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

MÉTODOS EFICACES DE BÚSQUEDA Y CORRECCIÓN DE DIRECCIONES DE PACIENTES COVID-19

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

MATÍAS NICOLÁS MARTÍNEZ PÉREZ

PROFESOR GUÍA:
FERNANDO ORDOÑEZ PIZARRO

MIEMBROS DE LA COMISIÓN:
RICHARD WEBER HAAS
FRANCISCO SUÁREZ SALAS

Este trabajo ha sido financiado por:
Proyecto COVID 0251 de la Agencia Nacional de Investigación y Desarrollo

SANTIAGO DE CHILE
2022

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL
POR: MATÍAS NICOLÁS MARTÍNEZ PÉREZ
FECHA: 2022
PROF. GUÍA: FERNANDO ORDOÑEZ PIZARRO

MÉTODOS EFICACES DE BÚSQUEDA Y CORRECCIÓN DE DIRECCIONES DE PACIENTES COVID-19

Las pandemias han sido eventos que han ocurrido a lo largo de la historia de la humanidad, unas más letales que otras, además de ir apareciendo en escenarios mundiales bastante diversos e incluso diametralmente diferentes. Actualmente, el mundo está viviendo la pandemia COVID-19, "una enfermedad causada por el nuevo coronavirus conocido como SARS-CoV-2" [1]. Al día 17 de julio de 2021, la cantidad total de casos en el país superó la cifra 1.5 millones y la cantidad de fallecidos superó la cifra de 30 mil [2].

En variadas situaciones tales como la investigación de epidemias, el mapeo de crímenes o sistemas de análisis, en los sistemas de manejo de desastre y riesgo o incluso en la ciencia política [3] es una parte esencial del trabajo lograr tener un seguimiento geográfico y visión espacial de lo que está sucediendo, por lo tanto es crucial lograr la localización de los sujetos de interés y por lo tanto que las direcciones ser geocodificadas, es decir, poder ser convertidas en una posición en coordenadas (latitud, longitud). El problema es que esto no siempre es posible debido a que la dirección con la que contemos del sujeto o sujetos podría no ser intepretable por los servicios de geocodificación a raíz de múltiples causas tales como: la dirección está mal escrita, la captación de la dirección desde las bases de datos no se logró de manera exitosa, la dirección no existe, la dirección está incompleta, la dirección no coincide con la forma en la cual aparece en los registros de los servicios de geocodificación, la dirección tiene información de más o de menos, etc.

Una forma de solucionar esto aplicando un método de corrección y normalización de las direcciones, por lo que se desea desarrollar un método de normalización de direcciones que permita geocodificar de manera eficiente y eficaz las direcciones de los pacientes covid del SSMSO. Se desea además desarrollar herramientas de visualización y realizar un análisis con variables sociodemográficas que permita entender la propagación del virus en el país.

Para lograr esto, se aplicaron una serie de procesos que principalmente consisten en: Normalización de direcciones, geocodificación de las direcciones normalizadas, asignación de manzana-entidad a paciente, asociar variables sociodemográficas y finalmente visualización espacial y temporal de los casos. Todos estos procesos pudieron ser llevados a cabo con buenos resultados. La normalización mejoró ampliamente el desempeño en la búsqueda de direcciones. Con un proceso de geocodificación confiable y bien logrado, fue posible utilizar esos resultados con mayor seguridad para los análisis y procesos siguientes, ya que la geocodificación es la base de la confiabilidad de los resultados que dependen de esta información.

A mis padres por su amor
Y a las lamentables víctimas de esta pandemia.

En la memoria

Agradecimientos

Agradezco a mis padres por darme la educación y apoyarme. A la Universidad de Chile por ser la cuna de mis conocimientos. A mis compañeros y profesores, sobre todo al profesor que me acompañó y guió en la realización de esta memoria, Fernando Ordoñez. A ANID por sus aportes para que proyectos como este puedan ser realizables. Al equipo del SSMSO por facilitar los datos y acoger el proyecto. A todas las personas u organizaciones referenciadas en esta memoria, en especial a Daniel Ponce Maripangui por facilitar su memoria.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción y Antecedentes

Las pandemias han sido eventos que han ocurrido a lo largo de la historia de la humanidad, unas más letales que otras, además de ir apareciendo en escenarios mundiales bastante diversos e incluso diametralmente diferentes. Actualmente, el mundo se siente en un período de madurez superior, un poco más preparado, menos indefenso y más estable, sobre todo porque ya han ocurrido y siguen ocurriendo grandes revoluciones tecnológicas tales como la industrial, la científica, la informática, la científico-técnica, etc., y además esta vez no nos encontramos en medio de una guerra mundial.

El mundo se siente casi en la cúspide del conocimiento y progreso en las sociedades, pero el nuevo coronavirus demostró que quizás no todo estaba tan bien madurado ni lo suficientemente preparado para enfrentar situaciones de emergencia como estas. Si bien al día de hoy se ha logrado un mayor control del virus (a través de medidas no siempre eficaces y por medio del conocimiento del mundo médico y el avance de la vacunación de la población) ha quedado al descubierto también la inequidad existente en el mundo, la implícita división entre países y personas ricas y pobres, lo cual ha traído consigo no sólo problemas de salud sino también económicos y sociales.

En este mundo moderno, pese a lo desesperanzador que pueda parecer la situación, no lo es tanto si se trabaja duro para poder combatir esta emergencia sanitaria de la mejor manera, utilizando los conocimientos, la inteligencia, todo lo que el humano ha logrado en su historia y utilizando las nuevas tecnologías que son el emblema de la modernidad que se vive. Si bien se dice que se vive en la era de la Transformación Digital, y se escuchan conceptos tales como inteligencia artificial, deep learning, big data, data science, machine learning, etc., entonces en una pandemia como la actual, ¿es esto aplicable? ¿ayuda de algo? ¿en qué se están aplicando? ¿se está realmente usando todas las nuevas herramientas que se tienen a la mano? Otro asunto necesario de analizar es ¿qué provecho se le está sacando a los datos con que se cuenta?

No obstante, un paso antes de definir qué herramientas se utilizarán (el medio), sería pensar ¿por dónde empezar? ¿desde qué arista se le hará frente? ¿qué impacto se quiere tener? ¿qué problema se atacará? ¿cómo? y ¿con qué?. Es decir, en otras palabras, ¿cuál es el fin? ¿por qué se está haciendo? y ¿qué se intentará solucionar? En el caso de esta memoria se quiere abordar la pandemia desde una perspectiva geográfica. La importancia de esto radica sobre todo en la necesidad imperiosa de detectar rápidamente los casos nuevos y los conglomerados de casos

de COVID-19, antes de que se produzcan brotes o se generalice la transmisión dado que los brotes de COVID-19 pueden crecer de forma rápida y exponencial. La vigilancia continua de la COVID-19 también es clave para conocer las tendencias y cambios epidemiológicas a lo largo del tiempo. Por lo tanto, los sistemas de vigilancia han de ofrecer una cobertura geográfica completa y se debe potenciar la vigilancia de los grupos de población vulnerables o de alto riesgo. Por consiguiente, ello exigirá una combinación de sistemas de vigilancia que comprenderá el rastreo de contactos en todos los niveles del sistema de salud, en el ámbito comunitario, en los entornos residenciales cerrados y en otros grupos vulnerables [4].

Sin embargo, para la realización de análisis geográfico para la investigación de pandemias o de otro tipo tales como: "mapeo de crímenes o sistemas de análisis, sistemas de manejo de desastre y riesgo y ciencia política" [3], el dato clave es contar con la dirección del sujeto y lograr que ésta pueda ser llevada convertida en una posición, es decir, en coordenadas (latitud, longitud), el problema radica en que esto no siempre es lograble debido a diversas razones tales como: la dirección con la que se cuente del sujeto o sujetos a analizar podría no ser intepretable correctamente por los servicios de geocodificación, ya sea porque la dirección estuviese mal escrita, la captación de la dirección desde las bases de datos no se logró de manera exitosa, la dirección no existe, la dirección está incompleta, la dirección no coincide con la forma en la cual aparece en los registros de los servicios de geocodificación o que la dirección tiene información de más o de menos, etc.

Por lo tanto, si no se logra hacer esta conversión de texto de la dirección hacia coordenadas, tampoco se puede ubicar al sujeto en un mapa, ni menos llegar a hacer cualquier análisis geodemográfico o de transmisibilidad, por lo tanto ¿qué técnicas se puede usar para resolver este problema? Luego de resuelto, ¿en qué podría servir esta información en la actual pandemia? ¿se podría lograr predecir/explicar el comportamiento del contagio o las posibles variables que podrían estar afectando? ¿hay variables geodemográficas que expliquen la transmisión?. Preguntas como estas intentarán ser respondidas y/o resueltas en el presente trabajo.

1.1. Definiciones y conceptos iniciales

Es necesario partir definiendo 2 conceptos básicos que son importante entender para el desarrollo de esta memoria: **dirección**, esta palabra tiene variados significados en el español, incluso algunos bastante distantes de la temática de esta memoria, pero el que nos interesa es el siguiente: "domicilio de una persona, la ubicación de un edificio[...]" [5]. El segundo concepto es **domicilio**, el cual "se utiliza para nombrar a la vivienda permanente y fija de una persona", y es además "un atributo que puede aplicarse a cualquier persona, tanto física como jurídica", también es necesario tener en cuenta que "el domicilio legal (es decir, aquel declarado ante el Estado) puede no coincidir con el domicilio real" [6].

En esta memoria cuando se habla de dirección o atributo "DIRECCION" se hace referencia en su mayoría a la dirección del domicilio legal de una persona natural (física), es decir, la dirección del lugar en que la persona reside de acuerdo a los registros que se tiene en el sistema público, que no siempre coincide con el domicilio real actual de una persona. De todas formas, se utilizarán indistintamente para hacer referencia a la dirección del domicilio del sujeto.

Además, otro concepto clave es **geocodificación** o **codificación geográfica**, o por su nombre en inglés *geocoding*, este proceso es "la conversión de direcciones basadas en texto hacia coordenadas geográficas" [7] (latitud, longitud), pero hay otros conceptos que se suelen confundir con esta idea: *georreferenciación*, *geolocalización* y *geoetiquetado*. "Se define la georreferenciación, como un proceso por el cual se dota de un sistema de referencia de coordenadas terreno a una imagen digital que originariamente se encuentra en coordenadas pixel. La geolocalización, en cambio, se define como la identificación de la ubicación de un dispositivo por ejemplo un radar, teléfono móvil o cualquier aparato tecnológico conectado a internet" [8]. Finalmente, el geoetiquetado "es el proceso de agregar información geográfica en los metadatos de archivos de imágenes, vídeos, sonido, sitios web, etc. que sirva para su georreferenciación" [9].

Además resulta importante definir otros conceptos relacionados al coronavirus y que podrían otorgar una mejor comprensión al leer la memoria. Las siguientes definiciones provienen del Manual operativo para las acciones de Trazabilidad y Aislamiento del Minsal [10]:

Con respecto al Tipo de Caso:

- "Caso sospechoso: Paciente que presenta un cuadro agudo con al menos dos de los síntomas compatibles con COVID-19: fiebre (37.8ºC o más), tos, disnea, dolor toráxico, odinofagia, mialgias, calofríos, cefalea, diarrea, o pérdida o disminución brusca del olfato (anosmia o hiposmia) o del gusto (ageusia o disgeusia). O bien, paciente con infección respiratoria aguda grave (que requiere hospitalización).
- Caso confirmado: Toda persona que cumpla la definición de caso sospechoso en que la prueba específica para SARS-CoV-2 resultó "positiva" (RT-PCR).
- Caso confirmado asintomático: Toda persona asintomática identificada a través de estrategia de búsqueda activa en que la prueba específica de SARS-CoV-2 resultó "positiva" (RT-PCR).
- Caso probable:
    1. Caso probable por resultado de laboratorio: paciente que cumple con la definición de caso sospechoso en el cual el resultado de la PCR es indeterminado, o bien tiene una prueba antigénica para SARS-CoV-2 positiva.
    2. Caso probable por nexo epidemiológico: persona que ha estado en contacto estrecho con un caso confirmado, y desarrolla fiebre (temperatura axilar mayor a 37.8 grados celsius) o al menos dos síntomas compatibles con COVID-19 dentro de los primeros 14 días posteriores al contacto. No será necesaria la toma de examen PCR para las personas que cumplan los criterios de caso probable. Si por cualquier motivo, un caso probable se realiza un examen confirmatorio y este resulta positivo, se considerará como caso confirmado. Por el contrario, si el resultado es negativo o indeterminado, se seguirá considerando caso probable.
    3. Caso probable por imágenes: caso sospechoso con resultado de RT-PCR para SARS-CoV-2 negativo pero que cuenta con una tomografía computarizada de tórax con imágenes características de COVID-19 según el informe radiológico.
    4. Caso probable por síntomas: persona que presenta pérdida brusca y completa del olfato (anosmia) o del sabor (ageusia) sin causa que lo explique.

5. Nota: Los casos probables se deben manejar para todos los efectos como casos confirmados: Aislamiento por 11 días a partir la fecha de inicio de síntomas, identificación y cuarentena de sus contactos estrechos y licencia médica."

Otros conceptos relacionados:

- "Aislamiento: Restricción de movimiento por 11 días que se aplica a los casos confirmados y probables.
- Cuarentena: Restricción de movimiento por 14 días que se aplica a los contactos estrechos.
- RT-PCR (sigla en inglés): Reacción en cadena de la polimerasa con transcriptasa inversa en tiempo real. Actualmente, la detección del virus se basa en esta técnica con muestras nasofaríngeas.
- Trazabilidad: Estrategia que permite identificar de manera continua a los contactos estrechos de un caso probable o confirmado.
- Trazadores: Personal de salud a cargo de la estrategia de trazabilidad en los centros de APS.
- Grupos de riesgo: Personas que poseen características que aumentan su probabilidad de padecer una enfermedad grave. Estas pueden ser: Personas mayores de 65 años, personas con obesidad, tabaquismo, diabetes, hipertensión arterial, cáncer activo, trasplantado, virus de la inmunodeficiencia humana (VIH), enfermedad autoinmune o tratamiento inmunosupresor por otra patología, personas postradas o personas con patologías respiratorias."

1.2. Pandemias

Una pandemia es una epidemia, es decir, "un incremento significativamente elevado en el número de casos de una enfermedad con respecto al número de casos esperados" [11], "que ocurre a una escala que cruza las fronteras internacionales y que generalmente afecta a personas a escala mundial" [12]. Sin embargo, una enfermedad o afección, por el hecho de estar extendida o que causa muchas muertes no es una pandemia, ya que debe tener un carácter infeccioso. Por ejemplo, el cáncer es responsable de muchas muertes, pero no se considera una pandemia, porque la enfermedad no es contagiosa (es decir, fácilmente transmisible) y tampoco es infecciosa [12].

Pese a lo preocupante que ha sido la covid-19, nuestra más reciente pandemia, han existido casos de pandemias más letales a lo largo de la historia de la humanidad. A continuación se describirán brevemente 5 de ellas: La Peste Negra, "la epidemia más devastadora de la historia de la humanidad, terminó con la vida de entre 75 y 200 millones de personas en el siglo XIV". La Viruela, "también bautizada en honor a las pústulas que provoca en la piel, fue una pandemia devastadora con una tasa de mortalidad de un 30 %, especialmente alta entre niños y bebés", se le asocia la muerte de 56 millones de personas durante el año 1520. La Gripe Española, "al contrario de lo que puede parecer debido a su nombre, mató a más de 40 millones de personas en todo el mundo entre los años 1918 y 1919 y no se inició

en España, sino más bien muchos científicos sitúan sus primeros casos en Estados Unidos en 1918". La Plaga de Justiniano, "existen evidencias que sugieren que la llamada Plaga de Justiniano se encuentra en cuarto lugar entre las más devastadoras, con cifras de mortalidad entre los 25 y los 50 millones de personas fallecidas", abarcando principalmente los años 541 y 542. VIH/SIDA, "desde su aparición en 1976 ha matado a 32 millones de personas, según la Organización Mundial de la Salud, y al día de hoy aún hay entre 31 y 35 millones conviviendo con la enfermedad, sobre todo en África" [13].

1.3. Covid-19

La COVID-19, de acuerdo con [1], es "la enfermedad causada por el nuevo coronavirus conocido como SARS-CoV-2". Según [14]: "La mayoría de las personas infectadas por el virus experimentarán una enfermedad respiratoria de leve a moderada y se recuperarán sin requerir un tratamiento especial. Sin embargo, algunas enfermarán gravemente y requerirán atención médica. Las personas mayores y las que padecen enfermedades subyacentes, como enfermedades cardiovasculares, diabetes, enfermedades respiratorias crónicas o cáncer, tienen más probabilidades de desarrollar una enfermedad grave. Cualquier persona, de cualquier edad, puede contraer la COVID-19 y enfermar gravemente o morir".

La primera noticia que recibió la Organización Mundial de la Salud sobre este nuevo virus ocurrió "[...] el 31 de diciembre de 2019, al ser informada de un grupo de casos de *neumonía vírica* que se habían declarado en Wuhan (República Popular China)". [1] Es importante destacar que en un comienzo, como fue mencionado anteriormente, el virus era sólo catalogado como una neumonía, y no como es ahora "una nueva cepa de coronavirus". Un mes más tarde, "el 30 de enero de 2020, el Director General de la Organización Mundial de la Salud (OMS) declaró que el brote de COVID-19 constituía una emergencia de salud pública mundial[...]". [15] Luego, "el 11 de marzo de 2020, tras la propagación del brote en un gran número de países de varios continentes, el Director General declaró, por consejo del Comité de Emergencias del Reglamento Sanitario Internacional, que la COVID-19 había adquirido el carácter de pandemia" [15].

En Chile, "el 3 de marzo se detectó en el Hospital de Talca el primer paciente de COVID-19 en el país, caso que fue notificado por el Instituto de Salud Pública (ISP) y el Laboratorio del Hospital Guillermo Grant Benavente luego de analizar la muestra del paciente mediante la técnica PCR". [16] Al día 17 de julio de 2021, la cantidad total de casos en el país superó la cifra 1.5 millones y la cantidad de fallecidos superó la cifra de 30 mil [2].

1.4. Contexto de la Investigación

La presente investigación se enmarca en el proyecto COVID 0251, financiado por la Agencia Nacional de Investigación y Desarrollo. El proyecto se titula "Sistema integrado de información para el seguimiento domiciliario de pacientes COVID-19 en servicios de salud", cuyo director es Richard Weber, Profesor Titular de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile. En este proyecto participaron investigadores del Departamento de Ingeniería Industrial y profesionales del Centro de Sistemas Públicos de esta unidad.

El objetivo general de este proyecto es: "aumentar la efectividad del seguimiento a los

pacientes COVID-19 confirmados, los casos sospechosos, los casos probables y sus contactos mediante una plataforma analítica escalable que integre diferentes fuentes de información, incluyendo la autodeclarada por pacientes, y entregue analítica relevante para potenciar la toma de decisiones". Es posible encontrar más información sobre el proyecto en el siguiente enlace: https://www.sistemaspublicos.cl/gproyecto/covid0251/

Dentro de este proyecto participaron como tesistas varios alumnos de la Universidad de Chile, en el cual cada uno trabajó en una herramienta que aportara desde un punto de vista distinto al logro del objetivo general. El alumno autor de esta tesis junto a Juan Pablo Alvarado Glena desarrollaron una aplicación en Streamlit a través de Python que permitiera normalizar y geocodificar de forma eficiente y eficaz las direcciones de los pacientes que estuviesen integrados a la plataforma de seguimiento, además de lograr la visualización espacial y temporal de los casos agregando nociones de riesgo clínico por unidades territoriales.

1.5. La Organización

Para el desarrollo de esta memoria se trabajó con datos proporcionados por el Servicio de Salud Metropolitano Sur Oriente (SSMSO). El SSMSO es una Red de Atención de Salud Pública del país, cuyo propósito es satisfacer las necesidades de salud de la población usuaria preferentemente del territorio Metropolitano Sur Oriente, es decir, de las comunas de Puente Alto, La Florida, San Ramón, La Granja, La Pintana, San José de Maipo y Pirque, a través de sus centros de atención primaria (Cesfam), hospitales, centros de salud mental comunitaria y SAPUS. Las comunas anteriormente nombradas comprenden el área bajo el nombre Suroriente de la Figura 1.1.

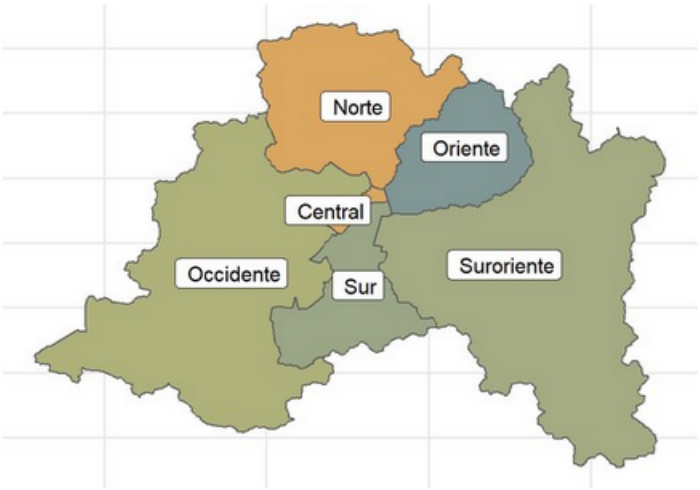


Figura 1.1: Mapa de divisiones del Servicio de Salud Metropolitano

1.6. Estructura de la Memoria

Esta sección tiene por objetivo explicar la estructura de los siguientes capítulos y secciones:

- **Introducción y Antecedentes**: En este capítulo se entregan una introducción al tema y nociones sobre lo que es una pandemia y contexto sobre el proyecto de investigación.

