–3 metros en zonas con errores de georeferenciación.

5.5.4 Consulta directa al WMS mediante `getFeatureInfo`

Para los predios que no logran emparejarse ni con tolerancia geométrica, se recurre al servicio `getFeatureInfo` del WMS del SII. Este servicio simula un clic sobre el mapa en las coordenadas del predio y devuelve el rol del predio que el servidor asocia a esa ubicación. A diferencia de `GetMap` (que devuelve una imagen), `getFeatureInfo` devuelve datos estructurados que permiten establecer la correspondencia directa entre coordenada y rol predial.

Un aspecto técnico relevante es que el endpoint del SII implementa `getFeatureInfo` exclusivamente mediante solicitudes HTTP POST con payload JSON, a diferencia de la especificación estándar del OGC que define la operación como GET (Open Geospatial Consortium, 2006). Este detalle de implementación, no documentado, fue descubierto empíricamente durante el desarrollo del pipeline.

5.5.5 Estrategias de fallback para predios sin coordenada válida

Tras el *spatial join* multi-pasada y la consulta `getFeatureInfo`, un residuo de predios sigue sin geometría. Para estos casos se aplican tres estrategias de fallback basadas en la estructura del catastro chileno:

Herencia por coordenada compartida: los predios en edificios y condominios comparten exactamente la misma latitud y longitud. Si al menos uno de los predios de un grupo con coordenada compartida tiene geometría asignada, todos los demás la heredan.

Polígono más cercano (*nearest-polygon*, 50 m): para predios con coordenada válida pero sin match, se asigna el polígono más cercano dentro de un radio de 50 metros. Dado que el 99% de los predios faltantes tienen un polígono a menos de 3 metros, esta estrategia opera con alta precisión.

Vecindad por manzana: para predios sin coordenada válida, se explota la convención del rol predial chileno (formato `MMMM-PPPP`, donde `MMMM` identifica la manzana y `PPPP` el predio). Dentro de una manzana, los predios están numerados en orden físico, por lo que se hereda la geometría del predio numéricamente más cercano que ya tenga geometría (con un límite de diferencia $\leq$10 en el número de predio para evitar asignaciones erróneas).

Capítulo 6

Descripción del Enfoque de Solución

Este capítulo documenta el pipeline de extracción catastral: su arquitectura, las decisiones de diseño adoptadas en cada fase, el código clave que implementa cada componente y los resultados intermedios observados. El pipeline transforma tres fuentes públicas del SII (archivo TXT semestral, API REST y servicio WMS) en un dataset predial nacional georreferenciado y analíticamente útil (GDAL/OGR contributors, 2024; Jordahl et al., 2024).

6.1 Arquitectura general del pipeline

El pipeline está organizado en siete fases secuenciales (Fase 0 a Fase 6), más una fase de validación (Fase 7). Cada fase produce artefactos intermedios que se almacenan en un objeto de almacenamiento en la nube (S3-compatible), lo que permite la reanudación automática ante interrupciones y el procesamiento paralelo por comuna. La Tabla 6.1 presenta el resumen de fases.

La infraestructura de cómputo consiste en una VPS Linux de 16 núcleos y 128 GB de RAM, un bucket S3-compatible en Hetzner (Núremberg), y 30 túneles WireGuard Mullvad montados como *network namespaces* de Linux. El stack de software utilizado es Python 3.11 con GDAL 3.8, GeoPandas 0.14, Rasterio 1.3 y boto3 para interacción con S3 (GDAL/OGR contributors, 2024; Jordahl et al., 2024; Gillies et al., 2024a). La Figura 6.1 esquematiza el flujo de datos entre fases.

Tabla 6.1: Fases del pipeline de extracción catastral

Fase	Nombre	Output en S3
0	Extracción predial (API SII)	CSV por comuna
1	Descarga cartográfica WMS	GeoTIFF por comuna (z19)
2	Vectorización raster	GeoPackage de polígonos
3	Dos pasadas (z19 + z17)	GeoPackage fusionado
4	Join espacial CSV–GeoJSON	GeoJSON con atributos
5	Completar roles faltantes	GeoJSON enriquecido
6	Consolidación final	Parquet/GeoJSON por comuna
7	QA y validación nacional	Reporte CSV de alertas

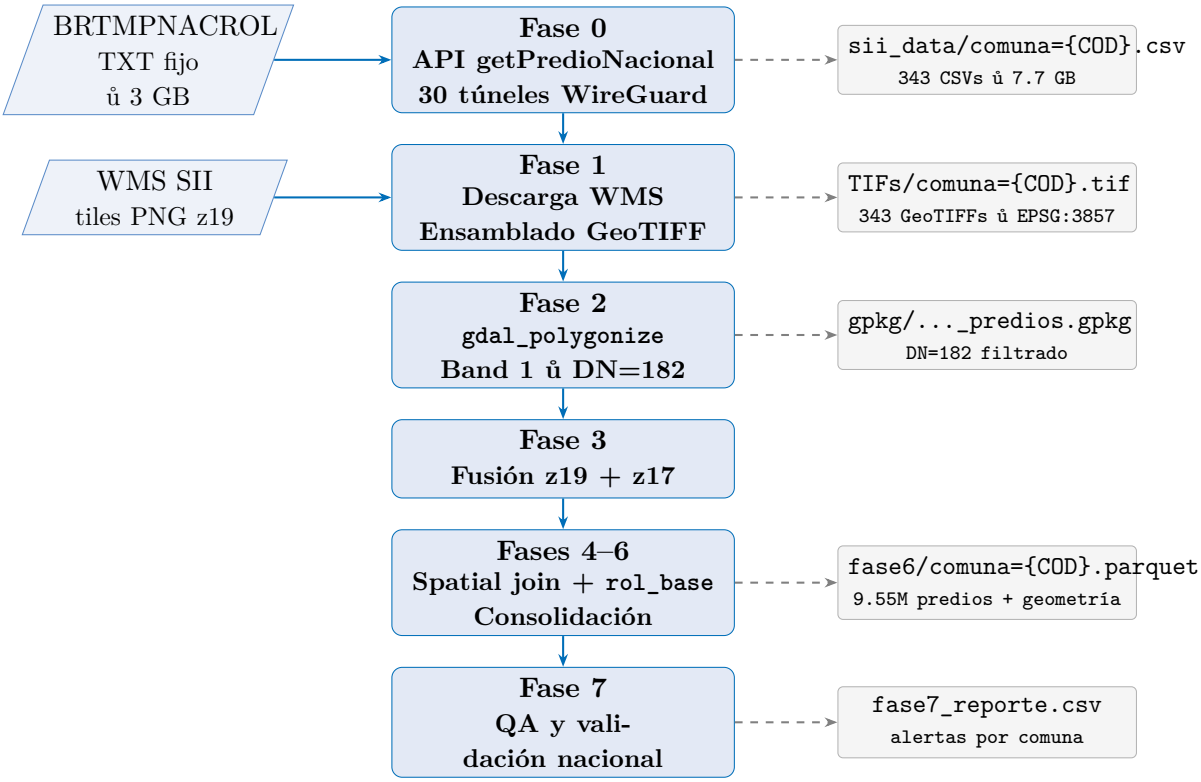

Figura 6.1: Arquitectura del pipeline de extracción catastral: flujo de datos entre las 8 fases. Flechas sólidas indican dependencias de procesamiento; flechas punteadas indican artefactos persistidos en S3. El mecanismo de *resume* permite reanudar el batch sin reprocesar fases ya completadas.

6.2 Fase 0: Extracción masiva de datos prediales

6.2.1 Parseo del archivo TXT de ancho fijo

El SII publica semestralmente el archivo `BRTMPNACROL_NAC_2025_2.txt`, denominado oficialmente *Rol Semestral de Contribuciones de Bienes Raíces* (Servicio de Impuestos Internos, 2025b). Es un archivo de texto de ancho fijo (encoding *latin-1*)

de aproximadamente 3 GB con una línea por predio, donde los campos no están delimitados por separadores sino que ocupan posiciones fijas dentro de un registro de 117 caracteres. Esta estructura es característica de los sistemas de procesamiento por lotes de entornos mainframe y constituye la primera barrera técnica: el archivo es ilegible sin conocer de antemano el mapa exacto de columnas, que solo consta en los manuales oficiales del SII (Servicio de Impuestos Internos, 2025b,a). El pipeline parsea este archivo por corte de posiciones, extrayendo para cada predio su identificación (comuna, manzana, predio), sus avalúos, su ubicación y su código de destino predial. El detalle completo de los 15 campos y el fragmento de código del parser se documentan en el Anexo A.1.

La clasificación de destino predial es especialmente relevante para el análisis posterior, ya que determina el uso del suelo declarado. El SII define 23 códigos de destino, resumidos en la Tabla 6.2.

Tabla 6.2: Tabla de destinos prediales según clasificación SII (Servicio de Impuestos Internos, 2025b)

Cód.	Destino	Cód.	Destino	Cód.	Destino
A	Agrícola	I	Industria	S	Salud
B	Agroindustrial	L	Bodega/Almacenaje	T	Transp./Telecom.
C	Comercio	M	Minería	V	Otros no consid.
D	Deporte/Recreación	O	Oficina	W	Sitio Eriazo
E	Educación/Cultura	P	Adm. Públ./Defensa	Y	Gallineros/chanch.
F	Forestal	Q	Culto	Z	Estacionamiento
G	Hotel/Motel				
H	Habitacional				

El resultado del parseo es una lista de roles (tuplas `manzana-predio`) por comuna, que constituye la cola de trabajo a distribuir en la extracción distribuida (Broxterman and Zhou, 2023).

6.2.2 La API `getPredioNacional`

La API REST del SII expone el endpoint `getPredioNacional` (POST), que devuelve la información completa de un predio dado su código de comuna, manzana y número de predio. La decisión metodológica clave aquí es que la API está diseñada para consulta unitaria (un predio por solicitud) y no ofrece endpoint de consulta masiva, lo que constituye la segunda barrera técnica: obtener el país completo exige emitir del orden de 9,5 millones de solicitudes individuales. La respuesta JSON entrega, por predio, su identificación, coordenadas geográficas (con los ejes latitud/longitud invertidos respecto de la convención habitual), avalúos fiscales, superficies y valores

comerciales de suelo por Área Homogénea. La estructura completa de la solicitud se detalla en el Anexo A.2.

6.2.3 Infraestructura de extracción distribuida

La escala de 9,5 millones de consultas unitarias, sumada al bloqueo automático de IPs que el SII aplica a quien consulta en volumen, hace inviable la extracción desde una sola dirección. La decisión de diseño fue distribuir la carga sobre 30 identidades de red independientes: 30 túneles WireGuard (proveedor Mullvad), cada uno montado como un *network namespace* aislado de Linux con su propia IP pública (Donenfeld, 2017). Cada túnel opera tres *workers* paralelos a 10 solicitudes por segundo, con un caudal sostenido cercano a 900 solicitudes por segundo en el pico. El elemento metodológicamente relevante es la *rotación automática de IP*: cuando un túnel detecta caída de su tasa de éxito, señal de que su IP fue bloqueada, el orquestador la reemplaza en caliente por un relay nuevo sin interrumpir a los demás. Con este mecanismo, el bloqueo por IP deja de ser una barrera y se convierte en un evento gestionable. La configuración de los namespaces, el diagrama de la arquitectura y un fragmento del log operativo se documentan en el Anexo A.3.

6.2.4 Resultados de la Fase 0

La ejecución del batch completo sobre las 343 comunas procesables de Chile (marzo 2026) arrojó los resultados que se presentan en la Tabla 6.3.

Tabla 6.3: Resultados de la Fase 0 (batch marzo 2026)

Métrica	**Valor**
Comunas procesadas	343
Predios extraídos	9.55 millones
Tiempo total	$\approx$ 76 horas
Rate efectivo	15–42 req/s
Rotaciones de IP	594
Fallos de comuna	0
Volumen almacenado en S3	7.7 GB (CSV)

6.3 Fase 1: Descarga cartográfica WMS

6.3.1 Clasificación de comunas: Tier A y Tier B

El análisis de los CSVs de Fase 0 reveló que 643.000 predios (6.7% del total) no tienen coordenadas geográficas en la API, principalmente predios agrícolas: 77 comunas tienen más del 30% de sus predios en esa condición. Esto plantea una decisión metodológica sobre qué tiles del WMS descargar. Usar exclusivamente el mapa de calor de coordenadas conocidas ahorraría descargas, pero dejaría sin cobertura justamente esas zonas agrícolas, cuyos polígonos sí existen en el WMS aunque el predio no tenga coordenada en la API.

La solución fue clasificar cada comuna en dos tiers según su proporción de predios agrícolas o sin coordenada, con un umbral de corte del 30%. Las comunas **Tier A** (mayoritariamente urbanas, con coordenadas densas) usan descarga selectiva guiada por el mapa de calor, con un ahorro del 50 al 80% de los tiles. Las comunas **Tier B** (alta proporción agrícola o sin coordenada) usan descarga del *bounding box* comunal completo, que garantiza cobertura a costa de descargar más. El árbol de decisión completo, con umbrales y ejemplos por tier, se detalla en el Anexo A.4.

6.3.2 Generación de GeoTIFFs georeferenciados

El servicio WMS del SII entrega tiles de 256×256 píxeles en formato PNG. Cada tile se solicita al nivel de zoom 19 de la proyección Web Mercator (EPSG:3857), lo que equivale a una resolución de aproximadamente 0.2986 m/px, derivada de la fórmula de georeferenciación:

$$p = \frac{2\pi R_{\oplus}}{2^z \cdot 256} \approx 0.2986 \text{ m/px} \quad (z = 19) \tag{6.1}$$

donde $R_{\oplus} = 6{,}378{,}137$ m es el radio ecuatorial de la Tierra en la proyección WGS84 (Open Geospatial Consortium, 2006).

Los tiles se solicitan al WMS del SII en el estilo cartográfico `PREDIOS_WMS_V0` y se ensamblan en un único GeoTIFF RGBA usando escritura *windowed* (Rasterio), lo que evita cargar la imagen completa en RAM. Santiago Centro, como ejemplo de referencia, genera un GeoTIFF de 14.336×13.568 px (13.3 MB), a partir de 2.968 tiles descargados en aproximadamente 7 minutos (Gillies et al., 2024a). La estructura de la URL de solicitud de un tile se documenta en el Anexo A.5.

6.3.3 Work stealing entre túneles

La distribución estática de tiles entre los 30 túneles implica que algunos terminan antes que otros y quedan ociosos mientras el más lento aún trabaja. El mecanismo de *work stealing* resuelve este desbalance: cuando un túnel termina, el orquestador le reasigna parte del trabajo pendiente del túnel más rezagado, omitiendo los tiles ya presentes en disco para evitar re-descargas. El impacto medido fue pasar de ≈23 a ≈48 tiles/s en la fase final de descarga de cada comuna.

6.4 Fase 2: Vectorización de polígonos prediales

6.4.1 Polygonización mediante GDAL Band 1

La vectorización de los GeoTIFFs se realiza con `gdal_polygonize`, que implementa el etiquetado de componentes conexas por 4-conectividad (GDAL/OGR contributors, 2024) sobre el canal rojo (Band 1) del GeoTIFF RGBA. El hallazgo que hace viable este paso es que, en el estilo cartográfico del SII, los píxeles interiores de cada predio tienen un valor de canal rojo constante de **DN=182**, mientras que bordes, calles y fondo tienen valores distintos y actúan como separadores naturales. Filtrar por ese valor separa los predios reales del resto del raster.

La reducción que produce el filtro es drástica. En Santiago Centro, la polygonización cruda genera 3.379.736 polígonos (763 MB); tras conservar solo los de DN=182 y descartar los menores a 10 mš (artefactos de anti-aliasing en los bordes), el resultado baja a 21.050 polígonos (29 MB), eliminando el 99.4% de los polígonos y el 96% del volumen sin perder más del 0,1% del área real. Un tercer paso de depuración por compactness (Sección siguiente) descarta los agujeros interiores. La Figura 6.2 resume el proceso; los comandos exactos de GDAL y `ogr2ogr` se documentan en el Anexo A.5.

