Universidad del Desarrollo
Facultad de Ingeniería

LA EVOLUCIÓN DE LA PANDEMIA A TRAVÉS DEL TIEMPO EN LA REGIÓN METROPOLITANA DE CHILE

POR: THAMARA LOPEZ PEREIRA

PABLO ROJAS TORRES

Capstone project presentado a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al grado académico de Magíster en Data Science

PROFESOR GUÍA:

DR. MAURICIO RENÉ HERRERA MARÍN

Diciembre 2022

SANTIAGO

AGRADECIMIENTO

Primero nos gustaría agradecer a nuestro asesor de este Capstone project el Dr. Mauricio René Herrera Marín de la Facultad de Ingeniería en la Universidad Del Desarrollo. La disposición del Prof. Herrera siempre fue la mejor cada vez que fue necesario.

Finalmente, expresamos nuestro más profundo agradecimiento a nuestras familias por brindarnos un apoyo incondicional y un aliento continuo a lo largo de nuestros años de estudio y durante el proceso de investigación y redacción de esta tesis. Este logro no hubiera sido posible sin ellos. Gracias.

TABLA DE CONTENIDO

Resumen

Los modelos de ajuste matemático internacionales de los primeros países infectados por el virus COVID 19 han puesto la pauta sobre medidas globales de mitigación y aislamiento social. Los pronósticos y modelos chilenos han variado según la evolución de la epidemia y el efecto de las medidas gubernamentales tomadas, estos modelos se han movido desde escenarios catastróficos en un inicio, quizás por considerar modelos usados en otras realidades, hasta un triunfalismo adelantado de control de la pandemia.

Este trabajo busca la caracterización de COVID 19 en el territorio nacional, específicamente para las principales comunas de la Región Metropolitana, esto mediante atributos internos como vacunación, estadísticas de fallecidos, casos activos entre otros.
La búsqueda de esta caracterización se basó en el uso de técnicas de modelación supervisada y no supervisada, entre los que se destacan:

- Modelos Supervisados: Redes Neuronales RNN, modelos de regresión Prophet y modelos de regresión multilineal ordinaria.
- Modelos No Supervisados: Cluesterización mediante K-Means y DBSCAN

A partir del resultado de estos estudios, y en la medida que se desarrolla este documento, se puede inferir que las hipótesis planteadas se cumplen globalmente en la mayor parte de sus declaraciones.

1. Introducción

La Organización Mundial de la Salud (OMS), declaró al COVID-19 una emergencia de salud pública de importancia internacional, el 11 de marzo de 2020. Hasta el 23 de septiembre de 2022, se han notificado a la OMS 611.421.786 casos confirmados de COVID-19, incluidas 6.512.438 muertes. Al 20 de septiembre de 2022, se han administrado un total de 12.640.866.343 dosis de vacunas. (World Health Organization, 2022).

La propagación del COVID-19, en el área Metropolitana de Santiago, siguió un patrón de concentración, presentándose en los sectores donde los determinantes sociales de la salud, asociados a la vivienda, eran mejores; en comparación a los sectores urbanos, donde la vivienda era un problema social complejo (José Francisco Vergara-Perucich, 2020).

En Términos generales, se considera que el impacto del COVID-19 resulta desproporcionado, en comunidades vulnerables, pues el acceso limitado a los servicios de salud y la vacunación genera alto riesgo de trasmisión de enfermedades (Monita Karmakar, 2021) (I.T. Peres, 2021).

Específicamente para Santiago de Chile, se encontraron asociaciones sólidas entre los resultados de COVID-19 y la situación socioeconómica, según los indicadores de salud y comportamiento (GONZALO E. MENA, 2021).

2. Trabajo Relacionado

Se seleccionaron alrededor de 30 artículos afines que parecieron relevantes, a partir de los cuales se realiza la primera selección de datos a utilizar y se continuó con la limpieza de los datos y respectivos análisis exploratorios, entre estos análisis se destacas los estadísticos básicos que se enmarcan para el uso de técnicas de modelación supervisada y no supervisada.

3. Hipótesis y Objetivos

3.1. Objetivo General

Como objetivo principal de este estudio, es analizar la evolución de la pandemia a través del tiempo, en la Región Metropolitana de Santiago, para profundizar en las causas de su rápida progresión y generar conocimiento a partir de distintos registros disponibles.

3.2. Objetivo Específicos

- Revisar el estado del arte, en base a las publicaciones de Covid-19 en las áreas de interés.
- Encontrar relaciones entre el índice de precariedad en las viviendas, la vacunación y la incidencia de contagios, a lo largo de la pandemia.
- Identificar las medidas que tuvieron mayor éxito para la contención del virus.
- Verificar cambios de comportamiento, tanto del virus como de la población, después o durante la pandemia.
- Búsqueda de un modelo que ajuste la evolución temporal de los casos COVID en escala comunal.

3.3. Hipótesis

- La incidencia de contagio está relacionada con la precariedad de viviendas, existiendo mayor cantidad de contagios, inclusive después de la vacunación, respecto a la media nacional.
- La movilidad explica, en cierta medida, el número de casos, por lo tanto, las restricciones de movilidad generan un impacto positivo en la contención del virus.
- Es posible predecir el comportamiento de la pandemia, a través de modelos no paramétricos.
- Existen diferencias clasificables, por modelos no supervisados, entre las comunas de la región metropolitana.

3.4. Plan de trabajo

Para la confección del plan de trabajo se realizó una carta Gantt con la estimación de los procesos que se deberían de realizar.

	Mes	Septiembre	Septiembre	Octubre	Octubre	Octubre	Octubre	Noviembre	Noviembre	Noviembre	Noviembre	Noviembre / Diciembre
	Nro semana	1	2	3	4	5	6	7	8	9	10	11
	Rango Semana	19-25	26 al 30	3 al 7	10 al 14	17 a 21	24 al 28	2 al 4	7 al 11	14 al 18	21 al 25	28 al 2
Actividad	Días	4	5	5	4	5	5	3	5	5	5	5
Focos de interés según hipótesis												
Preparación Información reuqerida - Entregable I												
Revisión Documentación planificación diagnóstico												
Búsqueda de información abierta												
Procesamiento de datos.												
Revisión razonabilidad metodológica de los datos utilizados												
Réplica de la metodología a datos empíricos.												
Implementación de modelos ajustados.												
Generación de kpis estadísticos utilizados en la industria.												
Desarrollo informe												
Desarrollo Presentación												
Preparación presentación final												

4. Datos y Metodología

4.1. Datos

Los datos utilizados fueron extraídos a partir de fuentes de información públicas, como el repositorio de github del ministerio de ciencia (https://github.com/MinCiencia/Datos-COVID19), en donde se obtuvieron datos como movilidad, vacunación y precariedad de vivienda. Además, se revisaron las principales bases de datos académicas, como Google Académico.

4.2. Metodología

La metodología adquirida en este estudio se divide en 3 etapas, la primer de ellas contempla el marco teórico a utilizar, razón por la cual se revisaron publicaciones de COVID-19; relacionadas principalmente a Movilidad, Vacunación y Precariedad de vivienda, de las principales bases de datos, como Google Académico. Finalmente, se seleccionaron alrededor de 30 artículos afines que parecieron relevantes, a partir de los cuales se basa la segunda etapa que tiene dice relación con la obtención y limpieza de los datos y respectivos análisis exploratorios, entre estos análisis se destacas los estadísticos básicos que se enmarcan en el uso de técnicas de modelación supervisada y no supervisada.

Como última etapa, y consolidando las dos etapas anteriores, se genera la búsqueda de patrones que pudieran caracterizar nuestro estudio, patrones que podrían ir desde modelos no supervisados hasta modelos predictivos que pudieran permitir una caracterización adecuada de eventuales coyunturas de salud como lo fue pandemia COVID-19.

4.3. Limpieza de datos.

Introducción

Se aborda Data Scrubbing mediante varios procesos destinados a mejorar la calidad de los datos, utilizando distintas herramientas y prácticas para eliminar los problemas de nuestro conjunto de datos.

Estos procesos se utilizaron para corregir o eliminar registros inexactos en una base de datos o conjunto de datos. En general, esto significa identificar y sustituir los datos o registros incompletos, inexactos, corruptos o irrelevantes.

Después de una limpieza de datos correctamente realizada, todos los conjuntos de datos deben ser coherentes y estar libres de errores. Esto es esencial para el uso y la etapa que continúa, la explotación de los datos. Sin la limpieza, es probable que los resultados de los análisis estén distorsionados y podían ofrecer un rendimiento deficiente.

Anomalías detectadas

Dentro de los problemas detectados en los datasets disponibles, se tuvieron que subsanar las siguientes anomalías:

- Homogenización de nombres de comuna entre datasets.
- Arreglo de datos anómalos, por ejemplo, en la serie de contagios acumulados, el número de casos del día posterior es inferior al día posterior.
- Relleno de datos para fin de semana, ya que, por ejemplo, en el destaset de contagios diarios de covid el fin de semana no existen registros y por lo tanto se

muestran acumulados el día lunes, por lo tanto se realizó la corrección repartiendo los datos en el fin de semana.

- Homogenización de fecha de inicio del dataset.
- Interpolación de datos faltantes en series de tiempo con datos faltantes.

4.4. Explotación de datos.

En esta etapa exploratoria, se intenta capturar la elocuencia de los datos principalmente por sus características gráficas y la evolución de su comportamiento mediante atributos a través del tiempo.

Tasa de incidencia acumulada por comuna cada 10,000 habitantes

La tasa de incidencia acumulada calculada, corresponde a los contagios acumulados a septiembre de 2022 entre la población de cada comuna y multiplicada por 10,000.

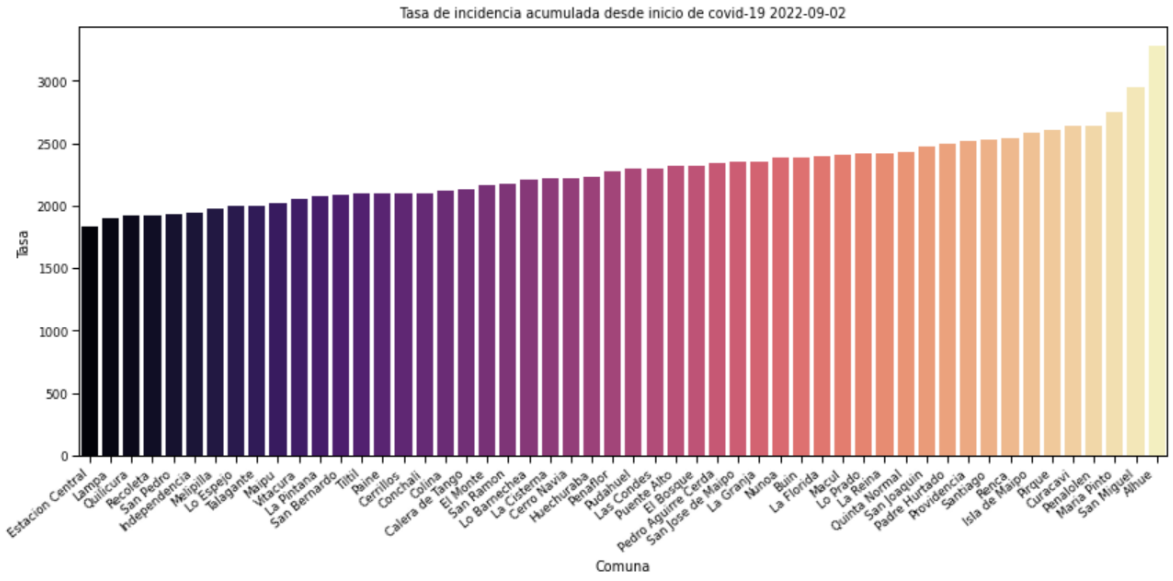


Figura 1 Tasa de incidencia para la Región Metropolitana

Tasa de incidencia por cada 10,000 habitantes y evolución a través del tiempo por comunas

Con la intención de entender cómo se propagó la pandemia en los primeros días, se graficó la tasa de incidencia por comuna en distintas fechas, como se muestra a continuación:

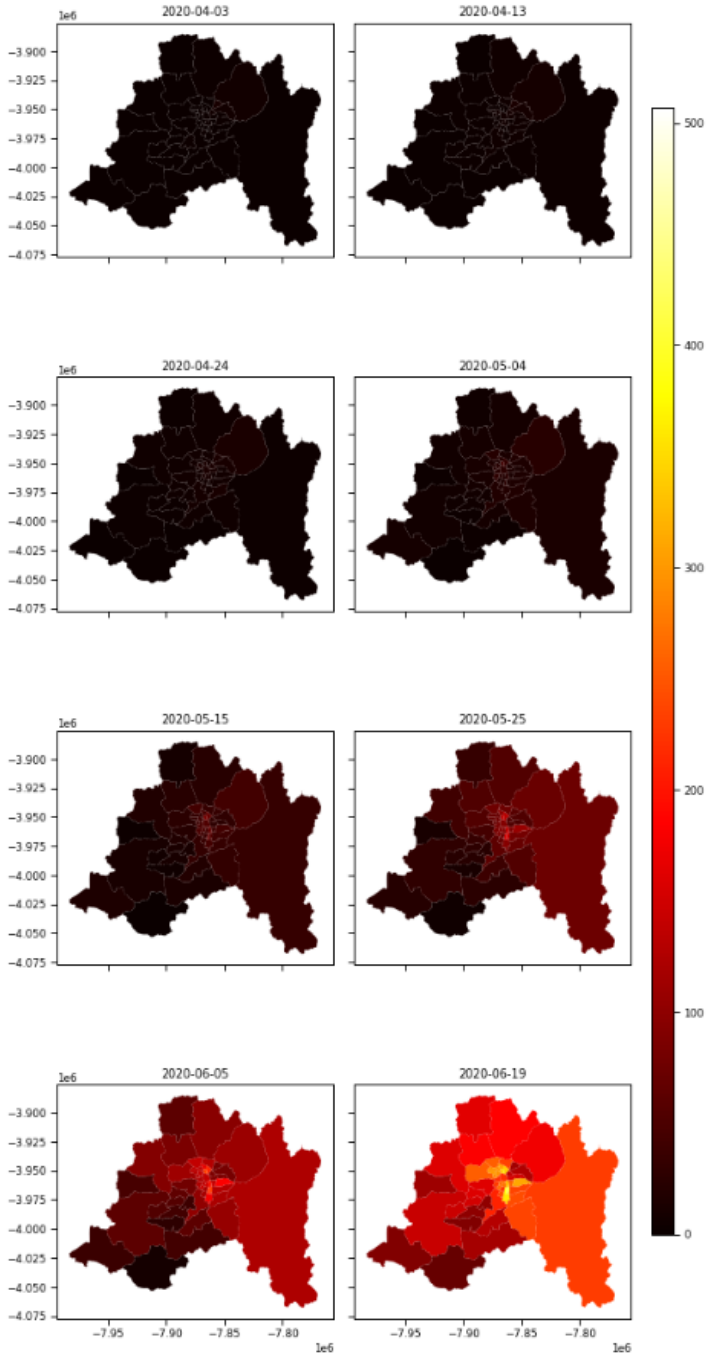


Figura 2 Propagación inicial de la pandemia en la Región Metropolitana

La evolución de la tasa de incidencia través del tiempo, sigue una tenencia natural de estado de pandémico, con contagios que crecen de manera exponencial en una primera etapa de maduración para luego estabilizarse.

Se puede observar a partir de la gráfica, una de las primeras e incipientes conclusiones que dice relación con la hipótesis de que comunas con mejores indicadores de precariedad, poseen menores niveles de contagios.

Autocorrelaciones

La autocorrelación es una representación matemática del grado de similitud entre una serie de tiempo determinada y una versión retrasada de sí misma en intervalos de tiempo sucesivos. Es conceptualmente similar a la correlación entre dos series de tiempo diferentes, pero la autocorrelación usa la misma serie de tiempo dos veces: una vez en su forma original y otra retrasada uno o más períodos de tiempo. (investopedia, 2022)

Autocorrelaciones evolutivas por comuna.

Para todas las comunas se muestra que existe una autocorrelación para los nuevos contagios semanales para cada comuna revisada, a modo de ejemplo se presentan gráficos para algunas comunas analizadas:

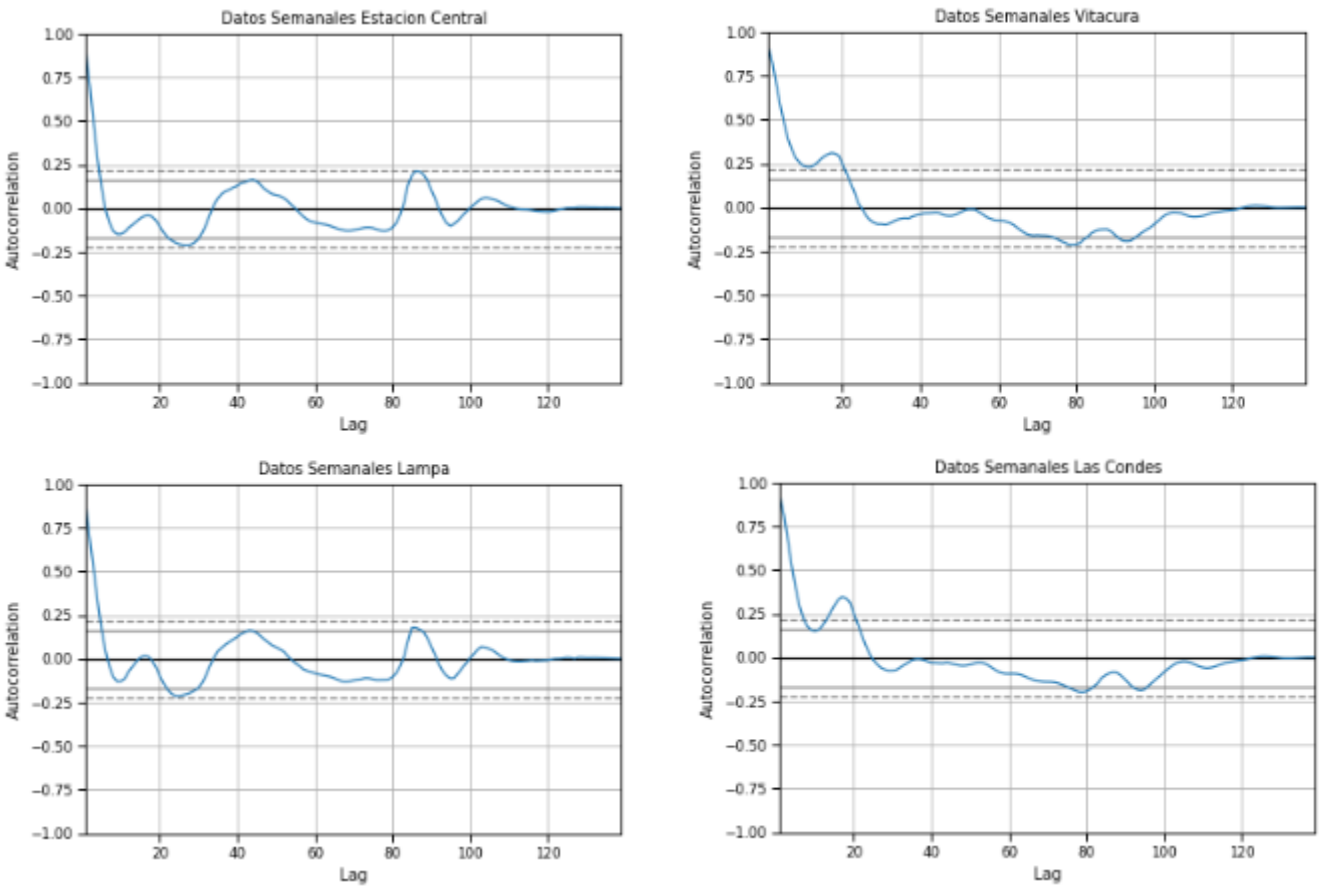


Figura 3 Autocorrelación

Autocorrelaciones consolidadas – distintos periodos.

