Análisis de la propagación del Covid-19 en la Ciudad de Santiago de Cali para el periodo Marzo del 2020 - Mayo del 2021

Daniel Andres Delgado Ordoñez
Bryan Steven Martinez Valencia

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2022

Análisis de la propagación del Covid-19 en la Ciudad de Santiago de Cali para el periodo Marzo del 2020 - Mayo del 2021

Daniel Andres Delgado Ordoñez
Bryan Steven Martinez Valencia

Trabajo de grado presentado como requisito parcial para optar al título de:
Estadístico

Director:
Ph.D. Jaime Mosquera Restrepo
Codirector:
M.Sc. David Arango

Universidad del Valle
Facultad de Ingeniería, Escuela de Estadística
Santiago de Cali, Colombia
2022

Dedicatoria

Este logro fue posible gracias a Dios por darme la Salud, la sabiduría y fuerza para afrontar cada etapa de este proceso. A mis padres que son el pilar fundamental de mi vida, gracias por su amor incondicional y sacrificio, gracias a ellos soy quien soy hoy en día. A mis abuelas, que desde el cielo me dan fuerzas para salir adelante. A mi novia por brindarme su apoyo, amor y compañía en momentos difíciles. A mis mejores amigos José y Carlos, por ayudarme a través de sus consejos. A todos y cada uno de las personas que hicieron parte para alcanzar esta meta tan importante en mi vida. ¡Muchas Gracias!

Daniel Andres Delgado Ordoñez

Inicial y principalmente a Dios, por haberme dado la vida, salud, fortaleza y sabiduría para llegar hasta este momento. A mi madre por ser la persona que creyó en mis capacidades desde el momento cero y me brindó su apoyo incondicional. Por último y no menos importante, agradezco esa maravillosa mujer que con mucho amor y paciencia estuvo ahí para enseñarme, aconsejarme y apoyarme a lo largo de este proceso.

Bryan Steven Martinez Valencia

Agradecimientos

Agradecemos a Dios por permitirnos lograr esta meta. A nuestros padres por brindarnos todo su apoyo incondicional, guiarnos y no permitir que desfallezcamos en cada una de las etapas de este largo proceso.

Agradecemos a la Universidad del Valle por abrirnos las puertas para alcanzar el título de Estadísticos, también agradecemos a la Escuela de Estadística y todo su cuerpo docente por brindarnos el conocimiento y las herramientas necesarias para formarnos como profesionales íntegros.

Agradecemos especialmente al profesor Jaime Mosquera Restrepo por su tiempo, sabiduría, interés y todo el apoyo que nos brindó. También, agradecemos la confianza depositada por la Secretaria Municipal de Salud de Santiago de Cali.

Agradecemos a todos los compañeros y amigos, que de alguna u otra manera fueron cruciales para alcanzar este título.

Resumen

Este trabajo de grado tiene como propósito identificar la relación entre el Número Efectivo de Reducción (R_t) y diversas características socioeconomicas asociadas a los barrios que componen la ciudad de Santiago de Cali para el periodo comprendido entre el mes de Marzo del 2020 y el mes de Mayo del 2021. Inicialmente se realiza un análisis exploratorio de los datos socieconomicos y de los registros de contagios en la ciudad. Posteriormente se realiza la estimación del indicador de propagación del virus denominado Número Efectivo de Reproducción. Finalmente se analiza la asociación entre la propagación y las condiciones socio económicas de los barrios, todo esto a través de la metodología de Mínimos Cuadrados Parciales (PLS).

Palabras clave: *Número efectivo de reproducción, Coronavirus, Mínimos Cuadrados Parciales, Propagación, Covid-19, análisis exploratorio espacial.*

Abstract

The purpose of this degree work is to identify the relationship between the effective number of reduction (R_t) and various socioeconomic characteristics associated with the neighborhoods that make up the city of Santiago de Cali for the period between the month of March 2020 and the month of May 2021. Initially, an exploratory analysis of the socioeconomic data and the records of infections in the city is carried out. Subsequently, the estimation of the virus propagation indicator is made; called the effective reproduction number. Finally, the associations that can occur between the propagation and the various socioeconomic conditions of the neighborhoods are analyzed, all this through the partial least squares (PLS) methodology.

Keywords: *Effective reproduction number, coronaviruses, partial least squares, propagation, exploratory spatial analysis.*

Contenido

Lista de Figuras

Lista de Tablas

1 Introducción

La enfermedad SARS-coV-2, más conocida como Covid-19, desde sus inicios en China ha venido evolucionando y se convirtió en un serio problema de salud pública para los gobiernos a nivel mundial. Todo esto debido a la velocidad de transmisión del virus y la poca preparación y/o capacidad que tenían las instituciones prestadoras de salud para atender la alta demanda de casos.

En este trabajo de grado se propone estudiar la propagación del virus a nivel local, con información de casos confirmados por Covid-19 que van desde el mes de Marzo del 2020 hasta el mes de Mayo del 2021, en la ciudad de Santiago de Cali. El presente trabajo tiene como objetivo principal modelar las relaciones entre variables socioeconómicas asociadas a los barrios como la cantidad de viviendas de uso residencial y comercial, número de viviendas que cuentan con los servicios de acueducto, alcantarillado, energía eléctrica; entre otras y el Número Efectivo de Reproducción a nivel de los 249 barrios en los cuales se divide la ciudad.

Para el desarrollo de este trabajo se cuenta con dos archivos de datos. El primero, suministrado por la Secretaria Municipal de Salud de Santiago de Cali que contiene los registros de personas confirmadas como positivas del virus Covid-19; en este archivo también se cuentan con variables como el genero de la persona, el barrio, la sintomatologia, el estado del caso (leve, moderado, severo, fallecido), la edad y si presenta comorbilidades. El segundo archivo con el que se cuenta para este trabajo es denominado características socioeconómicas de los barrios; extraído de los registros de la pagina oficial del Departamento Administrativo Nacional de Estadística (DANE), el cual contiene registros del último censo nacional de población realizado en el año 2018 y contiene variables asociadas a los barrios de la ciudad de Santiago de Cali, como la cantidad de viviendas de uso residencial y comercial, número de viviendas que cuentan con los servicios de acueducto, alcantarillado, energía eléctrica, entre otras.

Con la información contenida en el primer archivo de datos se procedió a realizar un análisis exploratorio, así como el conteo de los casos positivos de Covid-19 diario y mensual por barrios, comunas y en general de la ciudad de Cali, para posteriormente realizar la estimación del Número Efectivo de Reproducción de la enfermedad en cada uno de los escenarios mencionados anteriromente (comuas, barrios y la ciudad en general). Este indicador será el insumo para analizar de manera espacial el comportamiento de la propagación del virus mes a mes en la ciudad. Adicionalmente la estimación del indicador del Número Efectivo

de Reproducción por barrio también se utiliza como variable de respuesta en un modelo estadístico que pretende identificar si existe relación en la propagación del Covid-19 en la ciudad de Santiago de Cali con variables socioeconómicas asociados a los 249 barrios de la ciudad. Con el segundo archivo se realiza un análisis descriptivo de las variables asociadas a las condiciones de los barrios; las cuales serán predictores del modelo estadístico.

La estructura del documento se divide en 6 Capítulos. En el Capítulo 1 se realiza una breve introducción al fenómeno de estudio, mencionando el planteamiento del problema, la justificación y los objetivos del estudio, explicando las razones e importancia de este trabajo. En el Capítulo 2 se presenta una recopilación de estudios previos relacionados con la enfermedad SARS-coV-2 o Covid -19 en diferentes contextos a nivel mundial y también haciendo énfasis en Colombia. El Capítulo 3 está conformado por el Marco teórico de la investigación. En la primera parte se encuentra el marco conceptual, en el cual se incluyen definiciones sobre componentes epidemiológicos y en la segunda parte se describen las características más importantes de los métodos a implementar. En el Capítulo 4 se describe la metodología implementada. En el Capítulo 5 se describen los resultados obtenidos y finalmente, en el Capítulo 6 se presentan las conclusiones y recomendaciones para estudios posteriores.

1.1. Planteamiento del problema

El siglo XXI ha traído cambios en la sociedad a nivel socioeconómico y con ello surgen retos importantes para la salud pública, uno de estos retos ha sido la lucha contra enfermedades infecciosas de naturaleza transmisible, como el síndrome de la inmunodeficiencia adquirida (SIDA) o patologías que afectan el sistema respiratorio, cómo el síndrome respiratorio agudo (Bernabeu-Mestre et al., 2004). Para los entes gubernamentales es importante realizar estudios que permitan comprender el comportamiento de estas enfermedades, su nivel de propagación y evaluar la existencia de otros factores asociados que influyan en su comportamiento o propagación.

Las enfermedades de tipo infeccioso constituyen unos de los principales obstáculos para el progreso, disminuyen el desarrollo social, económico y muchas veces agudizan la desigualdad social dentro de los países. Un caso de este tipo de enfermedades ocurrió en el año 2002, un virus llamado síndrome respiratorio agudo severo (SARS), que inició en la provincia de Guangdong (China) y se expandió por 30 países alrededor del mundo ocasionando más de 800 personas fallecidas y más de 8.439 personas afectadas, lo que puso en alerta a diferentes entidades de salud a nivel mundial (Hincapié, 2007).

En la actualidad, el Covid-19 es una enfermedad de carácter infeccioso causada por una familia de coronavirus (SARS-Cov-2) que se ha propagado al rededor del mundo, de tal manera que para autoridades sanitarias como la Organización Mundial de la Salud (OMS)

la declaro como pandemia el 11 de Marzo del 2022. Este virus tiene sus inicios en el mes de Diciembre del 2019 en China, en la ciudad de Huwan, provincia de Hubei (Huang et al., 2020).

Según información extraída de (Datos Abiertos, 2022) en Colombia, desde el registro del primer caso de Covid-19 hasta el mes de Mayo del 2022 se tienen al rederor de 6.121.429 infectados de los cuales 139.989 son casos fallecidos. En el departamento del Valle del Cauca en el mismo intervalo de tiempo mencionado anteriormente se tienen 542.217 casos de los cuales 399.455 pertenecen a la ciudad de Cali y de los cuales 15.085 han fallecido.

Es relevante para la ciudad de Santiago de Cali, la cual es una de las principales ciudades de Colombia, contar con un estudio que permitiera conocer las características epidemiológicas, clínicas y virológicas del Covid-19; así mismo observar la dinámica de propagación dentro del territorio y esta cómo afecta a la población expuesta al virus. Por tal razón, esta investigación pretende identificar si existe relación entre la propagación del virus y factores socioeconómicos asociados a los barrios de la ciudad; identificando de manera espacial zonas o sectores de la ciudad donde la propagación tuvo un mayor impacto y analizando la evolución del Número Efectivo de Reproducción (R_t) que corresponde a la velocidad que una persona contagiada puede transmitir el virus a otra persona. Esto con el fin de proporcionar información relevante a las entidades municipales que están relacionadas con este problema de salud pública para que en futuras emergencias tengan herramientas que les permitan actuar de manera mas efectiva y salvaguardar la vida de la población.

Tenido en cuenta lo expuesto anteriormente surge una pregunta de investigación ¿Existe relación entre la velocidad de propagación del virus y factores socioeconómicos en los barrios de la ciudad de Cali?

1.2. Justificación

A nivel mundial muchos académicos e investigadores han contribuido al problema que se ha generado por el Covid-19, algunas de las contribuciones más importantes están en el uso de modelos epidemiológicos deterministas (Kucharski et al., 2020). Estas investigaciones se han centrado en analizar las dinámicas del fenómeno, esto con el fin de predecir nuevos casos de personas infectadas o fallecidas, obteniendo información relevante para los entes de control, donde se han podido tomar decisiones en materia de salud publica y manejar de forma adecuada la situación.

En la ultima década el análisis temporal y espacial ha sido importante en el área de la salud porque ha permitido analizar y hacer seguimiento a la incidencia de enfermedades como el Covid-19, este tipo de herramientas han evolucionado de manera rápida, ya que las pandemias se propagan de esa misma forma, lo cual da paso a que hayan muchas técnicas aun sin explorar y que se pueden aplicar en la salud publica.

En este tipo de virus el principal factor de contagio de la enfermedad es el ser humano debido que se da mediante micro partículas expulsadas por boca y nariz principalmente. Por lo cual, contraer este tipo de enfermedades es mucho más fácil debido a las relaciones sociales y económicas entre países y personas, lo que conduce a una aceleración en la propagación del virus. Por lo anterior, se propone analizar los casos positivos del Covid-19 a través del espacio y el tiempo, con el fin de identificar patrones de contagio que permitan observar en que zonas se presenta el contagio en mayor o menor medida.

Según lo reportado por el Instituto Nacional de Salud (INS), la ciudad de Santiago de Cali fue una de las ciudades con mayor tasa de incidencia diaria del país. Por lo tanto, es de gran interés conocer como se comporta la propagación del virus en diferentes sectores de la ciudad. Para esto existen indicadores para medir dicha propagación, entre ellos se encuentra el Número Efectivo de Reproducción que corresponde a la cantidad de personas que puede infectar un caso positivo.

Por lo cual en este trabajo se pretende realizar las estimaciones del Número Efectivo de Reproducción para los casos positivos de de Covid-19 en la ciudad de Santiago de Cali, esto con el fin de analizar el comportamiento de la propagación del virus en los diferentes sectores de la ciudad y junto con variables socioeconómicas asociadas a los barrios observar si existe relación entre la propagación del Covid-19 y aquellas variables mediante unos modelos estadísticos; estos modelos permitirán observar la dinámica y las características que influyen para que se presente una mayor o menor propagación del virus en los barrios de la ciudad y así en futuras epidemias las autoridades locales tengan herramientas para actuar y mitigar situaciones que pongan en riesgo la vida de la población.

1.3. Objetivos

1.3.1. Objetivo general

Analizar la propagación del virus Covid-19 en la ciudad de Santiago de Cali e identificar si existe relación con factores socioeconómicos asociados a los barrios de la ciudad de Santiago de Cali para el periodo Marzo del 2020 - Mayo 2021.

1.3.2. Objetivos específicos

- Estimar el Número Efectivo de Reproducción (R_t) para los casos confirmados por comunas, barrios y para la ciudad de Santiago en el periodo Marzo del 2020 - Mayo 2021.
- Realizar un análisis exploratorio espacial del Número Efectivo de Reproducción en las comunas de la ciudad de Santiago de Cali en el periodo Marzo del 2020 - Mayo 2021.

- Modelar la relación entre la velocidad de propagación del Covid-19 en tres instantes de tiempo y los factores socioeconómicos asociados a los barrios de la ciudad de Santiago de Cali para el periodo Marzo 2020 - Mayo 2021.

2 Antecedentes

En esta sección se presenta una revisión bibliográfica del estado del arte, que constituye varios desarrollos de investigación acerca de las enfermedades similares al Covid -19 en el ámbito nacional e internacional. También se indaga acerca de las metodologías estadísticas utilizadas en trabajos previos para el estudio del fenómeno de interés y la estimación del Número Efectivo de Reproducción.

A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics (Cori et al., 2013)

Desarrollaron una herramienta genérica y robusta en Microsoft Excel y el software estadístico R para estimar el Número Efectivo de Reproducción (R_t) en función del tiempo. Para la creación y evaluación de esta herramienta se analizaron 5 epidemias históricas que variaron en términos de transmisibilidad, intervalo de serie y tamaño de la población. Para cada día t de cada epidemia, se estimó el Número Efectivo de Reproducción (R_t). Llegaron a las siguientes conclusiones:

- Para la epidemia del Sarampion en Hagelloch, Alemania en 1861. El Número Efectivo de Reproducción R_t estimado inicialmente fue de 4.3, lo cual indicaba que se debían tomar medidas de inmediato. Posteriormente, el Número Efectivo de Reproducción R_t alcanzo un nivel máximo de 11.5. Finalmente, a partir de la semana 7 disminuyó hasta el final de la epidemia.

- Para la gripe pandémica en Baltimore, Maryland, 1918. Se obtuvo una estimación del Número Efectivo de Reproducción R_t inicial de 1.4, luego del paso de los días este indicador alcanzo el pico mas alto de esa epidemia a mediados de la semana 5 con un valor de 2.4. Posteriormente dicho valor disminuyo a un valor de 1 después de la séptima semana.

- Para la Viruela en Kosovo, 1972. En el brote pasa un tiempo considerablemente largo desde el primer caso y la estimación del R_t comenzó en la 4 semana con un valor de 3.4. Posterior a ese suceso empezó a aumentar esta estimación alcanzando un valor de 23.9 a mediados de la semana 6 y al finalizar la misma semana empezó a disminuir. En la octava semana disminuyó por debajo de 1 hasta finalizar la epidemia.

- Para el SARS en Hong Kong, 2003. Se tiene que en este brote se presentaron dos picos sucesivos. El primero, ocurrió en la semana 3 con una estimación del Número Efectivo de Reproducción R_t de 2.2. El siguiente pico ocurrió a finales de la semana 6, donde alcanzo un valor de 2.6. Posteriormente el Número Efectivo de Reproducción R_t disminuye su valor hasta el final del brote.
- Finalmente para la Influenza pandémica en una escuela en Pensilvania, 2009. Inicialmente la estimación de R_t se mantuvo de manera constante alcanzando un $R_t = 1.7$ y cayendo por debajo de 1 después de la semana 4. Esto se puede deber al impacto de las medidas de control.

Improved inference of time-varying reproduction numbers during infectious disease outbreaks (Thompson et al. (2019)

En esta investigación los autores proponen analizar las series temporales de incidencia de enfermedades y a partir de ellas estimar la distribución del intervalo de las serial. Esto lo realizan por medio del Número Efectivo de Reproducción (R_t) que permite evaluar la velocidad y dinámica de la propagación de cada una de las enfermedades, teniendo en cuenta los casos transmitidos e importados.

El trabajo realizó el estudio sobre cuatro situaciones:

- **Influencia H1N1 en una escuela de Nueva York (2009)**
- **Influencia H1N1 en una escuela en Pennsylvania (2009)**
- **Enfermedad por el virus del Ébola en Liberia (2014)**
- **MERS en Arabia Saudita (2014–2015)**

Los investigadores llegan a la conclusión que la cuantificación de la transmisibilidad de la enfermedad durante los brotes es crucial para diseñar medidas de control efectivas que permitan mitigar los efectos. Una de las características importantes dentro del marco de datos con el que contaban los investigadores fue los pares de datos de las personas infector / infectadas; de este modo permitió estimar la distribución del intervalo serial que corresponde a tiempo que transcurre desde que el momento que una persona contagiada de algún virus y esta contagie a otra y esta genere síntomas y el Número Efectivo de Reproducción R_t dependiente del tiempo conjuntamente.

Todo esto conllevo a obtener estimaciones más precisas de la transmisión, así como a una cuantificación más precisa de la incertidumbre asociada a estas estimaciones.

Por otro lado, el método planteado por los investigadores les permitió distinguir conjuntos de datos entre casos transmitidos e importados localmente para ser analizados adecuadamente.

Modelación Matemática de la Propagación del SARS-CoV-2 en la Ciudad de Bogota (Mejıa Becerra, 2020)

Modelaron de forma matemática la propagación del SARS-Cov-2 y evaluaron el impacto de las medidas de aislamiento preventivo obligatorio tomadas en la ciudad de Bogotá, Colombia. Se utilizaron datos globales y regionales de personas infectadas con las cuales se alimentó el modelo matemático compartimental SERI3RD para explicar la dinámica de transmisión del virus.

El modelo SERI3RD es un sistema dinámico determinístico que simula la transmisión y evolución de infecciones agudas, a través de este modelo se clasifica la población de estudio en ocho grupos:

- Susceptibles (S)
- Expuestos (E)
- Infecciosos asintomáticos (I_0)
- Infecciosos sintomáticos moderados (I_1)
- Infecciosos sintomáticos severos (I_2)
- Infecciosos sintomáticos críticos (I_3)
- Recuperados (R)
- Muertos (D)

Para calcular la propagación de la infección utilizaron el número reproductivo básico R_0, que indica el número promedio de individuos que pueden llegar a infectarse (casos secundarios) a partir del primer individuo infectado (caso primario) en una población completamente susceptible (Comincini Cantillo et al., 2021).

Al ser un modelo que trabaja con métodos numéricos y la teoría de sistemas dinámicos se plantean una serie de supuestos como:

- (S1) todas las personas de la población se comportan de la misma manera.
- (S2) todos los individuos tienen la misma probabilidad de contagiarse.
- (S3) todas las personas se relacionan entre sí de manera aleatoria.

Posteriormente se estiman los siguientes parámetros:

- (w) que indica el periodo en el cual se desarrolla el virus según el criterio del experto.
- (β_{1t}) es la tasa de transmisibilidad que cambia con el tiempo.
- (β_2 y β_3) es la tasa de transmisibilidad de un individuo en alguno de los grupos de infectados.
- (δ) es la probabilidad de que alguno de los individuos infectados se recupere sin empeorar su condición clínica.
- (ρ) es el tiempo medio de recuperación sin empeorar su condición clínica.
- (σ) es el tiempo medio de complicación de un individuo infectado con el virus.

Finalmente, los investigadores realizan el proceso de simulación en tres etapas. La primera, es un escenario sin cuarentena. El segundo escenario plantea una cuarentena hasta el 27 de abril 2020 y la tercera etapa es una cuarentena hasta el día 20 de junio del 2020.

Los autores concluyen que el modelo es conceptualmente adecuado e interpreta la dinámica de transmisión del virus y la incertidumbre causada por los diversos parámetros, es decir que sirve para realizar una evaluación cualitativa de las medidas tomadas, más no para pronosticar a futuro la cantidad de casos positivos.

Estimation of population infected by Covid-19 using regression Generalized logistics and optimization heuristics (Villalobos-Arias, 2020)

Desarrollan una propuesta para estimar la población contagiada por el Covid-19, sin utilizar los modelos clásicos como el SIR; este tipo de modelos fueron desarrollados por **Kermack y Mc-Kendrick en 1927**. Estos modelos estiman el número teórico de personas que son susceptibles de padecer el virus, el número de individuos infectados y el número de personas que ya no pueden transmitir la enfermedad; bien sean fallecidos o recuperados en una población a lo largo del tiempo (Mikler et al., 2005).Estos modelos presentan dificultades en la estimación de los parámetros si se cuenta con pocos datos.

Este estudio utilizó modelos de Regresión Logística Generalizada y se ajustaron los datos mediante curvas Gompertz, con la intención de estudiar el crecimiento de la población de infectados por el virus.

Los datos fueron extraídos de la página del European Centre for Disease Prevention and Control para el mes de marzo del 2020, tomando registros de países base como China y Corea del Sur principalmente.

Inicialmente, para el ajuste de las curvas se generaron dos situaciones para los registros de china y los registros de corea del sur; en la primera situación se ajustaron las curvas sin corregir los datos y en la segunda situación se ajustaron las curvas con los datos corregidos.

De lo anterior los investigadores obtienen que el comportamiento de la curva ajustada con los datos corregidos era mejor que la curva ajustada sin corregir los datos para los dos países mencionados. Adicionalmente, para las pruebas de predicción realizan mediciones del porcentaje de error relativo, generando una curva de prueba usando la función logística generalizada y la función Gompertz, observando menor porcentaje de error relativo cuando utilizaron la función Gompertz. Por lo tanto, esta función genera mejores predicciones.

Finalmente comparan estos dos métodos utilizando datos de contagiados en Costa Rica; donde obtienen que tanto la curva de la Regresión Logística Generalizada, como la curva Gompertz se adaptan de buena forma a las predicciones de contagiados por el virus a lo largo del tiempo.

Los investigadores concluyen que los modelos propuestos pueden ser utilizados por diferentes países para predecir el crecimiento en el número de contagiados por el virus y formular medidas para contener la propagación del virus.

Estimación del intervalo serial y número reproductivo básico para los casos importados de Covid-19 (Estrada-Alvarez et al., 2020)

Realizaron un estudio cuyo principal objetivo fue realizar la estimación del intervalo serial y el número reproductivo básico en la etapa de contención en la ciudad de Pereira, Colombia en el año 2020.

Este estudio fue cuantitativo de corte transversal sobre la forma en que se desarrolló la transmisión del virus en dicha ciudad; estos investigadores cuentan con información de campo de 12 casos que son confirmados con Covid-19 por un laboratorio a través de pruebas PCR-RT de casos importados y sus correspondientes casos secundarios confirmados; es decir con su red de contagios tanto familiares como sociales.

Para realizar la estimación y ajuste de la distribución del intervalo serial que corresponde a tiempo que transcurre desde que el momento que una persona contagiada de algún virus y esta contagie a otra y esta genere síntomas. Los investigadores utilizaron modelos paramétricos basados en las distribuciones Weibull, Log normal y Gamma. Posteriormente, los investigadores compararon los modelos utilizando el criterio de información Akaike (AIC); así mismo estas personas calcularon a través de máxima verosimilitud los parámetros de las distribución elegida o con mejor ajuste y su respectivo intervalo de confianza calculado por remuestreo, con dichos valores estimaron el valor de la media y el valor de la desviación estándar que fueron de 3.8 y 2.7 respectivamente para la distribución del intervalo serial.

Finalmente, los investigadores concluyen que los datos de las 12 personas infectadas por el virus se ajustan a una distribución Gamma para el intervalo serial con una media de 3.8 días y una desviación de 2.7 días; y un Número Efectivo de Reproducción con un valor de 1.7 con su respectivo intervalo de confianza. Los valores obtenidos para el Número Efectivo de Reproducción son inferiores a los encontrados en otros estudios; a lo largo del mundo al inicio del brote.

Como conclusiones de este estudio los autores afirman que un intervalo serial como el que se estimó en este estudio sugiere un periodo de transmisión mas rápido, es decir desde una etapa de pre sintomático, que según esto puede indicar que se alcanza un pico máximo a los 3.8 días en promedio, lo cual se debe tener en cuenta para posteriores investigaciones en las cuales se recomiendan tomar registros de red de contactos hasta 2 días antes de la fecha del caso inicial.

Análisis espacio-temporal del SARS-COV-2 en la ciudad de Santiago de Cali, Colombia (Cuartas et al., 2020)

Realizaron un estudio cuyo principal objetivo fue describir la distribución espacio temporal del Covid-19 en la ciudad de Santiago de Cali durante el primer mes de la epidemia. Este estudio fue realizado por métodos exploratorio de datos espaciales a través de un análisis de densidad de Kernel, donde verificaron la presencia de patrones espaciales por medio de la función K de Ripley.

Inicialmente los investigadores realizan un análisis exploratorio que comprende localización, distribución, asociación, interacción y evolución espacial de los casos en la ciudad; para esta parte del análisis los investigadores localizan cada caso por medio de puntos a través de métodos de suavización espacial kernel, que permitieron obtener información sobre la distribución y la densidad de los casos infectados. Posteriormente, realizaron un análisis de los patrones puntuales el cual les permitió identificar la concentración de puntos calientes (casos confirmados o probables) en diferentes sectores del área urbana de la ciudad. Finalmente obtienen como resultado que en el periodo de estudio se identifica que el mayor porcentaje de casos se concentra en el área sur de la ciudad, así como otras concentraciones en el oriente de la ciudad; estas en menor frecuencia.