6.4.2 Descubrimiento del comportamiento blank: umbral de escala en GeoServer

Durante el desarrollo se detectó un comportamiento no documentado del servidor GeoServer del SII: ciertos tiles descargados a zoom 19 llegaban completamente en blanco (sin geometría) aunque la zona sí tuviera predios. Este fenómeno ocurre en zonas de alta densidad predial donde el servidor sobrepasa un umbral interno de

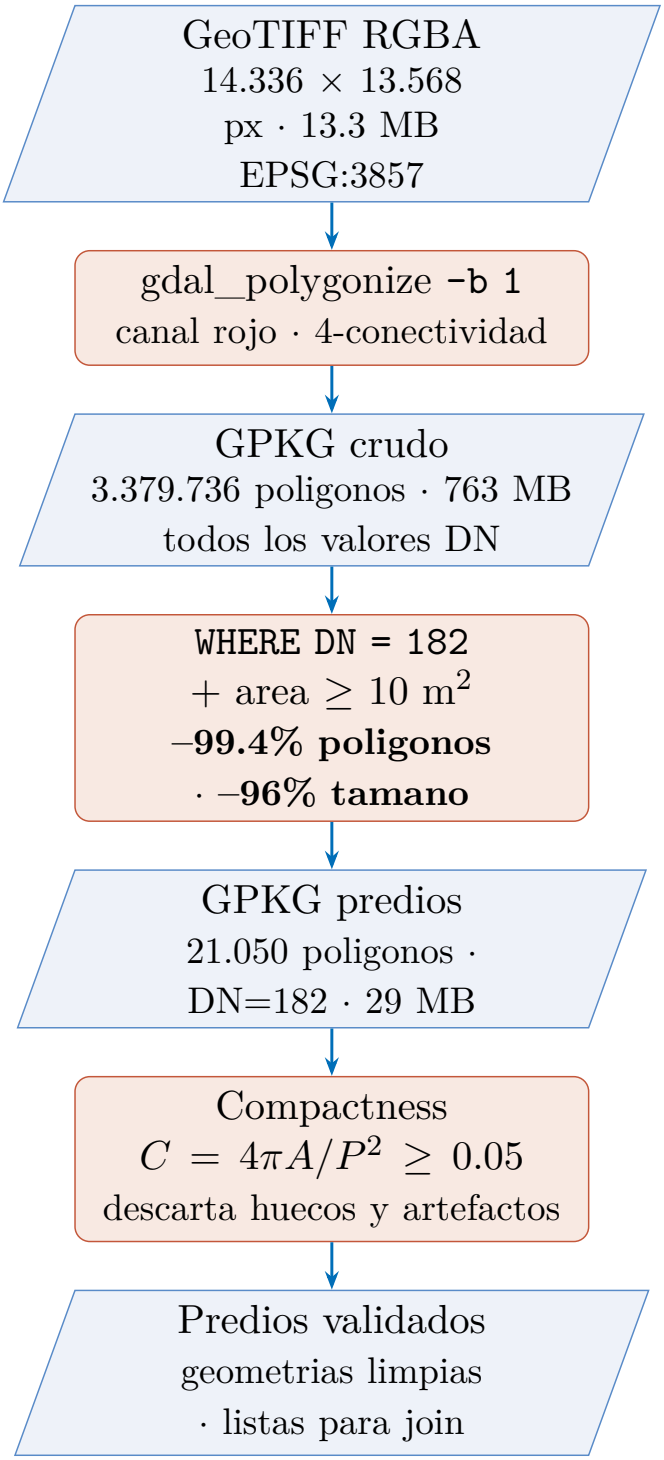

Figura 6.2: Proceso de vectorización en tres pasos: polygonización GDAL del canal rojo (Band 1), filtrado por valor DN=182 y área mínima, y depuración por índice de compactness. En Santiago Centro, el proceso reduce 3.38M polígonos crudos a ≈21.000 predios válidos.

renderizado y simplemente no entrega contenido.

La solución consistió en combinar dos resoluciones. Para cada zona con tiles blank

en z19, se descargó el mismo tile a zoom 17 (≈1.2 m/px). El zoom 17 tiene menor densidad por tile y siempre renderiza. Los polígonos obtenidos a z17 se reproyectan y fusionan con los de z19 en las zonas de cobertura faltante. En Santiago Centro, esta corrección recuperó **599 predios** que no aparecían en la descarga original a z19 (GeoServer Project Steering Committee, 2024).

6.4.3 Filtrado por compactness

La vectorización raster genera polígonos que incluyen no solo predios sino también los *agujeros* en el interior de predios con patios o huecos. Para distinguir estos agujeros (generalmente de forma irregular) de predios reales, se utiliza el índice de compactness o cociente isoperimétrico:

$$C = \frac{4\pi A}{P^2} \tag{6.2}$$

donde A es el área del polígono y P su perímetro. Un valor $C \to 1$ indica forma circular (compacta); los agujeros producidos por texto o artefactos de renderizado tienden a tener $C < 0.1$. El umbral de clasificación implementado en el pipeline es $C \geq 0.05$ para retener un polígono como candidato a predio válido.

6.5 Fase 3: Pipeline de dos pasadas (z19 + z17)

La fusión de las dos resoluciones se implementa en `3_dos_pasadas.py`. El algoritmo opera en dos etapas:

1. **Base z19**: se toman todos los polígonos vectorizados de la descarga a zoom 19. Estos representan la geometría de mayor precisión (≈30 cm/px) y son la fuente primaria.

2. **Complemento z17**: para cada zona del bounding box de la comuna que no esté cubierta por ningún polígono z19 (es decir, celdas de 256×256 px a z19 que resultaron blank), se inyectan los polígonos equivalentes obtenidos de la vectorización a zoom 17. La zona de cobertura de cada tile se verifica mediante una operación espacial de diferencia.

La estrategia es conservadora: los polígonos z19 nunca son reemplazados por los de z17. Solo se agregan los z17 donde z19 está vacío. Esto preserva la máxima precisión

disponible en la mayor parte del territorio, mientras garantiza cobertura en zonas donde el GeoServer no renderiza a z19 (GeoServer Project Steering Committee, 2024).

6.6 Fases 4 a 6: Join espacial y consolidación

6.6.1 Spatial join CSV–GeoJSON

El objetivo de esta etapa es asignar a cada predio del CSV de Fase 0 el polígono catastral que le corresponde en el GeoPackage de Fase 3. La clave de enlace es el `rol_base`: un identificador que apunta al terreno físico (lote o edificio) al que pertenece cada predio, independientemente de si ese terreno tiene una o múltiples unidades (Jordahl et al., 2024).

El join se realiza con GeoPandas mediante una operación de intersección punto-polígono: para cada predio con coordenadas válidas (`lat`, `lon`), se construye un punto geométrico y se busca el polígono vectorizado que lo contiene:

```
import geopandas as gpd
from shapely.geometry import Point

# Construir GeoDataFrame de puntos desde CSV
gdf_puntos = gpd.GeoDataFrame(
    df,
    geometry=[Point(lon, lat)
              for lon, lat in zip(df["lon"], df["lat"])],
    crs="EPSG:4326"
)
# Leer poligonos vectorizados
gdf_pols = gpd.read_file("predios.gpkg").to_crs("EPSG:4326")

# Spatial join: each point -> polygon that contains it
resultado = gpd.sjoin(
    gdf_puntos, gdf_pols,
    how="left", predicate="within"
)
```

Listing 6.1: Spatial join entre puntos de predios y polígonos vectorizados

6.6.2 El concepto de `rol_base`

Un mismo terreno catastral puede contener múltiples predios: un edificio de departamentos tiene un predio por unidad (DP 101, DP 102, etc.) más un predio `90xxx` que representa los bienes comunes del edificio completo. En el mapa WMS, el polígono dibujado corresponde al terreno completo, no a la unidad individual. Por tanto, el polígono se identifica con el rol del predio `90xxx`, denominado `rol_base`.

La detección automática de si un predio es lote simple o multi-unidad se hace agrupando todos los predios con coordenadas válidas por posición geográfica (redondeada a 10^{-6} grados, equivalente a $\approx$11 cm). Si en una posición hay un solo predio, es lote simple y su `rol_base` es su propio rol. Si hay múltiples predios en la misma posición, se lanza un escaneo de la API buscando el predio `90xxx` correspondiente, con variaciones ± 2 alrededor del candidato 90000 + predio_min.

6.6.3 Consolidación final (Fase 6)

El script `6_consolidar_final.py` genera el producto definitivo por comuna: un archivo Parquet con una fila por predio y geometría vectorial asignada. Los campos del polígono (`pol_area_m2`, `pol_tipo_predio`, `pol_n_roles`, `pol_n_roles_unitarios`, `pol_direccion_base`, `pol_destino_base`) se adjuntan a cada fila del CSV enriquecido, produciendo un dataset completamente plano sin necesidad de joins adicionales para su análisis.

6.7 Fase 7: QA y validación nacional

La validación de cobertura y calidad opera sobre el producto final de Fase 6. El script `7_qa_validacion.py` descarga cada archivo Parquet desde S3 y aplica una batería de verificaciones (GDAL/OGR contributors, 2024):

- **Integridad de filas**: número de predios en el CSV de Fase 0 igual al número de filas en el Parquet de Fase 6.
- **Cobertura del join**: porcentaje de predios con polígono asignado (`pct_con_geom`). El umbral de alerta es $< 80\%$.
- **Validez topológica**: ausencia de geometrías auto-intersectantes, verificada mediante `is_valid` de Shapely (Gillies et al., 2024b).

- **Bounds geográficos de Chile**: todos los centroides de polígonos deben caer dentro de los límites $-76° \leq \lambda \leq -65{,}5°$ (longitud) y $-56° \leq \phi \leq -17°$ (latitud).
- **Sistema de referencia**: CRS del GeoPackage debe ser EPSG:4326.

El reporte de salida es un CSV con una fila por comuna y todas las métricas anteriores (`/tmp/fase7_reporte.csv`), más un archivo de texto con las comunas que generaron alertas. Este proceso puede ejecutarse incrementalmente (`-resume`) sobre comunas nuevas sin reprocesar las ya validadas.

6.8 Diccionario de datos del producto final

El dataset resultante contiene campos provenientes de tres fuentes distintas, consolidados en una fila por predio. La Tabla 6.4 presenta los campos principales y su origen.

Tabla 6.4: Diccionario de datos del producto final (campos principales)

Campo	Origen	Descripción
Identificación		
`v`	Calculado	Clave única: `comuna|manzana|predio`
`rol`	Calculado	Rol: `manzana-predio`
`rol_base`	Fase 4	Rol del terreno/edificio padre
Atributos del TXT semestral		
`txt_direccion`	TXT SII	Dirección oficial del predio
`txt_avaluo_total`	TXT SII	Avalúo fiscal total (CLP)
`txt_cod_destino`	TXT SII	Código de destino (1 dígito)
Atributos de la API		
`lat`, `lon`	API SII	Coordenadas WGS84 del predio
`valorTotal`	API SII	Avalúo fiscal total vía API (CLP)
`supTerreno`	API SII	Superficie de terreno
`supConsMt2`	API SII	Superficie construida (mš)
`valorComercial_clp_m2`	API SII	Valor comercial suelo (CLP/mš)
`destinoDescripcion`	API SII	Destino en texto (HABITACIONAL, etc.)
`ah`	API SII	Código de Área Homogénea
`ah_valorUnitario`	API SII	Valor unitario del AH
Atributos del polígono (WMS vectorizado)		
`pol_area_m2`	Fase 2–3	Área del polígono catastral
`pol_tipo_predio`	Fase 4	`lote_simple` o `multi_unidad`
`pol_n_roles_unitarios`	Fase 4	Unidades en el edificio

Además de los campos de la Tabla 6.4, el dataset incluye datos del Observatorio de Mercado de Suelo Urbano del SII (prefijo `obs_`), que proveen series históricas de transacciones por Área Homogénea para los años 2022–2025, y datos de Estudio de Avalúos Comerciales (prefijo `cap__ah_muestra_*`) con el valor comercial de suelo por EAC (versiones 14 y 15). Este nivel de detalle histórico convierte al dataset en una herramienta para análisis de plusvalía y evolución de valores de suelo a escala de manzana.

6.9 Catastral.cl: plataforma de distribución

El pipeline de extracción catastral produce los datos; la plataforma **Catastral.cl** los hace accesibles. Este capítulo cierra describiendo la arquitectura de distribución construida sobre el dataset, dado que la plataforma es el resultado aplicado que concreta el objetivo central de la tesis: eliminar las barreras de acceso para cualquier actor, con independencia de su capacidad técnica.

6.9.1 Separación de dominios

La arquitectura separa físicamente dos responsabilidades en dos servidores distintos, decisión de diseño que aísla la carga de extracción de la carga de servicio al público. El **VPS extractor** ejecuta el pipeline de extracción y almacena los productos en almacenamiento de objetos S3 (Hetzner), con una ruta uniforme por archivo y comuna. El **VPS de plataforma** consume esos archivos mediante un pipeline ETL semestral y los sirve al público a través de Nginx y Cloudflare (reverse proxy y CDN), un backend FastAPI, una base de datos PostgreSQL/PostGIS y una SPA React. Esta separación permite que una re-ejecución completa del pipeline no afecte la disponibilidad de la plataforma en producción. El diagrama detallado de la arquitectura de dos dominios se documenta en el Anexo A.6.

6.9.2 Pipeline ETL de carga a PostgreSQL

Los archivos generados por el pipeline de extracción catastral se cargan en PostgreSQL a través de un pipeline ETL de 9 scripts (`01_download_csvs.py` a `09_load_geometries.p`
El proceso de carga tabular tarda aproximadamente 25 minutos; la carga de geometrías poligonales desde los GeoJSON de S3 (`09_load_geometries.py`) tarda ≈5 horas adicionales, ya que procesa archivos de hasta 11 GB usando `ijson` en

streaming para no saturar la RAM del servidor. El pipeline produce dos tablas principales:

- `catastro_actual`: 9.407.339 filas correspondientes al semestre vigente (2025-S2), con 40 columnas incluyendo la columna `geom geometry(Geometry, 4326)` que almacena los polígonos vectorizados de 5,67 millones de predios. Es la fuente de las estadísticas de cobertura del catálogo y la base sobre la que se generan los archivos enriquecidos por comuna.
- `catastro_historico`: 136.630.730 filas que cubren 16 semestres (2018-S1 a 2025-S2), con 13 columnas. Permite calcular la evolución del avalúo fiscal de cualquier predio en los últimos 7 años.

Los índices de PostgreSQL cubren búsqueda fuzzy de direcciones (GIN trigram), búsqueda espacial por coordenadas (GIST PostGIS), y filtros por comuna, destino, avalúo y superficie, lo que deja la base de datos preparada para incorporar consultas individuales de predios (búsqueda por dirección, detalle por rol, predios en un radio dado) como evolución futura de la interfaz pública.

6.9.3 API REST

El backend expone sus endpoints organizados en routers FastAPI. Los endpoints centrales se organizan en cuatro categorías:

Catálogo de comunas. Lista las 343 comunas disponibles con sus estadísticas de cobertura: porcentaje de predios con polígono, proporción urbana y agrícola, y peso de los archivos por formato.

Capas por comuna. Devuelve URLs prefirmadas S3 con expiración de 15 minutos para los archivos enriquecidos (CSV y Parquet) a usuarios autorizados mediante pago social.

Descargas históricas gratuitas. Lista 16 CSVs semestrales sin geometría (≈22,8 GB en total), descargables sin registro.

Pago social y administración. Validación de publicaciones de LinkedIn y emisión de tokens de acceso, *domain grants* organizacionales, y el panel de administración con métricas de uso.

6.9.4 Frontend y modelo de acceso

El frontend React implementa las páginas públicas bajo el dominio `catastral.cl`: la portada con las cifras globales del dataset, el catálogo de **Comunas** y la **Tienda** (343 comunas con estadísticas de cobertura y formatos de descarga), la sección de **Descargas** históricas gratuitas y la sección de **Metodología**, que documenta públicamente el pipeline y el diccionario de datos, además del panel privado de administración.

