



MOVILIDAD POBLACIONAL Y CONTAGIOS EN COLOMBIA DURANTE COVID-19

Un informe de la Universidad del Rosario

ANEXO METODOLÓGICO: MOVILIDAD POBLACIONAL Y CONTAGIOS EN COLOMBIA DURANTE COVID-19

Un informe de la Universidad del Rosario

Autores Principales

Felipe González-Casabianca
Universidad del Rosario

Alejandro Feged-Rivadeneira
Universidad del Rosario
alejandro.feged@urosario.edu.co

Colaboradores

Andrés Ángel
Universidad de Los Andes
ja.angel908@uniandes.edu.co

Vladimir Corredor
Universidad Nacional
vcorredore@unal.edu.co

María Carolina Salazar Pardo
Universidad del Rosario

Andrea Parra
Universidad del Rosario

Andrés Miguel Sampayo
Universidad del Rosario
andres.sampayo@urosario.edu.co

18 de julio de 2020

1. Introducción

Este informe hace parte de una serie de documentos elaborados para aportar en la toma de decisiones en salud pública relacionada con la epidemia de COVID-19 en Colombia. En este documento se presentan cambios en patrones de movilidad, a la fecha y acumulados, a partir de datos agregados. Los datos utilizados provienen del servicio GeoInsights, de Facebook.

Con base en los datos analizados, se busca identificar patrones de movilidad y establecer magnitudes de flujo de población entre y al interior de unidades funcionales definidas administrativamente o a través de esta herramienta. Para ello se utilizó teoría de redes sociales, y se calcularon las comunidades con base en los flujos observados. Este trabajo está inspirado en la idea de megaregiones[?], en la cual se utilizan viajes diarios para identificar grupos de municipios que componen unidades funcionales y/o administrativas. Otros ejemplos de trabajo teórico y empírico han abordado el tema de megaregiones para distintos contextos

2. Limitaciones

Este informe tiene como objetivo ser parte de un ecosistema de diagnósticos de la situación. El análisis presentado debe ser comparado con los resultados de diferentes métodos, el objetivo no es remplazarlos. La precisión de los datos de reportes diarios, por el cuello de botella que supone el diagnóstico, hace que los resultados estén sujetos al sesgo que muestran los datos. Debido a los datos a los que se tiene acceso, la geolocalización se hace de forma automática, sacrificando precisión por eficiencia. En Colombia, la epidemia apenas está en sus primeras etapas, hay pocos datos y las series de tiempo son relativamente cortas. A medida que aumente la información o se extienda a otros países, la

precisión del análisis debe aumentar. En la medida en que ocurran diferentes cambios en la epidemiología, también el análisis se volverá más robusto. Es de esperar que tenga errores, pero es una herramienta más que puede brindar información en tiempo real.

Los datos de cambio de movilidad entre y al interior de unidades espaciales, aparte de ser un indicador de la eficacia de las medidas de distanciamiento social, permiten establecer cual es su efecto en la progresión de la infección y su posible carga para el sistema de salud a nivel regional si están acoplados a datos epidemiológicos mínimos disponibles y confiables (e.g. número de casos confirmados, número de individuos expuestos por unidad espacial, número de fallecidos por causa de la infección, número de camas ocupadas de los servicios de hospitalización general, unidades de cuidado intermedio (UCIM) y unidades de cuidado intensivo (UCI) en Instituciones Prestadoras de Servicios de Salud – IPS de naturaleza privada y pública frente al total disponible y habilitado). Por ejemplo, los datos de casos confirmados dependen del número de pruebas diarias realizadas por unidad espacial, del momento en que se realiza la prueba y del tiempo que tardan los laboratorios de diagnóstico en proporcionar el resultado. Estos datos no necesariamente están disponibles.

3. Definiciones

Reducción significativa en movilidad: Valor-z debajo de 2 desviaciones estándar del observado para las semanas anteriores (similar a un canal endémico).

Megaregión, comunidad, unidad funcional: La modularidad es definida como una función del total de conexiones del subconjunto de una red entre nodos de la misma comunidad (y no conexiones con nodos externos a una comunidad)[?]. Dado que no puede ser evaluada como la simple fracción de aristas (pues derivaría en una comunidad compuesta por todos los nodos), se evalúa en comparación con redes aleatorias[?].