Otra información que permite evidenciar este estudio es que en las primeras semanas epidemiológicas los casos importados son mayores tanto en la zona sur; como en la zona Norte. Por otra parte, para complementar este estudio los investigadores realizan un análisis de tendencia de los casos, con lo cual identificaron que el comportamiento de los casos es creciente las primeras semanas, pero a partir de la implementación de las medidas de cuarentena, dicha tendencia decrece.

Como conclusión, determinan que el patrón espacial identificado puede estar influenciado por las medidas de aislamiento tomadas a nivel local y nivel nacional, además no se puede descartar el efecto que tiene el poco acceso de la población general a las pruebas diagnósticas para detectar el virus, los retrasos y represamientos para conocer los resultados de las mismas y aun los posibles sesgos por dificultades en la técnica de toma de la muestra o su conservación.

3 Marco teórico

En esta sección se presenta el Marco Teórico de la investigación. En la primera parte (Sección 3.1) se muestra el marco conceptual, en el cual se incluyen definiciones importantes del contexto epidemiológico del Covid-19 y la medición de su impacto. En la segunda parte (Sección 3.2) se detalla el marco teórico estadístico, donde se presenta la teoría fundamental para el estudio y su respectivo desarrollo matemático.

3.1. Marco conceptual

A continuación, se presenta algunos conceptos epidemiológicos referentes a la enfermedad Covid-19 que son relevantes para la investigación y a la medición de su impacto.

3.1.1. Coronavirus

Los coronavirus (CoVs) son virus ARN monocatenarios de sentido positivo, poseen envoltura, son altamente diversos y causan trastornos respiratorios, digestivos, hepáticos y neurológicos de severidad variable en un amplio rango de especies animales, incluyendo al ser humano, en quien pueden causar enfermedades graves (Cortés, 2020). Los CoVs se agrupan en cuatro géneros: Alfacoronavirus, Betacoronavirus, Gammacoronavirus y Deltacoronavirus. La primera descripción CoV humano fue hecha en 1965 por Tyrrell y se llamó así debido a las proyecciones desde su superficie que semejaban a una corona (Velázquez-Silva, 2020). Las primeras epidemias fueron causadas por el SARS-CoV (Severe Acute Respiratory Syndrome-CoV) en el 2002 y por el MERS-CoV (Middle East Respiratory Syndrome-CoV) en 2012 (Velázquez-Silva, 2020).

3.1.2. Covid-19

Según la definición de la Organización Mundial de la Salud (OMS), el Covid-19 es una enfermedad infecciosa causada por un virus llamado SARS-CoV-2, que cuenta con un promedio de incubación entre 7 y 14 días. Los síntomas asociados a esta enfermedad van desde tos seca, fiebre, dolor muscular, dificultad para respirar y en algunos casos puede presentar un cuadro clínico de deficiencia respiratoria aguda, choque séptico y sepsis (Huang et al., 2020).

Estudios anteriores de otras epidemias por este tipo de coronavirus afirman que es de origen animal; principalmente se cree que el primer escalón del virus es el murciélago. Sin embargo, todavía se realizan investigaciones para saber si hay un posible segundo huésped como lo puede ser el pangolín salvaje (Aylward et al., 2020).

3.1.3. Transmisión

Es el mecanismo conocido, bien sea de animal a animal, humano a humano o de animal a humano, que puede generar el contagio. En algunos estudios anteriores se encontró que la forma de transmisión de animal a humano se da únicamente mediante secreciones o contacto directo con el animal infectado; se piensa que la transmisión se debe a las relaciones de contacto con el aparato digestivo del animal. En estudios anteriores se conoce evidencia de tres casos; dos relacionados con perros en Hong-Kong y uno relacionado con un gato en Bélgica que presentan contagio por el virus. Una de las hipótesis que manejan es que la raíz del contagio puede haber sido por sus dueños; ya que estos establecían relaciones de contacto aun estando infectados (Knobler, 2004; Sit et al., 2020).

Por otra parte, la trasmisión de humano a humano es más preocupante y prende las alarmas al rededor del mundo; ya que se puede contagiar por secreciones o micro gotas infectadas que miden no más de 5 micras liberadas a través de la tos, estornudos, respiración o habla y pueden alcanzar una distancia de hasta 2 metros (Chin et al., 2020).

Aunque existen otros factores determinantes para la propagación del virus, uno de ellos es el contacto a través de las manos u otra extremidad, pues si están contaminadas y tiene algún tipo de contacto con la boca o nariz inmediatamente puede verse contagiado el individuo. Además de esto, algunas investigaciones afirman que este virus puede durar hasta 30 minutos en la superficie de algunos materiales, según el material y hasta las condiciones de temperatura pueden influir para que este sobreviva.

3.1.4. Incidencia

Se determina como el número de casos nuevos de una enfermedad en un periodo de tiempo determinado en una población especifica. La incidencia permite medir la velocidad a la que se producen casos nuevos en la población expuesta a la enfermedad (Diez-Fuertes et al., 2020).

3.1.5. Prevalencia

Determina el número de casos nuevos y existentes que se encuentran presentes en una población y lugar determinado en un periodo de tiempo; la prevalencia permite medir la magnitud de la enfermedad en la población expuesta a la enfermedad (Diez-Fuertes et al., 2020).

3.1.6. Definición conceptual del Número Efectivo de Reproducción (R_t)

El Número Efectivo de Reproducción (R_t), es el número promedio de casos secundarios de la enfermedad causados por un solo individuo infectado durante su período infeccioso (Cori et al., 2013). El seguimiento del R_t a lo largo del tiempo proporciona información sobre la eficacia de las intervenciones y sobre la necesidad de intensificar los esfuerzos de control que permiten reducir el R_t por debajo del valor umbral de 1 y tan cerca de 0 como sea posible (Cori et al., 2013).

En el caso del Covid-19 se ha estimado que un valor del $R_t = 1$ indica que la epidemia estaría siendo mitigada; es decir que en estos casos una persona contagiada no tendría la capacidad de infectar a otra (Grillo Ardila et al., 2020). Sin embargo, hay que aclarar que este valor es muy susceptible a cambios y se disminuye mediante medidas de aislamiento social e higiene en la población afectada.

3.1.7. Intervalo serial

Representa el tiempo medio que transcurre entre el inicio de la enfermedad de un caso primario y el inicio de la enfermedad de un caso secundario. Se dice que entre mayor sea el tiempo del intervalo serial más tiempo hay para actuar sobre el problema. Por el contrario, si el intervalo serial es pequeño se debe actuar con mayor inmediatez para mitigar los efectos de la enfermedad en la población expuesta (Peláez Sánchez and Más Bermejo, 2020).

3.1.8. Duración de la enfermedad

Indica el tiempo que permanecen los síntomas de la enfermedad; cuando esta se presenta en un estado leve, el tiempo por lo general en promedio de 2 semanas. Por el contrario, si los síntomas de la enfermedad son graves; la duración media pasa de 3 a 8 semanas hasta la recuperación o la muerte de la persona infectada (Aylward et al., 2020). Así mismo el tiempo promedio que tardan en aparecer los síntomas graves de la enfermedad en las personas contagiadas es de una semana aproximadamente.

3.1.9. Vulnerabilidad social

Se define como una serie de condiciones que presentan las personas en un lugar determinado. Por ejemplo, víctimas de los desastres naturales, las situaciones de marginalidad y delincuencia, la discriminación racial o de género, la exclusión social, los problemas de salud mental, etc. Todas estas situaciones generan espacios de vulnerabilidad, un clima o unas condiciones desfavorables que exponen a las personas a mayores riesgos, a situaciones de

falta de poder o control, a la imposibilidad de cambiar sus circunstancias y, por tanto, a la desprotección (Feito, 2007).

3.2. Marco teórico estadístico

En esta sección se presenta el marco teórico estadístico, el cual está divido en dos partes. En la primera parte se define de manera matemática Número Efectivo de Reproducción R_t y su forma de cálculo. La segunda parte está relacionada con los modelos estadísticos, iniciando con el modelo de regresión lineal, modelo de regresión lineal generalizado, hasta llegar a la regresión por mínimos cuadrados parciales (PLS).

3.2.1. Análisis Factorial Múltiple

El Análisis Factorial Múltiple (AFM) fue desarrollado por (Escofier and Pagès, 1992) y hace parte de las técnicas de reducción de dimensionalidad. Es un método factorial adaptado al tratamiento de tablas de datos en las que un mismo conjunto de individuos se describe a través de múltiples grupos de variables. Esta técnica permite analizar múltiples tablas formadas por grupos de variables de diferente naturaleza definidas sobre el mismo conjunto de individuos con la condición de que en cada grupo contengan variables de la misma naturaleza.

Inicialmente los datos están conformados por un conjunto de individuos descritos por varios grupos de variables; cada grupo de variables corresponde a una tabla (Escofier and Pagès, 1990). En cada grupo, las variables deben ser del mismo tipo.

Los símbolos I, J, K y K_j denotan tanto el conjunto como su cardinalidad. Una variable del grupo K_j se denota $v_k(k \in K_j)$. Se tienen individuos y variables con pesos; p_i es el peso asignado a los individuos i ($\sum_i p_i = 1$) y m_k el peso asociado a la variable v_k.

El Análisis Factorial Múltiple es fundamentado en los principios del Análisis de Componentes Principales (ACP) y posee tres etapas.

- **Etapa I:** Cada grupo de variables es asociado a una nube de individuos denominada "Nube Parcial", la cual se analiza de forma independiente por medio de un ACP, ACM o ACS según corresponda el tipo del grupo de variables.
- **Etapa II:** Se asigna un peso a cada uno de los grupos de variables con el fin de equilibrar la influencia de los grupos. Las variables asociadas a cada grupo de variables se les asocia el mismo peso buscando siempre conservar la misma estructura dentro del grupo. Ese peso es calculado en base al primer valor propio asociado a cada grupo

obtenido por medio del ACP, ACM o ACS de la etapa anterior.

$$\frac{1}{\lambda_{1k}}$$

- **Etapa III:** Se calcula un ACP general sobre la tabla global con ponderaciones sobre las variables.

Relación del AFM con el ACP

La mayoría de los métodos multivariados tienen como fin maximizar la inercia de los nuevos ejes ortogonales, para ello se plantea un sistema de maximización teniendo en cuenta la restricción de que los nuevos ejes sean de norma uno ($u_\alpha^{'} M u_\alpha = 1$)

La inercia en AFM para la nube de idividuos esta dada de la siguiente manera:

$$Inercia =^2 p_i(i, G) = \psi' N \psi$$

Donde N es el peso de los individuos, Z es la tabla yuxtapuesta o tabla global y $\psi_\alpha = Z_{global} M u_{alpha}$; siendo M la métrica de las variables, la cual será el inverso del primer valor propio en cada tabla k

$MaxI_\alpha = \psi'_\alpha N \psi_\alpha = u'_\alpha M Z' M u_\alpha$; $u_\alpha^{'} M u_\alpha = 1$

$$L(u_\alpha) = u'_\alpha M Z' N Z M u_\alpha - \lambda(u_\alpha^{'} M u_\alpha - 1)$$

$$L(u_\alpha) = u' M Z' N Z M u - \lambda u_\alpha^{'} M u_\alpha + \lambda_\alpha$$

$$\frac{\partial L}{\partial u_\alpha} = 2 Z' N Z M u_\alpha = 2(Z' N Z M u_\alpha - \lambda u_\alpha)$$

Igualando a cero se tiene el sistema de valores y vectores propios:

$$Z' N Z M u_\alpha = \lambda_\alpha M u$$

Coordenadas en R^p y en R^n

- **Individuos:** $\psi_\alpha = F_\alpha = Z M u_\alpha$
- **Variables:** $\varphi_\alpha = G_\alpha = M Z' N^{1/2} v_\alpha$

3.2.2. Definición matemática del Número Efectivo de Reproducción R_t

Esta teoría fue desarrollada por *Kermack and McKendrick (1927)* en su artículo titulado *"A Contribution to the Mathematical Theory of Epidemics"*; este indicador se define como el número promedio de casos producidos por un solo individuo infectado durante su periodo infeccioso. Este indicador es específico del tiempo y de la situación; se utiliza principalmente para caracterizar la transmisibilidad del patógeno durante una epidemia (Cori et al., 2013).

Se define el número total de casos incidentes que surgen en el periodo de tiempo t, como la suma del número de casos locales I_t^L y el número total de casos importados I_t^I (Thompson et al., 2019).

$$I_t = I_t^{local} + I_t^{importado} \tag{3-1}$$

Según (Cori et al., 2013) se define el número de reproducción dependiente del tiempo R_t, como la relación entre el número de nuevos casos infectados localmente I_t^{local} y el potencial de infección total Λ_t entre todos los individuos infectados en el momento t. Si existe una distribución de intervalo de serie w_s (s = 1,2,...), que representa la probabilidad de que surja un caso secundario en un periodo de tiempo s después del caso primario, entonces cada caso primario (local o importado) que aparece en un periodo de tiempo anterior $t - s$ contribuye a la infección total actual a un nivel relativo dado por w_s. Por lo tanto si se condiciona a w_s, se puede obtener Λ_t como:

$$\Lambda_{(t)} w_s = \sum_{s=1}^{t} (I_{t-s}^L + I_{t-s}^I) w_s = \sum_{s=1}^{t} I_{t-s} w_s \tag{3-2}$$

Dada una distribución de intervalo w_s, datos sobre el número de casos incidentes hasta un periodo anterior (I_0, I_1, ..., I_{t-1}) y el número de reproducción dependiente del tiempo R_t , se tiene que el número esperado de casos incidentes localmente infectados en el tiempo t es:

$$E(I_t^{local} \mid I_0, I_1, ..., I_{t-1}, w_s, R_t) = R_t \Lambda_t(w_s) \tag{3-3}$$

Según (Cori et al., 2013) se puede suponer que el número de casos locales en el tiempo t se extrae de una distribución de Poisson, entonces la probabilidad de observar incidentes locales I_t^{local} en el paso del tiempo es:

$$P(I_t^L / I_0, I_1, ...I_{t-1}, w_s, R_t) = \frac{(R_t \Lambda_t(w_s))^{I_t^L} exp(-R_t \Lambda_t(w_s))}{I_t^L!} \tag{3-4}$$

Según (Cori et al., 2013) se puede suponer que el número efectivo de reproducción es constante durante un período de tiempo $[t-\tau, t]$, con τ que representa la longitud de la ventana de tiempo sobre la cual se estima el R_t . Por lo tanto, la probabilidad de observar la incidencia local $I_t^L, I_{t+1}^L, ..., I_t^L$ durante este período de tiempo, dado el Número Efectivo de Reproducción R_t y condicional a los datos de incidencia anteriores es:

$$P(I_{t-\tau}^L, I_{t-\tau+1}^L, ...I_t^L \mid I_0, ...I_{t-\tau-1}, w_s, R_t) = \prod_{k=t-\tau}^{t} \frac{(R_t\Lambda_k(w_s))^{I_k^L} exp(-R_t\Lambda_k(w_s))}{I_k^L!} \tag{3-5}$$

Usando un enfoque Bayesiano se utiliza una distribución Gamma previa para el R_t como Cori et al. (2013), la distribución posterior del R_t dado los datos de incidencia pasada y condicionada a la distribución del intervalo de serie w_s, es:

$$P(R_t \mid I_0, I_1, ...I_{t-\tau}, I_{t-\tau+1}^L, ...I_t^L, w_s) \propto P(I_{t-\tau}^L, I_{t-\tau+1}^L, ...I_t^L \mid I_0, ...I_{t-\tau-1}, w_s, R_t)P(R_t) \tag{3-6}$$

$$= \left(\prod_{k=t-\tau}^{t} \frac{(R_t\Lambda_k(w_s))^{I_k^L} exp(-R_t\Lambda_k(w_s))}{I_k^L!}\right)\left(\frac{R_t^{a-1} exp(-\frac{R_t}{b})}{\Gamma(a)b^a}\right)$$

$$\propto R_t^{a+\sum_{k=t-\tau}^{t} I_k^L} exp\left(-R_t\left(\sum_{k=t-\tau}^{t} \Lambda_k w_s + \frac{1}{b}\right)\right) * \prod_{k=t-\tau}^{t} \frac{\Lambda_k(w_s)^{I_k^L}}{I_k^L!}$$

Donde se tiene que a y b son parámetros de forma y escala de la distribución Gamma previa. Se usa una distribución Gamma previa conjugada con la probabilidad Poisson, para obtener una formula analítica de la distribución posterior del R_t. De acuerdo con la ecuación 3-6, la distribución posterior para el R_t dado unos datos de incidencia y condicionada a la distribución del intervalo serial w_s, es una distribución Gamma con parámetros:

- **de forma:** $a + \sum_{k=t-\tau}^{t} I_k^L$
- **de escala:** $\frac{1}{\sum_{k=t-\tau}^{t} \wedge_k(w_s) + \frac{1}{b}}$

Finalmente, se dice que las estimaciones posteriores de R_t son un numero positivo mayor o igual a cero, por lo tanto si dicha estimación están por debajo de un valor de uno, indicaría que la epidemia está bajo control.

3.3. Modelación estadística

En la fase modelación estadística se plantea identificar e interpretar la relación entre la propagación del Covid-19 con los datos de la composición socioeconómica correspondiente a los barrios de la ciudad de Santiago de Cali. Para este proceso se cuenta con un conjunto de estimaciones correspondientes al Número Efectivo de Reproducción R_t para cada uno de los barrios; debido a los alcances de este trabajo de grado se considera modelar tres instantes de tiempo que corresponden a los tres picos de contagios identificados en el periodo de Marzo 2020 - Mayo 2021. Se hace necesario llevar esas estimaciones diarias de los picos a una media que generalice la información correspondiente al pico por lo cual se propone trabajar con la mediana de las estimaciones correspondientes a cada pico debido a la distribución asimétrica de los valores del R_t. También se cuenta con variables socioeconómicas asociadas a los barrios, todo esto con el fin entender la posible asociación entre ellas.

En general, el objetivo de los procedimientos de modelado de datos es encontrar una función estadística o algoritmo que permita relacionar las variables explicativas o predictores con la variable de respuesta. Para el caso de la modelación existen dos situaciones, una de ellas surge cuando la variable de respuesta es de naturaleza cualitativa; en ese caso se desarrolla un problema de clasificación. Por otro lado, si la variable de respuesta es de naturaleza cuantitativa se procede a realizar un análisis de regresión; siendo este el más apropiado para modelar el nivel de la propagación del Covid-19 en los barrios de la ciudad de Santiago de Cali.

Lo que se busca es encontrar metodologías que permitan identificar la relación entre la variable respuesta y las variables explicativas a partir de una función matemática. A continuación, se presenta una breve introducción de algunos métodos explorados para dar solución a la problemática planteada.

3.3.1. Modelo de Regresión Lineal

Los Modelos de Regresión Lineal tienen como fin explicar la relación de dependencia entre una variable de respuesta y un conjunto de variables explicativas o predictores. Una de las hipótesis más relevantes que se manejan bajo este tipo de regresiones tiene que ver con que la variable de respuesta (Y) puede ser modelada por la sumatoria del conjunto de variables independientes (X) ponderadas según su contribución y un factor que compone el error, el cual es aleatorio y está compuesto por todos los factores que no se pueden mantener bajo control (Novales, 2010); la estructura se muestra a continuación:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + e \tag{3-7}$$

Se tiene que la variable Y es el predictor; mientras que X_j, representa el conjunto de variables independientes, β_i representa el cambio esperado en la respuesta Y por cambio unitario en

X_i cuando todas las variables regresoras X_i se mantienen constantes y e es el componente aleatorio ligado al error.

Este tipo de modelos lineales deben cumplir con unos supuestos para presentar estimaciones insesgadas; según (Novales, 2010) estos son:

- **Correcta Especificación:** $E[e] = 0$
- **Homogeneidad:** $Var[e] = \sigma^2$
- **Normalidad:** $e \sim N(0, \sigma^2)$
- **Independencia:** $cov[e_i, e_j] = 0 \qquad \forall \;\; i \neq j$

3.3.2. Multicolinealidad

Siguiendo la idea planteada por Vargas and Rodríguez (1980); se puede decir que la multicolinealidad es un problema que surge cuando hay un par o varias variables predictoras están altamente correlacionadas.

Una de las principales dificultades de las estimaciones por mínimos cuadrados es la presencia de este fenómeno que representa un problema grave si su propósito es evaluar la contribución individual de las variables explicativas; esto es debido a que en presencia de multicolinealidad los coeficientes β_j tienden a ser inestables, es decir sus errores estándar presentan magnitudes indebidamente grandes.(Ramírez et al., 2005)

Algunas de las fuentes principales de multicolinealidad son:

- El método de recolección de datos que se empleó.
- Especificación del modelo.
- Un modelo sobre definido (más parámetros que datos).
- La naturaleza de las variables predictoras y su asociación.

Se tiene un modelo $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ y se toma una muestra aleatoria de tamaño n, si se estandarizan las variables entonces:

$$W_{i1} = \frac{X_{i1} - \overline{X}_1}{\sqrt{\sum_{i=1}^{n}(X_{i1} - \overline{X}_1)^2/n-1}} \qquad ; \qquad W_{i2} = \frac{X_{i2} - \overline{X}_2}{\sqrt{\sum_{i=1}^{n}(X_{i2} - \overline{X}_2)^2/n-1}}$$

El modelo estaría definido como: $Y = \beta_0 + \beta_1 W_1 + \beta_2 W_2 + \varepsilon$ y se tiene para la estimación de los $\beta' s$ los siguiente:

$$W = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \\ W_{31} & W_{32} \\ \vdots & \vdots \\ W_{n,1} & W_{n,2} \end{pmatrix} \quad ; \quad W^T W = \begin{pmatrix} \sum_{i=1}^n (W_{i1})^2 & \sum_{i=1}^n W_{i1} W_{i2} \\ \sum_{i=1}^n W_{i2} W_{i1} & \sum_{i=1}^n (W_{i2})^2 \end{pmatrix}$$

$\sum_{i=1}^n (W_{i1})^2 = \sum_{i=1}^n (\frac{X_{i1} - \overline{X}_1}{\sqrt{\sum_{i=1}^n (X_{i1} - \overline{X}_1)^2}})^2 = \frac{\sum_{i=1}^n (X_{i1} - \overline{X}_1)^2}{\sum_{i=1}^n (X_{i1} - \overline{X}_1)^2} = 1$

$\sum_{i=1}^n W_{i1} W_{i2} = \sum_{i=1}^n (\frac{X_{i1} - \overline{X}_1}{\sqrt{\sum_{i=1}^n (X_{i1} - \overline{X}_1)^2}})(\frac{X_{i2} - \overline{X}_2}{\sqrt{\sum_{i=1}^n (X_{i2} - \overline{X}_2)^2}})$

$\sum_{i=1}^n W_{i1} W_{i2} = \frac{\sum_{i=1}^n (X_{i1} - \overline{X}_1)(X_{i2} - \overline{X}_2)}{\sqrt{\sum_{i=1}^n (X_{i1} - \overline{X}_1)^2 (X_{i2} - \overline{X}_2)^2}} = r_{12} = r_{21}$

La matriz $W^T W$ corresponde a una matriz de correlaciones donde la matriz principal serán 1's ya que corresponde a la correlación de la variable W_i consigo misma, mientras que la diagonal secundaria corresponde a la correlación entre la variable W_1 y W_2. Entonces:

$$W^T W = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix} \quad ; \quad (W^T W)^{-1} = \frac{1}{1 - r_{12}^2} \begin{pmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{pmatrix} \quad ; \quad W^T \overrightarrow{Y} = \begin{pmatrix} \overrightarrow{W}_1^T \overrightarrow{Y} \\ \overrightarrow{W}_2^T \overrightarrow{Y} \end{pmatrix}$$

$$\widehat{\overrightarrow{\beta}} = (W^T W)^{-1} W^T \overrightarrow{Y} = \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix}$$

$$\widehat{\overrightarrow{\beta}} = \frac{1}{1 - r_{12}^2} \begin{pmatrix} \overrightarrow{W}_1^T \overrightarrow{Y} - r_{12} \overrightarrow{W}_2^T \overrightarrow{Y} \\ -r_{12} \overrightarrow{W}_1^T \overrightarrow{Y} + \overrightarrow{W}_2^T \overrightarrow{Y} \end{pmatrix} \tag{3-8}$$

De la Ecuación 3-8 se tienen las estimaciones de los parámetros $\widehat{\beta}_1$ y $\widehat{\beta}_2$ que corresponden a:

$\widehat{\beta}_1 = \frac{1}{1 - r_{12}^2} (\overrightarrow{W}_1^T \overrightarrow{Y} - r_{12} \overrightarrow{W}_2^T \overrightarrow{Y}) \quad ; \quad \widehat{\beta}_2 = \frac{1}{1 - r_{12}^2} (\overrightarrow{W}_2^T \overrightarrow{Y} - r_{12} \overrightarrow{W}_1^T \overrightarrow{Y})$

Si X_1 y X_2 están linealmente correlacionadas, entonces $|r_{12}| \approx 1$. Lo cual implica que:

$$\widehat{\overrightarrow{\beta}} = \frac{1}{1 - r_{12}^2} \begin{pmatrix} \overrightarrow{W}_1^T \overrightarrow{Y} - r_{12} \overrightarrow{W}_2^T \overrightarrow{Y} \\ -r_{12} \overrightarrow{W}_1^T \overrightarrow{Y} + \overrightarrow{W}_2^T \overrightarrow{Y} \end{pmatrix} = \frac{1}{1 - 1^2} \begin{pmatrix} \overrightarrow{W}_1^T \overrightarrow{Y} - r_{12} \overrightarrow{W}_2^T \overrightarrow{Y} \\ -r_{12} \overrightarrow{W}_1^T \overrightarrow{Y} + \overrightarrow{W}_2^T \overrightarrow{Y} \end{pmatrix} \to \infty$$

$Var(\widehat{\beta}_j) = C_{jj} \sigma^2 \quad ; \quad C = (W^T W)^{-1}$

$Var(\widehat{\beta_1}) = C_{11}\sigma^2 = \frac{1}{1-r_{12}^2}\sigma^2 = \frac{1}{1-1^2}\sigma^2 \to \infty$

$Var(\widehat{\beta_2}) = C_{22}\sigma^2 = \frac{1}{1-r_{12}^2}\sigma^2 = \frac{1}{1-1^2}\sigma^2 \to \infty$

$Cov(\widehat{\beta_1}, \widehat{\beta_2}) = C_{12}\sigma^2 = \frac{-r_{12}}{1-r_{12}^2}\sigma^2 = \pm\frac{1}{1-1^2}\sigma^2 \to \infty$

Llegando a la conclusión que entre más correlacionadas estén las variables puede dar pie a una mala interpretación del modelos e incorrectas inferencias.