La plataforma implementa un modelo de **pago social** con tres niveles de acceso:

1. **Acceso gratuito**: navegación de la portada y del catálogo de comunas con sus estadísticas de cobertura, documentación metodológica completa, y descarga sin registro ni pago de los 16 CSVs semestrales históricos sin geometría (series de avalúo 2018–2025).

2. **Acceso con pago social**: archivos enriquecidos por comuna con geometría vectorial en dos formatos (CSV y Parquet), listos para análisis en QGIS, PostGIS o Python/GeoPandas. El *precio* de estas capas no es pecuniario sino de reconocimiento: el usuario publica una mención de Catastral.cl en LinkedIn como condición para recibir el enlace de descarga, o declara haberlo hecho antes mediante la opción de buena fe *Ya compartí en LinkedIn antes*, que libera la descarga de inmediato. El objetivo es liberar el dato a quien lo necesita y a la vez generar visibilidad orgánica de la plataforma en la red profesional donde se concentra su público objetivo, de manera análoga a la atribución que requieren las licencias de software libre.

3. **Acceso organizacional**: *domain grants* que permiten que todos los emails de un dominio corporativo o institucional accedan automáticamente sin pasar por el flujo individual.

La autenticación *passwordless* por OTP (código de 6 dígitos vía Resend, TTL 10 min) está reservada exclusivamente para el administrador, accesible mediante la URL directa `/admin`; el botón de login no aparece en la navegación pública. Los usuarios públicos no tienen cuenta: el acceso a las capas vectoriales se gestiona mediante un token anónimo UUID almacenado en la tabla `share_tokens` y en una cookie `httpOnly` con TTL de 1 año, generado al validar la publicación de LinkedIn. El sistema incluye un mecanismo de recuperación por LinkedIn username en caso de pérdida de la cookie. Las descargas son URLs prefirmadas S3 con expiración de 15 minutos, sin posibilidad de redistribución masiva.

6.9.5 Cobertura y actualización semestral

De las 346 comunas de Chile, 343 tienen archivos completos disponibles en S3 y son las que se publican en la plataforma; la descarga pública por comuna se ofrece en formatos CSV y Parquet. Las tres comunas sin datos son Antártica (12202), Isla de Pascua (05201) y Juan Fernández (05104), territorios insulares o antárticos sin cobertura del WMS del SII a las resoluciones utilizadas. La actualización es semestral, sincronizada con la publicación del nuevo archivo `BRTMPNACROL` por el SII: el pipeline ETL (`05_run_all.py`) carga los datos tabulares en ≈25 minutos; la actualización de geometrías (`09_load_geometries.py`) requiere ≈5 horas adicionales para procesar los GeoJSON de S3.

Capítulo 7

Resultados, Casos de Uso e Impacto

Este capítulo presenta los dos resultados principales de esta tesis: el dataset predial nacional construido por el pipeline de extracción catastral, y la plataforma **Catastral.cl** que lo distribuye públicamente. Se evalúa la utilidad de ambos para la industria inmobiliaria mediante casos de uso concretos, y se examina el impacto en el acceso democrático a la información pública.

7.1 El dataset resultante

7.1.1 Cobertura nacional

El dataset producido por el pipeline de extracción catastral (batch de marzo de 2026, semestre 2025-2 del SII) cubre la totalidad del territorio catastral de Chile. La Tabla 7.1 resume las métricas de cobertura a nivel nacional.

La cobertura de polígonos vectoriales alcanza el 60,3% a nivel nacional. El 39,7% restante corresponde principalmente a predios rurales y agrícolas sin levantamiento cartográfico en el WMS del SII (los predios existen en el TXT semestral pero el servidor no publica su geometría a las resoluciones utilizadas), y a unidades dentro de edificios que comparten el polígono del lote padre. Esta cobertura refleja el máximo disponible en las fuentes públicas del SII, no una limitación del pipeline. Los predios sin polígono están presentes en el dataset con todos sus atributos tabulares intactos.

Tabla 7.1: Cobertura del dataset predial nacional

Métrica	**Valor**
Comunas cubiertas	343 de 346
Predios totales	9.55 millones
Predios con polígono vectorial	5.671.116 (60,3% del total)
Urbanos (U)	4.821.993 de 7.285.199 (66,2%)
Rurales (R)	765.565 de 2.018.772 (37,9%)
Habitacional (H)	3.874.400 de 5.987.385 (64,7%)
Agrícola (A)	269.727 de 998.053 (27,0%)
Predios sin polígono (sin levant. cartogr.)	3.736.223 (39,7%)
Resolución geométrica	$\approx$ 30 cm/píxel (zoom 19)
Sistema de referencia	EPSG:4326 (WGS84)
Formato de entrega	CSV y Parquet por comuna
Volumen total comprimido	$\approx$ 18 GB

7.1.2 Variables disponibles

Cada predio del dataset incluye campos de tres categorías distintas, según lo descrito en el Capítulo 3:

Los **atributos de identificación y localización** incluyen el rol predial único (`rol`), el rol base del terreno (`rol_base`), la dirección oficial según el TXT semestral y según la API, el código de comuna, la latitud y longitud del centroide, y la geometría del polígono del lote o edificio.

Los **atributos de valor** incluyen el avalúo fiscal total, afecto y exento en CLP (fuente: TXT semestral y API), el valor comercial de suelo en CLP/mš por Área Homogénea, el valor unitario del EAC vigente (Estudio de Avalúos Comerciales), y series históricas de valores de suelo para los años 2022–2025 (EAC 14 y EAC 15). Todos los valores monetarios se expresan en pesos chilenos nominales a la fecha de publicación de cada corte semestral del SII; dado que la fecha de corte es conocida, su conversión a unidades reajustables como la UF es directa, y la incorporación de columnas derivadas en UF, que facilitarían la comparación intertemporal de avalúos, se identifica como una mejora futura del dataset.

Los **atributos catastrales** incluyen la superficie de terreno, la superficie construida en mš, el código de destino predial en texto (`HABITACIONAL`, `COMERCIO`, `INDUSTRIA`, `AGRICOLA`, etc.), el tipo de predio (`lote_simple` o `multi_unidad`), el número de unidades en el edificio (`pol_n_roles_unitarios`), y el área del polígono vectorial calculada en mš.

7.1.3 Comparación con alternativas comerciales

La Tabla 7.2 compara Catastral.cl con las principales alternativas disponibles para acceder a datos prediales a escala nacional en Chile, evaluadas en cinco dimensiones: cobertura territorial, costo de acceso masivo, precisión geométrica, disponibilidad de atributos de valor y nivel de apertura.

Tabla 7.2: Comparación con fuentes alternativas de datos prediales en Chile

Fuente	Cobertura	Costo acceso masivo	Geometría	Nivel de acceso
Catastral.cl (esta tesis)	343 comunas, 9,4M predios	≈USD 150/mes infra	Vectorial (≈30 cm)	Dataset abierto, API, descarga bulk
SII WMS	343 comunas	Gratuito individual; sin bulk	Solo raster (PNG)	Consulta tile a tile; sin API masiva
SII portal web	346 comunas	Gratuito individual	Sin geometría	Consulta predio a predio
Operadores privados	Parcial	USD 2.000–6.000 / comuna	Variable	API privada con cuotas
Portales inmobil.	Parcial (avisos)	USD 50–500/mes	Puntual	Sin descarga masiva
CBR	Nacional (legal)	USD 200+/mes suscr.	Sin geometría	Portal web; sin API
SIG municipal	30–40 comunas	Variable / convenio	Alta donde disponible	Heterogéneo (PDF, SHP)

La literatura sobre datos abiertos documenta que el valor económico de liberar información pública se concentra justamente en las dimensiones que las alternativas fragmentan (World Bank, 2021; McKinsey Global Institute, 2013). El diferenciador principal de Catastral.cl es la combinación de cobertura nacional completa, geometría vectorial de alta resolución, atributos de valor de suelo del propio SII, y distribución a través de una plataforma web con descarga masiva. Las alternativas existentes ofrecen alguna de estas dimensiones pero no todas simultáneamente: el SII provee cobertura nacional pero sin formato vectorial ni descarga masiva; los operadores privados ofrecen datos estructurados pero a costos que restringen el acceso a actores con presupuesto; y los SIG municipales proveen alta precisión pero solo para una fracción del territorio (Open Knowledge Foundation, 2023; Janssen et al., 2012).

7.2 La plataforma Catastral.cl

7.2.1 Descripción y alcance

Catastral.cl es la interfaz pública del sistema desarrollado en esta tesis. Es el resultado aplicado que concreta el objetivo de democratización: transforma un dataset técnicamente producido por el pipeline en un servicio web consultable por cualquier actor, sin requerir conocimientos de programación, acceso a bases de datos ni software especializado.

La plataforma está operativa en producción bajo el dominio `catastral.cl`, servida desde un VPS dedicado en Helsinki (Hetzner, 32 cores, 122 GB RAM) con Nginx como *reverse proxy*, Cloudflare para DNS y CDN con certificado SSL Full Strict, FastAPI como backend, PostgreSQL 16 + PostGIS 3.5 como motor de datos, y una Single Page Application React 19 como frontend.

Figura 7.1: Página de inicio de `catastral.cl`: 9,5 millones de predios, 343 comunas, 112 variables y 9,1 millones de polígonos vectorizados. Disponible en `https://catastral.cl`.

7.2.2 Base de datos y cobertura

La base de datos PostgreSQL integra los productos de todas las fases del pipeline en dos tablas principales:

`catastro_actual` almacena 9.407.339 predios del período vigente (2025-S2) con 40 columnas, incluyendo la columna `geom geometry(Geometry, 4326)` que contiene los polígonos vectorizados cargados desde los GeoJSON de la Fase 6 del pipeline. Corresponde al producto de la Fase 0 enriquecido con coordenadas, superficies, valor comercial de suelo por Área Homogénea, código de destino predial y geometría poligonal. Los índices incluyen búsqueda fuzzy de direcciones (GIN trigram), índice espacial GIST sobre coordenadas y sobre la columna `geom` (`idx_catastro_actual_geom`), y B-tree sobre comuna, destino, avalúo y superficie.

`catastro_historico` almacena 136.630.730 filas que cubren 16 semestres entre 2018-S1 y 2025-S2, con 13 columnas. Esta tabla es la fuente de los 16 archivos semestrales publicados en la sección de descargas, y permite reconstruir la evolución del avalúo fiscal de cualquier predio a lo largo de 7 años, una perspectiva sin precedente en fuentes públicas chilenas.

7.2.3 Funcionalidades de la plataforma

La plataforma implementa tres niveles funcionales accesibles desde el mismo dominio:

Exploración gratuita del catálogo: cualquier usuario puede navegar sin registro la portada con las cifras globales del dataset (9,5 millones de predios, 343 comunas, 112 variables y 9,1 millones de polígonos) y el catálogo de comunas, que presenta para cada una el porcentaje de cobertura de polígonos, la proporción de predios urbanos y agrícolas, y el peso de los archivos disponibles. La sección de metodología documenta públicamente el proceso de construcción del dataset y el diccionario de datos completo.

Descargas históricas gratuitas: la sección `/descargas` ofrece los 16 archivos CSV semestrales (datos tabulares sin geometría, desde 2018 hasta 2025), descargables directamente sin registro ni pago, representando aproximadamente 22,8 GB de datos históricos.

Capas con geometría vectorial: la sección `/tienda` presenta el catálogo de 343 comunas con estadísticas por cada una: porcentaje de cobertura de polígonos (**C**), proporción de predios urbanos (**U**), proporción de predios agrícolas (**A**) y peso

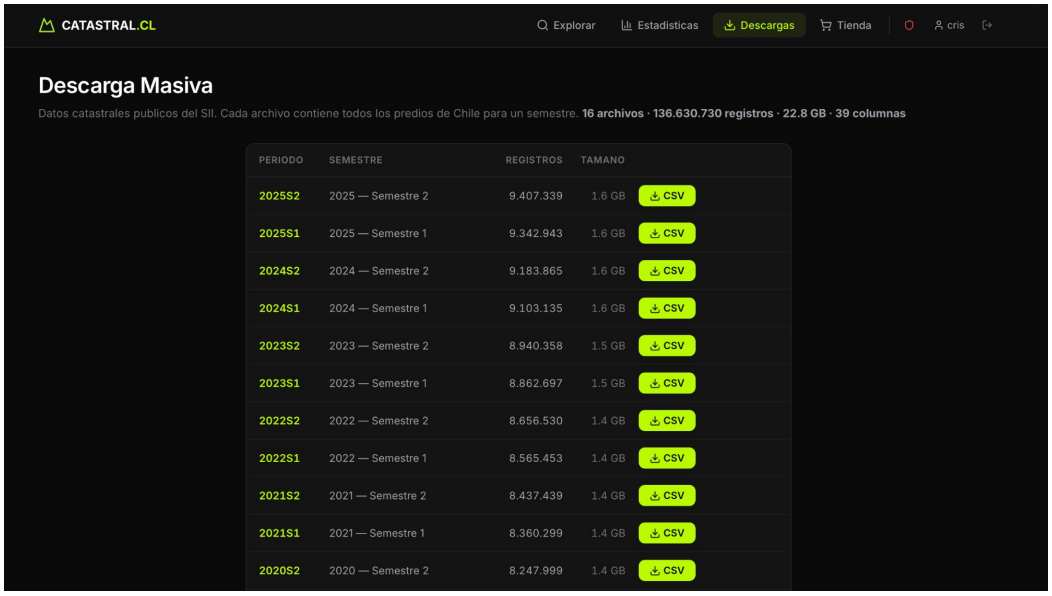

Figura 7.2: Sección `/descargas` de `catastral.cl`: 16 archivos CSV semestrales con 136,6 millones de registros en total, de acceso libre y gratuito sin registro.

del archivo de descarga (desde 0,7 MB en Torres del Paine hasta 11 GB en Las Condes, mediana ≈184 MB). El catálogo es ordenable por cobertura y por peso. Para cada comuna se ofrecen dos formatos de descarga: CSV plano y Parquet con geometría, compatibles con QGIS, PostGIS y Python/GeoPandas. El acceso no requiere pago monetario: el usuario realiza un **pago social**, una contraprestación de reconocimiento y no pecuniaria, publicando una mención de Catastral.cl en LinkedIn. El dato se libera completamente; el precio es visibilidad en la red profesional.

La cobertura es de 343 comunas con archivos completos disponibles en S3, publicadas en la plataforma. Las tres comunas sin datos son Antártica (12202), Isla de Pascua (05201) y Juan Fernández (05104), territorios sin cobertura WMS del SII a las resoluciones utilizadas por el pipeline.

7.2.4 Autenticación y pagos

La plataforma opera sin registro para usuarios públicos. La autenticación *passwordless* por OTP (código de 6 dígitos vía email, Resend, TTL 10 min) está reservada exclusivamente para el administrador, quien accede a través de la URL directa `/admin`; el botón de login no aparece en la navegación pública.

Para acceder a las capas con geometría vectorial, el usuario público completa el pago social siguiendo tres pasos: (1) copia el texto sugerido por la plataforma (que men-

ciona a `@crishernandezco` y `@tremen-tech`) y publica un post en LinkedIn; (2) pega la URL del post en el formulario; (3) el backend valida el formato del URL mediante expresión regular sobre dominios `linkedin.com` y crea un token UUID almacenado en la tabla `share_tokens` y en una cookie `httpOnly` con TTL de 1 año, lo que otorga descargas ilimitadas para las 343 comunas disponibles. El flujo contempla además la opción *Ya compartí en LinkedIn antes*, que libera las descargas de inmediato sobre la base de la declaración del usuario, sin exigir una nueva publicación. Si el usuario limpia sus cookies, puede recuperar el acceso ingresando su LinkedIn username: el backend busca el token en `share_tokens.linkedin_username` y restaura la sesión. Las descargas son URLs prefirmadas S3 con expiración de 15 minutos.