A partir de la gráfica se puede inferir tendencias de autocorrelaciones totales, tanto por comuna como en una vista consolidada a escala mensual.

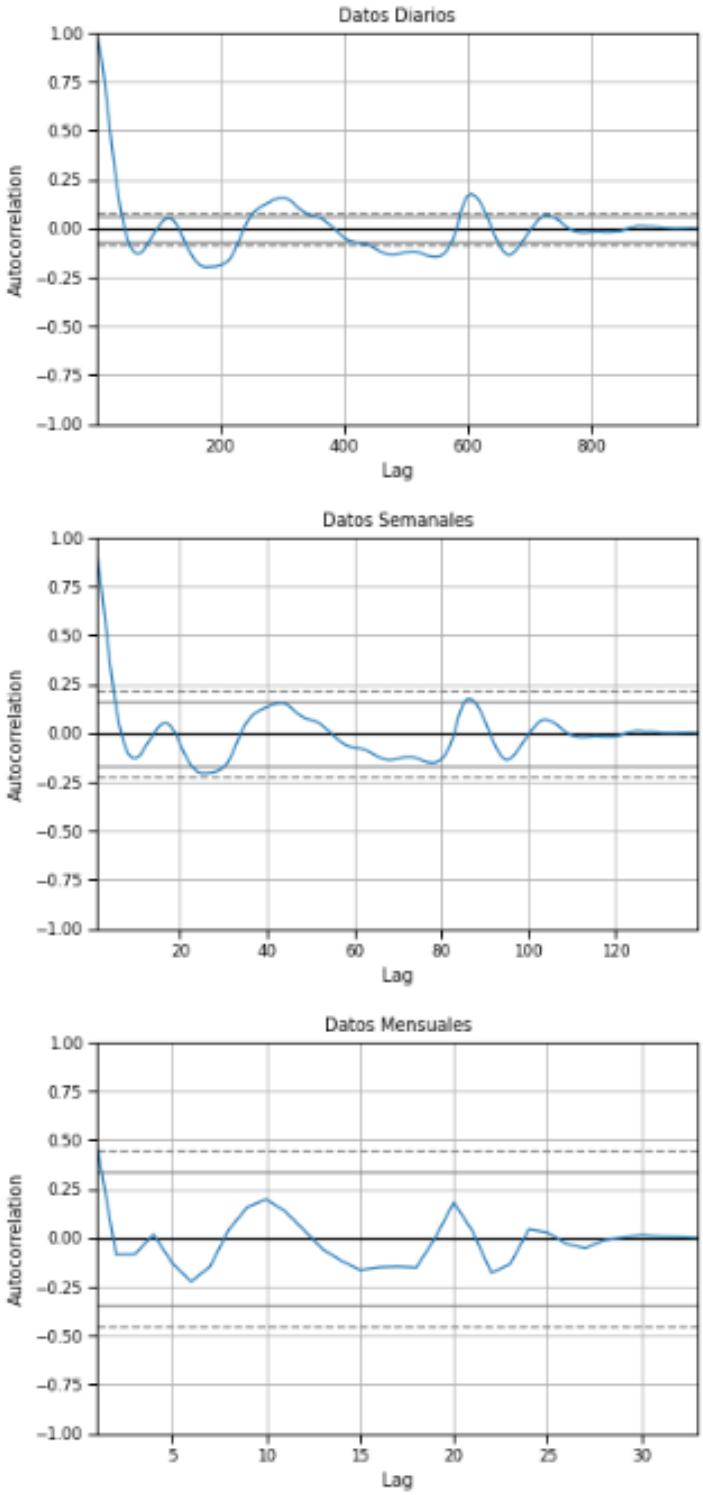


Figura 4 Autocorrelación consolidada

La normalización o Detrending Semanal

El componente estacional de la serie temporal se encuentra restando el componente de tendencia de los datos originales y luego agrupando los resultados y promediando, en este caso se utiliza el paquete de python statsmodels con su clase seasonal_decompose (statsmodels, 2022).

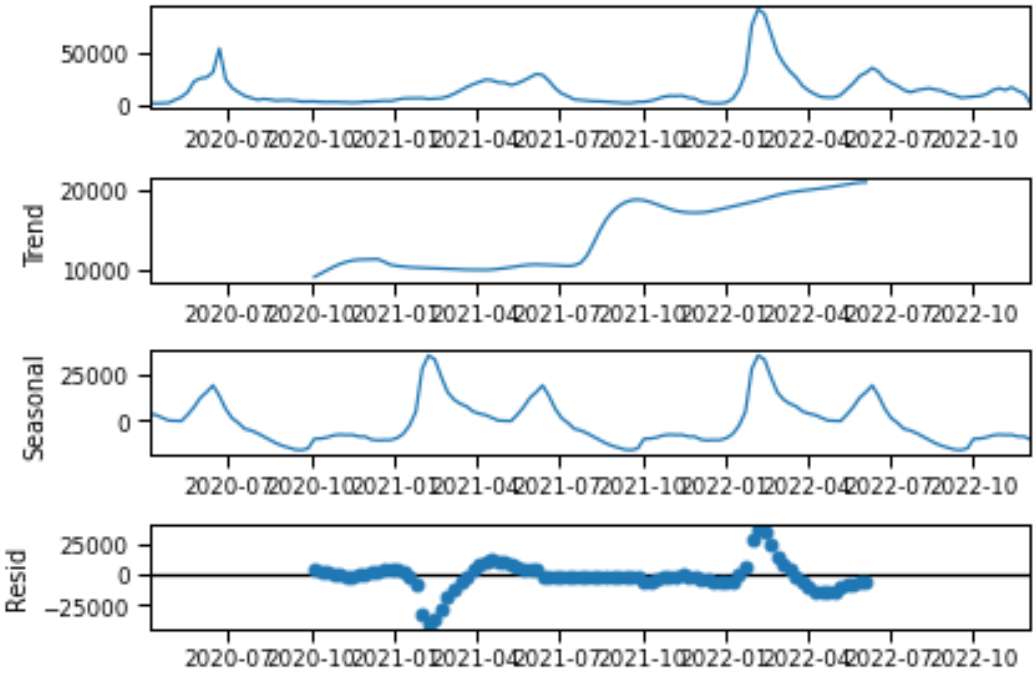


Figura 5 Descomposición estacional de contagios

La descomposición se puede interpretar de la siguiente manera:

- **Tendencia**: la dirección general de la serie durante un largo período de tiempo.
- **Estacionalidad**: un patrón distinto y repetitivo que se observa en intervalos regulares debido a varios factores estacionales. Puede ser mensual, semanal, etc.

- **Residual**: el componente irregular que consiste en las fluctuaciones en la serie temporal después de eliminar los componentes anteriores.

Correlaciones Cruzada

Cross Correlation – Movilidad

Para el cálculo de la correlación cruzada entre el índice de movilidad interno y los casos nuevos en cada comuna, nos encontramos con que todas las comunas de la región metropolitana presentan el mismo comportamiento, como se muestran algunos ejemplos a continuación:

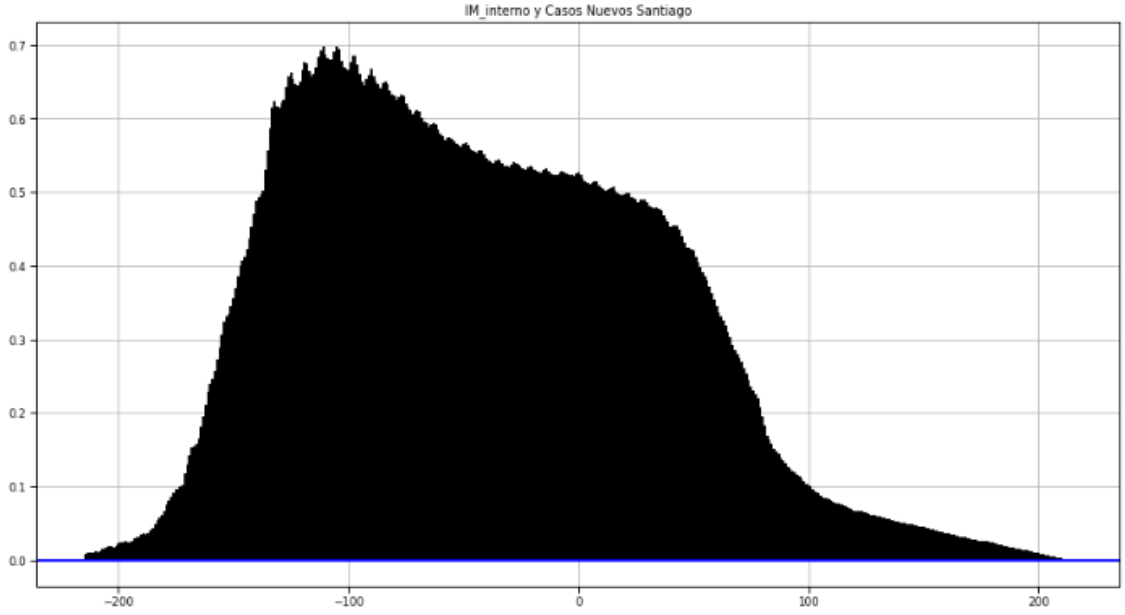


Figura 6 Índice de movilidad interno y casos nuevos en la comuna de Santiago

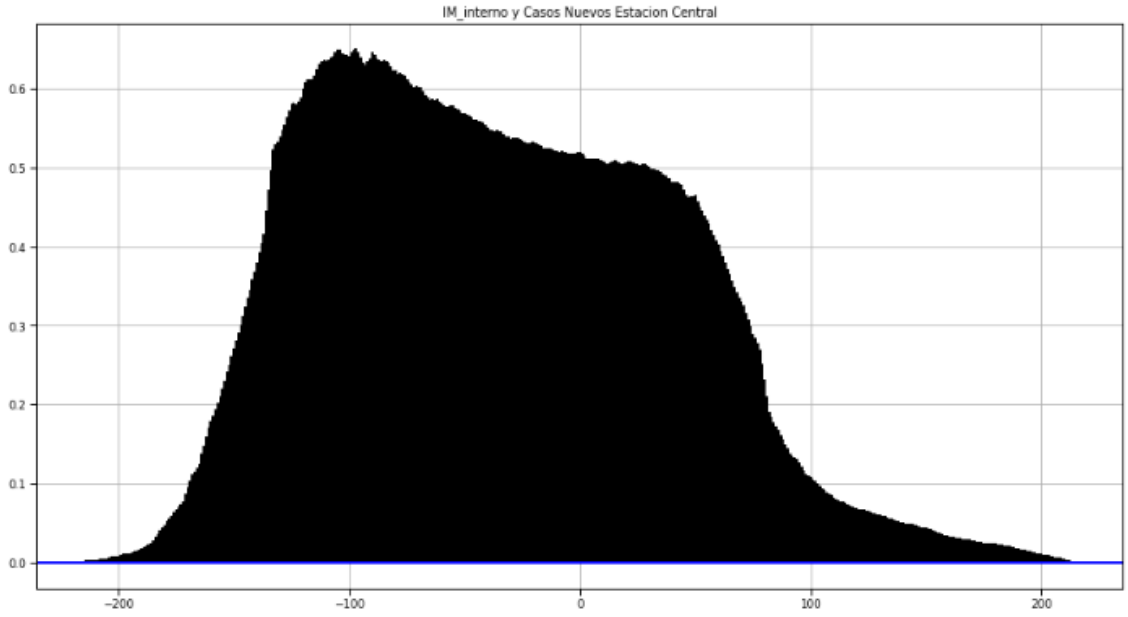


Figura 7 Índice de movilidad y casos nuevos en Estación Central

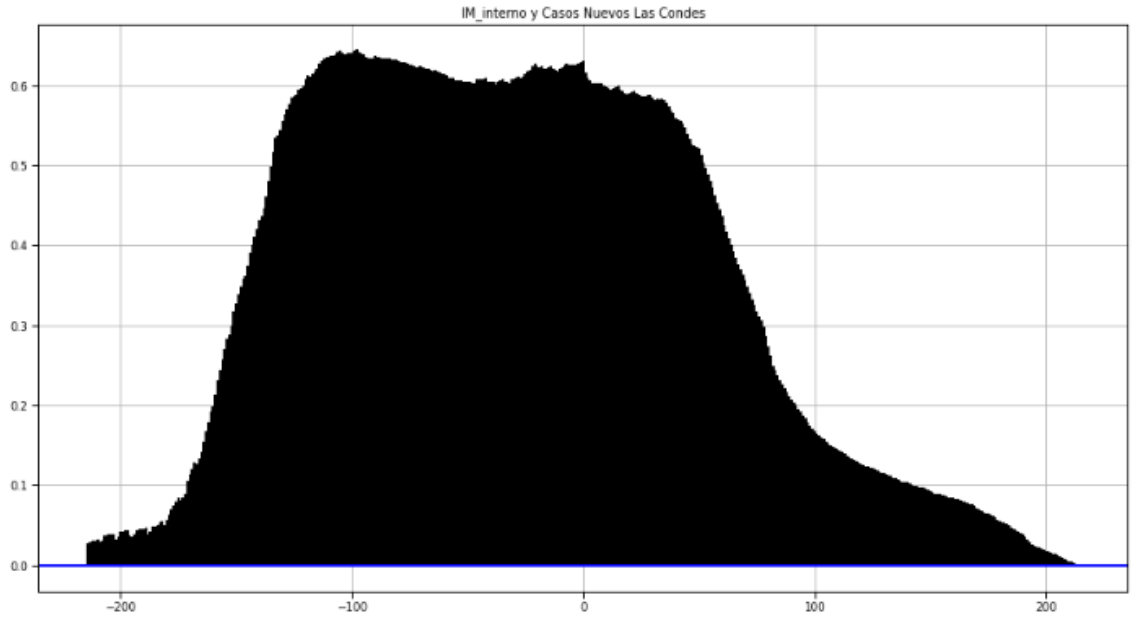


Figura 8 Índice de movilidad y casos nuevos en la comuna de Las Condes

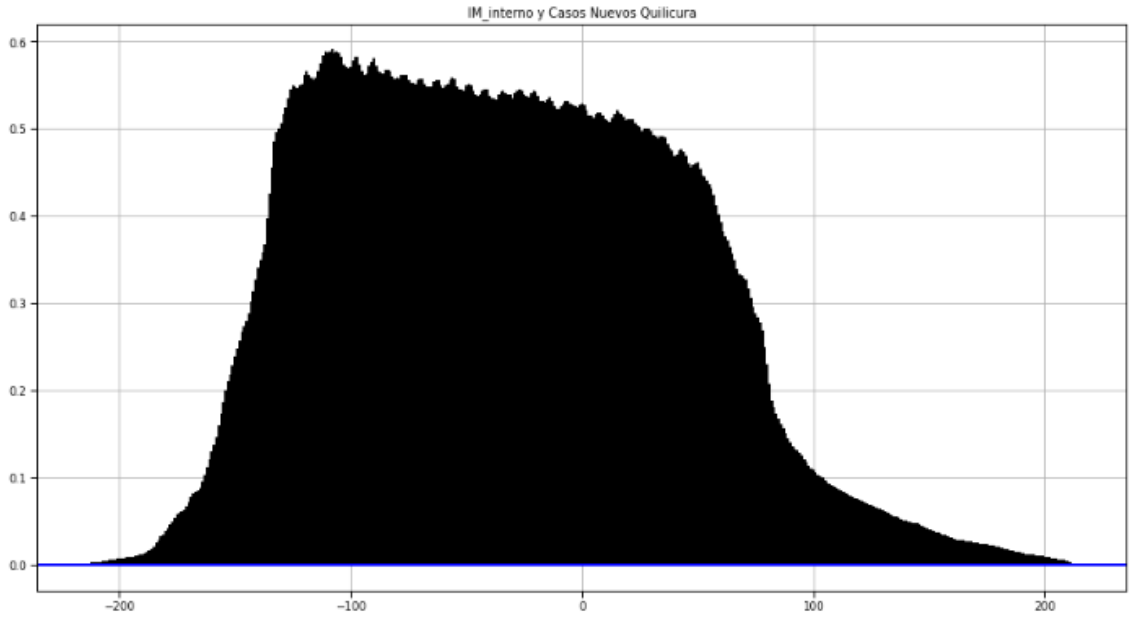


Figura 9 Índice de movilidad y casos nuevos en la comuna de Quilicura

Cross Correlation Vacunación – Nuevos Contagios

Al igual que lo mencionado anteriormente, todas las comunas de la región metropolitana presentan comportamiento similar, por lo que solo se muestran algunos gráficos a modo de ejemplo:

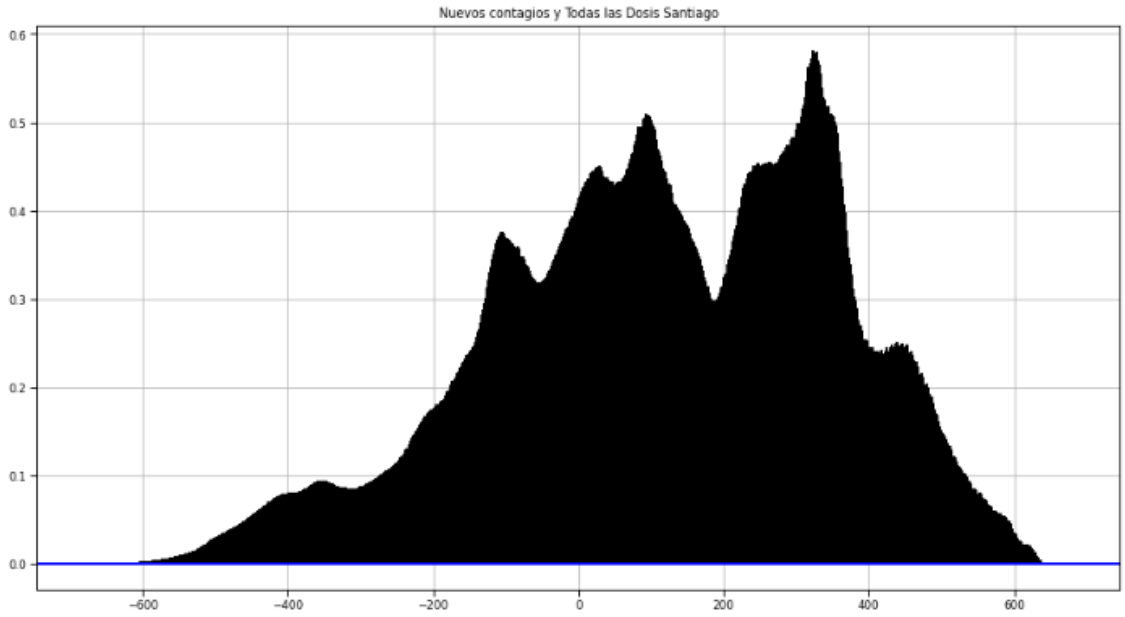


Figura 10 Nuevos contagios y todas las dosis suministradas en la comuna de Santiago