4. Métodos

4.1. Construcción de Unidades Espaciales de COVID-19 en Colombia

Con base en el algoritmo de clustering *walktrap*[?] (cuyo objetivo es identificar grupos de nodos cuyas conexiones entre sí sean mucho mayores que las conexiones con otros nodos), implementado con los datos de movilidad donde los municipios son los nodos y las aristas tienen como peso el número de viajes que se hacen entre estos, se identifican los grupos o unidades espaciales. Estas unidades son de vital importancia en el contexto de cuarentena flexible, pues son las unidades desde las cuales hay más riesgo de contagio entre sí.

Estas unidades se pueden pensar como las unidades regionales funcionales, de manera que los esfuerzos de control deben obedecer a estas dinámicas y no a las divisiones administrativas, pues en ocasiones municipios diferentes pueden hacer parte del mismo "grupo metropolitano", como Bogotá y Soacha, por mencionar solo un ejemplo.

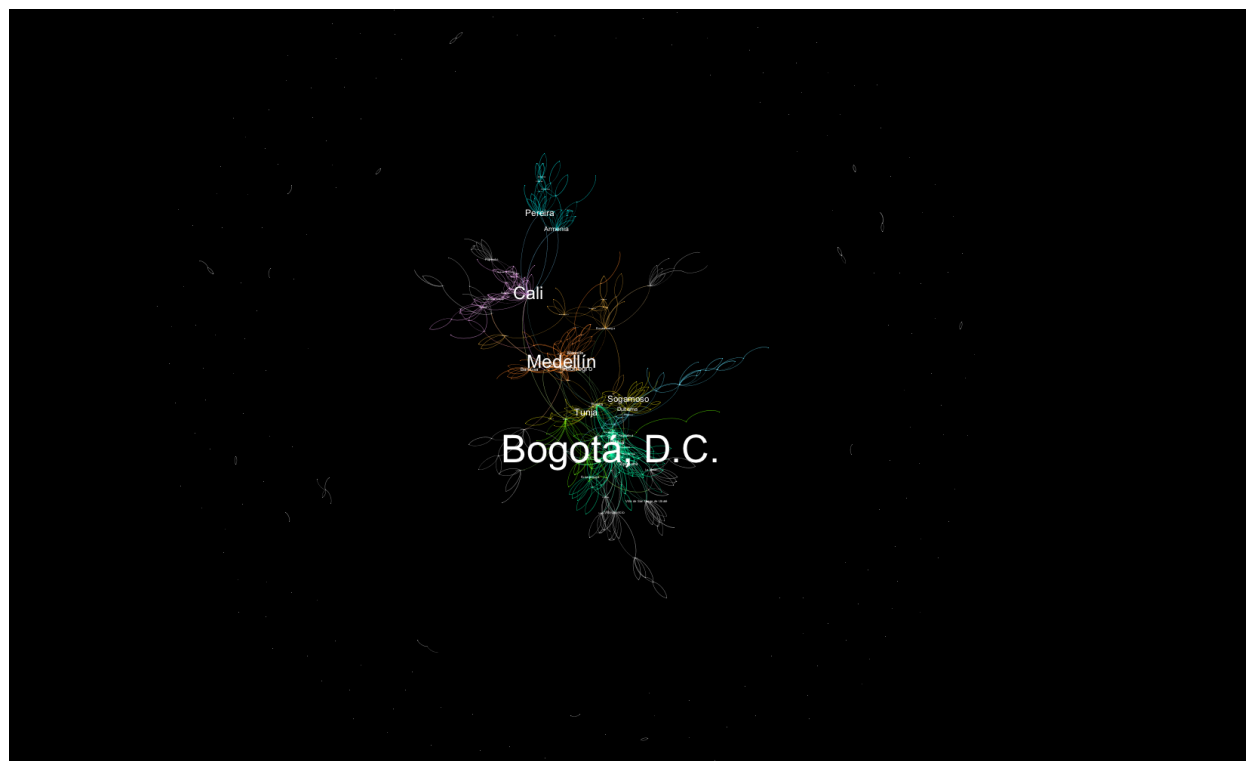


Figura 1: **Unidades espaciales de COVID-19 en Colombia.** Con base en los datos, se definen las regiones de control, identificadas con diferentes colores en esta gráfica. Las unidades espaciales están pensadas de tal manera que son las comunidades regionales observadas.