Para lo primero se explica qué es una pandemia, cuáles han sido las más letales en la historia, de qué se trata y cómo ha ido evolucionando la Covid-19 en el mundo y en el país. Para lo segundo, se explica el proyecto y la organización en el cual estuvo inserto el desarrollo de la memoria, y se expone también la herramienta de visualización desarrollada, que en su momento fue el entregable.

- **Definiciones de la memoria**: En este capítulo se relatan las razones por la cuales no se puede aplicar directamente geocodificación a las direcciones de los pacientes existentes. Además, se describen las hipótesis, el objetivo general, los objetivos específicos y la solución propuesta a estos problemas, alcances y resultados esperados del trabajo de memoria.

- **Marco Conceptual**: En este capítulo se exponen trabajos relacionados y las bases conceptuales que fundamentan los métodos de normalización de direcciones.

- **Metodología**: En este capítulo se explican las herramientas que existen para geocodificar direcciones y su funcionamiento. Además se presentan los datos y bases de datos utilizadas en el desarrollo de la memoria. Finalmente, se detalla las etapas/procesos que van recorriendo los datos para lograr los objetivos, por lo cual se describen: los métodos de normalización de direcciones, la geocodificación de direcciones normalizadas, la asignación de paciente a manzana censal y la visualización de casos.

- **Evaluación y Resultados**: En este capítulo se detallan 2 experimentos que se quiso llevar a cabo y sus resultados para evaluar el desempeño de la solución propuesta y los procesos/etapas asociadas.

- **Conclusiones**: En este capítulo se analizan el cumplimiento de los objetivos y se responden ciertas preguntas propuestas en la Introducción. Además se discuten los resultados y se compara con los obtenidos en literatura relacionada. También se señalan cosas que podrían haber quedado pendientes y que podrían tratarse en un trabajo futuro.

Capítulo 2

Definiciones de la memoria

2.1. El problema

Al analizar los detalles de esta memoria puede surgir una duda bastante válida: ¿por qué no bastaría solamente pasar las direcciones directamente a una API de geocodificación? Es posible, sin embargo lo real es que con los datos con los que se cuenta y la estructura de aquellos, no se lograrían los mejores resultados por diversos motivos, tales como: las direcciones no están escritas bajo un estándar, direcciones incompletas o con más de la información necesaria, existencia de abreviaturas, sistemas imprecisos de numeración (por ejemplo, existencia de más 1 número, lo cual confunde al momento de identificar el número de la calle), errores de escritura, la captación de la dirección desde las bases de datos no se logró de manera exitosa, la dirección no existe, la dirección no coincide con la forma en la cual aparece en los registros de los servicios de geocodificación, etc.

En la base utilizada en particular también aparece el problema de que lo señalado en el campo COMUNA no siempre es la comuna de la dirección, probablemente esté asociado a algún otro dato del paciente, como por ejemplo, la comuna del establecimiento de salud en el cual se atendió, pero se descubrió que en los casos en que existía un código postal la comuna asociada a este expresaba la comuna correcta. Hay que pensar que la dirección la está intentando interpretar una máquina con características de sistema automático, que aprende, pero no logra del todo razonar como lo haría un humano. Lo ideal para un geocodificador es que la dirección cuente con el nombre de la calle, el número de la casa/edificio y la ciudad/comuna, sobre todo cuando se hace una búsqueda por componentes para lograr una mayor precisión y menores errores, todo lo adicional puede confundir al localizador y entregar como resultado una localización no correcta o incluso lograr que la dirección no sea encontrada.

Además, si no se logra encontrar correctamente un paciente hay muchas consecuencias que esto puede acarrear sobre todo en el sistema de salud, como por ejemplo si sucede una emergencia se necesita acudir rápidamente al domicilio del paciente, por lo cual se pueden realizar 2 cosas: preguntarle al paciente su dirección, o acudir con la información de su domicilio que se tiene, ¿qué tal si el paciente no puede hablar? ¿qué tal si el paciente por nerviosismo no recuerda su dirección? Se necesita actuar rápido en una emergencia. O, relacionado con el covid, ¿cómo se podrá saber dónde están los casos si no se ha logrado localizar los casos? ¿cómo se enfrenta o identifica un brote si no se tiene un nivel de agregación menor a la comuna? ¿cómo se podría realizar análisis sociodemográficos si no se sabe dónde específicamente

viven los pacientes?

2.2. Hipótesis

Basado en ejemplos de la literatura la normalización de direcciones es lograble y puede mejorar los resultados de búsqueda [7] [17], además también se ha evidenciado que ciertos factores relacionados a la pobreza multidimensional, tales como: educación, salud, vivienda, trabajo, entre otros, son los que pueden hacer que catástrofes naturales y pandemias, como la COVID-19, pongan en mayor riesgo a un grupo de la población por sobre otro [18] [19] [20]. Por lo cual se proponen las siguientes hipótesis:

- Es posible el uso de un método de normalización sobre las direcciones que mejore el proceso de geocodificacion de éstas con respecto a si no se aplicase ninguno.
- Es posible encontrar una mayor ocurrencia de casos ante la presencia de mayor hacinamiento en los hogares.

2.3. Objetivo General

Como Objetivo General de esta Memoria se tiene: "Desarrollar un método de normalización de direcciones que permita geocodificar de manera eficiente y eficaz las direcciones de los pacientes covid del SSMSO. Se desea también desarrollar herramientas de visualización y realizar un análisis con variables sociodemográficas que permita entender la propagación del virus en el país".

2.4. Objetivos Específicos

Para poder lograr alcanzar el objetivo general, se tienen los siguientes objetivos específicos:

1. Desarrollar y probar métodos eficientes y eficaces de normalización/estandarización de direcciones.
2. Aplicar normalización a las direcciones de los pacientes confirmados y probables ingresados en la plataforma del SSMSO.
3. Geocodificar las direcciones normalizadas, revisar el desempeño de la búsqueda y analizar el desempeño y conveniencia de las APIs de geocodificación.
4. Asociar variables espaciales, temporales y sociodemográficas al caso.
5. Desarrollar herramientas de visualización que permitan ver la evolución de la pandemia.
6. Estudiar la relación entre variables sociodemográficas con respecto a la propagación del virus.

2.5. La Solución

Una solución básica podría ser poner a un humano a interpretar las direcciones una a una y buscarlas usando algún servicio de geocodificación, pero esto indudablemente no es

lo óptimo sobre todo cuando se trata de grandes cantidades de datos. Si se hiciera hay que pensar en el gran cansancio mental que se ocasionaría a la persona a cargo y segundo la gran inversión o pérdida de tiempo en que podría incurrirse, tiempo que en la salud es valioso. Por lo tanto, lo que se busca en esta memoria es idear un sistema inteligente que pueda emular o acercarse a lo que haría un humano con su razonamiento e interpretación, pero de forma automática.

En los siguientes capítulos se desarrollará esta idea y se llevará a cabo con cierto procesos que procesarán los datos y permitirán lograr una forma más eficaz y eficiente de geocodificar la direcciones. Además se demostrará un análisis sociodemográfico que es posible llevar a cabo.

Los procesos mencionados consisten en los siguientes principalmente: Normalización de direcciones, geocodificación de las direcciones normalizadas, asignación de manzana-entidad a paciente, asociar variables sociodemográficas y finalmente visualización espacial y temporal de los casos.

A modo de ejemplo de lo que se quiere lograr hacer se muestra a continuación en la Tabla 2.1 20 ejemplos de direcciones tomadas al azar de la base tal cual vienen, y cómo la interpretaría un humano, que es a lo que se espera se acerque la máquina al realizar la solución propuesta.

Tabla 2.1: Ejemplos de direcciones

DIRECCION (DESDE LA BASE)	COMUNA (DESDE LA BASE)	CALLE+NÚMERO (HUMANO)	COMUNA (HUMANO)
PASAJE PUNITAQUI 10559	LA PINTANA	PUNITAQUI 10559	LA PINTANA
BORODIN 01491 DEPTO 35	PUENTE ALTO	BORODIN 1491	PUENTE ALTO
AMERICO VESPUCIO	LA GRANJA	DIRECCIÓN INVÁLIDA	DIRECCIÓN INVÁLIDA
TONGOY 1065-C, 8820000 LA PINTANA, REGION SANTIAGO METROPOLITAN, CHILE	LA PINTANA	TONGOY 1065	LA PINTANA
CORDILLERA DE LA COSTA 2680	PEÑAFLOR	CORDILLERA DE LA COSTA 2680	PEÑAFLOR
josefa portales 11572	LA FLORIDA	JOSEFA PORTALES 11572	LA FLORIDA
CALLE EMILIANO FIGUEROA 8082, 8860000 SAN RAMON, REGION SANTIAGO METROPOLITAN, CHILE	SAN RAMON	EMILIANO FIGUEROA 8082	SAN RAMÓN
PEDRO AGUIRRE CERDA 10935 1	LA PINTANA	PEDRO AGUIRRE CERDA 10935	LA PINTANA
SANTA RAQUEL 4416 BLOCK10 DPTO12	LA FLORIDA	SANTA RAQUEL 4416	LA FLORIDA
LOS TOLOMIROS 6425, LA FLORIDA, REGIÓN METROPOLITANA, CHILE	LA FLORIDA	LOS TOLOMIROS 6425	LA FLORIDA
CALLE EL BOSQUE, 8150000 PUENTE ALTO, REGION SANTIAGO METROPOLITAN, CHILE	LA FLORIDA	DIRECCIÓN INVÁLIDA	DIRECCIÓN INVÁLIDA
AV AMERICO VESPUCIO 0769 DEPTO 11	LA GRANJA	AMÉRICO VESPUCIO 769	LA GRANJA
PASAJE PASEO LOS GUINDOS 3681, 8240000 LA FLORIDA, REGION SANTIAGO METROPOLITAN, CHILE	LA FLORIDA	PASEO LOS GUINDOS 3681	LA FLORIDA
CALLE SAN PEDRO 941, 8240000 LA FLORIDA, REGION SANTIAGO METROPOLITAN, CHILE	LA FLORIDA	SAN PEDRO 941	LA FLORIDA
SOTERO DEL RIO 515 LA FLORIDA, REGIÓN METROPOLITANA, CHILE	LA FLORIDA	SOTERO DEL RIO 515	LA FLORIDA
PASAJE SANTA ÁNGELA 11040, 8820000 LA PINTANA, REGION SANTIAGO METROPOLITAN, CHILE	LA PINTANA	SANTA ÁNGELA 11040	LA PINTANA
VICUÑA MACKENNA PONIENTE 6800, DPTO 405	LA FLORIDA	VICUÑA MACKENNA PONIENTE 6800	LA FLORIDA
CALLE IGNACIO ECHEVERRÍA 8203, 7970000 LA CISTERNA, REGION SANTIAGO METROPOLITAN, CHILE	LA CISTERNA	IGNACIO ECHEVERRÍA 8203	LA CISTERNA
PASAJE RENÉ SCHNEIDER 9123, 8860000 SAN RAMON, REGION SANTIAGO METROPOLITAN, CHILE	SAN RAMON	RENÉ SCHNEIDER 9123	SAN RAMÓN
CALLE MILLALONGO 79, 8240000 LA FLORIDA, REGION SANTIAGO METROPOLITAN, CHILE	LA FLORIDA	MILLALONGO 79	LA FLORIDA

2.5.1. Evaluación del desempeño de la solución

Para evaluar el desempeño de la solución propuesta se llevarán a acabo 2 Experimentos cuyos resultados se presentarán en el capítulo 5:

1. Probar, evaluar y comparar el desempeño de los distintos métodos de normalización de direcciones descritos en esta memoria y luego comparar la búsqueda de las direcciones resultantes utilizando la API de Geocodificación de Google Maps y la de Open Street Map

2. Asignación de pacientes a manzanas-entidades correspondientes, asociar variables sociodemográficas interesantes y analizar relación entre el hacinamiento con la ocurrencia de casos

Principalmente en cada uno de los experimentos se evalúa el desempeño de los distintos métodos presentados en esta memoria, analizar sus aciertos y errores, además de las ventajas y desventajas de uno frente a otro.

Además es importante señalar que la variable sociodemográfica que está en estudio es "el hacinamiento", definido como la cantidad de personas promedio que vive por vivienda. Por lo cual, se espera determinar si el hacinamiento determina una mayor ocurrencia de casos promedios, en otras palabras, si en sectores en donde la población vive más hacinada en promedio ocurren una mayor cantidad de casos, lo cual permitiría determinar si esta variable es determinante en la propagación del virus. Los resultados de esto también pueden ser encontrados como parte del segundo experimento desarrollado en el capítulo 5.

2.6. Alcances y Resultados Esperados

Al final del trabajo realizado en esta memoria, se espera que el alumno memorista haya logrado geolocalizar correctamente a gran parte (90 %) de los pacientes del SSMSO existentes con un domicilio válido, sobre todo con respecto a los Confirmados y Probables, ya que son 2 grupos importantes de observar debido a que ambos se consideran como confirmados con respecto a haber adquirido COVID. Con esto también se espera haber desarrollado un método o forma de automatización en la cual se logre que las direcciones de los pacientes puedan ser estandarizadas y encontradas con gran porcentaje de eficacia por los motores de búsquedas de geocodificación, debido a que esto lograría ser un aporte para agilizar la ubicación de pacientes, y como aporte para otros procesos/modelos. También se espera lograr construir una visualización temporal y espacial que pueda permitir ver el desplazamiento de la pandemia de forma clara y didáctica dentro de las comunas del SSMSO.

Además, se espera que con la ubicación de los pacientes lograda, pueda sevir como aporte para la discusión de saber si efectivamente indicadores como el hacinamiento explican la evolución y comportamiento de la pandemia en el país o explicar de mejor manera el contagio persona-persona, al menos en temas geográficos-temporales.

Capítulo 3

Marco Conceptual

3.1. Trabajos Relacionados y Referencias

En el transcurso del año 2021, el alumno de la Universidad de Chile Daniel Ponce Maripangui publicó su memoria para optar por el título de ingeniero civil industrial bajo el título "Sistema de búsqueda inteligente de direcciones para empresa de Distribución Postal" [17], cuyo problema consistió en "la asignación de códigos postales a una cantidad significativa de direcciones no normalizadas provenientes de clientes tipo empresas". Se propuso entonces como solución un sistema de búsqueda inteligente de direcciones postales que permitiese normalizar automáticamente grandes volúmenes de direcciones utilizando modelos de procesamiento del lenguaje natural.

La solución propuesta plantea un clasificador de direcciones para segmentar y etiquetar sus atributos (nombre de calle, número principal e información adicional). El clasificador, además, verifica si posee un número principal bien definido, si no lo posee la dirección se identifica como inválida y no se normaliza debido a que el código postal requiere reconocer un frente de cuadra para poder ser asignado. Con las direcciones etiquetadas, se aplica un modelo de coincidencia de texto utilizando la distancia de Levenshtein y el ratio de similitud de Levenshtein, generando un ranking con las tres direcciones más similares en base al puntaje ranking, para luego calcular el puntaje de selección. Finalmente, la dirección es normalizada si se cumplen los criterios de asignación. El modelo de coincidencia de texto con mejor rendimiento elimina abreviaturas que enuncian un tipo de calle (por ejemplo: PSJE o AVDA), calcula el puntaje ranking utilizando el ratio de similitud de Levenshtein y busca la coincidencia comparando la cadena completa de texto. Finalmente, se genera categorías de confiabilidad según el puntaje de selección para disminuir el error de la normalización.

El trabajo de memoria [17] se limitó a normalizar las direcciones a nivel de comuna y en zonas urbanas. Además, se entrenaron los modelos utilizando solo direcciones de la comuna de Quilicura y Santiago. Al normalizar envíos de Quilicura a través de coincidencia directa, se lograba un 20 % de normalización. Al aplicar la solución propuesta, se alcanzó sobre un 90 % de normalización con un error asociado menor a 5 %. Al utilizar las categorías de confiabilidad, se alcanzó sobre 80 % de normalización con un error asociado menor a 1 %.

Un punto importante a señalar es que el trabajo desarrollado en [17] no implicó una geocodificación de las direcciones normalizadas, sino más bien una asignación al código postal que

le correspondiere dado el algoritmo de coincidencia de texto comparada hacia las direcciones de la base con que se contaba. Por lo cual, en lo principal en que corresponden el trabajo presentado en esta memoria y en aquella es la normalización de las direcciones.

Anteriormente, en el año 2018 se publicó un artículo bajo el título "Address standardization using the natural language process for improving geocoding results"[7] de Dilek Küçük Matci e Uğur Avdan, en el cual se hace referencia al problema de que el uso del proceso de geocodificación no es posible aún en algunas áreas en donde podría servir como una herramienta efectiva, teniendo como causa varias razones tales como inconsistencias en los formatos de las direcciones, lo cual incluye sistemas imprecisos de numeración, errores de escritura, el uso de abreviaciones y una falta de datos que se refieran al proceso de geocodificación. Este estudio busca abordar estos problemas por la vía del proceso de estandarización. Para esto, utiliza un método que descompone las direcciones usadas como datos de entrada en la geocodificación, identifica los problemas de escritura y abreviaciones, y reorganiza las direcciones a través del Proceso de Lenguaje Natural (PLN).

Como datos de testeo son tomadas las direcciones de escuelas primarias en el distrito turco de Eskisehir. Primero el proceso de geocodificación se realiza en el conjunto de datos, usando tanto la API de geocodificación de Google como la API de geocodificación de ArcGIS. Entonces, las direcciones son reformateadas en 3 formatos de dirección con la aplicación de procesos de estandarización. La geocodificación se realiza en las direcciones re-formateadas y los resultados comparados a los resultados no-estandarizados. La estandarización usada se muestra para hacer una mejora significante en la precisión de los resultados de geocodificación. El método usado en este estudio es significante no sólo en aumentar la precisión del proceso de geocodificación, sino también en sostener su uso más amplio.

En este estudio sí se llevó a cabo un proceso de geocodificación de las direcciones normalizadas a diferencia de lo hecho en [17], y se evidencia una mejora significante en la precisión de los resultados de geocodificación, debido a que se observa que el proceso de estandarización mejoró en un 50 % la captura de coordenadas usando la API de geocodificación de Google y además provee una mejora en la proporción de coordenadas encontradas pasando de un 64.4 % a un 99.1 %.

También es importante mencionar un artículo publicado en CIPER bajo el nombre "Hacinamiento: la variable clave en la propagación del Covid-19 en el Gran Santiago" [20], en el cual se relata los resultados de un estudio cuyo objetivo era investigar sobre qué factores caracterizan la diseminación de la covid-19 en las comunas de la Región Metropolitana. Por lo tanto, realizaron un análisis estadístico en el que cruzaron algunos elementos relacionados a la pobreza con tasas de contagio y fallecimiento ocurridos en distintas comunas del Gran Santiago, con lo cual se pudo constatar que una variable clave es el hacinamiento, la cual correlaciona positiva y significativamente con la tasa de contagio y con la tasa de fallecidos. En el estudio consideraron la medición de la pobreza multidimensional por comuna en 5 dimensiones (educación, salud, vivienda, trabajo y seguridad social, entorno y redes) y además se miró con detención otros 5 indicadores que teóricamente se vinculan con mayor vulnerabilidad ante el virus: escolaridad, malnutrición, seguridad social, hacinamiento y estado de la vivienda. En el análisis la variable hacinamiento aparece como la que más correlaciona con tasa de contagio, según éste debido a que la diseminación del virus se facilita cuando se

comparten dormitorios y cuando es difícil aislar a un integrante del hogar que enferma. Tanto la carencia en el indicador de escolaridad como en seguridad social pueden relacionarse con una mayor exposición al virus. Hogares donde sus integrantes tienen trabajos informales o menores niveles de escolaridad, deben muchas veces, exponerse al virus saliendo a trabajar, a pesar de que las normas sanitarias lo prohíban, debido a que no cuentan con protección laboral o no pueden realizar teletrabajo. Por último, niños que presentan malnutrición enfrentan mayores riesgos frente a la enfermedad, ya sea por una desnutrición -sistemas inmunes más débiles- u obesidad -factor de riesgo-.

3.2. Procesamiento de Lenguaje Natural

"El procesamiento del lenguaje natural (PLN) es una ciencia que utiliza métodos de inteligencia artificial para comunicarse con el computador en lenguaje natural". [21] "El lenguaje natural es la lengua o idioma hablado o escrito por humanos para propósitos generales de comunicación". [22] Algunos temas de investigación del PLN son: la extracción de información, la indexación automática y el resumen o la generación de texto en los formatos deseados. En palabras simples, los sistemas de procesamiento de texto automático toman un texto en un formato específico y lo convierten en otro del formato deseado. [21]

Hay cinco elementos básicos de los sistemas de procesamiento del lenguaje natural en general: el analizador sintáctico, el diccionario (léxico), el analizador semántico, la base de conocimientos y el generador. [23] Sin embargo, los diccionarios son el elemento clave, siendo su función permitir determinar las características de las palabras, detectar el lenguaje, detectar palabras y expresiones y separar claves mediante procesos léxicos y sintácticos [24].