7.2.5 Modelo de distribución: pago social

El modelo de acceso de Catastral.cl distingue dos niveles de servicio. La navegación del catálogo y las descargas históricas tabulares (16 semestres) son de acceso libre y sin registro. Para las capas con geometría vectorial (CSV enriquecido y Parquet), el acceso requiere un *pago social*, una contraprestación de reconocimiento y no pecuniaria: la publicación de una mención de la plataforma en LinkedIn. Esta decisión de diseño responde a dos objetivos: eliminar la barrera económica para usuarios sin capacidad de pago (investigadores, municipios, emprendedores en etapa temprana) y generar difusión orgánica en la red profesional donde se concentra la audiencia objetivo. Esta decisión se inscribe en la discusión sobre modelos sostenibles de provisión de datos abiertos, donde la captura de valor no monetaria es una alternativa reconocida a la venta directa (Attard et al., 2015; Zuiderwijk and Janssen, 2014).

Conceptualmente, el pago social es análogo a la cláusula de atribución del software libre y de las licencias abiertas tipo Creative Commons BY: el activo se entrega completo y sin costo monetario, y la contraprestación es el reconocimiento público de la fuente. La analogía se extiende al mecanismo de verificación, que es deliberadamente laxo: la opción *Ya compartí en LinkedIn antes* libera las descargas sobre la base de la sola declaración del usuario. Esta decisión de diseño refleja que el objetivo del modelo es la difusión y el reconocimiento, no el control de acceso: la barrera efectiva para cualquier usuario es cercana a cero.

La hipótesis subyacente es que la demanda por datos catastrales en Chile se segmenta en dos grupos: un segmento profesional (consultoras, desarrolladores GIS, operadores de infraestructura) con disposición a pagar por el dato, y un segmento social (academia, sector público, periodismo de datos) con necesidad legítima pero

sin recursos para adquirir bases de datos comerciales. El modelo busca que ambos segmentos accedan al mismo dataset, donde el primero contribuye con visibilidad y el segundo se beneficia de la ausencia de barrera monetaria.

Tabla 7.3: Segmentación de usuarios según modelo de pago social de Catastral.cl

Tipo de usuario	**Perfil**	**Acceso**
Empresa de geocoding / GIS	Construye productos comerciales sobre el dato	Pago social (post LinkedIn)
Consultora inmobiliaria	Prospección, due diligence	Pago social (post LinkedIn)
Operador de infraestructura	Localización de activos a escala	Pago social (post LinkedIn)
Investigador / academia	Análisis territorial, tesis	Gratuito (descargas históricas)
Municipio / planificador	Gestión territorial pública	Gratuito o domain grant
Periodista de datos	Investigación de interés público	Gratuito
Ciudadano	Consulta de su predio	Gratuito

Desde la perspectiva de sostenibilidad, el modelo presenta una propiedad de rendimientos crecientes: a medida que crece el número de usuarios que realizan el pago social, la difusión orgánica reduce el costo de adquisición de nuevos usuarios, lo que potencialmente permite mantener el acceso libre sin comprometer la operación del sistema.

El pago social es el modelo de distribución vigente, pero no agota las vías de sostenibilidad de la plataforma. Una línea de desarrollo futuro es complementarlo con productos de mayor valor agregado, como la venta de informes específicos por predio o por cartera (informes de plusvalía, due diligence, prospección de ubicaciones), manteniendo el acceso al dato base bajo el modelo actual. Este esquema preservaría el objetivo de democratización, ya que el dato crudo permanece accesible, y monetizaría el análisis construido sobre él, que es donde se concentra la disposición a pagar del segmento profesional.

7.2.6 Rendimiento observado

Las consultas centrales de la capa de datos responden dentro de umbrales aceptables para uso interactivo, medidas sobre la base de datos de producción: la búsqueda por dirección con índice trigram tarda ≈200 ms; la búsqueda espacial por coordenadas con índice GIST, ≈50 ms; y la recuperación de polígonos vectoriales, ≈1,7 s para

50 polígonos cercanos usando filtro de bounding box GIST más `ST_DWithin`. Estas capacidades de consulta individual residen hoy en la capa de datos y no están expuestas en la interfaz pública, que se concentra en el catálogo y las descargas. Las estadísticas agregadas se precalientan al arranque (≈10–11 segundos) y se sirven desde caché con TTL de 1 hora a ≈13–14 ms. El endpoint `/api/health` responde en 15 ms como línea base del stack. La carga inicial de geometrías desde S3 (5,67 millones de polígonos procesados con `ijson` en streaming) tomó ≈5 horas en el VPS de la plataforma.

7.3 Evaluación de precisión geométrica

Dado que el dataset construido constituye la base para los análisis presentados en las secciones siguientes, resulta fundamental evaluar su precisión espacial. En particular, se busca cuantificar la fidelidad del proceso de vectorización a partir de imágenes WMS, para determinar si el dataset es suficientemente preciso para los casos de uso propuestos.

7.3.1 Metodología y alcance de la métrica

Para cada predio del dataset coexisten dos representaciones espaciales que el SII produce de forma independiente:

1. **Coordenada API** (`lat`, `lon`): entregada por el endpoint `getPredioNacional`, corresponde a la ubicación oficial registrada para cada predio.
2. **Centroide del polígono**: calculado a partir del polígono vectorizado desde el WMS del SII.

La distancia entre ambas, calculada mediante la fórmula de Haversine, es la métrica de evaluación. Es importante precisar su alcance: al comparar dos productos del mismo organismo (la coordenada tabular y la geometría cartográfica del SII), la métrica cuantifica la **consistencia inter-fuente** entre ambas representaciones, no la exactitud absoluta contra una fuente de verdad externa (*ground truth*), que no está disponible en formato vectorial abierto en Chile (véase la discusión de amenazas a la validez en el Capítulo 8). Una consistencia alta indica que la vectorización reproduce fielmente la posición que el propio SII asigna al predio. Este enfoque de evaluación contrasta con el de la literatura de extracción de límites catastrales por aprendizaje

profundo, que valida contra anotaciones manuales y reporta métricas de *recall* y precisión sobre la detección de bordes visibles (Xia et al., 2019).

Para superar la limitación de una evaluación sobre una única comuna, se construyó una muestra estratificada de cuatro comunas de perfiles y regiones distintos, sobre las que se dispone simultáneamente de coordenada API y polígono vectorizado, sumando $n = 305.119$ predios. La asociación entre cada predio y su polígono se realizó por contención espacial (*point-in-polygon*).

7.3.2 Resultados

La Tabla 7.4 reporta el error de consistencia por comuna y agregado. La distribución es fuertemente asimétrica a la derecha en todas las comunas: la mediana agregada es de 2,0 metros mientras la media se eleva a 4,5 metros por los percentiles superiores. La Tabla 7.5 desagrega el resultado agregado en umbrales acumulativos con sus intervalos de confianza al 95% (método de Wilson); el gran tamaño muestral produce intervalos muy estrechos (inferiores a ±0,3 puntos porcentuales).

Tabla 7.4: Consistencia inter-fuente por comuna: distancia entre coordenada API y centroide del polígono (método de la tesis, proceso 2025-S2)

Comuna	n	**Mediana**	**Media**	**p95**	**<5 m**	**<20 m**
Santiago (densa)	236.802	2,1 m	4,5 m	14,9 m	68,6%	97,3%
Independencia	54.645	2,0 m	5,2 m	17,1 m	63,8%	96,7%
Lebu (costa)	6.927	0,1 m	1,7 m	8,7 m	91,0%	98,2%
Aysén (patagónica)	6.745	0,2 m	2,7 m	17,0 m	85,1%	96,8%
Agregado	**305.119**	**2,0 m**	**4,5 m**	**15,8 m**	**68,6%**	**97,2%**

Tabla 7.5: Proporción de predios dentro de umbrales de consistencia (agregado, $n = 305.119$), con IC 95%

Umbral	**Predios**	**IC 95%**
< 5 m	68,6%	[68,5%, 68,8%]
< 10 m	84,9%	[84,7%, 85,0%]
< 20 m	97,2%	[97,2%, 97,3%]
< 50 m	99,9%	[99,9%, 100%]

El resultado más robusto, y consistente entre las cuatro comunas, es que el **97,2% de los predios presenta una consistencia inferior a 20 metros**. La mediana varía con la densidad predial: en comunas de altísima densidad (Santiago, Independencia) es de ≈2 metros y la proporción bajo 5 metros baja al 64–69%, mientras que en comunas menos densas (Lebu, Aysén) la mediana es sub-métrica y la proporción bajo 5 metros supera el 85%. Esta degradación en zonas densas es en parte un artefacto

del método de evaluación: cuando los predios son muy pequeños y contiguos, el punto de la coordenada API puede caer dentro del polígono de un predio vecino, lo que incrementa la distancia medida sin que ello implique un error de la vectorización en sí.

7.3.3 Benchmark: evolución del método

La evaluación anterior corresponde al método documentado en esta tesis (proceso 2025-S2). Una iteración posterior del pipeline (proceso 2026-S1), que reemplaza la descarga por GeoTIFF completo por una descarga en *supercells* con vectorización en bloques, permite un benchmark de la evolución del método. La Tabla 7.6 compara ambos sobre muestras independientes.

Tabla 7.6: Benchmark entre el método de la tesis y su iteración posterior

Dimensión	**Método tesis (2025-S2)**	**Iteración (2026-S1)**
Muestra de precisión (n)	305.119 (4 comunas)	523.439 (4 comunas)
Mediana de consistencia	2,0 m	0,8 m
Consistencia <5 m	68,6%	85,7%
Consistencia <20 m	97,2%	97,2%
Cobertura de polígonos	60,3%	99,96%
Cobertura de coordenadas	n/d	99,99%

La mejora inequívoca es la de **cobertura**: la iteración eleva la fracción de predios con polígono del 60,3% al 99,96% (9,57 millones de predios), reduciendo el residuo sin geometría del 39,7% al 0,04% (4.097 roles). En cuanto a la consistencia posicional, la mediana mejora de 2,0 a 0,8 metros; conviene notar, sin embargo, que las dos evaluaciones se realizaron sobre comunas distintas y con métodos de asociación predio-polígono distintos, por lo que parte de esa mejora aparente no es directamente atribuible al cambio de vectorización. La cobertura, en cambio, no está sujeta a este sesgo de comparación.

7.3.4 Interpretación y limitaciones estadísticas

La consistencia mediana de 2 metros y el 97% de predios bajo 20 metros son suficientes para los casos de uso de analítica urbana y *scoring* presentados en las secciones siguientes, donde las decisiones operan a escala de manzana (≈50–200 m). El dataset no es adecuado, en cambio, para usos legales o de demarcación formal de límites, que requieren precisión catastral (<0,1 m) obtenible solo a través del Conservador de Bienes Raíces.

La evaluación tiene tres limitaciones estadísticas que conviene explicitar. Primera, la métrica mide consistencia inter-fuente, no exactitud contra ground truth; una consistencia alta es condición necesaria pero no suficiente de exactitud, ya que un error compartido por ambas representaciones del SII no sería detectado. Segunda, la muestra de cuatro comunas, aunque grande en número de predios, es una muestra de conveniencia por disponibilidad y no un muestreo aleatorio del territorio nacional, por lo que los intervalos de confianza reportados describen la precisión de la estimación dentro de las comunas evaluadas, no su representatividad nacional. Tercera, el método de asociación por contención espacial introduce un sesgo al alza en la distancia medida en zonas de alta densidad predial, según se explicó. En conjunto, estas limitaciones aconsejan leer las cifras como una cota conservadora de la fidelidad del método, no como su valor exacto.

7.4 Casos de uso para la industria inmobiliaria

7.4.1 Análisis de plusvalía y valor de suelo

El campo `valorComercial_clp_m2` permite calcular el valor de mercado estimado de cualquier predio multiplicando su superficie de terreno (`supTerreno`) por el valor unitario del Área Homogénea. Al combinar esto con la geometría vectorial, es posible generar mapas de valor de suelo a escala de manzana para toda una comuna o región metropolitana, un insumo que la teoría de los mercados inmobiliarios identifica como determinante de la formación de precios y de la reducción de fricciones informacionales (DiPasquale and Wheaton, 1996; Broxterman and Zhou, 2023). Este análisis, que requería horas de trabajo manual con datos dispersos, se convierte en una consulta SQL sobre el dataset:

```sql
SELECT
    manzana,
    COUNT(*)                                    AS n_predios,
    AVG(valorComercial_clp_m2)                  AS
       valor_uf_m2_promedio,
    SUM(supTerreno * valorComercial_clp_m2) AS
       valor_total_manzana_clp
FROM predios
WHERE comuna = '13101'
  AND pol_tipo_predio = 'lote_simple'
  AND supTerreno > 0
```

```
GROUP BY manzana
ORDER BY valor_uf_m2_promedio DESC;
```

Listing 7.1: Valor de suelo por manzana en Santiago Centro

7.4.2 Prospección de predios por destino y superficie

El campo `destinoDescripcion` permite filtrar predios por uso actual (habitacional, comercio, industria, estacionamiento, etc.). Combinado con la superficie de terreno y el valor unitario de suelo, esto habilita consultas de prospección que antes requerían la intervención de un broker especializado con acceso a bases privadas:

```
SELECT
    nombreComuna,
    rol,
    direccion_sii,
    supTerreno,
    valorComercial_clp_m2,
    pol_area_m2
FROM predios
WHERE destinoDescripcion LIKE '%INDUSTRIA%'
  AND supTerreno BETWEEN 2000 AND 10000
  AND comuna IN ('13109','13110','13119','13131')
ORDER BY valorComercial_clp_m2 ASC;
```

Listing 7.2: Prospección de predios industriales en comunas de la RM

7.4.3 Identificación de suelo subutilizado

La combinación de `supTerreno` (superficie de terreno) y `supConsMt2` (superficie construida) permite calcular el índice de ocupación de cada predio (superficie construida / superficie de terreno). Predios con índice de ocupación bajo en zonas de alta densidad o alto valor son candidatos a proyectos de densificación. Esta es una capacidad analítica de alto valor para fondos de inversión inmobiliaria (REIT), oficinas de planificación urbana y desarrolladores, y que no existía como consulta ejecutable sobre datos públicos.

7.4.4 Due diligence con datos objetivos

Para operaciones de adquisición inmobiliaria comercial (oficinas, bodegas, locales), el dataset permite verificar de forma independiente el avalúo fiscal, la superficie de terreno, el destino predial oficial y la geometría del lote antes de contratar una tasación. En el contexto de Newmark Chile descrito en el Capítulo 1, esto significa que un broker puede entregar a su cliente, en cuestión de minutos, un informe de due diligence sobre cualquier predio de Chile con datos directamente extraídos del SII, sin depender de bases de datos privadas ni de información de segunda mano.

7.4.5 Caso aplicado: inteligencia de localización para estaciones de servicio

Los cuatro casos de uso anteriores son analíticos: ilustran qué tipo de consultas habilita el dataset. La plataforma **combustible.tremen.tech** es un caso de uso real y operativo que demuestra que el dataset producido por el pipeline de extracción catastral puede ser la capa base de una aplicación comercial de inteligencia geoespacial completa.

El problema que resuelve

La prospección de ubicaciones para nuevas estaciones de servicio (bencineras) en Chile se realiza tradicionalmente mediante *scouts* en terreno que recorren zonas durante 6 a 12 semanas, evaluando visualmente cada sitio. El proceso es lento, subjetivo, incompleto (es imposible cubrir todos los predios disponibles a escala nacional) y no reproducible entre consultores. La plataforma reemplaza este proceso por un pipeline automatizado de aproximadamente 30 minutos que analiza los 9,5 millones de predios del país.

Cómo consume el dataset de la tesis

El dataset del pipeline de extracción catastral es la capa base de la plataforma: sin él, no existiría la capacidad de filtrar masivamente por superficie, destino y localización geográfica a escala nacional. La Tabla 7.7 muestra los campos del dataset que la plataforma utiliza directamente.