Este tipo de problemas son muy frecuentes en las áreas de economía y mercadeo, un claro ejemplo es el caso donde están presentes variables asociadas al ingreso familiar y los áctivos; las cuales posiblemente tendrán una alta correlación.

Existen dos casos de multicolinealidad:

- **Exacta:** Ocurre cuando dos variables tienen una correlación casi perfecta; es decir su correlación toma valores cercanos a 1 o -1, lo que ocasiona que al estimar los parámetros por mínimos cuadrados ordinarios, el determinante de $(X^TX)^{-1}$ sea igual a cero y con esto $(X^TX)^{-1}$ no existe. Por lo tanto, esto imposibilita la estimación de los parámetros β.
- **Parcial:** Se presenta cuando una variable predictiva tiende a ser combinación lineal de la otra; lo cual genera una alta correlación entre ellas, pero no al nivel de la multicolinealidad exacta. Este fenómeno puede ocasionar que un modelo de regresión parezca ser significativo; incluso puede presentar un coeficiente R^2 alto, pero el problema real ocurre en la matriz de covarianzas de β; ya que presentará varianzas excesivamente grandes, haciendo que los coeficientes estimados sean muy sensibles ante pequeños cambios en los datos.

Identificación de la multicolinealidad

Existen diversas técnicas para identificar la presencia de multicolinealidad, entre esas se encuentran:

Matriz de correlaciones

Una forma muy práctica de determinar el grado de colinealidad es la construcción de una matriz de correlación. Las variables se colocan en filas y en columnas y sus intercepciones deben presentar el coeficiente de regresión lineal de Pearson (Pereira González, 2010). Adicionalmente Mason and Perreault Jr (1991) recomienda que sea eliminada una de las variables que tenga un coeficiente de correlación mayor a 0.8 con otras.

Factores de Incremento de Varianza (VIF)

Otra práctica muy usual consiste en regresar cada columna de X sobre las restantes; un R^2 muy elevado en una o más de dichas regresiones evidencia una relación lineal aproximada entre la variable tomada como regresando y las tomadas como regresores(Tusell, 2011).

Llamemos R_i^2 al R^2 resultante de modelar X_i sobre las restantes columnas de X. Se define el factor de incremento de varianza (variance inflation factor) $VIF(i)$ así(Tusell, 2011) :

$$VIF(i) = \frac{1}{1 - R_i^2}$$

Valores de $VIF(i)$ mayores que 10 (equivalentes a $R_i^2 > 0.90$) se consideran indicativos de multicolinealidad afectando a X_i junto a alguna de las restantes columnas de X (Tusell, 2011).

Finalmente, lo que se recomienda para evitar este tipo de inconvenientes en los estudios es tener información confiable; es decir tener un proceso generador de datos bien estructurado, así como utilizar métodos de selección de variables qué a su vez permiten extraer variables redundantes que pueden generar sesgos en el modelo.

3.3.3. Modelo de Mínimos Cuadrados Parciales (PLS)

La regresión por mínimos cuadrados parciales es una metodología multivariante, versátil, que permite modelar las relaciones entre una o más variables y así dar explicación a un fenómeno de interés.

La regresión por mínimos cuadrados parciales (PLS) fue introducida por (Wold, 1975) para que fuese aplicada inicialmente en las ciencias económicas y sociales. Sin embargo, gracias a diferentes contribuciones este tipo de metodología ha ganado espacio en otro tipo de áreas; donde se realizan estudios que tienen muchas variables como predictores, los cuales presentan inconvenientes de multicolinealidad y pocas observaciones de estudio.

Según Vega-Vilca and Guzmán (2011) la idea inicial de los mínimos cuadrados parciales (PLS) fue heurística; por este motivo todavía se desconocen algunas propiedades. Sin embargo, algunos de los nuevos hallazgos los han realizado Helland (1988) con la estructura del modelo, Höskuldsson (1988) presentando los métodos, Stone and Brooks (1990) entre otros.

En general, se puede decir que la regresión por Mínimos Cuadrados Parciales (PLS) tiene dos pasos principales. Según Vega-Vilca and Guzmán (2011), uno de ellos es transformar la matriz de variables predictoras (X_i) de tamaño $n*p$ con ayuda del vector de respuesta Y

de tamaño n*1. Este procedimiento genera una matriz de componentes no correlacionados denotada como $T = (T_1, ..., T_p)$ de tamaño n*p, la cual es denominada como componentes PLS.

El segundo paso permite calcular el modelo de regresión estimado usando el vector de respuesta original y como variables predictoras los componentes encontrados en el primer paso. Finalmente, se puede decir que la idea de utilizar la metodología de mínimos cuadrados parciales (PLS) es reducir la dimensionalidad, con la garantía de que las primeras componentes ortogonales obtenidas mejoran la predicción de la variable de respuesta (Márquez Ruiz, 2017).

Según (Gaviria Peña, 2016), algunos de los factores importantes que motivan al uso de este tipo de metodologías son:

1. El modelo de regresión PLS es un potente método de regresión lineal, que considera la multicolinealidad en las variables explicativas y acepta un número muy grande de variables.

2. El modelo resultante predice la(s) respuesta(s) a partir de un conjunto de variables linealmente dependientes $x_1, x_2, , x_n$.

3. Durante el desarrollo del modelo, un relativo número pequeño de componentes PLS son calculados y utilizados para la regresión.

4. El número de componentes PLS determina la complejidad de el modelo y puede ser optimizado para tener un alto rendimiento en la predicción.

El PLS, se basa en los conceptos del Análisis de Componentes Principales y la Regresión, desarrollados inicialmente por ***Herman Wold*** en 1975 (Vega-Vilca and Guzmán, 2011). El método PLS permite adaptarlo en situaciones donde se presentan datos faltantes, multicolinealidad y mayor número de individuos que variables (González Rojas, 2016).

Normalización de los datos

Este tipo de método permite transformar variables que están medidas en diferentes escalas a una sola escala común, esto con el fin de que dichas variables puedan se comparables entre si. Existen diversas formas de normalizar los datos. Según (Gaviria Peña, 2016) dos las formas mas utilizadas en la literatura son:

Forma 1

Esta forma consiste en restar para cada una de las variables su media y dividir por la raíz cuadrada de la suma de los cuadrados de las desviaciones de su media.

$$y_i = \frac{y_i - \bar{y}}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad Para \;\; i = 1, 2, ..., n \tag{3-9}$$

$$x_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}} \qquad Para \;\; i = 1, 2, ..., n \qquad j = 1, 2, ..p \tag{3-10}$$

Forma 2

La segunda forma consiste en restar para cada una de las variables su media y dividir por la raíz cuadrada de la suma de los cuadrados de las desviaciones de su media dividido por $n1$.

$$y_i = \frac{y_i - \bar{y}}{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}} \qquad Para \;\; i = 1, 2, ..., n \tag{3-11}$$

$$x_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}{n-1}}} \qquad Para \;\; i = 1, 2, ..., n \qquad j = 1, 2, ..p \tag{3-12}$$

Los coeficientes de regresión de las variables x_{ij} de la primera normalización (Ecuación 3-18) y los coeficientes de regresión de las variables x_{ij} de la segunda normalización (Ecuación 3-18); son los mismos. Esta afirmación puede encontrarse en (Delfa and Calleja, 2003)

Regresión PLS1

Se conocen dos tipos de Regresión PLS. El primero de ellos denominado PLS 1, este método se utiliza cuando se tiene una sola variable de respuesta y un conjunto de p variables explicativas Wold (1975).

La Regresión que se quiere estimar viene dada por:

$$Y = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + e \tag{3-13}$$

La Ecuación 3-13 presenta el modelo de Regresión PLS en el cual solo se tiene una variable de respuesta Y y un conjunto de variables explicativas X_1,X_2,...,X_p, que pueden estar correlacionadas y presentar sobre dimensionalidad; es decir $p>n$. Es necesario que la matriz X y el vector Y estén estandarizados. Con ello se busca obtener componentes ortogonales $t = Xw$ que estén altamente correlacionados con Y en el espacio de las variables predictoras, tal que se obtenga la siguiente Regresión:

$$Y = c_1 t_1 + c2 t_2 + ... + c_p t_p + Y_H \tag{3-14}$$

Para posteriormente, utilizar el desdoblamiento de las $t = f(x)$ (presentado en la Ecuación 3-15), que son combinación lineal de las X_i gracias a w, estimar el modelo expresado en la Ecuación 3-13

$$\beta_{PLS} = W(P^T W)^{-1} C \tag{3-15}$$

Siguiendo la idea de la metodología de análisis de componentes principales, lo que se busca es proyectar cada vector X_i sobre un vector w, teniendo en cuenta que $w'w = 1$.

Lo mencionado se puede observar en la siguiente gráfica (Wold, 1975):

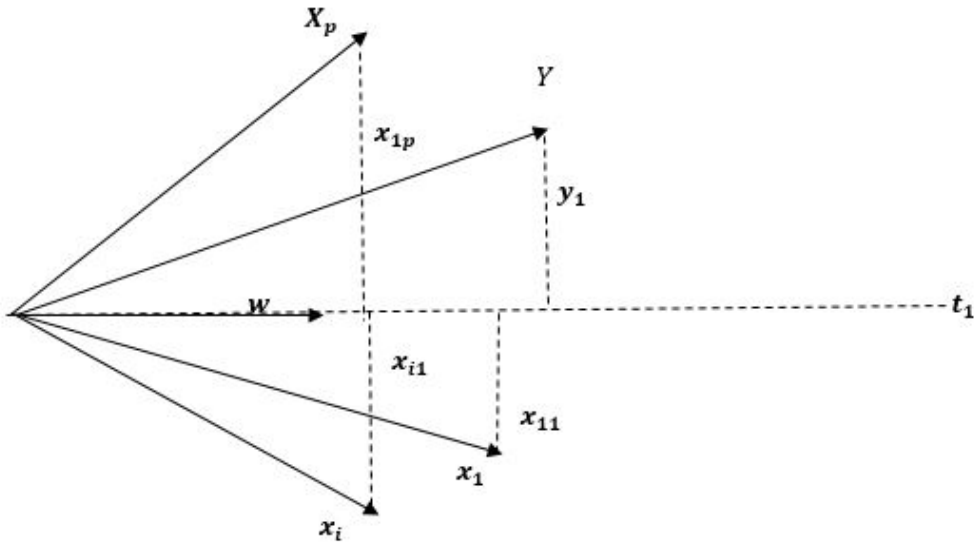

Figura 3-1: Proyecciones de los vectores X_i y el vector Y sobre un vector w

En la Figura **3-2** se puede observar que forman un triangulo rectángulo entre el vector Y, los residuos y_1 y la componente principal t_1. El triangulo formado se puede observar a continuación:

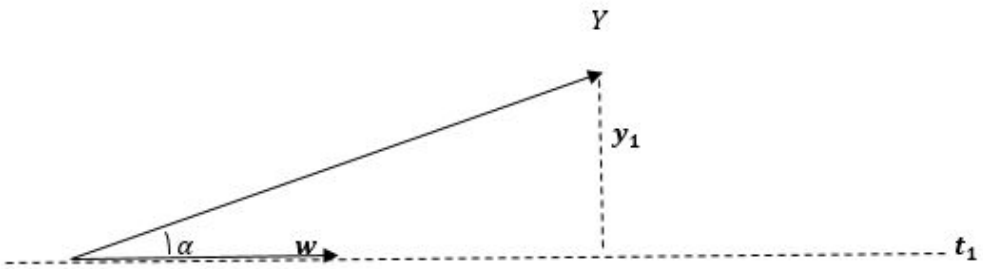

Figura 3-2: Triangulo rectángulo formado por Y, y_i y la componente t_1.

Si utilizamos el teorema de Pitagoras y despejamos la componente t_1, se obtiene lo siguiente:

$$Y^2 = t_1^2 + y_1^2 \Rightarrow Y^2 - y_1^2 = t_1^2$$

Lo que se busca con esto es minimizar la proyección y_1^2 y maximizar la componente t_1^2 de forma geométrica, todo esto se consigue haciendo que la componente t_1 se encuentre altamente correlacionada con la variable Y. En el plano geométrico de la Figura **3-2**,el coseno equivale a la correlación, es decir que si el ángulo de α es de 90° grados, entonces $\cos(\alpha)$=0, lo que indicaría que Y es ortogonal a w, es decir que es independiente de w.

Pero si α es igual a 0°, entonces el $\cos(\alpha)$=1; esto indicaría que Y está superpuesto sobre w, por lo tanto Y tiene una correlación perfecta con w. Dicho esto, al maximizar el coseno o la correlación es lo mismo que maximizar t_1, entonces se tiene que:

$$t_1^2 = (Cor(t_1, Y))^2 = (Cor(Xw, Y))^2 \tag{3-16}$$

La covarianza entre variables estandarizadas se establece como:

$$Cov(X,Y) = \sum_{i=1}^{n}(\frac{(x_i-\bar{x})}{\sigma_x} - 0)(\frac{(y_i-\bar{y})}{\sigma_y} - 0) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sigma_x\sigma_y} = Cor(X,Y)$$

Utilizando las propiedades de la covarianza y asumiendo una constante k, se puede demostrar que:

$$Cov(kX, Y) = kCov(X, Y)$$

Aplicando esto en la Ecuación 3-16 se obtiene:

$$t_1^2 = (Cor(Xw, Y))^2 = (Cov(Xw, Y))^2 = (w'Cov(X, Y))^2$$

Ahora, sea V una matriz de tamaño p*1, del vector de covarianzas X e Y, es decir ($V = X'Y$), entonces se tiene:

$$(w'Cov(X, Y))^2 = (w'V)^2 = w'VV'w$$

La función lagrangiana ϕ que maximiza $w'VV'w$, esta sujeta a la restricción de ortogonalidad de w. La función ϕ es:

$$\phi = w'VV'w - \lambda(w'w - 1)$$

Por lo tanto, si se deriva con respecto a w se tiene:

$$\frac{\partial}{\partial w}\phi = 2VV'w - 2\lambda w = 0$$

Lo que arroja como resultado:

$$VV'w - \lambda w \tag{3-17}$$

De la Ecuación 3-17 tenemos que w y λ son el vector y valor propio respectivos de VV'. También se tiene que VV' es simétrica.

Ahora, si multiplicamos la Ecuación 3-17 a ambos lados por w', se obtiene:

$$w'VV'w - \lambda \tag{3-18}$$

Complementando esta idea, también se puede multiplicar la Ecuación 3-17 a ambos lados por V' de tal modo que se obtiene $(V'VV'w - \lambda V'w) = 0$, luego de realizar una factorizacion queda la siguiente expresión $(V'V - \lambda)V'w = 0$. De esto se puede observar dos situaciones:

$$(V'V - \lambda) = 0$$

$$V'w = 0$$

No obstante, se descarta la segunda situación debido a que V es el vector que se busca maximizar y w por su naturaleza, no puede ser cero. Por esto $(V'V - \lambda) = 0 \Rightarrow \lambda = V'V$ o lo que seria igual:

$$\lambda = \|V\|^2 \tag{3-19}$$

De la Ecuación 3-21 se puede decir que:

$$\lambda^2 = (V'V)(V'V) = \lambda \|V\|^2$$

Por lo tanto, se obtiene:

$$\frac{V'}{\|V\|} V'V \frac{V'}{\|V\|} = \lambda \tag{3-20}$$

Finalmente, tenemos que al igualar las Ecuaciones 3-18 y 3-20 daría lo siguiente:

$$\frac{V'}{\|V\|} V'V \frac{V'}{\|V\|} = w'VV'w \tag{3-21}$$

Por lo tanto, el vector w que maximiza la covarianza es:

$$w = \frac{V}{\|V\|} = \frac{XY}{\|XY\|}$$

De esta forma la primera componente será determinada como $t_1 = Xw$ y posterior se realiza la regresión de Y sobre la componente t_1, de la siguiente forma:

$$Y = c_1 t_1 + Y_1 = c_1 w_1 1 x_1 + c_1 w_{12} x_2 + ... + c_2 w_{1p} x_p + Y_1$$

Donde el coeficiente de regresión está representado por $c_1 = Y^T t_1$, el vector de residuos es Y_1, además se calcula P que representa el coeficiente de regresión de X_j sobre t_1, es decir que $P = X^T t_1$, y se actualiza la matriz X y el vector Y.

Con esto se puede calcular una segunda componente en el caso en que el poder explicativo de la regresión $Y = t_1$ no sea suficiente, en ese caso se procede de manera similar con el calculo a estimar el modelo $Y = t_1 + t_2$.

El desdoblamiento de las componentes $t = f(x)$ para presentar el modelo de Regresión por Mínimos Cuadrados Parciales (PLS) con los coeficientes de regresión estimados en función de la variables iniciales, se realiza con la Ecuación 3-15. Este procedimiento se realiza hasta encontrar el numero de componentes necesarios, aunque en la literatura se puede encontrar que diferentes autores plantean que el numero de componentes es dos o tres, ya que después de allí el poder explicativo del modelo no va aumentar de manera significativa (González Rojas, 2016).

Regresión PLS2

La regresión PLS2 se entiende como una extinción del PLS1, en el caso que se tiene que explicar un conjunto de variables $y_1, y_2, ..., y_q$ mediante de un conjunto de variables explicativas $x_1, x_2, ..., x_p$. Entonces, como afirma en (Delfa and Calleja, 2003) la regresión PLS2 consiste en aplicar el método de componentes principales de un conjunto de variables $x_1, x_2, ..., x_p$, bajo la condición que estas expliquen de la mejor manera el conjunto de variables $y_1, y_2, ..., y_q$.

Para el caso del PLS2 se construyen matrices X y Y, donde X tiene columnas formadas por vectores x_i con $i = 1, 2, ..., p$, mientras que Y esta formada por vectores columna y_k con $k = 1, 2, .., q$. Por lo tanto, cada vector de x_i y y_k pertenece al espacio R^n, mientras que las matrices X y Y pertenecen a los espacios vectoriales $R_{n\times p}$ y $R_{n\times q}$.

Mencionado lo anterior, se puede afirmar que la regresión PLS2 consiste principalmente en realizar proyecciones simultaneas de ambos espacios sobre hiperplanos de menor dimensión. Las coordenadas de estos puntos generadas en los hiperplanos constituyen elementos de las matrices T y U.

Finalmente, se puede decir que la bajo la metodología de la regresión PLS2 se pueden alcanzar dos objetivos básicos. El primero es maximizar la correlación entre el conjunto de variables $x_1, x_2, ..., x_p$ y $y_1, y_2, ..., y_q$. El segundo es aproximar mediante hiperplanos lo mejor posible a los espacios vectoriales generados por $x_1, X_2, ..., x_p$ y $y_1, y_2, ..., y_q$, es decir que se recoja toda la información posible.

Aspectos matemáticos de la regresión

En la regresión PLS2 los datos de las filas que componen las matrices X y Y provienen de n individuos, donde X contiene las p características, mientras que Y describe las q propiedades. Asumiendo que las matrices X y Y están centradas, el objetivo de la regresión PLS2 es determinar una relación lineal:

$$Y = XB + E$$

entre las variables x y y, usando una matriz B de orden $p\times q$ de coeficientes de regresión y una matriz de errores E. Mientras que en el PLS1 todo esto queda reducido a tener $y = X\beta + e_1$, en lugar de determinar directamente esta relación, se tiene que las matrices X y Y son modeladas mediante el uso de variables latentes en base a los modelos de regresión:

$$X = TP^T + E_X \quad \text{y} \quad Y = UQ^T + E_Y$$

Con sus respectivas matrices de errores E_X y E_Y. Las matrices de scores T y U, junto con las matrices de pesos P y Q tiene un numero de a columnas, donde $a \leq min(n, q, p)$ es el número de componentes PLS. Si tenemos que u_j, t_j, q_j y p_j denotan la j-ésima columna de U, T, P y Q, entonces se tiene la siguiente relación entre estas:

$$u_j = d_j t_j + h_j$$

Donde h_j son los residuales y d_j son los parámetros de la regresión. Si la relación entre u_j y t_j es fuerte; es decir h_j es pequeño, entonces tenemos que los x-scores de la primera componente pueden predecir bien los y-scores y en consecuencia predicen bien y-datos. En la regresión PLS2 se consideran varias componentes, por lo tanto :

$$U = TD + H$$

Donde D es una matriz diagonal que en su diagonal se encuentran los elementos $d_1, d_2, ..., d_a$, mientras que H es una matriz residual compuesta por las columnas h_j.

El objetivo de la regresión PLS2 es maximizar la covarianza entre los dato x-scores y los y-scores, como el problema de la maximización no es único, entonces se hace necesario una restricción en los vectores escores, tal que $\|t\| = \|u\| = 1$.

Continuando, se obtiene los vectores scores por medio de las proyecciones de las matrices X y Y en los vectores de cargas. como en (Delfa and Calleja, 2003) los vectores de carga utilizados sera w para las x variables y c para las y variables. Por lo tanto, $t = Xw$ y $u = Yc$, con esto el problema de maximización sera:

$$\text{MAX } cov(Xw, Yc)$$

$$s.a \begin{cases} \|t\| = \|Xw\| = 1 \\ \|u\| = \|Yc\| = 1 \end{cases} \tag{3-22}$$

Donde tenemos que la *cov* denota la covarianza simple. Las soluciones a este problema de maximización son t_1 y u_1, de esta forma se pueden obtener los siguientes scores, los cuales se calculan de manera similar pero deben adicionarse nuevas restricciones, estas restricciones están relacionadas a la ortogonalidad de los scores previos, es decir que $t_j^T t_k = 0$ y $y_j^T u_k = 0$ para $1 \leq j \leq k \leq a$.

Finalmente, cuando en el problema de optimización de la Ecuación 3-22 se toma la covarianza simple, se obtiene:

$$\text{MAX } t^T u = (Xw)^T (Yc) = w^T X^T Y_c$$

$$s.a \begin{cases} \|t\| = \|Xw\| = 1 \\ \|u\| = \|Yc\| = 1 \end{cases} \tag{3-23}$$

Las soluciones para w y c se pueden calcular a partir de la descomposición en valores singulares de X^TY, entre las posibles direcciones de los vectores w y c, la solución mas óptima al problema de la Ecuación 3-23 se alcanza con los vectores w_1 y c_1, que corresponden al valor singular mas grande de X^TY (Höskuldsson, 1988).

Algoritmo PLS2

Se tiene un conjunto de variables explicativas $x_1, x_2, ..., x_p$ y un conjunto de variables respuesta $y_1, y_2, ..., y_q$, a partir de dichas variables se construyen las matrices X y Y, donde X tiene como columnas los vectores x_i con $i = 1, 2, ..., p$, mientras que Y tiene como columnas los vectores y_k con $k = 1, 2, ..., q$. Dicho esto se tiene los siguientes pasos:

1. Se construyen las matrices X_0 y Y_0 que están conformadas por variables centradas de las variables es predictoras y las variables de respuesta respectivamente.

2. se construye una combinación lineal u_1 de las columnas de Y_0 y una combinación lineal t_1 de las columnas de X_0, de modo que estas maximicen $cov(u_1, t_1)$. De esta manera se obtiene dos nuevas variables u_1 y t_1 lo mas correlacionadas posibles y que permiten resumir de la mejor manera la información contenida en las matrices Y_0 y X_0.

3. Luego se construye una regresión lineal simple del conjuntó de variables explicativas y también del conjunto de variables respuesta sobre la componente t_1

$$X_0 = t_1 p_1^T + X_1$$

$$Y_0 = t_1 q_1^T + Y_1$$

   Donde p_1 y q_1 son los vectores de coeficientes de regresión.

   Se repite la etapa anterior, pero remplazando las matrices Y_0 y X_0 por las nuevas matrices Y_1 y X_1, con esto se obtienen nuevas componentes t_2 y u_2 que permiten maximizar la $cov(t_2, u_2)$. A partir de estas componentes se obtiene por medio de regresión lineal simple:

$$X_1 = t_2 p_2^T + X_2$$

$$Y_1 = t_2 q_2^T + Y_2$$

   Por lo tanto, se puede deducir que:

$$X_0 = t_1 p_1^T + t_2 p_2^T + X_2$$

$$Y_0 = t_1 q_1^T + t_2 q_2^T + Y_2$$

Todo este proceso se repite hasta que las componentes $t_1, t_2, ..., t_h$ puedan explicar de la mejor manera a Y_0. De la siguiente descomposición:

$$Y_0 = t_1 q_1^T + t_2 q_2^T + ... + t_h q_h^T + Y_h$$

Se deduce la ecuación para el PLS2 (Höskuldsson, 1988):

$$y_k* = \hat{\beta}^*_{k,0} + \hat{\beta}^*_{k,1} x_1 + \hat{\beta}^*_{k,2} x_2 + ... + \hat{\beta}^*_{k,p} x_p \quad \text{para} \quad k = 1, 2, ..., q$$

4 Metodología

En este sección se describe la metodología utilizada para desarrollar el presente trabajo de grado. Este componente metodológico está dividido en dos partes: La primera hace referencia a la descripción de la población de estudio, la zona de estudio, descripción de los conjuntos de datos y un breve análisis exploratorio espacial. La segunda parte contiene todo lo referente al análisis estadístico, el cual fue realizado en todo el periodo de estudio de forma continua, es decir de Marzo 2020 - Mayo 2021 la cual se subdivide en el Análisis exploratorio, donde se describen algunos comportamientos previos de los datos, la estimación del Número Efectivo de Reproducción (R_t) a nivel de la ciudad de Santiago de Cali en general y por comunas, donde se calcula este indicador con los datos contenidos en el archivo de contagios.

Finalmente, se tiene la Modelación Estadística, en la cual se realiza a partir de la estimación del Número Efectivo de Reproducción R_t a nivel de barrios, pero en este caso se trabajó con los tres picos medianos identificados en el periodo de estudio (marzo 2020 - Mayo 2021)por barrio como variables respuesta. Posteriormente con ayuda extraída del Departamento Administrativo Nacional de Estadística (DANE) asociarle información socioeconómica de los barrios y así por medio de la metodología de Mínimos Cuadrados Parciales 2 (PLS2) por medio de tres modelos (uno para cada pico) identificar la relación existente entre la propagación del Covid-19 con variables socioeconómicas asociadas a los barrios de la ciudad de Santiago de Cali.

4.1. Archivo de datos

Este trabajo se realizó basado en la información de los registros de los casos confirmados de Covid-19 en la ciudad de Santiago de Cali. Es decir aquellas personas visitantes o residentes de la ciudad que hayan contraído y notificado el virus del Covid-19 en el periodo Marzo 2020 - Mayo 2021.