3.3. Clasificación de Texto

"El proceso de clasificación de texto consiste en asignar automáticamente categorías o etiquetas predefinidas a texto libre" [25]. Además, la identificación de cada elemento que compone la dirección es crucial para la posterior aplicación de algoritmos que permitan trabajar con la información de la cadena de texto ingresada. El desempeño del anterior proceso afectará la capacidad de los algoritmos de coincidencia de texto aplicados para poder diferenciar una dirección de otra [17].

Existen varias técnicas que se pueden utilizar para la clasificación de texto, pero la técnica utilizada en la memoria es la siguiente:

Clasificación basada en reglas lógicas: Este tipo de modelos aplican un conjunto de reglas lógicas definidas por el usuario y usan información contextual para asignar etiquetas a subconjuntos de palabras. Una de estas reglas podría ser: "Si una palabra es seguida por un número y hay un único número en la oración entonces, es un nombre de calle". Aunque a lo anterior además habría que verificar que el número no sea parte del nombre de una la calle, lo cual puede formar parte de otra regla lógica. El rendimiento de este tipo de técnicas depende de la estructura del conjunto de datos en el cual son aplicados, si es simple y con pocos atributos el rendimiento suele ser bueno. Además es necesario señalar que como no se

utiliza un conjunto de datos de entrenamiento, el modelo es considerado de aprendizaje no supervisado [17].

Existen otras técnicas de clasificación tales como Modelos de Deep Learning o Etiquetado estocástico, cuyos detalles pueden ser revisados en [17], sin embargo no se profundizará en ellos debido a que no se utilizan en el desarrollo de esta memoria.

3.4. Conceptos cartográficos relevantes

Existen algunos conceptos cartográficos interesantes de definir para esta memoria:

- **Meridianos**: Son las líneas imaginarias que atraviesan paralelamente al eje sobre el que gira la Tierra. Es decir, a modo de simplificación, dividen al planeta verticalmente, desde el Polo Norte al Polo Sur [26].

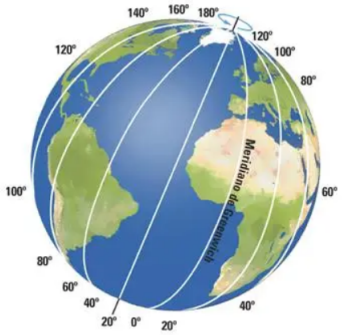


Figura 3.1: Meridianos

- **Paralelos**: Son las líneas imaginarias de la Tierra que atraviesan perpendicularmente al eje sobre el que gira la tierra. O dicho de otra forma más simple, cortan al planeta horizontalmente [26].

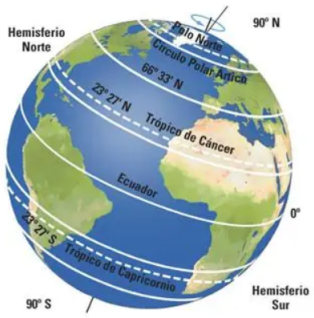


Figura 3.2: Paralelos

- **Latitud**: Es una medida que expresa la distancia angular entre la línea ecuatorial y un punto determinado de la Tierra, medida a lo largo del meridiano en el que se encuentra dicho punto [27].

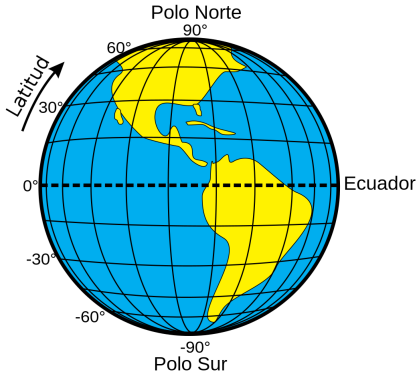


Figura 3.3: Latitud

- **Longitud**: Es una medida que expresa la distancia angular entre un punto dado de la superficie terrestre y el meridiano base, medida a lo largo del paralelo en el que se encuentra dicho punto [28].

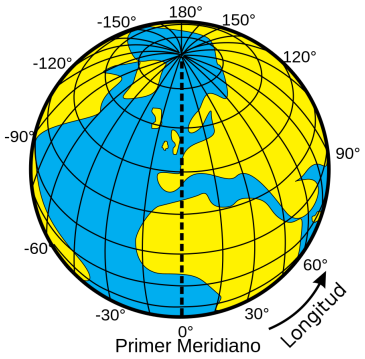


Figura 3.4: Longitud

- **Polígono**: Es un área cerrada (figura de muchos lados) que representa la forma y ubicación de entidades homogéneas como países, estados, regiones, provincias, comunas, manzanas, parcelas, tipos de suelo y zonas de uso del suelo. En otras palabras, un polígono es la serie de puntos geográficos que contienen a una entidad territorial [29].

Figura 3.5: Polígonos de manzanas-entidades en la Región Metropolitana

- **Centroide**: En una figura geométrica, sea línea, superficie o figura tridimensional, el centroide es su centro geométrico. Sería el punto donde coinciden los hiperplanos (según las dimensiones de la figura geométrica) que dividen a la figura en partes de igual momento. Sería su centro de simetría [30].

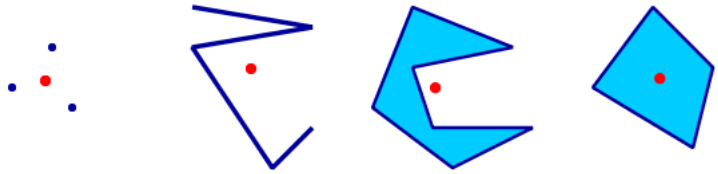


Figura 3.6: Centroide de distintas figuras

3.5. Hacinamiento

Un concepto clave que es utilizado más adelante es "el hacinamiento", el cual se refiere principalmente a la situación que se enfrenta cuando "la cantidad de seres humanos que habitan o que ocupan un determinado espacio es superior a la capacidad que tal espacio debería y puede contener, de acuerdo a los parámetros de comodidad, seguridad e higiene". [31] Tanto en la CASEN como en el CENSO el hacinamiento se mide como la razón entre el número de personas residentes en la vivienda y el número de dormitorios de la misma, considerando piezas de uso exclusivo o uso múltiple. En la CASEN se muestran como el porcentaje de viviendas que corresponden a: sin hacinamiento (menos de 2.5 personas por dormitorio), con hacinamiento medio (2.5 a 3.4 personas por dormitorio), con hacinamiento alto (3.5 a 4.9 personas por dormitorio) y con hacinamiento crítico (5 o más) y sin datos. [32] En el CENSO, también se expresa con el porcentaje de viviendas, pero con las categorías de: indeterminado, hacinamiento medio (2.5 y 4.9 personas por dormitorio) y hacinamiento crítico (5 o más personas por dormitorio) [33].

Desde ambas fuentes se puede obtener tal índice, pero en el caso de la Encuesta CASEN no existe una desagregación más específica como lo son las manzanas-entidades, sino que está construida en mayores niveles de agregación, además de no contar con los datos de todos los hogares, sino que más bien, como dice su nombre es una encuesta, y no es tan exhaustiva como lo es el CENSO. En el caso del CENSO la variable existe a nivel de manzana-entidad, pero existe un no despreciable de viviendas con hacinamiento indeterminado, y además de que si bien existe la variable de la cantidad de viviendas, no existe la variable de cantidad de dormitorios (al menos públicamente) como para hacer el cálculo de forma propia. Por lo cual, en esta memoria se hizo una aproximación, que se mostrará más adelante, basada en la definición de hacinamiento, pero asumiendo de cierto modo una uniformidad en el tamaño de las viviendas por simplicidad.

Capítulo 4

Metodología

4.1. Marco Referencial

4.1.1. Herramientas para la Geocodificación

Como fue descrito en el capítulo 1, la geocodificación, codificación geográfica o *geocoding*, es la manera de convertir una dirección que se encuentra en texto hacia coordenadas geográficas (latitud, longitud), lo cual nos permite visualizar aquella dirección en un mapa. Para lograr esto existen diversas herramientas construidas por distintas proveedores tales como Google Maps, OpenStreetMap, Bing Maps, ArcGIS, etc. Para lograr interactuar con estas plataformas sobre todo cuando se trata de varias consultas simultáneas (ya que por lo general en este tipo de trabajos no es lo ideal ni óptimo ir realizando una consulta una por una corregida por un humano, por la gran cantidad de tiempo que esto exigiría cuando se trata de cientos o miles de direcciones) se puede utilizar la ayuda de ciertos lenguajes de programación, por ejemplo: Python, porque "es un lenguaje sencillo de leer y escribir debido a su alta similitud con el lenguaje humano. Además, se trata de un lenguaje multiplataforma de código abierto y, por lo tanto, gratuito, lo que permite desarrollar software sin límites. Con el paso del tiempo, Python ha ido ganando adeptos gracias a su sencillez y a sus amplias posibilidades, sobre todo en los últimos años, ya que facilita trabajar con inteligencia artificial, big data, machine learning y data science, entre muchos otros campos en auge" [34].

Para lograr la interacción de Python con las plataformas antes mencionadas se debe actuar a través de APIs (interfaz de programación de aplicaciones), donde cada proveedor cuenta con sus propias APIs y teniendo cada una sus propias limitaciones y beneficios sobre todo cuando se habla de un uso gratuito de éstas. Por ejemplo, para usar la API de Google Maps "primero se necesitará crear una cuenta en Google Cloud, e ingresar la información de una tarjeta de crédito/débito" de uso internacional, esto permitirá la creación de una key personal para el uso del servicio. "A pesar de que este servicio es pagado, Google otorgará $200 dólares en créditos gratuitos (con fecha de vencimiento) cuando se cree la cuenta por primera vez tomando la prueba gratuita. Esto significa que se puede hacer alrededor de unas 40.000 llamadas a su API de geocodificación antes de que se haga algún cargo a la tarjeta". Luego de terminado el período de prueba los créditos otorgados desaparecen, pero se mantiene un límite gratuito de consultas antes de ejercer un cargo. Mientras no se alcance este límite, la cuenta no tendrá cargos. En contraste, "la API de OpenStreetMap es totalmente gratuita,

pero es más lenta y menos precisa que la de Google Maps" [35].

Aquí una diferencia fundamental a considerar es que como se dijo hace instantes atrás Google Maps necesita una tarjeta asociada y se pueden generar cargos si se superan ciertos límites, por lo cual pese a sus ventajas en búsqueda, no es lo ideal si se piensa en no ser el único usuario que esté haciendo uso de la API, lo ideal es que sea de uso personal, porque así se puede llevar un uso más medido de la cantidad de búsquedas, y no llevarse sorpresas de grandes cobros a la persona dueña de la tarjeta, en este sentido es más seguro el uso de la API de Open Street Map, ya que se evitan sorpresas indeseadas y además se evita también el proceso de la creación de una cuenta y una key, pero se debe soportar una búsqueda mucho más limitada sobre todo en búsquedas por unidad de tiempo y en "inteligencia", incluso se puede dar el caso de que si se hace la comparación muchas de las direcciones buscadas no puedan ser encontradas por OpenStreetMap mientras sí pueden ser encontradas por Google Maps, así que aquí se juega bastante con un trade-off. Una alternativa a esto es hacer un uso mixto de ambas opciones priorizando la búsqueda por OpenStreetMap, y así evitar sobrepasar los límites gratuitos de Google Maps y logrando un menor tiempo total que solamente haciendo búsquedas con OpenStreetMap.

4.2. Datos

4.2.1. Datos provenientes del SSMSO

El SSMSO, a través de la gestión de sus funcionarios, facilitó el acceso a bases de datos necesarias para este proyecto. Con regularidad se recibía una base en formato excel que contenía los casos anonimizados registrados en la plataforma de seguimiento hasta ese día o los casos acumulados en una ventana de tiempo definida.

La Plataforma de Seguimiento es una herramienta cuyo objetivo es documentar el seguimiento de las personas infectadas con Covid-19 en la red de establecimientos del SSMSO junto con entregar información significativa para la toma de decisiones en salud. Por lo tanto, permitía:

1. Realizar el seguimiento de usuarios ingresados
2. Registrar y actualizar antecedentes de salud de usuarios y sus contactos según evolución clínica.
3. Exportar listas de trabajo de las personas en seguimiento.
4. Obtener estadísticas diarias de la situación sanitaria por subred, comunas y establecimientos.

Dentro del excel exportado desde la plataforma de seguimiento se contaba con 2 hojas escenciales para el análisis de esta memoria las cuales consistían en la hoja de PERSONAS cuyos campos más relevantes eran:

- Tipo de ingreso (Confirmado Indice o Secundario, Probable Indice o Secundario, Sospechoso, Descartado Indice o Secundario y Contacto) que definía el status del caso con respecto a la covid.

- Fecha de inicio de cuarentena.
- Fecha de inicio de sus síntomas.
- Fecha de la toma del examen PCR.

La segunda hoja importante para esta memoria es la de DIRECCIONES, la cual cuenta con:

- DIRECCIÓN (obtenida de diferentes fuentes).
- COMUNA (obtenida de la plataforma) de los pacientes.

La base utilizada para este trabajo es una muestra incremental del seguimiento de casos entre las fechas 22 de Abril de 2020 y 28 de Junio de 2021.

4.2.2. Censo y PreCenso

Para el análisis de la propagación del virus en la población se quiso investigar si ciertas variables sociodemográficas podrían mostrar algún efecto en el comportamiento del contagio, para lo cual luego de lograda la geocodificación del domicilio de cada paciente, se le asociaría a éste los datos sociodemográficos relacionados con su entorno cercano. Para esto era necesario de alguna forma extraer variables relevantes del contexto del paciente, por lo cual se decidió utilizar información recolectada por el Censo sobre los habitantes del país. "El Censo es el conteo y caracterización de todas las viviendas y habitantes del territorio nacional en un momento determinado" [36] y en Chile esto es llevado acabo por el INE (Instituto Nacional de Estadísticas).

El Censo más cercano a nuestra fecha se realizó el año 2017, siendo un censo de hecho [37], contando con una etapa previa el año 2016, la cual se denomina PreCenso. "Para el levantamiento del Censo del año 2017 y lograr una desagregación de los datos censales se utiliza en primer lugar la División Política Administrativa del país, de carácter legal y luego la división censal, que es de ámbito operativo, la que permite obtener una desagregación a nivel de microdato.

- División Política Administrativa (DPA): Permite el cumplimiento de los objetivos, tanto de gobierno como de la administración del Estado, y se encuentra normada mediante la Constitución Política. De acuerdo con la actual División Política Administrativa, el país está conformado por 16 regiones, 56 provincias y 346 comunas. Para efectos legales, es la Comuna la unidad básica de administración del territorio. Para efectos operativos, está subdividida en unidades territoriales menores que permiten una mejor organización precensal y censal.
- División geográfica censal: El territorio comunal se divide en distritos, los que pueden ser urbanos, rurales o mixtos. A su vez, en el área urbana se reconocen Zonas censales y en el área rural, Localidades. Las Zonas censales se componen de manzanas y las Localidades de entidades, las cuales a su vez tienen categorías." [38]

En la Figura 4.1 se puede apreciar con más detalle las divisiones existentes relatadas anteriormente.

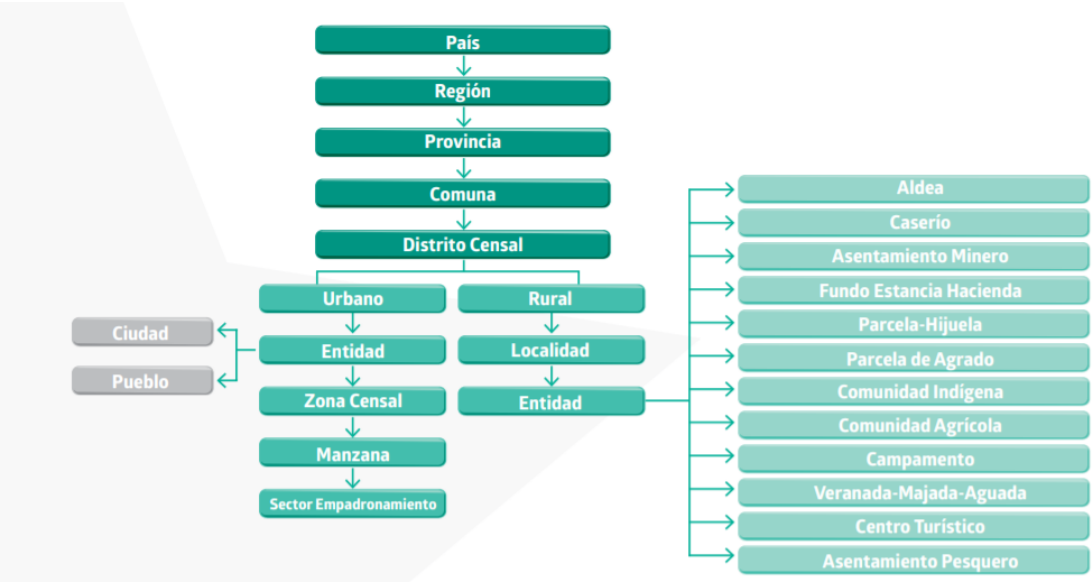


Figura 4.1: Division del territorio tomada de [37]

En específico, en este trabajo se quiso hacer un análisis utilizando la unidad mínima de agregación, para obtener resultados más específicos. Por lo cual, se hizo uso de las manzanas-entidades, aunque principalmente las manzanas, debido a que estas conforman las zonas urbanas. Una manzana censal es la "unidad geográfica básica con fines estadísticos que conforman zonas censales en áreas urbanas" [38] y contiene "un grupo de viviendas contiguas o separadas, edificios, establecimientos y/o predios, delimitados por rasgos geográficos, culturales y naturales" [37]. Es una división similar a una cuadra, aunque en ocasiones puede corresponder a una agrupación de ellas. En la Figura 4.2 se puede visualizar varias manzanas-entidades de la Región Metropolitana para tener una noción aproximada de su forma y estructura.

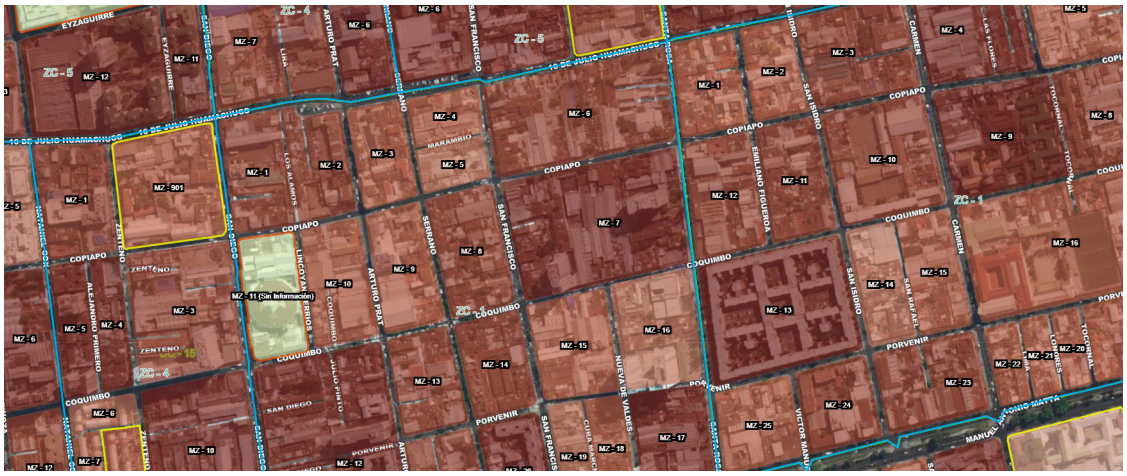


Figura 4.2: Algunas manzanas-entidades de la RM en el mapa

La base de datos del Censo tiene la siguiente estructura: Cada fila incluye la información de una manzana-entidad distinta, en total a nivel país la base incluye 180.499 manzanas-entidades (151.912 manzanas y 28.587 entidades). La Región Metropolitana, que es donde se centra el análisis, cuenta con 51.070 manzanas-entidades (48.860 manzanas y 2.210 entidades). Cada manzana-entidad cuenta con una serie de variables sociodemográficas que describen

e intentan resumir la situación de los habitantes de aquella manzana-entidad. Algunas de estas variables son: Número total de personas, Total de hombres, Total de mujeres, Total de personas por rangos de edades, Total de personas migrantes, Total de personas que se consideran pertenecientes a un pueblo indígena u originario, Total de viviendas, Cantidad de hogares, variables descriptivas del tipo de vivienda (tipo, materialidad, agua, etc). Pero, de acuerdo a lo investigado la base del Censo no cuenta con la información cartográfica incluida, al menos en la versión que existe públicamente y puede ser descargada del siguiente sitio http://www.censo2017.cl/microdatos/.

