UNIVERSIDAD NACIONAL DE INGENIERÍA

FACULTAD DE INGENIERÍA CIVIL

TESIS

APLICACIÓN DE TEORÍA DE GRAFOS PARA LA OPTIMIZACIÓN DE REDES HIDRÁULICAS EMPLEANDO LOS MÉTODOS DE CLUSTERING JERÁRQUICO Y ALGORITMOS GENÉTICOS

PARA OBTENER EL TÍTULO PROFESIONAL DE INGENIERO CIVIL

ELABORADO POR

JAIME HUAMANI RAMOS
ID: 0009-0006-6222-5733

ASESOR

Ph.D. JUAN WALTER CABRERA CABRERA
ID: 0000-0002-7490-7807

LIMA- PERÚ

2025

Huamani Ramos, Jaime
E-mail: jhuamani@uni.pe
Celular: 932192719

Dedicatoria

A mi padre Raúl Huamani C., por enseñarme la importancia del orden y la limpieza en el lugar del trabajo.

A mi madre Clara Ramos D., por creer en mi en los momentos más difíciles de mi vida y nunca dejar que me rinda.

A mi hermana Haydee Huamani R., quien me apoyo en el desarrollo de esta tesis y darme mi primer cuarto propio de mi vida.

A mi alma mater la Universidad Nacional de Ingeniería mi eterno agradecimiento por ayudarme a vivir mi vida profesional de la mejor manera.

AGRADECIMIENTOS

Al Ingeniero Sabino Pompeyo Basualdo Montes por hacer realidad el desarrollo del presente trabajo de investigación de tesis, por su apoyo en la realización de este tema de investigación, demostrando ser un gran profesional y ser humano.

A mi cuñado Diego Martin Montoya Martínez por ser mi ejemplo de cómo un verdadero profesional se abre camino en el ámbito laboral y ser un gran proveedor de su hogar.

A mi ídolo de toda la vida Cristiano Ronaldo, que con sus acciones me enseñó a darle el respeto que se merece a la profesión que escogí, no rendirse cuando la vida te da golpes duros y sobre todo a siempre buscar ser el mejor de todos.

ÍNDICE

Resumen

La sectorización de Redes de Distribución de Agua Potable (RDAP) es una técnica de gestión que consiste en subdividir la red en zonas homogéneas, con el objetivo de optimizar el control de aspectos clave como el caudal, las reparaciones, la detección de fugas y la calidad del servicio. En los últimos años, se han desarrollado diversas metodologías para llevar a cabo esta subdivisión, entre las que destacan el método de áreas, el método de agrupamiento y el método de cierre controlado de válvulas. Sin embargo, no todas estas metodologías se fundamentan en bases matemáticas sólidas, lo que puede limitar su efectividad y aplicabilidad en ciertos contextos.

El método de detección de comunidades en redes sociales aplicado en el presente trabajo abre una forma de sectorizar una Red de Distribución de Agua Potable basándose en un sustento matemático. El primer paso consiste en identificar las líneas de conducción principal utilizando el Concepto de Caminos más Cortos, propio de la teoría de grafos. En el segundo paso se subdivide la red basándose en el Clustering Jerárquico, que es un algoritmo de detección de comunidades en redes sociales. El tercer paso consiste en calcular la cantidad de roturas detectadas aplicando la Simulación Monte Carlo y volumen de fugas de fondo usando el Modelo Conceptual en cada subdivisión. Para el cuarto paso, con todo lo anterior mencionado se logra optimizar el Conjunto de Entrada y Válvulas de Cierre (CEVC) haciendo uso de los algoritmos genéticos para evitar óptimos locales y encontrar el óptimo global.

El método de sectorización propuesto representa una solución viable para los desafíos a largo plazo en la gestión de redes de agua. En su primer año de implementación, este enfoque ha demostrado generar un ahorro significativo en términos de caudal y costos de reparaciones, superando los gastos iniciales de implementación por un margen de S/. 235.69, equivalente al 0.12% del costo total en el primer año de implementación.

Abstract

The sectorization of Drinking Water Distribution Networks (RDAP) is a management technique that consists of subdividing the network into homogeneous zones, with the aim of optimizing the control of key aspects such as flow, repairs, leak detection and quality of service. In recent years, various methodologies have been developed to carry out this subdivision, among which the area method, the grouping method and the controlled valve closure method stand out. However, not all of these methodologies are based on solid mathematical foundations, which may limit their effectiveness and applicability in certain contexts.

The method of detecting communities in social networks applied in this work opens a way to sectorize a Drinking Water Distribution Network based on a mathematical basis. The first step is to identify the main conduction lines using the Concept of Shortest Paths, typical of graph theory. In the second step, the network is subdivided based on Hierarchical Clustering, which is an algorithm for detecting communities in social networks. The third step consists of calculating the number of breaks detected by applying the Monte Carlo Simulation and the volume of background leaks using the Conceptual Model in each subdivision. For the fourth step, with all of the above mentioned, it is possible to optimize the Inlet and Shut-Off Valve Assembly (CEVC) using genetic algorithms to avoid local optima and find the global optimum.

The proposed sectorization method represents a viable solution to long-term challenges in water network management. In its first year of implementation, this approach has proven to generate significant savings in terms of flow and repair costs, exceeding the initial implementation expenses by a margin of S/. 235.69, equivalent to 0.12% of the total cost in the first year of implementation.

Prólogo

Uno de los grandes problemas de las redes de agua potable (RDAP) en nuestro país es el elevado volumen de pérdidas, problema que se agrava en la dificultad inherente de detectar la ubicación de dichas pérdidas y la afectación a la población que implica cortar temporalmente el abastecimiento de agua en la zona para hacer el mantenimiento respectivo. Esta situación conlleva a replantear la metodología de diseño convencional y buscar alternativas que permitan prever y controlar de manera más eficiente la ocurrencia de este problema. En este contexto, la presente investigación propone el uso del Método de Detección de Comunidades en Redes Sociales para sectorizar redes de distribución de agua potable basado en la aplicación de Teoría de Grafos para la Optimización de redes hidráulicas empleando los métodos de Clustering Jerárquico y Algoritmos Genéticos en una red con poco tiempo de operación. Para este fin, se seleccionó una red hidráulica ubicada en la zona norte de Lima, correspondiente a parte de los distritos de Callao, Ventanilla Y San Martin de Porres, operativa y con poco tiempo de haber sido implementada.

El método propuesto consta de cuatro etapas: primero, identificar las líneas de conducción principal utilizando el concepto de “caminos más cortos”; segundo, subdividir la red aplicando el algoritmo de Clustering Jerárquico; posteriormente, estimar la cantidad de roturas detectadas aplicando Simulación Monte Carlo; y, finalmente, optimizar el Conjunto de Entrada y Válvulas de Cierre (CEVC) aplicando algoritmos genéticos para evitar óptimos locales y encontrar el óptimo global. Los resultados encontrados muestran un notable ahorro de cerca de S./200 000 anuales que, si bien puede parecer un monto pequeño respecto a los costos del proyecto, aplicado a su vida útil justifica ampliamente su implementación. De esta manera, el trabajo presentado abre un nuevo abanico de posibilidades dentro del estudio y optimización de redes de agua potable, rompiendo los esquemas tradicionales y aplicando técnicas estadísticas con un sólido sustento matemático.

Ph.D. Juan Walter Cabrera Cabrera
Asesor

Lista de tablas

Lista de figuras

Lista de símbolos y siglas

RDAP	**:**	Red de Distribución de Agua Potable
WWC	**:**	World Water Council
MCA	**:**	Metros Columna de Agua
UOC	**:**	Unidad Operativa de Control
IWA	**:**	International Water Association
SMC	**:**	Simulación Monte Carlo
AG	**:**	Algoritmos Genéticos
WWC	**:**	World Water Council
IWA	**:**	International Water Association
PNUMA	**:**	Programa de las Naciones Unidas para el Medio Ambiente
CMC	**:**	Camino Mas Corto
BFS	**:**	Breadth First Search
DFS	**:**	Depth First Search
FIFO	**:**	First in First Out
PAM	**:**	Partición Alrededor de Medoides
DIANA	**:**	Distance Geometry Algorithm for NMR Applications
CCCF	**:**	Coeficiente de Correlación Cofenetica
EPA	**:**	Environmental Protection Agency
INP	**:**	Input
AS	**:**	Ancho de Silueta

Capítulo I: Introducción

1.1 Generalidades

1.1.1 Antecedentes Referenciales

El crecimiento poblacional ocasiona el incremento del tamaño de las ciudades en todo el mundo y por ende aumenta la demanda del recurso hídrico. Las ciudades deben contar con redes de agua potable proporcionales a su tamaño, administradas por una entidad encargada de la distribución, mantenimiento y monitoreo del servicio de agua potable. (Quivera, 2009)

Desde una perspectiva técnica, los problemas asociados a las Redes de Distribución de Agua Potable (RDAPs) se pueden agrupar en cuatro áreas principales: fugas y agua no contabilizada; integridad estructural de la red; calidad del agua que se distribuye; y la fiabilidad y precisión de la base de datos de los sistemas de distribución. En relación con el primer aspecto, el control de pérdidas ha sido una preocupación desde la construcción de las primeras redes. (Pilcher et al., 2007)

Debido a los altos volúmenes de pérdidas en las redes de agua potable, Campbell (2017) propone la sectorización de las redes hidráulicas tomando en cuenta diferentes criterios. Para tal fin propone el uso de algoritmos de detección de comunidades en redes sociales, como el Clustering jerárquico, algoritmo de detección multinivel o Método Louvain y la detección de comunidades a través de Caminos aleatorios. El método fue aplicado en Managua-Nicaragua, obteniéndose una reducción de las pérdidas de hasta un 12.5% en caudal.

En la ciudad de Huacho-Perú se ha empleado el criterio de dividir a la red por áreas, con un valor máximo de 300 Ha y un estimado de 400 a 4 000 usuarios. En la ciudad del Callao-Perú, el método consistió en dividir la red en áreas menores a 3 $km2$, con un solo punto de entrada y otro en caso de emergencia, con presiones entre 15 a 50 MCA (metros columna de agua) y utilizaron las avenidas como límites de sector. Se instalaron circuitos con tuberías de gran diámetro para aumentar la presión en las tuberías (Vegas, 2012).

Según un estudio realizado por la Organización Mundial de la Salud (2020), la implementación de la sectorización puede reducir las pérdidas de agua, que en redes no sectorizadas pueden alcanzar hasta un 30% del caudal total, disminuyendo a cifras tan bajas como el 17% tras su aplicación. Además, se ha observado un incremento en la eficiencia energética, con mejoras que oscilan entre el 8% y el 10% en el consumo de energía para bombeo. Otro beneficio significativo es la capacidad de regular las presiones en los diferentes sectores, lo cual no solo optimiza el uso del recurso hídrico, sino que también mejora la calidad del servicio al garantizar un suministro continuo y adecuado a los usuarios (OMS, 2020).

1.1.2 Descripción del problema de investigación

En los últimos 40 años el buen uso y distribución del agua potable es fundamental para garantizar la prosperidad de las futuras generaciones de la humanidad, además que la población sigue creciendo de manera exponencial en el mundo al igual que sus requerimientos.

En países en vías de desarrollo se estima que se pierde entre un 40% y 50% (entre fugas de fondo, roturas, conexiones clandestinas, etc.) del agua entregada, lo cual significa que para que un usuario tenga acceso a 1 m³ de agua tiene que entregarse a la red 2 m³ de agua potable. Se calcula que se pierden alrededor de 26.7 millones de m³ que representa un total 5.9 miles de millones de dólares americanos en costos marginales. Esta situación seria aún más grave si se contabilizaran los costos de externalidades tales como el costo por reparación de daños entre otros. (Kingdom, Liemberger, & Philippe, 2006)

Si se resolviera el problema anteriormente explicado, se podría alimentar hasta 90 millones de personas adicionales (WWC, 2009) en países de ingresos bajos y medianos, ahorrando 11 mil millones de m³ al año. (IWA (International Water Association), 2000) En los países desarrollados, se estima que la pérdida de agua varía en más del 15 %, y las proyecciones en todo el mundo son preocupantes. (Kingdom, Liemberger, & Philippe, 2006) Un claro ejemplo es que para el año 2025, dos tercios de la población mundial sufrirá estrés hídrico. (PNUMA, 2019) Cabe mencionar que el llegar a un 0% de pérdidas en las redes de distribución de

agua potable es ilusorio. Sin embargo, al año 2020 se han publicado técnicas novedosas y equipos sofisticas que permiten disminuir las pérdidas de agua potable entre las cuales se tiene la subdivisión de las redes en subsectores, métodos automáticos de cierre de válvulas, uso de grabadores acústicos y búsqueda sonora. (Pilcher, y otros, 2007)

En el Perú, el porcentaje de agua no facturada se calcula mediante la diferencia entre el volumen producido y facturado (cobrado) de agua potable y dividido entre el volumen de agua producida, la SUNASS (2023) elaboro un informe en donde se mostró que entre los años 2021 y 2022 el porcentaje de agua no facturada aumento un 2.27% a nivel nacional. Si bien la empresa Sedapal hizo mejoras en los micromedidores en todas sus redes, esto no pudo contrarrestar que el agua facturada disminuyera en las demás empresas prestadoras del servicio. La figura N° 1 muestra que el porcentaje de agua no facturada supera el 30%, esto indica que por cada 10 m³ entregados a una red de agua potable se pierden 3 m³.