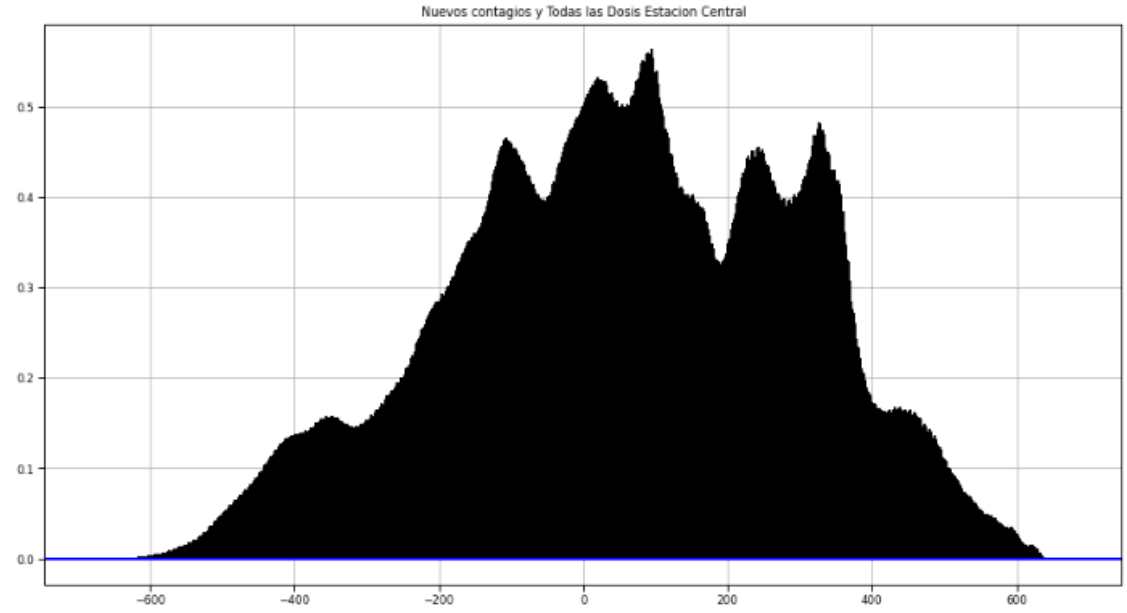


Figura 11 Nuevos contagios y todas las dosis administradas en la comuna de Estación Central

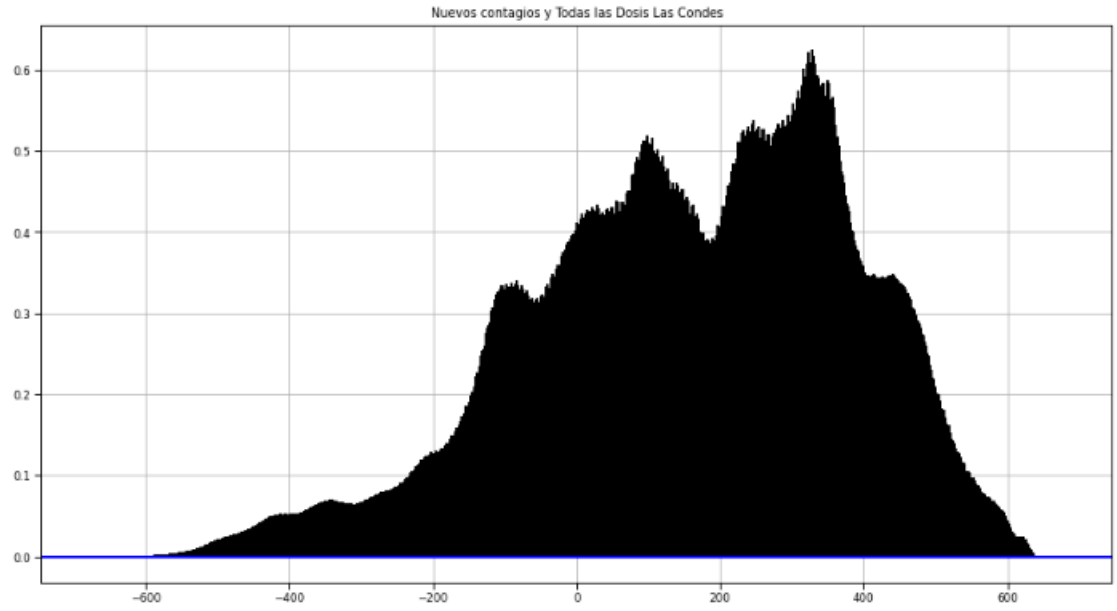


Figura 12Nuevos contagios y todas las dosis administradas en la comuna de Las Condes

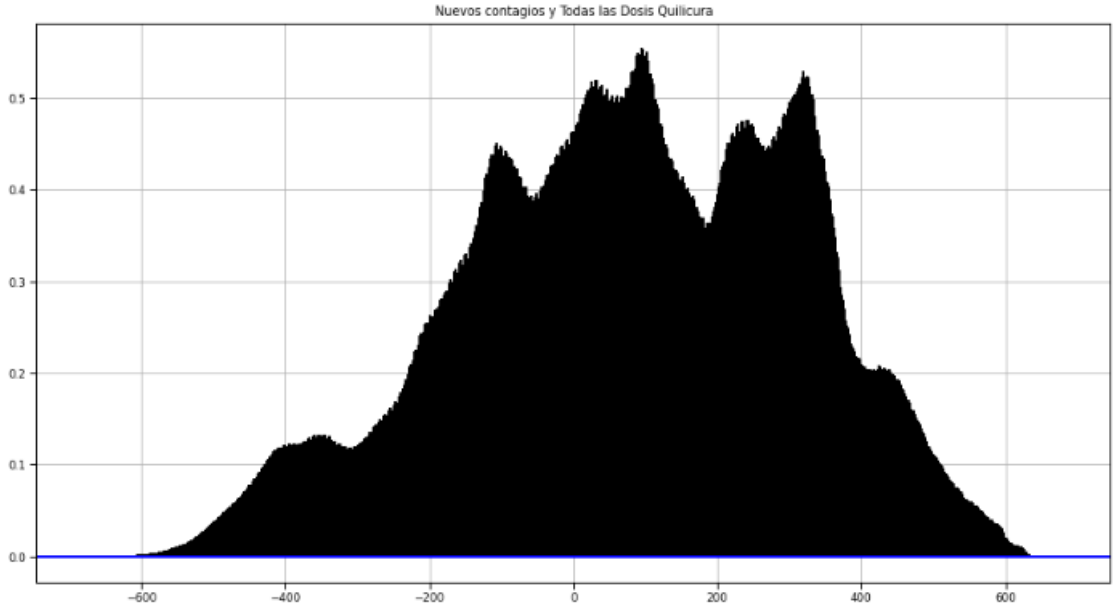


Figura 13Nuevos contagios y todas las dosis administradas en la comuna de Quilicura

En síntesis, se puede ver que la correlación entre contagios y vacunación comienza a bajar paulatinamente hasta los 6 meses, en donde se vuelve a aplicar otra dosis de la vacuna, donde vuelve a subir esta correlación. Esto en primera instancia lo podemos

explicar que a medida que se vacuna gente, la correlación entre más vacunados se vuelve inversa a los contagios (es decir, mayores vacunados implican menores contagiados), sin embargo, esto a modo preliminar de determinación.

Análisis de confinamiento (Plan paso a paso y contagios)

Para la confección del gráfico de análisis de confinamiento y contagios, se consideró la información oficial disponible, se mantuvo la clasificación del dataset obtenido y se realizó un análisis para para cada comuna, si bien los resultados por comuna son similares, se puede obtener información interesante por parte del gráfico, a modo de ejemplo a continuación se muestra la situación de la comuna de “Lo Espejo”:

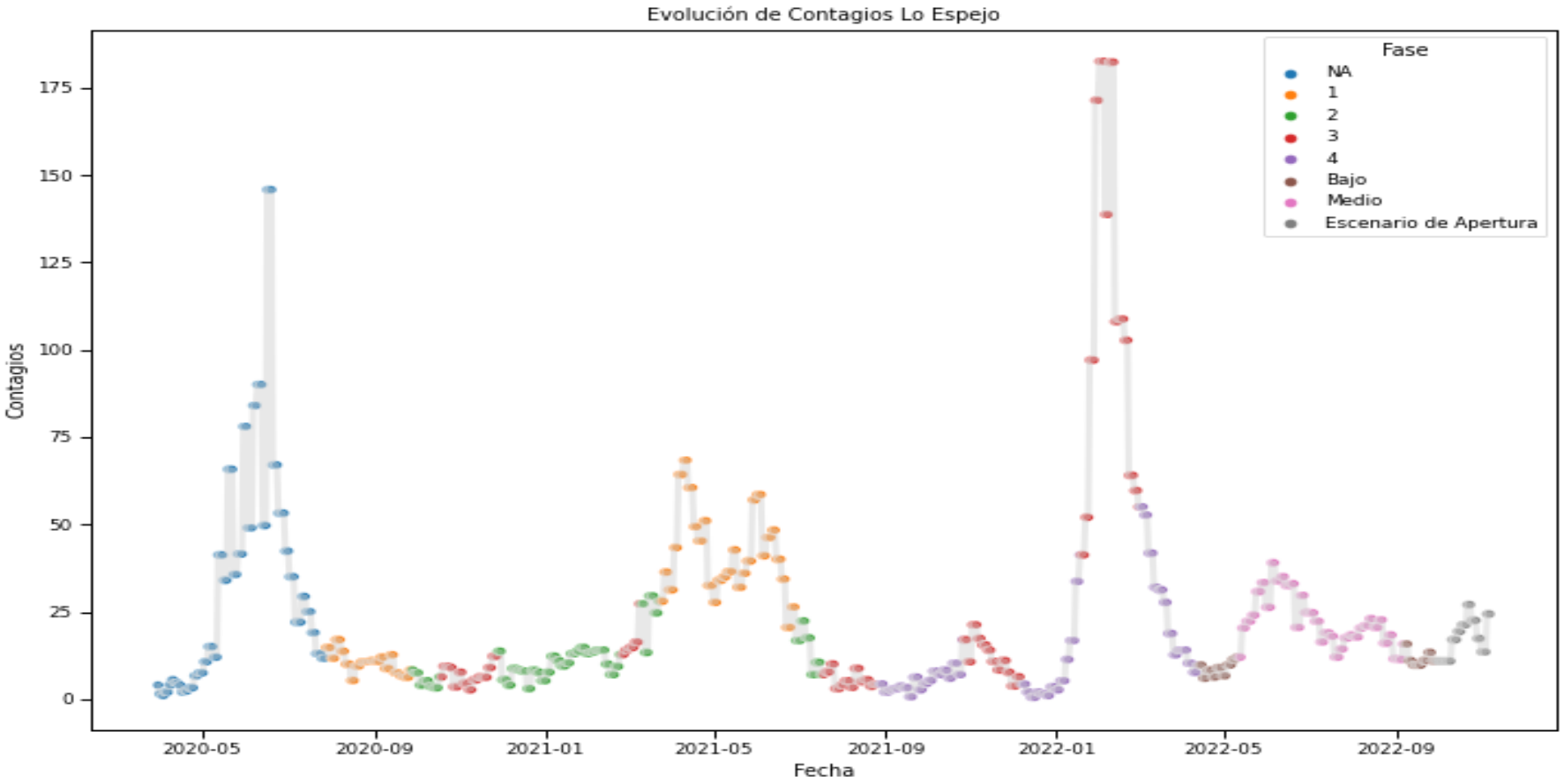


Figura 14Evolución de contagios y Fase en la comuna de Lo Espejo.

Los periodos clasificados como “NA” corresponde a No Aplica, debido a que todavía el plan paso a paso no estaba en funcionamiento, sin embargo, se observan algunas situaciones interesantes:

- En mayo de 2021 se puede ver que los cambios de Fase no tuvieron mayor influencia en los cambios de ratio de contagio, por lo menos en primera instancia. Ya que se ve que desde Fase 2 se pasó a Fase 3, sin embargo, tuvo que ser rápidamente devuelta a Fase 2 para pasar a Fase 1 y enfrentar un alza considerable en los contagios.
- En enero 2022 ya no se aplicaron medidas en base a contagios, ya que a pesar de ser un de las olas más fuertes de la pandemia, la fase se mantuvo en “3” durante todo el periodo, indicando a nuestro criterio que posiblemente fueron medidas económicas y sociales las predominantes en la medida.

Análisis de Movilidad (Plan paso a paso y movilidad)

Para el análisis de movilidad y el plan paso a paso, se encontró la dificultad de que el plan paso a paso posee datos desde agosto de 2020 aproximadamente y los datos de movilidad se encuentran disponibles hasta noviembre de 2020, por lo tanto, la serie a revisar no es tan extensa, si bien se podría rellenar manualmente la información, escapa de los alcances del presente proyecto. Al igual como se mencionó anteriormente los periodos clasificados como “NA” corresponde a No Aplica, debido a que todavía el plan paso a paso no estaba en funcionamiento. Sin embargo, de la información revisada se destaca en primera instancia lo siguiente:

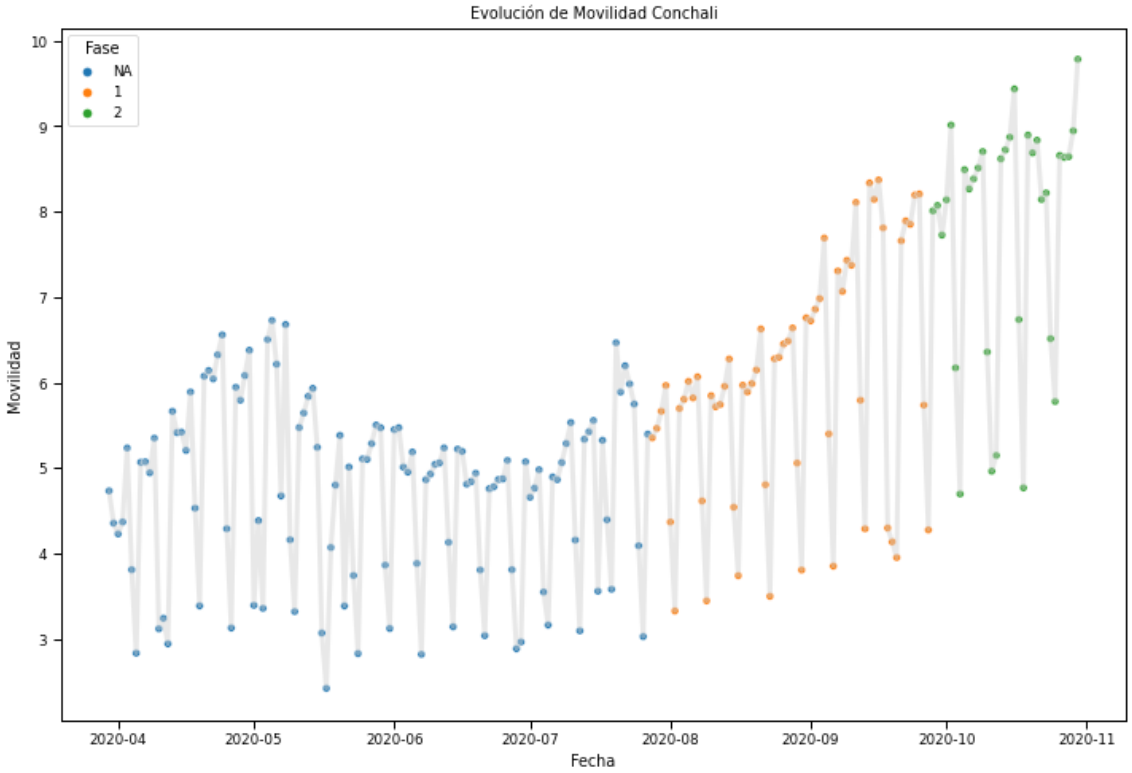


Figura 15 Movilidad y plan paso a paso en la comuna de Conchalí

En la comuna de Conchalí, evidenciamos que a medida que la Fase 1 aumenta en duración, la movilidad externa va en aumento, este fenómeno se reproduce en comunas como Independencia y La Reina:

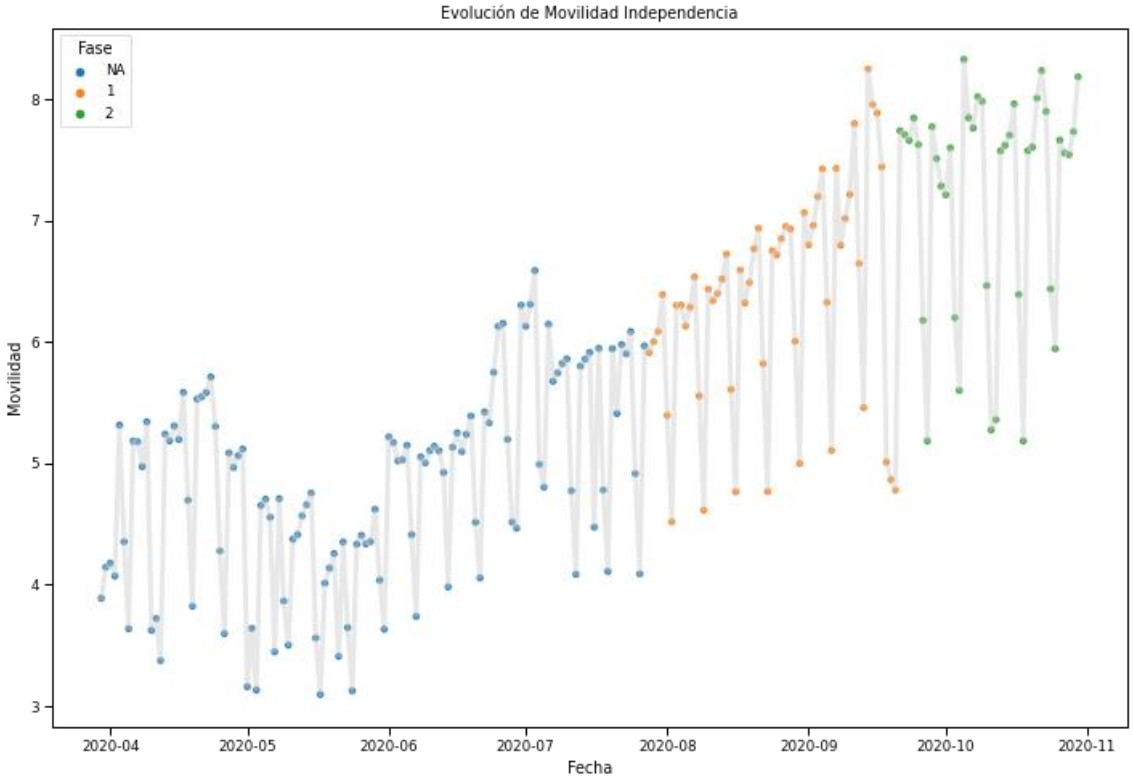


Figura 16 Movilidad y plan paso a paso en la comuna de Independencia

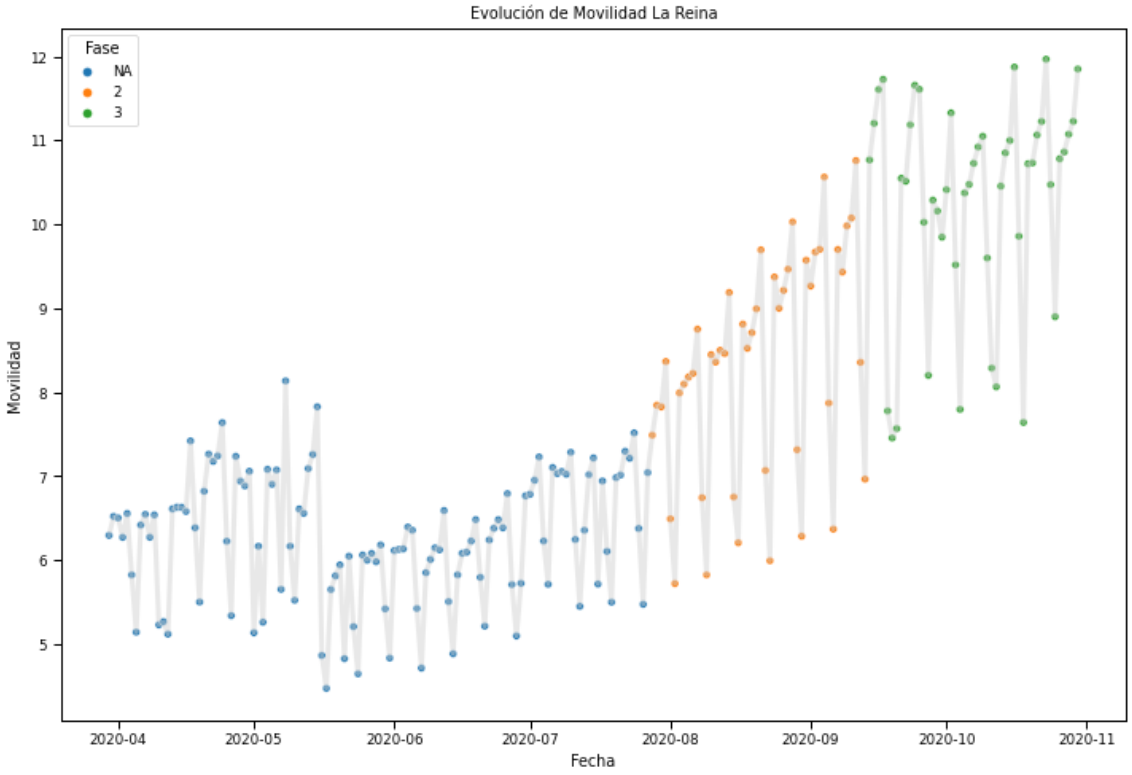


Figura 17 Movilidad y plan paso a paso en la comuna de La Reina

Además, se observa que luego de pasar a Fase 2, se observa una estabilización de la movilidad externa en estas comunas.

Por otra parte, en comunas como La Pintana, se observa una baja efectividad de la medida de confinamiento, como se muestra continuación:

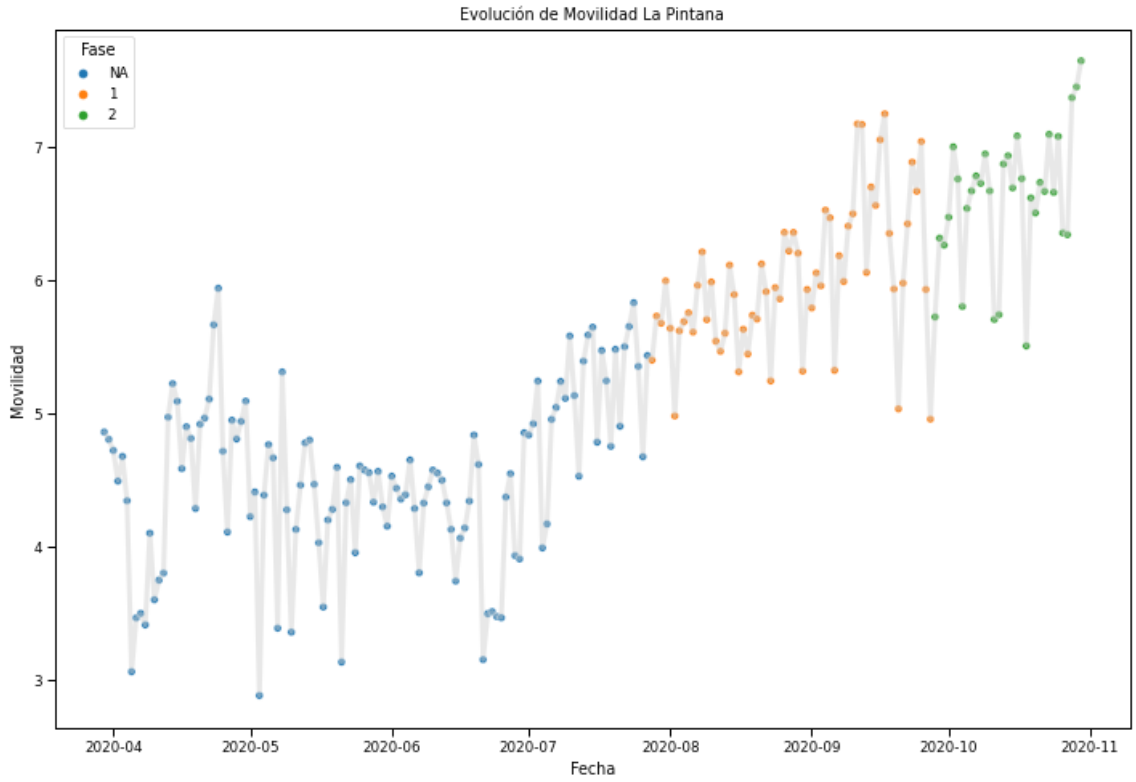


Figura 18 Movilidad y plan paso a paso en la comuna de Conchalí

Análisis de Contagios vs Muertes

En primera instancia se inicia generando una nube de palabras en base al total de contagios y al total de muertes podemos por comuna, es decir que si una comuna posee mayores muertes o contagios se verá en mayor tamaño el nombre de la comuna. Este proceso se realizó con la ayuda del paquete WordCloud disponible para Python. Los resultados pueden verse a continuación:

Figura 19 Nube de palabras en base a muertes por covid

Figura 20 Nube de palabras en base a contagios por covid

Posteriormente se procedió a generar dos mapas de calor, en el primero se muestran los casos nuevos que existieron por cada comuna, en color más oscuro se pueden ver la mayor cantidad de contagios, en la siguiente imagen, se graficó la cantidad de muertes por comuna. Ambos gráficos se encuentran a escala mensual (contagios y muertes acumuladas).

Figura 21 Mapa de calor con contagios para la comuna

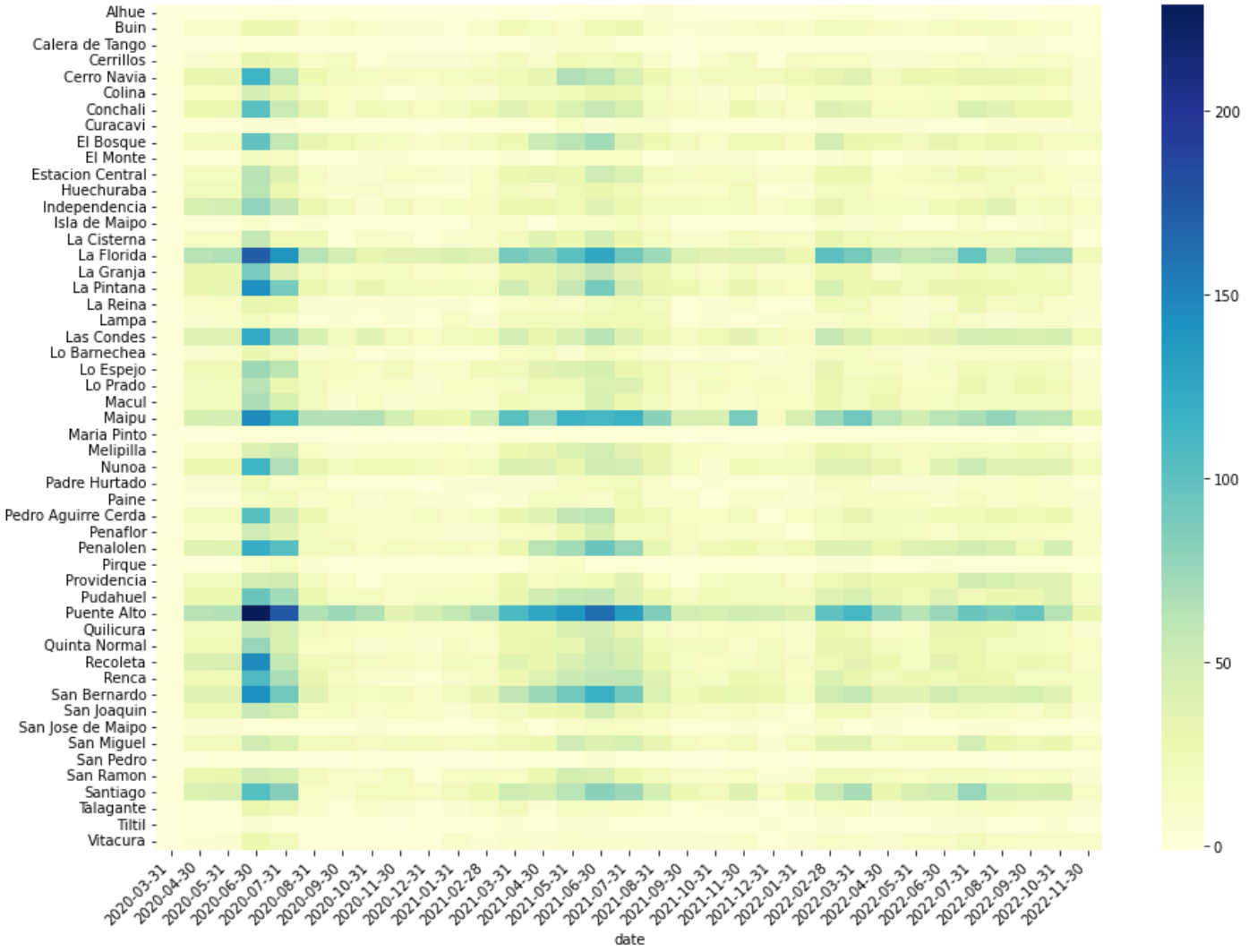
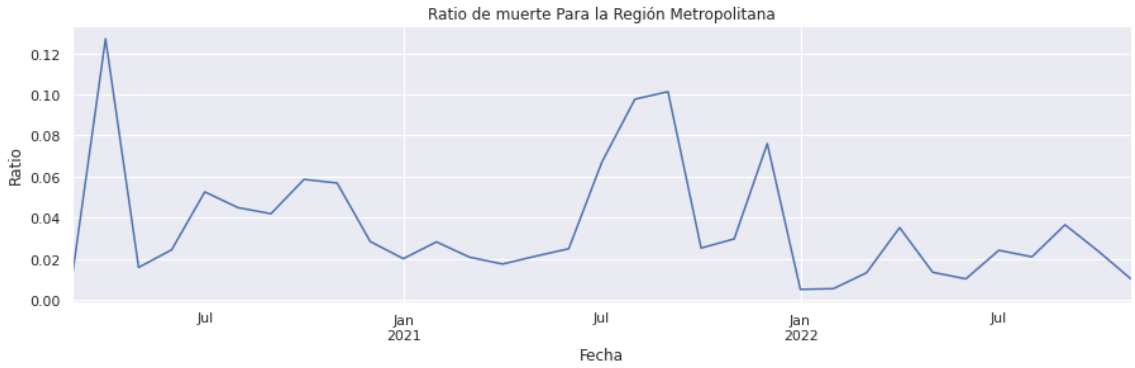


Figura 22 Mapa de calor con muertes por comuna

Se puede ver que el ratio de muerte aumenta a medida que aumentan los contagios, comportamiento que parece lógico, sin embargo, de igual manera se decide graficar el ratio de muerte para la región metropolitana (muertes mensuales dividido en contagios mensuales):



Se comprueba que posee una leve tendencia a la baja, sin perjuicio a ello, se ha mantenido relativamente constante a través del tiempo en términos generales (se debe considerar que no existe diferenciación de edad en este gráfico).

El detalle del ratio promedio mensual de muerte para cada comuna, es el siguiente:

Comuna	Ratio promedio mensual	Comuna	Ratio promedio mensual
Alhue	0.097	Las Condes	0.033
San Jose de Maipo	0.080	Paine	0.033
Lo Espejo	0.054	Maipu	0.033
La Pintana	0.054	Macul	0.033
San Ramon	0.053	San Bernardo	0.032
Conchali	0.052	Cerrillos	0.032
Cerro Navia	0.049	Talagante	0.032
Independencia	0.048	Quinta Normal	0.031
Pedro Aguirre Cerda	0.045	Penalolen	0.030
La Granja	0.044	Puente Alto	0.029
Penaflor	0.042	Pudahuel	0.028
Recoleta	0.042	La Reina	0.028
Lo Prado	0.041	Maria Pinto	0.028
San Joaquin	0.041	San Pedro	0.027

Melipilla	0.040	Buin	0.027
La Cisterna	0.040	Estacion Central	0.026
La Florida	0.039	Lampa	0.025
El Bosque	0.037	Calera de Tango	0.024
Isla de Maipo	0.036	Vitacura	0.024
El Monte	0.035	Colina	0.024
Renca	0.035	Quilicura	0.023
Nunoa	0.035	Padre Hurtado	0.023
San Miguel	0.035	Curacavi	0.023
Huechuraba	0.034	Pirque	0.021
Tiltil	0.034	Santiago	0.017
Providencia	0.034	Lo Barnechea	0.017

Por otro lado, si bien los gráficos anteriores en primera instancia no consideran la población, llama poderosamente la atención que comunas como Puente Alto, poseen alta cantidad de muertes durante toda la pandemia, no obstante, al realizar el mismo ejercicio cada 10,000 habitantes podemos ver que se homogenizan bastantes los gráficos, como se muestra a continuación:

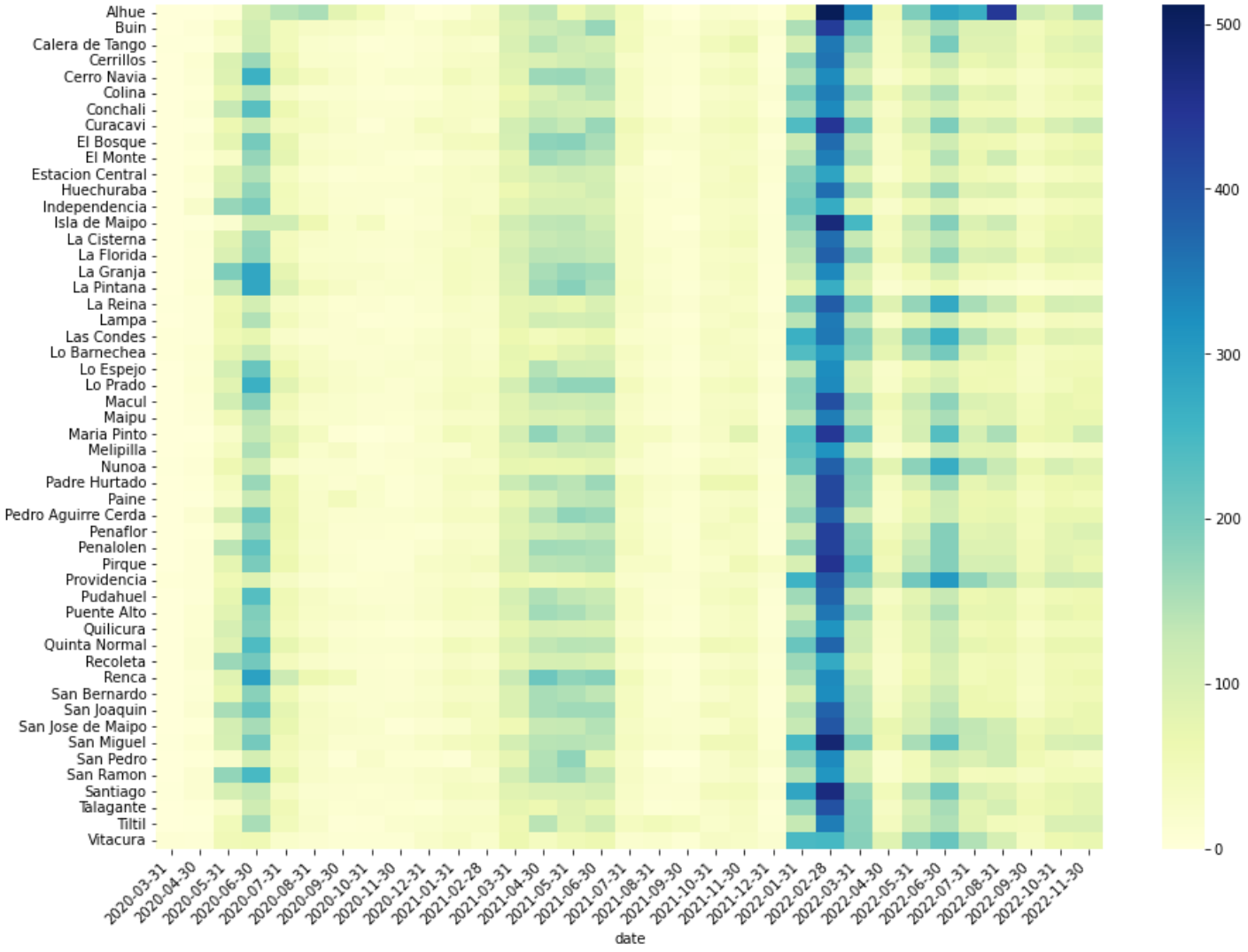


Figura 23 Mapa de calor de contagios cada 10,000 habitantes entre comunas

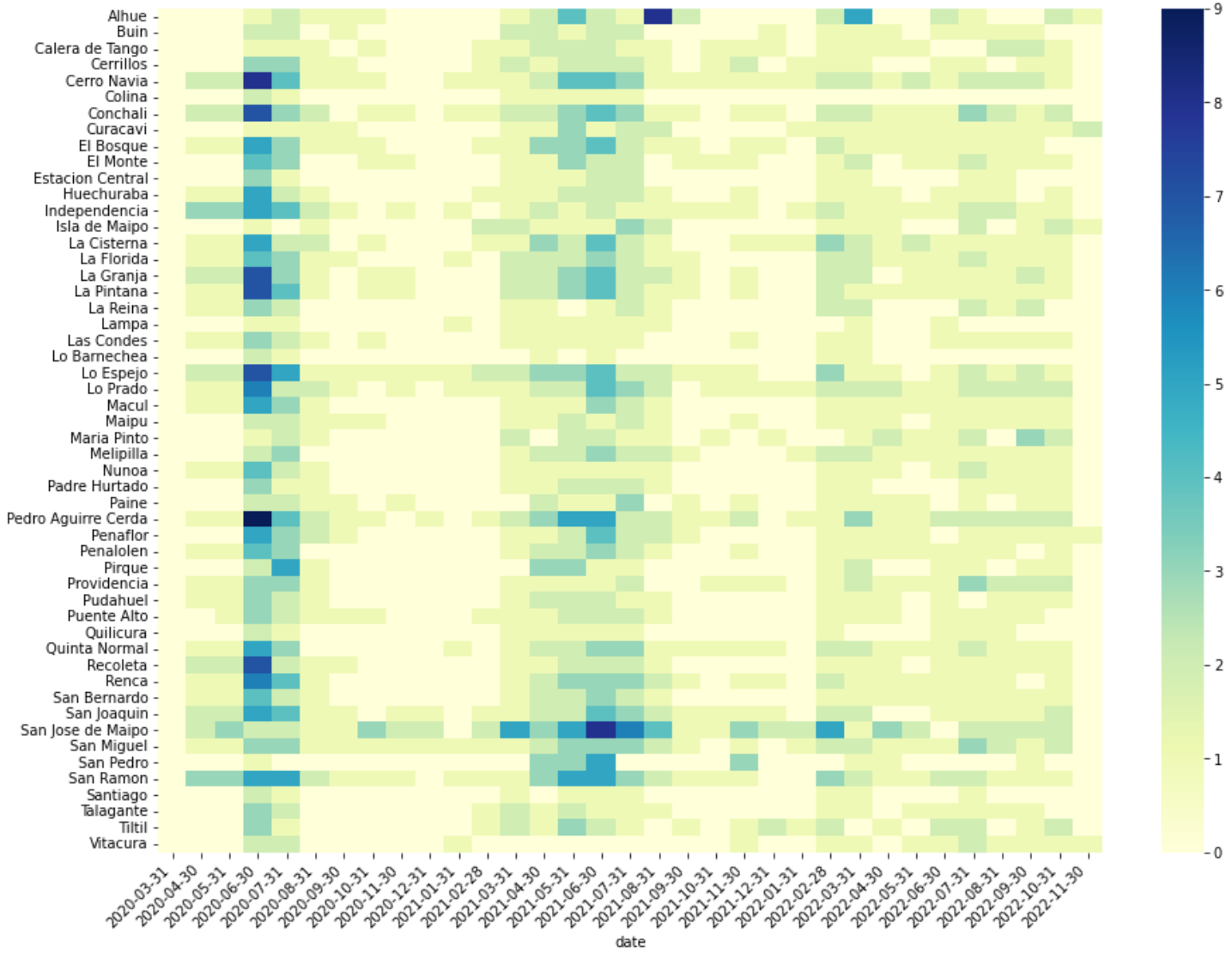


Figura 24 Mapa de calor de distancia con muertes cada 10,000 habitantes entre comunas

Ocupando las muertes cada 10,000 habitantes para cada comuna, se puede apreciar que las máximas muertes se producen en las comunas Pedro Aguirre Cerda, Cerro Navia, San José de Maipo, Conchalí, La Granja, Lo Espejo, La Pintana, Recoleta y Alhue.

Con relación a las muertes durante el tiempo y la edad, se puede observar que la mayoría de las comunas poseen un comportamiento similar, como se muestra a continuación:

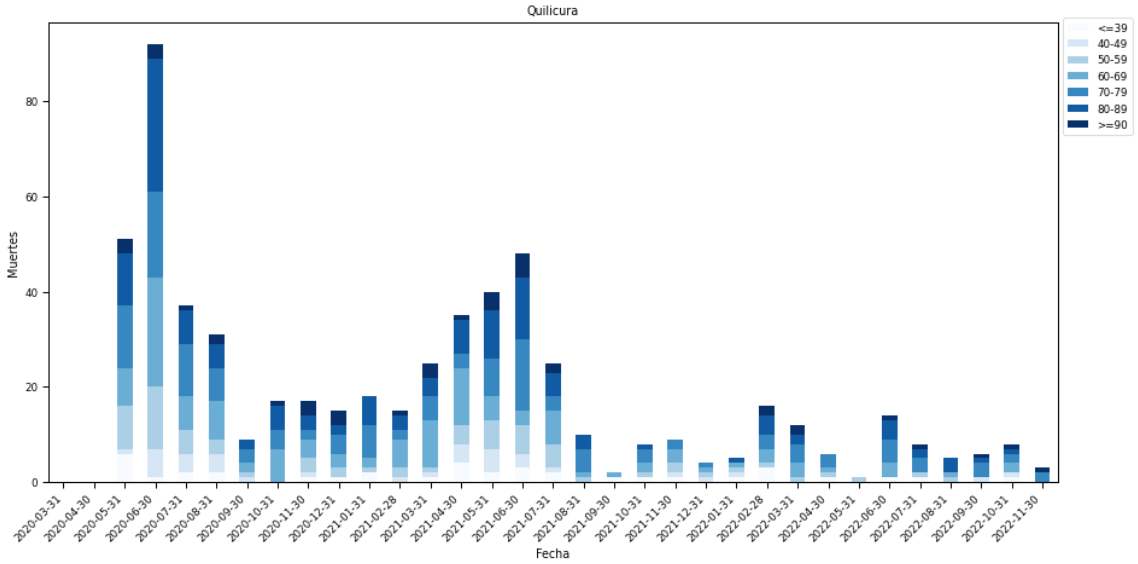


Figura 25 Muertes por edad durante el tiempo comuna de Quilicura

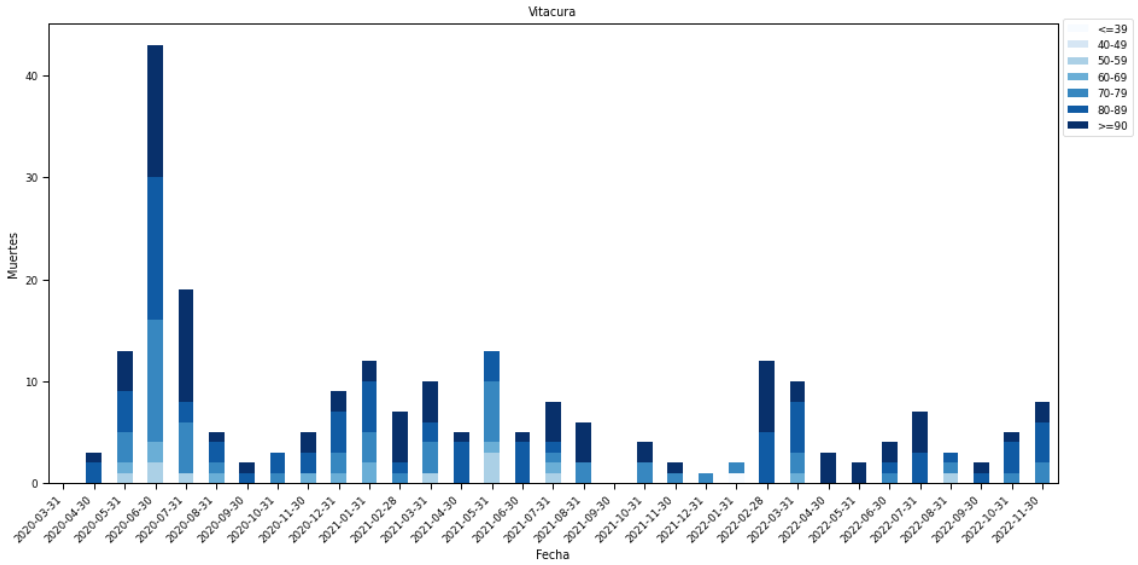


Figura 26 Muertes por edad durante el tiempo comuna de Vitacura

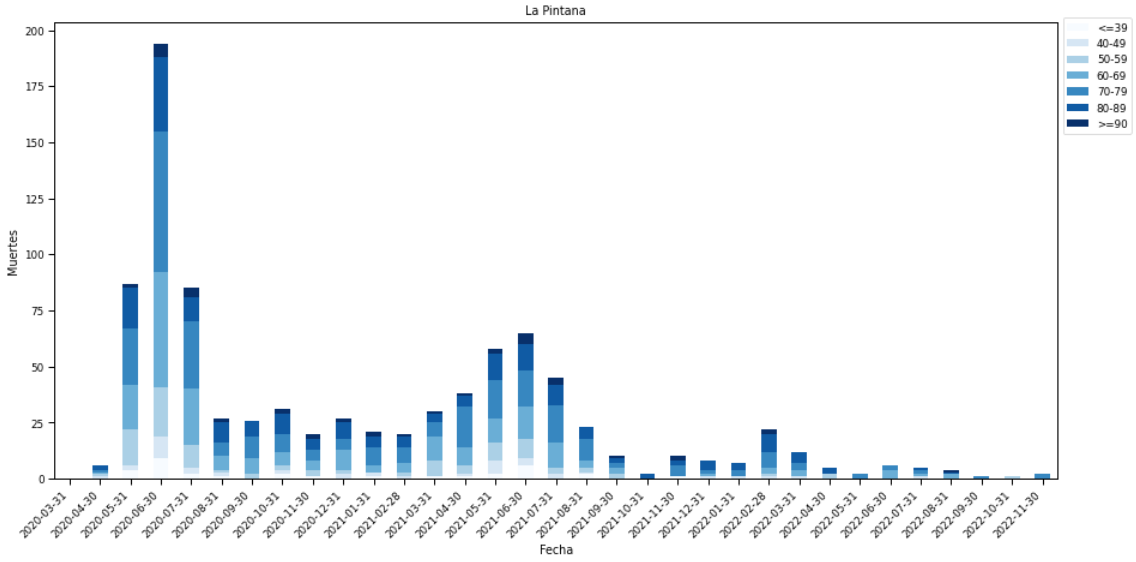


Figura 27 Muertes por edad durante el tiempo comuna de La Pintana

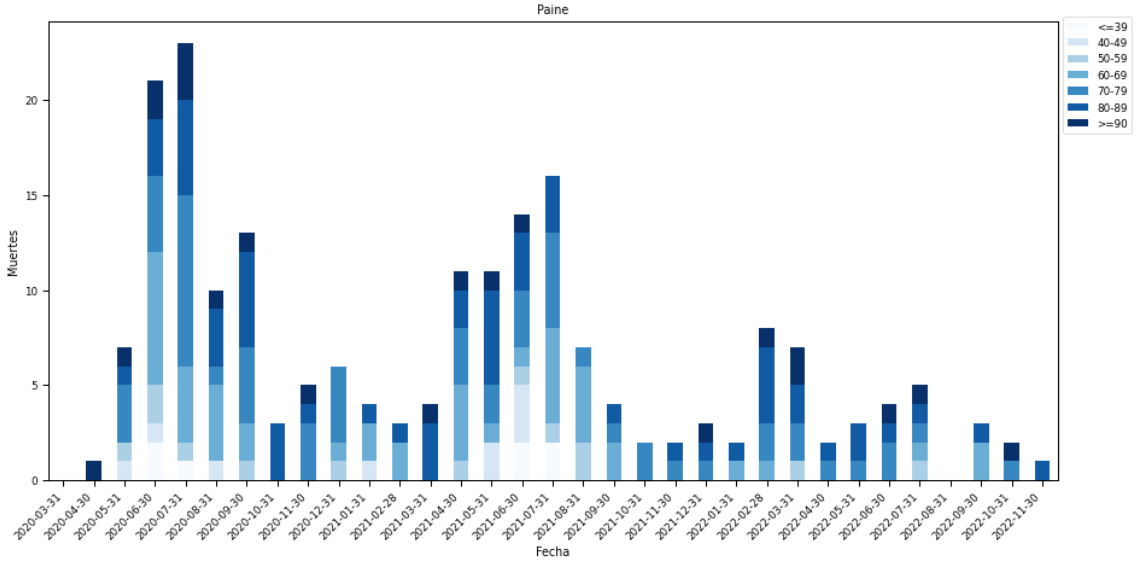


Figura 28 Muertes por edad durante el tiempo comuna de Paine

Se puede ver claramente que las muertes se concentran en los mayores de edad para la mayoría de las comunas, además se observa una tendencia a la baja sobre la cantidad de muertes desde findes del año 2021.

5. Resultados

5.1. Agrupamiento de Comunas

Para el agrupamiento de comunas se consideran principalmente dos fuentes de datos, la serie de tiempo de contagios y una caracterización generada para estos propósitos. En el caso del agrupamiento por contagio, se decidió elegir esta metodología de acuerdo con lo observado en el análisis exploratorio, pues se observan diferencias entre comunas, que en primera instancia generan un indicio de agrupamiento.

En el caso de la caracterización, se ocupar los siguientes parámetros para realizar la clasificación:

- Ingreso per cápita por comuna
- Personas hogar promedio por comuna
- Inmigrantes
- Total de hogares hacinados
- Índice de precariedad de la vivienda
- Muertes
- Vacunados
- Contagios acumulados

Distancia Euclidiana

Caracterización mediante la distancia Euclidiana y el número de contagios por comuna.

Para la caracterización mediante la distancia euclidiana, en primera instancia se calculó la distancia del número de contagiados por comuna, como se puede observar en el siguiente gráfico:

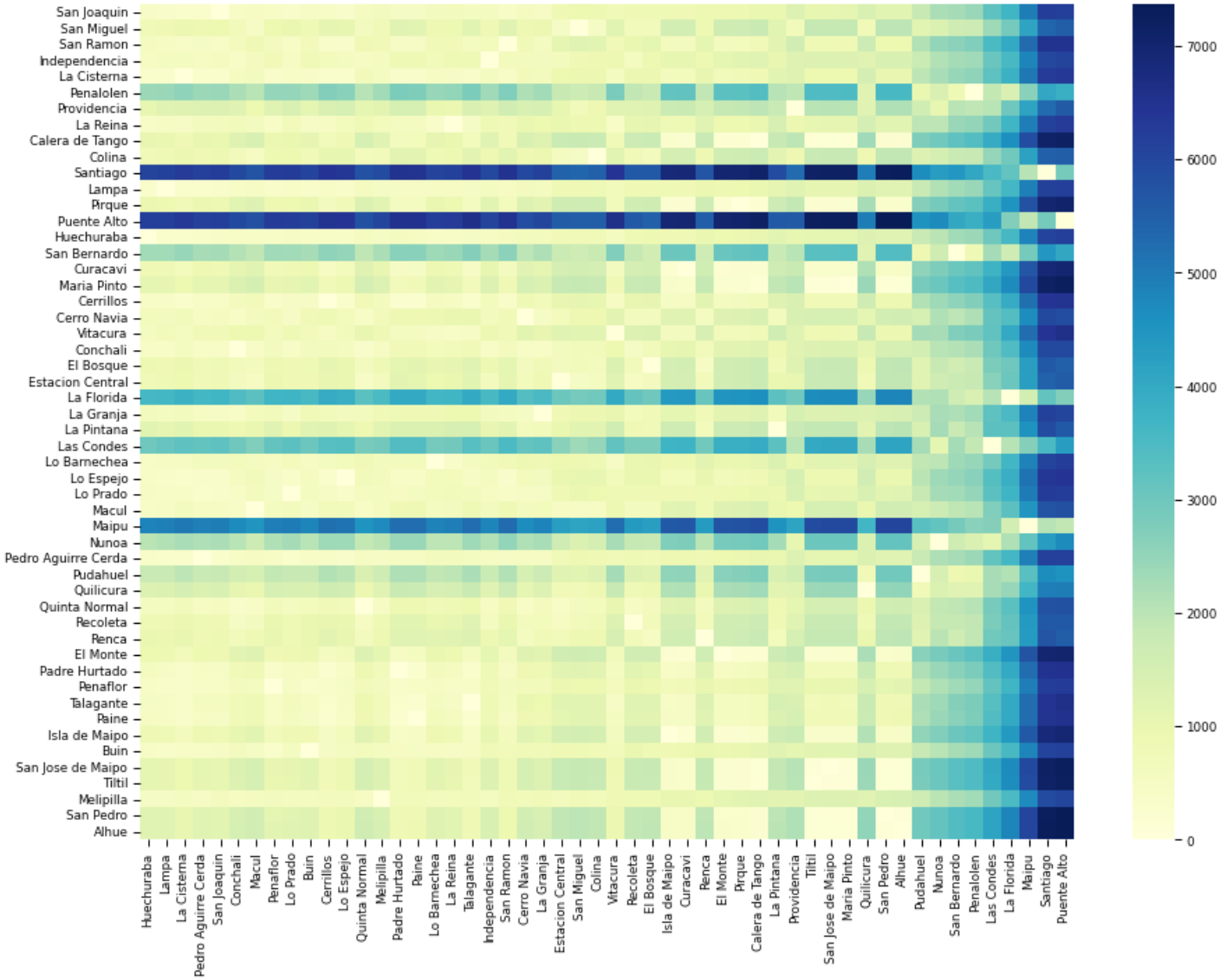


Figura 29 Mapa de calor de distancia euclidiana de contagios entre comunas

A modo de poder visualizar de mejor manera la información, se utiliza el promedio de distancia de cada comuna, como se puede ver en la siguiente tabla:

Comuna	Distancia Promedio a Otras Comunas	Comuna	Distancia Promedio a Otras Comunas
Puente Alto	5765	Colina	1401
Santiago	5729	San Miguel	1395
Maipu	4529	Estacion Central	1378
La Florida	3437	La Granja	1338
Las Condes	3061	Cerro Navia	1327
Penalolen	2433	San Ramon	1324
San Bernardo	2363	Independencia	1310
Nunoa	2242	Talagante	1308
Pudahuel	1999	La Reina	1303
Alhue	1780	Lo Barnechea	1301
San Pedro	1769	Paine	1298
Quilicura	1694	Padre Hurtado	1295
Maria Pinto	1687	Melipilla	1282
San Jose de Maipo	1684	Quinta Normal	1274
Tiltil	1670	Lo Espejo	1260
Providencia	1669	Cerrillos	1250
La Pintana	1628	Buin	1249
Calera de Tango	1620	Lo Prado	1244
Pirque	1555	Penaflor	1241
El Monte	1527	Macul	1239
Renca	1488	Conchali	1234
Curacavi	1487	San Joaquin	1223
Isla de Maipo	1476	Pedro Aguirre Cerda	1214
El Bosque	1415	La Cisterna	1212
Recoleta	1414	Lampa	1206
Vitacura	1406	Huechuraba	1203

Por otro lado, podemos observar lo mismo gráficamente.