En este trabajo de grado se consideraron todos los registros de los casos diarios confirmados reportados por la Secretaria de Salud Municipal de la ciudad de Santiago de Cali, que estuvieran localizados dentro del área metropolitana de la ciudad, con fechas de confirmación comprendidas entre el 10 de Marzo del 2020 y el 30 de Mayo del 2021. Además de esto los

casos confirmados deben contar con la información de localización (coordenadas geográficas dentro del perímetro urbano, ver Figura **4-1**).

A continuación se describen aquellos criterios de inclusión y exclusión que se tuvieron en cuenta en la construcción del conjunto de datos para la elaboración de este trabajo:

Criterios de inclusión en el archivo de datos

- Personas de cualquier edad, sexo, síntomas y estado del caso.
- Personas que se encuentren dentro del área metropolitana de la ciudad de Santiago de Cali.
- Aquellos registros que contaran con la dirección del hogar.

Criterios de exclusión en el archivo de datos

- Se excluyeron los individuos que no se encontraban en el área metropolitana de Cali.
- Los individuos que no contaban dirección registrada o barrio
- Los individuos que no contaban con la fecha del inicio de los síntomas.

4.2. Limitación geográfica del estudio

Esta investigación se realizó en la ciudad de Santiago de Cali, la capital del departamento del Valle del Cauca, Colombia. Esta ciudad se compone de 15 corregimientos en la zona rural, 22 comunas y 249 barrios en el perímetro urbano. La información extraída de (Planeación Municipal , 2020) se puede observar en la Figura **4-1**.

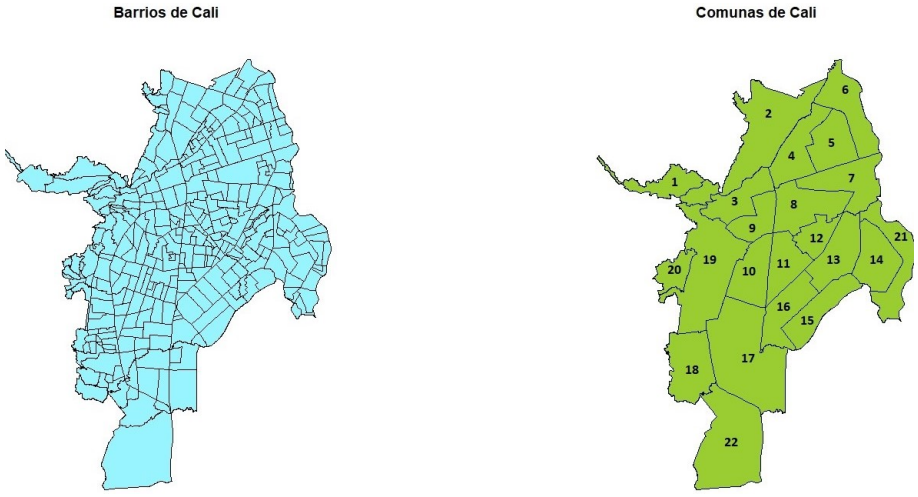

Figura 4-1: Municipio de Santiago de Cali dividido por barrios (Derecha) y Municipio de Santiago de Cali dividido por Comuna (izquierda).

La ciudad está localizada entre la cordillera occidental y la cordillera central de los Andes; además tiene más de 2.383.392 de habitantes en el área urbana según la Alcaldía de Santiago de Cali (2021), siendo la tercer ciudad más poblada de Colombia.

4.2.1. Conjuntos de datos

Para este estudio se contó con dos archivos de datos. El primer archivo de datos fue proporcionado por la Secretaría de Salud Municipal de Santiago de Cali. Este archivo de datos se le denominó registros de contagios diarios, el cual inicialmente contaba con un total de 198.556 registros confirmados de casos positivos por Covid-19, comprendidos entre el periodo del 10 de marzo del 2020 hasta el 30 de mayo del 2021.

Al archivo de registros de contagios diarios se le realizó una limpieza, donde se identifican 696 registros que no contaban con la información de la fecha de inicio de síntomas y otros 46.634 registros que correspondían a casos confirmados que se encontraban fuera del área urbana de la ciudad de Santiago de Cali. Finalmente, este archivo de datos queda con un total de 152.226 registros de casos confirmados validos para la elaboración de este trabajo de grado.

En la Tabla **4-1** se muestran las variables contenidas en este archivo de datos y una breve descripción de cada una de ellas:

Variable	**Descripción**
Sexo	Sexo del individuo (Masculino o Femenino)
Edad	Edad del individuo
Fecha de inicio de síntomas	Fecha en la que el individuo manifiesta que inició los síntomas
Barrio	Barrio donde reside el individuo en la ciudad de Cali
Estado del caso	Estado del individuo al momento de la prueba (leve, moderado, grave o severo)
Comorbilidades	Si el individuo presentaba o no comorbilidades (se encontró un 0,7 % de datos faltantes)
Síntomas	Si el individuo era sintomático o asintomático (se identificó un 1,1 % de datos faltantes)
Comuna	Comuna donde reside el individuo en la ciudad de Cali

Tabla 4-1: Descripción de las variables contenidas en el archivo de datos de registros de contagios diarios de Covid-19 en la ciudad de Santiago de Cali.

El segundo archivo de datos con el que se contó para este trabajo de grado fue extraído de la pagina oficial del Departamento Administrativo Nacional de Estadística (DANE). Este

archivo contiene información de las condiciones socioeconómicas del último censo nacional de población y vivienda realizado por el DANE en el año 2018.

En la Tabla **4-2** se muestran las variables contenidas en este archivo de datos y una breve descripción de cada una de ellas:

Variable	**Descripción**
Barrio	Nombre del barrio de la ciudad de Cali
Área	Área del barrio en m^2 de la ciudad de Cali
Uso vivienda residencial	Conteo de viviendas de uso residencial
Uso vivienda comercial	Conteo de viviendas de uso comercial
Viviendas	Cantidad de viviendas por barrio
Hogares	Cantidad de hogares por barrio
Habitantes	Cantidad de personas que viven en el barrio
Viviendas con energía	Conteo de viviendas que cuentan con energía eléctrica
Viviendas con acueducto	Conteo de viviendas que cuentan con acueducto
Viviendas con alcantarillado	Conteo de viviendas que cuentan con alcantarillado
Viviendas con recolección de basuras	Conteo de viviendas que cuentan con recolección de basuras
Viviendas con gas natural	Conteo de viviendas que cuentan con gas natural
Viviendas con internet	Conteo de viviendas que cuentan con internet
Edad	Conteo de habitantes por rangos de edad (0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, mayores de 80)
Hombres	Conteo de hombres por barrio
Mujeres	Conteo de mujeres por barrio
Nivel educativo preescolar	Conteo de habitantes con un nivel máximo educativo alcanzado de preescolar
Nivel educativo secundaria	Conteo de habitantes con un nivel máximo educativo alcanzado es secundaria
Nivel educativo técnico, tecnológico o profesional	Conteo de habitantes con un nivel máximo educativo alcanzado es técnico, tecnológico o profesional
Estrato moda	Estrato modal por barrio
IPM	Índice de pobreza multidimensional

Tabla 4-2: Descripción de variables asociadas a los datos de la composición socio - económica de los barrios de la Ciudad de Santiago de Cali.

Para el presente trabajo se realiza una depuración y solo se dejan registros con información asociada a la ciudad de Santiago de Cali. Este archivo de datos se le denominó como composición socioeconómica de los barrios de la ciudad de Santiago de Cali. En el están contenidas variables asociadas a los barrios como: cantidad de viviendas de uso comercial, cantidad de viviendas de uso residencial, viviendas con servicios de energía, acueducto, alcantarillado, etc; que permiten caracterizar las condiciones en que se encuentran los habitantes dentro de los barrios de la ciudad.

Adicionalmente, con la información contenida en el archivo de datos de las condiciones socioeconómicas de los barrios de la ciudad de Santiago de Cali, se construyeron unas variables que permite resumir dicha información, reducir parámetros a estimar y evitar la posible multicolinealidad en el modelo; estas variables se presentan a continuación en la Tabla **4-3**:

Variable	**Descripción**
Prop. uso residencial	Proporción de viviendas del barrio de uso netamente residencial
Prop. uso comercial	Proporción de viviendas del barrio de uso netamente comercial
Uso comercial / uso residencial	Cantidad de viviendas de uso residencial sobre las viviendas de uso comercial
Indicador de servicios básicos	Porcentaje de viviendas que cuentan con agua, luz, alcantarillado y recolección de basuras
Indicador de servicios complementarios	Porcentaje de viviendas que cuentan con Gas natural e internet
Indicador de escolaridad básica	Porcentaje de personas que alcanzaron primaria y secundaria
Indicador de escolaridad superior	Porcentaje de personas que alcanzaron el pregrado o un postgrado
Razón de masculinidad	Cantidad de hombres sobre la cantidad de mujeres
Índice de Juventud	Personas con edad menor a 19 años sobre las mayores a 60
Densidad poblacional	Densidad de población por barrio, calculado como el total de habitantes por 1000 sobre el área del barrio

Tabla 4-3: Descripción de variables construidas para la modelación

Adicionalmente me menciona de donde surgieron los indicadores contenidos en la Tabla **4-3**

- **Indicador de servicios básicos:** Este indicador fue propuesto por nosotros y corresponde a la cantidad de viviendas que contaban con todos servicios básicos como lo es el agua, luz, alcantarillado y recolección de basuras sobre la cantidad de viviendas del barrio.
- **Indicador de servicios complementarios**:Este indicador fue propuesto por nosotros y corresponde a la cantidad de viviendas que contaban con ambos servicios complementarios como lo es el internet y el gas domiciliario sobre la cantidad de viviendas del barrio.
- **Indicador de escolaridad básica:** Este indicador fue propuesto por nosotros y corresponde a la cantidad de personas que alcanzaron la secundaria completa con respecto al total de habitantes del barrio.
- **Indicador de escolaridad superior:** Este indicador fue propuesto por nosotros y corresponde a la cantidad de personas que alcanzaron algún estudio posterior a la secundaria como lo es una carrera técnica, tecnológica, profesional o algún tipo de postgrado con respecto al total de habitantes del barrio.
- **Índice de juventud:** Revisando literatura diversos autores proponen diversos indicadores de juventud pero debido a las limitaciones que se presentaba en la estructura de los datos, se propuso esta extensión contemplando los menores a 19 años sobre los mayores a 60 años.

Para ilustrar lo dicho anteriormente se presenta la siguiente figura:

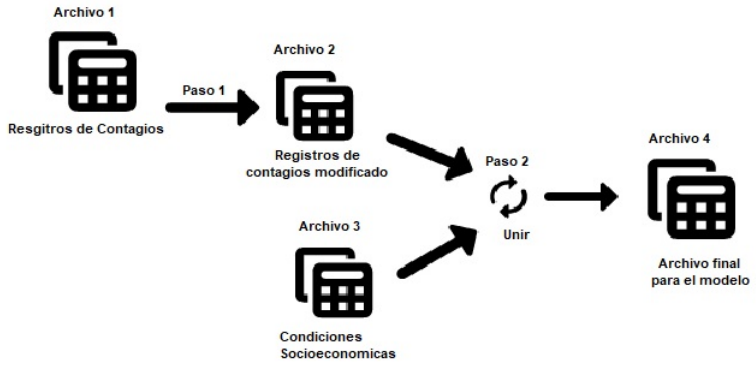

Figura 4-2: Ilustración del proceso de georreferenciación

La Figura **4-2** ilustra el proceso de configuración del archivo final, con el cual se realiza la modelación. Se cuentan con los siguientes archivo de datos:

- **Archivo de datos 1:** Este archivo contiene los registros de individuos contagiados por el virus Covid -19, con todas las variables que nos proporciono la Secretaria de Salud Municipal.
- **Archivo de datos 2:** Este archivo contiene los registros de individuos contagiados por el virus Covid -19 referenciados por barrio.
- **Archivo de datos 3:** Este archivo contiene los registros de las condiciones socioeconómicas de los barrios de la ciudad de Santiago de Cali.
- **Archivo de datos 4:** Este archivo contiene las estimaciones del R_t y las condiciones socioeconómicas de los barrios, con el cual se realiza la modelación.

Inicialmente se tiene un archivo de datos de contagios (Archivo 1) con el cual a través de las variables latitud y longitud se gerreferencia cada registro (paso 1) en su correspondiente barrios dentro de la ciudad, con estos se obtiene otro archivo de datos de contagios el cual ya cuenta con información sobre el barrio donde se presenta el caso (Archivo 2).

Posteriormente, con el nuevo archivo de datos de contagios (Archivo 2) se hace análisis espacial y temporal del contagio, así como la estimación del R_t por barrio y en la ciudad. Después de obtener las estimaciones del R_t por barrio, se realiza un cruce (Paso 2) con el archivó de datos de condiciones socioeconómicas (archivo 3) mediante la variable barrio que se encuentra presente en los dos archivos (archivo 2 y archivo 3).

Finalmente, se obtiene un archivo de datos final con 249 registros correspondientes a los barrios de la ciudad, este archivo contiene las estimaciones de R_t para los tres periodos de estudio y las condiciones socioeconómicas asociadas a esos barrios, con este archivo final s realiza la modelación.

4.3. Georreferenciación

Para la georreferenciación se hizo uso del archivo de datos de registros de contagios diarios suministrado por la Secretaría de Salud Municipal de Cali. Así, cada registro contenido en este archivo de datos fue georreferenciado dentro del perímetro urbano de la ciudad de Santiago de Cali (ver Figura **4-1**), por medio de coordenadas geográficas planas contenidas en el archivo de datos y medidas en las variables latitud y longitud para cada uno de los registros de casos confirmados en la ciudad.

Cabe resaltar que no todos los registros contaban con información sobre las coordenadas o no estaban registradas de forma correcta lo cual conllevó a que dichos registros con información errónea sean ubicados a través de la información presente en la variable barrio del mismo archivo de datos y así obtener un archivo de datos con un total de 152.226 casos confirmados de Covid-19.

4.4. Análisis estadístico

En esta sección se describe el análisis estadístico utilizado en la elaboración de este trabajo de grado. La primera parte está compuesta por el análisis exploratorio de los dos archivos de datos, donde se presentan algunas características y condiciones de las personas contagiadas por el virus y el conjunto de datos de que contiene la información socioeconómica de los barrios en la ciudad de Santiago de Cali.

En la segunda parte se realiza la estimación del Número Efectivo de Reproducción (R_t), utilizando el archivo de datos de registros de contagios diarios suministrado por la Secretaria de Salud Municipal. Para esta estimación se hace uso de la variable fecha de confirmación, la cual se estima el número de casos diarios en el periodo Marzo 2020 - Mayo 2021, con dicha información se realiza la estimación del R_t a nivel de la ciudad, barrio y comuna. posteriormente, con los valores de la estimación se procede a realizar un análisis exploratorio espacial a nivel de comuna en la ciudad de Santiago de Cali.

Finalmente, se realiza la Modelación Estadística del Número Efectivo de Reproducción (R_t) por barrio como variable respuesta en tres instantes de tiempo, los cuales hacen referencia a los picos presentados en el periodo de estudio Marzo 10 2022 hasta Mayo 30 del 2022.

Por medio de un Modelo de Minimos Cuadrados Parciales 2 (PLS2) fueron calculados los tres modelos asociados a los tres instantes de tiempo ya mencionados. Cabe resaltar que esta metodologia se caracteriza por hacer la estimación simultanea de las variables respuesta (en este caso 3). Los resultados estadísticos fueron obtenidos principalmente con el software R Core Team (2019), utilizando algunos de los paquetes como: Epiestim, Ggplot2, geoforest, MASS entre otros.

4.4.1. Análisis exploratorio

En este trabajo se realizó un análisis descriptivo univariado con las variables más relevantes (cualitativas y cuantitativas) contenidas en los dos conjuntos de datos, a través de gráficos de barras, cajas, tendencia para las variables cuantitativas y tablas de frecuencia para las diferentes variables cualitativas.

Por otra parte, a través de estadísticas como la media, mediana, desviación estándar, máximo y mínimo, se realizó un análisis de las variables cuantitativas. Complementariamente, se realizaron mapas de densidad de los casos confirmados de Covid-19 mes a mes, mapas de densidad de casos acumulados del virus en el periodo estudiado (Marzo 2020- Mayo 2021) de forma continua y la razón de la propagación del virus mes a mes (casos confirmados por comuna sobre la cantidad de habitantes de la comuna) en la ciudad, todo esto para observar el comportamiento de la propagación del virus en la zona urbana de la ciudad de Santiago de Cali.

4.4.2. Estimación del Número Efectivo de Reproducción (R_t)

Siguiendo la metodología implementada por Cori et al. (2013) a partir de la información suministrada por la Secretaria de Salud Municipal de Santiago de Cali, contenida en el archivo de datos de registros de contagios diarios se realizó un conteo de casos diarios durante el periodo de Marzo 2020 - Mayo 2021 ya que es el insumo fundamental junto con la distribución del intervalo serial para realizar la estimación del Número Efectivo de Reproducción. Con la ayuda del software estadístico (R Core Team, 2019) y el paquete "*EpiEstim*" realiza la estimación del Número Efectivo de Reproducción R_t suministrandole la serie de casos confirmados por día y la información de la media y la desviación estándar en días del intervalo serial que como se menciono en la Sección 4 fue extraído de *Estrada-Alvarez et al. (2020)*.

Distribución del Intervalo Serial

En este trabajo de grado no se contó con registros de casos confirmados, con una red de contactos identificada, la cual permitiría obtener información sobre los casos primarios y los casos secundarios del virus, es decir obtener fechas, cantidades de personas portadoras del virus, posibles contagiados por dicho caso primario y tiempo de incubación del virus. Por lo tanto, sin dicha información se vuelve complejo saber o determinar el periodo de incubación del virus en cada persona infectada. Es así como se vuelve necesario tener información sobre la distribución del intervalo serial.

Mediante estos parámetros se determina el tiempo medio entre el inicio de síntomas de un caso primario y un caso secundario, para posteriormente realizar la estimación del Número Efectivo de Reproducción (R_t). Mencionado esto, para determinar los valores de los parámetros para la media y la varianza de la distribución del intervalo serial, se recurre a los diferentes estudios realizados que se encuentran en la literatura.

Algo que hay que tener en cuenta, es que entre más altos los valores de dichos parámetros, la forma del intervalo serial este será más amplio y de esta forma se tendría más tiempo para actuar y tomar medidas de control sobre la enfermedad.Caso contrario pasa si los

valores de los parámetros son pequeños; entonces la forma del ancho del intervalo serial seria más reducida, porque el tiempo de propagación es mucho menor.

Uno de esos estudios fue realizado por *Viego et al. (2020)* donde lo que hicieron fue estimar el periodo de incubación e intervalo serial del virus Covid-19 en la ciudad de Bahía Blanca durante Marzo-Mayo del 2020 con una muestra de de pacientes sintomáticos. Adicionalmente se tiene el estudio realizado por *Estrada-Alvarez et al. (2020)* que consistió en estimar el intervalo serial y el número efectivo de reproducción del Covid-19 entre casos importados durante la fase de contención en Pereira, Colombia, en Marzo del 2020.

Por lo tanto, en este trabajo de grado se utilizaron las estimaciones realizadas por *Estrada-Alvarez et al. (2020)*; debido a que se considera que estas estimaciones de los parámetros de la distribución del intervalo serial son mas coherentes al comportamiento local, por ser realizadas con información regional. La estimación de la distribución del Intervalo Serial se ajustaron varios modelos paramétricos como lo fueron la distribuciones Weibull, Log-normal y Gamma; luego, se compararon entre sí utilizando el criterio de información de Akaike; posteriormente se calcularon las estimaciones por máxima verosimilitud de los parámetros de la distribución elegida con mejor ajuste con su respectivo intervalo de confianza, calculados por remuestreo *Estrada-Alvarez et al. (2020)*.

Llegando a la conclusión que la distribución a la que más se ajusta es una Gamma con parametros estimados por máxima verosimilitud que vienen dados por α =1.96 (IC95 %1.1-5.9) y β =0.51 (IC95 % 0.26 - 1.7) bajo la siguiente paramtrización *Estrada-Alvarez et al. (2020)*:

$$f(\frac{x}{\alpha}, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

Finalmente, Se estimó un intervalo serial con un valor promedio de 3,8 días y una desviación estándar de 2,7 días*Estrada-Alvarez et al. (2020)*.

Después de establecer los parámetros de la distribución del intervalo serial, se continua con la estimación del Número Efectivo de Reproducción (R_t) con la ayuda el software estadístico R Core Team (2019) y el paquete $''EpiEstim''$. Para el cálculo de este indicador solo se tiene en cuenta los casos diarios y no la cantidad acumulada; siguiendo la metodología implementada por *Cori et al. (2013)*, donde a través del enfoque bayesiano realizan la estimación del Número Efectivo de Reproducción (R_t). En dicho estudio se asume que los casos nuevos registrados a lo largo de los días siguen una distribución de probabilidad Poisson y complementariamente; también se asume una distribución Gamma como distribución previa para el R_t.

Después de realizar la estimación del Número Efectivo de Reproducción (R_t) por comunas y barrios, el primero con el fin de realizar un análisis exploratorio espacial para identificar zonas de la ciudad donde su propagación fuera más rápida o tendencias y el segundo se establece como variable de respuesta dependiente para el modelo utilizado; esto es debido a que si se trabajaba el modelo a nivel de comunas solo se contaba 21 observaciones para la construcción del modelo. Posteriormente junto a otras variables socioeconómicas que fueron extraídas del archivo de datos de los barrios de la ciudad de Santiago de Cali se construye un modelo identificar la relación de la propagación del Covid-19 y variables socioeconómicas.

Modelo general de la situación

Los datos de los incidentes nuevos modifican el valor que puede tomar el R_t en la fecha en que se estima, todo esto asumiendo que el valor del R_t hoy esta relacionado con el valor previo del día anterior R_{t-1}, por tal motivo se diferentes autores como (Bettencourt and Ribeiro, 2008) proponen el uso de la regla de bayes para calcular las estimaciones día a día del valor del R_t. El teorema de bayes es el siguiente:

$$P(R_t \mid k) = \frac{P(k \mid R_t)P(R_t)}{P(k)} \tag{4-1}$$

En la Ecuación 4-1 se entiende que si hay k nuevos casos confirmados, entonces la probabilidad del R_t es igual a la probabilidad de haber observado k nuevos casos dado un valor del R_t, multiplicada por la probabilidad previa $P(R_t)$ y dividida entre la probabilidad de observar k casos.

Al realizar las iteraciones para cada nuevo día, se hace uso de la probabilidad del día anterior $P(R_{t-1})$ como la probabilidad previa $P(R_t)$. Según (Bellot, 2020) se puede asumir que la distribución del R_t es gausiana centrada al rededor de R_{t-1}, tal que:

$$P(R_t \mid R_{t-1}) = N(R_{t-1}, \sigma) \tag{4-2}$$

Donde σ es un parámetro que se estima posteriormente, aplicando esta ecuación en la estimación del primer día se obtiene:

$$P(R_1 \mid k_1) \propto P(R_1).L(R_1 \mid k_1) \tag{4-3}$$

La Ecuación 4-3 permite interpretarse como la probabilidad de R_1 dado que se observaron k_1 casos es proporcional a la probabilidad de observar el valor R_1 multiplicado por la verosimilitud de haber observado k_1 casos dado que R_t haya adoptado el valor R_1. Así, se tiene una secuencia que puede ser estimada para el segundo día como:

$$P(R_2 \mid k_1, k_2) \propto L(R_2 \mid k_2) = \sum_{R_1} P(R_1 \mid k_1).P(R_2 \mid R_1).L(R_2 \mid k_2) \tag{4-4}$$

Ahora, el problema se traslada a encontrar una función de máxima verosimilitud $L(k_t \mid R_t)$ que permita estimar la probabilidad de estimar k casos nuevos dado un valor del R_t. Mencionado esto, (Bellot, 2020) afirma que dada una tasa media de recurrencia λ de nuevos casos por día, la distribución de probabilidad de observar k nuevos casos, puede ser modelada por una distribución Poisson.

$$P(k \mid \lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{4-5}$$

Estos debido a que la distribución de Poisson modela la variabilidad del número de nuevos casos al rededor del número de casos esperados λ. Ahora para conectar λ con el R_t se muestra la siguiente ecuación, obtenida por (Bettencourt and Ribeiro, 2008), que se usa para reparametrizar la función de verosimilitud:

$$\lambda = k_{t-1}.e^{Y(R_{t-1})} \tag{4-6}$$

En la Ecuación 4-6 Y es el reciproco del intervalo serial de la enfermedad, el cual es el tiempo medio entre el reporte de dos casos, en los cuales el segundo es derivado del primero. Finalmente, Se multiplica la verosimilitud por la probabilidad previa (que es la verosimilitud del día anterior) para obtener la probabilidad posterior del R_t.

4.5. Modelación

Como se ha presentado a lo largo del documento, se cuenta con dos conjuntos de datos, el primero contiene los registros de los contagios diarios notificados en la Ciudad de Santiago de Cali, con los cuales se estimó el Número Efectivo de Reproducción (R_t) a nivel de barrios; esta estimación fue utilizada como variable de respuesta.

El segundo archivo cuenta con la composición socioecónomica de los barrios de la ciudad de Santiago de Cali; las variables seleccionadas de este archivo fueron la cantidad de hombres por barrio, la cantidad de mujeres por barrio, el conteo de viviendas que cuentan con servicios (energía, acueducto, alcantarillado, recolección de basura, gas, internet), conteo de viviendas

de uso (residencial, comercial), conteo de hogares, conteo de viviendas, densidad poblacional, edad por rangos (0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, mas de 80), grado de escolaridad (primaria, secundaria, media, profesional,posgrado).

Adicionalmente, se construyeron otras variables como índice de juventud, indicador de escolaridad, cociente entre viviendas de uso residencial y comercial; donde la descripción de cada uno de dichos indicadores esta en la Tabla **4-3** y la forma de construcción en la Sección 4. Vale la pena aclarar que la información socioeconómica asociada a los barrios es fija ya que fue tomado del ultimo censo realizado por el Departamento Administrativo Nacional de Estadística (DANE) en el año 2018; como se presneta en la Tabla **4-4**.

Para este trabajo se cuenta con dos estructuras de datos, una de ellas es fija (Condiciones socioeconómicas) y la otra es variable en el tiempo (R_t). Por lo tanto, esto agrega un nivel de complejidad importante que debido al alcance de un trabajo de grado no se abordara en la parte del modelo estadístico todo el periodo de estudio (Marzo 2020 - Mayo 2021) de forma continua; sino que se considera únicamente modelar la propagación de virus mediano por barrio en los 3 picos de contagios identificados en el periodo de estudio (Marzo 2020 - Mayo 2021).