Sin embargo, para solucionar esta situación existe la opción de utilizar entonces la cartografía recolectada por el PreCenso, que fue la etapa previa al Censo, y que tenía como objetivo "dividir con criterio técnico la totalidad del territorio nacional en sectores censales o sectores de empadronamiento censal, es decir, fracciones del territorio que pudieran constituir la carga de trabajo de un censista en el día del Censo" [37]. Pero, se debe utilizar con precaución debido a que "durante el 2017, se incorporaron las manzanas que mediante un proceso de actualización en terreno generan una fusión, división o la creación de una nueva entidad. La actualización de estas manzanas, para efectos de coherencia y concordancia espacial de la información, en algunos casos implicó modificaciones a otras capas" [39]. De hecho, se hizo el ejercicio de comparar cuántas manzanas-entidades coincidían a través del identificador de éstas en ambas bases dentro de la Región Metropolitana y sólo ocurrieron 47.752 coincidencias, quedando fuera 3.318 manzanas-entidades, lo que significa que un 6 % de las manzanas-entidades no existen ni fueron encontradas en la otra base.

Por tal motivo, se optó por utilizar como fuente la Plataforma de Datos Geoestadísticos del INE (https://ine-chile.maps.arcgis.com/apps/dashboards/e8292e6a13814b6b8bcfd3415ef4eb02), en donde se observó que se podía acceder a todas las manzanas-entidades actualizadas e incluso con ciertas variables sociodemográficas ya asociadas. Se hizo de nuevo el ejercicio de comparar la coincidencia de manzanas-entidades, y esta vez coincidieron 51.012, sólo quedando fuera 58 con respecto a la base del Censo. Esta base contaba con las siguientes variables (algunas heredadas del Censo y otras nuevas):

- Número total de personas
- Total de hombres
- Total de mujeres
- Total de personas mayores de 65 años
- Cantidad de hogares
- Total de viviendas
- Cantidad de hogares allegados
- Cantidad de núcleos allegados, hacinados e independientes
- Cantidad de viviendas según tipo (Casa o Departamento)
- Porcentaje de viviendas con hacinamiento (crítico, medio e indeterminado)

Además de contar como anteriormente se mencionó con información cartográfica de cada manzana-entidad, es decir, con el polígono que conformaba cada una de estas unidades censales. Es importante señalar para evitar confusiones que de aquí en adelante se preferirá utilizar el término manzana-entidad, aunque también se utilizará manzana como sinónimo, debido a que representan la gran mayoría de las manzanas-entidades.

4.3. Etapas

4.3.1. Métodos de normalización de direcciones

4.3.1.1. Método de la palabra Gatillante

Este método fue desarrollado por el alumno autor de esta memoria en conjunto con Juan Pablo Alvarado Glena. Básicamente el funcionamiento de este método de normalización fue construído sobre identificar desde qué palabra (uso de un diccionario de palabras para identificar esto) la dirección comenzaba a otorgar información adicional que no aportara a la estructura deseada: "Nombre de Calle + Número de Calle". En otras palabras, la dirección se mantendría sólo hasta antes de que se encuentre con una "palabra gatillante" que indique que ya no se le está aportando a la normalización o estructura deseada, sino que más bien se está agregando obstáculos al posterior proceso de geocodificación. Por lo tanto, el trabajo fue descubrir cuáles son esos gatillantes.

Pero primero, para poder llevar a cabo la normalización y posterior geocodificación de las direcciones normalizadas se comenzó preprocesando los datos, siguiendo los siguientes pasos:

1. Crear el rango de fechas de los casos que se quieren geocodificar.
2. Tomar desde la tabla de PERSONAS todos los casos cuyo formato de fecha de inicio de cuarentena estuviese correcto y dentro del rango requerido.
3. Eliminar todos los casos cuyo Tipo de Ingreso no fuese de interés geocodificar. Por ejemplo, si el objetivo fuera localizar sólo a pacientes contagiados con covid, deberíamos quedarnos sólo con los cuyo Tipo de Ingreso haya sido Confirmado o Probable, ya sea como caso Índice o Secundario.
4. Quedarse sólo con el identificador del paciente, la dirección y la comuna.
5. Eliminar duplicados.

Todo este proceso tiene como objetivo disminuir la cantidad de búsquedas innecesarias, ya que luego de este proceso sólo se cuenta con los casos que realmente se necesita geocodificar y se evita a la vez buscar a un paciente más de 1 vez, debido a que en la base de datos se puede encontrar con que un paciente aparezca más de 1 vez, no porque se haya duplicado, sino porque el paciente pudo haber requerido llevar a cabo como caso probable o confirmado una cuarentena en más de 1 ocasión debido a que probablemente se contagió nuevamente, por lo tanto se considera como un caso diferente, ya que si bien es el mismo paciente las fechas difieren. Luego, se comenzó con el proceso de normalización, cuyo objetivo como se dijo anteriormente era dejar el texto de las direcciones lo más directamente interpretable para los servicios de geocodificación, apuntando a que todas estuvieran normalizadas/estandarizadas

llegando a la estructura “Nombre de Calle + Número de Calle”. Para esto se siguieron los siguientes pasos:

1. Poner el texto de las direcciones completamente en mayúscula, para evitar tratar con la misma palabra más de una vez cuando sólo difiera en combinación de mayúsculas y minúsculas.

2. Eliminar carácteres y expresiones prescindibles (éstas fueron determinadas probando principalmente qué expresiones empeoraban la búsqueda, sobre todo pensando en geocodificadores que no fueron en su origen diseñados para buscar direcciones en español o expresiones usadas localmente en Chile). Ej: #, “Pasaje”, “Calle”, “Número”, “Avenida”, entre otras y sus variaciones y abreviaciones posibles.

3. Eliminar información adicional que desviase la estructura propuesta, es decir, borrar toda la información adicional que siguiese a ciertas expresiones como “Departamento”, “Block”, “Población”, “Edificio”, “Villa”, “Sin Número”, “Casa”, “Teléfono”, “Comuna”, entre otras y sus abreviaciones.

4. Eliminar espacios entre palabras mayores a 1, o si existiesen al inicio o final de la dirección, estos espacios pueden ocurrir con el proceso de eliminación de los pasos anteriores.

5. Varias direcciones contenían un código postal y comuna asociados, por lo tanto, si es que existiese, se conservaría el nombre de la comuna. Esto pues se consideraría como una aproximación más exacta de la comuna a la que pertenece la dirección, ya que no siempre coincidía con la entregada en el campo COMUNA, y a simple inspección correspondía con mayor certeza a la comuna de la dirección.

6. Eliminar de datos extras escritos dentro de paréntesis.

7. Eliminar códigos postales.

8. Añadir la comuna indicada en el campo COMUNA a las direcciones que no contasen anteriormente con la comuna proveniente del código postal y que el campo COMUNA no contuviese un valor nulo o inexistente.

9. Eliminar carácteres especiales o que hayan quedado como vestigios de las transformaciones anteriores. Ej: Comas, puntos y comas, puntos, signos de pregunta, y casos indeseados que hayan pasado los pasos anteriores de eliminación.

10. Eliminar espacios adicionales.

Por lo tanto, con ese proceso podríamos llevar una dirección del estilo “CALLE VILLARRICA 7980 (CASA AMARILLA, CON UN AUTO BLANCO AFUERA), 8780000 LA GRANJA, REGION SANTIAGO METROPOLITAN, CHILE” a su forma normalizada “VILLARRICA 7980”, y sumarle la COMUNA dada por la base si es que no se contaba con la comuna heredada del código postal, lo que sería más fácilmente interpretable para las APIs.

4.3.1.2. Método Clasificador Reglas Lógicas

Este es el modelo desarrollado en [17]. De todos modos, para aplicarlo a la base de datos y tipos de datos con los cuales se contaba, se desarrolló una versión con ciertas mejoras y adaptaciones. Cabe destacar que para la construcción del código que permitió aplicar el método se contó solamente con las instrucciones relatadas en [17], ya que el código para llevarlo a la práctica fue 100 % elaboración propia, pero basado en gran parte en las indicaciones contenidas en la memoria refereciada.

Se comenzó efectuando primero el preprocesamiento de datos propuesto en el método anterior. Luego se siguió con las indicaciones, las adaptaciones y mejoras agregadas detalladas a continuación.

El modelo propuesto es desarrollado con el lenguaje de programación Python. Utiliza un conjunto de sentencias condicionales if, else y elif para verificar las reglas lógicas y se recorren las cadenas de texto utilizando ciclos for.

Se tiene como objetivo que el modelo separe las direcciones en la siguiente estructura: Nombre de calle, número de calle. Como supuesto base se asume que las direcciones poseen un número principal. La lógica del clasificador se basa en encontrar el número principal dentro de las diferentes direcciones. En caso de no existir un número principal bien definido, la dirección se etiqueta como inválida y no se normaliza.

El primer paso es convertir toda la columna de direcciones a string, para evitar futuros problemas con el tratamiento del texto. Luego eliminar ciertas palabras que podían contaminar la identificación del número de la calle tales como comas, puntos y ciertas expresiones que anteceden al número como CASA, Nº, NÚMERO y sus variantes incluyéndolas con y sin tildes. Esto se hace para que un caso que se relatará a continuación que habla de si "2 números quedaban juntos" funcione de buena manera.

A continuación, el clasificador crea un arreglo donde cada elemento es el es cada una de las palabras y números sin considerar los espacios en blanco entre ellos. Luego a partir de aquel arreglo se crean dos nuevos donde uno contiene sólo las palabras y el otro sólo los números, además en otros 2 arreglos más se guardan las posiciones que ocupaban inicialmente en el primer arreglo.

Como regla general se asigna al campo nombre de calle, todas las palabras existentes hasta el primer número que aparezca, si es que existe un solo número, para esto es necesario fijarse en la posición que ocupa el primer y único número de la dirección y se guarda como calle la unión de todos los elementos del primer arreglo con los respectivos espacios que debiesen existir entre cada uno.

Si hay más de un número, entonces, puede ser parte de un número secundario que indica el departamento, oficina o block, o puede que el primer número sea parte del nombre de la calle. En el caso de que existan dos números seguidos, se considera el primero como parte del nombre de calle y el segundo como principal (casos "pasaje 4 722" ó "calle 2 5859"). Si no hay dos números seguidos se considera el primero como número principal.

Para realizar una correcta clasificación, se trabaja con dos diccionarios: un diccionario con palabras que anteceden el nombre de calle, llamado Pre_calle y otro con palabras que hacen referencia al tipo de edificación o lugar, llamado Tipo. A continuación, se detalla el contenido de cada diccionario.

Pre_Calle: “PASAJE”, “PJE.”, “PJE”, “PSJE.”, “PSJE”, “PJ”, “PJ.”, “PJ/”, “PJE/”, “PSJ/”, “AVENIDA”, “AVDA”, “AVDA.”, “AV.”, “AV”, “CALLE”, “C/”

Tipo: “BLOCK”, “BLOCK.”, “BL”, “BL.”, “B/”, “BL/”, “BLCK”, “B0”, “B1”, “B2”, “B3”, “B4”, “B5”, “B6”, “B7”, “B8”, “B9”, “DPTO”, “DPTO.”, “DEP”, “DEP.”, “DPT”, “DPT.”, “DP”, “DP.”, “DEPTO”, “DEPTO.”, “D/”, “DEP/”, “DEPTO/”, “D-”, “VILLA”, “VLLA”, “V.”, “V”, “V/”, “VI”, “VI.”, “POB”, “POB.”, “POBLACION”, “POBLACIÓN”, “P/”, “PB”, “PB.”, “PB/”, “POBL”, “POBL.”, “POBL/”, “POBLACIN”, “TORRE”, “COND”, “COND.”, “CONDOMINIO”, “EDIF”, “E/”, “EDIF.”, “E.”, “E”, “LOTE”, “LOTEO”.

Estos diccionarios se utilizan para crear las sentencias condicionales, ya que entregan información semántica de las direcciones.

Después el clasificador identifica si una dirección es válida o no. Para esto, en primer lugar, verifica si existe al menos un número asignado a Número de calle y, en caso de que no exista, se etiqueta como dirección inválida. En segundo lugar, se verifica si la dirección posee el indicativo “S/N”, el cual indica que no existe numeración principal para esa ubicación. En tercer lugar, se verifica que en el nombre de calle no posea indicativos del tipo de vivienda, ya que esto implica que el número identificado como principal puede deberse a una numeración interna del tipo de edificación, es decir si el nombre de calle posee palabras como “BLOCK” o “DPTO” se etiqueta como dirección inválida.

Finalmente se corrigen algunas palabras y abreviaciones usuales y se eliminan las palabras contenidas Pre_Calle si existiesen en la cadena de texto resultante sólo cuando esa palabra no sea relevante, por ejemplo en casos como PASAJE SANTA ANA 65, pero no en casos donde la palabra sea parte del nombre de la calle como en el caso de PASAJE 3 518, donde el nombre de la calle es PASAJE 3, ya que si se quitara PASAJE quedaría el nombre 3, lo cual podría llevar a confusiones.

4.3.2. Geocodificación de Direcciones Normalizadas

Con la dirección ya normalizada por alguno de los 2 métodos, se puede llevar a cabo la geocodificación de la dirección y con ello la localización del domicilio del paciente, siguiendo los siguientes pasos:

1. Pasar las direcciones normalizadas a las APIs de geocodificación, prefiriendo la búsqueda por componentes para mayor precisión y menos errores. Lo que se hizo entonces fue pasar primero la búsqueda por la API Nominatim de OpenStreetMap (Se prefirío a Nominatim es un poco más completa en funciones que Photon, por ejemplo, posee búsqueda estructurada, es decir, se puede ir otorgando las distintas partes de una dirección por separado y así evitar la imprecisión, mayores detalles pueden ser revisados en [40]) para evitar cargos por la gran cantidad de direcciones a buscar, las direcciones que no fuesen encontradas por la API se les pasaría a la API de Google Maps.

2. Se considera que una dirección fue encontrada incorrectamente cuando la dirección era encontrada fuera de la región metropolitana de Chile, corroborando por ejemplo que la ubicación del mapa encontrada no quedara fuera de ciertos límites latitud, longitud coherentes.

3. Para recoger los resultados de la mejor forma se creó 4 campos: latitud, longitud, motor por el cual fue encontrado (OSM o Google Maps) y el status final de la búsqueda ("Encontrado correctamente" o "No encontrado")

4.3.3. Asignación de paciente a Manzana Censal

Con la geocodificación lograda el siguiente paso es la asignación del paciente a la manzana censal correspondiente a su domicilio, para así poder asociar variables geo-demográficas del entorno al paciente. Para esto se pensó en 2 formas:

1. **Asignación por distancia del paciente a la manzana más cercana**: Este método consistía en asociar al paciente a la manzana que estuviese más cercana. Para lograr esto se tenía que definir primero con respecto a qué punto de la manzana se haría el cálculo de cercanía. Teniendo en cuenta que se contaba con el polígono de la manzana, es decir, cada uno de los puntos de los bordes que componían la figura, se decidió utilizar el centroide del polígono, lo cual no es tan complicado de calcular ya que existe en Python una función para aquello. Con los centroides calculados lo siguiente sería calcular la distancia del paciente a los centroides de todas las manzanas y asignarle el que estuviese más cercano. Para calcular la distancia entre puntos se decidió utilizar la Fórmula del Semiverseno, la cual permite calcular la distancia entre 2 puntos definidos por su latitud y longitud sobre una superficie esférica, lo cual es útil si se considera una aproximación esférica a la forma de la Tierra. Este cálculo también puede ser llevado a cabo por una función de Python, aunque si se desea también se puede programar. Sin embargo, podría surgir la duda de por qué se utiliza esto y no se usa simplemente la distancia euclidiana entre las longitudes y latitudes de cada punto. Esto se explica por la esfericidad de la Tierra, las distancias se miden por sobre la superficie de la Tierra y no pasando por el interior de esta. Además de que la distancia euclidiana no nos permitiría tener noción del valor de la distancia. A continuación se describe la Fórmula del Semiverseno:

$$\text{semiversin}\left(\frac{d}{R}\right) = \text{semiversin}(\varphi_1 - \varphi_2) + \cos(\varphi_2)\cos(\varphi_1)\ \text{semiversin}(\lambda_2 - \lambda_1) \quad (4.1)$$

donde:

d es la distancia entre dos puntos (sobre un círculo máximo de la esfera)

R es el radio de la esfera, que en este caso sería el radio medio de la Tierra

φ_1 es la latitud del punto 1 en radianes

φ_2 es la latitud del punto 2 en radianes

λ_1 es la longitud del punto 1 en radianes

λ_2 es la longitud del punto 2 en radianes

Además, se debe tener en cuenta que:

$$\text{semiversin}(\theta) = \text{sen}^2\left(\frac{\theta}{2}\right) \tag{4.2}$$

Y, si se llama "h" al semiversin $\left(\frac{d}{R}\right)$ y se reemplaza en 4.1, entonces si se aplica la inversa del semiverseno se podría obtener el valor de "d" de la siguiente manera:

$$d = R\ \text{semiversin}^{-1}(h) = 2R \arcsin\left(\sqrt{h}\right) \tag{4.3}$$

Por consiguiente, si "h" tomara el valor de la parte derecha de 4.1 y si se tiene en cuenta 4.2, no existiría problema en calcular el valor de "d" ya que se cuenta con todo lo necesario si también se considera 4.2. Por lo tanto, si se utiliza 4.3 se tendría lista la forma de calcular la distancia entre 2 puntos sobre la Tierra. Hay que tener en consideración que la magnitud y unidad de medida de la distancia que dará como resultado dependerá de en que unidad se otorgue el radio de la Tierra "R", en otras palabras si se otorga el radio en kilómetros el resultado también estará en kilómetros.

2. **Asignación por encontrarse al interior del polígono formado por la manzana**: Este método consiste en verificar si las coordenadas que definen la localización del domicilio del paciente se encuentra dentro o no de alguno de los polígonos de las manzanas existentes. Esto también puede ser logrado con una función de Python, la cual entregará un valor booleano acerca de si se encuentra o no el paciente dentro de alguno de los polígonos de las manzanas.

Ambos métodos permiten una asignación del paciente a una manzana, y ambos cuentan con sus ventajas y desventajas. El primero no deja a ningún paciente sin asignar, a todos los asigna a la manzana cuyo centroide esté más cercano a ellos, pero esto puede tener a la vez sus consecuencias negativas, debido a que la manzana más cercana podría estar lo suficientemente lejana para de ningún modo corresponder el paciente a ella, esto se puede dar en el caso en que la geocodificación del paciente falló y fue localizado en una posición errónea donde ninguna manzana estuviese cerca, por lo cual si se permiten estos casos se estaría asignando una persona a una manzana a la cual no pertenece, para solucionar esto se podría poner un límite entre la distancia del paciente y la manzana para ser asignada a ella, aquí juega un punto a favor el saber el valor de las distancias, ya que por ejemplo se podría poner la condición de que el paciente no fuese asignado a ninguna manzana si la más cercana que se encuentre estuviese a más de 100 metros de él, y con eso se podría poner atención a aquellos casos que no fueron asignados y revisar su proceso de geocodificación.

Existe otra razón por la cual pueda surgir el caso anterior, la cual sería que las manzanas que se estén considerando no sean todas las que cubran el rango de búsqueda querido/requerido o que la manzana no haya existido en el 2017. Otro caso desfavorable sería el hipotético caso en el que se tuviese alrededor de un paciente una manzana muy grande y otra muy pequeña, el método de asignación a manzana más cercana asignará por lo general a la manzana más pequeña debido a que su centroide estará más cercano al paciente, pese a que el paciente pueda encontrarse al interior del polígono de la más grande. También hay que considerar que un centroide no siempre está dentro de los límites del polígono de la manzana, que podría afectar negativamente en la asignación de algunos casos.

Con respecto al segundo, si bien parece una verificación simple, debido a que sólo se debe revisar si el paciente está dentro o no del polígono de alguna manzana, su desventaja radica en que por construcción de datos las manzanas no incluyen las calles, en otras palabras entre cada manzana existe un espacio vacío que la separa de las otras, lo cual puede implicar que si un paciente fue localizado muy cercano a la calle o en la calle, este quedará fuera de la manzana ya que no estará dentro de sus límites, pese a que lógicamente y racionalmente pertenezca a ella, por lo cual se terminaría con una cantidad no menor de personas no asignadas debido a que una persona sea localizada por los servicios de geocodificación en la calle no es algo poco común, sobre todo porque existen casas que están en los bordes de las manzanas, o incluso se puede dar debido a que en el proceso de geocodificación la calle fue encontrada, pero no el número de la casa, por lo cual el localizador posicionaría al paciente en la calle.

En conclusión, resultaría útil usar ambos métodos aprovechando sus ventajas y desventajas. Por ejemplo, primero comenzar asignando a los pacientes con el segundo método y así nos aseguramos que todos los asignados efectivamente pertenezcan a la manzana, y los que no fueron asignados procesarlos con el primer método, el cual los asignará a la manzana que tengan más cercana y así se solucionaría el problema con los pacientes localizados muy cercanos a la calle, sin olvidar poner el límite pertinente para que no sean asignados a manzanas lejanas. A continuación, en la Figura 4.3 se muestra una ejemplificación los casos desfavorables que fueron mencionados en los párrafos anteriores. Ahí el punto rojo simboliza la ubicación encontrada del paciente luego de geocodificar su domicilio, los puntos amarillos son los centroides de cada manzana-entidad que lo rodea y las líneas azules simbolizarían las distancias calculadas.