Tabla 7.7: Campos del dataset predial consumidos por la plataforma Combustible

Campo (dataset catastral)	**Uso en Combustible**
`lat`, `lon`	Geolocalización del predio candidato en el mapa
`supTerreno`	Filtro de superficie: 1.200–5.000 mš
`destinoDescripcion`	Filtro por destino: ERIAZO o COMERCIO
`txt_direccion`	Identificación del predio en tabla y popups
`rol`	Identificador único de cada candidato
`valorComercial_clp_m2`	Contexto de valor de suelo para evaluación económica

Pipeline de 10 pasos: del catastro al ranking

El pipeline toma como entrada el dataset del pipeline de extracción catastral y lo cruza con cinco fuentes adicionales (OpenStreetMap, panel GPS de movilidad, API de precios CNE, Google Routes API e imágenes satelitales HERE Maps) en diez etapas organizadas en dos fases:

La **Fase A** aplica cinco filtros progresivos que reducen los 9,5 millones de predios iniciales a aproximadamente 3.600 candidatos: filtro de superficie y destino (SII), uso de suelo permitido según el Instrumento de Planificación Territorial, exclusión por proximidad a equipamientos sensibles (colegios, hospitales, cementerios, iglesias a menos de 50 m), compatibilidad con el patrón vial de las 2.017 bencineras existentes, y umbral mínimo de densidad de dispositivos GPS en radio de 500 m.

La **Fase B** enriquece los candidatos con cuatro métricas adicionales: extrapolación censal de población, clasificación satelital automática mediante CLIP ViT-B-32 para descartar predios con estructuras existentes, precios actuales de combustible de la API de la CNE, y conteo vehicular estimado mediante el modelo de tráfico BPR (*Bureau of Public Roads*) con datos de Google Routes API en modo `TRAFFIC_AWARE`.

Modelo de scoring

El score final de cada candidato combina cuatro componentes normalizados intra-regionalmente, cuyos pesos fueron determinados mediante dos criterios complementarios:

$$S = 0{,}60 \cdot S_{\text{veh}} + 0{,}20 \cdot S_{\text{precio}} + 0{,}15 \cdot S_{\text{ruta}} + 0{,}05 \cdot S_{\text{mov}} \tag{7.1}$$

donde S_{veh} es el flujo vehicular estimado (veh/hora) mediante el modelo BPR, S_{precio} es el precio promedio de combustible en zona de 5 km (mayor precio implica mayor margen potencial), S_{ruta} es el tipo de vía OSM ponderado por el patrón de las bencineras existentes, y S_{mov} es la densidad de dispositivos GPS únicos como *proxy* de actividad humana.

El componente S_{ruta} (15%) se calibró empíricamente a partir del patrón orgánico de localización de las 2.017 estaciones de servicio existentes en Chile: cada bencinera fue matcheada a su segmento OSM más cercano, se contó la distribución por `fclass`, y los pesos de cada categoría vial se normalizaron proporcional a esa distribución real. Los tipos de vía con mayor concentración de estaciones reciben mayor peso; los tipos sin presencia empírica son excluidos del scoring.

Los pesos de los tres componentes restantes (tráfico vehicular 60%, precio zona 20%, movilidad GPS 5%) constituyen una decisión de diseño basada en la lógica sectorial del negocio de combustibles: el flujo vehicular es el determinante primario del volumen de ventas de una estación de servicio (una estación sin tráfico no es viable independientemente de las demás variables), el precio de zona captura el margen potencial, y la movilidad GPS opera como indicador complementario de actividad. Los pesos fueron refinados iterativamente a lo largo del desarrollo del sistema, pasando de un modelo inicial con dos variables (tipo de vía y edificación) a la formulación de cuatro componentes presentada aquí. Una calibración estadística formal de estos pesos requeriría datos de ventas reales por estación, que no están disponibles públicamente en Chile.

El componente de mayor peso, el conteo vehicular (60%), se estima invirtiendo la fórmula BPR:

$$\frac{V}{C} = \left(\frac{v_{\text{libre}}/v_{\text{tráfico}} - 1}{0{,}15} \right)^{0{,}25} \tag{7.2}$$

donde v_{libre} y $v_{\text{tráfico}}$ son las velocidades sin y con congestión obtenidas de Google Routes API, y C es la capacidad vial HCM según el tipo de vía (motorway: 4.400 veh/h; primary: 1.600; secondary: 800; residential: 300).

Dashboard interactivo

El resultado del pipeline es un dashboard React accesible en `combustible.tremen.tech` (protegido por Cloudflare Access) que presenta los 50 mejores candidatos por región en un mapa Leaflet con tiles satelitales HERE Maps y una tabla ordenable con

20 columnas (scores, métricas de tráfico, precios de combustible por tipo). La exportación de candidatos genera un bundle ZIP con un archivo Excel y un KMZ para Google Earth, con los placemarks coloreados por score.

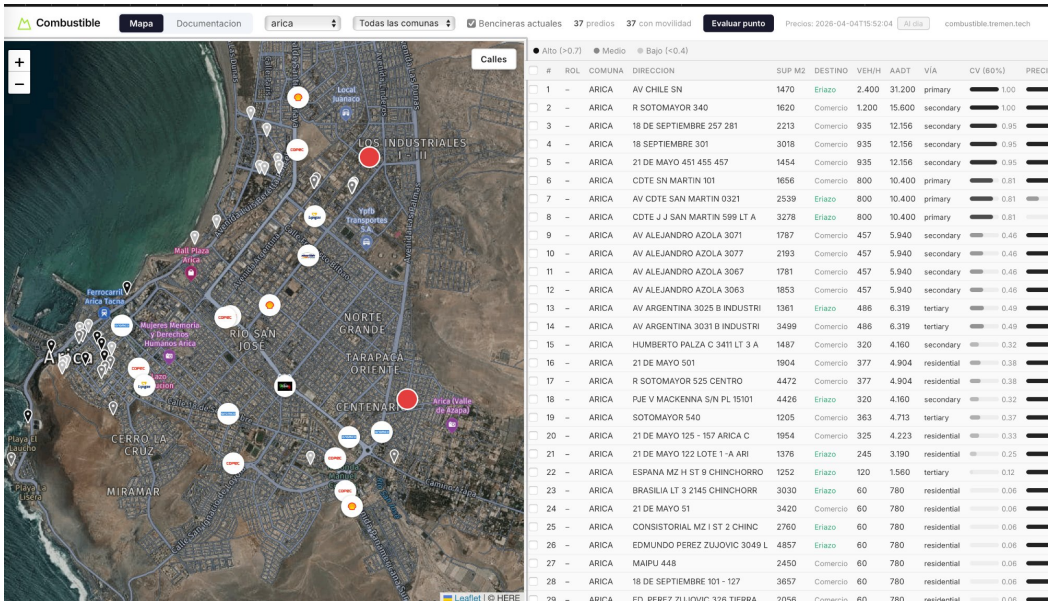

Figura 7.3: Dashboard de `combustible.tremen.tech`: mapa Leaflet con los predios candidatos para la comuna de Arica (37 predios, coloreados por score) y tabla interactiva con métricas de tráfico, superficie, destino y valor de suelo. Disponible en https://combustible.tremen.tech.

Evaluación de puntos arbitrarios

Además del análisis masivo de predios catastrados, la plataforma incorpora una funcionalidad de evaluación bajo demanda: el botón **Evaluar Punto** permite al usuario ingresar las coordenadas geográficas de cualquier ubicación (latitud y longitud) o cargar un archivo KML/KMZ proveniente de Google Earth o un GPS externo. El sistema ejecuta entonces el pipeline completo de 10 pasos sobre esa ubicación específica, devolviendo en tiempo real, mediante Server-Sent Events, el score compuesto, las métricas de tráfico BPR, el tipo de vía OSM, la densidad de dispositivos GPS y el precio de combustible en zona de influencia de 5 km.

Esta funcionalidad es especialmente relevante para puntos que no están en el catastro del SII: por ejemplo, un terreno en proceso de subdivisión, un predio con rol reciente que no ha sido procesado aún, o una ubicación hipotética que el cliente quiere evaluar antes de iniciar el proceso de compra. El análisis de evaluación individual tarda entre 15 y 45 segundos dependiendo de la latencia de las APIs externas (Google Routes, CNE) y produce el mismo informe exportable en Excel y KMZ que el análisis

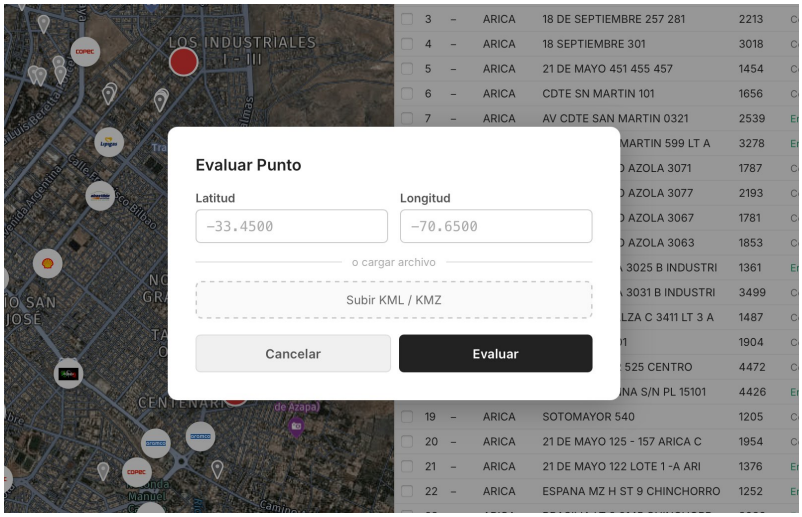

Figura 7.4: Modal de **Evaluar Punto** en `combustible.tremen.tech`: el usuario ingresa coordenadas o sube un archivo KML/KMZ y recibe el análisis completo del pipeline en tiempo real.

masivo. Esto convierte la plataforma en una herramienta de soporte a la toma de decisiones tanto para análisis exploratorios a escala nacional como para la validación de oportunidades específicas en negociación.

Conexión directa con los casos de uso de la tesis

La plataforma Combustible es la demostración empírica de los tres casos de uso descritos en las subsecciones anteriores: la prospección por destino y superficie (filtro inicial ERIAZO/COMERCIO, 1.200–5.000 mš) corresponde exactamente a la consulta SQL de la Sección 4.2.2; la identificación de suelo subutilizado (predios eriazos en zonas de alto tráfico) corresponde a la Sección 4.2.3; y el informe exportable en Excel y KMZ con scores cuantitativos es la forma concreta del due diligence con datos objetivos descrito en la Sección 4.2.4. Sin el pipeline de extracción catastral y el dataset predial nacional, esta plataforma simplemente no podría existir.

7.5 Impacto público y democratización de la información

7.5.1 Acceso antes y después de Catastral.cl

La Tabla 7.8 contrasta la situación de distintos actores antes de este trabajo y después de la disponibilidad de la plataforma Catastral.cl.

Tabla 7.8: Cambio en el acceso a datos prediales según tipo de actor

Actor	Antes	Después (Catastral.cl)
Broker inmobiliario	Bases privadas, costo elevado	Descargas históricas gratuitas + descarga enriquecida por comuna
Investigador académico	Sin datos a escala nacional	9,4M predios + 136M filas históricas consultables vía API
PyME / emprendedor	Sin acceso práctico	Dataset descargable en CSV/-Parquet vía pago social
Periodista de datos	Sin geometría catastral	Mapas verificables con polígonos de 30 cm de precisión
Municipio / planificador	WMS no analizable	Dataset SQL-consultable + descarga Parquet lista para QGIS
Ciudadano	Portal predio a predio (manual)	Búsqueda por dirección con mapa, historial de avalúos 7 años
Institución / universidad	Acceso ad hoc, sin escala	Domain grant institucional, acceso para todos los emails del dominio

El impacto es estructuralmente asimétrico: los actores más beneficiados son quienes tenían menor capacidad de pagar por datos privados. Un investigador de una universidad regional o un periodista de datos de un medio independiente pasan de no tener acceso a disponer del mismo dataset que una consultora inmobiliaria de primer nivel, con la diferencia de que ahora lo pueden consultar en un navegador sin escribir una línea de código (Open Knowledge Foundation, 2023; Biblioteca del Congreso Nacional de Chile, 2008).

7.5.2 Adopción real y ahorro cuantificado

Más allá de la comparación estructural, la operación de la plataforma en producción permite cuantificar tres dimensiones del impacto con datos reales.

Adopción medida. Al momento de esta evaluación, la plataforma registra **86.666 descargas** de capas de datos, realizadas a partir de **4.637 pagos sociales** (publicaciones en LinkedIn), lo que arroja un promedio de 18,7 descargas por cada pago social. Estas cifras convierten la afirmación de democratización en un hecho medible: 86.666 entregas de datos que, bajo cualquiera de los modelos de comercialización vigentes en el mercado (Sección 2.2.4), habrían tenido un costo monetario, y que en Catastral.cl tuvieron un costo pecuniario de cero. El modelo, además, no generó ingresos monetarios: su retorno es enteramente reputacional, coherente con el diseño de pago social.

Ahorro cuantificado. La Tabla 7.9 estima el costo de obtener el dataset predial nacional bajo cada modelo de mercado, frente al costo de obtenerlo mediante la plataforma. Los valores en UF se aproximan a USD con una paridad referencial de 1 UF ≈ USD 40.

Tabla 7.9: Costo de obtener el dataset predial nacional según vía de acceso

Vía de acceso	Base de cálculo	Costo (USD)
Datasets estructurados por comuna	346 comunas × USD 2.000–6.000	692.000–2.076.000
Servicio de scraping (precio de mercado)	(346/60) × USD 5.000	≈28.800
Consulta unitaria SaaS (1 UF / 50 consultas)	9,5M predios / 50 × 1 UF	≈7.600.000
Catastral.cl (esta tesis)	Pago social; ≈USD 150/mes de operación	**0 (usuario)**

El ahorro para el usuario final es total, y el costo de operación del sistema completo (≈USD 1.800 anuales) es entre dos y cuatro órdenes de magnitud inferior al costo de mercado de adquirir el mismo universo de datos, según el modelo de comparación. Esta magnitud de ahorro es consistente con las estimaciones de la literatura sobre el valor liberado por la apertura efectiva de datos públicos (McKinsey Global Institute, 2013; Zuiderwijk et al., 2019).

Impacto en el mercado inmobiliario. El valor del dato se materializa con nitidez en la prospección de localización, uno de los casos de uso del dataset (Sección 7.4.5). En la práctica del sector, el servicio de prospección de una ubicación se cobra en torno al 4% del valor total de la propiedad evaluada, y no se realiza para propiedades de valor inferior a unas 10.000 UF; el piso de mercado de una prospección se sitúa, por tanto, en torno a 400 UF (≈USD 16.000). El dataset producido en esta tesis reduce el insumo de datos de ese servicio, antes obtenido con semanas de trabajo en terreno, a una consulta de minutos sobre los 9,5 millones de predios del país, y hace económicamente viable prospectar a escala masiva un análisis que antes solo

se justificaba caso a caso sobre propiedades de alto valor.

7.5.3 Observaciones cualitativas sobre la recepción

A la evidencia cuantitativa anterior se suman observaciones cualitativas recogidas tras la publicación de la plataforma en redes profesionales. Estos comentarios espontáneos no constituyen una evaluación formal, pero complementan las métricas de adopción con los perfiles de la demanda.

Se identificaron tres categorías recurrentes en los comentarios recibidos. La primera agrupa a profesionales con necesidades técnicas concretas: desarrolladores de productos geoespaciales que reportaron la inestabilidad del WMS del SII y la ausencia de servicios WFS como barreras para integrar datos catastrales en flujos de trabajo SIG, y consultores que necesitaban los datos para construir herramientas de geocodificación y validación de direcciones.

La segunda categoría corresponde a actores del sector público e investigación que expresaron demanda por datos abiertos. Un caso notable es el de un investigador en planificación urbana que declaró haber recurrido al Consejo de la Transparencia, mecanismo legal de último recurso bajo la Ley 20.285 (Biblioteca del Congreso Nacional de Chile, 2008), como vía para obtener los polígonos catastrales que esta tesis extrae del WMS.