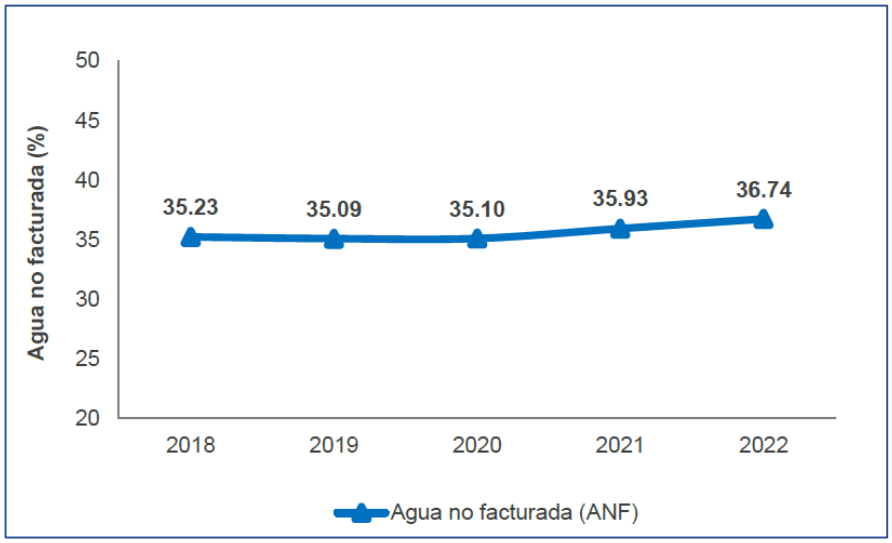

Figura N° 1. Evolución del agua no facturada, 2016-2022

(Fuente: SUNASS, 2023)

La figura N° 2 muestra el porcentaje de agua no facturada de las principales empresas prestadoras comparadas con Sedapal, la cual tiene el menor porcentaje de agua no facturada desde el año 2018 al 2022.

La empresa Grande 1 tiene un porcentaje de agua no facturada del 44.48% en el año 2022, la empresa Grande 2 tiene un porcentaje de agua no del 39.65% en el año 2022, la empresa Mediana tiene un porcentaje de agua no facturada de 41.73% y la empresa Pequeña tiene un porcentaje de agua no facturada del 39.25% en el año 2022. El porcentaje de agua no facturada en los últimos 5 años varía entre 26.50% y 48.05%, siendo la empresa Pequeña la que más agua no factura.

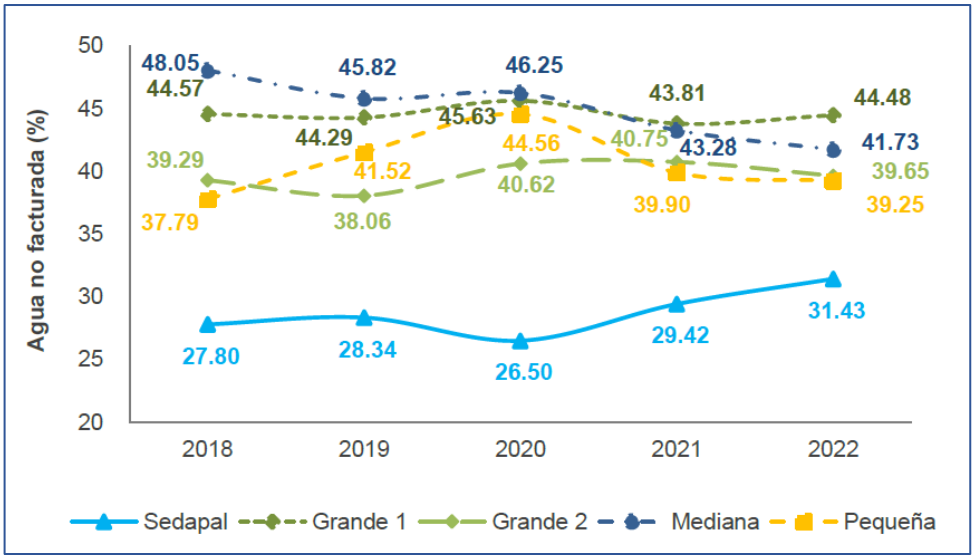

Figura N° 2. Evolución del agua no facturada por grupo de Empresa Prestadora, 2018-2022
(Fuente: SUNASS, 2023)

En donde las Empresas prestadoras Grande 1, Grande 2, mediana y pequeña representa a las siguientes EPS en el Cuadro N°1:

Cuadro N°1: Clasificación de las EPS

Grande 1	Grande 2	Mediana	Pequeña
SEDALIB S.A.	EPS SEDALORETO S.A.	EMAPISCO S.A.	EPSSMU S.A.
EPS GRAU S.A.	EPS SEDA AYACUCHO S.A.	EPS EMAPAT S.A.	EMAPA PASCO S.A.
	EMAPA SAN MARTÍN S.A.	EPS SELVA CENTRAL S.A.	EMAPAVIS S.A.
	EPS SEMAPACH S.A.	EPS MOQUEGUA S.A.	EMAPAB S.A.
	EPS SEDAJULIACA S.A.	EMAPA HUARAL S.A.	EPS AGUAS DEL ALTIPLANO S.R.L.
	SEDAM HUANCAYO	EPS ILO S.A.	
		EPS BARRANCA S.A.	

(Fuente: Elaboración propia)

1.2 Objetivos

1.2.1 Objetivo General

Aplicar la teoría de grafos para la optimización de redes hidráulicas empleando los métodos de Clustering jerárquico en la sectorización y algoritmos genéticos para la ubicación de válvulas de cierre y puntos de abastecimiento.

1.2.2 Objetivos Específicos:

- Seleccionar una línea de aducción para la red de distribución en estudio, tomando como criterio el análisis de los caudales de mayor demanda y el concepto del Camino más Corto propio de la teoría de grafos.
- Definir la cantidad de subsectores que debe tener la red de distribución de agua potable con la técnica de Clustering Jerárquico.
- Optimizar la ubicación de las válvulas de cierre y puntos de abastecimiento usando algoritmos genéticos, incluyendo la representación de los eventos de fugas en cada subsector empleando la Simulación Montecarlo.
- Comparar los beneficios económicos de los costos marginales del resultado de la sectorización propuesta contra el modelo convencional que se encuentra en funcionamiento de una red de distribución existente.

1.3 Formulación de la hipótesis

1.3.1 Hipótesis General

El uso de la teoría de Grafos en la propuesta de sectorización de una red existente, empleando los métodos de Clustering Jerárquico (algoritmo de detección de comunidades en redes sociales) y Algoritmos Genéticos, permitirá optimizar la red de distribución de agua potable.

1.3.2 Hipótesis Específicos

- La selección de una línea de aducción en la red de distribución, basada en el análisis de los caudales de mayor demanda y utilizando el concepto del Camino más Corto de la teoría de grafos, optimizará la eficiencia del suministro y reducirá las pérdidas hidráulicas en el sistema.
- La aplicación de la técnica de Clustering Jerárquico permitirá identificar y definir un número óptimo de subsectores en la red de distribución de agua potable, basado en la demanda y características geográficas de los nodos

y tuberías, lo que resultará en una gestión más eficiente y uniforme del recurso hídrico.

- La optimización de la ubicación de las válvulas de cierre y los puntos de abastecimiento mediante algoritmos genéticos, junto con la representación de eventos de fugas a través de la Simulación Montecarlo, resultará en un sistema de distribución de agua más eficiente y resiliente, disminuyendo las pérdidas y mejorando la capacidad de respuesta ante emergencias.
- La sectorización propuesta de la red de distribución de agua potable generará beneficios económicos al reducir las pérdidas de agua potable en comparación con el modelo convencional actualmente en funcionamiento, lo que permitirá una gestión más eficiente y sostenible del recurso hídrico.

Capitulo II: Marco teórico y conceptual

2.1 Teoría de grafos

La teoría de grafos es parte de las matemáticas discretas que se ocupa de los arreglos de objetos discretos (números enteros, funciones, conjuntos, proporciones, grafos, etc.) que se encuentran separados unos a otros. En otras palabras, un grafo es una estructura de datos no lineal que puede tomar formas complejas.

Un grafo $G = V, E$ es un conjunto de elementos conformados por nodos y vértices (dependiendo del tipo de grafo) que están unidos por enlaces, arcos o aristas que permiten representar relaciones binarias entre los elementos de un conjunto. (Thulasiraman & Swamy, 1992) Los arcos son representados por un par de elementos (v_i, v_j) donde v_i y v_j son los vértices unidos por el arco. La estructura elemental de un grafo está conformada por una arista y dos vértices unidos entre si (conocido como singletón). Pese a esto, los grafos pueden tomar formas muy complejas. Al tener una gran variedad de características y propiedades, consecuentemente tienen un gran número de formas de clasificarlos.

Una primera forma de clasificar los grafos se basa en donde comienzan y terminan los vértices de las aristas (Ver Figura N°3), pudiendo ser:

- Adyacentes; cuando un par de aristas tienen un vértice en común
- Paralelas; cuando un par de vértices comparten una arista
- Lazo; cuando los vértices de partida y llegada son lo mismo

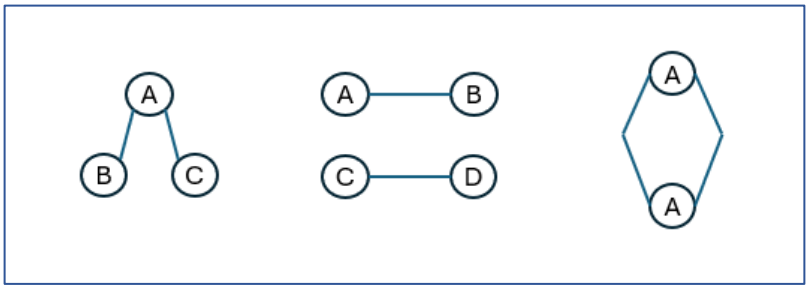

Figura N° 3. Tipos de grafos según su inicio y fin.
(Fuente: Elaboración propia)

Otro tipo de clasificación toma en cuenta si los puntos de partida y llegada están definidos, en este caso se tiene:

2.1.1 Grafos dirigidos

Un dígrafo o grafo dirigido, es un par $G = (V, U)$, donde $U \subset V\ x\ V$ son los arcos del dígrafo. En los dígrafos existe orientación en los arcos, es decir, $u = (x, y) \neq (y, x)$. (Ver Figura N° 4)

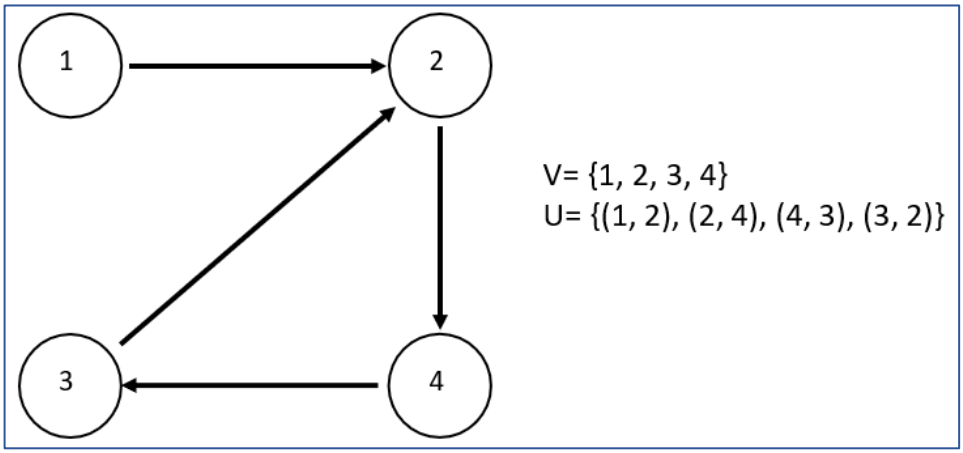

Figura N° 4. Ejemplo de grafos dirigidos
(Fuente: Campbell, 2017)

Y dentro de los grafos dirigidos se pueden clasificar en:

- Grafos simétricos. – aquí se encuentran los grafos en donde las aristas direccionadas tiene otra arista en dirección contraria.
- Grafos dirigidos acíclicos. -aquí se encuentran los grafos dirigidos acíclicos.
- Grafos tipo torneo. – se obtendrá seleccionando una dirección en específico sin faltar a ningún enlace en un grafo completo.

2.1.2 Grafos no dirigidos

Los grafos no dirigidos como su nombre lo indica, no tienen dirección o también conocidos como pares no ordenados. Topológicamente son representados por puntos conectados por aristas que no tienen dirección y su representación matemática es de la siguiente manera:

$G = (V, E)$ es no dirigido si para cada $v, w \in V \rightarrow (V, W) \in E \Leftrightarrow (w, v) \in E$

2.1.3 Características de los grafos

Entre las principales características de los grafos tenemos los siguientes referentes a su forma:

a) Conectividad: La conectividad de los grafos es una medida de su robustez y cohesión, refiriéndose al número mínimo de elementos (vértices o aristas) que deben ser eliminados para desconectar el grafo.
b) Embebido en Superficies: Se refiere a la representación de un grafo sobre una superficie, como una esfera o un plano, de manera que las aristas no se crucen entre sí, excepto en vértices.
c) Grado de los Vértices: Es una medida fundamental que indica la cantidad de aristas que inciden en ese vértice.
d) Dualidad: Es un concepto fundamental en la teoría de grafos que se refiere a la relación entre un grafo planar y su grafo dual.