Información Socioeconómica					
Barrio	**V1**	**V2**	**V3**	$\cdots$	**V54**
Barrio 1	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$\cdots$	$X_{1,54}$
Barrio 2	$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	$\cdots$	$X_{2,54}$
Barrio 3	$X_{3,1}$	$X_{3,2}$	$X_{3,3}$	$\cdots$	$X_{3,54}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ldots$	$\vdots$
Barrio 249	$X_{249,1}$	$X_{249,2}$	$X_{249,3}$	$\cdots$	$X_{249,54}$

Tabla 4-4: Estructura de la base socioeconomica de los barrios de Santiago de Cali

La intención principal de esta sección es construir unos modelos estadísticos que reproduzcan la relación entre la propagación del virus Covid-19 y variables asociadas a características de los barrios de la ciudad de Santiago de Cali; donde se quiere identificar si estas variables afectan de forma similar la propagación del virus Covid-19 y adicionalmente dicha afectación es la misma en los tres instantes de tiempo o si se presentan variaciones en cada uno de los periodos.

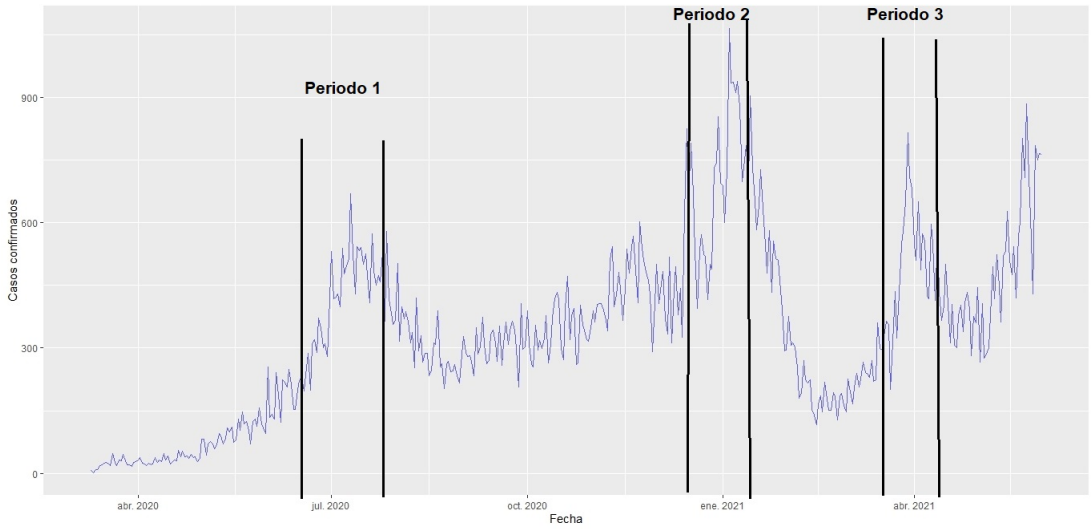

Figura 4-3: Serie del número de casos diarios confirmados en el Periodo Marzo 10 del 2020 a Mayo 30 del 2021 en la Ciudad de Santiago de Cali

Los tres instantes de tiempo seleccionados fueron elegidos teniendo en cuenta los picos presentes en la Figura **4-3**, estos picos son:

- **Tiempo 1:** 15 Junio a 14 de Julio del 2020.
- **Tiempo 2:** Enero 2021.
- **Tiempo 3:** Abril 2021.

El primer periodo de tiempo viene dado por el intervalo 15 de Junio al 14 de Julio, en este caso se denomina primer pico de la pandemia para los datos de contagios obtenidos, donde se tuvo un incremento en la propagación del virus y por ello el Número Efectivo de Reproducción presentó unos valores por encima de 2.

El segundo periodo de tiempo es a inicios del año 2021 exactamente en el mes de enero, donde se presenta nuevamente un incremento en el numero de casos confirmados diarios del virus, este segundo pico posiblemente fue generado por los diferentes acontecimientos sociales presentados en la ciudad, algunos de estos acontecimientos fueron: la final del fútbol profesional y celebraciones de fin de año.

Finalmente, el tercer periodo de tiempo contempla el mes de Abril del 2021, donde se evidencia un incremento el el número de casos diarios del virus, este nuevo incremento pudo ser ocasionado por eventos de tipo cultural como la semana santa que se presta para la salida de personas a lugares de descanso para dicha semana.

Para la estimación del Número Efectivo de Reproducción R_t se realizó un conteo del número de casos confirmados diarios en cada uno de los barrios que componen la ciudad de Santiago de Cali, esto para cada uno de los tres periodos seleccionados, así se obtienen un aproximado de 30 estimaciones del R_t para cada barrio y en cada periodo de tiempo.

Posteriormente, se hace necesario resumir ese indicador que estaba de forma diaria por cada barrio. Es decir, resumir dicha información en un solo indicador con el fin de tener una medida que contenga toda la información de la propagación de cada barrio y cada periodo de tiempo a estudiar.

Como se menciona, surgió la necesidad de resumir las estimaciones dirías de los barrios en los tres periodos de tiempo. Por lo tanto, a continuación se presenta la distribución de la mediana y la distribución promedio del Número Efectivo de Reproducción R_t en cada uno de los tres periodos de tiempo.

Estimación del R_t			
Barrio	**T1**	**T2**	**T3**
Barrio 1	$R_{t1,1,1}$	$R_{t1,2,1}$	$R_{t1,3,1}$
Barrio 1	$R_{t1,1,2}$	$R_{t1,2,2}$	$R_{t1,3,2}$
Barrio 1	$R_{t1,1,3}$	$R_{t1,2,3}$	$R_{t1,3,3}$
⋮	⋮	⋮	⋮
Barrio 1	$R_{t1,1,30}$	$R_{t1,2,30}$	$R_{t1,3,30}$
Barrio 2	$R_{t2,1,1}$	$R_{t2,2,1}$	$R_{t2,3,1}$
Barrio 2	$R_{t2,1,2}$	$R_{t2,2,2}$	$R_{t2,3,2}$
Barrio 2	$R_{t2,1,3}$	$R_{t2,2,3}$	$R_{t2,3,3}$
⋮	⋮	⋮	⋮
Barrio 2	$R_{t2,1,30}$	$R_{t2,2,30}$	$R_{t2,3,30}$
⋮	⋮	⋮	⋮
Barrio 249	$R_{t249,1,30}$	$R_{t249,2,30}$	$R_{t249,3,30}$

Tabla 4-5: Estructura del conjunto de datos de la estimación del R_{tijk} por barrios.

La Figura **4-5** hace una representación de la estructura del conjunto de datos del Número Efectivo de Reproducción en los tres periodos de tiempo y en cada uno de los barrios de la ciudad.

En la Figura **4-4** se presenta la distribución de los dos indicadores que fueron consideras en este trabajo de grado, en este caso no se nota una diferencia importante entre la media y la mediana de las estimaciones del Número Efectivo de Reproducción, por lo cual se decide trabajar con el R_t mediano por barrio.

En la Tabla **4-6** se presenta la estructura de las variables dependientes o también llamadas variables respuesta, donde cada Y_r corresponde al R_t mediano de los tres instantes de tiempo mencionados anteriormente.

La Tabla **4-7** que corresponde a la matriz de variables independientes o predictoras donde cada x_p corresponde a cada una de las variables socioeconomicas asociadas a los barrios; estas son las encargadas de explicar a partir de una expresión matemática la relación de la propagación del Covid-19 y variables socioeconomicas. Cabe recordar que cada fila de las matrices Y y X corresponden a los 249 barrios que componen la ciudad de Santiago de Cali.

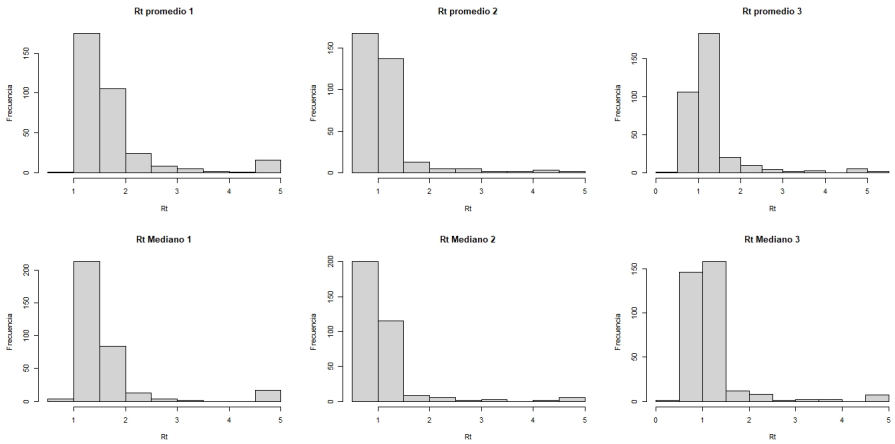

Figura 4-4: Distribución de los indicadores de tendencia central(promedio y mediana) del Número Efectivo de Reproducción R_t para cada instante de tiempo

y_1	y_2	y_3
R_{t1}	R_{t1}	R_{t1}
R_{t2}	R_{t2}	R_{t2}
R_{t3}	R_{t3}	R_{t3}
$\vdots$	$\vdots$	$\vdots$
R_{t249}	R_{t249}	R_{t249}

Tabla 4-6: Matriz de respuestas Y

X_1	X_2	X_3	$\cdots$	X_{16}
$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$\cdots$	$x_{1,16}$
$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$\cdots$	$x_{2,16}$
$x_{3,1}$	$x_{3,2}$	$x_{3,3}$	$\cdots$	$x_{3,16}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
$x_{249,1}$	$x_{249,2}$	$x_{249,3}$	$\cdots$	$x_{249,16}$

Tabla 4-7: Matriz de variables independientes X

Posterior al análisis exploratorio de los datos, con el fin de identificar la metodología apropiada para la modelación de este fenomeno inicialmente se exploraron varias alternativas metodológicas por ejemplo se realizó un análisis de componentes principales para seleccionar unas variables previas que posteriormente fueron incluidas para construir tres modelos de regresión lineal (uno para cada instante de tiempo). Posteriormente, este análisis se complementa con la selección de las variables basados en el criterio de información AKAIKE mediante el método Backward, la cual permitió elegir las variables más significativas para el modelo.

A pesar de que estos resultados no se presentan en este documento, se hace necesario destacar que este modelo no cumplió con las validaciones de supuestos de normalidad e independencia de los datos aun después de realizar transformaciones lineales, lo cual conlleva a una alternativa invalida.

Buscando otra alternativa, se procede a construir tres modelos lineales generalizados (GLM), donde los resultado de estos modelos tampoco serán presentados en este trabajo de grado debido a que el conjunto de datos (R_t mediano en los tres periodos de tiempo) no seguían una distribución de probabilidad conocida de acuerdo a los resultados de las pruebas Anderson Darling y Kolmogórov-Smirnov en las cuales se rechaza las hipótesis de que los datos siguen una distribución de probabilidad conocida.

Finalmente y luego de la implementación de las metodologías anteriores se establece que la metodología mas apropiada es la modelación por PLS2; debido a la estructura de las variables respuesta (tres variables respuesta) ya que este permite la corrección de multicolinealidad y utiliza la correlación entre las variables como fuentes predictoras del modelo.

5 Resultados

En este capitulo se presentan los resultados de la investigación; cabe aclarar que en las Secciones 5.1, 5.2 y 5.3 se analizarón los datos de forma continua, es decir en todo el periodo de estudio (Marzo 2020 - Mayo 2021). En la Sección 5 se implementa la modelación pero únicamente en los tres picos identificados en el periodo de estudio (Marzo 2020 - Mayo 2021).

En la Sección 5.1 se desarrolla el análisis descriptivo asociados a las diferentes variables de los casos confirmados de Covid-19; es decir la composición de cada individuo confirmado por Covid-19 en la ciudad de Santiago de Cali.

En la Sección 5.2 se realiza la estimación del Número Efectivo de Reproducción R_t. Para ello se utilizó como insumo el archivo de datos suministrado por la Secretaría de Salud Municipal, desde el cual se realiza el conteo de los casos diarios y con la ayuda del software estadístico (R Core Team, 2019) y el paquete "*EpiEstim*" realiza la estimación del Número Efectivo de Reproducción R_t suministrandole la serie de casos confirmados por día y la información de la media y la desviación estándar en días del intervalo serial que como se menciono en la Sección 4 fue extraído de *Estrada-Alvarez et al. (2020)*.

Posteriormente en la Sección 5.3 se obtienen las estimaciones del R_t por comunas y a nivel de la ciudad de Santiago de Cali para el análisis espacial exploratorio y se plasman en mapas del perímetro urbano de la ciudad de Santiago de Cali y así observar el comportamiento espacial del contagio en el tiempo.

Finalmente, en la Sección 5.4 se vuelve a estimar el Número Efectivo de Reproducción pero a nivel de barrios analogamente como se contruyó en la Seccion 5.2, con el fin de construir los 3 modelos para identificar la relación entre la propagación del Covid-19 con variables sosioeconómicas asociadas a los barrios en los tres picos de pandemia en el periodo de Marzo 2020- Mayo 2021 por medio de la metodología de Mínimos Cuadrados Parciales 2 (PLS2).

5.1. Análisis descriptivo

La Figura **5-1** resume las gráficas que presentan la edad como variable principal con respecto a otras variables, en la gráfica (a) de la Figura **5-1** se presenta la distribución de la edad por sexo de los casos confirmados. De los 152.226 casos confirmados registrados en la hoja de datos por Covid-19, el 52 % corresponden a hombres y el 48 % corresponde a mujeres.

En general, el comportamiento del número de casos confirmados es similar tanto en hombres como en mujeres, siendo las edades entre los 25 y 35 años las que mayor número de casos confirmados presentan para los dos géneros.

En la Figura **5-1** (gráfica c y d) se muestra la distribución por edad y la presencia de comorbilidades, esto se realiza con el fin de identificar si hay una relación entre dichas variables; se puede evidenciar que el 22.1 % de los casos confirmados presenta al menos una comorbilidad y estas son más frecuentes en las personas cuya edad está alrededor de los 60 años de edad, es decir la población adulta mayor; cabe resaltar que en todos los rangos de edad se presentan comorbilidades.

5.1.1. Análisis descriptivo del archivo del número de contagios diarios

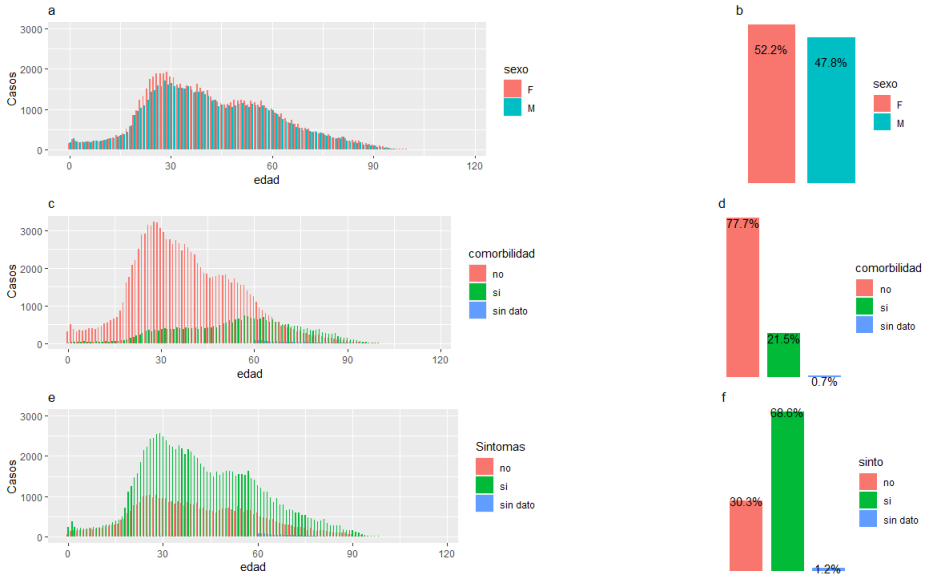

Figura 5-1: (a) Dist. de los casos según la edad y el sexo, (b) Participación porcentual de casos positivos por sexo, (c) Dist. de los casos según la edad y comorbilidad, (d) Participación porcentual de casos positivos por comorbilidad,(e) Dist. de los casos según la edad y los síntomas, (f) Participación porcentual de casos positivos por síntomas

Finalmente, la Figura **5-1**(gráficas e y g) muestran la distribución por edad y síntomas asociados a los contagiados. En esta gráfica se evidencia mayor frecuencia de casos asociados a personas reportadas como sintomáticas (69 %), que personas asintomáticas (29 %) respecto a su edad; adicionalmente se evidencia una mayor frecuencia de casos con síntomas en las edades de 25 a 35 años.

En la Figura **5-2** se observa el comportamiento de la edad de los casos confirmados mes a mes desde el 10 de Marzo del 2020 hasta el 30 de Mayo del 2021. En general se puede decir la edad de los individuos contagiados es similar en cada uno de los meses trascurridos. Algo que también se evidenció en la Figura **5-2** es que la enfermedad era frecuente en personas más jóvenes.

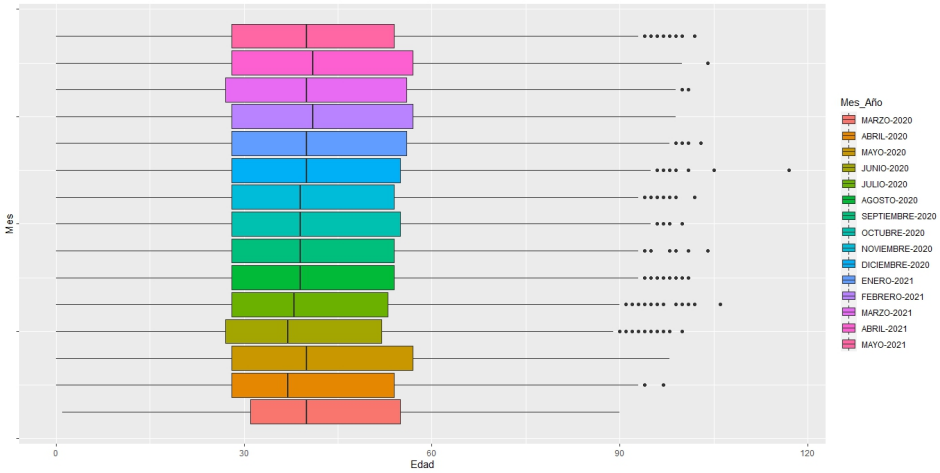

Figura 5-2: Distribución de la edad de los casos positivos de Covid-19 en los meses de estudio (Marzo 2020 -Mayo 2021)

La Figura **5-3** presenta la distribución de la edad de los casos confirmados y el estado de gravedad del caso confirmado. En este caso se observa que la enfermedad se expresa de forma mas severa en personas de mayor rango de edad, esto se evidencia en la gráfica ya que en su mayoría las personas fallecidas son adultos que se encuentran en edades superiores a los 64 años.

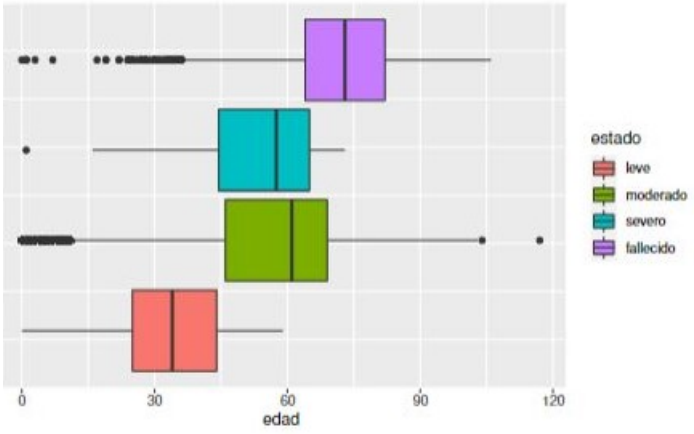

Figura 5-3: Diagrama de cajas de la edad de los casos confirmados en el tiempo según estado de salud, para el periodo de estudio (Marzo 2020 - Mayo 2021)

Por otra parte, en la gráfica **5-3** también se observa que las personas con síntomas leves de la enfermedad, ninguna de ellas sobrepasa los 60 años de edad. Finalmente, de manera general se puede afirmar que esta enfermedad se vuelve más letal si la edad de la persona es mayor.

Representación espacial del contagio

A continuación, en la Figura **5-4**, se muestra a través de un mapa la cantidad de casos de Covid-19 acumulados en todo el periodo de estudio desde el 10 de Marzo del 2022 hasta el 30 de Mayo del 2021 en cada una de las 22 comunas de la ciudad de Santiago de Cali.

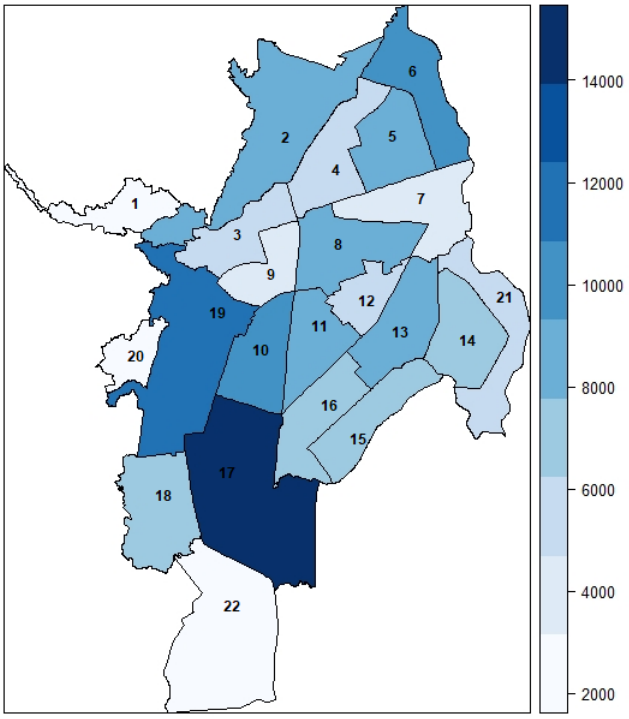

Figura 5-4: Casos confirmados acumulados por comunas para la ciudad de Santiago de Cali en el periodo (Marzo 2020 - Mayo 2021)

En la Figura **5-4** se puede identificar que la Comuna 17 es la que mayor casos confirmados presento a lo largo del periodo de análisis (Marzo 2020 - Mayo 2021) acumulando más de 14.000 casos de Covid-19.

Seguido se encuentran las Comunas 19, 2, 6 y 10 en donde el número de casos acumulados ronda los 12.000 y 13.000, estas comunas presentan similitud en cuanto a la distribución de personas por metro cuadrado, ya que en ellas se encuentran barrios que en su mayoría están compuestos por cuadras donde predomina el acceso netamente peatonal, lo cual permite que la interacción o el contacto entre personas se presente de manera más frecuente y esto pueda estar relacionado con el aumento en la propagación de casos por Covid-19.

Continuando con el análisis, la Figura **5-4**; también permite evidenciar de manera particular que la Comuna 22 (zona inferior - color blanco), que comprende Ciudad Jardín y sus alrededores, es la que menor número de casos acumulados presentan en todo el periodo de estudio y a su vez es una de las comunas con mayor extensión de área.

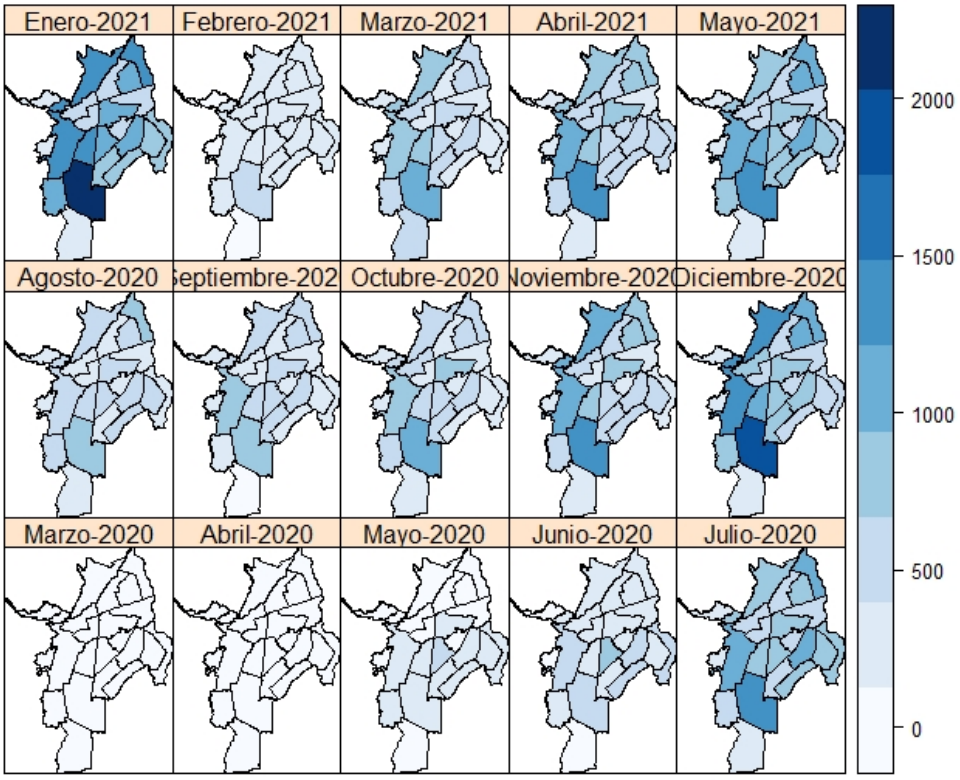

Figura 5-5: Casos confirmados por mes en las comunas de la ciudad de Santiago de Cali

La Figura **5-5** ilustra la evolución de los casos confirmados a través del tiempo desde el mes de Marzo del 2020 hasta el mes de Mayo del 2021. En esta figura se muestra que desde el inicio de la pandemia la Comuna 17 es donde se condensan la mayoría de casos por Covid-19. Este resultado se explica debido a que el virus llegó de forma importada y las personas que pueden tener relaciones y/o mayor contacto con las personas provenientes del extranjero son aquellas de estrato más alto; adicionalmente es la comuna que más casos positivos por mes reportaba como se puede ver en la Figura **5-5**

Otro aspecto relevante que se puede deducir en la Figura **5-5** es la efectividad de la cuarentena y las medidas tomadas por las autoridades locales como el pico y cédula, cierre de

establecimientos nocturnos, clausura de eventos masivos entre otros, en los primeros meses ya que a lo largo del periodo se nota que no hay un incremento desproporcionado de los casos en la mayoría de las comunas de la ciudad.