País	Lugar	Movimiento	Casos	Fecha Inicio Movimientos	# Unidades de Control
Colombia	País Entero	Si	Si	2020-04-02	246
Colombia	Bogotá	Si	Si	2020-04-12	494
USA	Indiana	Si	Si	2020-04-04	91
USA	Illinois	Si	Si	2020-04-05	94
USA	Kentucky	Si	Si	2020-04-02	96
USA	Florida	Si	Si	2020-04-09	64
USA	Vermont	Si	Si	2020-04-06	14
USA	Oregon	Si	Si	2020-04-28	34
Italia	País Entero	Si	Si	2020-03-23	29
Perú	País Entero	Si	Si	2020-03-25	160
Australia	New South Wales	Si	Si	2020-05-03	56
Alemania	País Entero	Si	Si	2020-03-25	223
Brasil	País Entero	Si	Si	2020-03-31	2318
Chile	País Entero	Si	Si	2020-03-23	100

Cuadro 1: Distintos lugares integrados en el esquema de análisis.

La mayoría de estos repositorios ya tienen los casos geocodificados, pero existen algunos que únicamente tienen asociado a cada caso el lugar específico (municipio, región, ciudad) en el que ocurrió, como sucede para Colombia, Alemania y Australia. Para estos lugares se hace una geocodificación automática, asignando las coordenadas geográficas del lugar que permitirá cruzar esta información con la información de movilidad de Facebook GeoInsights.

La extracción de estos conjuntos y construcción de los datos unificados se hace a diario, descargando automáticamente los datos de movilidad de Facebook, los reportes de casos confirmados de los repositorios públicos de cada lugar y geocodificando los sitios de ocurrencia para los lugares que no se encuentran precomputados. A medida que se encuentran disponibles más lugares, estos se integrarán en el esquema de extracción.

5.1. Modelos de Proyección

Estos modelos permiten proyectar los casos de diferentes unidades hasta una semana en el futuro, utilizando el movimiento interno, movimiento externo de unidades de control adyacentes y el número de casos confirmados por unidad. A continuación se presentan los detalles en la construcción de los datos y el esquema de entrenamiento para una unidad espacial.

5.1.1. Estructura de los Datos

A partir de los datos de movilidad de Facebook GeoInsights y los reportes de casos históricos, se construye el siguiente conjunto para cada día d en la unidad de control i (una unidad de control cualquiera):

- **Número de Casos Interno** (C_i^d): el número de casos confirmados en el día.
- **Movimiento Interno** (M_{ii}^d): la suma de todas las observaciones de movimiento que comienzan y terminan en dentro de la unidad espacial.
- **Número de Casos Externo** (CE_i^d): el número de casos confirmados en el día para el resto de unidades espaciales.

$$CE_i^d = \sum_{j \neq i} C_j^d$$

- **Movimiento Externo** (ME_i^d): el movimiento del resto de unidades espaciales del país hacia la unidad en cuestión.

$$ME_i^d = \sum_{j \neq i} M_{ij}^d$$

- **Día de la Epidemia** (δ_i^d) el número actual de la epidemia, tomando como el día 0 la fecha en que se presentó el primer caso en la unidad espacial analizada. Por ejemplo, si se tiene una unidad espacial A y B con primer caso reportado: 27 de Marzo y el 2 de Abril respectivamente, si d es el día 3 de Abril entonces: $\delta_A^d = 7$ y $\delta_B^d = 1$.