Tratándose de la propiedad del grado, la conectividad de los vértices es una de las más importantes; y es conocido como el grado o valencia del grafo. Esta propiedad consiste en el conteo de la cantidad de aristas que inciden en un vértice (un bucle se considera como dos aristas). Si se tuviera el caso en que ninguna arista incida en uno de los vértices de un grafo se considera como un vértice aislado o vértice con valencia igual a 0. Otro caso es cuando en un vértice solo incide una arista y por ello tendrá un grado o valencia de 1, son comúnmente nombrados vértice hojas y sus aristas incidentes con también llamadas aristas pendientes.

También se puede incluir al vértice que se encuentra unido a todos los vértices de un grafo mediante al menos una arista y cuenta con un grado igual a $n-1$, es llamado como vértice pendiente. De todo lo expuesto con anterioridad, se concluye que los vértices de un singletón se les conoce como vértices pendientes por su simplicidad. En esta misma línea, la definición de un grafo completo es cuando un vértice está conectado mediante varias aristas a otros vértices y estos a su vez tienen como punto de partida otros vértices. Siguiendo con este caso, la cantidad de aristas que inciden sobre todos los vértices es $n-1$, donde n es la cantidad de vértices que conforman el grafo. Dicho número indica el grado de cada vértice. El grado como propiedad se puede extrapolar, en donde un grafo regular

tiene un grado igual al grado de cada uno de sus vértices. (los nodos de un grafo regular tienen igual cantidad de aristas) (Ver Figura N° 5)

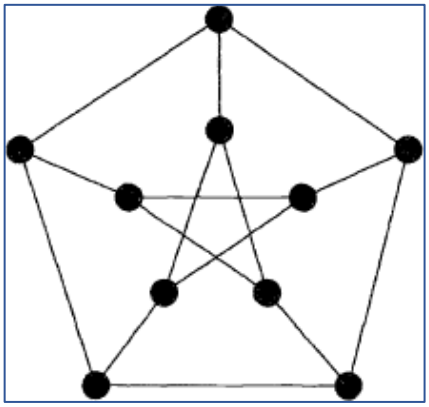

Figura N° 5. Grafo completo en donde el grado de todos los vértices es igual a 3
(Fuente: Campbell, 2017)

Los grafos irregulares en cambio consideran que el vértice con más aristas determina el grado del grafo y el vértice con la menor cantidad de aristas conectados será el grado mínimo. En los grafos, las aristas también cuentan con propiedades o características. Otra forma de clasificar los grafos es considerando el peso de las aristas, en donde si el peso de las aristas es diferente de 1 es un grafo pondera y caso contrario, un grafo no ponderado.

Actualmente es de gran interés de estudio la accesibilidad entre los vértices de un grafo y este se estudia a través de recorridos por los vértices y aristas del grafo. Como los grafos son conjuntos finitos de aristas y vértices los recorridos forman cadenas, las cuales pueden ser:

a) Cadenas abiertas: aquellos grafos en donde los vértices iniciales y finales no coinciden.
b) Cadenas cerradas: aquellos grafos en donde los vértices iniciales y finales coinciden.

Las cadenas también pueden ser clasificadas como caminos y ciclos. En los caminos no se repiten aristas o vértice en el mismo; y en los ciclos solo se repiten los vértices final e inicial.

Entre los ciclos más conocidos tenemos al Hamilton y Euler. En cualquiera de los casos la cantidad de aristas que forman la longitud del grafo los ciclos se relacionan con el concepto de diámetro del grafo. La máxima longitud requerida para llegar de un vértice a cualquier otro es igual al diámetro del grafo.

2.1.4 Representación Matricial de Grafos

Un grafo puede ser representado de manera matricial (matriz cuadrada $n\,x\,n$) donde n es el número de aristas. Se parte de un grafo no ponderado, su matriz de adyacencia estará conformada por ceros (cuando dos vértices no están conectados) y unos (cuando dos vértices están conectados).

En la Figura N° 6 se puede apreciar la representación gráfica de un grafo no ponderado.

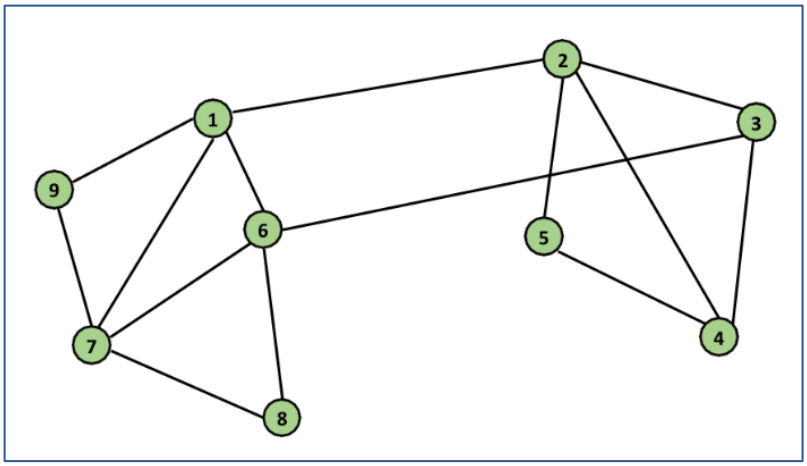

Figura N° 6. Grafo ejemplo
(Fuente: Elaboración propia)

En la Tabla N° 1 se tiene la representación gráfica de un grafo y su respectiva matriz de adyacencia.

Wij = 1 si i y j son adyacentes
Wij = 0 si i y j no son adyacentes

Tabla N° 1. Representación matricial del grafo ejemplo no ponderado.

	1	2	3	4	5	6	7	8	9
1	0	1	0	0	0	1	1	0	1
2	1	0	1	1	1	0	0	0	0
3	0	1	0	1	0	1	0	0	0
4	0	1	1	0	1	0	0	0	0
5	0	1	0	1	0	0	0	0	0
6	1	0	1	0	0	0	1	1	0
7	1	0	0	0	0	1	0	1	1
8	0	0	0	0	0	1	1	0	0
9	1	0	0	0	0	0	1	0	0

(Fuente: Elaboración propia)

2.1.4.1 *Matriz de adyacencia (W_{ij}) del grafo ejemplo*

Los valores de la diagonal de la matriz son 0 ya que el grafo representado no cuenta con bucles como aristas. A continuación, se describirán diferentes formas de representar la relación de los vértices entre sí.

- Matriz de disimilaridad. - indica la diferencia que existe entre las características de los vertidos o nodos de un grafo y se miden mediante distintas distancias métricas como la Euclidiana, Manhattan, etc.
- Matriz de grado “D”. - esta matriz contiene información referente al grado de los vértices de un grafo. Los elementos de la matriz en donde i<>j no representan un vértice se le considerara igual a cero. (Ver Ecuación 1):

$$D_{ij} = \begin{cases} 0, & \text{si } i \neq \text{j} \\ deg(Vi), & \text{si } i \neq \text{j} \end{cases} \qquad \text{(Ecuación 1)}$$

- Matriz de afinidad A.- en la matriz se tiene los pesos de las aristas del grafo, donde los elementos de la matriz iguales a 1 representan el valor del peso

$$A_{ij} = S_{ij} \text{ ; sí } i \text{ y } j \text{ están conectados}$$

$$A_{ij} = 0 \text{ ; sí } i \text{ y } j \text{ no están conectados}$$

En la figura N° 7 se puede apreciar la representación gráfica de un grafo ponderado y en la tabla N° 2 la representación matricial.

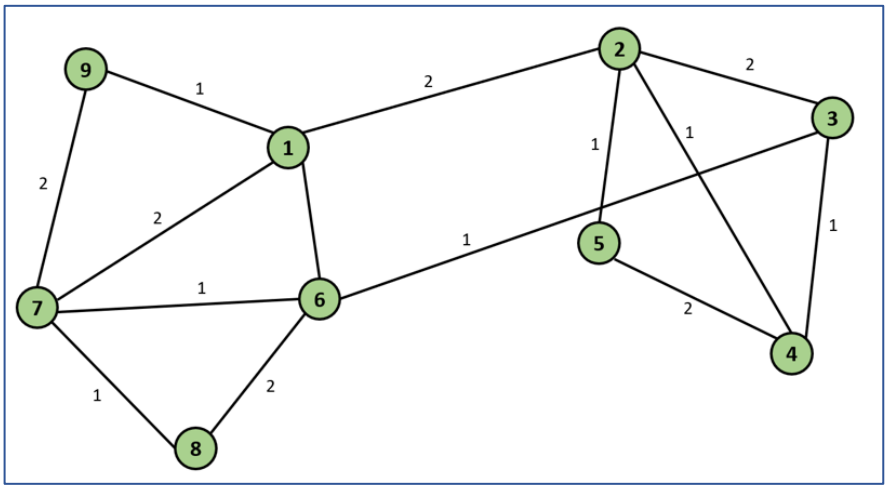

Figura N° 7. Grafo ponderado.
(Fuente: Elaboración propia)

Tabla N° 2. Matriz de afinidad $A = A_{ij}$ del grafo de la ilustración.

	1	2	3	4	5	6	7	8	9
1	0	2	0	0	0	1	2	0	1
2	2	0	2	1	1	0	0	0	0
3	0	2	0	1	0	1	0	0	0
4	0	1	1	0	2	0	0	0	0
5	0	1	0	2	0	0	0	0	0
6	1	0	1	0	0	0	1	2	0
7	2	0	0	0	0	1	0	1	2
8	0	0	0	0	0	2	1	0	0
9	1	0	0	0	0	0	2	0	0

(Fuente: Elaboración propia)

La matriz Laplaciana L.- es el resultado de restar la matriz A y la D, tal como se puede apreciar en la Ecuación 2.

$$L = D - A$$

$$L_{ij} = \begin{cases} deg(Vi), & \text{si } i = \text{j} \\ -1, & \text{si } i \neq \text{j y vi es adyacente a vj} \\ 0, & \text{en cualquier otro caso} \end{cases} \qquad \text{(Ecuación 2)}$$

La matriz no normalizada L quedaría como se tiene a continuación:

Tabla N° 3. Matriz Laplaciana $L = L_{ij}$ del grafo

	1	2	3	4	5	6	7	8	9
1	4	-1	0	0	0	-1	-1	0	-1
2	-1	4	-1	-1	-1	0	0	0	0
3	0	-1	3	-1	0	-1	0	0	0
4	0	-1	-1	3	-1	0	0	0	0
5	0	-1	0	-1	2	0	0	0	0
6	-1	0	-1	0	0	4	-1	-1	0
7	-1	0	0	0	0	-1	4	-1	-1
8	0	0	0	0	0	-1	-1	2	0
9	-1	0	0	0	0	0	-1	0	2

(Fuente: Elaboración propia)

Si se quiere normalizar se utilizará la siguiente regla de correspondencia:

$$L_{ij}^{*} = \begin{cases} 1 \text{ si } i = j \text{ y } L_{ii} \neq 0 \\ -\dfrac{1}{\sqrt{L_{ii} L_{jj}}} \; i \neq j \text{ y } v_i \text{ es adyacente a } v_j \\ 0 \text{ en cualquier otro caso} \end{cases} \qquad \text{(Ecuación 3)}$$

2.1.5 Concepto de Caminos más Cortos en Grafos

En el transporte de materias primas el tiempo es uno de los factores más importantes en su desarrollo, debido a esto se buscan las posibles soluciones para minimizarlo. En la teoría del flujo de mínimo coste se encuentra el problema del Camino más Corto propio de la teoría de grafos. (More, 1959)

Para el grafo dirigido $G = \{V, E\}$,se tiene a los nodos V y arista E representado a continuación en la Figura N° 8:

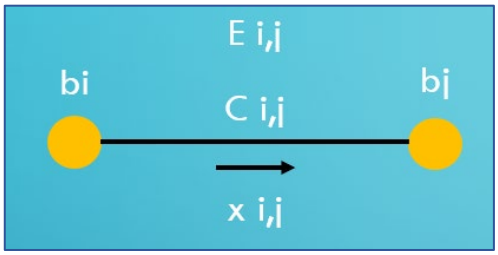

Figura N° 8. Esquema de grafo dirigido

(Fuente: Elaboración propia)

- E (i, j) = conjunto de enlaces dirigidos (par ordenado)
- b (i, j) = valor entero por nodo
- x (i, j) = flujo que circula por una arista
- C (i, j) = peso por unidad que circula

Los b (i) son valores enteros pertenecientes a Vi en V, se tendrá lo siguiente:

- b(i) > 0: nodo de abastecimiento
- b(i) = 0: nodo de transporte
- b(i) < 0: nodo de demanda

$$min_P \sum_{(i,j)\in P} C_{ij}x_{ij} \qquad \text{(Ecuación 4)}$$

Sujeto a:

$$\sum_{\{j:(i,j)\in P\}} x_{ij} - \sum_{\{j:(i,j)\in P\}} x_{ji} = b(i), \forall i$$

$$I_{ij} \leq x_{ij} \leq u_{ij}, \forall i,j \qquad \text{(Ecuación 5)}$$

I_{ij} y u_{ij} son las restricciones del flujo y P la ruta aleatoria entre los vértices del grafo. (Ahuja, Magnanti, & Orlin, 1993)

- b (i) = 1: es un nodo fuente
- b (j) = -1: es un nodo sumidero
- b (k) = 0: es el resto de los nodos

En las restricciones mostradas líneas arriba I_{ij} y u_{ij} son conocidas como el balance de masa estableciendo que la demanda del nodo es igual a la diferencia del flujo de salida y entrada. A continuación, se tienen los siguientes casos:

- Nodo de abastecimiento: cuando la diferencia entre el flujo de salida y entrada es positiva.
- Nodo de demanda: cuando la diferencia entre el flujo de salida y entrada es negativa.