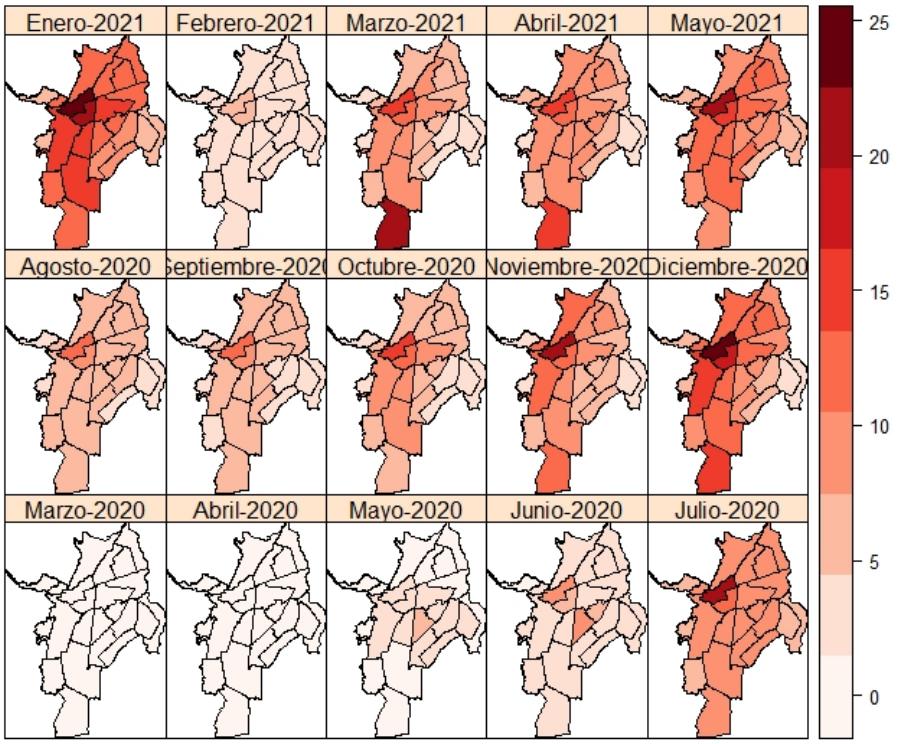

Figura 5-6: Tasa de contagio de Covid-19 en la ciudad de Santiago de Cali por comuna y mes

La Figura **5-6** representa la velocidad de contagio por cada mil habitantes. En este caso, los primeros 5 meses de estudio (Marzo - Julio del 2020) no se observa una diferencia importante de la infección entre las 22 comunas de la Ciudad de Santiago de Cali. Sin embargo, al finalizar Julio se nota un incremento de la velocidad de infección en la Comuna 3 (centro de la ciudad) que presenta un área pequeña en comparación a las otras 21 comunas de las ciudad.

Continuando con el análisis de la Figura **5-6**, se observa que desde el mes de Agosto hasta el mes de Diciembre del 2020 hay leves incrementos en la velocidad de infección o propagación entre comunas, especialmente en los meses de Septiembre a Diciembre.

En estos meses las zonas más afectadas, según muestra la gráfica son las comunas ubicadas en el centro de la ciudad; este comportamiento posiblemente se debe a que en la zona la

densidad de población es mayor y que en esta parte de la ciudad, por sus características comerciales está concentrada en mayor medida la fuerza laboral y comercial de la ciudad.

Finalmente, para el periodo comprendido entre los meses de Enero - Mayo del 2021 se evidencia mayor velocidad de propagación con respecto a los meses anteriores. Es importante resaltar que en estos meses ve afectada de manera fuerte la Comuna 22, en la zona sur de la ciudad. Sin embargo, en esta zona las condiciones de persona por área son distintas a las de la zona centro, ya que en este lugar las necesidades de los habitantes vienen dadas por otras circunstancias. Es decir que posiblemente la infección se propagó porque este segmento de población estuvo en otras actividades, como lo son los viajes ya sea de negocios o turismo.

5.1.2. Análisis descriptivo de la composición socioeconómica de los barrios de la ciudad de Santiago de Cali

En esta sección se presenta la composición de la población de la ciudad de Santiago de Cali; mostrando variables de entorno tanto sociales como demográficas (número de personas en diferentes rangos de edad, número de hogares, número de viviendas, viviendas con servicios básicos, servicios complementarios, escolaridad, etc).

La información que se presenta a continuación está contenida en el archivo de datos extraído de (DANE, 2018). Cabe aclarar que los datos vienen consolidados a nivel de manzanas, pero para este trabajo se realizó limpieza, depuración y consolidación de los registros a nivel de barrio.

Para la transformación de las variables del archivo de datos de la composición socioeconómica de los barrios, que inicialmente se encontraba a nivel de manzana, se hizo uso del programa Mapinfo en el cual se tenia la cartografía de la ciudad de Santiago de Cali, esta cartografía contenía información a nivel de manzana, barrio y comuna.

Con esta información se procede a realizar un cruce por el id de manzana entre el archivo de datos de la composición socioeconómica y la información del programa Mapinfo, después de ese cruce se le agrega la variable barrio contenida en Mapinfo al archivo de datos de la composición socioeconómica de los barrios, y así finalmente este archivo se empalmó con la información del archivo de datos suministrado por la Secretaria Municipal de Salud (archivo de casos confirmados) que también tenia una variable del barrio del caso registrado y finalmente obtener un solo archivo con el cual se realiza este trabajo.

A continuación se presentan algunas características de manera descriptiva de la población de la ciudad de Santiago de Cali.En la Figura **5-7** se observa que el porcentaje de personas con edades menores a los 30 años es mayor en la mayoría de los barrios de la ciudad, alcanzando una proporción del 35 % en algún barrio de la ciudad. Por otra parte, los porcentajes de personas que se encuentran dentro de los rangos de edad mayores a 40 años en los barrios

de la ciudad es menor, esto pueden indicar que en general en los barrios de la ciudad la población joven predomina sobre la adulta.

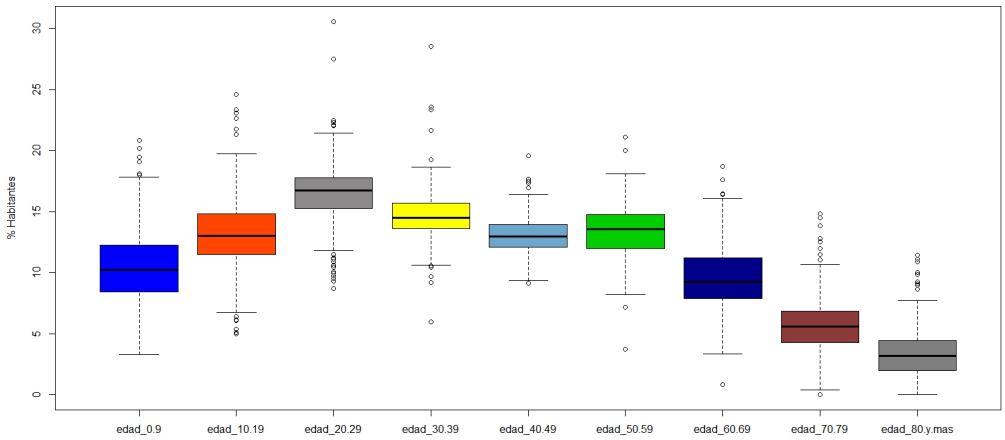

Figura 5-7: Boxplot del porcentaje de personas por rangos de edad en los barrios.

Complementando este análisis, es importante mencionar que hay barrios donde se presenta una gran población entre los 20 y 39 años de edad. Adicionalmente, la edad mediana de los casos contagiados en la ciudad de Cali es superior en edades inferiores a los 59 años y posteriormente desciende hasta una participación del 3 % en edades de 80 y más años en los barrios de la ciudad.

Indicador	Mínimo	Máximo	Media	Mediana	Varianza	Desviación
Servicios Basicos	12	100	97,69	99,53	0,63	7,96
Servicios Complementario	27,39	100	70,88	74,41	3,57	18,90

Tabla 5-1: Indicador del porcentaje de cobertura de servicios públicos (Definidos en la Sección 4)

En la Tabla **5-1** se presentan dos indicadores construidos para medir el porcentaje de servicios públicos domiciliarios; estos están divididos en dos grupos. El primer grupo corresponde a los servicios básicos que están compuestos por agua, energía eléctrica, acueducto, alcantarillado y recolección de basuras, y el segundo son los servicios complementarios que están compuestos por la conexión a internet y el gas domiciliario.

Según, la Tabla **5-1** se evidencia que gran parte de los barrios de la ciudad de Cali cuentan con cobertura de servicios básicos domiciliarios alta, superando en promedio el 90 % de barrios con cobertura total de servicios básicos.Por otra parte, para el caso del indicador de servicios complementarios se puede observar que presenta una disminución en el acceso a

dichos servicios, además se presenta una alta variabilidad entre barrios, lo que indica que la accesibilidad a esos servicios es diferente entre los barrios.

Es decir que, en la ciudad hay un sector que presenta dificultades en cuanto a cobertura de servicios como la conexión a internet, lo cual puede haber dificultado que se tenga una mayor información sobre la prevención del virus y como consecuencia, en dichas zonas se presento una mayor cantidad de casos confirmados.

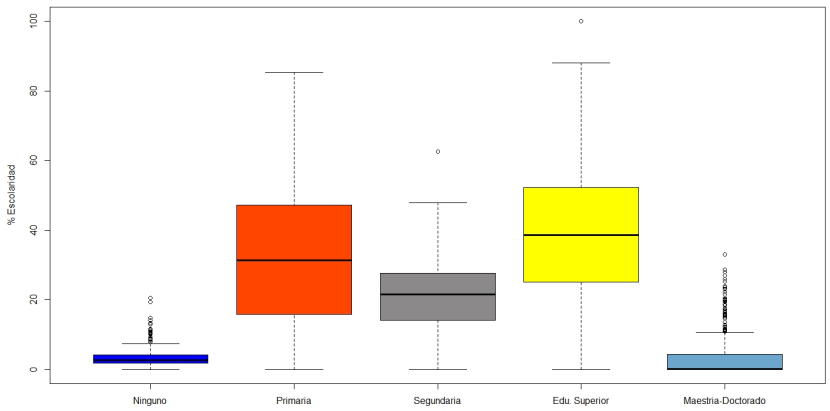

Figura 5-8: Distribución del porcentaje de escolaridad en los barrios de la Ciudad de Santiago de Cali.

La Figura **5-8** presenta la distribución del grado de escolaridad en los barrios de la ciudad de Cali, donde se puede identificar que hay barrios con un alto porcentaje de personas que cuentan en educación superior. Sin embargo, también se presenta gran variabilidad entre barrios que alcanzan la educación superior. Algo que se observa en la Figura **5-8** es que hay varios barrios donde la participación de personas con postgrados es superior, adicionalmente el 50 % de los barrios que alcanzan postgrado es aproximadamente el 2 % de los habitantes del barrio.

Es que la variabilidad en el porcentaje de personas que presentan postgrado es baja, al igual que en el porcentaje de personas que no presentan algún grado de escolaridad; es decir que en este caso el comportamiento es similar entre barrios.

Tipo de uso de vivienda	Mínimo	Máximo	Media	Mediana	Varianza	Desviación
Uso residencial	18	100	81	85,31	1,89	13,75
Uso comercial	0	22	3	2,06	0,12	3,42

Tabla 5-2: Resumen descriptivo del porcentaje de uso de vivienda en los barrios de la Ciudad de Cali.

En la Tabla **5-2** se presenta el porcentaje de viviendas destinadas al uso residencial y al uso comercial. Se puede evidenciar que en general la mayoría de los barrios de la ciudad de Cali usan las viviendas de manera residencial, mientras que en cambio en promedio el 3 % de las viviendas en los barrios de Cali son de uso comercial.

5.2. Estimación del Número Efectivo de Reproducción (R_t)

A continuación se presenta la información asociada a los casos positivos del virus reportados en el periodo de estudio (Marzo 2021 - Mayo 2022) en la ciudad de Santiago de Cali. Con esta información se obtuvo la estimación diaria del Número Efectivo de Reproducción (R_t).Para obtener dichas estimaciones se utilizó uso de la herramienta R Core Team (2019) y en particular del paquete *Epiestim* desarrollado por (Cori et al., 2013), que permite estimar estos indicadores de la enfermedad.

5.2.1. Distribución del Intervalo Serial

Son diversos los casos a nivel internacional donde se utilizan diferentes valores para la media y la desviación de la distribución del intervalo de serie. Para el presente trabajo se asumen los valores sugeridos por **Estrada-Alvarez et al. (2020)** en el estudio denominado *Estimación del intervalo serial y número reproductivo básico para los casos importados de Covid-19*, el cual se reporta en los antecedentes de este trabajo; donde se estima un intervalo serial con un valor promedio de 3.8 días y una desviación estándar de 2.7 días **Estrada-Alvarez et al. (2020)**.

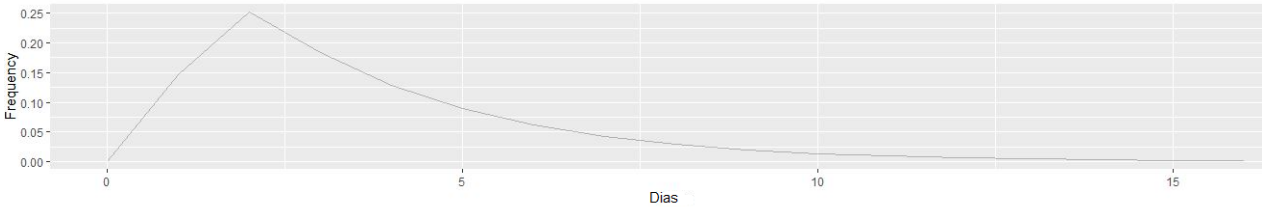

Figura 5-9: Distribución de frecuencias del intervalo serial de los casos de Covid-19 en la ciudad de Santiago de Cali para el periodo Marzo 2021 - Mayo 2022

En la Figura **5-9** se observa que la mayor probabilidad de propagación o contagio entre una persona que tiene el virus (caso principal) y otra persona (caso secundario) que fue contagiada por ese caso primario se da antes de los primeros 5 días.

Todo esto puede indicar que si se toman medidas de prevención por parte de los organismos de control y se identifican de manera rápida los casos ya sean sintomáticos o asintomáticos

en los primeros días y se establecen medidas como el aislamiento de la persona contagiada, se puede disminuir la probabilidad de contagio en la población general, ya que se establece un cerco para aquellos portadores del virus; evitando el contacto con posibles receptores y así mermar la propagación en la ciudad.

Con los resultados mencionados anteriormente, se realiza la estimación del Número Efectivo de Reproducción, los cuales se muestran a continuación:

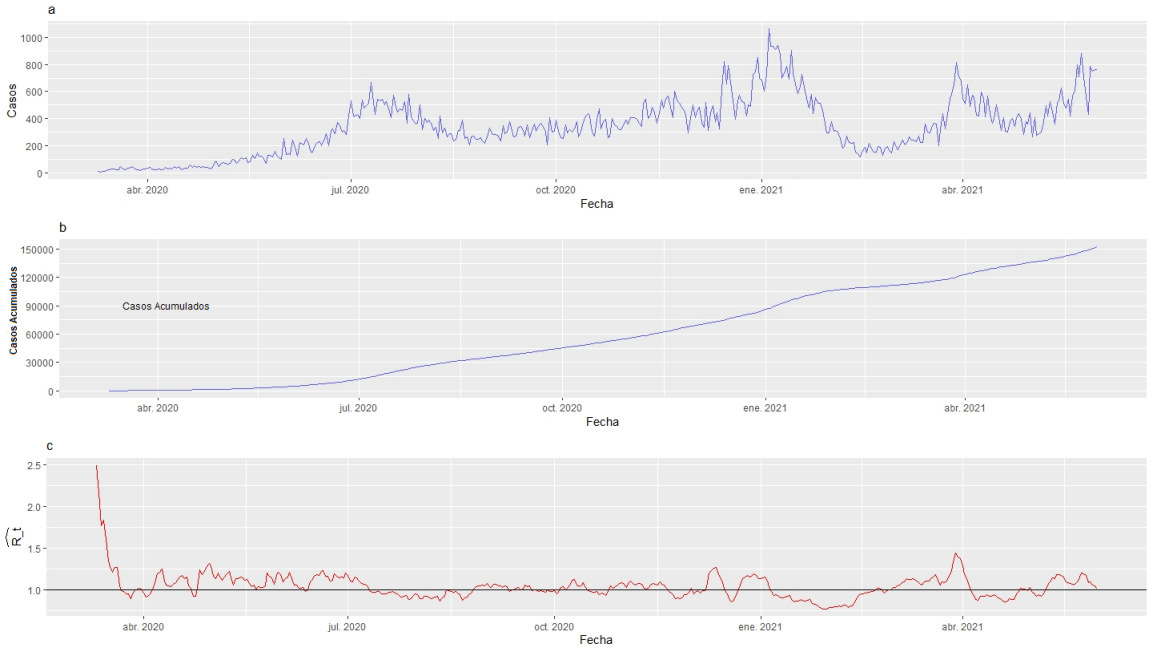

Figura 5-10: (a) Casos confirmados diarios en el tiempo de estudio en la ciudad de Santiago de Cali. (b) Casos acumulados en el tiempo de estudio en la ciudad de Santiago de Cali. (c) Estimación del Número Efectivo de Reproducción para la ciudad de Santiago de Cali. El periodo de estudio contenido en cada gráfica es de Marzo 2020 - Mayo 2021

La Figura **5-10** condensa la información de los casos diarios (incidencia) de las personas contagiadas por el virus en la ciudad de Santiago de Cali (a), así como los casos diarios acumulados (b) y la estimación del Número Efectivo de Reproducción $(\widehat{R_t})(c)$. Se evidencia que el número de casos confirmados al inicio de la pandemia son bajos, con valores inferiores a los 200 casos diarios, particularmente en los meses de Marzo, Abril y Mayo del 2020. Es decir, en estos meses el virus se propagó de manera lenta en la ciudad. Al llegar el mes de Junio y Julio se evidencia un incremento en la cantidad de casos confirmados del virus, esto pudo ser debido a la flexibilidad en las medidas de contención dadas por el Gobierno Nacional en conjunto la Alcaldía de Santiago de Cali, donde aún se mantenía la emergencia y el aislamiento preventivo obligatorio, pero se implementó paulatinamente la salida de menores

de 17 años a espacios abiertos, la realización de actividad física y medidas para retomar el comercio en la ciudad; todo lo anterior conllevó a generar el primer pico de la pandemia.

Posteriormente, se evidencia que los casos disminuyen un poco y se mantiene constante hasta finales del año 2020. A inicio del año 2021 se presenta el inicio del segundo pico de contagios, aquí se pudo evidenciar las consecuencias de las celebraciones de fin de año que son comunes en nuestra ciudad, adicional a eso también se presentó la final del fútbol colombiano generando aglomeraciones.

En los meses de Febrero y Marzo del año 2021 se nota una disminución importante de contagios que se puede atribuir a las medidas de mitigación tomadas por la Alcaldía de la Ciudad de toques de queda los fines de semana, reduciendo la facilidad de reuniones familiares nocturnas. Finalmente, la gráfica (a) muestra el inicio del último pico de casos identificados en este trabajo, que fue en el mes de Abril del 2021, atribuido posiblemente a las actividades semana santa, donde muchos ciudadanos viajan a otros lugares en un periodo de receso vacional.

Algo a resaltar es que esta enfermedad, al ser un virus que se trasmite a través del contacto con una persona infectada, su comportamiento es creciente en el tiempo; al menos hasta que se tomen diferentes medidas de mitigación por parte de los organismos estatales y finalmente será decreciente cuando se logre establecer una alternativa como la vacunación masiva.

Finalmente, la Figura **5-10** (c), evidencia el comportamiento de la propagación del Covid-19; inicialmente se tiene una estimación alta para el mes de marzo alcanzando valores superiores a 5, lo cual indica que a inicio de la llegada del virus a la ciudad por cada individuo con el virus contagiaba a 5 personas.

Por lo tanto, como se evidencia en la Figura **5-5**, este virus inicialmente se contagiaba en zonas donde se puede tener contacto con personas del extranjero, que son personas con mayor capacidad adquisitivo (sur y parte del oeste de la ciudad) y por consecuente las personas con las que reside, entre estas se tienen personas de oficios varios que por lo general provienen de sectores de la ciudad con estratos más bajos (oriente de la ciudad).

También hay que mencionar que en los meses iniciales se tomaron diferentes medidas de control por el gobierno local para mitigar la propagación. A partir de las primeras medidas restrictivas (toque de queda, cierre de establecimientos nocturnos, cancelación de eventos masivos, entre otras) tomadas por los entes de control del municipio se observa una disminución de este indicador para el mes de Junio, es decir que pasó a tomar valores entre 0.7 y 1.5, lo cual indica una disminución de la propagación del virus. También hay que decir que al igual que en el gráfico de la incidencia de casos diarios, este muestra una alza de la estimación en los picos de Diciembre (2020), Enero(2021) y Abril (2021) como se había mencionado anteriormente.

5.3. Análisis exploratorio espacial de la velocidad de propagación

En la Figura **5-11** se muestra una comparación de la estimación del Número Efectivo de Reproducción para las 22 comunas de la ciudad de Cali (azul claro) frente a la estimación general de la ciudad (color azul oscuro). En este caso se evidencia un comportamiento similar entre algunas de las comunas y la estimación general.

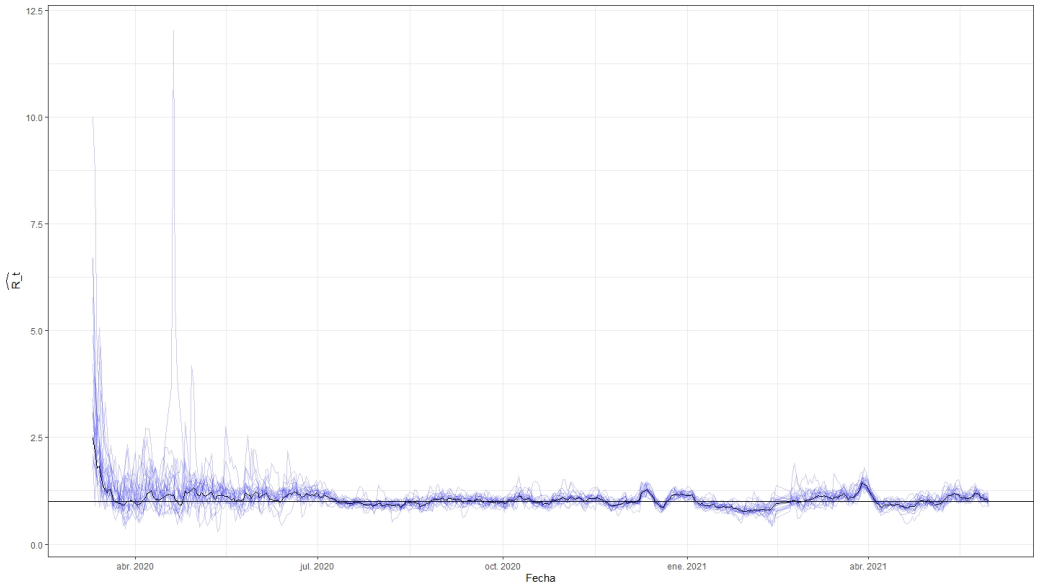

Figura 5-11: Comparación en la estimación del R_t por comunas (líneas azul claro) y la estimación para toda la ciudad (línea azul oscuro)

Observando el comportamiento de las estimaciones del R_t, tanto para comunas como en los barrios, se puede identificar que al inicio de la pandemia, es decir en el mes Marzo, Abril, Mayo, Junio y Julio del 2020 se presentó un pico prolongado de contagios que llega a valores superiores a 2, este comportamiento se da gracias a que en algunos barrios la propagación del virus, a pesar de ser baja, presentó instantes de aceleración, por lo tanto la estimación de forma general para la ciudad no es estable; en conclusión se puede decir que en estos meses se da una primera fase de propagación en la ciudad.

Posterior a las medidas de control tomadas por la administración local (toque de queda, pico y cédula, cancelación de eventos masivos, etc), en el mes de Agosto se ve un descenso drástico del R_t hasta llegar a valores menores a 1 a mediados del mes de Septiembre, por tanto se puede decir que se tiene una segunda fase de mitigación de la propagación del virus en la ciudad y son consecuencias de las acciones tomadas a nivel local.

Continuando con el análisis, en los meses de Octubre, Noviembre y Diciembre del 2020 se observa nuevamente un leve incremento en el Número Efectivo de Reproducción con valores

que oscilan entre 1 y 1.5 a nivel de ciudad y a nivel de comunas, para luego establecer un segundo pico en el mes de Enero del 2022, pico generado posiblemente por los diferentes acontecimientos que dieron lugar en la ciudad como fueron la final del Fútbol Colombiano, la Feria de Cali y las celebraciones de fin de año; todo esto desencadenó un incrementó de casos y una mayor propagación del virus al inicio del año 2022.

Finalmente, la Figura **5-11** muestra que, posterior al segundo pico se presenta una disminución en la propagación en los meses de Febrero y Marzo, dado que se toman nuevamente medidas restrictivas de movilidad. Para finalizar, el periodo de estudio se da un incremento en el mes de Abril como consecuencia de las interacciones de la Semana Santa y descenso en los meses posteriores.

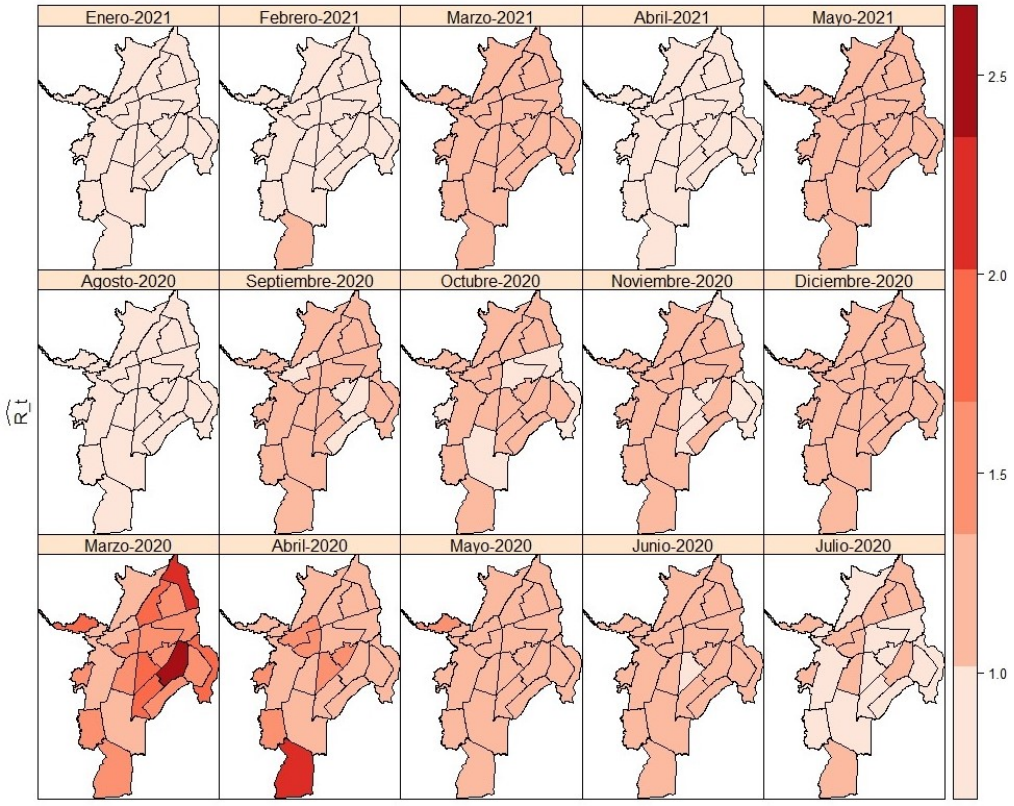

Figura 5-12: Comportamiento del Número Efectivo de Reproducción (R_t) promedio mes a lo largo del periodo de estudio Marzo 2020 - Mayo 2021 en la Ciudad de Santiago de Cali.