La tercera categoría agrupa a profesionales del sector privado interesados en aplicaciones comerciales del dataset, desde consultoras inmobiliarias hasta operadores de infraestructura.

Estas observaciones, aunque exploratorias, son consistentes con el diagnóstico de las tres barreras planteado en el Capítulo 1 y sugieren que la demanda abarca perfiles heterogéneos. Una evaluación rigurosa del impacto de la plataforma requeriría un estudio de usuarios con metodología formal (encuestas estructuradas, entrevistas en profundidad o análisis de logs de uso), lo cual queda fuera del alcance de esta tesis pero constituye una línea de trabajo futuro relevante.

7.5.4 Escalabilidad del modelo a América Latina

El método desarrollado no es exclusivo de Chile ni del SII. Varios países de América Latina publican sus catastros prediales en servicios WMS similares (OGC-compatible) con restricciones análogas: el dato existe, es técnicamente público, pero no está disponible en formato analítico. Perú (COFOPRI), Colombia (IGAC) y México

(INEGI) cuentan con servicios WMS de catastro que potencialmente son vectorizables con el mismo pipeline, adaptando los parámetros de capa, estilo y valor DN. La arquitectura de 30 túneles WireGuard con rotación automática de IP es igualmente aplicable a cualquier API que imponga *rate limiting* por dirección IP (Donenfeld, 2017).

La condición técnica necesaria para replicar el método es que el servidor GeoServer (o equivalente) use un valor de píxel distinto para el interior de los predios respecto al fondo del mapa. Esta condición se cumple en todos los servidores WMS evaluados como referencia, dado que es el comportamiento por defecto del estilo SLD de GeoServer al renderizar capas catastrales.

Capítulo 8

Discusión, Limitaciones y Conclusiones

Este capítulo discute los resultados del Capítulo 7 a la luz de la literatura internacional, presenta las limitaciones técnicas y metodológicas del trabajo y las amenazas a su validez, deriva las conclusiones y expone las líneas de investigación y desarrollo futuro.

8.1 Discusión

Esta sección sitúa los resultados obtenidos en el contexto de la literatura previa, en tres frentes: el valor económico de los datos abiertos, la asimetría de información en mercados inmobiliarios y los métodos automatizados de extracción de geometría catastral. A partir de ese contraste se identifican las fortalezas del enfoque y sus amenazas a la validez.

8.1.1 El resultado frente a la literatura de datos abiertos

La literatura sobre datos abiertos gubernamentales documenta de manera consistente una brecha entre la publicación formal del dato y la generación efectiva de valor. Zuiderwijk et al. (2019), sobre la base de 168 respuestas relativas a 156 iniciativas de datos abiertos en distintos niveles de gobierno, encuentran un desajuste sistemático entre los objetivos declarados de las políticas de apertura y los resultados efectivamente alcanzados, y atribuyen ese desajuste, entre otras causas, a barreras de usabilidad y de calidad del dato publicado. Este trabajo aporta evidencia em-

pírica concordante desde un ángulo distinto: el caso del catastro del SII muestra que la brecha no es solo de adopción por el lado de la demanda, sino que puede residir enteramente en el formato de publicación por el lado de la oferta. El dato catastral chileno cumple formalmente con la política de transparencia (Biblioteca del Congreso Nacional de Chile, 2008) y, sin embargo, era inutilizable para el análisis masivo por razones estrictamente técnicas. En términos de Attard et al. (2015) y Janssen et al. (2012), el caso ilustra que las barreras de *usabilidad*, *comprensibilidad* y *calidad* pueden ser tan excluyentes como la ausencia total de publicación, y que su superación es una condición previa a cualquier realización de valor.

Respecto de la magnitud de ese valor, las estimaciones macro de la literatura (World Bank, 2021; McKinsey Global Institute, 2013) sitúan el potencial de los datos abiertos en varios puntos del PIB. Este trabajo no valida ni refuta esas cifras agregadas, que exceden su alcance, pero sí aporta una estimación micro y verificable para un dato específico: el costo de la inaccesibilidad del catastro chileno documentado en la Sección 2.2.4. La contribución metodológica es mostrar que ese valor macro, habitualmente estimado de forma agregada y difícil de auditar, puede descomponerse y medirse dato por dato a partir de los costos de acceso que hoy enfrentan los actores del mercado.

8.1.2 El resultado frente a la teoría de asimetría de información

El marco teórico de esta tesis se ancla en la teoría de asimetría de información (Akerlof, 1970; Stiglitz, 2000), según la cual la distribución desigual de información entre las partes de una transacción degrada la eficiencia del mercado. La literatura empírica reciente confirma este mecanismo en mercados inmobiliarios concretos: Li and Chau (2024) muestran que, en el mercado secundario de Hong Kong, los compradores no locales (peor informados) pagan una prima del 2,8% sobre compradores locales por unidades con atributos observables equivalentes, lo que cuantifica el costo directo de la asimetría informacional para el actor desinformado. Este trabajo se ubica en el lado de la intervención sobre esa asimetría: la apertura del dato catastral es, en términos de la teoría, una reducción de la brecha informacional que la teoría predice debería mejorar la eficiencia del mercado. Es importante ser preciso sobre el alcance de esta afirmación, y aquí se conecta con la observación de que el trabajo propone y estima, pero no demuestra, el impacto: esta tesis construye la infraestructura que habilita la reducción de la asimetría y estima su valor potencial, pero no mide econométricamente el efecto sobre precios de transacción, lo que requeriría un

diseño cuasi-experimental posterior a la adopción de la plataforma. La contribución es hacer medible y accesible la información cuya asimetría la teoría identifica como fuente de ineficiencia; la cuantificación del efecto de equilibrio queda como pregunta empírica abierta.

8.1.3 El resultado frente a los métodos de extracción de geometría catastral

En el plano técnico, la literatura reciente sobre extracción automatizada de límites prediales se concentra en la aplicación de aprendizaje profundo sobre imágenes de alta resolución (UAV o satelitales), especialmente en países en desarrollo sin catastro digital. Xia et al. (2019), en un caso de estudio en Busogo (Ruanda) con redes totalmente convolucionales sobre imágenes UAV, reportan una precisión de 0,79 pero un *recall* de solo 0,37 y un F-score de 0,50 en la detección de límites catastrales visibles. Este contraste ilumina la naturaleza distinta del problema abordado en esta tesis y constituye su principal fortaleza técnica comparada: mientras el enfoque de aprendizaje profundo debe *inferir* dónde están los límites prediales a partir de rasgos visuales del terreno (muros, cercos, cambios de textura), con la incertidumbre que ello implica, el método de esta tesis *recupera* límites que el propio Estado ya ha delineado y publica en su servicio cartográfico, solo que en formato raster no descargable. El problema, por tanto, no es de detección sino de acceso, y esa diferencia explica por qué la cobertura alcanzable es sustancialmente mayor: no hay error de inferencia, sino fidelidad de extracción respecto de una fuente autoritativa. La implicancia metodológica es que, en jurisdicciones donde el Estado ya mantiene un catastro digital pero lo publica de forma opaca, la re-extracción desde el servicio oficial domina a la re-detección desde imágenes, tanto en cobertura como en validez del resultado.

8.1.4 Amenazas a la validez

La caracterización sistemática anterior permite explicitar las amenazas a la validez de las conclusiones, en las tres dimensiones habituales.

La **validez interna** se refiere a si el dataset representa fielmente la realidad catastral. La principal amenaza es que el producto hereda cualquier error de la fuente: si el SII registra mal un avalúo, una superficie o una geometría, el pipeline reproduce ese error sin corregirlo. El sistema de QA (Fase 7) verifica la consistencia interna

pero no valida contra una fuente de verdad externa, por lo que la corrección de los datos está condicionada a la corrección del SII. Esta amenaza se acota, no se elimina, observando que el SII es la fuente autoritativa del catastro fiscal en Chile: no existe un *ground truth* más autoritativo contra el cual contrastar.

La **validez externa** se refiere a hasta dónde generalizan los resultados. La amenaza es doble: dentro de Chile, el sesgo geográfico documentado en la Sección siguiente limita la representatividad en zonas rurales; fuera de Chile, la transferibilidad depende de las tres condiciones formuladas en la sección de aprendizajes (dato público, barrera técnica y no legal, demanda insatisfecha). Las conclusiones sobre factibilidad técnica generalizan a cualquier catastro con servicio cartográfico compatible con OGC (Open Geospatial Consortium, 2006), pero las conclusiones sobre impacto económico son específicas del mercado y la estructura de acceso chilenos.

La **validez de constructo** se refiere a si las métricas miden lo que dicen medir. La métrica de *cobertura* mide la fracción de predios con geometría extraída, no la exactitud de esa geometría, que se reporta por separado en la evaluación de precisión. La métrica de *precisión* de 30 cm describe la resolución del instrumento (el píxel del WMS), no un error validado contra deslindes legales. La distinción es relevante porque un lector podría interpretar la cobertura como garantía de exactitud, cuando ambas dimensiones son independientes y se reportan por separado.

8.2 Limitaciones

8.2.1 Sesgos del proceso de extracción

Cada decisión técnica del pipeline actúa como un filtro sobre el universo de predios, y lo que cada filtro deja fuera no se distribuye al azar en el territorio ni entre tipos de predio. Corresponde, por tanto, caracterizar los sesgos de manera sistemática:

- **Filtro DN=182 y predios con colores especiales.** Algunos predios aparecen en el WMS con colores distintos al estándar: los casos documentados incluyen predios en litigio, predios de organismos del Estado y predios con observación administrativa. Están presentes en el mapa visual pero no son capturados por el filtro de polygonización, quedando sin geometría. Su fracción es pequeña (estimada en menos del 1% del total), pero el sesgo tiene estructura: subrepresenta sistemáticamente el patrimonio fiscal y los predios en conflicto, lo que puede ser relevante en análisis de propiedad estatal o de

zonas con alta densidad de predios públicos.

- **Coordenada API desplazada en predios de gran superficie.** Como se documentó en la evaluación de precisión (Capítulo 7), en predios simples de gran superficie la coordenada del SII registra el punto de acceso o un vértice en lugar del centroide. Esto introduce un sesgo espacial que afecta principalmente los análisis de distancia y pertenencia en predios agrícolas y semi-rurales grandes.

- **Estrategia diferenciada Tier A / Tier B.** Las comunas con alta proporción agrícola o de predios sin coordenada se procesaron con una estrategia de descarga distinta, por lo que la calidad de cobertura no es homogénea entre comunas y está correlacionada con la ruralidad.

- **Unidades en copropiedad.** Las unidades dentro de edificios heredan el polígono del lote padre, de modo que los análisis por unidad individual deben usar los atributos tabulares (superficie construida, avalúo por rol) y no la geometría, que representa al terreno y no a la unidad.

- **Sesgo geográfico agregado.** El efecto conjunto de lo anterior es que el 39,7% de predios sin polígono en la versión documentada no se distribuye aleatoriamente: se concentra en comunas rurales y agrícolas donde el SII no ha realizado levantamiento cartográfico digital. Comunas como Guaitecas, Torres del Paine y San Juan de la Costa presentan más del 80% de sus predios sin polígono, mientras que el Gran Santiago supera el 97% de cobertura. Cualquier análisis espacial derivado de esta versión del dataset, incluyendo los casos de uso del Capítulo 7, tiene una representatividad significativamente menor en zonas rurales y de baja densidad poblacional, y sus conclusiones no deben extrapolarse a esas zonas sin considerar esta limitación.

Estos sesgos, reales en la versión del dataset que documenta esta tesis (proceso 2025-S2), demostraron ser abordables en iteraciones posteriores del método. Al cierre de esta versión (julio de 2026), el proceso 2026-S1 en producción alcanza una cobertura de coordenadas de 99,99% y de polígonos de 99,96% (9.573.069 predios), combinando tres mecanismos: herencia semestral de geometría entre períodos (90,9% de los casos), reasignación por *point-in-polygon* contra el período anterior (7,6%) y re-vectorización WMS dirigida del residuo (1,5%). El residuo duro se reduce a 4.097 roles (0,04%), concentrado en cuatro comunas sin capa WMS propia, territorios insulares y predios que el visor del SII no dibuja. La lección metodológica es que

el sesgo de cobertura no es estructural sino acumulativo: cada iteración semestral hereda lo ya resuelto y concentra el esfuerzo de extracción en el residuo, con lo que la brecha converge hacia los casos genuinamente ausentes de las fuentes públicas.

8.2.2 Dependencia del SII como riesgo estructural

El sistema depende de tres piezas que el SII controla unilateralmente, y conviene analizar por separado el radio de daño de un cambio en cada una. Si cambia la estructura del TXT semestral, se rompe el parser de ancho fijo: la adaptación es acotada (reconstruir el mapa de columnas) y el histórico ya descargado no se ve afectado. Si cambia el esquema de la API de consulta individual, se pierde el mecanismo de refresco de coordenadas para predios nuevos, pero los atributos tabulares y las geometrías existentes sobreviven. Si cambia el estilo cartográfico del WMS o el valor DN=182, se invalida el filtro de polygonización y es necesario re-descubrir el nuevo valor; el procedimiento de descubrimiento empírico está documentado en el Capítulo 6, por lo que el costo es de re-ejecución, no de re-diseño. Esta última dependencia es inherente al método de vectorización raster y no tiene solución definitiva sin acceso a los datos vectoriales originales del SII (GeoServer Project Steering Committee, 2024).

La evidencia histórica de estabilidad es favorable: el formato del rol semestral y los endpoints públicos del SII se han mantenido estables desde al menos 2018, y los 16 semestres procesados comparten la misma estructura de ancho fijo. El método de extracción que esta tesis documenta ha operado de forma continua durante aproximadamente cinco años sin requerir cambios de diseño: en todo ese período no se registró ningún cambio de formato, de estilo cartográfico ni de esquema de API que rompiera el pipeline.

El riesgo mayor no es técnico sino legal: que los datos dejen de ser públicos. Ese escenario requeriría una modificación normativa contraria a la Ley de Transparencia (Biblioteca del Congreso Nacional de Chile, 2008), lo que lo convierte en un cambio institucional mayor y de baja probabilidad. E incluso en ese escenario extremo, el cuerpo histórico ya extraído (16 semestres, 2018–2025, complementado con los procesos posteriores) constituye por sí solo una base de análisis con valor propio, sobre la cual es posible construir proxies y proyecciones aunque la fuente se cerrara hacia adelante.

8.2.3 Reproducibilidad y validez del método

Es necesario distinguir entre la replicabilidad del método y la reproducibilidad exacta del resultado. El método es replicable: los parámetros de tuning (el valor DN, los umbrales de *compactness*, las tolerancias del join espacial) están documentados en los Capítulos 5 y 6, y un investigador independiente puede ejecutar el pipeline completo. El resultado exacto, en cambio, no es reproducible: el WMS del SII renderiza únicamente el catastro vigente, sin versiones históricas de la geometría, por lo que cada ejecución produce la fotografía del semestre en curso y no la misma fotografía que documenta esta tesis. El dataset debe entenderse como una serie de instantáneas semestrales de un instrumento que la fuente controla, no como un experimento repetible en sentido estricto.

Las condiciones mínimas de validez del método son dos, y ambas son verificables en cualquier momento: que el SII publique coordenadas de referencia por predio, y que exista cartografía raster de los predios en su visor público. Mientras ambas se cumplan, el método es aplicable; los cambios de formato, estilo o esquema alteran parámetros de configuración, no el diseño del pipeline.

Por último, el sistema de QA implementado (Fase 7) verifica la consistencia interna del pipeline (integridad de filas, validez topológica de polígonos, contención dentro de los bounds geográficos de Chile), pero no valida contra una fuente de verdad externa (*ground truth*). La cifra de cobertura describe el resultado del proceso, no garantiza la corrección geométrica de cada polígono. Una validación rigurosa requeriría comparar contra geometrías oficiales del Conservador de Bienes Raíces o del MINVU, que no están disponibles en formato vectorial abierto. La inspección visual selectiva sobre el WMS muestra que los polígonos vectorizados coinciden con los bordes renderizados, lo cual es esperable dado que el pipeline extrae exactamente la información visual del servicio; la fuente de error geométrico principal sigue siendo la resolución de 30 cm/píxel, que introduce una incertidumbre de ± 1 píxel en los bordes.