- Nodo de transporte: cuando la diferencia entre el flujo de salida y entrada es igual a cero.

Existe un límite de flujo utilizado para modelar restricciones impuestas sobre el rango de operaciones de un caudal dado (Campbell, 2017). El problema de Camino más Corto es una aplicación del problema descrito líneas arriba y consiste en enviar una partícula del nodo fuente en dirección al nodo objetivo para hacer un mapeo de los posibles caminos solución. Para resolver dicho problema se diseñaron diferentes algoritmos y entre los más conocidos se tienen los algoritmos PRIM, DFS, BFS y Dijkstra.

2.1.5.1 Algoritmo de búsqueda en Amplitud (BFS)

Publicado por E.F More (1950) y Lee (1961) respectivamente. El algoritmo BFS sirve como base para los algoritmos Dijkstra y PRIM, este registra las distancias (considerando el mínimo número de enlaces y pesos) desde un nodo R en dirección al nodo inmediatamente conectado a este y así sucesivamente hasta alcanzar todos los nodos de la red. Después de ello se deberá escoger otro nodo R y repetir el proceso hasta alcanzar todos los nodos de la red (Ver Figura N° 9). (Campbell, 2017).

```
Entrada G y R
Se inicializan todos los v en G, marcándolos como no
visitados
Para cada v en G
{
Estado = No visitado
Distancia = infinita
Padre = Null
}
Se crea un saco S
Introducir R en S
Mientras S no esté vacío
{
      Introducir v  en S
{
Introducir todos los nodos adyacentes v' en S
Sacar el primer v' de S
Introducir todos los nodos v'' adyacente a v'en el saco
 }
}
Salida: árbol de profundidad del G
```

Figura N° 9. Pseudocódigo 1 describe el algoritmo BFS. grafo dirigido

(Fuente: Campbell, 2017)

En donde el llenado de datos esta dictado por el principio FIFO, esto quiere decir que el primer nodo R que se envía en la fila uno es también la primera que sale del algoritmo. El Pseudocódigo 1 es la base para crear el algoritmo en cualquier lenguaje de programación.

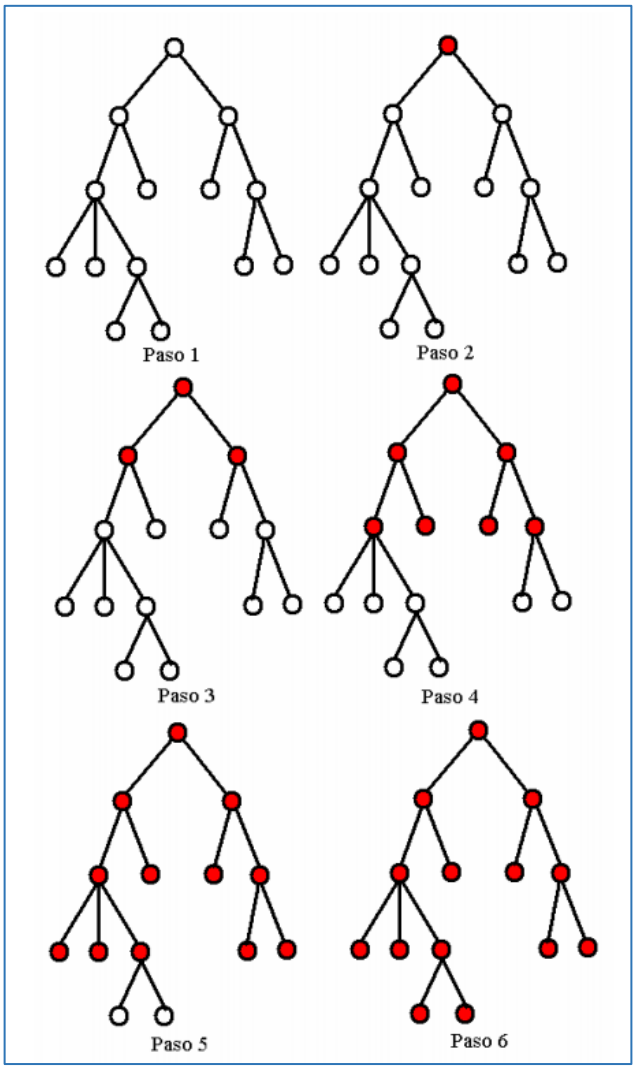

Figura N° 10. Secuencia seguida por el algoritmo BFS para recorrer
(Fuente: Leon Escobar & Giraldo Suarez, 2005)

2.1.5.2 Algoritmo de búsqueda en profundidad DFS

Por sus siglas en ingles el algoritmo DFS consiste en realizar un mapeo más profundo del grafo. Este algoritmo no espera a conectar todos los nodos y después cambiar de origen, a diferencia del algoritmo BFS una vez encontrado todos los nodos inmediatamente conectados al nodo R cambia de origen y se repite el proceso hasta mapear todos los nodos de la red (Ver Figura N° 11 y 12).

```
Entrada G y R
Se inicializan todos los v en G, marcándolos como no
visitados
Para cada v en G
{
Estado = No visitado
Distancia = infinita
Padre = Null
}
Se crea un saco S
Introducir R en S
Mientras S no esté vacío
{
      Introducir v  en S
{
Introducir todos los nodos adyacentes v' en S
Sacar el primer v' de S
Introducir todos los nodos v'' adyacente a v'en el saco
 }
}
Salida: árbol de profundidad del G
```

Figura N° 11. El Pseudocódigo 2 describe el algoritmo DFS

(Fuente: Campbell, 2017)

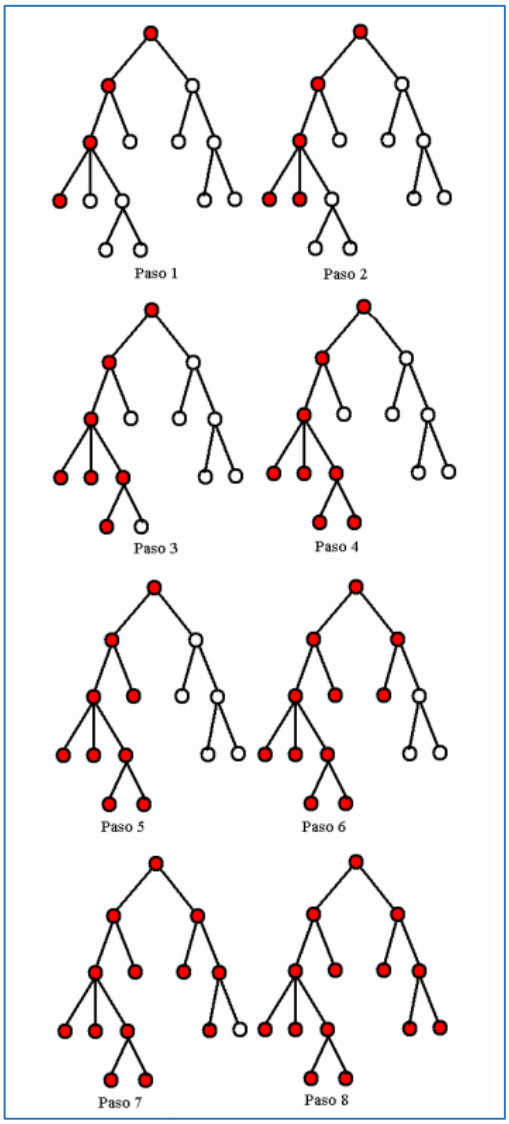

Figura N° 12. Secuencia seguida por el algoritmo DFS para recorrer todo el grafo
(Fuente: Leon Escobar & Giraldo Suarez, 2005)

2.1.5.3 Algoritmo de Dijkstra

El Algoritmo de Dijkstra tiene como objetivo encontrar la distancia más corta de un vértice inicial a cualquier otro vértice. Al aplicar este algoritmo se formará un árbol individual con los enlaces, dicho árbol inicia del nodo R (seleccionado aleatoriamente) y empieza a mapear todos los nodos del grafo. Cabe mencionar que el algoritmo almacena los enlaces tratando de minimizar el peso acumulado de estos.

Para la construcción del árbol individual mencionado líneas arriba se requiere de un método óptimo para seleccionarlo. En el Pseudocódigo 3, el grado conectado y el nodo R son los datos que se necesitan para incrementar el tamaño del árbol. En el mismo momento que se esté ejecutando el algoritmo, los datos que estén en el árbol se guardan en G y los que quedan fuera se guardan en Q basándose en una propiedad clave. La propiedad clave es el mínimo peso de los enlaces conectados a algún nodo del árbol (Ver Figura N° 13).

```
Entrada G y R
Se crea un fila de prioridad Q
Se incrusta R en Q
Mientras Q no esté vacío
         Se saca un elemento de la fila y se denomina u
         Si u == VISITADO se saca otro elemento de Q
         Se marca u como visitado
         Para cada v' adayacente a u
       Si v'== NO VISITADO
       Relajación {u, v, w}

Relajación (actual, adyacente , peso)
Si distancia [ actual ] + peso < distancia [ adyacente]
Distancia [adyacente]  = distancia [ actual ] + peso
Se agrega adyacente a Q

Salida: CMC entre dos nodos.
```

Figura N° 13. El Pseudocódigo 3 describe el algoritmo Dijkstra
(Fuente: Campbell, 2017)

2.1.6 Teoría de formación de clústeres

El ser humano tiene la capacidad intrínseca de agrupar los objetos a su alrededor guiados por las características que comparten. Los clústeres se definen como conjuntos de elementos con características similares entre sí o en pocas palabras son conglomerados de elementos que se forman por sus propiedades semejantes. (Kaufman & Rousseeuw, 1990)

Para clasificar clústeres podemos partir del resultado que se forma al principio de su creación. Las clasificaciones son:

a) Jerarquía de clústeres. – La formación de clústeres se basan en dos principios, la anidación y desanidación. La anidación requiere partir de un singletón hasta llegar a un subgrafo que contenta a todos los elementos del grafo. La desanidación comienza con una matriz donde estén todas las observaciones del clúster y como principio tenemos al algoritmo de desagrupación que empieza con el proceso de partición hasta alcanzar al singletón.
b) Partición de clústeres. - Es la clasificación más simple de todas y consiste que particionar el conjunto, pero dejando a cada elemento de este en un subconjunto disjunto.

Por otro lado, tenemos a los clústeres excluyentes que se clasifican en:

a) Clústeres exclusivos. – tenemos a los clústeres que contienen elementos en un subconjunto que no tienen otros subconjuntos de la misma partición. Como su nombre lo menciona, contienen elementos exclusivos.
b) Clústeres no disjuntos. – son los clústeres en donde un elemento puede existir en diferentes clústeres al mismo tiempo.
c) Clústeres borrosos. - aquí los elementos pueden o no pertenecer a un subconjunto, su peso se relaciona con su estado después de la partición. Cuando el valor es 1 significa que pertenece a un clúster y cuando es 0 significa que no pertenece a ningún clúster. (Abonyi & Balázs, 2007)

Finalmente tenemos a la clasificación en función de las metas de cada partición de clústeres.

a) Clústeres basados en prototipos. Se forman alrededor de un centro como lo puede ser un medoide o centroide, los cuales son considerados como prototipos.
b) Clústeres basados en grafos. – son los más importantes para el presente tema de investigación, se forman a partir de grafos que contienen aristas y nodos.
c) Clústeres basados en densidad. – se localizan las zonas con mayor grado de concentración de elementos.

Típicamente para el Clustering Jerárquico, los conglomerados se representan mediante diagramas o mejor conocidos como dendrogramas (Ver Figura N°14). El

eje Y representa los valores de disimilaridad de los conglomerados de elementos que van en el eje X.