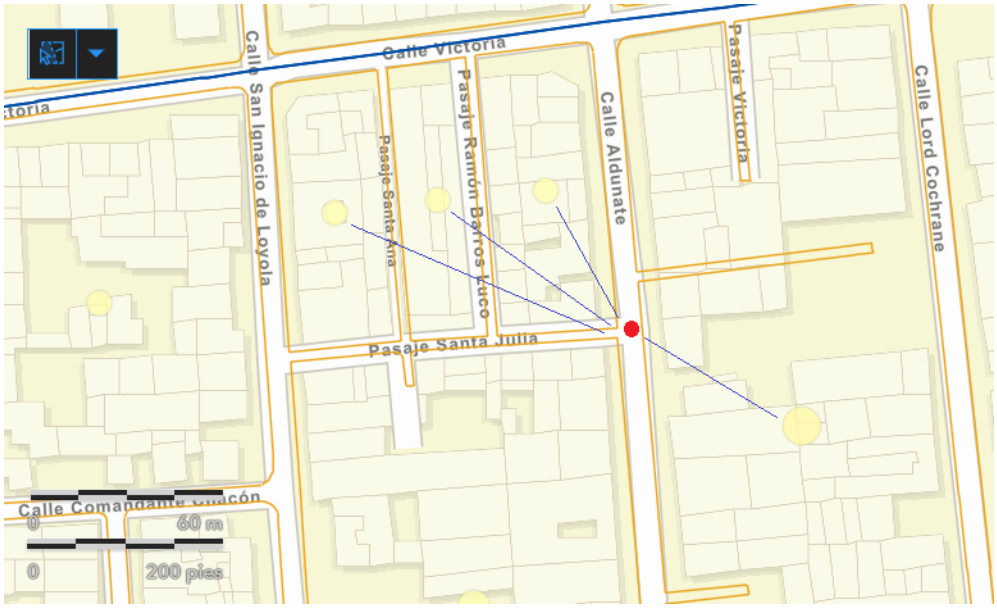


Figura 4.3: Ejemplo de caso desfavorable

4.3.4. Visualización de Casos

Esto ya fue de cierta forma relatado en el capítulo 1 en la sección Contexto de la Investigación, debido a que formó parte del entregable que se hizo hacia el SSMSO. Se desarrolló una

aplicación en Streamlit a través de Python que permitiera estandarizar y geocodificar de forma eficiente y eficaz las direcciones de los pacientes que estuviesen integrados a la plataforma de seguimiento, además de lograr la visualización espacial y temporal de los casos.

En ese entregable se aplicó como método de normalización de direcciones el Método de la Palabra Gatillante, ya que fue el que primero se desarrolló, pero de todas formas, tanto el Método de la Palabra Gatillante como el Método Clasificador Reglas Lógicas podían ser aplicados. Al tener los casos de covid históricos geocodificados de antemano y asociados a su fecha de ocurrencia, se podía agregar un nuevo caso a los anteriores teniendo como fecha de ocurrencia la otorgada por el usuario. Los casos que ya estaban geocodificados estaban asociados a su fecha de inicio de cuarentena debido a que era el campo más íntegro y más representativo de una posible fecha de inicio del contagio.

Con lo anterior listo, se podían ver todos los casos que rodeaban a este nuevo caso tanto temporal como espacialmente, permitiendo así a través de la visualización poder analizar también la existencia de brotes focalizados en algún área de la región metropolitana, por lo que para ello se graficó con un gradiente de colores la diferencia de días en la ocurrencia de cada uno de los casos ya geocodificados con respecto a la fecha ingresada por el usuario. Lo mencionado puede ser visto aplicado en la Figura 4.4.

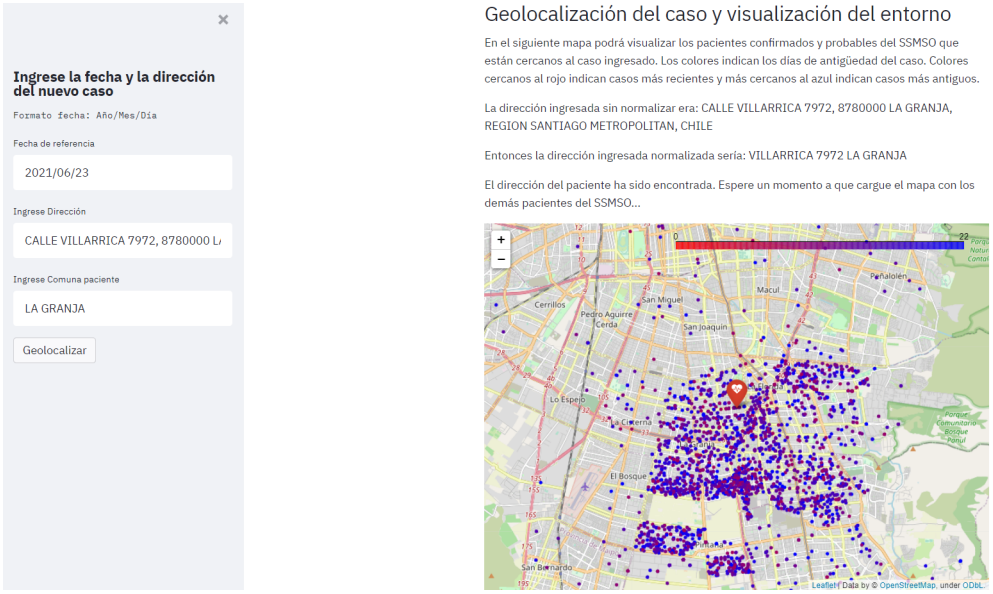


Figura 4.4: Caso de ejemplo geocodificado y su entorno

Capítulo 5

Evaluación y Resultados

Para esta memoria se quiso llevar a cabo 2 "experimentos":

1. **Probar, evaluar y comparar el desempeño de los distintos métodos de normalización de direcciones descritos en esta memoria y luego comparar la búsqueda de las direcciones resultantes utilizando la API de Geocodificación de Google Maps y la de Open Street Map**: El objetivo de este experimento era comparar el desempeño en la correción y normalización de direcciones de distintos métodos y verificar cuál entregase los mejores resultados, utilizando tanto la API de Google Maps como la de Open Street Maps. Las 3 formas a comparar serían: "Método de la Palabra Gatillante" de Juan Pablo Alvarado y el alumno autor de esta memoria, el "Método de Clasificación Reglas Lógicas" de Daniel Ponce Maripangui adaptado y por último una búsqueda de las direcciones tal y cual vienen de la base de datos.

   Entonces, para probar cada uno de estos métodos se decidió aplicarlos a las primeras 100 direcciones de la hoja de DIRECCIONES, sin importar el Tipo de Ingreso que tuviesen. Se decidió utilizar las 100 primeras por simplicidad para el proceso, para este experimento sólo se quiso tomar una muestra, es verdad que tal vez tomar una muestra aleatoria puede ser más beneficioso para evitar sesgo, en este caso se ignoró ese posible sesgo teniendo en cuenta que hay baja probabilidad que las 100 primeras entradas de direcciones estén altamente relacionadas, también se optó por las primeras 100 por un sentido de facilidad en la reproducibilidad del experimento sobre todo para verificar los resultados. Además es necesario señalar que se prefirió hacer una búsqueda en las APIs de geocodificación a través de la separación por componentes, es decir, entregando un campo que especificara el nombre y número de la calle y otro con la ciudad/localidad, para lograr una búsqueda más precisa y focalizada y que los sistemas de búsqueda interpretaran de forma más correcta la dirección, ya que se podría dar el caso por ejemplo de que un nombre de calle coincida con el nombre de una localidad/ciudad y podría otorgar inexactitud o confusión, así entonces se asegura que la búsqueda de la calle y número sean buscadas sólo dentro de la localidad entregada en la otra componente. Esto sólo era lograble correctamente en los métodos que aplicasen normalización y corrección, debido a que generalmente las direcciones sin tratar el campo de DIRECCIÓN contendría información extra al nombre y número de calle, que si se ingresara dentro de esa componente probablemente llevaría a mayores errores, debido a que se estaría buscando por nombre de calle algo que pudiese contener una localidad en su interior o datos "basura" que nada tengan que ver con la componente señalada, provocando que los servicios no

encuentren la componente ya que no existiría, por lo cual en el caso de direcciones no procesadas, se entregaría a las APIs la dirección completa del campo de DIRECCIÓN sumada al campo de COMUNA sin separar en componentes.

Otro punto importante a aclarar en este experimento era la forma en la que se lograba verificar que una búsqueda fuera exitosa o correcta, lo cual definiría la forma de comparar los desempeños y verificar los resultados. Para esto se verificaron 3 puntos principales: "La dirección que fue buscada fuese válida", "la dirección fuese encontrada" y "el resultado de la búsqueda fuese similar o cercano a una búsqueda manual por un humano".

Para verificar el primer punto, se revisaba que la dirección procesada tuviese al menos un nombre de calle y un número de calle. Esta verificación estaba ya hecha en el caso del método de Método Clasificador Reglas Lógicas, ya que al funcionar como un clasificador de texto parte del procesamiento era identificar la existencia de una cadena de string que pudiese ser considerado como nombre de calle y un número que pudiese ser considerado como el número de la calle, y si alguno de los 2 no existiese se consideraba inválida la dirección. Para el caso del otro método, como no funcionaba como un clasificador de texto ni tampoco había una identificación de las partes de la dirección sólo se podía realizar con una revisión manual, lo mismo para el caso en que no se realizaba normalización. Como son sólo 100 direcciones no resultaba tan complicado revisar una por una las direcciones procesadas, por lo cual se procedió de esa manera.

Para verificar el segundo punto sólo bastaba verificar que luego de haber pasado por el proceso de geocodificación un par latitud, longitud hubiese sido entregado como resultado, si no, se consideraba como no encontrada.

Para el último punto lo que se hizo fue primero buscar manualmente las 100 direcciones utilizando los servicios de búsqueda de Google Maps y Open Street Map, priorizando en la búsqueda los resultados entregados por Google Maps y verificando que la posición dada por el buscador estuviese correcta, es decir, la ubicación en el mapa correspondiese con seguridad a la dirección buscada con el razonamiento de un humano, debido a que el objetivo de esta memoria es encontrar un método en el cual un humano no esté buscando la posición del paciente manualmente 1 por 1 y asimilarse al razonamiento e interpretación inteligente que haría un humano, por lo cual esas posiciones encontradas se tomarían como referencia, o en otras palabras, "lo que deberían entregar los métodos". Y, la forma de comparar qué tan cerca estuviesen de la referencia se simplificó a comparar la distancia euclidiana del punto encontrado por el método hacia el punto de referencia. Se prefirió aquí el uso de la euclidiana antes de la del semiverseno por las siguientes razones: simplicidad y rapidez en la forma de calcular, y además que considerando que las distancias entre puntos deberían ser menores a las distancias hacia los manzanas debido a que se está haciendo la búsqueda y comparación sobre la misma dirección sólo tratada de distinta manera, por lo tanto no afectaría significativamente la curvatura de la Tierra en las distancias entre puntos, además no se necesita la noción de distancia en metros o kilómetros debido a que sólo se quiere saber quién está más cerca y no cuánto más cerca específicamente. Entonces, se razonaría de la siguiente forma: mientras más

cercano esté el punto a lo determinado manualmente por un humano mejor habría sido el desempeño. Los resultados de este experimento se muestran en la Tabla 5.1:

Tabla 5.1: Resultados del Primer Experimento

Fuente	Ítem de desempeño evaluado	Método Clasificador Reglas Lógicas	Método de la palabra Gatillante	Sin Estandarizar
OSM	Encontradas por el método	70	70	23
	Encontradas por el método y que eran válidas luego de procesadas	66 (94.28 %)	57 (81.42 %)	14 (60.86 %)
	Distancia promedio hacia la ubicación encontrada manualmente	0.0292	0.0286	0.0434
	Distancia mínima hacia la ubicación encontrada manualmente	0.00003	0.00003	0.00003
	Distancia máxima hacia la ubicación encontrada manualmente	0.22	0.22	0.205
Google Maps	Encontradas por el método	85	100	97
	Encontradas por el método y que eran válidas luego de procesadas	78 (91.76 %)	82 (82 %)	81 (83.5 %)
	Distancia promedio hacia la ubicación encontrada manualmente	0.00901	0.01047	1.542
	Distancia mínima hacia la ubicación encontrada manualmente	0.0000001	0.0000001	0.0000001
	Distancia máxima hacia la ubicación encontrada manualmente	0.22	0.22	62.187

Nota: Para interpretar la tabla de mejor manera es que en la tabla el porcentaje que aparece entre paréntesis indica a qué porcentaje equivale el número al costado con respecto a la categoría superior

Lo primero que es necesario notar es que al inspeccionar las 100 direcciones 1 a 1 manualmente por un humano, sólo 88 eran válidas, las demás no lo eran debido a que no contaban con un número de calle, no contaban con un nombre de calle o ninguna de ambas. Y, al buscar estas direcciones que eran válidas sólo 82 pudieron ser encontradas correctamente, debido a que esas 6 de diferencia si bien contaban con algo que pudiese ser considerado calle y número, no eran precisas, es decir, si se buscaban no se podía encontrar una ubicación correcta ni única, probablemente estaban mal escritas o incompletas, o no escritas de la misma forma en la que aparecen en los mapas. Es importante tener en cuenta esto, debido a que cualquier resultado mejor que este sería sospechoso, ya que lo más probable es que si bien esté entregando un resultado este esté erróneo.

El primer resultado notable es que el método de Método Clasificador Reglas Lógicas, de las direcciones encontradas un gran porcentaje de ellas tenían sentido, es decir, eran válidas debido a que contaban con una calle y un número con sentido (aquí se excluyen resultados con nombre o números de calle sin sentido al razonamiento de un humano), es decir, la gran mayoría de las encontradas por los servicios de geocodificación partieron de la base de que la dirección que se les entregó tenía sentido, esto es una buena consecuencia directa de la construcción del método debido a que no son ingresadas direcciones inválidas, es decir, que no poseen calle ni número, en el geocodificador.

En cambio, el otro método logró un porcentaje de direcciones encontradas válidas menor, pero mayor al 80 % Esto es principalmente consecuencia de que no se verificaba que la dirección contara con calle ni número, por lo tanto si bien logró una mayor cantidad de encontradas varias de esas encontradas no tenían sentido (recordar que cualquier valor sobre las 82 es un potencial objeto de dudas).

El desempeño de la búsqueda sin utilizar método de normalización tuvo no tan buenos resultados, sobre todo al emprender la búsqueda de la dirección con Open Street Map. Si bien con Google Maps varias fueron encontradas, varias de ellas desde su origen no eran realizables, esto pues la dirección desde sus inicios en al menos 18 casos no estaba correcta. El desempeño de estos resultados encontrados se aclarará en los siguientes

puntos.

El segundo resultado notable con respecto a la búsqueda es que con Open Street Map los resultados estuvieron en promedio más cercanos a los determinados por un humano con el Método de la palabra Gatillante, pero al utilizar Google Maps, los resultados se invertían, ya que los resultados entregados por el método de clasificación de texto estaba más cercano a la referencia. Cabe decir, que las comparaciones de distancias entre puntos sólo se pudieron hacer con respecto a las direcciones que habían sido encontradas manualmente, es decir, con las 82, ya que en las demás no existe certeza de un resultado correcto o con sentido. Ambos de todas formas lograron un mejor desempeño con respecto a utilizar las direcciones sin tratar, ya que ambos se encuentran más cercanos utilizando tanto Google Maps como Open Street Map.

El tercer resultado notable es que el valor minimo de la distancia a la dirección más cercana a la referencia en todos los casos fue el mismo, debido probablemente a que esa dirección no necesitaba una normalización ni corrección, ya que ya estaba correcta y lista para ser interpretada por los geocodificadores, por lo mismo, todos los métodos probablemente la dejaron tal cual estaba.

El cuarto resultado notable es que el valor de la distancia máxima hacia la referencia, en ambos métodos coincidió, pero con respecto a la búsqueda a través de Google Maps de la dirección sin tratar la máxima distancia hacia la referencia fue un valor lejano de lo aceptable debido a que si bien la dirección fue encontrada, fue asignado a un punto sin sentido y que no correspondía a la ubicación real, esto fomentado por la existencia de información basura.

Finalmente, un quinto resultado notable es que la geocodificación de Google Maps logró mejores resultados que la de Open Street Map con respecto a ambos métodos, debido a su "mayor inteligencia", pero nuevamente aquí entra en acción que el servicio de Google Maps si bien es mucho más rápido y certero, la cantidad de búsqedas gratuitas es limitada y se puede incurrir en costos si se supera la cuota gratuita, además es necesario asociar el servicio con una tarjeta visa o mastercard para que funcione. En cambio, el otro es más lento, menos preciso, pero en ningún momento se tiene el riesgo de incurrir en un cobro, ni es necesario ingresar una tarjeta, lo cual lo hace ideal para el uso de más de una persona, ya que el de Google Maps debería ser esencialmente personal para no llevarse sorpresas de cobros a la persona de la tarjeta asociada. También es necesario señalar que este análisis está pensado cuando se hacen búsquedas de muchas direcciones, y que superen la cuota gratuita de Google Maps, ya que de no ser el caso no afecta demasiado.

En resumen, ambos métodos de normalización demostraron un mejor desempeño con respecto a no realizar ningún tipo de procesamiento a las direcciones, cada uno cuenta con sus ventajas y desventajas, pero a juicio del memorista, el que da un resultado más preciso y con menos errores o falsos encontrados para este caso sería utilizar en conjunto el método de clasificación de texto mejorado y adaptado y la API de Google Maps.

2. **Asignación de pacientes a manzanas-entidades correspondientes, asociar va-**

riables sociodemográficas interesantes y analizar relación entre el hacinamiento con la ocurrencia de casos: En la base de PERSONAS se cuenta con casos que van desde el 22 de Abril de 2020 hasta el 28 de Junio de 2021, es decir, aproximadamente 1 año y 2 meses. El objetivo era asignar a los casos confirmados en covid (de acuerdo al Manual Operativo para las acciones de Trazabilidad y Aislamiento del Minsal [10] y que además se encuentra en el primer capítulo de esta memoria, cuentan como confirmados los casos que hayan sido declarados como Confirmados o Probables), localizarlos y asignarlos a la manzana-entidad que les correspondía para así poder obtener variables sociodemográficas asociadas al entorno del paciente. La variable que se optó por analizar fue la de hacinamiento, que si bien existía una construida en la base de datos, no era muy tratable y fácil de usar para estos fines, por lo cual se decidió definir el "hacinamiento" dentro de una manzana como la cantidad total de personas que viven en la manzana dividida por la cantidad de viviendas existentes (no se divide por el área de las manzanas debido a que esto se vería distorcionado por la coexistencia de tanto casas que es una sola vivienda como edificios con muchas viviendas) estas 2 variables existen dentro de la base del Censo.

Es importante recordar que la base de datos del Censo cuenta con lo recopilado el año 2017, a lo cual al día de hoy pudo haber cambiado bastante, ya que ya han pasado más de 4 años, por lo cual puede darse la situación de que en cada manzana vivan más o menos personas que en 2017, o que viviendas que existían en ese momento ya no existan o que se hayan construido nuevas, pero se considera una aproximación y lo más cercano que podemos en datos tener hoy a la realidad, ya que no ha sido llevado a cabo otro Censo. Luego al tener esta variable de hacinamiento se esperaba comparar con la ocurrencia de casos, y así poder determinar si por ejemplo, un mayor hacinamiento explica una mayor ocurrencia de casos.

El primer paso para lograr este experimento es tomar a los pacientes que aparecieran en la hoja de PERSONAS con TIPOINGRESO Confirmado y Probables (tanto Índices como Secundarios). Si un paciente individualizado por su identificador único IDPER apareciera en más de 1 ocasión en la hoja, se consideraría como un nuevo caso si es que la fecha (FECHAINICIOCUARENTENA) fuese distinta. Luego se buscaría cada uno de estos pacientes en la hoja de DIRECCIONES y se les asignaría la dirección de su domicilio correspondiente y si se diera el caso en que la dirección del paciente no existiese, ese paciente no sería considerado, ya que un paciente sin dirección no permitiría la ejecución de los siguientes pasos. Luego, las direcciones de los domicilios de los pacientes serían procesadas bajo alguno de los métodos de normalización de direcciones (ya que ayudan a la obtención de mejores y buenos resultados según lo visto en el experimento anterior). En este caso se prefirió utilizar el método mejorado y adaptado de clasificación de texto debido a que entregaba mejores resultados y menor cantidad de encontrados incorrectos/falsos según lo visto en el Expeimento 1, pero en este caso en la parte de la geocodificación de las direcciones procesadas se preferiría utilizar un trabajo conjunto de ambas APIs, priorizando la búsqueda por Open Street Map (como fue mencionado inicialmente en la memoria) para evitar la generación de costos, es decir, se buscaría la totalidad de las direcciones a través de Nominatim de Open Street Map y las no encontradas buscarlas nuevamente por la API de geocodificación de Google Maps, esto es necesario hacer debido a que en este experimento se está trabajando con una cantidad relevante de direcciones a diferencia del experimento anterior donde sólo eran 100.