La Figura **5-12** representa el comportamiento espacial de las estimaciones del Número Efectivo de Reproducción a nivel de comunas por mes durante el periodo de estudio (Marzo del 2021 hasta Mayo 2022), en cada una de las 22 comunas de la ciudad de Cali. El indicador

graficado corresponde al valor medio de la serie R_t estimada por mes en cada una de las comunas de Santiago de Cali.

En términos generales se observa un comportamiento similar en las comunas; en el primer mes la mayoría de comunas presentan un Número Efectivo de Reproducción que toma valores entre 1.5 y 2.5, siendo las Comunas 3, 13 y 16 (color rojo oscuro) las que presentan mayor foco de propagación del virus. Posteriormente se tiene una disminución en los meses de Abril, Mayo y Junio, que son meses donde se estaban estableciendo algunas medidas de control como lo son el pico y cédula, toque de queda y cancelación de eventos masivos; a partir de que se toman estas medidas el resultado se observa en los meses de Agosto, Septiembre, Octubre y Noviembre con un descenso de la propagación con valores entre 0.5 y 1.5, lo cual indicaría que las medidas fueron efectivas y se entra en un periodo de mitigación.

Finalmente, en Diciembre del 2020 se da un incremento paulatino a nivel de ciudad lo cual desencadena nuevamente una serie de medidas menos drásticas que al inicio de la pandemia, pero que de igual forma contribuyen a que en el mes de Febrero del año 2022 se disminuya la propagación del virus.

Sin embargo, pese a estas medidas nuevamente se da un incremento en el mes de Marzo y Mayo del 2021, que está cobijado por las diferentes actividades realizadas como la Semana Santa, que es un periodo donde se espera que las personas se aglomeren o salgan de la rutina diaria y no se tomen cuidados contra la enfermedad y por esto puede ser ese leve incremento en los tres últimos paneles de la Figura **5-12**.

5.4. Modelación Estadística

En la Sección 5.2, se realizó el análisis de la propagación del Covid-19 a nivel de comunas en la ciudad de Santiago de Cali con el fin de identificar generalidades y/o comportamiento del virus en la ciudad de Santiago de Cali. Para la modelación estadística las unidades de estudio fueron los barrios, con el fin de identificar características más específicas, ya que una comuna está compuesta de diferentes barrios y en algunos casos esos barrios no son homogéneos entre ellos (habitantes, estrato, extensión territorial entre otros). Adicionalmente, al trabajar a nivel de comunas solo se cuenta con 21 observaciones para la construcción del modelo. Hay que aclarar que el objetivo para este caso en particular más que predecir, es describir la relación que puede existir entre las características socioeconomicas de los barrios y la propagación del virus.

La parte de la Modelacion Estadística se desarrolla un Análisis Factorial Múltiple (AFM) previo a la modelación con el fin identificar una primera aproximación de relación entre grupos de variables (los grupos se describen a continuación) desde un enfoque multivariado y posteriormente se implementa un Modelo de Mínimos Cuadrados Parciales 2 (PLS2) que

utiliza las características socioeconómicas del barrios como predictor de la velocidad de propagación del virus en los barrios.

5.4.1. Análisis Factorial Múltiple

Como ya se menciono en la Sección 5.4 se realizó un Análisis Factorial Multiple (AFM) como un analisis previo a la modelación para identificar posibles relaciones entre grupos de variables entre los mismos grupos o con las variables repuesta (en este caso los tres instantes de tiempo ya mencionados) de forma previa.

Inicialmente se describen los cinco grupos de variables construidos para este análisis; cabe resaltar que la información de cada uno de estos grupos está consolidado a nivel de barrio.

- R_t : Número Efectivo de Reproducción mediano para los tres instantes de tiempo.
- **Densidad:**Variables asociadas a la composición poblacional de los barrios como lo son el área, la cantidad de habitantes, viviendas, etc
- **Vivienda:** Variables que proporcionan información del uso que se le da a la vivienda como es el uso comercial o residencial.
- **Edad:** Son valores asociados a la distribución de la población por rangos de edad, incluyendo el índice de juventud.
- **Educación:** Indicadores del nivel de educación alcanzada por la población (educación básica y educación superior)

La Figura **5-14** evidencia que existe una alta relación entre el Número Efectivo de Reproducción R_t mediano y las variables asociadas a la composición poblacional de los barrios de la ciudad de Santiago de Cali; así mismo estos dos grupos de variables presentan una mayor contribución en el segundo eje factorial. Adicionalmente se identifica que se presenta mucha relación entre los diferentes grupos de edad y los niveles de educación alcanzada, las cuales se encuentran en la parte inferior derecha de la gráfica, estos dos grupos de variables presentan mayor relación en la primara dimensión factorial.

Finalmente, se evidencia que el porcentaje de varianza explicada por al formar estos grupos es mayor al 50 %; esto puede ser debido a que entre mayor edad posiblemente puedan alcanzar estudios superiores y en edades inferiores solo alcanzan niveles básicos de escolaridad. También el grupo asociado al tipo de viviendas del barrio presenta algo de asociación a los grupos de edad y educación. Finalmente estos tres grupos de variables (vivienda, edad y educación) no son relacionados de forma multivariada con el número efectivo de reproducción.

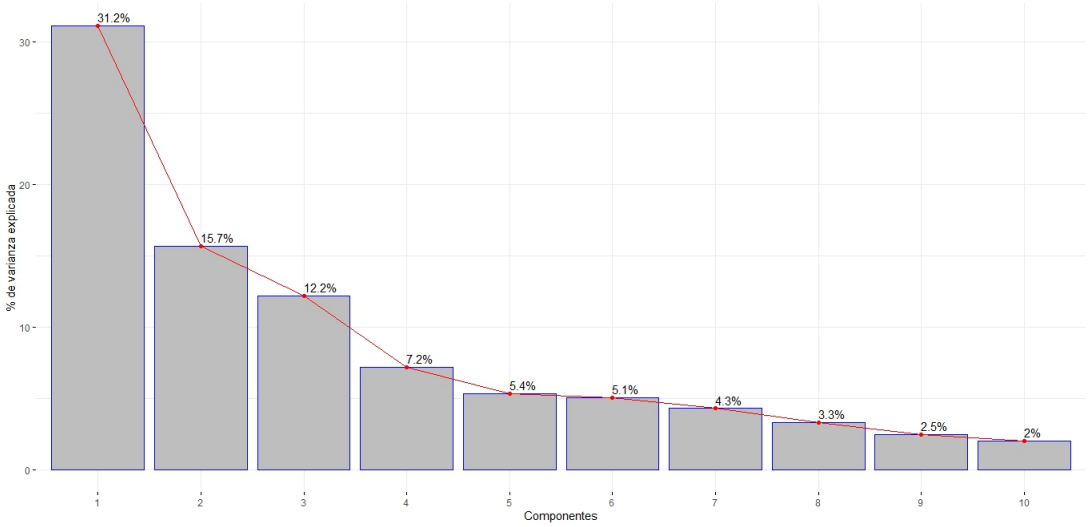

Figura 5-13: Porcentaje de varianza explicado de cada componente principal

En la Figura **5-13** se presenta el porcentaje de varianza explicada por cada una de las componentes principales. Se considera trabajar con las primeras dos componentes que representa el 47 % de la variabilidad de los datos y hace mucho más fácil su interpretación.

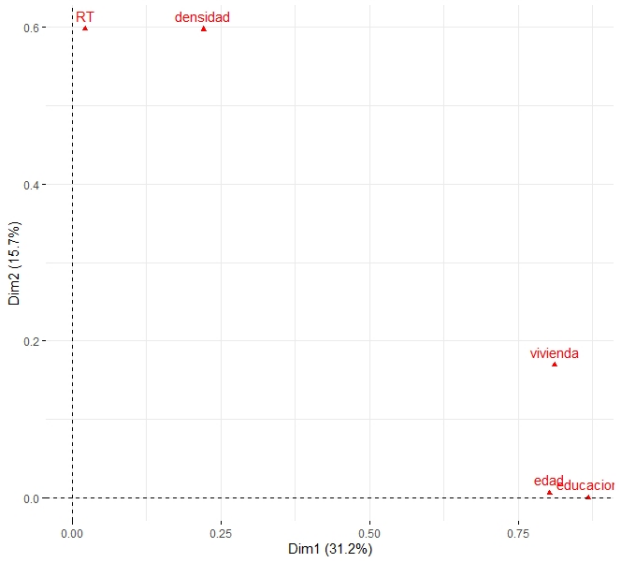

Figura 5-14: Análisis Factorial Múltiple por grupos de variables

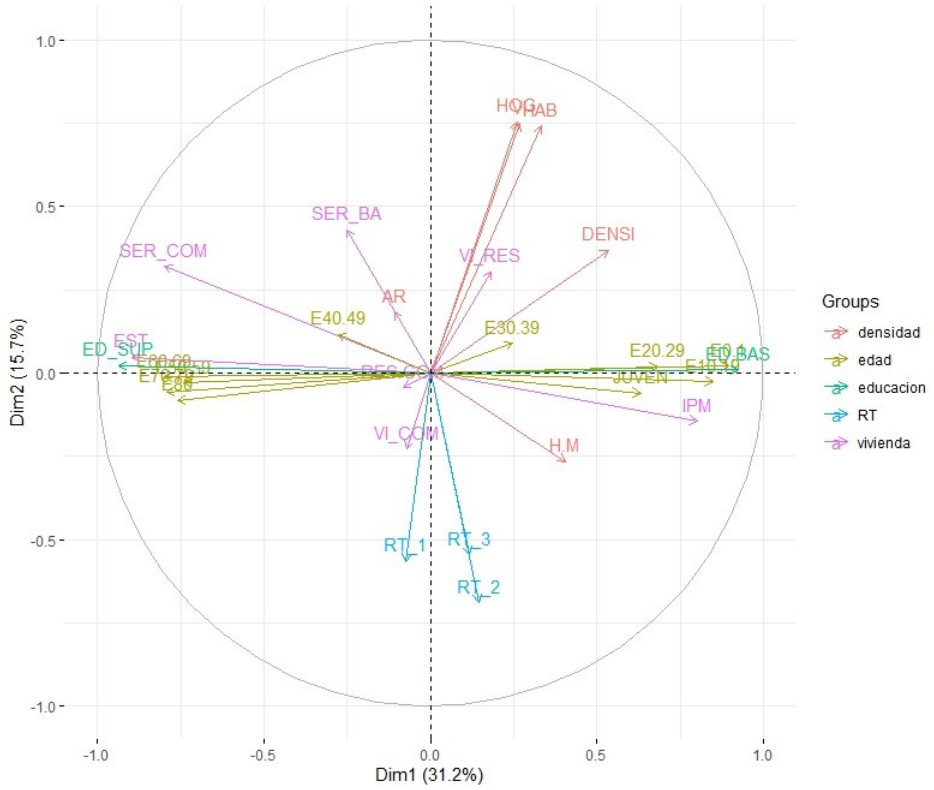

Figura 5-15: Respresentación de las dos primeras componentes principales de cada grupo por su correlación con las dos primeras componentes principales del AFM con las variables del estudio

Con el fin de identificar qué tipo de relaciones existen entre las variables con las cuales fueron construidos los grupos del AFM se presenta la Figura **5-15** con la proyección de las variables en el plano. Inicialmente se evidencia que la relación entre el Número Efectivo de Reproducción y variables asociadas a la densidad poblacional es de forma negativa, es decir que a mayor cantidad de habitantes menor es el R_t mediano en el barrio.

En mismo sentido, la asociación presentada en la Figura **5-14** de los grupos de variables edad y educación se corrobora en la Figura **5-15** donde se evidencia una estrecha relación entre edades inferiores a los 29 años y el indicador de educación básica, adicional a estas variables también se encuentra una relación con el índice de pobreza monetaria y tiene mucho sentido ya que este indicador suele ser muy influenciado por barrios (en este caso) donde predominan niños.

Paralelo a estas variables anteriormente mencionadas pero en sentido contrario se encuentran las edades superiores a 50 años asociados al indicador de educación superior y también la variable estrato que indica que en estratos superiores predominan barrios con alta población adulta con estudios superiores.

Por último, las variables asociadas a la composición de la vivienda están dispersas a lo largo de los clúster mencionados anteriormente; las viviendas de uso comercial está relacionada positivamente con el R_t, es decir que en barrios donde predomine este tipo de viviendas el R_t será más alto. Por otro lado, en aquellos barrios donde predominen las viviendas de uso netamente residencial el R_t será menor.

Descripción de las variables

A continuación se presentan las variables que se van a utilizar en la construcción del modelo.

Variables de respuesta

- R_t Mediano 1: Corresponde a la estimación de la mediana del Número Efectivo de Reproducción en el periodo comprendido entre el de Junio al 14 de Julio del 2020.
- R_t Mediano 2 : Corresponde a la estimación de la mediana del Número Efectivo de Reproducción en el periodo Enero de 2021.
- R_t Mediano 3: Corresponde a la estimación de la mediana del Número Efectivo de Reproducción en el periodo Abril 2021.

Estas tres variables serán utilizadas como respuesta en el modelo.

Variables explicativas

Esta variables representan las condiciones socio económicas de los barrios de la ciudad.

- **Indicador porcentaje de educación básica:** Porcentaje de personas con educación primaria y secundaria
- **Indicador porcentaje de educación superior:** Porcentaje de personas con educación técnica, tecnológica, universitaria y postgrados.
- **Indicador servicios básicos:** Porcentaje de viviendas que cuentan con los servicios de Energía, agua, alcantarillado y recolección de basuras.
- **Indicador servicios complementarios:** Porcentaje de viviendas que cuentan con los servicios de Internet y Gas.
- **Indicador de viviendas uso residencial:** Porcentaje de viviendas las cuales su uso es netamente residencial.
- **Indicador de viviendas uso comercial:** Porcentaje de viviendas las cuales su uso es netamente comercial.

- **Indicador uso residencial / uso comercial:** Cociente entre el porcentaje de las viviendas de uso residencial y comercial.
- **Razón de masculinidad:** Cociente entre la cantidad de hombres sobre mujeres.
- **Índice de juventud:** Cociente entre la cantidad de personas menores de 19 años sobre las personas mayores de 60 años.
- **IMP:** Índice de pobreza multidimensional promedio.
- **Estrato moda :** Representa el estrato modal.
- **Densidad poblacional:** Densidad de población por barrio, calculado como el total de habitantes por 1000 sobre el área del barrio.
- **Habitantes:**Cantidad de habitantes.
- **Cont_hogares:** Cantidad de hogares.
- **Cont_viv:** Cantidad de viviendas.
- **Grupos de edades** Corresponden a 9 variables que cuentan con la proporción de habitantes de rangos de edad que aumentan de a 10 años. Es decir, el primer grupo es de 0 años hasta los 9 años y así incrementan hasta los 79 años de edad y la ultima variable recoge lo que tienen edades superiores a 80 años.

Estas 25 variables son las que conforman las variables predictoras para el modelo.

Para la Modelación del Número Efectivo de Reproducción se presenta un Modelo de Mínimos Cuadrados Parciales 2 (PLS2), debido a las características multivariantes de las variables respuesta y dificultades presentadas por los datos, como lo son la multicolinealidad, la correlación entre las diferentes variables predictoras, entre otras.

Este tipo de modelos, suele ser mas flexibles en la validación de supuestos que no brindan las metodologías tradicionales, toman esas relaciones existentes entre variables predictoras y aprovechan eso para explicar el comportamiento de las variables dependientes.

5.4.2. Resultados del Modelo PLS2

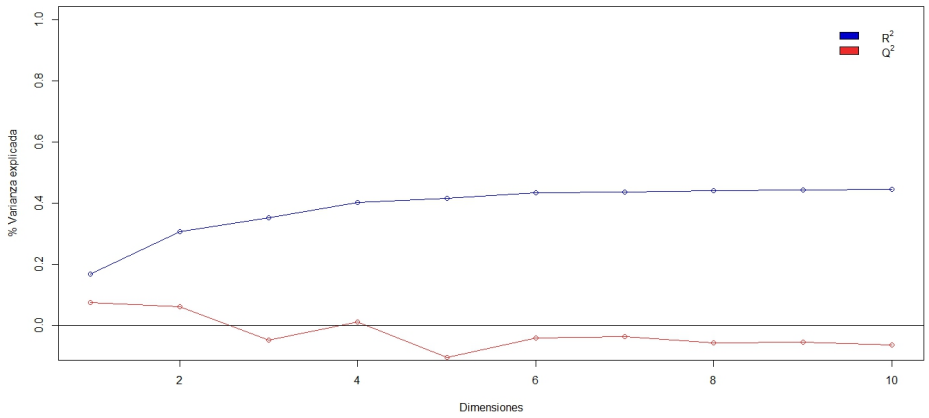

Figura 5-16: Comparación entre el R^2 y el Q^2 con cada una de las componentes

En la Figura **5-16** se presentan dos indicadores para la elección de componentes; en la parte superior se encuentra el Q^2 que corresponde a la variación total que puede predecir mediante cada uno de los componentes basado en validación cruzada para comprobar la bondad del modelo, en este se puede observar que con las primeras dos componentes alcanza un Q^2 de 0.06 y posterior a la segunda componente este indicador comienza a descender.

Por otro lado, la Figura **5-16** también permite evidenciar el R_y^2 que es considerado como un R^2 tradicional pero este hace referencia al nivel explicativo de cada componente a emplear; es decir la varianza explicada pero en la proyección en el plano factorial. En este caso se evidencia que las primeras 4 componentes son donde mayor nivel explicativo presenta y posteriormente su aumento no es significativo.

De acuerdo a lo mencionado anteriormente, se toma la desición de trabajar con las dos primeras componentes principales, aunque el R_y^2 sugiere que se puede trabajar hasta con 4 componentes el Q^2 se vuelve negativo lo que indica que el modelo se esta sobreajustando. Por lo dicho anteriormente y por el principio de parsimonia que sugiere que si existen dos modelos (en este caso se puede ver con componentes) es preferible elegir aquel que tenga mayor facilidad de interpretación.

La Figura **5-17** evidencia el comportamiento entre variables en el plano factorial, se tiene que en la parte superior derecha del plano se encuentran las variables relacionadas con los tres momentos o picos del Número Efectivo de Reproducción (R_t), de las cuales el R_{T2} y R_{T3} están más relacionados entre sí, este comportamiento puede estar ligado a que la propagación ya estaba más avanzada y había más presencia de casos en mayor cantidad de barrios en comparación al primer pico. Mientras que en el momento del R_{T1} los casos eran

más dispersos entre los barrios y por lo tanto las estimaciones de este indicador era más variable. Adicionalmente, también se tiene el cociente entre viviendas de uso residencial y viviendas de uso comercial por barrio, donde en aquellos barrios donde predominaran las viviendas de uso residencial probablemente presentaba mayor propagación del virus.

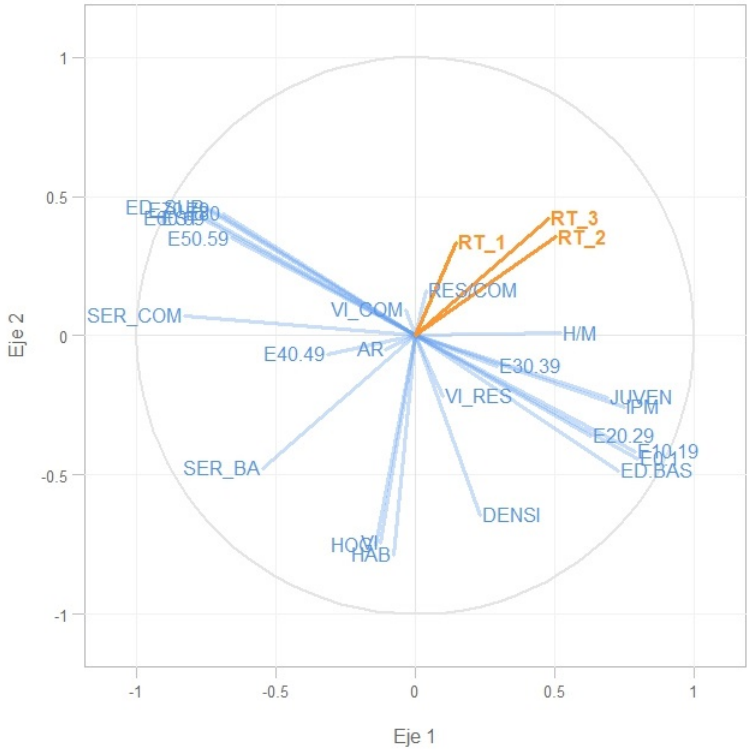

Figura 5-17: Representación de las variables utilizadas en la Regresión PLS2 en el plano factorial

Continuando con el análisis, se tiene una división entre los rangos de edad, en este caso los rangos de edad de la población joven y adultos hasta los 40 años se encuentran muy relacionados con el índice de juventud, el índice de pobreza monetario y el indicador de educación básico; esto cobra mucho sentido ya que al ser población joven (edades inferiores a los 30 años) puede que no hayan terminados sus estudios universitarios o no pudieran acceder a ellos ya que esa relación va acompañada del índice de pobreza monetaria que regularmente suele ser alta en presencia de muchos niños y/o hacinamiento.

Paralelamente a este grupo de barrios pero en dirección contraria se encuentran los rangos de edad de las personas adultas (mayor a 40 años) se encuentra relacionado con el indicador de servicios complementarios y educación superior, estos son barrios con un índice de pobreza

monetario bajo por ende pueden alcanzar estudios superiores al bachiller y tener acceso a servicios complementarios.

Adicionalmente, las variables anteriores se encuentran relacionadas con el estrato que indica que en barrios donde el estrato es superior predominan personas con edades superiores y con estudios universitarios (también se podría atrever a afirmar que esta población no suele tener muchos hijos). Este comportamiento mencionado ayuda a darse cuenta que existen barrios dentro de la ciudad donde la población joven predomina sobre la población adulta y viceversa, es decir que en cuanto a esta característica los barrios también presentan diferencias en su población.

Finamente, se identifica que el indicador de servicios básicos (energía eléctrica, alcantarillado, agua potable, recolección de basuras) se encuentra en dirección contraria la propagación del virus, esto hace pensar que la propagación debería afectar en mayor medida a los barrios donde se cuenta con menos acceso a dichos servicios básicos, ya que al ser el Covid-19 una enfermedad contagiosa a través del contacto, factores como la falta del agua juegan un papel fundamental para prevenir el contagio. Adicionalmente al rededor del indicador de servicios básicos también se encuentran asociadas variables poblacionales como lo son los habitantes, viviendas y hogares que es curioso ya que se cree que el virus se propaga más en zonas de mayor población.

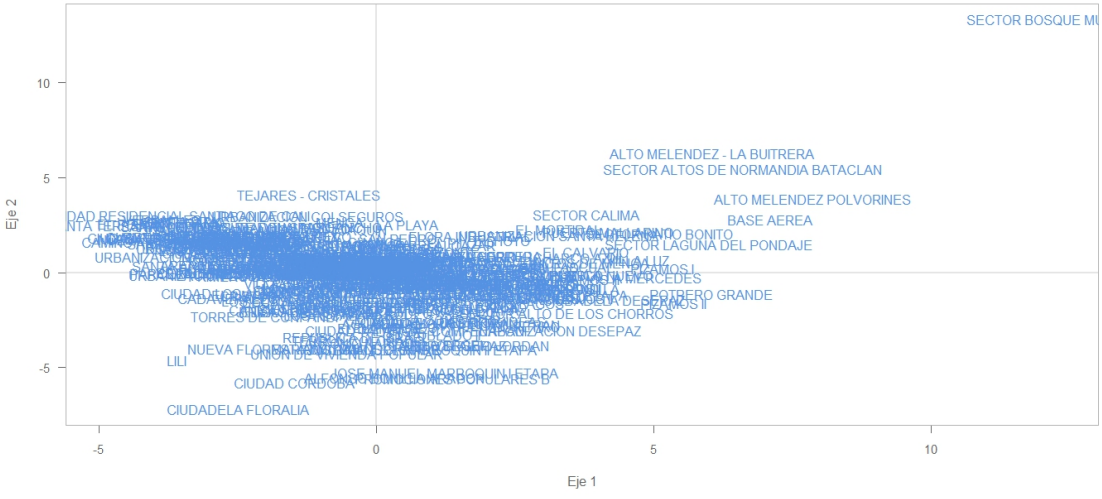

Figura 5-18: Representación de los barrios de la Ciudad de Santiago de Cali en el plano factorial del modelo PLS2

La Figura **5-18** presenta la representación de los barrios de la ciudad en el plano factorial, en este caso se observa que barrios como Bosque Municipal, Alto Melendez, Alto Normandia, Polvorines y la Base Aérea tienen características similares, así mismo estos barrios se encuentran en dirección de las variables relacionadas con la propagación del virus de los instantes 2 y 3 como se observa en la Figura **5-17**, mientras que los barrios como Calima, el Morichal, el Pondaje están más relacionados con el periodo de propagación 1, complementando esta información en estos barrios los indicadores de servicios básicos como el agua y alcantarillado son menores, ya que algunos de los barrios mencionados están en sectores de ladera.

Barrios como Potrero Grande, Decepaz, Pizamos II los chorros, entre otros están compuestos con caracteristicas similares entre los barrios y con una propagación del virus inferior, además que tienen como denominador que su población está compuesta en mayoría por personas menores de 40 años; algunos de estos barrios son barrios de reubicación otorgados por la Alcaldía de la ciudad de Cali para aquellas personas que vivían en las laderas, invasiones o estaban en condiciones de peligro por lo cual son lugares donde la presencia de niños es alto.

Mientras que barrios Colseguros, Santa Anita y otras urbanizaciones presentan una propagación mucho menor y tiene como denominador que cuentan con una población adulta mayor, con unos servicios básicos altos y que bajo estas condiciones pueden tener medidas de prevención mucho mejores a los demás barrios ya mencionados.