8.2.4 Precisión geométrica de 30 centímetros

La resolución de $\approx$30 cm/px en zoom 19 implica que los polígonos vectorizados tienen una precisión máxima de 1 píxel, equivalente a 30 cm en el terreno. Para análisis a escala de predio esto es más que suficiente: el error es menor que el espesor de un muro. Sin embargo, para aplicaciones que requieren precisión catastral legal (deslindes, subdivisiones, expropiaciones), los polígonos de este dataset no

reemplazan a los planos oficiales del Conservador de Bienes Raíces (CBR). Son una aproximación de alta calidad útil para análisis, no un sustituto del instrumento legal.

8.3 Conclusiones

Esta tesis abordó el problema de la inaccesibilidad práctica de los datos catastrales del SII chileno mediante el diseño e implementación de un pipeline de extracción y una plataforma de distribución. Los resultados cuantitativos obtenidos son: 9,55 millones de predios con atributos tabulares completos, 5,67 millones con polígono vectorial de ≈30 cm de precisión (60,3% de cobertura nacional), 16 semestres de series históricas de avalúo (136 millones de filas), y 343 comunas disponibles para descarga en formatos analíticos a través de la plataforma Catastral.cl. A partir de estos resultados se derivan las siguientes conclusiones.

Primera conclusión: las barreras de acceso son técnicas, no legales, y son superables mediante ingeniería de datos. La hipótesis planteada en el Capítulo 1 queda confirmada. La Ley de Transparencia (Biblioteca del Congreso Nacional de Chile, 2008) garantiza formalmente el acceso a los datos catastrales, pero la barrera efectiva reside en los formatos de publicación: texto de ancho fijo sin estructura, API unitaria sin endpoint masivo, y geometría disponible únicamente en formato raster. Superar estas barreras requirió combinar parseo de formato propietario, extracción distribuida con 30 túneles WireGuard y vectorización raster con GDAL. Este resultado indica que la inaccesibilidad del dato catastral chileno es un problema de ingeniería, no de política pública, y que la solución es técnicamente reproducible.

Segunda conclusión: la distribución del dato es condición necesaria para su impacto. La existencia de un dataset almacenado en un sistema de archivos no constituye, por sí sola, democratización del acceso. La plataforma Catastral.cl cumple la función de transformar el resultado técnico del pipeline en un servicio consultable sin requerir conocimientos de programación ni software especializado. De forma complementaria, la plataforma Combustible (`combustible.tremen.tech`) demuestra que el mismo dataset puede funcionar como capa base de aplicaciones analíticas especializadas, en este caso un sistema de prospección de ubicaciones para estaciones de servicio que reduce el tiempo de análisis de semanas a minutos.

Tercera conclusión: la metodología es potencialmente transferible a otros catastros de la región. Varios países de América Latina publican catastros prediales mediante servicios WMS compatibles con OGC, entre ellos Perú (COFOPRI),

Colombia (IGAC) y México (INEGI) (Open Geospatial Consortium, 2006). La arquitectura técnica desarrollada (túneles WireGuard, polygonización GDAL, spatial join con GeoPandas, API FastAPI, base de datos PostGIS) es independiente del SII chileno y aplicable a cualquier servidor WMS que utilice valores de píxel diferenciados por tipo de elemento cartográfico. La validación de esta transferibilidad requeriría estudios específicos por país, considerando las particularidades de cada infraestructura catastral.

Cuarta conclusión: el trabajo constituye un argumento empírico a favor de la apertura de datos catastrales. Si el Estado chileno publicara los datos catastrales directamente en formato vectorial abierto (GeoJSON, GeoPackage o Parquet georreferenciado), el esfuerzo técnico documentado en esta tesis resultaría innecesario. El costo marginal de publicar un archivo con los polígonos prediales es bajo para una institución que ya dispone de la información en formato vectorial internamente. El beneficio potencial para investigadores, planificadores urbanos, pequeñas empresas y ciudadanos sería significativo (Janssen et al., 2012; Attard et al., 2015). En este sentido, el pipeline de extracción catastral y la plataforma Catastral.cl pueden interpretarse como evidencia de que existe demanda no satisfecha por datos abiertos catastrales, y como un argumento técnico a favor de que el SII actualice sus estándares de publicación (Biblioteca del Congreso Nacional de Chile, 2008; Open Knowledge Foundation, 2023).

8.4 Aprendizajes para la analítica de negocios sobre datos públicos

Más allá de los resultados específicos del caso catastral, este trabajo deja lecciones generalizables para futuros desarrollos de analítica de negocios basada en datos públicos. Se formulan aquí como aprendizajes transferibles, junto con las condiciones bajo las cuales la metodología puede aplicarse a dominios distintos del mercado inmobiliario.

Primer aprendizaje: la brecha entre publicidad formal y usabilidad analítica es un espacio de creación de valor sistemáticamente subexplotado. Los marcos legales de transparencia garantizan el acceso al dato, pero no su formato. El resultado es una categoría de activos públicos que nadie explota: demasiado técnicos para el usuario general, demasiado laboriosos para el analista individual, y sin incentivo comercial para el intermediario que ya lucra con la opacidad. El

primer paso metodológico ante un dato público candidato es una auditoría de barreras en tres dimensiones: el formato de publicación (£es legible por máquina?), el mecanismo de consulta (£admite extracción masiva?) y la representación de la información (£los datos espaciales o estructurados están disponibles como tales, o solo como imágenes y documentos?). Esa auditoría, que en esta tesis corresponde al diagnóstico de las tres barreras del Capítulo 2, define el costo de ingeniería del proyecto antes de escribir una línea de código.

Segundo aprendizaje: el costo de superar las barreras se paga una sola vez y el activo resultante habilita múltiples verticales de negocio. El pipeline de extracción catastral se construyó una vez, pero sobre el dataset resultante operan ya tres aplicaciones de naturaleza distinta: la plataforma de distribución Catastral.cl, el sistema de inteligencia de localización para estaciones de servicio (`combustible.tremen.tech`) y la plataforma de analítica tributaria Valori. Esto cambia la economía de la decisión: el esfuerzo de ingeniería de datos no debe evaluarse como costo de un análisis particular, sino como inversión en un activo reutilizable cuyo valor crece con cada vertical que se construye sobre él. Para la analítica de negocios, la implicancia es que la pregunta correcta no es "£justifica este análisis el costo de extracción?" sino "£cuántos análisis distintos habilita el activo una vez construido?".

Tercer aprendizaje: la distribución del dato es tan determinante como su extracción, y la barrera monetaria no es la única forma de sostenerla. Un dataset sin interfaz de acceso no genera impacto, como se estableció en la segunda conclusión. El modelo de pago social de Catastral.cl agrega una lección adicional: es posible distribuir un activo de datos sin barrera monetaria financiando la difusión con reconocimiento, de manera análoga a la atribución en el software libre. Para emprendimientos de datos en etapa temprana, el reconocimiento social de los usuarios puede ser más valioso que sus pagos: cada mención convierte a un usuario en canal de adquisición de los siguientes.

Cuarto aprendizaje: la transferibilidad de la metodología depende de tres condiciones verificables. La metodología completa (auditoría de barreras, extracción masiva, estructuración, distribución con modelo de acceso) es aplicable a un dominio distinto cuando se cumplen tres condiciones: (i) el dato es formalmente público y no contiene datos personales protegidos, de modo que la barrera sea efectivamente técnica y no legal; (ii) las barreras identificadas son de formato, consulta o representación, superables con métodos estándar de ingeniería de datos; y (iii) existe demanda insatisfecha por el dato estructurado, con disposición a pagar

en dinero o en reconocimiento. En Chile cumplen estas condiciones, entre otros, los registros de compras públicas de Mercado Público, las publicaciones del Diario Oficial, los datos operacionales de transporte público y las estadísticas agregadas de salud. En el plano regional, la tercera conclusión identificó los catastros de Perú, Colombia y México como candidatos directos dentro del mismo dominio. Dado un dato público candidato, estas tres condiciones funcionan como criterio de decisión para evaluar si el patrón desarrollado en esta tesis aplica.

8.5 Trabajo futuro

8.5.1 Validación muestral con datos catastrales de referencia

Una línea de trabajo futuro relevante sería realizar una validación muestral sobre 5–10 comunas donde existan datos catastrales vectoriales de referencia (por ejemplo, municipalidades que hayan digitalizado sus planos reguladores), cuantificando métricas como el IoU (*Intersection over Union*) entre los polígonos del pipeline y los polígonos de referencia. Asimismo, un estudio futuro podría cuantificar la correlación entre la cobertura del pipeline y variables socioeconómicas comunales para caracterizar formalmente el sesgo de selección descrito en la sección anterior.

8.5.2 Sat-Catastral: validación del uso de suelo mediante series de tiempo satelitales

Una limitación del dataset actual es que refleja la clasificación de uso de suelo declarada por el SII, sin verificación independiente de si dicha clasificación corresponde al uso real del predio. Esta brecha es especialmente relevante en predios agrícolas, donde la clasificación SII determina la carga tributaria y donde la discrepancia entre el uso declarado y el uso efectivo puede ser significativa.

Una línea de trabajo futuro es **Sat-Catastral**: un sistema que cruza cada predio del dataset con series de tiempo de imágenes satelitales de alta resolución, aplicando modelos de visión computacional para determinar si el uso real del suelo coincide con el destino predial registrado. El enfoque propuesto utiliza **GeoSAM** (*Geospatial Segment Anything Model*), implementado como plugin de QGIS (https://plugins.qgis.org/plugins/GeoOSAM/), para segmentar automáticamente los

objetos presentes en cada predio (cultivos, construcciones, suelo desnudo, cuerpos de agua) y comparar los resultados con la descripción de destino del SII. La hipótesis es que una fracción no despreciable de los predios clasificados como agrícolas presenta en imagen satelital patrones inconsistentes con uso agropecuario activo (por ejemplo, suelo urbanizado, instalaciones industriales o abandono), lo que podría indicar errores de clasificación con implicancias tributarias.

En cuanto a las fuentes de imágenes, además de los programas públicos de observación terrestre (Sentinel-2, Landsat), una integración con proveedores comerciales como **Planet** (https://www.planet.com), cuya constelación de satélites propios captura imágenes diarias de la superficie completa del planeta con resoluciones de 3 a 5 metros, permitiría elevar la frecuencia de monitoreo de mensual a diaria. Sobre el dataset predial esto habilitaría detección temprana de cambios de uso de suelo, construcciones no declaradas y actividad en predios clasificados como eriazos, transformando el catastro estático del SII en una capa de inteligencia territorial con actualización continua.

8.5.3 Valori: impugnación de avalúos ante el SII mediante analítica catastral

El dataset predial nacional que esta AFE produce es también la base de una aplicación comercial de analítica tributaria. La plataforma **Valori** (https://valori.tremen.tech) utiliza los atributos de avalúo fiscal, superficie, destino predial y valor comercial de suelo por área homogénea para identificar predios que pagan contribuciones de bienes raíces en exceso respecto de lo que correspondería según los parámetros del propio SII.

El mecanismo es el siguiente: cuando el avalúo fiscal de un predio supera el valor que resulta de aplicar los coeficientes oficiales del SII a sus características físicas (superficie, destino, área homogénea), el propietario puede impugnar el avalúo ante el organismo y solicitar su rectificación. Si la impugnación prospera, el SII debe devolver las contribuciones pagadas en exceso con efecto retroactivo de hasta tres años. Valori automatiza la detección de estos casos sobre el universo completo de predios del dataset, priorizando aquellos con mayor probabilidad de impugnación exitosa y mayor monto de devolución potencial. Este caso de uso ilustra de forma concreta el impacto económico directo que tiene la disponibilidad del dato catastral en formato analítico: sin el dataset producido por esta AFE, el análisis sería inviable a escala nacional.

8.5.4 Integración futura con el CBR

El Conservador de Bienes Raíces publica escrituras, inscripciones y planos de subdivisión en formato PDF. La integración de esta fuente (extracción de planos técnicos + OCR de datos de inscripción) con el dataset del SII permitiría construir un grafo de propiedad que vincule el predio físico con el historial de transacciones, la cadena de dominio y el propietario actual. Este es el paso natural siguiente para un sistema de inteligencia inmobiliaria completo basado exclusivamente en fuentes públicas chilenas.

Apéndice A

Detalles de implementación

Este anexo reúne el detalle técnico de programación e infraestructura del pipeline descrito en el Capítulo 6. Se traslada aquí para no interrumpir la lectura conceptual del cuerpo principal, y se organiza en el mismo orden que las fases del pipeline.

A.1 Estructura del archivo TXT de ancho fijo

El archivo `BRTMPNACROL_NAC_2025_2.txt` tiene un registro de 117 caracteres por predio, con campos en posiciones fijas (sin delimitadores). La Tabla A.1 muestra la estructura completa de los 15 campos definidos en el manual oficial del SII.

El parseo se realiza por corte de posiciones sobre cada línea:

```python
def parse_line(line: str) -> dict:
    """Extrae campos de una linea del archivo BRTMPNACROL."""
    return {
        "comuna":  line[0:5].strip(),
        "anio":    line[5:9].strip(),
        "semestre":line[9].strip(),
        "manzana": line[57:62].zfill(5),
        "predio":  line[62:67].zfill(5),
        "txt_avaluo_total": line[81:96].strip(),
        "txt_cod_destino":  line[116].strip(),
    }
```

Listing A.1: Parseo de línea TXT de ancho fijo

Tabla A.1: Estructura completa del archivo BRTMPNACROL según manual oficial del SII (Servicio de Impuestos Internos, 2025b)

Nř	Campo	Pos.	Ancho	Descripción
1	`comuna`	1–5	5	Código CONARA de la comuna
2	`anio`	6–9	4	Año del proceso de rol de cobro
3	`semestre`	10	1	Semestre (1 o 2)
4	`ind_aseo`	11	1	"A" si la cuota incluye tarifa de aseo
5	*(espacios)*	12–17	6	Campo sin información
6	`txt_direccion`	18–57	40	Dirección predial
7	`manzana`	58–62	5	Número de manzana
8	`predio`	63–67	5	Número de predio dentro de la manzana
9	`cod_serie`	68	1	Serie: `A`=Agrícola, `N`=No agrícola
10	`cuota_trimestral`	69–81	13	Contribución neta trimestral (CLP)
11	`txt_avaluo_total`	82–96	15	Avalúo fiscal total (CLP)
12	`txt_avaluo_exento`	97–111	15	Avalúo exento del impuesto territorial (CLP)
13	`anio_exencion`	112–115	4	Año término exención (2055 = indefinida)
14	`cod_ubicacion`	116	1	Ubicación: `R`=Rural, `U`=Urbana
15	`txt_cod_destino`	117	1	Código de destino predial (ver Tabla 6.2)

A.2 Estructura de la solicitud a la API `getPredioNacional`

El endpoint `getPredioNacional` (POST) devuelve la información completa de un predio dado su código de comuna, manzana y número de predio.

La respuesta JSON incluye campos de identificación del predio (`existePredio`, `nombreComuna`, `direccion`), coordenadas geográficas (`ubicacionX` y `ubicacionY`, que el SII designa con los ejes invertidos: `X` contiene latitud y `Y` contiene longitud), avalúos fiscales (`valorTotal`, `valorAfecto`, `valorExento`), superficies (`supTerreno`, `supConsMt2`), y datos de Área Homogénea con valores comerciales de suelo en CLP/mš.