A continuación, luego de tener la localización de los domicilios de los pacientes queda asignarlos a la manzana-entidad correspondiente, lo cual se realiza de la forma relatada en el capítulo anterior, es decir, los pacientes cuya localización se encontrase al interior de un polígono de alguna de las manzana-entidades sería asignado a aquella, y en el caso de que no se encontrase al interior de los límites de ninguna debido a que el paciente probablemente quedó localizado en el espacio entre manzanas (calles sobre todo), sería asignado a la manzana-entidad más cercana a su ubicación, es decir, como fue definido en el capítulo anterior, la manzana cuyo centroide estuviese más cercano, aunque a diferencia de lo planteado en el capítulo anterior se decidió por no poner un límite a la distancia de asignación, debido a que al realizar el ejercicio, se verificó que las distancias de las manzanas más cercanas a estos casos no asignados no eran tan grandes, además de considerar que si bien una manzana-entidad muchas veces corresponde a lo que sería una cuadra, no siempre es así, a veces es un conjunto de ellas, o un espacio más grande de lo que se tiene como referencia de una cuadra, considerando que las cuadras no tienen un tamaño definido y fijo, al menos en Chile.

Ya teniendo cada caso asignado a su manzana correspondiente, se podía proceder a realizar el análisis de la relación planteada en un comienzo. Para esto era necesario, hacer el análisis por manzanas, por lo cual ese sería el nuevo nivel de agregación. Lo que se hizo fue agrupar por cada una de las manzanas hacer un conteo de la cantidad de casos que ocurrieron dentro de cada una de ellas y traer del censo las variables de cantidad de viviendas y la cantidad de personas para cada una también. Lo siguiente fue la construcción de la variable que definiera "el hacinamiento" y la ocurrencia de casos. Para el "hacinamiento" bastaba con dividir la variable del total de personas por la cantidad de viviendas para cada manzana, así tendríamos la cantidad de personas por vivienda en promedio para la manzana, de aquí se excluyó todas las manzanas-entidades cuya cantidad de personas hubiese sido 0, debido a que no haría sentido considerando que hay casos asignados a aquellas manzanas (puede ser que en el 2017 no hayan tenido moradores, o las viviendas aún no estuviesen construidas). Como lo que se quería comprobar era un efecto promedio y no puntual, se decidió agrupar las manzanas para formar rangos, es decir, por ejemplo, si la variable de "hacinamiento" estuviese entre los valores de 0 a 15 personas/vivienda, si se agrupara por grupos de 0.1, el primer conjunto representado serían manzanas cuya cantidad de personas por vivienda fuese entre 0 y 0.1 personas/vivienda, luego el siguiente sería de 0.1 a 0.2 personas/vivienda y así hasta llegar al máximo. Los rangos que no contuviesen a ninguna manzana simplemente no se representarían debido a que si se agregasen no tendría sentido ya que no existen casos asociados.

Las últimas decisiones anteriores están justificados en que si no se agrupa las manzanas habrán demasiados puntos y en el caso en que 2 manzanas o más compartieran la misma cantidad de personas por vivienda se tendría 2 valores o más que querrían explicar la ocurrencia de casos, entonces habría que tomar una decisión, cuál de todos esos valores tomaríamos como el que representara la ocurrencia de casos, además tampoco se cumpliría que aquello representara una función debido a que para el valor de una variable independiente (que sería la variable de personas por vivienda), habrían más de 2 resultados posibles para la variable dependiente (que sería la ocurrencia de casos), aunque se podría promediar todos aquellos valores, pero se prefirió agrupar por rangos las

manzanas debido a que el valor del rango también representaría el valor del promedio de las manzanas en el rango y mitigaría las variaciones que podrían ocurrir entre 2 puntos contiguos debido a un error en la asignación de la manzana al caso o el proceso de geocodificación o incluso las diferencias a la realidad actual debido a la antigüedad del censo.

Lo último para proceder al análisis sería cómo se define correctamente "la ocurrencia de casos". Si no se reflexionara demasiado se pensaría que bastaría con contabilizar la cantidad de casos totales que ocurrieron para cada agrupación de manzana, pero el problema de esto es que no se puede ignorar de que en las manzanas que cuenten con mayor población existe una tendencia a ocurrir mayor cantidad de casos en ella. Para visualizar esto, imaginemos que existe una manzana con sólo 2 personas, y otra con 1000 personas, y en ambas ocurrieron 2 casos, ¿cómo las comparamos justamente? Porque la ocurrencia de 2 casos en una manzana de 2 es un caso muy diferente a 2 casos en una manzana donde viven 1000 personas. Por lo cual se decidió crear una nueva variable que viniese a responder a este problema y que no llevara ese sesgo comparativo. Esta variable se definiría de la siguiente manera: Cantidad de casos ocurridos dentro de la agrupación de las manzanas dividido en la población total también de la agrupación. Entonces la variable vendría a representar la cantidad de casos promedio que ocurrieron por habitante. Y, si se aplicase al ejemplo que fue señalado, el primer caso esta variable sería 1 y en la otra 2/1000, es decir, 0.002, de lo cual se puede deducir que si se supone un contagio comunitario, en el segundo caso fue bastante más controlado, debido a que pese a que se presenta una mayor cantidad de población, la cantidad de casos ocurridos fue considerablemente menor con respecto a la cantidad de personas que vivían.

Para verificar los resultados de lo señalado en los párrafos anteriores se generaron los siguientes 2 gráficos:

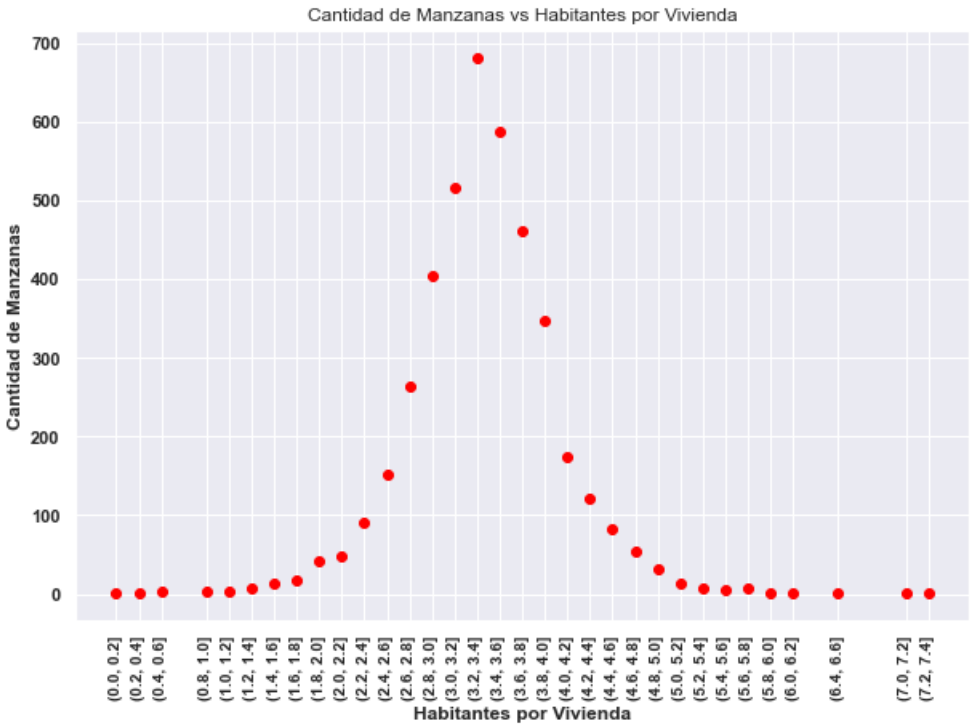


Figura 5.1: Cantidad de manzanas por rango de habitantes por vivienda con separación de 0.2 y máximo de 8

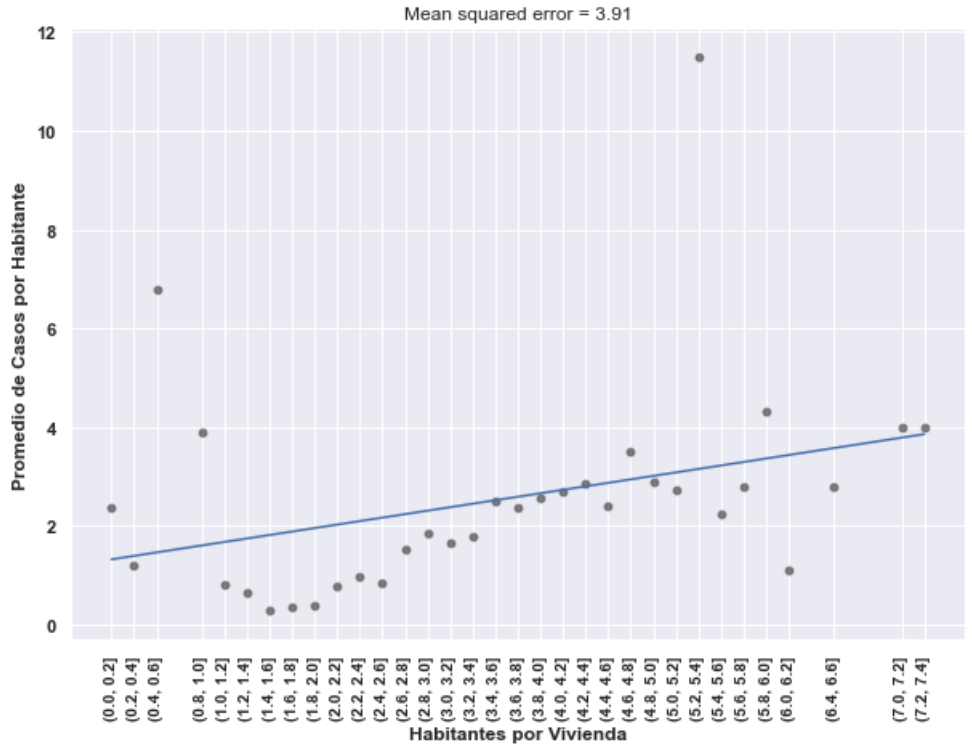
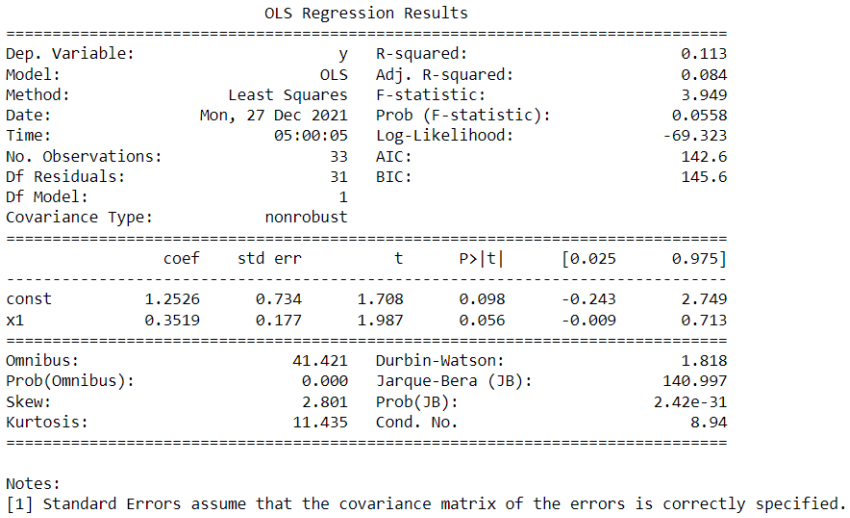


Figura 5.2: Promedio de casos por habitantes vs habitantes por vivienda



OLS Regression Results

Dep. Variable:	y	R-squared:	0.113
Model:	OLS	Adj. R-squared:	0.084
Method:	Least Squares	F-statistic:	3.949
Date:	Mon, 27 Dec 2021	Prob (F-statistic):	0.0558
Time:	05:00:05	Log-Likelihood:	-69.323
No. Observations:	33	AIC:	142.6
Df Residuals:	31	BIC:	145.6
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.2526	0.734	1.708	0.098	-0.243	2.749
x1	0.3519	0.177	1.987	0.056	-0.009	0.713

Omnibus:	41.421	Durbin-Watson:	1.818
Prob(Omnibus):	0.000	Jarque-Bera (JB):	140.997
Skew:	2.801	Prob(JB):	2.42e-31
Kurtosis:	11.435	Cond. No.	8.94

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Figura 5.3: Resultados de la regresión lineal

La Figura 5.1 tiene por objetivo representar la cantidad o conteo de manzanas por cada rango formado cuando se aplicase una separación entre uno y otro de 0.2 habitantes por vivienda. Se puede apreciar que la agrupación de las manzanas bajo los criterios mencionados en los párrafos anteriores muestra un comportamiento similar a una campana de gauss o normal, lo cual señala que la mayoría de las manzanas se acumularon en los valores medios/promedios de la variable del eje x y la minoría en los valores extremos, lo cual era esperable y natural. Existen manzanas con más de 8 personas por vivienda, pero eran casos irrelevantes, aislados y bastante alejados del promedio, por lo cual se decidió clasificarlos como datos atípicos y no considerarlos.

Se probó con una distancia de 0.1 en el eje x, pero se observó que 0.2 permitía que cada agrupación contara con mayor cantidad de manzanas y una mejor aproximación al comportamiento promedio, permitiendo visualizar de mejor manera el comportamiento natural de los datos. De todas formas, en Anexos se adjuntan resultados y comportamiento con otras separaciones. Finalmente en la Figura 5.2 se puede apreciar el comportamiento de la relación de la ocurrencia de casos y "el hacinamiento". Se ajustó además una regresión lineal que pudiese ilustrar de mejor manera la relación. Se puede observar al simplemente mirar el gráfico que en general mientras más personas vivan en promedio en una vivienda es mayor la ocurrencia de casos, en otras palabras, en las viviendas con mayor cantidad de hacinamiento, debido a que existen más miembros que conviven bajo el mismo techo, es más probable a que en promedio se presente una mayor proporción de casos, probablemente causado por la convivencia en espacios comunes, cerrados, de forma diaria y reducidos respecto a la cantidad de personas.

Vale aclarar que en este análisis no fue considerado el tamaño de cada vivienda, como había sido comentado en la sección de Hacinamiento del capítulo 3, debido a que no es un dato que esté disponible en el Censo, y que además complejizaría el cálculo. En la Figura 5.3 se pueden observar mayores detalles y resultados sobre la regresión lineal, que si se mira el R-cuadrado y el ajustado, se puede observar que la regresión lineal permitiría explicar aproximadamente un 10 % de la variabilidad de los datos, y si se mira la variable se puede observar el comportamiento positivo en el coeficiente ya que

es mayor a cero, que evidentemente se podía ver con el comportamiento creciente de la curva en el gráfico, pero también se podría observar que la variable resultaría no significativa, ya que por obvias razones sabemos que exante que "el hacinamiento" no es la única variable que influye en los contagios y propagación del virus, por si sólo la variable no puede explicar por completo el comportamiento. Además hay que considerar que quizás forzar un comportamiento lineal a la curva puede ser un supuesto bastante fuerte. También es importante mencionar, que distintas separaciones (las longitudes de los rangos), pueden cambiar el comportamiento de los promedios. Para visualizar estos últimos puntos en anexos se adjuntan los gráficos de este experimento con otra combinación de separaciones y realizando ajustes no lineales a la curva.

También es interesante señalar cómo fue evolucionando el proceso entero. Se partió con 15574 casos (Confirmados y Probables entre las fechas indicadas). Luego, al mantener sólo los casos cuyas búsquedas de su dirección resultaron de forma satisfactoria (ya que de otra forma lo siguiente no tendría sentido) quedaron 13886 casos. Luego al momento de asignarles la manzana-entidad correspondiente, sólo 7295 casos (lo que equivale al 52.5 %) pudieron ser asignados por encontrarse adentro de alguno de los polígonos, el resto tuvo que ser asignado a la manzana con el centroide más cercano. La distancia máxima de un caso a un centroide de manzana asignada fue de 482 metros, lo cual no se considera tan grave si es que la manzana es grande, hay que recordar que no hay un tamaño fijo para las manzanas. El resto de detalles se encuentran en los gráficos, ya que el siguiente proceso es el de agrupación de manzanas.

Capítulo 6

Conclusiones

6.1. Conclusiones generales

El desarrollo de esta memoria tenía como objetivo principal desarrollar un método de normalización de direcciones que permitiese geocodificar de manera eficiente y eficaz direcciones provenientes de distintas fuentes de datos sin una estructura definida ni fija, para así lograr geocodificar estas direcciones y finalmente localizar de forma correcta y precisa a los pacientes del SSMSO para poder realizar análisis posteriores. Lo cual fue logrado, ya que se aplicaron métodos de normalización que permitieron una mejora en el proceso de geocodificación, permitiendo además analizar el efecto de una variable sociodemográfica como lo es el hacinamiento (evidenciado en el capítulo de Evaluación y Resultados). También se pudo lograr desarrollar una herramienta de visualización y normalización de direcciones que fue un entregable hacia el SSMSO. Todo este proceso no sólo sirvió como componente a análisis mencionados en esta memoria, sino que también como aporte para los análisis de algunos de los demás memoristas partícipes de este proyecto, que necesitaban esencialmente que los pacientes fueran localizados correctamente. En efecto, en la memoria de Juan Pablo Alvarado se hará referencias a este trabajo.

Con un proceso de geocodificación confiable y bien logrado, era posible utilizar esos resultados con mayor seguridad para los análisis y procesos que continuaron, debido a que aquel proceso era la base de la confiabilidad de los siguientes. Con la geocodificación lograda de buena forma se puede localizar una gran base de pacientes, y analizar por ejemplo si la cantidad de personas por vivienda (hacinamiento) afectaría o no en la ocurrencia de casos. Lo que podría utilizarse con las debidas precauciones, para focalizar los esfuerzos y prestar más atención en ciertos puntos de mayor riesgo de contagio.

Por lo tanto, se puede evidenciar que una herramienta como la presentada en esta memoria y entregada al SSMSO puede ser aplicada y aportar en la correcta localización de casos covid-19. Sin embargo, debido a que el entregable fue bastante cercano al final del proyecto, no se pudo observar realmente su utilización por parte del SSMSO. Algo que podría haber logrado una mayor apreciación de la utilidad sería haber finalizado y entregado la aplicación en una etapa más temprana del proyecto, y ejercido una mayor y más clara difusión de los beneficios de ésta, en conjunto con la realización de reuniones posteriores para evaluar su uso y dudas.

Luego del trabajo realizado, se pudo observar que el proceso de geocodificación y entendi-

miento del funcionamiento interno de los sistemas de localización con sus ventajas y desventas actuales puede ser un trabajo arduo, pero bastante útil en problemas de la vida real, como lo es la atención de problemáticas mundiales como lo son los temas de salud, crímenes, política, accidentes o catástrofes. Sobre todo es importante en aportar a una respuesta más rápida y certera con el apoyo de la automatización y un sistema más eficaz.

Finalmente, a modo de reflexión la recomendación técnica y de política pública que se le recomendaría a las autoridades locales y nacionales con respecto a los datos geográficos de los habitantes del país es que lo ideal sería que se implementara un sistema de recolección de direcciones a través de un sistema estructurado, es decir, con campos tales como Calle, Número de calle y Localidad, y quizás además un campo adicional que recolecte información extra (pero no tan relevante para los servicios de geocodificación) tales como la población o los detalles que el paciente quiera otorgar, ya que esto facilitaría el proceso de geocodificación de las direcciones y agilizaría las tareas que requieran localizar al paciente, ya que en ese momento ya no sería necesario aplicar por ejemplo métodos de normalización que deban procesar previamente las direcciones para lograr resultados más correctos.

6.2. Discusión y Trabajo Futuro

Los procesos necesarios en esta memoria se llevaron a cabo con buenos resultados en general, pero quedaría pendiente analizar ciertas situaciones, por ejemplo: cuánto sería el efecto en los resultados finales de aquellas personas que no pudieron ser localizadas, ya sea porque su dirección estaba incompleta, incorrecta o no existía en la base, esto último sobre todo debido a que si bien tanto la hoja de DIRECCIONES como de PERSONAS estaban dentro del mismo archivo excel, tienen orígenes distintos: La hoja de PERSONAS proviene del seguimiento de casos de pacientes covid del SSMSO, pero la hoja de DIRECCIONES incluye información de los pacientes recopiladas de otras fuentes, evidencia de esto es la baja estructura y gran variabilidad en que se encontraban escritas las direcciones. También, queda analizar en cuánto afectaría a los resultados finales la diferencia que pueda existir entre la ubicación real del paciente y del producto del proceso de geocodificación. Además, otro punto necesario a darle una mirada sería que se está asumiendo una cierta "estaticidad" de los pacientes, ya que en esta memoria se hace un análisis sobre los domicilios de los pacientes, pero los contagios no siempre se dan dentro del hogar, o en el entorno cercano al hogar.