A partir de este conjunto de datos diarios, se construyen dos valores adicionales para el modelo de proyección. Estos valores también se encuentran asociados a la unidad espacial en cuestión para cada uno de los días de la epidemia:

- **Flujo Interno** (FI_i^d): La multiplicación de la cantidad de casos internos en el día por el movimiento interno de la unidad. Esta variable pretende capturar la “fuerza” interna de la epidemia en la unidad de control para un día en particular.

$$FI_i^d = C_i^d M_{ii}^d$$

- **Flujo Externo** (FE_i^d): La suma de la multiplicación de la cantidad de casos y el movimiento hacia la unidad de control en cuestión, del el resto de unidades del país. Esta variable pretende capturar la “fuerza” entrante de la epidemia hacia la unidad de control para un día en particular.

$$FE_i^d = \sum_{j \neq i} C_j^d M_{ji}^d$$

El modelo implementado utilizará estas variables para los anteriores 5 días, para encontrar el número de casos 8 días en el futuro.

$$X_i^d = \{C_i^d, M_{ii}^d, CE_i^d, ME_i^d, FI_i^d, FE_i^d, \delta_i^d\}$$

$$(X_i^d, X_i^{d-1}, \dots, X_i^{d-4}) \mapsto C_i^{d+8}$$

Los valores de 5 y 8 fueron ajustados según los datos.

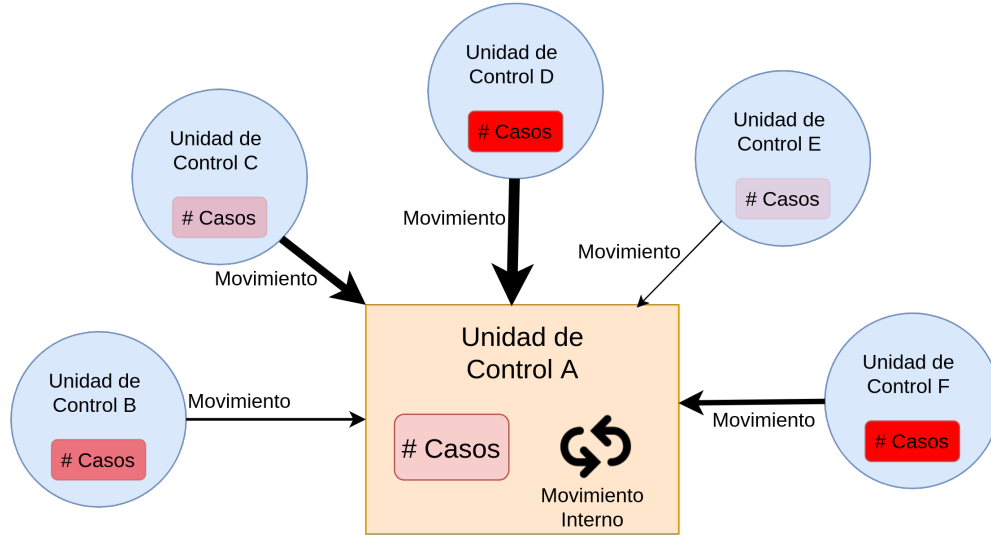


Figura 3: Funcionamiento modelo de proyección. A partir del movimiento y los casos para las distintas unidades espaciales alrededor, se construyen los datos de entrada al modelo para una unidad en particular.

5.1.2. Entrenamiento

A partir del conjunto de unidades de control para los distintos lugares de los que se tiene información (ver tabla 1), se seleccionan las 20 unidades con curvas epidémicas más similares a la unidad espacial que se quiere predecir. La similitud entre curvas epidémicas se hace utilizando una técnica llamada Dynamic Time Warping [?], únicamente comparando las partes de las curvas que coinciden en los días de la epidemia, esto quiere decir que lugares donde la epidemia esta menos avanzada que la de la unidad serán descartadas y las curvas de los lugares donde la epidemia esta mas avanzada serán recortadas.

El conjunto de series más similares a la unidad de control se actualiza cada vez que se ejecuta la predicción, permitiendo adaptarse en parte a los cambios que presenten.

Los datos contruidos para la unidad espacial serán unidos con los de sus unidades más similares para tener el conjunto final de datos para la proyección.

5.1.3. Modelo

El modelo de predicción utiliza una regresión lineal de mínimos cuadrados con regularización L2. Se usa un esquema de validación cruzada para encontrar el mejor parámetro del peso de la regularización.