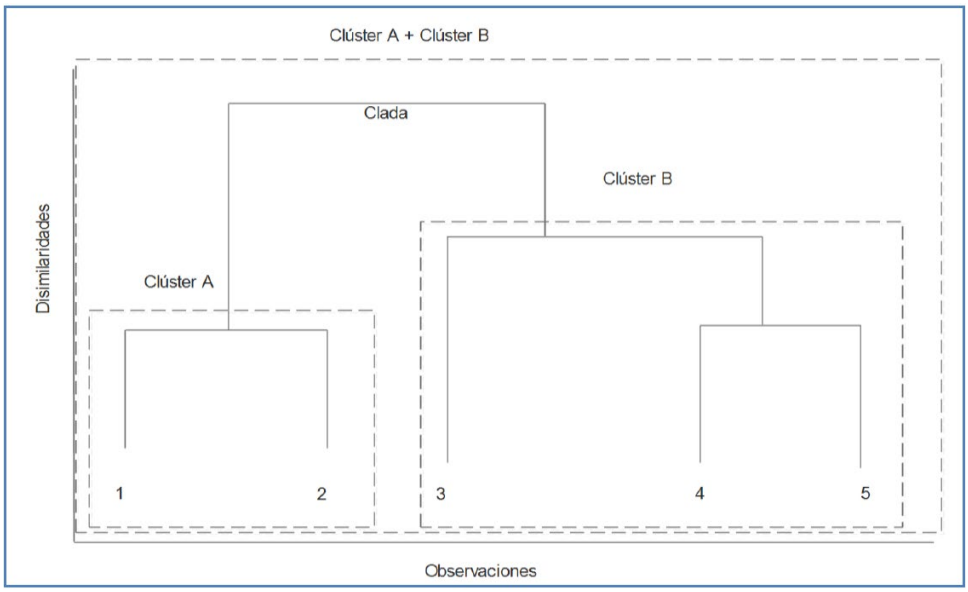

Figura N° 14. El Pseudocódigo 4 describe el algoritmo Prim
(Fuente: Campbell, 2017)

2.1.6.1 Métodos de evaluación de la calidad de clústeres

Existen diferentes tipos de inconvenientes cuando se trata de la partición de grafos en clústeres y uno de estos problemas es en cuantas partes se debe dividir el grafo. Para un escenario en donde se aplique el Clustering Jerárquico es particularmente importante ya que se genera una partición junto a una jerarquía de clústeres. Para las técnicas de clústeres no jerárquicos el número de particiones es no a priori y uno tiene que determinar cuál será el óptimo para un caso en específico.

Para poder determinar la cantidad de clústeres más adecuado se formularon un conjunto de medidas que permitan resolver el problema ya mencionado y estas son las medidas internas y externas que pasaremos a describir brevemente.

- Medidas Internas

Son un conjunto de parámetros (Ancho de Silueta, conectividad, y el índice de Dunn) que se estiman a partir de 3 características fundamentales: la conectividad,

la compactación y la separación. La conectividad es la explicación de porque algunos nodos pertenecen a un clúster, pero a otros no, la compactación se fundamenta en la homogeneidad de los clústeres formados y la separación que, como su nombre lo dice, se basa en la separación de los clústeres. (Brock, Pihur, & Datta, 2008)

La conectividad se basa en medir la correlación global de cada uno de los elementos contenidos en un mismo clúster. (Handl, Knowles, & Kell, 2005) En la ecuación 6 se expresa el valor de la conectividad en donde la comparación entre un elemento i de un clúster con otro elemento j de un clúster vecino producirá un valor igual a $1/j$. La cantidad de vecinos seleccionados será L y la cantidad de elementos de un clúster serán N. El valor de la ecuación 6 varía entre cero e infinito. En donde el valor más optimo será el más cercano a cero.

$$Conn(\sigma) = \sum_{i=1}^{N} \sum_{j=1}^{L} \chi_{i,nn_{i(j)}} \qquad \text{(Ecuación 6)}$$

El Ancho de Silueta (AS) se calcula mediante la combinación de la característica de cohesión y separación. Resulta en un promedio de valores AS correspondiente a cada conjunto de elementos en el interior de un clúster. El valor del AS estará entre 1 y -1, siendo este último el peor valor esperado. Se calcula mediante la ecuación 6.

$$AS(i) = \frac{b_i - a_i}{\max(b_i, a_i)} \qquad \text{(Ecuación 7)}$$

Al dividir la distancia mínima entre dos clústeres y la distancia máxima dentro de los clústeres (diámetro de un clúster) se obtiene el índice de Dunn. Se calcula mediante la Ecuación 8. (Dunn, 1974)

$$D(C') = \frac{\min\limits_{C_k, C_l \in C', C_k \neq C_l} \left(\min\limits_{i \in C_k, j \in C_l} dist(i,j) \right)}{\max\limits_{C_m \in C'} diam\,(C_m)} \qquad \text{(Ecuación 8)}$$

En donde $diam(C_m)$ es la máxima distancia entre dos observaciones en un mismo clúster C_m . Tomando en cuenta que la distancia dentro y entre clústeres diferentes puede ser infinita da como resultado que el valor de $D(C')$ este entre cero e infinito. A mayor valor mejor será la partición que se tenga.

2.1.7 Software utilizado

Para el análisis de clusterización se propone utilizar el software R, con su paquete ClValid que sirve para calcular la cantidad de clústeres óptimo para una base de datos dada. Este packages internamente aplica todas las medidas anteriormente descritas (medidas internas y externas). (R Core, 2015) El paquete ClValid también permite aplicar otras técnicas de clustering como lo son: Jerárquico, k-means, PAM, CLARA, DIANA, etc. A continuación, haremos una breve descripción de dichas técnicas con excepción del método jerárquico el cual será descrito en la Subsección 2.3.

- k-means: Este método requiere ingresar un número k (número de centroides) que actúan como puntos en donde los demás elementos se van agrupando para que cada elemento del conjunto pertenezca a un clúster. Cuando todos los elementos ya formen parte de un clúster los centroides se vuelven a reasignar. Este proceso se repite hasta alcanzar un cierto grado de estabilidad de los clústeres. (Campbell, 2017)
- PAM: Para este método también es necesario ingresar el número de clústeres deseados a priori. Este algoritmo a diferencia de anterior trabaja con medoides, pero el proceso que continua es similar a kmeans. En la última fase del algoritmo se mejora la calidad de las particiones intercambiando los elementos seleccionados con el resto de los elementos. (Campbell, 2017)
- CLARA: Cuando se tiene una base de datos de gran envergadura pero que no es tan grande como para aplicar Big Data se puede utilizar el algoritmo Clara (Kaufman & Rousseeuw, 1990). Es un algoritmo basado en los principios de k-medoides, pero para una base de datos más grande, este funciona agrupando una muestra de la base datos para posteriormente designar los elementos que restaron a los nuevos clústeres creados. (Campbell, 2017)

2.2 Grafos de redes sociales y detección de comunidades

Una sociedad es un conjunto de individuos que cohabitan un mismo territorio regidos por un determinado esquema de organización, estos a su vez interactúan y comparten lazos económicos, culturales o políticos. Partiendo de este concepto se puede explicar que las redes sociales virtuales son grafos especializados en representar la interacción de un conjunto de individuos. Un individuo en una sociedad cumple un rol y genera un aporte. En relación con eso, un individuo puede ser un profesor dictando una clase en una universidad, una abeja polemizando flores o un nodo en una red de agua potable. Considerando el último ejemplo, el aporte de un nodo podría ser la cota, demanda, coordenadas geográficas o carga hidráulica. (Campbell, 2017)

En los últimos 30 años, con el desarrollo de la computación e informática, se han desarrollado herramientas de visualización y análisis muy complejas como:

- Gephi (Bastian, Heyman, & Jacomy, 2009)
- Pajek (Bataglj & Mrvar, 2002)
- Graphviz (Bilgin, Ellson, Gansner, & North, 2017)
- Librería Igraph en R (Csardi & Nepusz, 2017)

Siendo las herramientas más comercializadas la librería Igraph en R y Gephi por su gran versatilidad, llegando a utilizarse en campos tan diversos como la biología, medicina, estadística e ingeniería (Csardi & Nepusz, 2017; Bastian, Heyman, & Jacomy, 2009)

Dicha iniciativa se centra en el interés por estudiar la forma dentro de las redes sociales para comprender la función y organización de los elementos que conforman la comunidad. En particular, el concepto que más interesa es la detección de comunidades o conglomerados dentro de estas redes. El reconocimiento de estas comunidades es crucial para el desarrollo del presente trabajo, ya que permitirá identificar módulos funcionales que no son evidentes a simple vista. Al analizar las conexiones, podríamos encontrar elementos que están fuertemente conectados con otros, grupos que interactúan solo entre sí, o incluso

elementos aislados que no están conectados a ninguno. Además, podríamos identificar aquellos que actúan como puentes entre diferentes grupos, lo que enriquece nuestra comprensión de la estructura subyacente de la red.

Los subgrupos en redes sociales se forman bajo sus propios criterios y ello les brinda una identidad. Para facilitar la comprensión de la idea de comunidades en redes sociales se trabaja en su expresión matemática. De esta manera, hoy en día, es común encontrar aplicaciones de la idea de comunidades en campos como el financiero, el inter, la minería de datos, etc. Una comunidad es un subconjunto de elementos formados por la presencia de una mayor densidad de cantidad de enlace (Steinhaeuser & Chawla, 2010). Así es como tendremos la idea de representar una red social mediante un grafo que a su vez tendrá subgrafos (comunidades o subgrupos). (Campbell, 2017)

Existen cuatro criterios para describir el concepto de estructura comunitaria y estas son:

a) Mutualidad de enlaces: en donde los pares de miembros en todos los subconjuntos se escojan entre sí.
b) Cercanía o accesibilidad: en donde todos los miembros estén conectados con el resto de los miembros. Aquí es donde aparece la importancia de las conexiones entre miembros de la comunidad.
c) Frecuencia de enlaces: se refiere a un subgrafo que contiene n nodos, los cuales son adyacentes a referido n-k nodos, donde k es mayor que cero.
d) Frecuencia relativa de enlaces: aquí se utilizan las densidades de los enlaces en cada subgrafo para hacer una comparación entre sí. Esto hace que la densidad sea relativa con respecto a los subgrupos más densos.

En resumen, la detección de comunidades aparte de considerar las características de los nodos también tiene presente las características de los enlaces de cada individuo, es decir que las comunidades resultantes se forman tanto de características de los individuos como de la interacción entre ellos. Aquí es donde

se encuentra la gran diferencia con el clustering tradicional que solo consideran las características de los nodos.

2.2.1 Concepto de Modularidad

En la actualidad el indicador de Modularidad es la medida más usada para evaluar la calidad de la partición de redes sociales en subgrupos. Fundamentalmente es la comparación de densidades dentro y fuera de cada munidad del grafo contra la densidad esperada si los enlaces fueran distribuidos de forma aleatoria en el grafo. Los autores parten de la base en que la distribución de grados es una propiedad intrínseca de la red y por ello, sugieren un prototipo nulo en función a este principio. Es lo que a este modelo estadístico se le otorga un prototipo nulo que especifique que es lo que se espera del azar (Newman, 2003). Para calcular la Modularidad (Q) de una partición determinada (P) se aplica la ecuación 9:

$$Q(P) = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{d_i d_j}{2m} \right] \sigma(C_i, C_j) \qquad \text{(Ecuación 9)}$$

En donde:

- A_{ij}: Es la matriz de adyacencia del grafo
- d_i: Es el grado del nodo i
- d_j: Es el grado del nodo j
- m: Número de aristas o links
- $\frac{d_i d_j}{2m}$: Según el modelo nulo, corresponde al número esperado de aristas entre los nodos i y j
- La función δ cambia a un valor igual a uno si es que los nodos i y j están en la misma comunidad y caso en caso contrario es igual a cero.

En un principio la definición de modularidad requiere que cada vértice pertenezca a una sola comunidad (esto excluye el solapamiento). Según Fortunato & Barthélemy (2007) demostraron matemáticamente que la optimización del coeficiente de modularidad sufre de un límite de resolución. Esto quiere decir que es incapaz de detectar comunidades de un cierto tamaño menor que un límite que viene determinado por el número de enlaces del grafo y su patrón de unión.

2.2.2 Metodología de sectorización basada en la Detección de Comunidades en Redes Sociales

La metodología de sectorización basada en la detección de comunidades en redes sociales se centra en dividir las redes de abastecimiento de agua potable en sectores utilizando criterios matemáticos para una gestión eficiente. Esta metodología utiliza algoritmos de detección de comunidades en redes sociales, como Clustering Jerárquico, Algoritmo de Detección Multinivel y Detección de Comunidades a través de Caminos Aleatorios. Se busca optimizar la red definiendo sectores y puntos de abastecimiento, considerando aspectos económicos y la detección de fugas futuras. (Ver Figura N° 14-A).

A su vez, se emplean técnicas de optimización heurísticas como Algoritmos Genéticos, Optimización de Enjambre de Partículas (Algoritmo Louvain) y Optimización de Enjambre de Agentes (Algoritmo Waltrap) para mejorar la sectorización y reducir fugas. Esta metodología equilibra la calidad del suministro con aspectos económicos, optimizando el coste/beneficio de la sectorización. (Ver Figura N° 14-B).