Sin embargo, algo que es importante señalar es que los resultados presentados en este trabajo son concordantes con respecto a lo que puede ser encontrado en otros estudios/artículos tales como el artículo de CIPER que fue relatado en la sección de Trabajos Relacionados y Referencias del capítulo 3, en donde se habla del hacinamiento como variable clave para explicar la propagación de la covid-19.

Es relevante de todas maneras señalar las diferencias que existen entre aquel estudio y el llevado a cabo en esta memoria. Aquel estudio se llevó a cabo utilizando datos de la encuesta CASEN (Encuesta de Caracterización Socioeconómica Nacional) lo cual les permitió indagar sobre la pobreza multidimensional y otras variables, pero sólo a un nivel de agregación mayor al llevado a cabo en esta memoria. La CASEN presenta resultados generales con nivel de agregación mínimo a nivel comuna, en cambio, en la memoria se quiso ir a un nivel de agregación menor, como lo son las manzanas-entidades. Otra diferencia es que la medida

de hacinamiento utilizada en el estudio fue la de “El número de personas en el hogar por dormitorio de uso exclusivo es mayor o igual a 2.5”, en cambio, en esta memoria se quiso definir el hacinamiento como “La cantidad de personas promedio que viven por vivienda” dentro de la manzana-entidad. De todas formas, se llegaron a conclusiones similares.

A futuro sería interesante indagar sobre el efecto en el contagio de otras variables y el efecto conjunto de ellas, ya que en este trabajo sólo se indagó sobre el efecto de una, o incluso sería importante realizar un análisis a nivel país, ya que esté sólo se centró en la Región Metropolitana. Variables importantes serían: la escolaridad, rangos de edades, cantidad de personas migrantes, entre otras. Aunque lo ideal sería contar primero con una base de datos que contenga estas variables más actualizada y que represente más fielmente la realidad actual del país, por lo cual habría que esperar una realización de otro Censo.

Bibliografía

[1] O.M.S., "Información básica sobre la covid-19," 2020, https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19 (visitado el 31 de Diciembre de 2021).

[2] MINSAL, "Covid-19: Se reporta una positividad de 3 % en las últimas 24 horas a nivel nacional," 2021, https://www.minsal.cl/wp-content/uploads/2021/07/CP-Reporte-Covid-17072021.pdf (visitado el 31 de Diciembre de 2021).

[3] Davis, C. y Fonseca, F., "Assessing the certainty of locations produced by an address geocoding system," Geoinformatica, vol. 11, p. 103–129, 2007, https://doi.org/10.1007/s10707-006-0015-7 (visitado el 31 de Diciembre de 2021).

[4] O.M.S., "Vigilancia de salud pública en relación con la covid-19: orientaciones provisionales, 7 de agosto de 2020," 2020, https://apps.who.int/iris/handle/10665/334000 (visitado el 31 de Diciembre de 2021).

[5] Porto, J. P. y Gardey, A., "Definicion.de: Definición de dirección," 2021, https://definicion.de/direccion/ (visitado el 31 de Diciembre de 2021).

[6] Porto, J. P. y Merino, M., "Definicion.de: Definición de domicilio," 2014, https://definicion.de/domicilio/ (visitado el 31 de Diciembre de 2021).

[7] Küçük Matci, D. y Avdan, U., "Address standardization using the natural language process for improving geocoding results," Computers, Environment and Urban Systems, vol. 70, pp. 1–8, 2018, https://doi.org/10.1016/j.compenvurbsys.2018.01.009 (visitado el 31 de Diciembre de 2021).

[8] DataCentric, "Cómo distinguir entre geolocalización y georeferenciación," 2018, https://www.datacentric.es/blog/geomarketing/diferencia-entre-geolocalizacion-y-georeferenciacion/ (visitado el 31 de Diciembre de 2021).

[9] Wikipedia, "Geoetiquetado — wikipedia, la enciclopedia libre," 2020, https://es.wikipedia.org/w/index.php?title=Geoetiquetado&oldid=124516951 (visitado el 31 de Diciembre de 2021).

[10] MINSAL, D. D. E., "Manual operativo para las acciones de trazabilidad y aislamiento," 2020, https://coronavirus.achs.cl/docs/default-source/default-document-library/ordinario-4152-minsal.pdf?sfvrsn=18d655e_0 (visitado el 31 de Diciembre de 2021). Versión 2.0.

[11] Wikipedia, "Epidemia — wikipedia, la enciclopedia libre," 2021, https://es.wikipedia.org/w/index.php?title=Epidemia&oldid=140053804 (visitado el 31 de Diciembre de 2021). [Internet; descargado 27-diciembre-2021].

[12] Dumar, A. M., Swine flu: What you need to know. Place of publication not identified:

Brownstone Books, 2009, http://www.worldcat.org/oclc/401165992 (visitado el 31 de Diciembre de 2021).

[13] Garay, C. C., "Las cinco pandemias más letales de la historia de la humanidad," 2021, https://www.nationalgeographic.es/historia/2020/11/cinco-pandemias-mas-letales-de-historia-de-humanidad (visitado el 31 de Diciembre de 2021).

[14] O.M.S., "Coronavirus," 2020, https://www.who.int/es/health-topics/coronavirus#tab=tab_1 (visitado el 31 de Diciembre de 2021).

[15] O.M.S., "Necesidades, percepciones y demandas de las comunidades: instrumento de evaluación comunitaria: módulo de la serie de evaluaciones de la capacidad de los servicios de salud en el contexto de la pandemia de covid-19: orientaciones provisionales, 5 de febrero de 2021," 2021, https://apps.who.int/iris/handle/10665/340296 (visitado el 31 de Diciembre de 2021).

[16] MINSAL, "A seis meses del primer caso de covid-19, el 93 % de los pacientes se han recuperado," 2020, https://www.minsal.cl/a-seis-meses-del-primer-caso-de-covid-19-el-93-de-los-pacientes-se-han-recuperado/ (visitado el 31 de Diciembre de 2021).

[17] Ponce Maripangui, D., "Sistema de búsqueda inteligente de direcciones para empresa de distribución postal," 2021, http://repositorio.uchile.cl/handle/2250/181650 (visitado el 31 de Diciembre de 2021).

[18] Barraza, R., Barrientos, R., Díaz, X., Pleitez, R., y Tablas, V., "Covid-19 y vulnerabilidad: una mirada desde la pobreza multidimensional en el salvador," 2020, https://www.latinamerica.undp.org/content/rblac/es/home/library/crisis_prevention_and_recovery/covid-19-y-vulnerabilidad--una-mirada-desde-la-pobreza-multidime.html (visitado el 31 de Diciembre de 2021).

[19] Alkire, S., Dirksen, J., Nogales, R., y Oldiges, C., "Multidimensional poverty and covid-19 risk factors: A rapid overview of interlinked deprivations across 5.7 billion people," 2020, https://ophi.org.uk/b53/ (visitado el 31 de Diciembre de 2021).

[20] Simunovic, A. T. y Urquiza, N. F., "Hacinamiento: la variable clave en la propagación del covid-19 en el gran santiago," 2021, https://www.ciperchile.cl/2020/10/17/hacinamiento-la-variable-clave-en-la-propagacion-del-covid-19-en-el-gran-santiago/ (visitado el 31 de Diciembre de 2021).

[21] Liddy, E. D., "Enhanced text retrieval using natural language processing," Bulletin of the American Society for Information Science and Technology, vol. 24, no. 4, pp. 14–16, 1998, https://doi.org/10.1002/bult.91 (visitado el 31 de Diciembre de 2021).

[22] Wikipedia, "Lenguaje natural — wikipedia, la enciclopedia libre," 2019, https://es.wikipedia.org/w/index.php?title=Lenguaje_natural&oldid=121210722 (visitado el 31 de Diciembre de 2021).

[23] Adalı, E., "Doğal dil İşleme," Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi, vol. 5, pp.–, 2016, https://dergipark.org.tr/en/pub/tbbmd/issue/50642/643345 (visitado el 31 de Diciembre de 2021).

[24] Chowdhury, G. G., "Natural language processing," Annual Review of Information Science and Technology, vol. 37, no. 1, pp. 51–89, 2003, https://doi.org/10.1002/aris.1440370103 (visitado el 31 de Diciembre de 2021).

[25] Wang, Y. y Wang, X.-J., "A new approach to feature selection in text classification," en

2005 International Conference on Machine Learning and Cybernetics, vol. 6, pp. 3814–3819 Vol. 6, 2005, 10.1109/ICMLC.2005.1527604 (visitado el 31 de Diciembre de 2021).

[26] Ezequiel, "Cuál es la diferencia entre paralelos y meridianos.", https://epicentrogeografico.com/2018/10/cual-es-la-diferencia-entre-paralelos-y-meridianos/ (visitado el 31 de Diciembre de 2021).

[27] Wikipedia, "Latitud — wikipedia, la enciclopedia libre," 2022, https://es.wikipedia.org/w/index.php?title=Latitud&oldid=142073343 (visitado el 31 de Diciembre de 2021).

[28] Wikipedia, "Longitud (cartografía) — wikipedia, la enciclopedia libre," 2022, https://es.wikipedia.org/wiki/Longitud_(cartografÃŋa) (visitado el 31 de Diciembre de 2021).

[29] Resources, A., "Tres representaciones fundamentales de capas de información geográficar.", https://resources.arcgis.com/es/help/getting-started/articles/026n0000000n000000.htm (visitado el 31 de Diciembre de 2021).

[30] Serra, B. R., "Centroide.", https://www.universoformulas.com/matematicas/geometria/centroide/ (visitado el 31 de Diciembre de 2021).

[31] Bembibre, C., "Hacinamiento," 2022, https://www.definicionabc.com/social/hacinamiento.php (visitado el 31 de Diciembre de 2021).

[32] de Desarrollo Social y Familia, M., "CaracterizaciÓn socioeconÓmica," 2020, https://datasocial.ministeriodesarrollosocial.gob.cl/fichaIndicador/649/1 (visitado el 31 de Diciembre de 2021).

[33] INE, "Ine - plataforma de datos geoestadisticos.", https://ine-chile.maps.arcgis.com/apps/dashboards/e8292e6a13814b6b8bcfd3415ef4eb02 (visitado el 31 de Diciembre de 2021).

[34] Santander, U., "Python: qué es y por qué deberías aprender a utilizarlo," 2021, https://www.becas-santander.com/es/blog/python-que-es.html (visitado el 31 de Diciembre de 2021).

[35] Selvaraj, N., "Geocoding in python: A complete guide," 2021, https://www.natasshaselvaraj.com/a-step-by-step-guide-on-geocoding-in-python/ (visitado el 31 de Diciembre de 2021).

[36] INE, "¿qué es el censo?.", http://www.censo2017.cl/que-es-el-censo/ (visitado el 31 de Diciembre de 2021).

[37] INE, "Memoria del censo 2017," 2018, https://www.censo2017.cl/memoria/descargas/memoria/libro_memoria_censal_2017_final.pdf (visitado el 31 de Diciembre de 2021).

[38] INE, "Base cartográfica censal -alcances y consideraciones para el usuario," 2018, http://www.censo2017.cl/servicio-de-mapas/descargas/mapas/alcances-base-cartografica-censo2017.pdf (visitado el 31 de Diciembre de 2021).

[39] INE, "Publicación cartográfica – ine 2017," 2017, https://www.ide.cl/index.php/sociedad/item/1740-cartografia-precenso-2016-region-metropolitana-de-santiago (visitado el 31 de Diciembre de 2021).

[40] Geoapify, "Nominatim vs photon geocoder," 2019, https://www.geoapify.com/nominatim-vs-photon-geocoder (visitado el 31 de Diciembre de 2021).

Anexo A

Otras Regresiones/Ajustes y Tamaños de Rango

A continuación, se muestran los resultados del segundo experimento del capítulo de Evaluación y Resultados, pero con distinto largo de los rangos del eje x o separaciones como les llamó, y además otros tipos de ajustes/regresiones:

A.1. Separación de 0.1

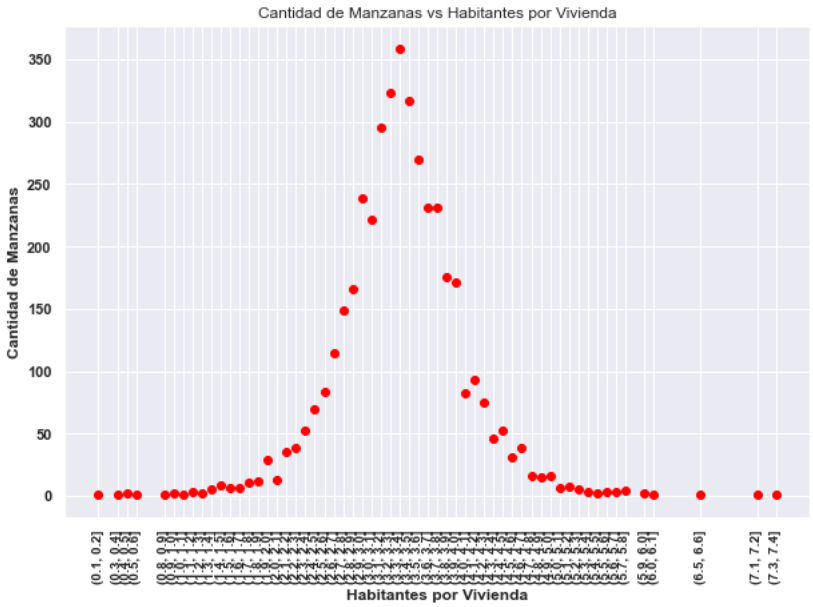
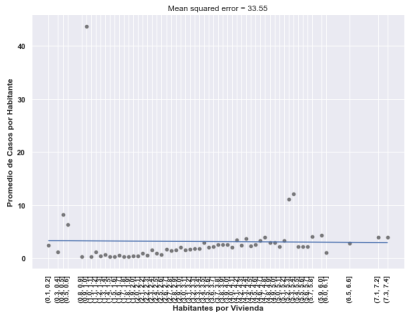
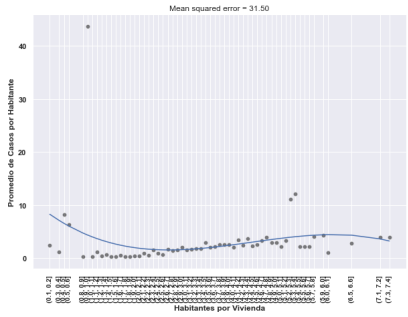


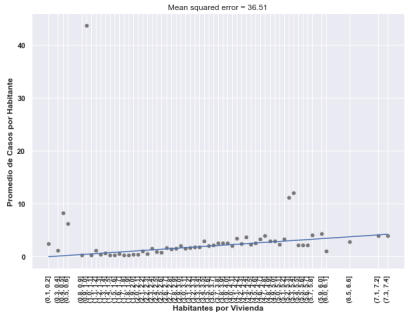
Figura A.1: Cantidad de manzanas por rango de habitantes por vivienda



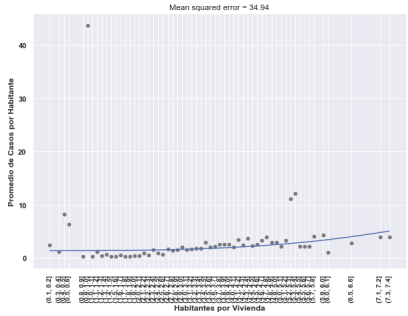
(a) Regresión Lineal



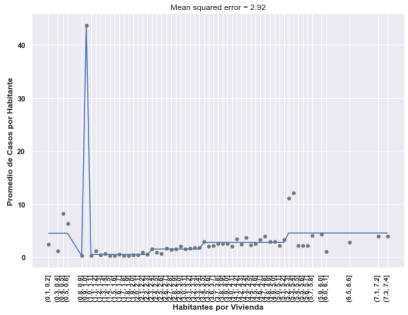
(b) Regresión Lineal con E.P. de 3er grado



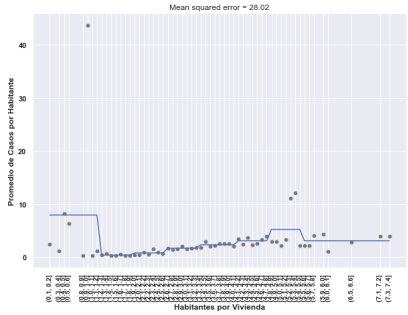
(c) SVM con Kernel Lineal



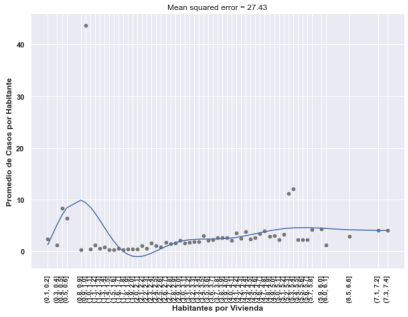
(d) SVM con Kernel Polinominal de 3er grado



(e) Regresión Árbol de Decisión



(f) Regresión Lineal con Transformador N.L. KBinsDiscretizer



(g) Regresión Lineal con Transformador N.L. Nystroem

Figura A.2: Promedio de casos por habitantes vs habitantes por vivienda

A.2. Separación de 0.2

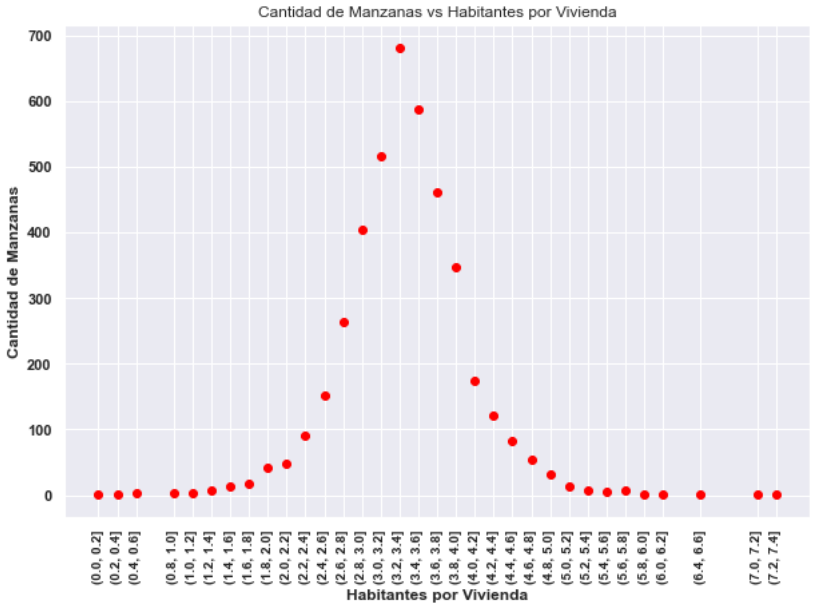
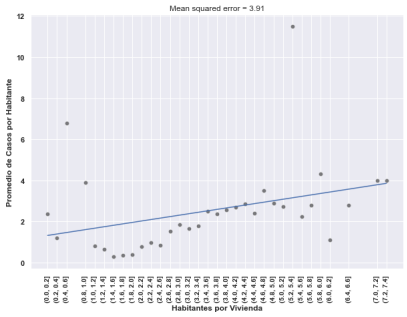
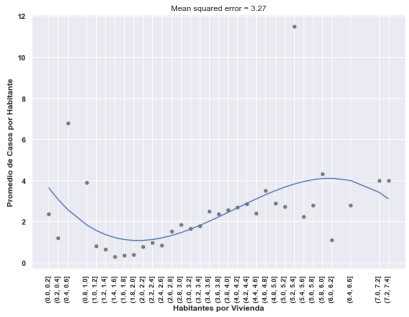


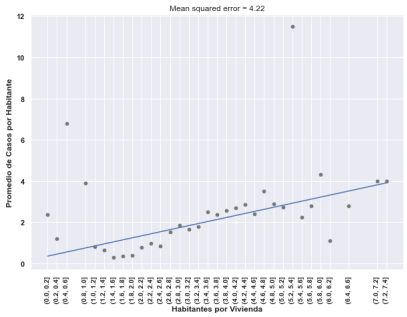
Figura A.3: Cantidad de manzanas por rango de habitantes por vivienda



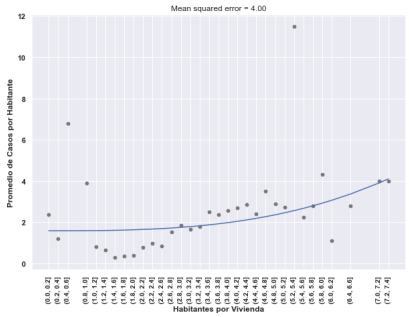
(a) Regresión Lineal



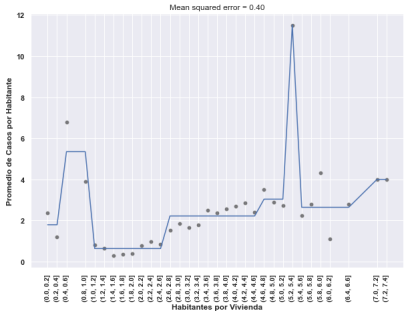
(b) Regresión Lineal con E.P. de 3er grado