5.2. Análisis Topológico de Datos (ATD)

El análisis topológico de datos, es un conjunto de técnicas y modelos que permiten inferir información sobre conjuntos de datos utilizando su estructura subyacente. Para este reporte, se utilizó nuestra propia implementación del algoritmo *Mapper* [?], disponible en etapa beta en [github](#). Una de las virtudes de este método es que captura los fenómenos locales y globales mejor que métodos similares, como algoritmos que analizan grupos o que reducen las dimensiones de los datos (puede interpretarse como una combinación de ambos). Ejemplos exitosos del uso de *Mapper* incluyen la identificación de subgrupos de pacientes que padecen enfermedades crónicas y transmisibles, grupos de patógenos [?, ?, ?, ?, ?, ?, ?, ?], patrones de votación [?], configuraciones de equipos deportivos [?], entre otros. Para obtener más detalles sobre los métodos, consulte [?, ?].

Puntualmente, los datos analizados son grafos que se pueden generar a partir de las unidades espaciales y su estado en cada día de la epidemia. El esquema es el siguiente:

- Para un día en particular, se filtran únicamente los casos confirmados y el movimiento en esa fecha.
- A cada unidad espacial se le asocian sus casos correspondientes creando un nodo en el grafo.
- Los nodos son unidos por aristas con el peso correspondiente al movimiento entre cada par de nodos. Adicionalmente, cada nodo se une a sí mismo con el valor de su movimiento interno para la fecha.

Por lo tanto, se obtiene un grafo no dirigido pesado para cada día de la epidemia.

Para la distancia entre cada uno de estos grafos se calculó la métrica de Wasserstein sobre la distribución de casos a nivel nacional, donde la distancia entre cada unidad de control corresponde a la distancia espacial desde su punto central. Por último, se proyecta cada grafo al espacio bidimensional de la suma de su movimientos internos y la suma de sus movimientos externos.

Utilizando *Mapper*, se identifican agrupaciones (clusters) de distribuciones espaciales de casos confirmados en diferentes intervalos de movilidad. Dado que el objetivo es caracterizar la relación entre movimiento interno-externo y casos confirmados, este método identifica distribuciones de casos espaciales similares por intervalos de movilidad. La intuición de este análisis es que se identifican perfiles de casos a nivel espacial (diarios y acumulados), y se asocian con perfiles de movilidad. Eso nos da una idea de perfiles de movimiento observados y sus respectivos perfiles de casos confirmados, lo cual es el objeto de este informe. Por ejemplo, se puede observar si patrones de movilidad relativamente similares corresponden a distribuciones de casos similares, o si por el contrario, se observan patrones de casos confirmados diferentes y variados con perfiles de movilidad relativamente similares.

5.3. Análisis topológico de datos

Las siguientes figuras (4- 5) muestran los resultados del espacio de proyección de los grafos diarios y el 1-esqueleto obtenido al ejecutar el algoritmo *Mapper* (ver sección 5.2).