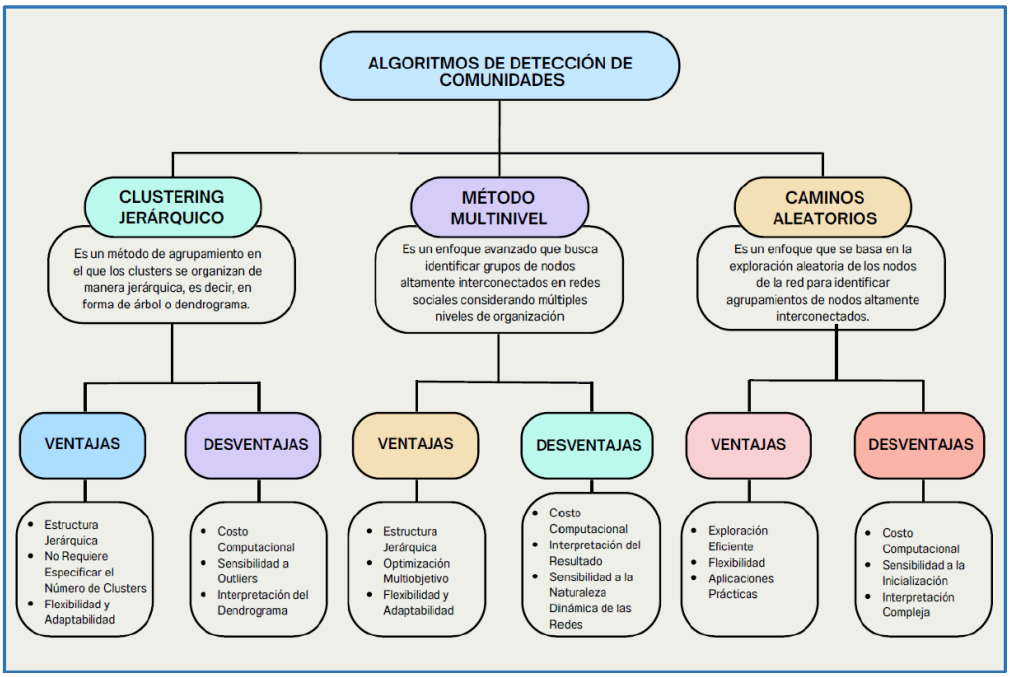

Figura N° 14-A. Métodos de detección de comunidades en redes sociales

(Fuente: Elaboración Propia)

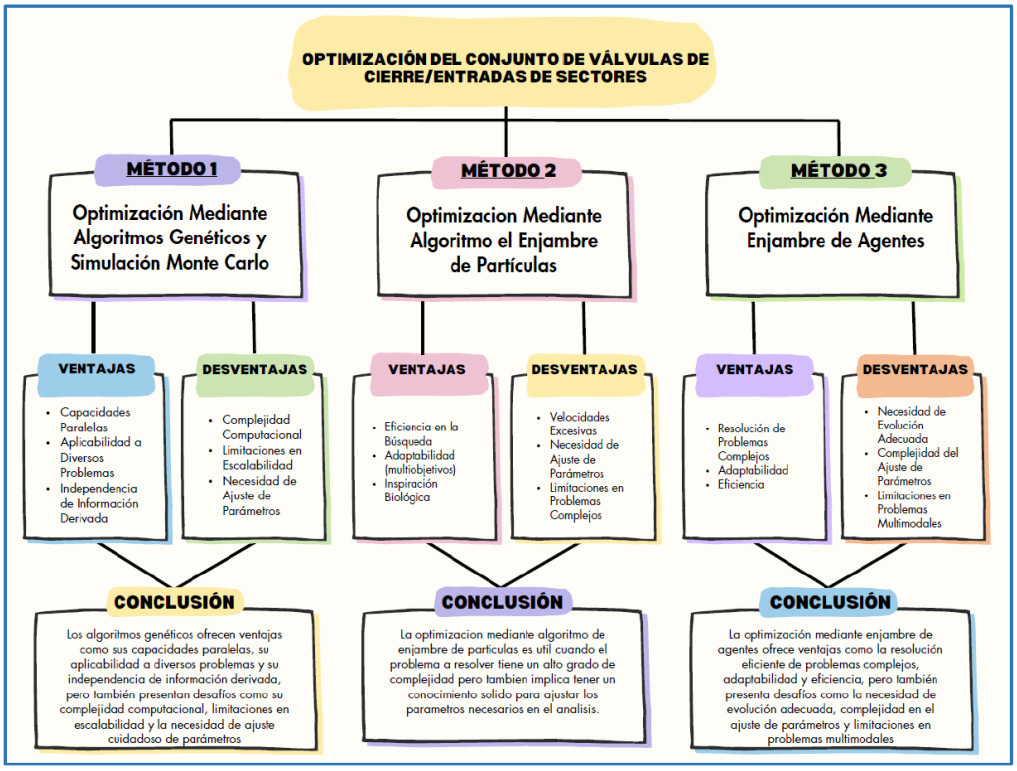

Figura N° 14-B. Métodos de Optimización del conjunto de válvulas de cierre/entrada de sectores
(Fuente: Elaboración Propia)

El clustering jerárquico es un método de análisis de datos que busca formar clústeres con una jerarquía en cada grupo. El clustering jerárquico pertenece al grupo de métodos de aprendizaje automático no supervisado los cuales comparten el objetivo general de exploración de datos. Dicho método forma arboles binarios que, sucesivamente se fusionan unos a otros en función del valor de similaridad. (Han, Kamber, & Pei, 2006) La diferencia con otros métodos que forman clústeres es que no es necesario ingresar un número a priori de clústeres para la partición, en cambio, el número de clústeres se estima a través del dendrograma resultante. (Manning, Raghava, & Schutze, 2008) A continuación, se mostrarán los pasos a seguir para aplicar el método de Clustering Jerárquico.

2.2.3 Etapas del Clustering Jerárquico Aglomerativos

2.2.3.1 Etapa 1: Matriz de Disimilaridad

Comenzando con un conjunto finito de datos discretos que cuentan con un conjunto de características de variables aleatorias. Se forma la matriz $n * n$ en donde n es el número de caso del conjunto de datos (en nuestro caso es el número de nodos de la red). Con ayuda de la matriz se evalúa la disimilaridad en parejas

del conjunto de datos usando las diferentes medidas métricas (Manhattan, Euclidiana y Gower) las cuales tendrán una breve descripción más adelante. En RDAP los elementos (nodos y tuberías) cuentan con más de una variable o característica, lo cual indica que harán comparaciones entre las características individuales.

- Distancia euclidiana. – resulta de la raíz cuadrada de la suma de las diferencias de los valores de disimilaridad al cuadrado.

$$d_E(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \qquad \text{(Ecuación 10)}$$

$$d_E(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad \text{(Ecuación 11)}$$

- Distancia Manhattan. – distancia entre cada par de casos siguiendo un camino tipo malla.

$$d = \sum_{i=1}^{n} |X_i - Y_i| \qquad \text{(Ecuación 12)}$$

- Distancia Gower. – distancia entre una base de datos mixto, puede ser la combinación de variables cualitativas y cuantitativas.

$$d_{ij}^2 = 1 - s_{ij}$$

$$s_{ij} = \frac{\sum_{h=1}^{p1} \frac{1 - |x_{jh} - x_{jh}|}{G_h} + a + \alpha}{p_1 + (p_2 - d) + p_3} \qquad \text{(Ecuación 13)}$$

Donde:

- p_1 = # de variables cuantitativas continuas
- p_2 = # de variables cualitativas
- α = # de coincidencias
- G_h = rango (o recorrido)

- a = # de coincidencias (1,1)
- d = # de coincidencias (0,0)

2.2.3.2 Etapa 2: Aglomeraciones de Casos en Clústeres

Una vez un par de clústeres haya dejado de ser singletones en función a la comparativa entre variables se procede en volver a hacer la comparativa entre el primer clúster formado con las demás variables. La característica utilizada para hacer las comparaciones depende del método aglomerativos que se haya utilizado al principio. Entre los más comunes se cuenta con los métodos de agrupación por promedio, individual, completa y la basada en centroides. Para hondar más afondo sobre los métodos de agrupación procedemos a explicar cada uno de ellos mencionados anteriormente y partiremos de una matriz n*n que contiene valores de disimilaridad entre pares de casos (matriz de disimilaridad), se procede a ubicar el par con la mínima distancia en base al método seleccionado de los cinco arriba mencionados. Es así como se forma el primer clúster que a continuación se compara con los elementos restantes. Se actualiza la matriz conforme se formen nuevos clústeres y así sucesivamente hasta alcanzar a todos los elementos en un mismo clúster. (Campbell, 2017)

a) Método de Aglomeración Promedio

   El método de Aglomeración Promedio se basa en comparar el promedio de los elementos de un clúster entre el promedio de los elementos pertenecientes a un clúster distinto.

$$D_{KL} = \frac{1}{n_k n_l} \sum_{ie C_K} \sum_{ie C_l} d(x_i x_j) \qquad \text{(Ecuación 14)}$$

   En donde $d(x_i x_j)$ representa la distancia entre 2 elementos de un mismo clúster y k, l son diferentes clústeres.

b) Método de Aglomeración Centroide

   Para el método de Aglomeración Centroide es necesario utilizar la semejanza entre sus centroides, estos a su vez son los vectores de medias de las variables medidas sobre los individuos del clúster (Cuadras, 2014).

El centroide de cada nodo corresponde a la distancia Euclidiana para cada par de nodos de los sectores. (Campbell, 2017)

$$D_{KL} = \left|\left|\text{prom}X_k - \text{prom}X_l\right|\right|^2 \qquad \text{(Ecuación 15)}$$

c) Método Ward o de Mínimo Incremento de Suma de Cuadrados

Para este método en cada etapa se unen los dos clústeres para los cuales se tenga el menor incremento en el valor total de la suma de los cuadrados de la diferencia dentro de cada clúster.

$$E = \sum_{i=1}^{n_k}\sum_{j=1}^{n}({x_{ij}}^k - m^k)^2 \qquad \text{(Ecuación 16)}$$

En donde:

- ${x_{ij}}^k$: en un clúster k se tiene a elemento j con una variable es la variable i
- m^k : centroide del clúster k
- E_k : sumatoria del cuadrado de los errores del clúster k

2.2.3.3 Etapa 3: Representación del Clúster Jerárquico Aglomerativo

El proceso jerárquico para la construcción de un dendrograma se puede ver con mayor detalle en la Subsección 2.1.4.1

2.2.3.4 Etapa 4: Selección de Métodos a Emplear

Como se mencionó anteriormente se tienen diferentes formas de medir la disimilaridad (métricas, Euclidiana, Manhattan, Gower) y a su vez diferentes maneras de unir los clústeres. Para una misma base de datos se tienen diferentes resultados dependiendo el camino y método que uno escoja. Para elegir el método de medida del valor de disimilaridad y la unión de los clústeres se introduce el concepto de Coeficiente de Correlación Cofenética (CCCF) (Everitt, Landau, Leese, & Stahl, 2011), Este coeficiente es muy reconocido por los estadísticos para medir la fiabilidad en la que un dendograma conserva las distancias intra-

clúster o, dicho de otra forma, el grado de representación de las características de un conjunto de datos.

El CCCF se define como la correlación entre la matriz de disimilaridad inicial y la matriz ultramétrica, que guarda en su interior todos los pares de casos con el valor de disimilaridad utilizado para la agrupación inicial. (Sokal & Michener, 1958)

La ecuación 17 muestra la forma de calcular el índice CCCF.

$$r_{xy} = \frac{\sum xy - \left({}^{1}/_{n}\right)(\sum x)(\sum y)}{\{[\sum x^2 - \left({}^{1}/_{n}\right)(\sum x)^2][\sum y^2 - \left(\frac{1}{n}\right)(\sum y)^2]\}^{{}^{1}/_{2}}} \qquad \text{(Ecuación 17)}$$

En donde x representa los valores de disimilaridades en pares S_{ij} de la matriz de disimilaridad, n el número de casos de la base de datos e y los valores de altura de la matriz de disimilaridad. (Campbell, 2017) En la ecuación 18 se puede apreciar que el valor del CCCF varía entre 0 y 1, pero se ha llegado a la conclusión que si el valor es mayor a 0.8 se considera aceptable la técnica de clustering jerárquico. (Romesburg, 2004)

$$0 < r_{xy} < 1 \qquad \text{(Ecuación 18)}$$

2.2.4 Ejemplo de Clustering Jerárquico

Partiendo de una base de datos X que contengan n casos, se inicia el proceso formando la matriz de disimilaridad de parejas de casos X_{ij}. Tendremos una matriz cuadrada nxn, cada valor de la matriz debe cumplir con la siguiente característica.

$$D = \left(X_{ij}\right), 1 \leq X_{ij} \leq n \qquad \text{(Ecuación 19)}$$

La primera partición será de la siguiente manera: $x = \{1\} + \{2\} + \cdots + \{n\}$. Esto quiere decir que los clústeres iniciales son los propios elementos individuales de los datos x. Se inicia la comparación de las características de los datos individuales encontrando un par con la menor disimilitud y, por ende, estos datos tendrán una mayor cercanía. Es así como encontraremos el primer par de datos conglomerados.

$$\{i\}\, U\, \{j\} \rightarrow \{i, j\} \qquad \text{(Ecuación 20)}$$

El valor de disimilaridad del primer conglomerado es el valor inicial de la matriz ultramétrica y representa la altura del enlace S'_{ij}. El siguiente paso es hacer la comparación del primer conglomerado con el resto de los datos de la matriz para seguir agrupándolos.