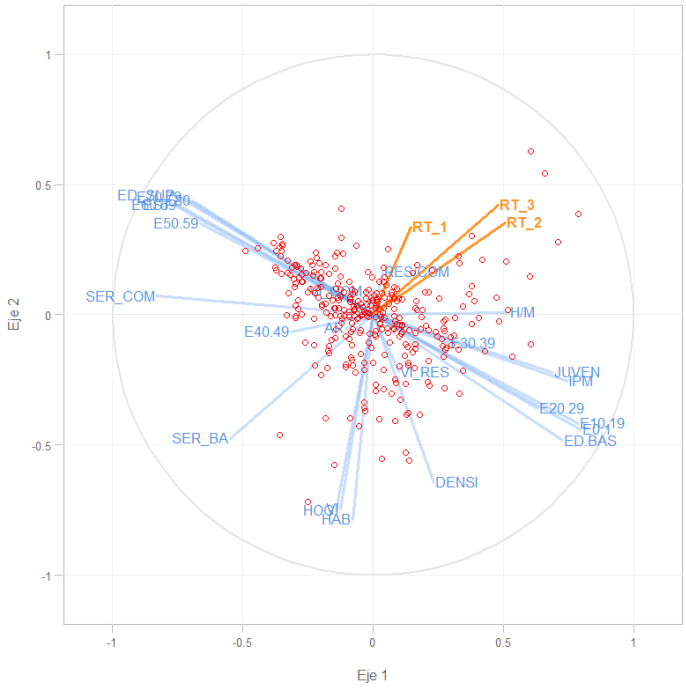

Figura 5-19: Representación simultánea de las variables y los barrios de la ciudad de Santiago de Cali en el plano factorial del modelo PLS2

En la Figura **5-19** se muestra la representación superpuesta de las variables y los barrios en el plano factorial, el este caso el Número Efectivo de Reproducción (R_t) están correlacionados por evidente razón ya que es el mismo indicador pero en 3 momentos de tiempo. En la parte superior izquierda encuentran los barrios de edades superiores a los 50 años y el indicador de Educación Superior; es decir en edades más altas las personas alcanzan unos altos niveles de educación y estos se ubican en barrios como: Santa Teresita, El Peñon, Versalles, Nueva Tequendama y Santa Monica; los cuales son barrios en los que se presentan mejores condiciones de vida.

Por otro lado, en la parte inferior (en sentido contrario del primer momento establecido) se observa relación entre las variables hogares, habitantes y viviendas; esto puede dar un indicio de que inicialmente los casos no se propaga en los barrios mas poblados como se suele pensar, si no que la propagación del virus suele ser mas alta en aquellos barrios donde su población es baja. Es decir que en los barrios donde se pudo propagar fueron los que contaban con personas de ingresos medios altos, los cuales presentan menor densidad poblacional y además pudieron ser más propensos a tener contacto con personas provenientes del extranjero y propagaron el virus paulatinamente en los barrios con menor densidad de población.

Todo lo mencionado anteriormente puede ser influenciado por las relaciones laborales, personales y sociales que contribuyeron a que los barrios mas poblados sigan esta dinámica; como por ejemplo: Ciudadela Floralia y Ciudad Cordoba, que son barrios donde se sabe por su composición de viviendas (viviendas sin calles principales en cada frente de las viviendas) presentan una mayor población por metro cuadrado.

Seguido, se evidencia una relación inversa entre los momentos 2 y 3 del R_t mediano con el Indicador de servicios básicos que se complementa con lo mencionado anteriormente de que inicialmente se propagaba en zonas con mejores ingresos y posteriormente al paso del tiempo se propaga a otros sectores de la ciudad; en este caso se encuentra el barrio Valle del Lili que cuenta con buen indicador de servicios básicos y una baja propagación del virus.

Por último, en la parte central (cerca al origen) se encuentran las variables de edades entre 30 y 49 años, porcentaje de vivienda residencial, porcentaje de vivienda comercial, el cociente entre estas dos variables y el área que corresponden a aquellas variables que no tienen tanta capacidad explicativa sobre el Número Efectivo de Reproducción. Adicionalmente, se puede evidenciar que la mayoría de los barrios están agrupados en el centro del plano factorial que corresponden a barrios promedio sin atipicidades en sus variables.

Ahora, junto a las variables respuesta están los barrios Alto Melendez, Base Aérea y Normandia que corresponden a barrios con un alto nivel de propagación en los diferentes picos analizados, Similarmente se encuentran los barrios como Calima, Santa Helena y el Pondaje que presenta una propagación media a comparación de los barrios anteriores.

En la Tabla **5-3** se presenta la matriz de importancia para la construcción de las proyecciones en las componentes 1 y 2; esta tabla correspode a un valor cuantitativo para medir la importancia de cada variable en la construcción de cada componente principal. En este caso, las variables que más contribuyeron a la creación de ambas componentes simultáneamente fueron los indicadores de servicios básicos y complementarios.

Seguidos de dos variables correspondientes densidades poblacionales como lo son la cantidad de viviendas y hogares que corrobora lo presentado en el Análisis Factorial Múltiple presentado en la Figura **5-14** donde el bloque correspondientes a las densidades tienen mucha relación con el Número Efectivo de Reproducción.

	Componente 1	Componente 2
SERVICIOS-BASICOS	2.92	2.58
SERVICIOS-COMPLEMENTARIOS	1.60	1.21
HOGARES	1.06	1.10
VIVIENDAS	1.05	1.11
ESTRATO-MODAL	1.01	1.01
HABITANTES	0.99	1.12
EDAD10-19	0.83	1.06
EDU-SUPERIOR	0.76	1.02
EDU-BASISCA	0.71	1.00
IPM	0.96	0.98
EDAD60-69	0.89	0.92
INDICE-JUVENTUD	0.87	0.82
HOMBRES/MUJERES	0.84	0.64
EDAD20-29	0.83	0.78
EDAD50-59	0.82	0.79
EDAD70-79	0.74	0.93
EDAD80-O-MAS	0.73	0.86
EDAD40-49	0.71	0.54
VIVI-RESIDENCIALES/VIVI-COMERCIALES	0.69	0.91
EDAD30-39	0.53	0.40
DENSIDAD-POBLACIONAL	0.36	0.92
VIVI-COMERCIALES	0.26	0.28
VIVI-RESIDENCIALES	0.16	0.12
AREA	0.15	0.17

Tabla 5-3: Matriz de importancia de las variables que componen el modelo PLS

	β_{i1}	β_{i2}	β_{i3}
AREA	0.06	-0.02	0.06
VIVIENDAS	-0.14	-0.09	-0.12
HOGARES	-0.08	-0.06	-0.08
HABITANTES	-0.05	-0.04	-0.06
DENSIDAD-POBLACIONAL	-0.10	0.13	0.10
HOMBRES/MUJERES	0.18	-0.09	0.03
VIVI-RESIDENCIALES	0.13	-0.09	0.01
VIVI-COMERCIALES	0.02	0.00	-0.01
VIVI-RESIDENCIALES/VIVI-COMERCIALES	0.29	0.13	0.13
SERVICIOS-BASICOS	-0.12	-0.65	-0.63
SERVICIOS-COMPLEMENTARIOS	0.05	-0.02	-0.03
IPM	0.03	-0.01	-0.06
ESTRATO-MODAL	0.02	-0.06	-0.05
EDAD0-1	0.02	-0.06	-0.06
EDAD10-19	-0.25	0.05	-0.05
EDAD20-29	0.07	0.10	0.15
EDAD30-39	0.20	0.04	0.08
EDAD40-49	0.07	-0.08	-0.03
EDAD50-59	-0.15	-0.06	-0.05
EDAD60-69	-0.02	-0.06	-0.07
EDAD70-79	0.11	0.02	0.03
EDAD80-O-MAS	0.08	0.03	0.03
INDICE-JUVENTUD	-0.01	-0.07	-0.04
EDU-BASISCA	0.11	0.01	0.01
EDU-SUPERIOR	-0.06	0.01	0.02

Tabla 5-4: Coeficientes de los modelos de Regresión PLS

Antes de proceder con la interpretación de cada uno de los modelos construidos, se quiere dar la claridad que la interpretación de los coeficientes de la metodología PLS2 se interpreta de forma análoga a la de los modelos de regresión lineal convencionales. Adicionalmente dar la claridad que solo se abordara la interpretación de los coeficientes (variables) con mayor influencia tanto positiva como negativamente a la propagación del Covid-19 (R_t).

Con la información presentada en la Tabla **5-4** se puede afirmar que en el primer pico de pandemia aquellos barrios donde predominan las viviendas de uso residencial con respecto a las de uso comercial, las edades entre 30-39 años y predominaran los hombres tenían mayor propagación del virus. Complementariamente aquellos barrios con edades entre 10-19, 50-59 años y mayor cantidad de viviendas menor propagación presentaron en el primer pico.

En el segundo y tercer pico se evidencia que ahora si la densidad poblacional ya empieza a ser parte importante de la propagación ya que a medida que aumenta la densidad poblacional en los barrios se espera que la propagación sea mas alta.

Por ultimo, para los tres momentos se identifican ciertas variables que contribuyen de forma similar; Por ejemplo entre más viviendas y mayor proporción de servicios básicos tengan los barrios menor es la propagación del R_t. Por el contrario en barrios donde se presentan mayor cantidad de viviendas residenciales que comerciales hace que se tenga una mayor propagación.

En la Ecuación 5-1 se presenta la estructura general del modelo estandarizado construido con todas las variables consideradas para el mismo. Hay que tener en cuenta que es un modelo de regresión con tres ecuaciones de pronostico; donde los coeficientes asociados a cada pronostico se presentan en la Tabla **5-4**

Estructura del modelo de regresión

$$[\mathrm{R}_t]_{kj} = \beta_1(AREA) + \beta_2(VIVIENDAS) + \beta_3(HOGARES) + \beta_4(DEN - POBLA) + \beta_5(HOM/MUJ) + \beta_6(VI - RES) + \beta_7(VI - COMER) + \beta_8(VI - RES/VI - COMER) + \beta_9(SERV - BS) + \beta_{10}(SERV - COMP) + \beta_{11}(IPM) + \beta_{12}(ESTRATO - MODAL) + \beta_{13}(EDAD0 - 9) + \beta_{14}(EDAD10 - 19) + \beta_{15}(EDAD20 - 29) + \beta_{16}(EDAD30 - 39) + \beta_{17}(EDAD40 - 49) + \beta_{18}(EDAD50 - 59) + \beta_{19}(EDAD60 - 69) + \beta_{20}(EDAD70 - 79) + \beta_{21}(EDAD80 - O - MAS) + \beta_{22}(IND - ENVEJ) + \beta_{23}(EDU - BASI) + \beta_{24}(EDU - SUPER) \quad (5\text{-}1)$$

$k = 1, 2, 3$ Instantes de tiempo
$j = 1, 2, 3, ..,249$ Barrios

6 Conclusiones, limitaciones y recomendaciones

6.1. Conclusiones

Las enfermedades de carácter contagioso como el Covid-19, han afectado a toda la población mundial en diferentes periodos del tiempo de una u otra manera, tiempo atrás se contaba con poca información sobre el comportamiento de los virus de este tipo. Es por eso, que era complejo controlar o mitigar este tipo de situaciones en la población, debido a su rápida forma de transmisión y propagación entre seres humanos.

Inicialmente en la parte temporal se encontraron tres picos de contagios importantes en todo el periodo Marzo del 2020 - Mayo 2021, estos picos fueron en los meses de Junio-Julio 2020, Enero 2021 y Abril 2021, se puede decir que en dichos picos la propagación del virus aumento alcanzando más de 500 casos diarios en la ciudad.

Algo importante por mencionar es que cada uno de estos picos, lo antecede una acción y/o evento importante lo cual pudo ser causal de cada uno de los picos presentados a continuación. El primer pico que fue en los meses de Junio -Julio 2020, la poca información y las leves medidas de prevención frente al virus que recién estaba iniciando en la ciudad de Santiago de Cali. El segundo pico fue en el mes de Enero 2021, aquí viene a desbordarse el contagio debido a dos acontecimientos importantes, el primero fue la final del fútbol profesional colombiano, la cual tuvo participación con uno de los equipos insignia de la ciudad y el segundo evento fue las fiestas de fin de año, que por costumbres permiten reuniones en las casas. Finalmente, el tercer pico en el mes de Abril fue posterior a la celebración de la semana Santa, que por tradición las personas se van a diferentes sectores y regresan posiblemente con el virus.

Continuando con la parte temporal, si bien fueron tres picos de contagio los identificados en el periodo de estudio, cada uno de estos picos se comporto de manera diferente. El primer pico en los meses de Junio-Julio 2020 afecto principalmente a la comuna 17 ubicada en el sector sur de la ciudad, lo cual hace pensar que el virus llego de manera importada, ya que la gran mayoría barrios que están en esa comuna son estrato 5 y 6 y se puede considerar que tienen un poder adquisitivo menor y pudieron haber sido contagiados fuera de la ciudad e incluso fuera del país. En el segundo pico se vieron más afectadas las zonas del centro y

centro oriente de la ciudad, posiblemente porque para el segundo pico, asumiendo que las hipótesis de que las personas fueron contagiadas de forma externa (contacto con personas del exterior y en zonas de mayor capacidad adquisitiva) y debido a las relaciones con las personas encargadas de los oficios varios y/o seguridad que se cree que la mayoría de estas personas residen en los barrios que componen el distrito de Aguablanca y estas personas siguieron con la red de contagios hacia esa zona. Finalmente, en el ultimo pico los contagio fueron similares en casi todas las comunas de la ciudad.

Ahora, en la parte espacial se videncia que el Número Efectivo de Reproducción fue muy similar en todas las comunas a lo largo del periodo de estudio, sin embrago, para los meses de marzo 2020 y abril 2020 se evidencia una estimación mas alta en las comunas 6, 13 y 22 como se observa en la Figura **5-12**, estas estimaciones tiene valores entre 2 y 2.5 lo cual indicaría que para dichos sectores la propagación era alta. Por otra parte en los meses de mayo, junio, julio, septiembre, octubre, noviembre, diciembre del 2020 y mayo 2021 estas estimaciones tuvieron valores entre 1 y 1.5, lo cual indicaba que la velocidad de propagación del virus estaba siendo mitigada, posiblemente a las diferentes medidas adoptadas por las autoridades locales, entre ellas el pico y cédula, el cierre de establecimientos nocturnos, la prohibición de espectáculos públicos; entre otros.

La estimación del Número Efectivo de Reproducción (R_t) es una herramienta informativa, que permite encender una luz de alerta a los diferentes organismos de control sobre cual es el probable desarrollo de la propagación en las diferentes zonas y así mismo según el nivel de este tomar acciones de mitigación para combatir este virus.

En la modelación estadística se evidenció que los barrios de la ciudad no presentaban características homogéneas en cuanto a variable socioeconómicas como servicios complementarios (Internet y Gas natural), nivel de escolaridad, el tipo de vivienda; bien sea de carácter comercial o carácter residencial, también en algunos barrios es mayor la población adulta que en otros, todos estos factores de una u otra forma pudieron haber influido en que la propagación del virus en cada uno de los tres picos escogidos para este trabajo haya afectado diferentes barrios y no a todos por igual. Adicionalmente, se creía que este virus se propaga más en sectores con mayor población y se pudo evidenciar que no es así. Sino que este puede verse más relacionado con la falta de acceso a servicios básicos que permitieran en una constante desinfección de manos y otros protocolos de seguridad. En términos generales dependiendo del momento de la pandemia hubieron factores socioeconómicos que impactaban más que otros.

Finalmente, de forma general se puede decir que este trabajo permite tener un panorama del comportamiento de la propagación del virus en los barrios y comunas de la ciudad de Santiago de Cali, así mismo sienta una base de los factores socio económicos presentes en los barrios que pueden estar influenciando la velocidad del contagios, todo esto sienta un precedente para realizar futuras investigaciones o que se empleen acciones desde los entes locales para mitigar

futuros acontecimientos de salud pública que pueden afectar a la población de la ciudad. También es importante mencionar que la metodología y archivos empleados en este trabajo de grado quedan a disposición pública para realizar futuras investigaciones o emplearlas en casos similares.

6.2. Limitaciones

Inicialmente, cuando la enfermedad del Covid-19 ingreso a Colombia, mas exactamente cuando inicio en la ciudad de Santiago de Cali, los entes de control no contaban con información precisa o no tenían una metodología especifica para obtener los datos de los contagios diarios (incidencia); eran muy incipientes los protocolos para la toma de muestras de los posibles casos de contagio. Esto generó algunas complicaciones como el desface en las fechas de confirmaciones de los casos (error en el diligenciamiento de las fechas (se encontraron casos confirmados desde antes que se reportara el primer caso confirmado en la ciudad)). Entre otros aspectos que también eran importantes. También algunas características de las hojas de datos presentaban inconsistencias y datos faltantes; lo cual no permitió obtener unos resultados 100 % precisos sobre el comportamiento.

Ahora en la parte de las variables sociodemográficas no se contaba con información, por lo que se debió extraer de fuentes externas que no estaban actualizadas al año correspondiente de estudio. También, al este estudio centrarse en una enfermedad nueva, no se contaba con antecedentes claros y metodologías para la modelacion de la enfermedad, lo cual introdujo un poco de dificultad al utilizar métodos tradicionales y se implementaron metodologías poco comunes para obtener resultados confiables.

6.3. Recomendaciones

Para futuras investigaciones relacionadas con temas de salud pública es importante implementar un sistema de información estandarizado, que permita obtener información más precisa y confiable, evitando la pérdida de información de los registros de casos confirmados. Ahora, con respecto a la información de las características socioeconómicas de los barrios de la ciudad es importante contar con información más actualizada, ya que posiblemente esto puede influir en los resultados obtenidos y así tener un panorama mas actual de la composición de los barrios.

Es importante que se sigan teniendo en cuenta este tipo de metodologías como lo es el PLS, ya que permiten aprovechar relaciones existentes entre variables con alto grado correlación y así evitar excluir dicha variable, que al final de cuentas aporta información sobre el comportamiento del virus.

Implementar otras metodologías como lo son los modelos de regresión poisson siempre y cuando se quiera modelar un solo tiempo y debido a que en este caso se abordaron 3 periodos y la intención de este trabajo de grado era analizarlo simultáneamente. Finalmente seria interesante realizar un modelo que permita tener en cuenta los cambios en el tiempo, que no fue considerado en este trabajo de grado ya que hacia muy extenso debido a que se había trabajado con la depuración de la base de datos, estimación del Número Efectivo de Reproducción y no se consideró abordar el estudio de esa forma; lo mencionado anteriormente se propone con el fin de estudiar todo el periodo de la pandemia y así observar otro tipo de diferencias o relaciones que puedan existir y que esto contribuya a una mejor explicación de la propagación de este tipo de enfermedades.

Bibliografía

Alcaldía de Santiago de Cali (2021). Información geográfica. [Web; accedido el 19-12-2021].

Aylward, B., Liang, W., et al. (2020). Report of the who-china joint mission on coronavirus disease 2019 (covid-19). *WHO-China Jt Mission Coronavirus Dis 2019*, 2019:16–24.

Bellot, D. (2020). NÃEfectivo de ReproducciÃdel COVID-19 en Bolivia. *Revista Boliviana de FÃsica*, 37:31 – 40.

Bernabeu-Mestre, J. et al. (2004). Epidemias y globalización: nuevos y antiguos retos en el control de las enfermedades transmisibles.

Bettencourt, L. M. A. and Ribeiro, R. M. (2008). Real time bayesian estimation of the epidemic potential of emerging infectious diseases. *PLOS ONE*, 3(5):1–9.

Chin, A. W., Chu, J. T., Perera, M. R., Hui, K. P., Yen, H.-L., Chan, M. C., Peiris, M., and Poon, L. L. (2020). Stability of sars-cov-2 in different environmental conditions. *The Lancet Microbe*, 1(1):e10.

Comincini Cantillo, E., Wilches Visbal, J. H., and Saraví, F. D. (2021). Factores epidemiológicos r0 y re durante la covid-19:¿ qué son y en qué difieren? *rev. cuid.(Bucaramanga. 2010)*, pages e1393–e1393.

Cori, A., Ferguson, N., Fraser, C., and Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *American journal of epidemiology*, 178.

Cortés, M. E. (2020). Coronavirus como amenaza a la salud pública. *Revista médica de Chile*, 148(1):124–126.

Cuartas, D. E., Arango-LondoÃ, D., GuzmÃ¡n-Escarria, G., MuÃ, E., Caicedo, D., Ortega, D., FandiÃ-Losada, A. A., Mena, J., Torres, M., Barrera, L., and MÃ, F. A. (2020). AnÃ¡lisis espacio-temporal del SARS-COV-2 en Cali, Colombia. *Revista de Salud PÃ*, 22.

DANE (2018). Censo nacional de población y vivienda. [Web; accedido el 20-01-2022].

Datos Abiertos (2022). Casos positivos de covid-19 en colombia. [Web; accedido el 05-06-2022].

Delfa, J. and Calleja, J. (2003). *Regresión PLS en las ciencias experimentales*. Línea 300. Editorial Complutense, S.A.

Diez-Fuertes, F., Perez-Gomez, B., Alvarez-del Arco, D., de Coronavirus, G. d. A. C., et al. (2020). Glosario de terminos epidemiologicos.

Escofier, B. and Pagès, J. (1990). Analyses factorielles simples et multiples: Objectifs. *méthodes et interpretation*, 1:284.

Escofier, B. and Pagès, J. (1992). Análisis factoriales simples y múltiples: objetivos, métodos e interpretación. *España: Universidad del País Vasco1992*.

Estrada-Alvarez, J. M., Ospina-Ramírez, J. J., Hincapié-Acuña, M., and Gómez-González, M. d. P. (2020). Estimación del intervalo serial y número reproductivo básico para los casos importados de covid-19. *Revista de Salud Pública*, 22(2).

Feito, L. (2007). Vulnerabilidad. In *Anales del sistema sanitario de Navarra*, volume 30, pages 07–22. SciELO Espana.

Gaviria Peña, C. A. (2016). Regresión por mínimos cuadrados parciales pls aplicada a datos variedad valuados. *Escuela de Estadística*.

González Rojas, V. M. (2016). Inter-battery factor analysis via pls: The missing data case. *Revista Colombiana de Estadística*, 39(2):247–266.

Grillo Ardila, E. K., Santaella-Tenorio, J. A., Guerrero, R., and Bravo, L. E. (2020). Mathematical model and COVID-19. *Colombia MÃ*, 51.

Helland, I. S. (1988). On the structure of partial least squares regression. *Communications in statistics-Simulation and Computation*, 17(2):581–607.

Hincapié, Doracelly Ospina, J. (2007). Bases para la modelación de epidemias: el caso del síndrome respiratorio agudo severo en canadá. *Revista de Salud Pública*, 9:117–128.

Höskuldsson, A. (1988). Pls regression methods. *Journal of chemometrics*, 2(3):211–228.

Huang, X., Wei, F., Hu, L., Wen, L., and Chen, K. (2020). Epidemiology and clinical characteristics of covid-19. *Archives of Iranian medicine*, 23(4):268–271.

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721.

Knobler, S. (2004). Institute of medicine (us) forum on microbial threats, institute of medicine (us) board on global health: Learning from sars: preparing for the next disease outbreak. In *Workshop summary. Washington, DC: National Academies Press*.

Kucharski, A. J., Russell, T. W., Diamond, C., Liu, Y., Edmunds, J., Funk, S., Eggo, R. M., Sun, F., Jit, M., Munday, J. D., et al. (2020). Early dynamics of transmission and control of covid-19: a mathematical modelling study. *The lancet infectious diseases*, 20(5):553–558.

Márquez Ruiz, C. (2017). Modelo de regresión pls.

Mason, C. H. and Perreault Jr, W. D. (1991). Collinearity, power, and interpretation of multiple regression analysis. *Journal of marketing research*, 28(3):268–280.

Mejıa Becerra, J. D. (2020). Modelación matemática de la propagación del sars-cov-2 en la ciudad de bogotá segunda versión. Documento no publicado oficialmente.

Mikler, A. R., Venkatachalam, S., and Abbas, K. (2005). Modeling infectious diseases using global stochastic cellular automata. *Journal of Biological Systems*, 13(04):421–439.

Novales, A. (2010). Análisis de regresión. *Universidad Complutense de Madrid: Madrid, Spain*, page 116.

Peláez Sánchez, O. and Más Bermejo, P. (2020). Brotes, epidemias, eventos y otros términos epidemiológicos de uso cotidiano. *Revista Cubana de Salud Pública*, 46:e2358.

Pereira González, A. (2010). Análisis predictivo de datos mediante técnicas de regresión estadística.

Planeación Municipal (2020). Mapas de división administrativa. [Web; accedido el 14-05-2020].

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramírez, G., Vasquez, M., Camardiel, A., Perez, B., and Galindo, P. (2005). Detección gráfica de la multicolinealidad mediante el h-plot de la inversa de la matriz de correlaciones. *Revista Colombiana de Estadística*, 28(2):207–219.

Sit, T. H., Brackman, C. J., Ip, S. M., Tam, K. W., Law, P. Y., To, E. M., Yu, V. Y., Sims, L. D., Tsang, D. N., Chu, D. K., et al. (2020). Infection of dogs with sars-cov-2. *Nature*, 586(7831):776–778.

Stone, M. and Brooks, R. J. (1990). Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(2):237–258.

Thompson, R., Stockwin, J., van Gaalen, R., Polonsky, J., Kamvar, Z., Demarsh, P., Dahlqwist, E., Li, S., Miguel, E., Jombart, T., et al. (2019). Improved inference of time-varying reproduction numbers during infectious disease outbreaks. *Epidemics*, 29:100356.

Tusell, F. (2011). Análisis de regresión. introducción teórica y práctica basada en r. *Adolescence. An age of opportunity*.

Vargas, A. and Rodríguez, I. (1980). Multicolinealidad. *Revista Colombiana de Estadística*, 1(2).

Vega-Vilca, J. C. and Guzmán, J. (2011). Regresión pls y pca como solución al problema de multicolinealidad en regresión múltiple. *Revista de Matemática Teoría y Aplicaciones*, 18(1):09–20.

Velázquez-Silva, R. I. (2020). Historia de las infecciones por coronavirus y epidemiología de la infección por sars-cov-2. *Revista Mexicana de Trasplantes*, 9(S2):149–159.

Viego, V., Geri, M., Castiglia, J., and Jouglard, E. (2020). Período de incubación e intervalo serial para covid-19 en una cadena de transmisión en bahía blanca (argentina). *Ciência & Saúde Coletiva*, 25:3503–3510.

Villalobos-Arias, M. (2020). Estimation of population infected by covid-19 using regression generalized logistics and optimization heuristics. *arXiv: Populations and Evolution*.

Wold, H. (1975). Soft modelling by latent variables: the non-linear iterative partial least squares (nipals) approach. *Journal of Applied Probability*, 12(S1):117–142.