Progresión de Movimientos

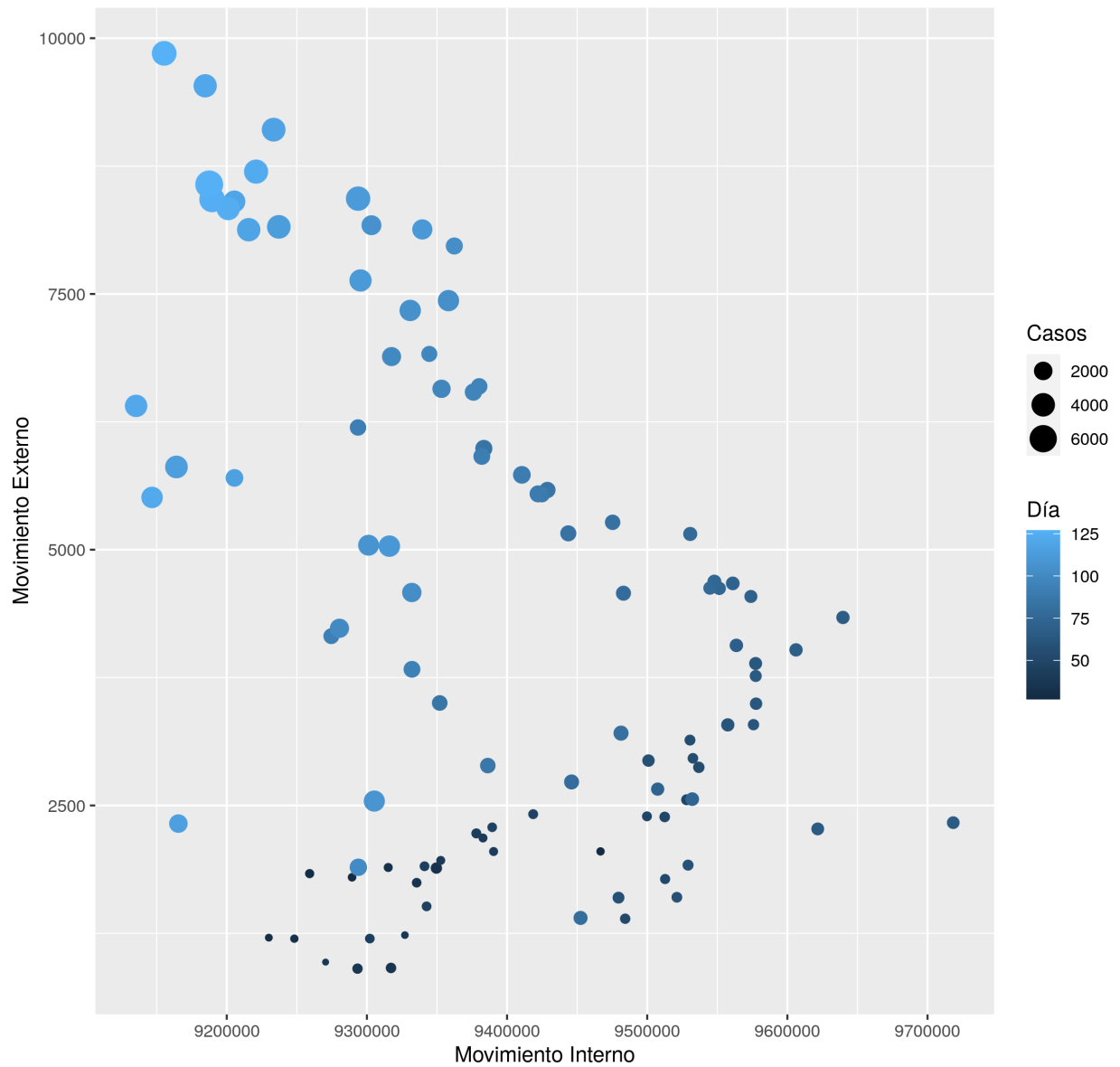


Figura 4: **Diagrama de dispersión de la movilidad interna/externa de las regiones.** Cada círculo representa un mapa; el tamaño representa el número de casos reportados para ese día, el color representa el día. Los ejes representan la movilidad interior (eje-x) y exterior (eje-y). La primera muestra corresponde a abril 4.