$$S'_{k(ij)} = f\left(S_{k(i)}, S_{k(j)}\right) \text{ en donde } k \neq i \text{ o } j \qquad \text{(Ecuación 21)}$$

En donde $S_{k(ij)}{}'$ representa el valor nuevo de la matriz $n * n$ que se compara con todos los elementos (k) versus el clúster $\{i, j\}$. Para obtener el valor mencionado se selecciona una función f (promedio, centroide, completa, individual). (Campbell, 2017)

Una vez culminado el paso anterior se repetirá la selección del menor valor de la matriz de disimilaridad y concluirá en el momento que todas las observaciones estén contenidas en un clúster igual a todo el conjunto de datos.

$$x = \{1, 2 \ldots, n\}. \qquad \text{(Ecuación 22)}$$

Para poder comprender mejor los conceptos anteriormente explicados se muestra el siguiente grafo como ejemplo aplicativo. (Ver Figura N° 15)

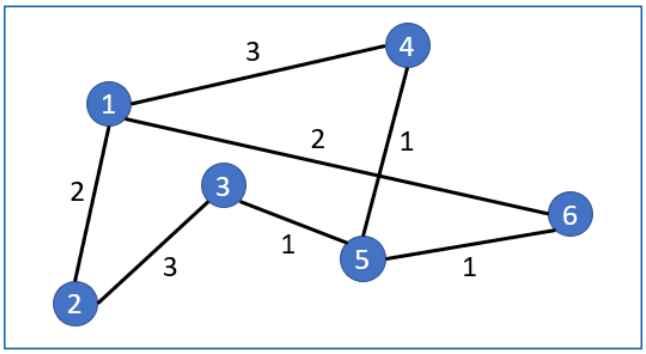

Figura N° 15. Representación típica del grafo ejemplo

(Fuente: Elaboración propia)

Tabla N° 4. Características del grafo ejemplo

ID	Característica 1	Característica 2
1	1	3
2	7	9
3	5	1
4	3	4
5	1	2

(Fuente: Elaboración propia)

Inicialmente, se tiene una primera partición x_1 en la que cada uno de los casos constituye un clúster.

$$x_1 = \{1\} + \{2\} + \{3\} + \{4\} + \{5\} \qquad \text{(Ecuación 23)}$$

El cálculo de la disimilaridad por pares entre casos se aplicará la métrica de distancia Euclidiana. (Ver Tabla N° 5)

Tabla N° 5. Matriz de disimilaridad del grafo ejemplo

ID	1	2	3	4	5
1	0.00				
2	8.49	0.00			
3	4.47	8.25	0.00		
4	2.24	6.40	3.61	0.00	
5	1.00	9.22	4.12	2.83	0.00

(Fuente: Elaboración propia)

De la matriz de disimilaridad resultante se resalta que los casos 1 y 5 son los que tienen menor valor de disimilaridad (altura $S'_{ij} = 1.00$), es así como obtenemos el primer clúster formado por {1, 5}. El siguiente paso es comparar el clúster formado con el resto de los casos $(k = 2, 3, 4)$ mediante el promedio.

Por tanto, se compara este nuevo clúster con el resto de los casos $(k = 3, \dots, n)$ a través de una función promedio.

$$\{1-5\}, 2 = [1-2], [5-2] = Promedio\ (8.49, 9.22) = 8.86$$

$$\{1-5\},3 = [1-3],[5-3] = Promedio\ (4.47, 4.12) = 4.30$$

$$\{1-5\},4 = [1-4],[5-4] = Promedio\ (2.24, 2.83) = 2.54$$

En la tabla N° 6 se muestra la matriz resultante $n * n$ de disimilaridad en donde el mínimo valor es 2.54:

Tabla N° 6. Matriz de disimilaridad con cuatro clústeres

ID	{1-5}	2	3	4
{1-5}	0.00			
2	8.86	0.00		
3	4.30	8.25	0.00	
4	2.54	6.40	3.61	0.00

(Fuente: Elaboración propia)

Como se puede apreciar en la tabla N° 6 el menor valor de disimilaridad (altura $S'_{ij} = 2.54$) está entre el clúster $\{1-5\}\ y\ 4$. Se comparan ahora el nuevo clúster con los demás casos.

$$\{1-5-4\},2 = [1-5.2],[4-2] = Promedio\ (8.86, 6.40) = 7.63$$

$$\{1-5-4\},3 = [1-5,3],[4-3] = Promedio\ (4.30, 3.61) = 3.96$$

En la tabla N° 7 se muestra la matriz resultante $3 * 3$ de disimilaridad del grafo ejemplo.

Tabla N° 7. Matriz de disimilaridad con tres clústeres

ID	{1-5-4}	2	3
{1-5-4}	0.00		
2	7.63	0.00	
3	3.96	8.25	0.00

(Fuente: Elaboración propia)

El proceso se repite con el menor valor de disimilaridad de la tabla N° 7 (altura $S'_{ij} = 3.96$) está entre el clúster $\{1-5-4\}\ y\ 3$. Con el clúster resultante hacemos la comparación con el caso 3.

$$\{1-5-4-3\}, 2 = [1-5-4, 2], [3-2] = Promedio\ (7.63, 8.25) = 7.94$$

Finalmente se forma el clúster que contenga a todos los casos mostrado en la tabla N° 8:

Tabla N° 8. Matriz de disimilaridad con dos clústeres

ID	{1-5-4-3}	2
{1-5-4-3}	0.00	
2	7.94	0.00

(Fuente: Elaboración propia)

Finalmente, la matriz ultramétrica resultante se muestra en la Tabla N° 9

Tabla N° 9. Matriz ultramétrica del grafo ejemplo

ID	1	2	3	4	5
1	0.00				
2	7.94	0.00			
3	3.96	7.94	0.00		
4	2.54	7.94	3.96	0.00	
5	1.00	7.94	3.96	2.54	0.00

(Fuente: Elaboración propia)

2.3 Criterios hidráulicos para sectorización

En el presente trabajo se busca optimizar la posición de las válvulas cierre y reductoras de presión en una red hidráulica en un esquema de sectorización. En la presente sección se hará una descripción de los criterios hidráulicos que se deben tomar en cuenta.

2.3.1 Índice de Resiliencia

Según (Creaco, Fortunato, Franchini, & Mazzola, 2013) en una red hidráulica el cambio de demanda o caudal producida por el fallo de una tubería genera una pérdida de energía que perjudica al servicio brindado. En la situación en donde la demanda sea abastecida con una presión exacta, esta puede no ser la suficiente

al momento de sufrir un fallo en alguno de sus componentes. Es así como las redes hidráulicas deben contar con un exceso de energía para contrarrestar algún fallo en la red, garantizando la continuidad del servicio.

En principio una red hidráulica transfiere energía desde la fuente (o fuentes) a través del circuito conductor (conjunto de tuberías y nodos) a un usuario final. En tanto, la energía entregada desde la fuente no es igual a la energía recibida por el usuario. La pérdida de energía es debido a la fricción en las tuberías y accesorios. Todini (2000) planteo una serie de ecuación que describen la distribución de energía entregada y recibida en la red.

La potencia de entrada P_{inp} de la red es igual a la suma de las potencias entregadas a los usuarios finales P_{out} y la potencia de operación P_{int} (perdidas por fricción en las tuberías y accesorios). (Ver ecuación 24)

$$P_{inp} = P_{out} + P_{int} \qquad \text{(Ecuación 24)}$$

Otra forma de expresar P_{inp} en función de la potencia que es suministrada por los bombeos $(j = 1, \cdots, n_p)$ y embalses $(e = 1, \cdots, n_e)$. (Ver ecuación 25)

$$P_{inp} = \sum_{e=1}^{n_e} Q_e \times H_e + \sum_{j=1}^{n_p} P_j \qquad \text{(Ecuación 25)}$$

La potencia entregada es expresada como la magnitud real en donde se utiliza la altura piezométrica real H_j, otra forma de expresarla es mediante la altura piezométrica máxima H_j^* requerida para satisfacer una presión mínima. (Ver ecuación 26)

$$P_{out}^{\text{real}} = \sum_{j=1}^{n_n} Q_j * H_j \qquad \text{(Ecuación 26)}$$

$$P_{out}^{\text{máx}} = \sum_{j=1}^{n_n} Q_j * H_j^* \qquad \text{(Ecuación 27)}$$

Como la potencia de entrada es igual a la potencia operacional más la potencia de entrega, entonces se puede expresar mediante la resta de ambas potencias la potencia de entrega. Al poder expresar la potencia de entrega en términos de máxima y real altura piezométrica, la potencia operacional también puede ser expresada de la siguiente manera. (Ver Ecuación 29)

$$P_{int}^{\text{real}} = P_{inp} - P_{out}^{\text{real}}$$

$$P_{int} = P_{inp} - P_{out} \qquad \text{(Ecuación 28)}$$

$$P_{int}^{\text{máx}} = P_{inp} - P_{out}^{\text{máx}}$$

Creaco et al., (2013) presento un índice de resiliencia (Ir) para evaluar la eficiencia energética de las RDAPs. Este índice describe la capacidad de una red hidráulica a reaccionar y superar estados atípicos (fallos o fugas). En términos generales se calcula restando la unidad con la relación de la potencia operacional real en la red con respecto a la potencia operacional máxima requerida para satisfacer una presión mínima.

$$Ir = 1 - \frac{P_{int}^{\text{real}}}{P_{int}^{\text{máx}}} \qquad \text{(Ecuación 29)}$$

Luego, sustituimos la potencia operacional real y la potencia operacional máxima para satisfacer una presión mínima con la Ecuación 31 y 32 respectivamente.

$$Ir = 1 - \frac{[\sum_{i=1}^{n_e}(Q_e * H_e)_i + \sum_{i=1}^{n_p} P_i] - \sum_{j=1}^{n_n} Q_J * H_J}{[\sum_{i=1}^{n_e}(Q_e * H_e)_i + \sum_{i=1}^{n_p} P_i] - \sum_{j=1}^{n_n} Q_J * H_j^*} \qquad \text{(Ecuación 30)}$$

$$Ir = 1 - \frac{\sum_{j=1}^{n_n} Q_J * (H_J - H_J^*)}{[\sum_{i=1}^{n_e}(Q_e * H_e)_i + \sum_{i=1}^{n_p} P_i] - \sum_{j=1}^{n_n} Q_J * H_j^*} \qquad \text{(Ecuación 31)}$$

Es así como se pudo conocer que tan confiable es una red hidráulica ante un fallo en alguno de sus elementos. Un valor optimo es igual a 0.5. Por otro lado, si el valor es menor entonces estaremos encontrando una red poco confiable y que no responda bien a un fallo en alguno de sus elementos, caso contrario si el valor es

cercano a la unidad ya que estaríamos en presencia de una red sobredimensionada.

El índice de desviación del índice de resiliencia propuesto por Di Nardo et al. (2013) permite evaluar la disminución del índice de resiliencia con respecto a su estado inicial antes de un esquema de sectorización. Dicho índice se calcula mediante la Ecuación 33. (Di Nardo, Di Natale, Santonastaso, Tzatchkov, & Alcocer Yamanaka, 2013)

$$I_{rd} = \left(1 - \frac{I^*_r}{I_r}\right) \times 100 \qquad \text{(Ecuación 32)}$$

En donde:

- I_{rd} : desviación del índice de resiliencia
- I^*_r : índice de resiliencia de la RDAP con un esquema de sectorización
- I_r : índice de resiliencia de la RDAP con el esquema original

2.3.2 Uniformidad de Presiones

Según Araque y Saldarriaga (2005), al minimizar el índice de resiliencia que representa la relación entre la energía disipada con la configuración inicial dada respecto a la energía optima disipada, se logra el estado de uniformidad de presiones. La energía optima disipada se puede definir como la cantidad de energía de puede disiparse en cada una de las tuberías de la RDAP que la conforman. El grado de uniformidad de presiones en una RDAP está definida en la Ecuación 34.

$$CU = \frac{\sum_{j=1}^{n} P_j}{n * \text{máx}[P_j]} \qquad \text{(Ecuación 33)}$$

En donde:

- n : número de nodos de la red
- P_j : presión en cada uno de los nodos de la red

Analizando el coeficiente de uniformidad en el contexto del Perú, se identifica la existencia de una cantidad de RDAP antiguas que no cuentan con un alto nivel de inversión o con escasez regular de agua este concepto de presión mínima de servicio no se cumple o es muy difícil implementarlo.