A.3 Infraestructura de 30 túneles WireGuard

Cada uno de los 30 túneles WireGuard (proveedor Mullvad) se monta como un *network namespace* independiente de Linux (Donenfeld, 2017), de modo que cada proceso de scraping tenga su propia pila de red aislada con IP pública diferente.

```
1 # Crear namespace para el tunel i
```

```
payload = {
    "rolBusqueda": {
        "numero_manzana": "00250",
        "numero_predio":  "00001",
        "codigo_comuna":  "13101"
    },
    "servicios": [
        {"nombre": "sii:BR_CART_AH_MUESTRAS",
         "style":  "AH_MUESTRA_EAC_15_2025",
         "eac": 15, "eacano": 2025}
    ]
}
response = requests.post(URL_API, json=payload, timeout=8)
```

Listing A.2: Estructura de request a la API getPredioNacional

```
ip netns add vpn${i}
# Levantar interfaz WireGuard dentro del namespace
ip netns exec vpn${i} wg-quick up wg${i}
# Ejecutar scraper dentro del namespace
ip netns exec vpn${i} python3 0_get_sii.py \
    --comuna 13101 --workers 3 --rps 10 \
    --chunk ${i} --total-chunks 30
```

Listing A.3: Creación de network namespace y activación de túnel

Cada túnel opera 3 *workers* paralelos con una tasa de 10 requests/segundo, lo que entrega un caudal sostenido de aproximadamente 900 req/s en el pico. Cuando un túnel detecta que la tasa de éxito cae (señal de bloqueo de esa IP), el orquestador (`batch_30ns.sh`) descarga un nuevo relay del JSON de Mullvad y reconfigura el túnel en el lugar, sin interrumpir los demás (*rotación automática de IP*). La Figura A.1 ilustra la arquitectura.

```
[START] 13101 (272449 predios, ~9081 per chunk) - 14:30:00
  14:30:15 [30 alive] ok=1200 fail=2 (80.0 r/s, 15s)
  [BURNED] T14 (ok stuck at 3421)
  [ROTATE]  T14 -> jp-tyo
[DONE]  13101 (7812s, 272449 rows, 34.8 r/s, 23 rot)
```

Listing A.4: Fragmento del log de la Fase 0

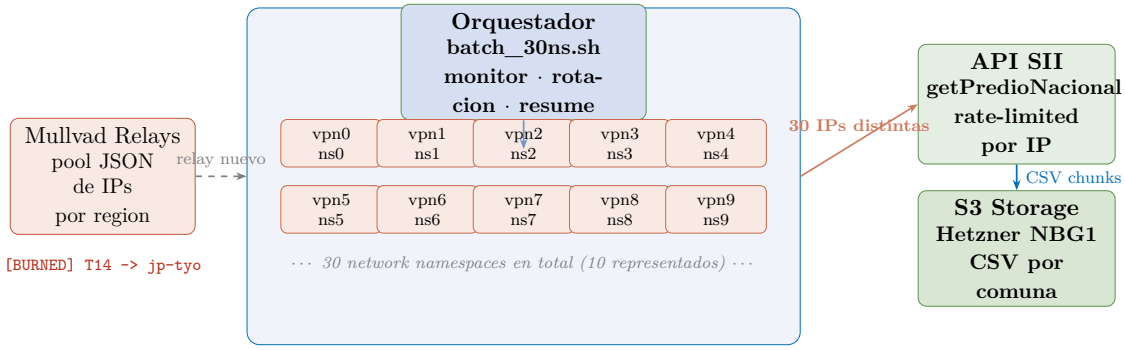

Figura A.1: Arquitectura de extracción distribuida con 30 túneles WireGuard. Cada túnel opera como un *network namespace* independiente con IP pública propia. El orquestador monitorea la tasa de éxito cada 15s y rota IPs quemadas desde el pool de Mullvad.

A.4 Clasificación de comunas en Tier A y Tier B

La Figura A.2 detalla el árbol de decisión que separa las comunas en dos estrategias de descarga WMS según su proporción de predios agrícolas o sin coordenada, con el umbral de corte del 30% y ejemplos representativos de cada tier.

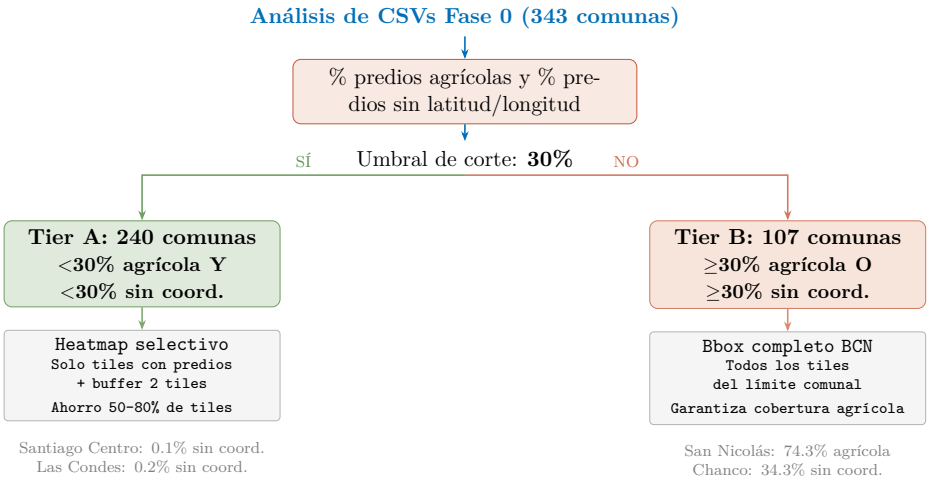

Figura A.2: Clasificación de comunas en Tier A y Tier B según proporción de predios sin coordenadas o agrícolas. Cada tier aplica una estrategia de descarga WMS distinta para garantizar cobertura sin descargar tiles innecesarios.

A.5 Solicitud de tiles WMS y polygonización con GDAL

Cada tile de 256×256 píxeles se solicita al servicio WMS del SII en formato PNG a zoom 19 (Web Mercator, EPSG:3857):

```
https://www4.sii.cl/mapasui/services/ui/wmsProxyService/call
  ?layers=sii:BR_CART_SANTIAGO_CENTRO_WMS
  &styles=PREDIOS_WMS_V0
  &srs=EPSG:3857
  &width=256&height=256
  &format=image/png
  &bbox=-7913137.9,{min_y},-7912625.9,{max_y}
```

Listing A.5: URL de solicitud de un tile WMS al SII

La vectorización de los GeoTIFFs ensamblados se realiza con `gdal_polygonize` sobre el canal rojo (Band 1), donde los píxeles interiores de cada predio tienen valor DN=182 en el estilo `PREDIOS_WMS_V0`:

```
gdal_polygonize.py input.tif -b 1 -f GPKG output.gpkg predios
    DN
```

Listing A.6: Polygonización del GeoTIFF usando GDAL Band 1

El resultado crudo incluye todos los polígonos del raster. El filtrado por valor DN=182 reduce el resultado a los predios reales:

```
ogr2ogr -f GPKG predios.gpkg crudo.gpkg \
  -sql "SELECT geom, DN FROM predios WHERE DN = 182" -nln
     predios
```

Listing A.7: Filtrado de polígonos con DN=182

En Santiago Centro, este filtrado reduce de 3.379.736 polígonos (763 MB) a 21.050 polígonos (29 MB), eliminando el 99.4% de los polígonos y el 96% del volumen. El proceso completo con sus tasas de reducción se ilustra en la Figura 6.2 del cuerpo principal.

A.6 Arquitectura de dos dominios de la plataforma

La Figura A.3 detalla la separación física entre el VPS extractor (que ejecuta el pipeline y almacena en S3) y el VPS de plataforma (que consume vía ETL semestral y sirve a través de Nginx, FastAPI, PostgreSQL/PostGIS y una SPA React).

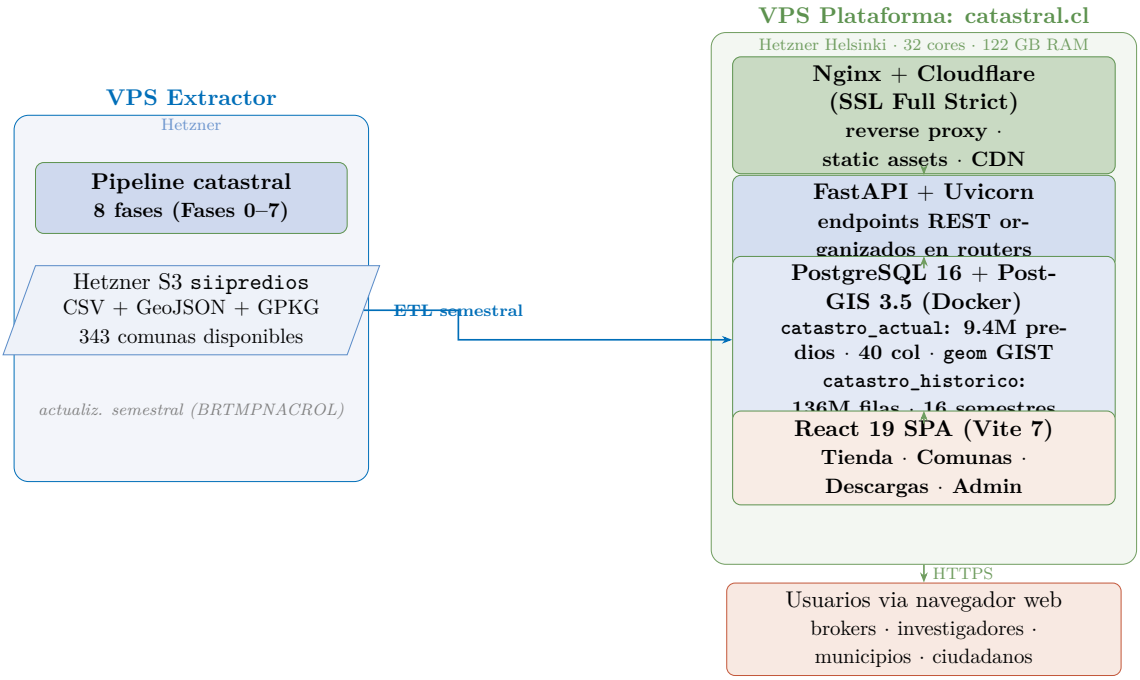

Figura A.3: Arquitectura de dos dominios de la plataforma Catastral.cl. El VPS extractor (izquierda) ejecuta el pipeline de extracción catastral y almacena los productos en S3. El VPS de plataforma (derecha) los consume vía un pipeline ETL semestral y los sirve a través de Nginx, FastAPI, PostgreSQL/PostGIS y una SPA React bajo el dominio `catastral.cl`.

Referencias

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500.

Attard, J., Orlandi, F., Scerri, S., and Auer, S. (2015). A systematic review of open government data initiatives. *Government Information Quarterly*, 32(4):399–418.

Biblioteca del Congreso Nacional de Chile (2008). Ley nº 20.285: Sobre acceso a la información pública. https://www.bcn.cl/leychile/Navegar?idNorma=276363. Publicada en el Diario Oficial el 20 de agosto de 2008.

Broxterman, D. and Zhou, T. (2023). Information frictions in real estate markets: Recent evidence and issues. *Journal of Real Estate Finance and Economics*, 66(2):203–298.

CEPAL (2022). Datos y hechos sobre la transformación digital: informe sobre los principales indicadores de adopción de tecnologías digitales. Comisión Económica para América Latina y el Caribe. https://www.cepal.org/es/publicaciones/46766-datos-hechos-la-transformacion-digital-informe-principales-indicadores-adopcion.

Cetl, V., Šamanović, S., Bjelotomić Oršulić, O., and Lisec, A. (2023). Building a cadastral map of Europe through the INSPIRE and other related initiatives. *Land*, 12(7):1462.

Deininger, K. and Feder, G. (2009). Land registration, governance, and development: Evidence and implications for policy. *The World Bank Research Observer*, 24(2):233–266.

DiPasquale, D. and Wheaton, W. C. (1996). *Urban Economics and Real Estate Markets*. Prentice Hall, Englewood Cliffs, NJ.

Donenfeld, J. A. (2017). WireGuard: Next generation kernel network tunnel. In

Proceedings of the Network and Distributed System Security Symposium (NDSS). San Diego, CA. Disponible en wireguard.com/papers/wireguard.pdf.

European Data Portal (2020). The economic impact of open data: Opportunities for value creation in Europe. https://data.europa.eu/en/doc/economic-impact-open-data-opportunities-value-creation-europe. Comisión Europea. Accedido: 2026-03-15.

GDAL/OGR contributors (2024). GDAL: Geospatial data abstraction library. OSGeo. https://gdal.org. Accedido: 2026-01-10.

GeoServer Project Steering Committee (2024). GeoServer: Open source server for sharing geospatial data. https://geoserver.org. Accedido: 2026-01-10.

Gillies, S. et al. (2024a). Rasterio: Geospatial raster I/O for Python programmers. https://rasterio.readthedocs.io. Versión 1.3. Accedido: 2026-01-10.

Gillies, S. et al. (2024b). Shapely: Manipulation and analysis of geometric objects. PyPI / GitHub. https://shapely.readthedocs.io. Accedido: 2026-01-10.

Janssen, M., Charalabidis, Y., and Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, 29(4):258–268.

Jordahl, K., Van den Bossche, J., Fleischmann, M., Wasserman, J., McBride, J., Gerard, J., et al. (2024). GeoPandas v0.14.3. Zenodo. https://zenodo.org/records/10601680. DOI: 10.5281/zenodo.10601680.

Li, L. and Chau, K. W. (2024). Information asymmetry with heterogeneous buyers and sellers in the housing market. *Journal of Real Estate Finance and Economics*, 68(1):138–159.

Linux man-pages project (2024). network_namespaces(7) Linux manual page. https://man7.org/linux/man-pages/man7/network_namespaces.7.html. Documentación oficial del kernel Linux. Accedido: 2026-01-10.

McKinsey Global Institute (2013). Open data: Unlocking innovation and performance with liquid information. https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information. Estimación: USD 3–5 billones anuales en valor potencial global.

Nasser, A. and Concha, G. (2013). Datos abiertos: un nuevo desafío para los gobiernos de la región. Comisión Económica para América Latina y el Caribe (CEPAL). https://repositorio.cepal.org/handle/11362/7331. Série Políticas Sociales, Nř 183.

Open Geospatial Consortium (2006). OpenGIS® web map service (WMS) implementation specification, version 1.3.0. https://www.ogc.org/standards/wms/. OGC Document 06-042.

Open Knowledge Foundation (2023). The open data handbook. https://opendatahandbook.org/. Accedido: 2026-01-10.

Redman, T. C. (2008). *Data Driven: Profiting from Your Most Important Business Asset*. Harvard Business School Press, Boston, MA.

Servicio de Impuestos Internos (2025a). Estructura de archivo para detalle catastral de bienes raíces. Documento técnico oficial. Disponible en https://www4.sii.cl/sismunInternet6/?caller=DETALLE_CAT_Y_ROL_COBRO (requiere autenticación con RUT y clave SII). Accedido: 2026-03-01.

Servicio de Impuestos Internos (2025b). Estructura de archivo para rol semestral de contribuciones de bienes raíces. Documento técnico oficial. Disponible en https://www4.sii.cl/sismunInternet6/?caller=DETALLE_CAT_Y_ROL_COBRO (requiere autenticación con RUT y clave SII). Accedido: 2026-03-01.

Stiglitz, J. E. (2000). The contributions of the economics of information to twentieth century economics. *The Quarterly Journal of Economics*, 115(4):1441–1478.

Williamson, C. R. and Kerekes, C. B. (2011). Securing private property: Formal versus informal institutions. *The Journal of Law and Economics*, 54(3):537–572.

World Bank (2021). World development report 2021: Data for better lives. https://www.worldbank.org/en/publication/wdr2021. Washington, DC: World Bank. DOI: 10.1596/978-1-4648-1600-0.

Xia, X., Persello, C., and Koeva, M. (2019). Deep fully convolutional networks for cadastral boundary detection from UAV images. *Remote Sensing*, 11(14):1725.

Zuiderwijk, A. and Janssen, M. (2014). Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly*, 31(1):17–29.

Zuiderwijk, A., Shinde, R., and Janssen, M. (2019). Investigating the attainment of open government data objectives: Is there a mismatch between objectives and results? *International Review of Administrative Sciences*, 85(4):645–672.