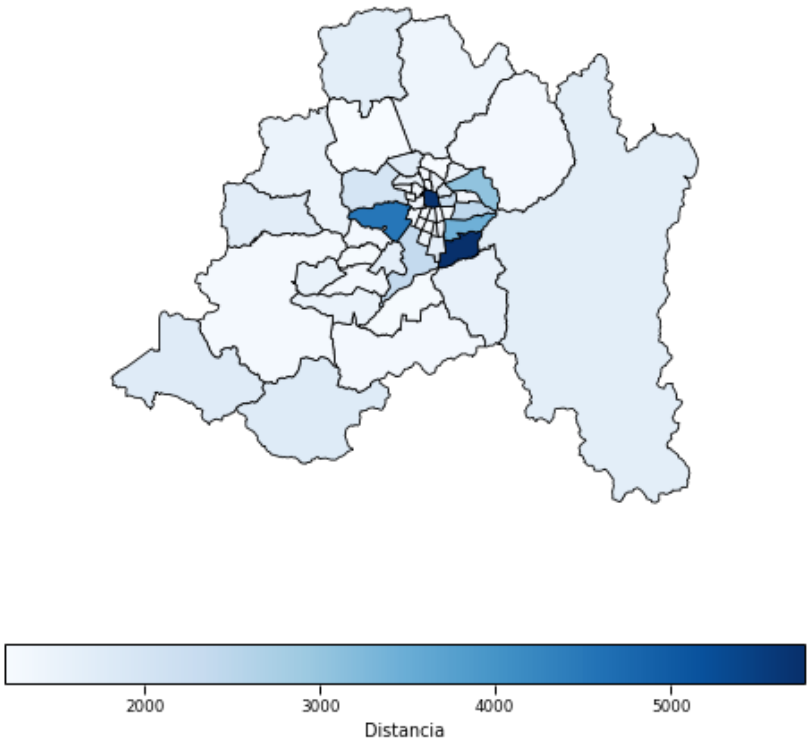


Figura 30 Distancia promedio Euclidiana por contagios

Caracterización mediante la distancia Euclidiana y el número de contagios por comuna por 10,000 habitantes.

Buscando otras relaciones, se realizó la ponderación por cada 10,000 habitantes, esto se puede ver en el siguiente mapa de calor:

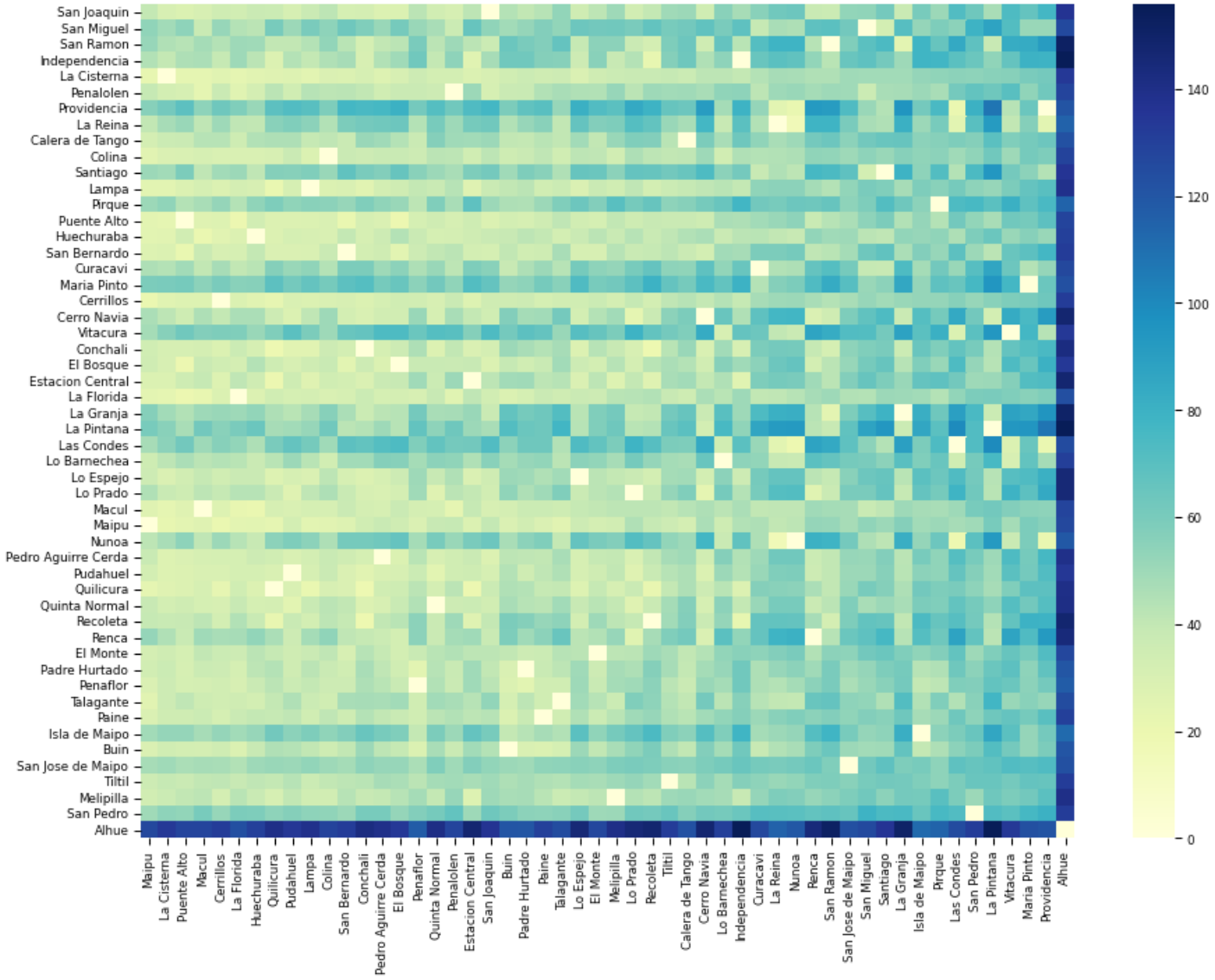


Figura 31 Mapa de calor de distancia euclidiana de contagios entre comunas cada 10,000 habitantes

Revisando los promedios de las distancias:

Comuna	Distancia Promedio a Otras Comunas por 10,000 habitantes	Comuna	Distancia Promedio a Otras Comunas por 10,000 habitantes
Alhue	130	El Monte	47
Providencia	66	Lo Espejo	47
Maria Pinto	65	Paine	47
Vitacura	64	Talagante	47
La Pintana	64	Padre Hurtado	47
San Pedro	62	Buin	46
Las Condes	61	San Joaquin	46
Pirque	59	Estacion Central	46
Isla de Maipo	58	Penalolen	46
La Granja	58	Quinta Normal	46
Santiago	58	Penaflor	45
San Miguel	57	El Bosque	45
San Jose de Maipo	57	Pedro Aguirre Cerda	45
San Ramon	55	Conchali	44
Renca	55	San Bernardo	44
Nunoa	55	Colina	43
La Reina	55	Lampa	43
Curacavi	54	Pudahuel	42
Independencia	54	Quilicura	42
Lo Barnechea	52	Huechuraba	41
Cerro Navia	51	Cerrillos	40
Calera de Tango	51	La Florida	40
Tiltil	51	Macul	40
Recoleta	49	Puente Alto	40
Lo Prado	49	La Cisterna	40
Melipilla	48	Maipu	40

Gráficamente esto se puede ver representando en el siguiente mapa:

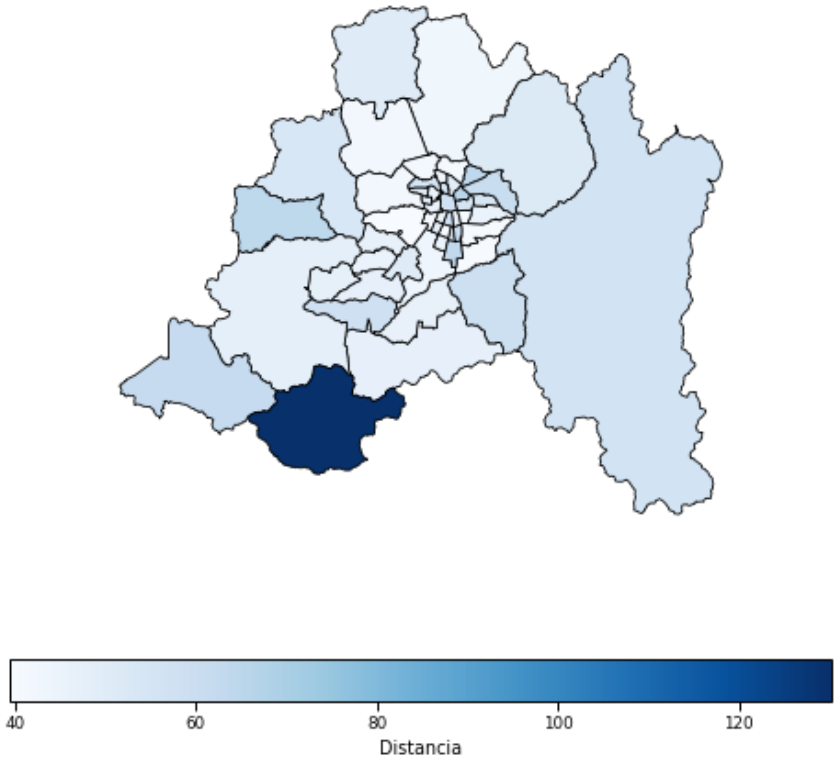


Figura 32 Distancia euclidiana promedio por contagios cada 10,000 habitantes

Dynamic Time Warping (DTW)

El algoritmo DTW permite realizar un alineamiento óptimo entre dos secuencias de vectores de distinta longitud mediante programación dinámica. De dicho alineamiento se obtiene una medida de distancia entre los dos patrones temporales (Giorgino, 2009).

Agrupación mediante contagios

Para los efectos de agrupar mediante la serie de tiempo, se calculó la distancia a cada comuna, como se muestra en el siguiente mapa de calor:

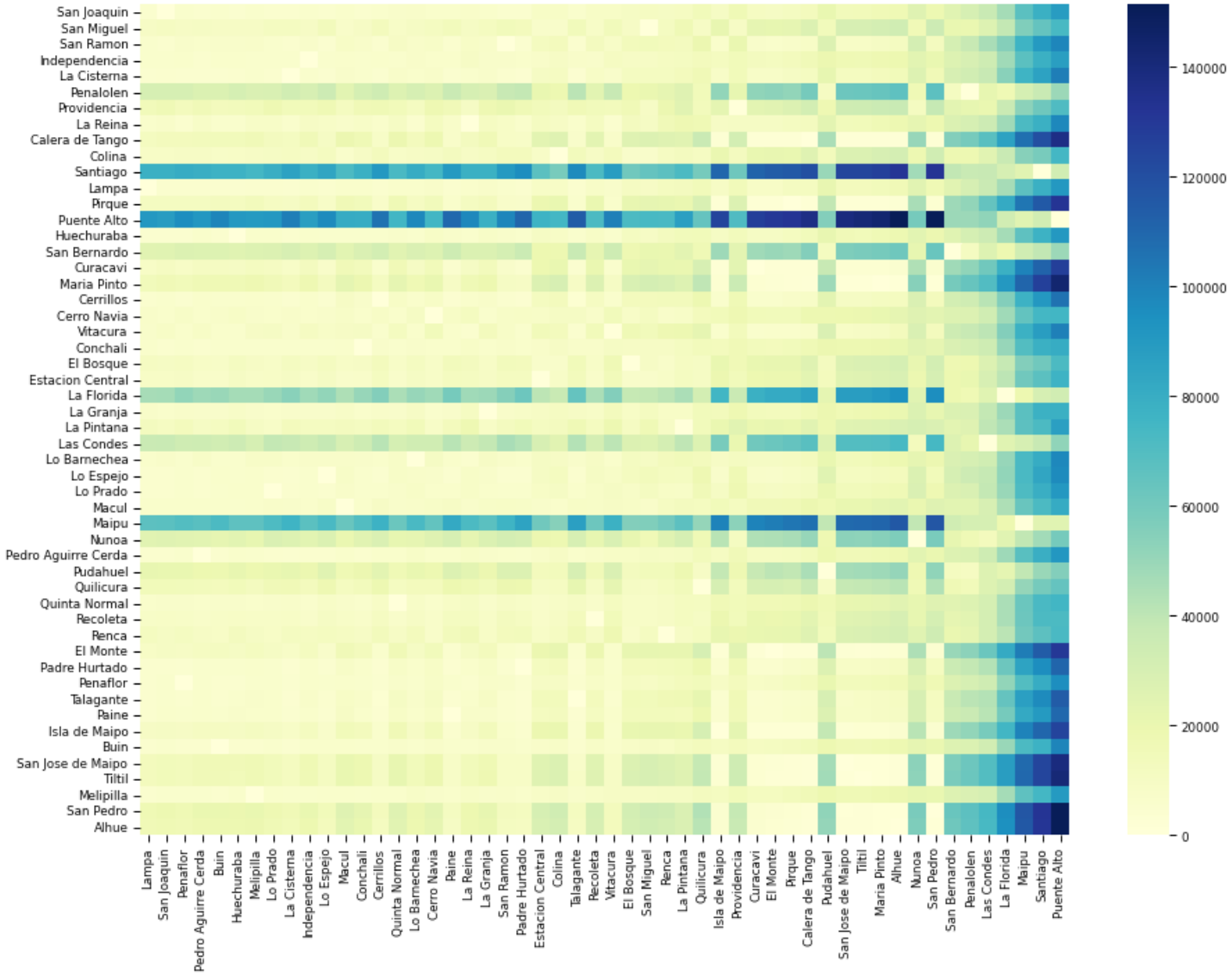


Figura 33 Mapa de calor de distancia DTW de contagios entre comunas

A efectos de simplificar la visualización, se pueden ver los promedios de distancia de cada comuna en la siguiente tabla:

Comuna	Distancia Promedio	Comuna	Distancia Promedio
Puente Alto	89424	Recoleta	17910
Santiago	79171	Colina	17800
Maipu	68370	Talagante	17752
La Florida	53030	Estacion Central	17583
Las Condes	39584	Padre Hurtado	17215
Penalolen	33800	San Ramon	17196
San Bernardo	31734	La Granja	17033
San Pedro	30485	Paine	16978
Nunoa	30416	La Reina	16967
Alhue	29469	Cerro Navia	16860
Maria Pinto	27774	Lo Barnechea	16754
Tiltil	26687	Quinta Normal	16656
San Jose de Maipo	26492	Cerrillos	16603
Pudahuel	26021	Conchali	16419
Calera de Tango	25227	Macul	16412
Pirque	22810	Lo Espejo	16202
El Monte	22160	Independencia	16132
Curacavi	21923	La Cisterna	16072
Providencia	21850	Lo Prado	15992
Isla de Maipo	21715	Melipilla	15983
Quilicura	21704	Huechuraba	15869
La Pintana	20842	Buin	15785
Renca	19320	Pedro Aguirre Cerda	15727
San Miguel	18594	San Joaquin	15706
El Bosque	18094	Penaflor	15697
Vitacura	18080	Lampa	15406

Lo que visualmente se observa como:

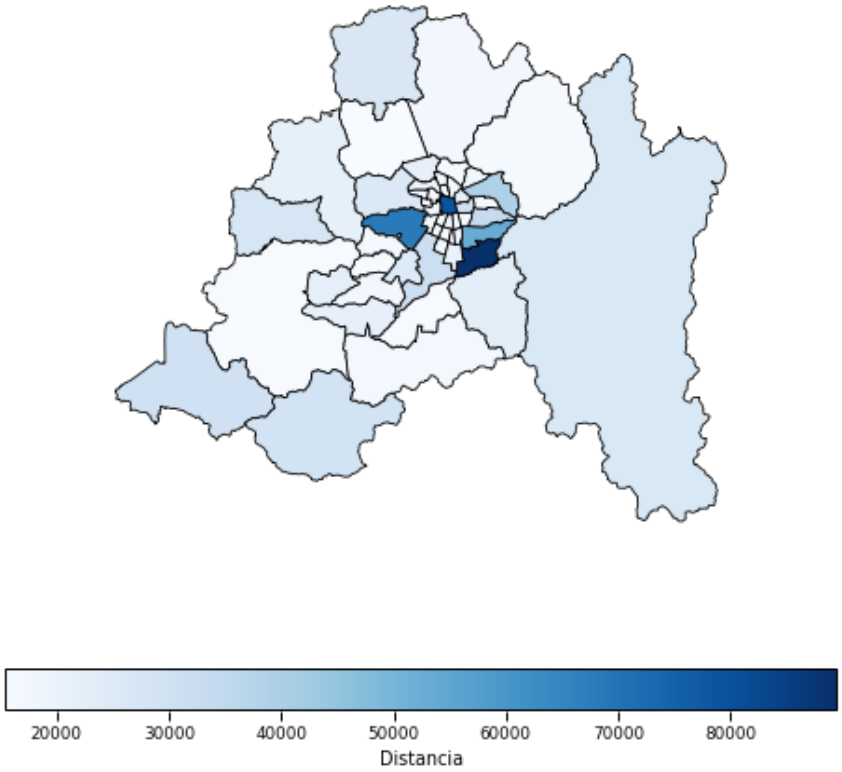


Figura 34Mapa de distancia DTW de contagios entre comunas

Agrupación mediante contagios cada 10,000 habitantes

Buscando otras relaciones, se realizó la ponderación de los contagios por cada 10,000 habitantes, esto se puede ver en el siguiente mapa de calor:

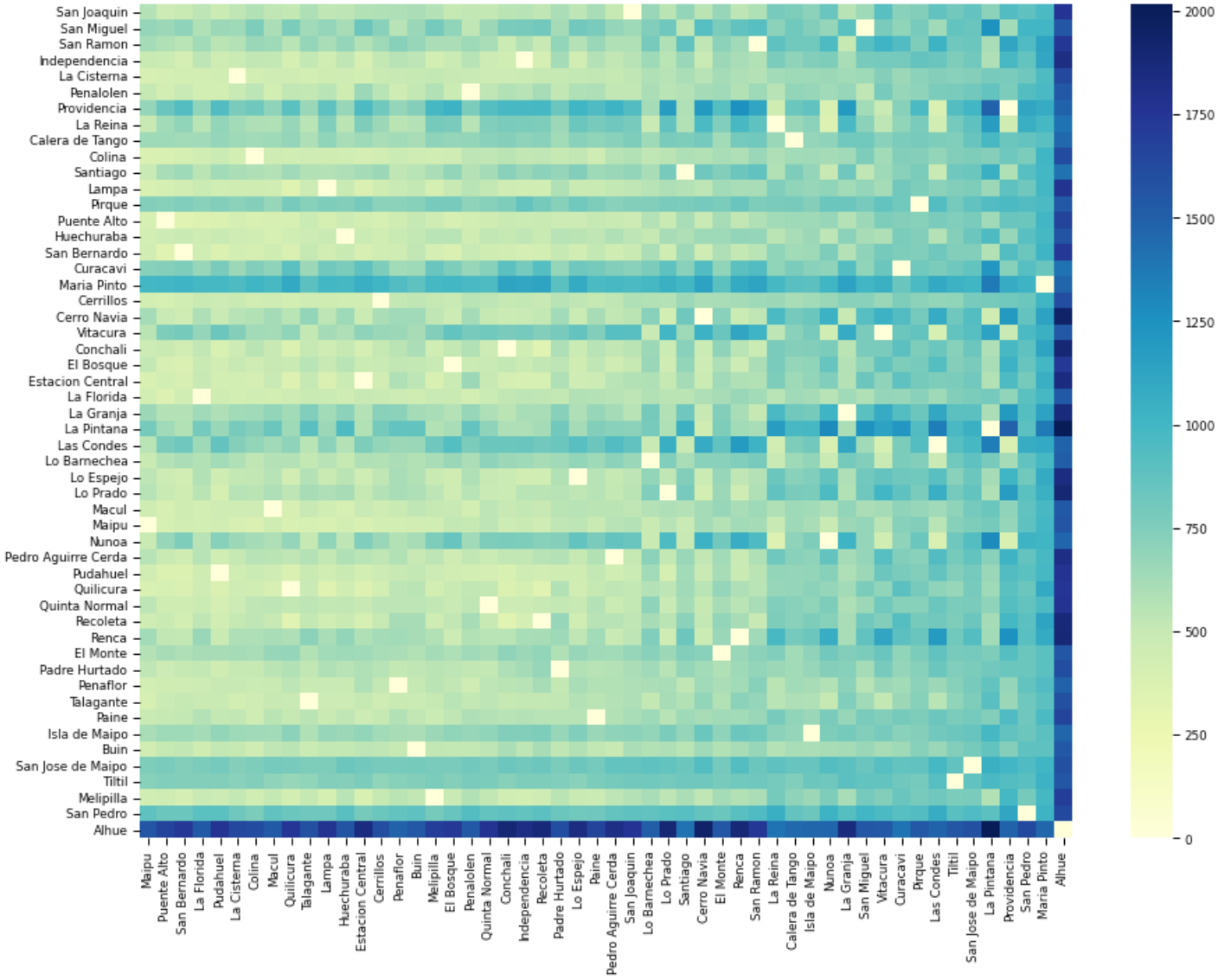


Figura 35 Mapa de calor de distancia DTW de contagios entre comunas cada 10,000 habitantes

A efectos de tener una mejor visualización, se utiliza el promedio de las distancias por comuna, como se muestra a continuación:

Comuna	Distancia	Comuna	Distancia
Alhue	1621	Paine	637
Maria Pinto	995	Lo Espejo	635
San Pedro	894	Padre Hurtado	632
Providencia	875	Recoleta	624
La Pintana	848	Independencia	619
San Jose de Maipo	840	Conchali	616
Las Condes	793	Quinta Normal	612
Tiltil	790	Penalolen	611
Pirque	785	Melipilla	609
Curacavi	780	El Bosque	607
Vitacura	771	Buin	602
San Miguel	748	Penaflor	597
La Granja	738	Estacion Central	596
Nunoa	736	Cerrillos	593
Isla de Maipo	722	Huechuraba	590
Calera de Tango	717	Lampa	585
La Reina	717	Quilicura	581
San Ramon	715	La Florida	581
Renca	713	Macul	581
El Monte	694	Talagante	580
Cerro Navia	685	Colina	579
Santiago	680	La Cisterna	578
Lo Prado	664	Pudahuel	577
Lo Barnechea	655	San Bernardo	568
San Joaquin	644	Puente Alto	560
Pedro Aguirre Cerda	640	Maipu	555

Gráficamente, se puede ver representada en la siguiente figura:

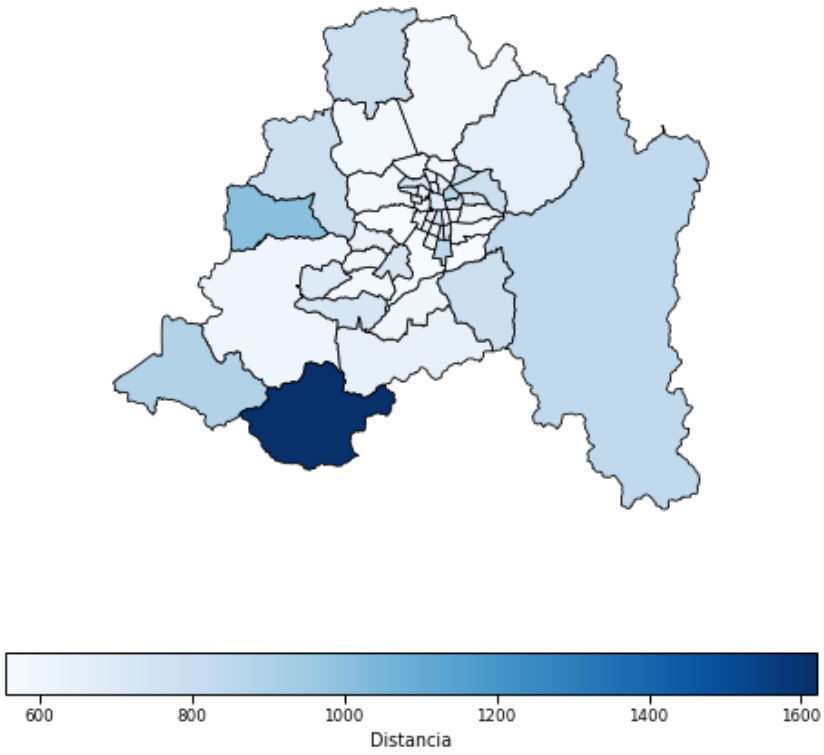


Figura 36Mapa de distancia DTW de contagios entre comunas cada 10,000 habitantes

K-Means

K-means es un algoritmo de clasificación no supervisada (clusterización) que agrupa objetos en k grupos basándose en sus características (unioviedo, 2022).

Para evaluar la calidad de agrupamiento de clustering, se utiliza el coeficiente de Silueta (o Silhoutte) identificando cuál es el número óptimo de agrupamientos (labredes, 2022).

Agrupación mediante el promedio de contagio mensual

En primera instancia se realiza el análisis del coeficiente de Silhouette, para entender cuál sería el número óptimo de clustering, como se muestra en la siguiente tabla:

Análisis del coeficiente de Silhouette

Silhouette analysis para K-Means clustering	
n_clusters	**Mean Silhouette score**
2	0.66
3	0.43
4	0.35
5	0.32
6	0.35
7	0.35
8	0.33
9	0.34
10	0.27

Al momento de generar la clasificación con (n=2) y mediante la utilización de componentes principales, podemos ver el siguiente gráfico de agrupamiento:

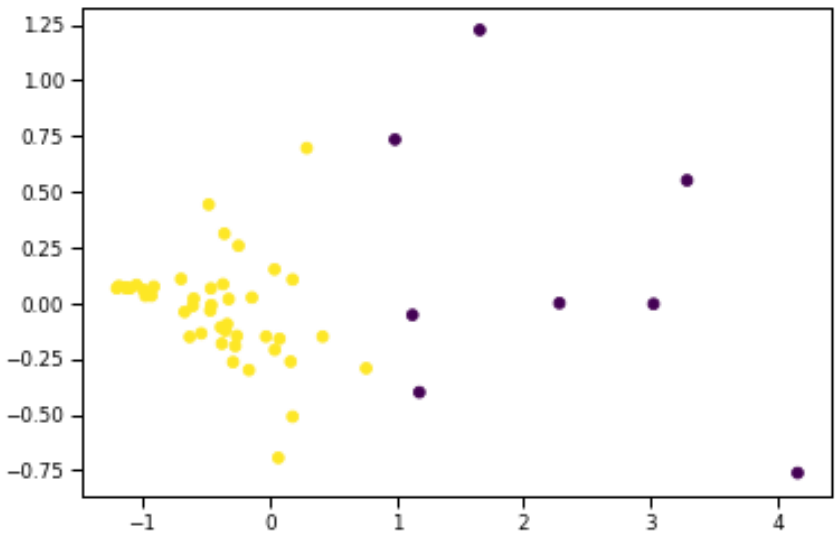


Figura 37 Representación de agrupamiento de comunas mediante componentes principales

Finalmente se puede ver la clasificación para cada una de las comunas mediante este método:

Comuna	Categoría	Comuna	Categoría
San Joaquin	1	Pedro Aguirre Cerda	1
San Miguel	1	Pudahuel	1
San Ramon	1	Quilicura	1
Independencia	1	Quinta Normal	1
La Cisterna	1	Recoleta	1
Providencia	1	Renca	1
La Reina	1	El Monte	1
Calera de Tango	1	Padre Hurtado	1
Colina	1	Penaflor	1
Lampa	1	Talagante	1
Pirque	1	Paine	1
Huechuraba	1	Isla de Maipo	1
Curacavi	1	Buin	1
Maria Pinto	1	San Jose de Maipo	1
Cerrillos	1	Tiltil	1
Cerro Navia	1	Melipilla	1
Vitacura	1	San Pedro	1
Conchali	1	Alhue	1
El Bosque	1	Penalolen	0
Estacion Central	1	Santiago	0
La Granja	1	Puente Alto	0
La Pintana	1	San Bernardo	0
Lo Barnechea	1	La Florida	0
Lo Espejo	1	Las Condes	0
Lo Prado	1	Maipu	0
Macul	1	Nunoa	0

Gráficamente, se puede revisar la asignación a cada comuna:

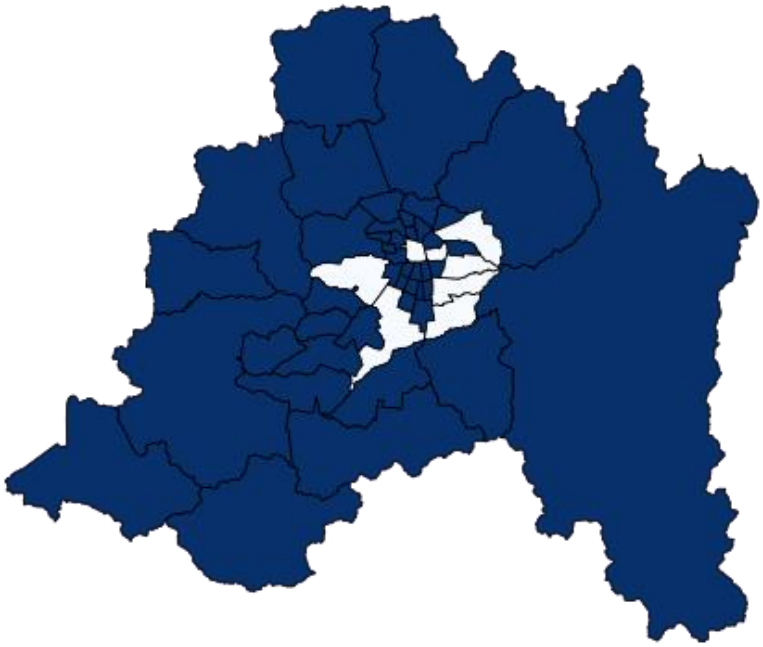


Figura 38 Representación de agrupamiento

Mediante la caracterización de cada comuna

Utilizando la misma formación anterior, podemos revisar que el coeficiente de Silhouette, mantiene el número de clústeres óptimos de 2.

Silhouette analysis para K-Means clustering	
n_clusters	**Mean Silhouette score**
2	0.43
3	0.29
4	0.28
5	0.24
6	0.25
7	0.21
8	0.22
9	0.17
10	0.17

Por medio de componentes principales, podemos revisar la clasificación efectuada:

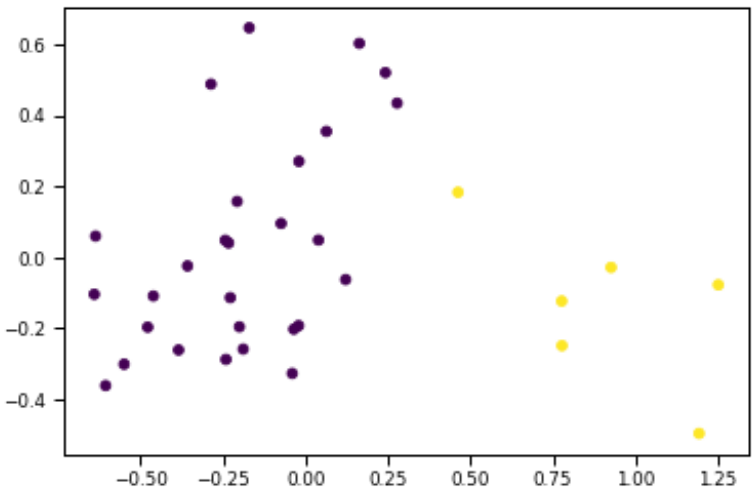


Figura 39 Representación de agrupamiento de comunas mediante componentes principales

Finalmente, la asignación para cada comuna se puede ver en la siguiente tabla:

Comuna	**Categoría**	**Comuna**	**Categoría**
Providencia	1	Conchali	0
La Reina	1	El Bosque	0
Vitacura	1	Estacion Central	0
Las Condes	1	La Florida	0
Lo Barnechea	1	La Granja	0
Nunoa	1	La Pintana	0
San Joaquin	0	Lo Espejo	0
San Miguel	0	Lo Prado	0
San Ramon	0	Macul	0
Independencia	0	Maipu	0
La Cisterna	0	Pedro Aguirre Cerda	0
Santiago	0	Pudahuel	0
Puente Alto	0	Quilicura	0
Huechuraba	0	Quinta Normal	0
San Bernardo	0	Recoleta	0
Cerrillos	0	Renca	0
Cerro Navia	0	Penaflor	0

Gráficamente, se puede observar lo siguiente:

Figura 40 Representación gráfica del agrupamiento

DBScan

Esta técnica de Clustering espacial basada en la densidad de aplicaciones con ruido. Encuentra muestras de núcleo de alta densidad y expande grupos a partir de ellas. Esta técnica es buena para datos que contienen grupos de densidad similar (scikit-learn, 2022).

Agrupación mediante el promedio de contagio mensual

En este caso, se lograron encontrar 3 grupos de comunas, como se muestra en la siguiente tabla:

Comuna	Categoría	Comuna	Categoría
Providencia	1	Lo Prado	0
La Reina	1	Macul	0
Vitacura	1	Maipu	0
Las Condes	1	Pedro Aguirre Cerda	0
Nunoa	1	Pudahuel	0

San Joaquin	0	Quilicura	0
San Ramon	0	Quinta Normal	0
Independencia	0	Recoleta	0
La Cisterna	0	Renca	0
Penalolen	0	El Monte	0
Calera de Tango	0	Penaflor	0
Colina	0	Talagante	0
Santiago	0	Paine	0
Lampa	0	Buin	0
Puente Alto	0	Melipilla	0
Huechuraba	0	San Miguel	-1
San Bernardo	0	Pirque	-1
Cerrillos	0	Curacavi	-1
Cerro Navia	0	Maria Pinto	-1
Conchali	0	Lo Barnechea	-1
El Bosque	0	Padre Hurtado	-1
Estacion Central	0	Isla de Maipo	-1
La Florida	0	San Jose de Maipo	-1
La Granja	0	Tiltil	-1
La Pintana	0	San Pedro	-1
Lo Espejo	0	Alhue	-1

Utilizando componentes principales podemos ver la clasificación:

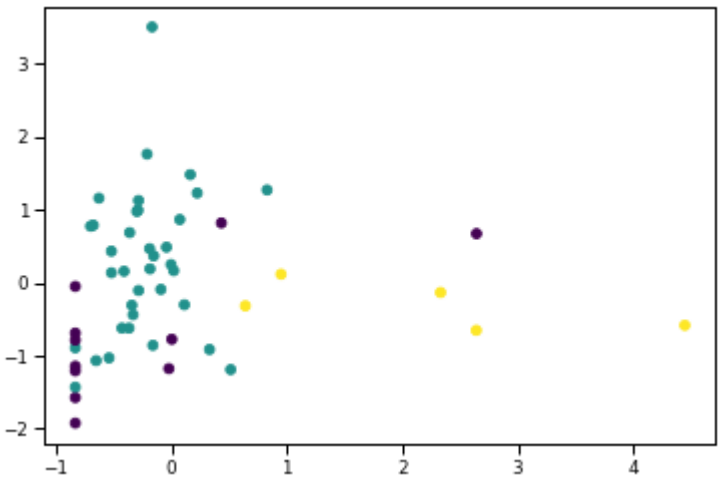


Figura 41 Representación de agrupamiento de comunas mediante componentes principales

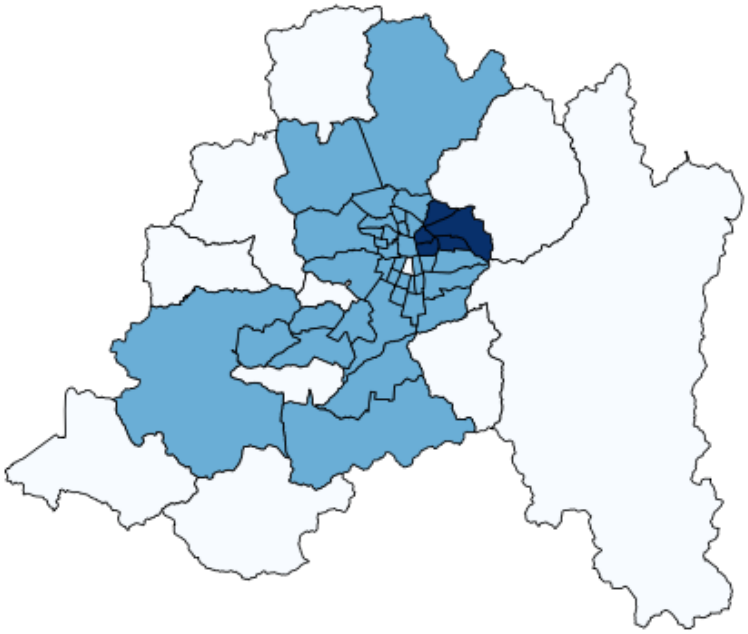


Figura 42 Representación de agrupamiento

Mediante la caracterización de cada comuna

Utilizando los parámetros de caracterización de las comunas, en las que se dispone información, se pudo llegar a obtener la siguiente calificación:

Comuna	Categoria	Comuna	Categoria
Providencia	-1	Cerro Navia	0
La Reina	-1	Conchali	0
Santiago	-1	El Bosque	0
Puente Alto	-1	Estacion Central	0
Vitacura	-1	La Florida	0
Las Condes	-1	La Granja	0
Lo Barnechea	-1	La Pintana	0
Maipu	-1	Lo Espejo	0
Nunoa	-1	Lo Prado	0
San Joaquin	0	Macul	0
San Miguel	0	Pedro Aguirre Cerda	0
San Ramon	0	Pudahuel	0
Independencia	0	Quilicura	0
La Cisterna	0	Quinta Normal	0
Huechuraba	0	Recoleta	0
San Bernardo	0	Renca	0
Cerrillos	0	Penaflor	0

Revisando la representación gráfica:

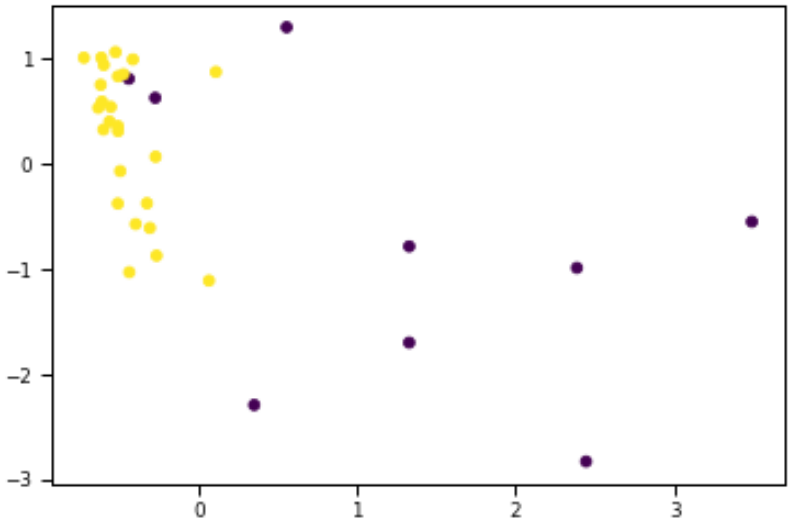


Figura 43 Representación de agrupamiento de comunas mediante componentes principales

Revisando la representación geográfica de la representación podemos ver:

Figura 44 Representación gráfica del agrupamiento

5.2. Pronóstico

Recurrent Neural Network

Red neuronal recurrente (RNN) es un tipo de red neuronal donde la salida del paso anterior se alimenta como entrada al paso actual. La característica principal y más importante de RNN es el estado oculto, que recuerda alguna información sobre una secuencia (stanford, 2022).