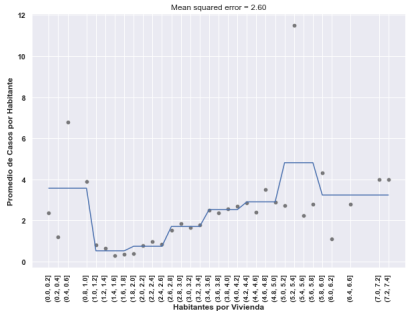
(c) SVM con Kernel Lineal



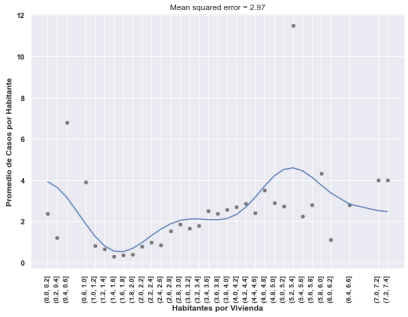
(d) SVM con Kernel Polinominal de 3er grado



(e) Regresión Árbol de Decisión



(f) Regresión Lineal con Transformador N.L. KBinsDiscretizer



(g) Regresión Lineal con Transformador N.L. Nystroem

Figura A.4: Promedio de casos por habitantes vs habitantes por vivienda

A.3. Separación de 0.3

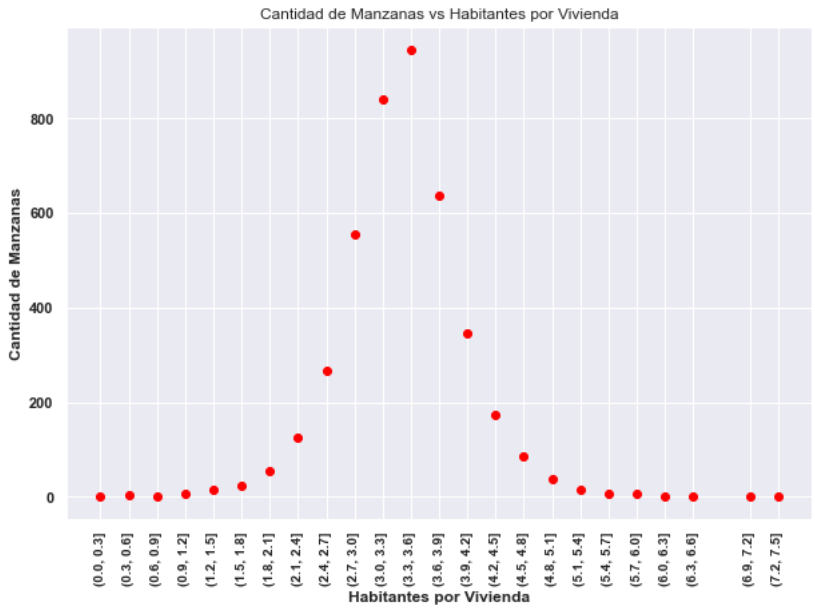
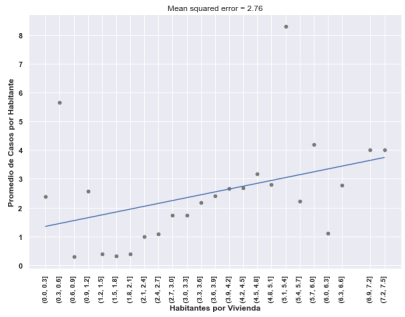
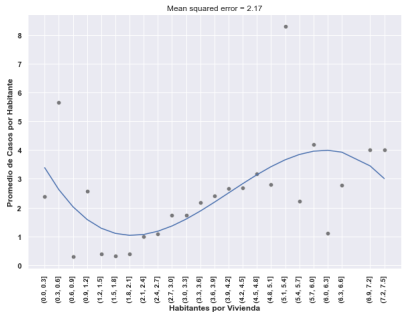


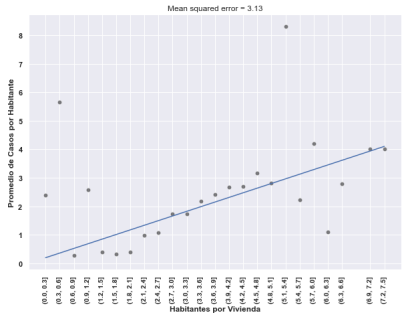
Figura A.5: Cantidad de manzanas por rango de habitantes por vivienda



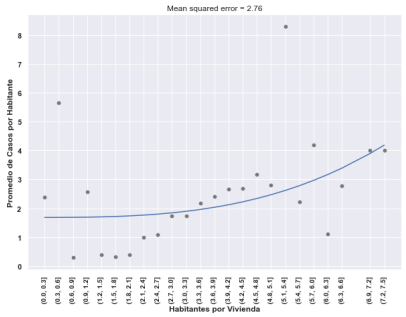
(a) Regresión Lineal



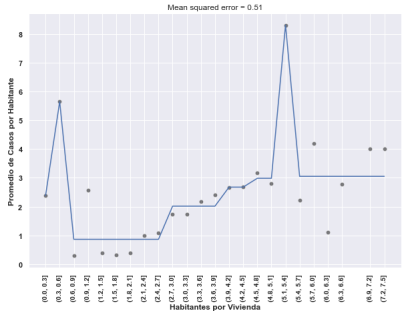
(b) Regresión Lineal con E.P. de 3er grado



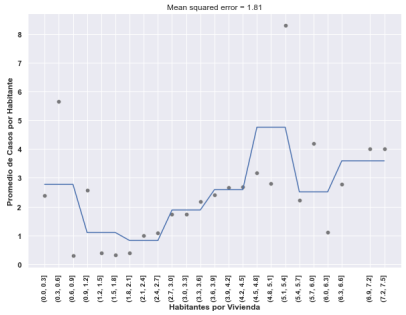
(c) SVM con Kernel Lineal



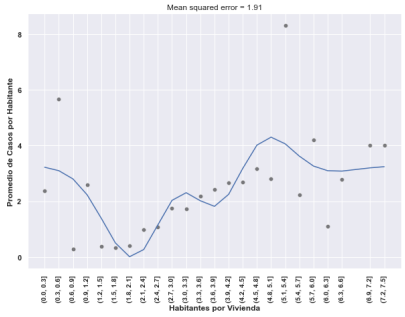
(d) SVM con Kernel Polinominal de 3er grado



(e) Regresión Árbol de Decisión



(f) Regresión Lineal con Transformador N.L. KBinsDiscretizer



(g) Regresión Lineal con Transformador N.L. Nystroem

Figura A.6: Promedio de casos por habitantes vs habitantes por vivienda

A.4. Separación de 0.4

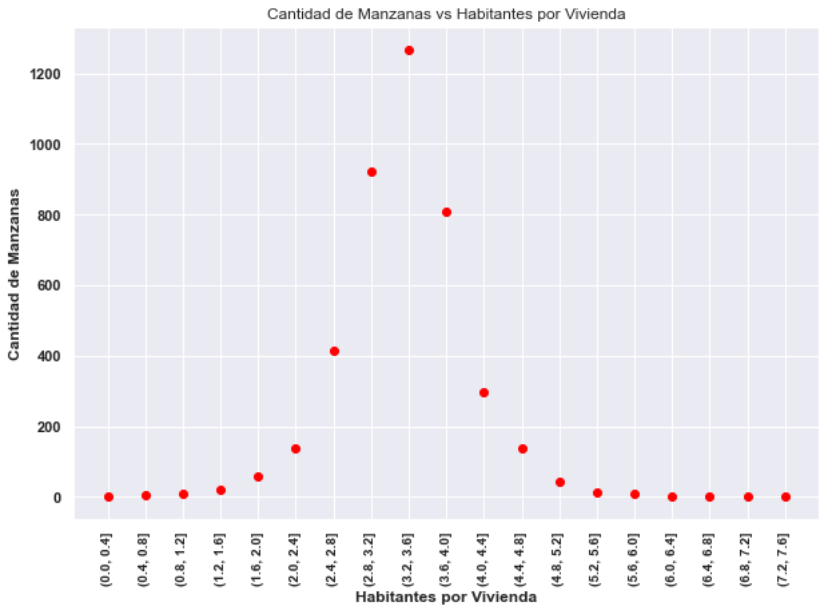
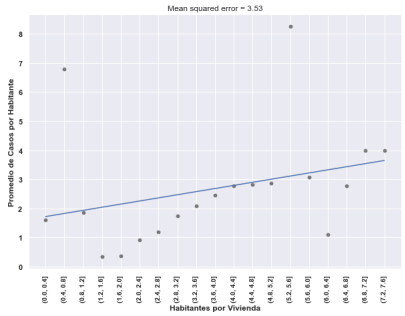
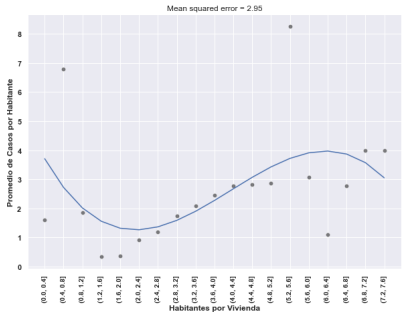


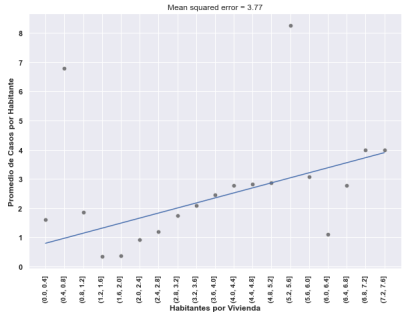
Figura A.7: Cantidad de manzanas por rango de habitantes por vivienda



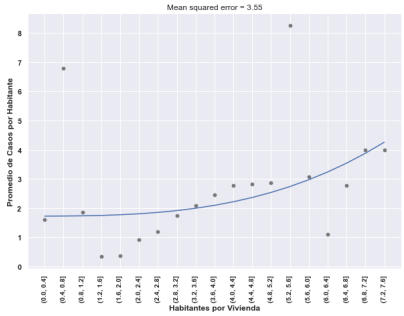
(a) Regresión Lineal



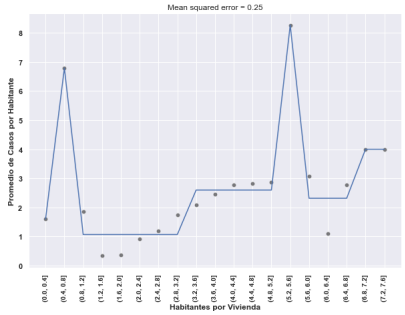
(b) Regresión Lineal con E.P. de 3er grado



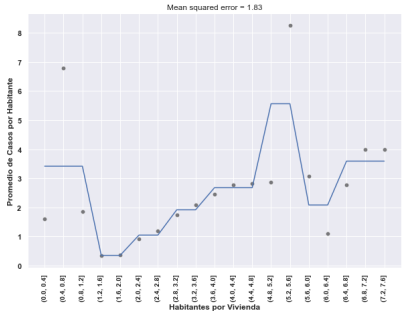
(c) SVM con Kernel Lineal



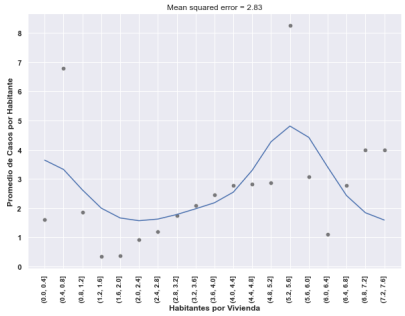
(d) SVM con Kernel Polinominal de 3er grado



(e) Regresión Árbol de Decisión



(f) Regresión Lineal con Transformador N.L. KBinsDiscretizer



(g) Regresión Lineal con Transformador N.L. Nystroem

Figura A.8: Promedio de casos por habitantes vs habitantes por vivienda

A.5. Separación de 0.5

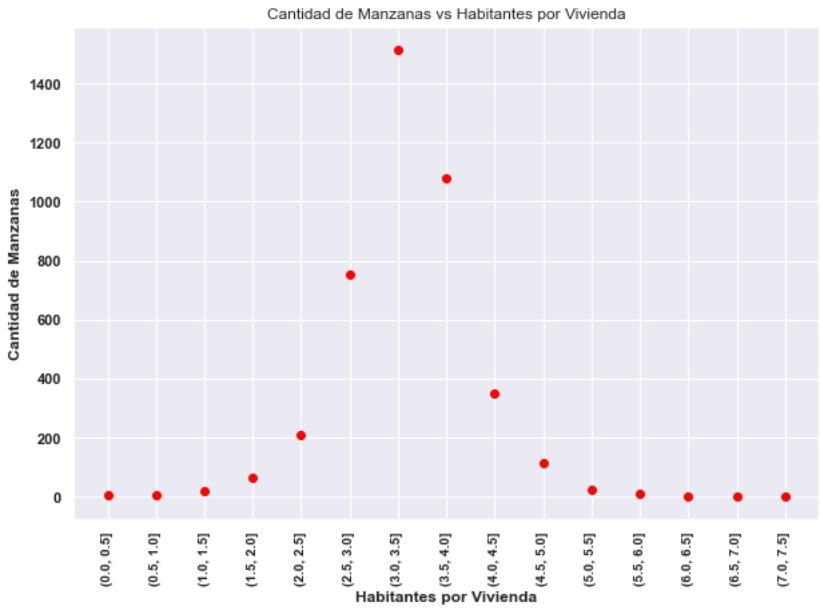
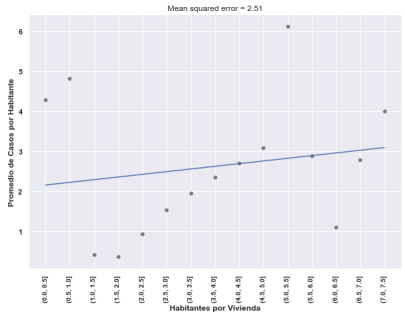
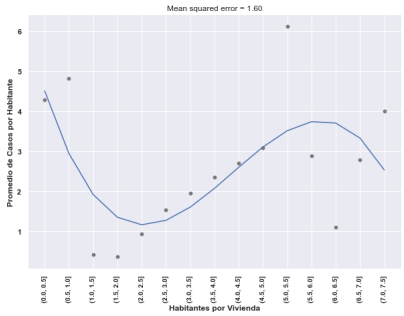


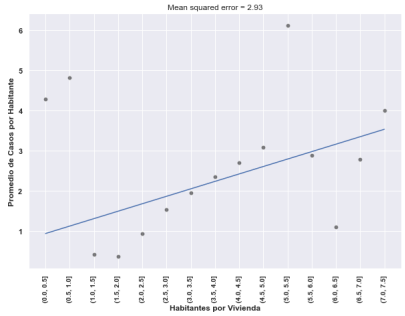
Figura A.9: Cantidad de manzanas por rango de habitantes por vivienda



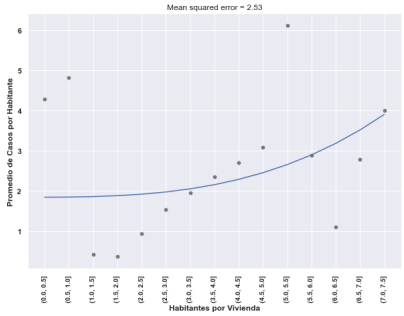
(a) Regresión Lineal



(b) Regresión Lineal con E.P. de 3er grado



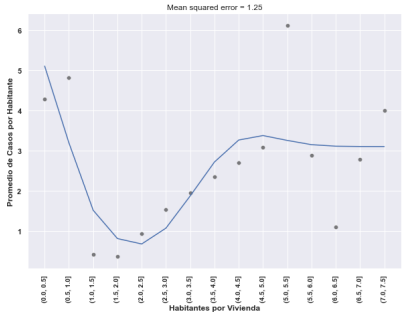
(c) SVM con Kernel Lineal



(d) SVM con Kernel Polinominal de 3er grado

(e) Regresión Árbol de Decisión

(f) Regresión Lineal con Transformador N.L. KBinsDiscretizer



(g) Regresión Lineal con Transformador N.L. Nystroem

Figura A.10: Promedio de casos por habitantes vs habitantes por vivienda

A.6. Separación de 1

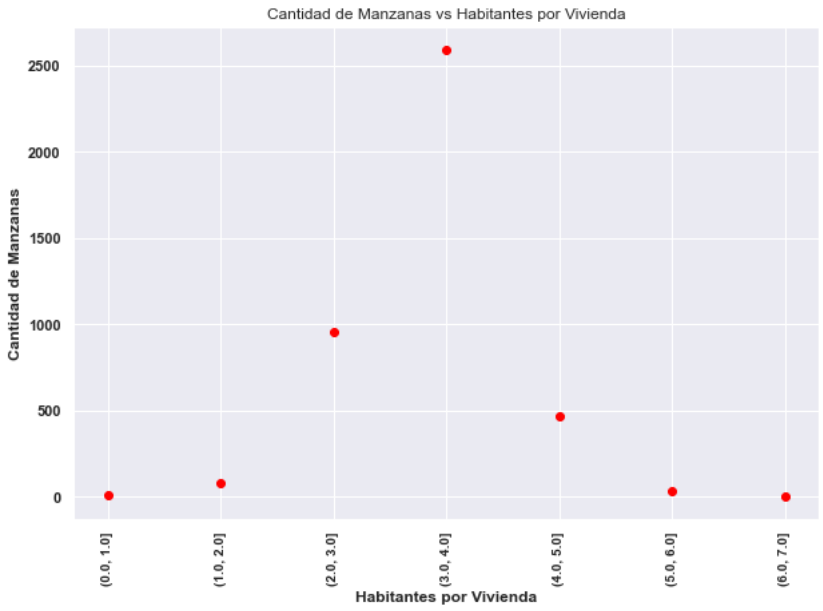
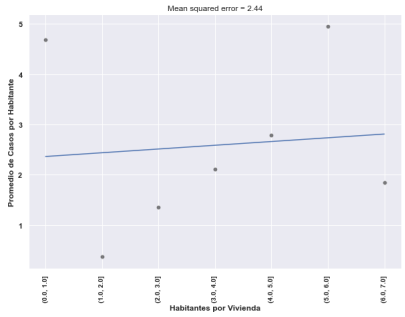
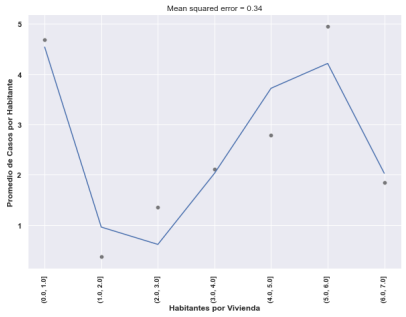


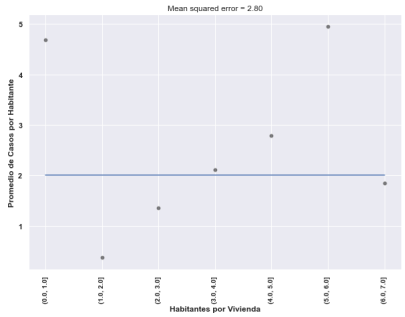
Figura A.11: Cantidad de manzanas por rango de habitantes por vivienda



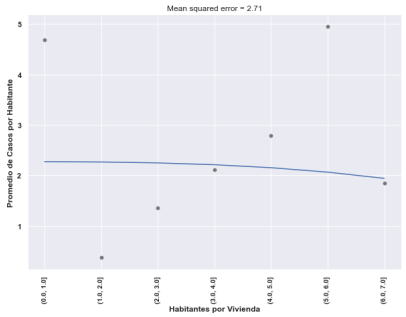
(a) Regresión Lineal



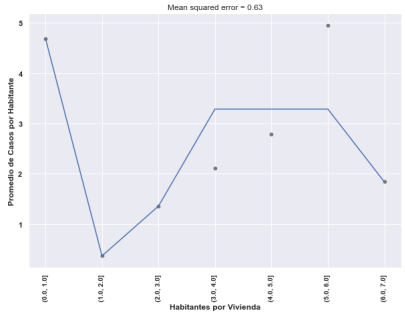
(b) Regresión Lineal con E.P. de 3er grado



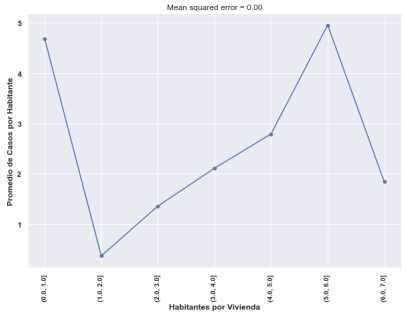
(c) SVM con Kernel Lineal



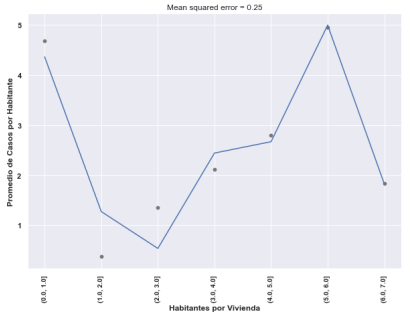
(d) SVM con Kernel Polinominal de 3er grado



(e) Regresión Árbol de Decisión



(f) Regresión Lineal con Transformador N.L. KBinsDiscretizer



(g) Regresión Lineal con Transformador N.L. Nystroem

Figura A.12: Promedio de casos por habitantes vs habitantes por vivienda