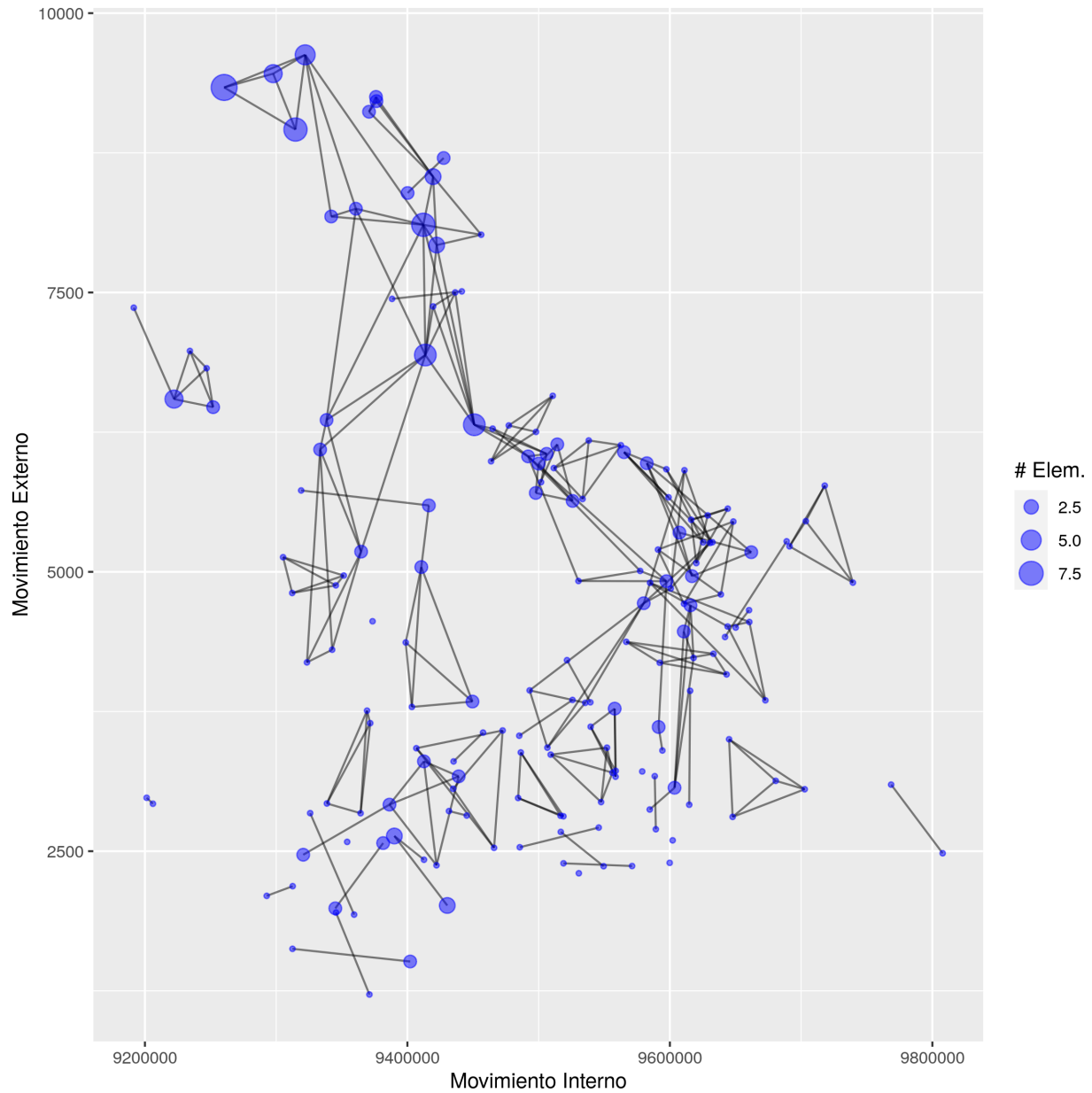


Figura 5: **Análisis Topológico de Datos (ATD)**. 1-esqueleto que representa agrupaciones (*clustering*) sobre estas dos dimensiones (movimiento interno y externo), con la distancia Wasserstein entre distribuciones espaciales de casos. Podemos observar que, a pesar de que la epidemia está en sus primeras etapas en Colombia, se aprecian ya grupos distintos de perfiles de movilidad y casos confirmados. Más interesante todavía, estos grupos pueden tener valores similares de movimiento interno y externo, pero la distribución espacial de casos es diferente y por eso se identifican diferentes grupos con valores similares. Esto quiere decir que perfiles de movilidad similares pueden conllevar distribuciones espaciales de casos diferentes.

6. Agradecimientos

Recibimos comentarios y valiosa retroalimentación de Mauricio Santhttps://www.overleaf.com/project/5f0f3ac54be2ad0001d7579aos-Vega (Universidad de Los Andes), Jaime Urrego (Universidad del Rosario) Diego Echeverri (Universidad Iceci-Unvialle), Pablo Cárdenas (MIT), Angélica Knudson Ospina (Universidad Nacional) y Manuela Carrasquilla (Harvard University).