Metodología

Para la confección del modelo se utilizó solamente la variable de contagios, se realizó un modelo para cada una de las comunas. El modelo contiene dos capas SimpleRNN y 3 capas densas entrenadas con el optimizador Adam para 100 épocas, para el modelo se utilizó el "promedio móvil del nuevo número de casos". El modelo se entrenó y validó con la intención de pronosticar 10 días en el futuro, por lo tanto, la validación se realizó con este número de días. Este modelo está basado en el trabajo desarrollado por George Saavedra (Saavedra, 2022).

Entrenamiento

Las métricas de entrenamiento pueden ser vistas a continuación:

Métrica	Promedio de valor	Máx. de valor	Mín. de valor
Bias train set	0.14	5.50	-3.97
MAE train set	1.17	5.63	0.10
RMSLE train set	0.05	0.09	0.03

A modo de ejemplo se incluye una imagen del modelo para la comuna de Vitacura, en donde se muestra la gráfica de validación del modelo.

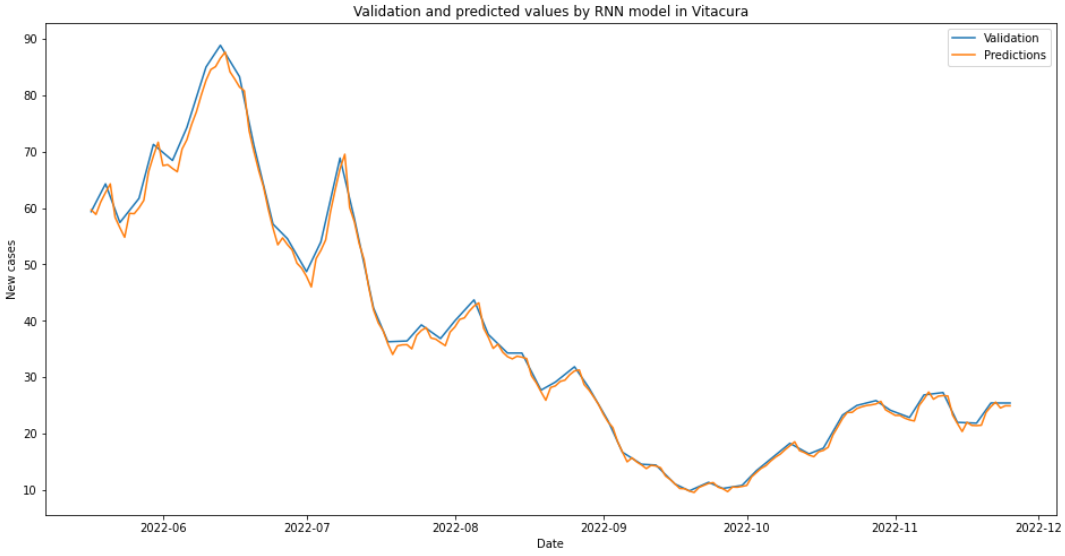


Figura 45 Validación y pronóstico de contagios en la comuna de Vitacura

Pronóstico

Las métricas de pronóstico pueden ser vistas a continuación:

Métrica	Promedio de valor	Máx. de valor	Mín. de valor
Bias validation set	0.07	1.95	-0.97
MAE validation set	1.29	7.78	0.12
RMSLE validation set	0.04	0.08	0.02

Se generaron las proyecciones para cada una de las comunas, a modo de ejemplo se incluye la proyección de contagio para la comuna de Vitacura.

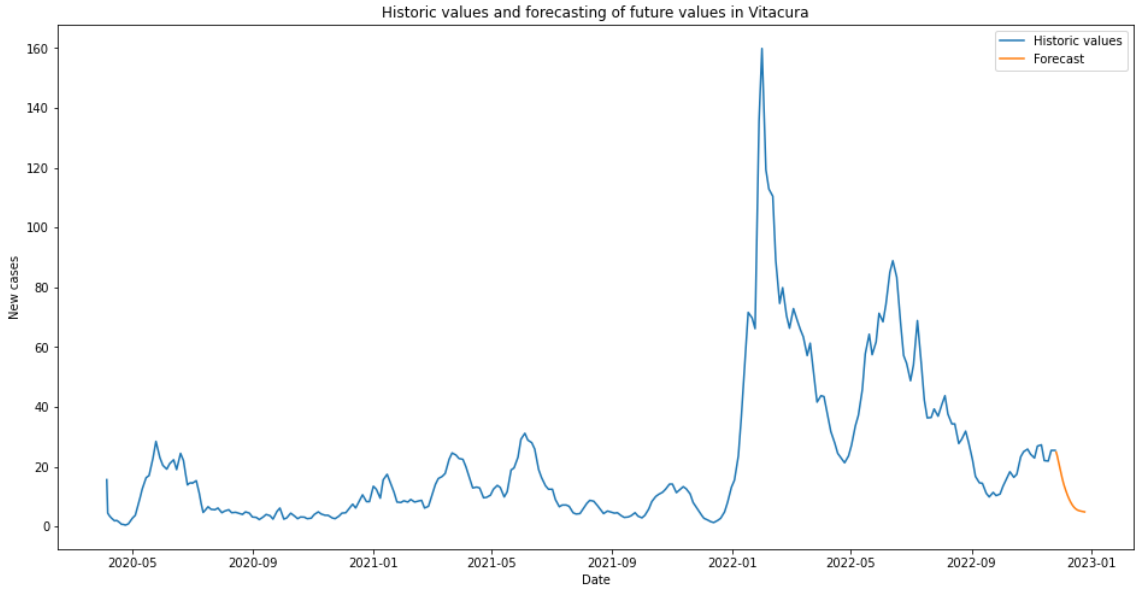


Figura 46 Pronóstico de contagios en Vitacura

El modelo es capaz de predecir con bastante precisión lo que sucederá en los próximos 10 días de la pandemia para cada una de las comunas.

Prophet

Metodología

Para la confección del modelo se utilizó solamente la variable de contagios, se realizó un modelo para cada una de las comunas. El modelo considera feriados y está construido con una estacionalidad semanal. El modelo se entrenó y validó con la intención de pronosticar 10 días en el futuro, por lo tanto, la validación se realizó con este número de días. Prophet incluye funcionalidad para la validación cruzada de series de tiempo para medir el error de pronóstico utilizando datos históricos. Esto se hace seleccionando puntos de corte en el historial y, para cada uno de ellos, ajustando el modelo usando datos solo hasta ese punto de corte. Luego podemos comparar los valores pronosticados con los valores reales (Facebook, 2022).

Entrenamiento

Las métricas de entrenamiento, obtenidas a través de la validación cruzada disponible en el paquete, pueden ser vistas resumidas para cada comuna a continuación:

Variable	min	max	prom
horizon	10 days	10 days	10 days
mse	9.39	15354.40	1320.17
rmse	3.06	123.91	27.58
mae	2.17	107.93	22.37
mape	0.09	0.87	0.40
mdape	0.09	0.76	0.38
smape	0.09	1.31	0.48
coverage	0.00	0.77	0.33

A modo de ejemplo se incluye una imagen del modelo para la comuna de Las Condes, en donde se muestra la gráfica de validación del modelo.

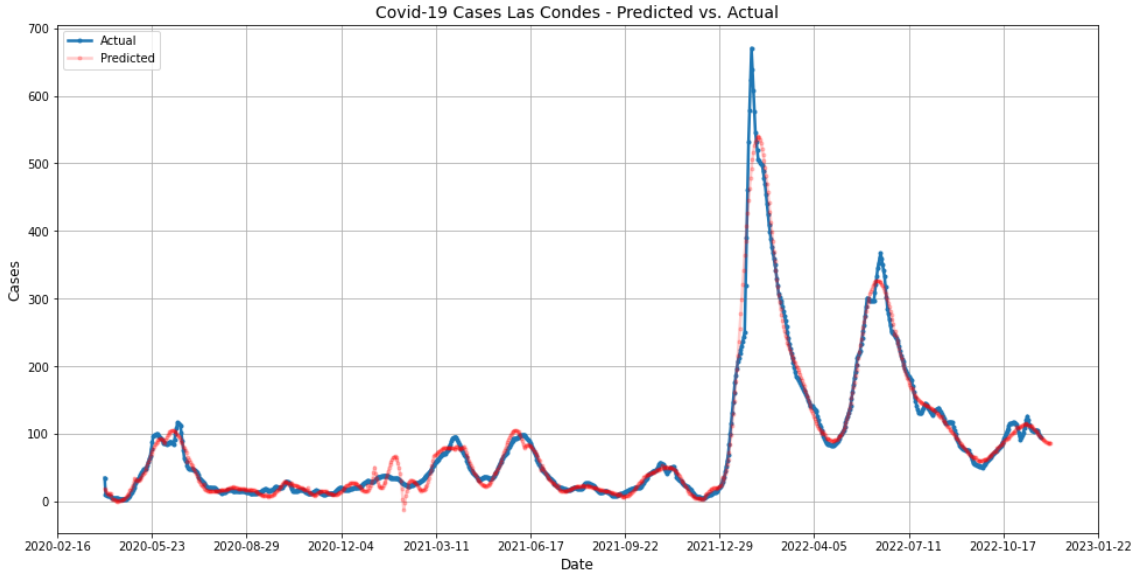


Figura 47 Casos actuales y modelación mediante Prophet

Cómo se puede observar, existen comunas donde el modelo se pudo calibrar bastante bien, sin embargo, para otras comunas el modelo es bastante malo y posiblemente requiera más tiempo de calibración para cada comuna.

Pronóstico

Debido a que el modelo ajusta relativamente bien para algunas comunas, pero para otras no funciona de la mejor manera, se decidió descartar la metodología, por lo complejo de la calibración.

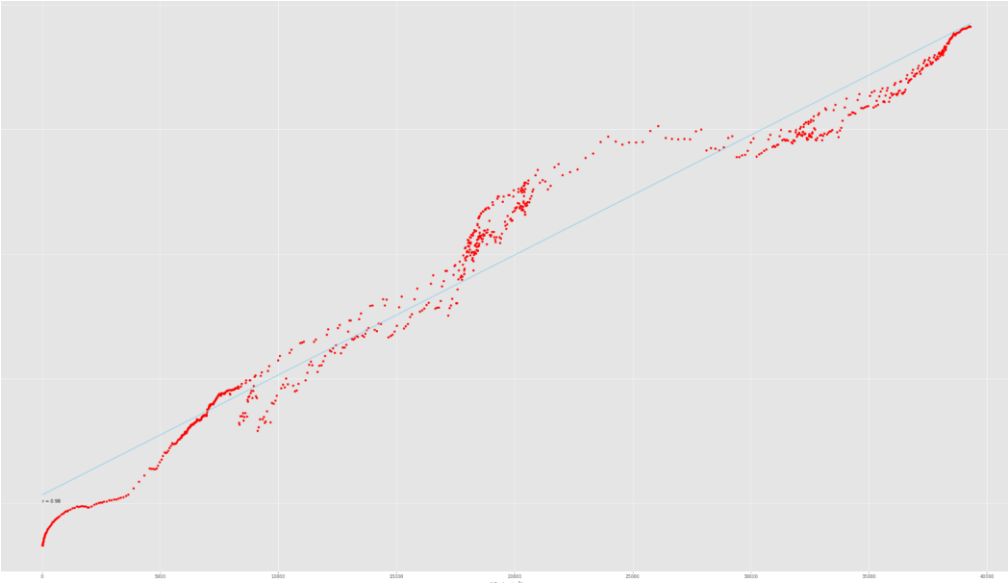
Modelo de Regresión Multilineal Ordinaria

Metodología

Para la confección del modelo se utilizaron las variables casos activos, muertos y vacunados; este modelo se realizó para cada una de las comunas por separado. El modelo considera feriados y está construido con una estacionalidad semanal. El modelo se entrenó y validó con la intención de pronosticar futuros contagios, por lo tanto, la validación se realizó sin número de días como los supuestos de modelos anteriores.

Métrica	Valor
R2	0.963
R2 Ajustado	0.963
R test	0.98

A modo de ejemplo se muestran la salida de Estación Central.



$$\boldsymbol{Contagio}_{\boldsymbol{Estación\ Central}} = 63{,}08 * muertos - 1{,}46 * vacunados + 2.77 * casos\ activos - 3.382{,}08$$

Pronóstico

Esta metodología se ajusta de manera global y razonablemente bien para la mayor parte de las comunas, y para otras funciona lo suficiente. Si bien esta metodología por la bonanza de sus indicadores no debería descartarse, no recoge cambios sutiles que pudieran hacer critico decisiones político-económicas a partir de sus resultados tan globales y poca sensibilidad.

6. Conclusiones

- Es factible realizar pronósticos a nivel comunal como se ha verificado en el presente trabajo.
- Como siguiente etapa se podría dejar disponible al público los análisis generados de forma automática y en tiempo real para ayudar a la toma de decisiones.
- Los modelos generados a partir de supervisión se ajustan bastante bien a la predicción, sin perjuicio a ello, solo el método de redes neuronales RNN recoge de manera razonable la sensibilidad de información.
- A criterio de los autores y por las metodologías revisadas, la mejor agrupación para las comunas de Santiago corresponde a dos grupos, sin embargo, se podrían generar otro tipo de agrupaciones incluyendo otro tipo de datos o inclusive sacrificar performance con otras agrupaciones dependiendo del uso que se le desee dar a la agrupación.
- La metodología que mejor caracteriza la pandemia en la Región Metropolitana es la no supervisada, en particular K-Means se comporta bastante bien, en donde se puede inferir que dada la agrupación de salida, la principal hipótesis de este trabajo es aceptada y respaldada por estos análisis, es decir, la incidencia de contagio si está relacionada con la precariedad de viviendas o extrapolando al estrato económico de la comuna, existiendo mayor cantidad de contagios, inclusive después de la vacunación, respecto a la media nacional.
- En el caso del índice de movilidad, se puede apreciar que es poco efectivo en ciertas comunas y por lo tanto se debe de reconsiderar la aplicación de esta medida, por ejemplo, con solo confinamiento de los escolares o medidas similares. Además, se muestra que no es factible mantenerlo en el tiempo.

- A partir de la declaración del punto anterior, en una próxima emergencia sanitaria, se debería poder establecer políticas de restricciones a la población utilizando el input de clasificación, de tal manera de no contribuir a la expansión desproporcionada de infectados en comunas precarias.
- Hubiera sido interesante caracterizar los primeros días de la pandemia, sin embargo, hasta el momento no fue posible obtener la información, en primera instancia, según lo descrito en el mismo repositorio de github, previo al 15 de abril de 2020 los informes epidemiológicos del Ministerio de Salud no entregaban datos de confirmados notificados en comunas con bajo número de casos, para proteger la identidad de las personas contagiadas.

Bibliografía

Facebook. (1 de 12 de 2022). *facebook*. Obtenido de facebook: https://facebook.github.io/prophet/docs/diagnostics.html

Giorgino, T. (2009). Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. *Journal of Statistical Software*, 1-24.

GONZALO E. MENA, P. P. (2021). Socioeconomic status determines COVID-19 incidence and related mortality in Santiago, Chile. *SCIENCE Vol 372*, Issue 6545.

I.T. Peres, L. B. (2021). Sociodemographic factors associated with COVID-19 in-hospital mortality in Brazil. *Public Health Volume 192*, 15-20.

investopedia. (15 de 11 de 2022). *investopedia*. Obtenido de investopedia: https://www.investopedia.com/terms/a/autocorrelation.asp

José Francisco Vergara-Perucich, J. C.-P.-N. (2020). The Spatial Correlation between the Spread of COVID-19 and Vulnerable Urban Areas in Santiago de Chile. *Critical Housing Analysis Volume 7 | Issue 2*, 21-35.

labredes. (18 de 11 de 2022). *labredes*. Obtenido de labredes: http://www.labredes.unlu.edu.ar/sites/www.labredes.unlu.edu.ar/files/site/data/bdm/coeficiente-silueta.pdf

Monita Karmakar, P. M. (2021). Association of Social and Demographic Factors With COVID-19 Incidence and Death Rates in the US. *JAMA Netw Open*.

Robert Dales, C. B.-V.-M. (2021). The association between air pollution and COVID-19 related mortality in Santiago, Chile: A daily time series analysis. *Environmental Research Volume 198*, 111284.

Saavedra, G. (11 de 11 de 2022). *kaggle*. Obtenido de kaggle: https://www.kaggle.com/code/georgesaavedra/best-covid-19-forecasting-in-us-uk-and-chile

scikit-learn. (20 de 10 de 2022). *scikit-learn*. Obtenido de scikit-learn: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

stanford. (11 de 12 de 2022). *stanford.edu*. Obtenido de stanford.edu: https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks

statsmodels. (11 de 10 de 2022). *statsmodels*. Obtenido de statsmodels: https://www.statsmodels.org/stable/index.html

unioviedo. (15 de 11 de 2022). *unioviedo*. Obtenido de unioviedo: https://www.unioviedo.es/compnum/laboratorios_py/kmeans/kmeans.html

World Health Organization. (25 de 9 de 2022). *WHO*. Obtenido de https://covid19.who.int/

Y. Yao, J. P. (2020). Temporal association between particulate matter pollution and case fatality rate of COVID-19. *Environmental Research Volume 189,* 109941.

ANEXO