2.3.3 Uniformidad de características

La similitud de ciertas características de los nodos de una RDAP puede ser útil para evaluar la uniformidad de los sectores. Según Alvisi & Franchini (2014), se pueden utilizar la demanda de los nodos, el grado de similitud entre los nodos se calcula mediante la desviación estándar de la demanda total o también la cota de cada sector como se puede apreciar en la Ecuación 35. Como es de esperar, valores altos de la desviación estándar reflejarían una baja calidad de la sectorización tanto para demanda o cotas. Por un lado, se tiene a la demanda que en realidad no sería una característica fiable para la sectorización por diferentes motivos como, por ejemplo; encontrarse un nodo un valor atípico de demanda dentro de un sector (hospitales o industria dentro de una zona urbana) ya que esta es un valor arbitrario y que no necesariamente represente la realidad geográfica de las redes. (Alvisi & Franchini, 2014)

En cambio, emplear las cotas es más factible como indicador de uniformidad, ya que la realidad topográfica de una RDAP es mejor representada por la cota de los nodos. (Ver Ecuación 34)

$$DEC = \frac{\sqrt{\frac{\sum_{i=1}^{N_n}(q_i - q_{av})^2}{N_n}}}{q_{av}} \qquad \text{(Ecuación 34)}$$

En donde:

- DEC : desviación estándar de característica (cota o demanda)
- qav : promedio de la característica de referencia (cota o demanda)

2.4 Optimización mediante algoritmos genéticos y simulación Monte Carlo: Predicción de nuevas fugas mediante sectorización

2.4.1 Descripción de Algoritmos Genéticos

Los Algoritmos Genéticos (AG) son métodos adaptativos que están basados en el proceso genético de los organismos vivos, estos pueden ser usados para resolver problemas de optimización y búsqueda de soluciones. Según Darwin en su postulado de la selección natural (1859), a lo largo de las generaciones las poblaciones evolucionan en la naturaleza con los principios de la selección natural (la supervivencia del más fuerte).

Los Algoritmos Genéticos imitan el proceso de creación de soluciones, reproduciendo y descartando soluciones óptimas para solucionar problemas del mundo real. La evolución de las soluciones depende de la codificación de estas. A continuación, se hará una descripción detallada del proceso de funcionamiento de los Algoritmos Genéticos (Palisade, 2010):

a) Inicialización (población inicial): se crea una población inicial de manera aleatoria según el tipo que solución que necesitamos encontrar. Esta población depende únicamente del usuario, si bien una población inicial pequeña puede generar soluciones malas, una población inicial muy grande puede ser un peso computacional muy elevado. Es por lo que el usuario debe implementar una población inicial lo suficientemente grande como para dotar de diversidad y evitar una convergencia apresurada al algoritmo genético.
b) Evaluación de costes (aptitud): cada individuo de la población inicial es evaluado por una función objetivo para ver qué tan buena es la aptitud de este ante los requerimientos establecidos. La formulación de la función objetivo es de suma importancia ya que esta determinara la superficie de las posibles soluciones del problema.
c) Selección: este procedimiento selecciona a los individuos con los mejores resultados de la evaluación de aptitud con el fin de preservar sus mejores características en las próximas generaciones. Se basa en un criterio de evolución que genera una medida de calidad para los cromosomas en el contexto del problema. Las técnicas de selección más utilizadas son: la

técnica de la ruleta, la selección por rango y la selección de estado estacionario.

d) Cruzamiento: esta etapa de los Algoritmos Genéticos puede entenderse como la reproducción sexual en la naturaleza para crean nuevos individuos que hereden las mejores características de los padres. Lo que se espera lograr es obtener una población más adaptada a la función objetivo en la nueva generación. En general, esta etapa se considera como la fuerza conductora de los AG.

e) Mutación: en un proceso basado en la selección natural se necesita incluir algo de aleatoriedad la genética de la población, caso contrario, todas las soluciones estarían incluidas en la población inicial, lo cual generaría óptimos locales y nos desviarían del optimo global. La mutación marca pequeños cambios de manera aleatoria a los genomas individuales. Así es como se mejoran las habilidades de los AG para encontrar soluciones más cercanas al optimo global de un problema.

f) Iteración/convergencia: con la creación de una nueva generación, esta toma y se repite dicho proceso desde el paso 2 hasta alcanzar la condición de finalización o la cantidad de iteraciones establecidas inicialmente.

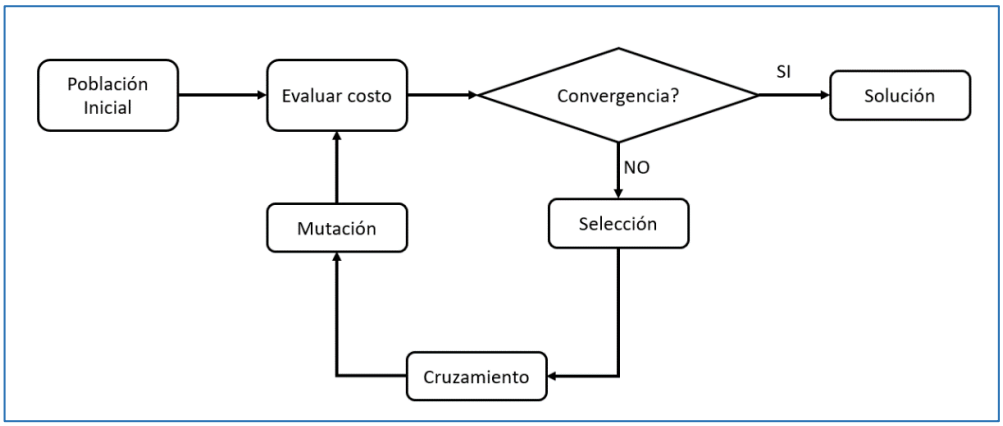

Figura N° 16. Representación típica del grafo 1

(Fuente: Elaboración propia)

2.4.2 Optimización Mediante Algoritmos genéticos y simulación Monte Carlo

La aplicación RiskOptimizer es una herramienta desarrollada por la compañía Palisade® como una extensión del programa Excel con fin de resolver problemas que la herramienta Solver no pueda analizar. Dicha aplicación está orientada a analizar e incluir factores que contienen incertidumbre durante el proceso de

optimización, realiza análisis de riesgo utilizando simulaciones para mostrar múltiples resultados posibles e indica que probabilidades hay de que dichos eventos sucedan.

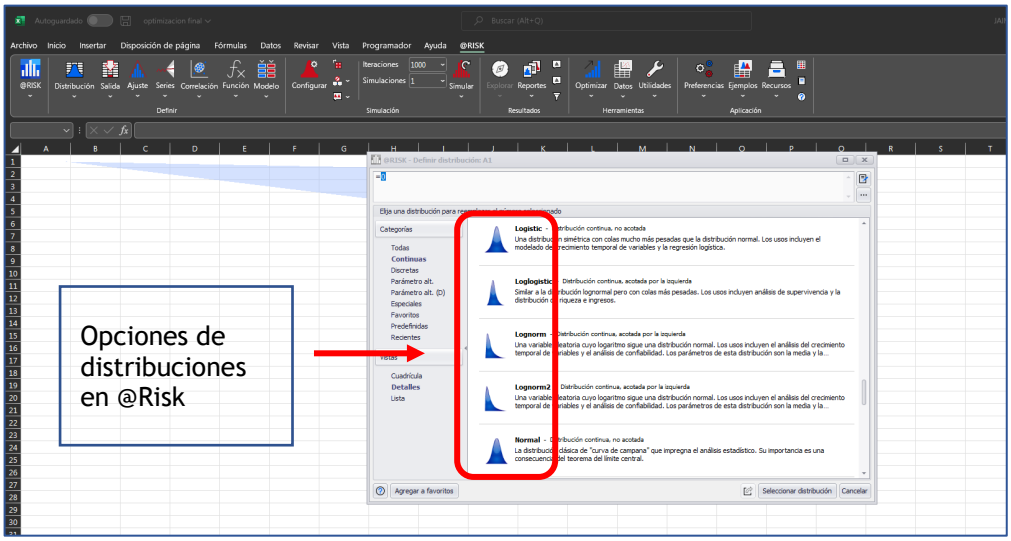

Figura N° 17. Aplicación RiskOptimizer en Excel
(Fuente: Elaboración propia)

2.4.2.1 Descripción del Método de Simulación Monte Carlo

La simulación Monte Carlo es un conjunto de métodos basados en la generación de valores aleatorios, que se agregan como entrada en un problema dado. Los valores aleatorios se generan de forma artificial tratando de emular el proceso de muestreo en una población inicial. Es así como un problema se evalúa como el resultado de una suposición y el resultado estadístico de repetir este experimento varias veces con valores seleccionados al azar partiendo de un rango probabilístico preestablecido.

El método de la Simulación Monte Carlo se basa en la Ley de los Grandes Números, el cual nos indica el resultado de realizar el mismo experimento una gran cantidad de veces. Esta Ley establece que el promedio de todos los resultados generados de una gran cantidad de iteraciones debe ser próximo al valor esperado y mientras la cantidad de iteraciones crece, la diferencia entre el valor esperado y el promedio se va acortando. Se puede inferir que dicha Ley es válida cuando la cantidad de iteraciones muy grande.

En la teoría de las probabilidades la ley de los grandes números se divide en dos versiones; Ley Fuerte de los Grandes Números y la Ley Débil de los Grandes Números. No obstante, ambas comparten el mismo concepto, el cual consiste que el promedio de un número de iteraciones converge a un valor esperado. (Ver Ecuación 35 y 36)

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_1) \quad \text{(Ecuación 35)}$$

$$\bar{X}_n \to \mu \; para \; n \to \infty \quad \text{(Ecuación 36)}$$

En donde:

- $\overline{X}_n$: promedio de un número n de iteraciones realizadas
- μ : valor esperado

El método de SMC puede dar como resultado lo que puede pasar y la probabilidad de que pase un evento. A diferencia de los análisis determinísticos, este tiene la ventaja de poder definir que variables tienen mayor peso sobre un problema. Esto resulta importante cuando se tienen que realizar análisis adicionales en donde destaca su simplicidad, escalabilidad y flexibilidad. El método de SMC es capaz de reducir problemas complejos en un conjunto de acontecimientos básicos representados por una serie de iteraciones y al emplearlo con modelos estocásticos ayuda a los algoritmos a escapar de óptimos locales en búsqueda del optimo global.

2.4.3 Simulación Monte Carlo en Redes de Abastecimiento de Agua Potable para detectar futuras roturas.

En el desarrollo de modelos hidráulicos de RDAPs existe un alto grado de incertidumbre con respecto a parámetros cuyo valor es difícil de estimar tales como la cantidad de posibles roturas, el volumen de agua potable que se pierde por diferentes tipos de roturas, el tiempo que demora la reparación de roturas o la frecuencia de inspección de una red para detectar roturas.

En el año 2007, se planteó una forma de predecir la cantidad de roturas esperadas (Thornton & Lambert., 2007) y la reducción de fugas al gestionar presiones en una red hidráulica que se basa en 110 sectores de 10 países diferentes. Una forma de poder calcular la cantidad de perdidas inevitables en una red de agua se formuló el Modelo Conceptual propuesto por Thornton & Lambert (2006, 2007) en donde se establece las Pérdidas Reales Inevitables. (Ver Ecuación 37)

$$UARL(l/día) = \left(1.8 * L_m + 0.8 * N_c + 25 * L_p\right) * P \quad \text{(Ecuación 37)}$$

En donde N_C es la cantidad de conexiones; L_m es la longitud de LC; P representa la presión en cada subsector y L_p es la longitud de las redes de distribución. Las consideraciones requeridas para aplicar la Ecuación 37 es tener una red con un alto grado de cuidado y mantenimiento, o en su defecto de una red relativamente con poco tiempo de funcionamiento. La cantidad de roturas en las LCP es 13/100 km/año (no se consideran la reparación de válvulas e hidrantes) y un valor esperado de 3/1000 conexiones/año, en donde tampoco se consideran las fugas pequeñas en los medidores y en caños (Lambert, Brown, Takizawa, & Weimer, 1999). Estas consideraciones fundamentan debido a que una red al tener poco funcionamiento soporta sobrepresiones y, por ende, la cantidad de roturas es muy baja.

Para simular la ocurrencia y detección de eventos de roturas, se debería establecer distribuciones de probabilidad triangular que estén en función de la longitud de las tuberías tanto de la red secundaria como la línea de conducción principal. Para ello también es necesario el porcentaje de detección que está conformado por un valor mínimo, un máximo y un valor más probable. Dicho porcentaje de detección puede asignarse mediante juicio de expertos tomando en consideración su longitud o de una base datos histórica de RDAP similares a la analizada. En el presente trabajo se consideró el porcentaje de detección igual tanto para fugas reportadas como fugas no reportadas debido a que estas últimas se localizan geográficamente por medios acústicos. (Lambert, Brown, Takizawa, & Weimer, 1999)

2.5 Representación de redes de abastecimiento de agua potable como grafos de redes sociales

Los modelos matemáticos de RDAP se han convertido en una herramienta indispensable en la gestión de estas. Por consiguiente, en la presente investigación se utilizará el modelo Epanet, creado por el departamento de seguridad de la Agencia de Protección Ambiental de EE. UU. EPA.

Los fundamentos básicos que maneja este software se componen de nodos (tanques, reservorios, extremos de tuberías) y enlaces (tuberías, bombas, cámara de válvulas); esto hace que la topología de las redes de agua descritas anteriormente.

Un modelo de Epanet se inicia estableciendo los nodos de consumo y abastecimiento (reservorios, tanques, etc.) como los vértices del grafo, y las tuberías, válvulas y bombas como las aristas. Una vez se asignan las características de los nodos (cota, demanda, coordenadas geográficas) y tuberías (diámetro, rugosidad, material, longitud), se procede a correr el programa y se obtiene la magnitud y dirección del caudal en el grafo. Como resultado se tendrá un dígrafo ponderado en donde los nodos contienen como como características: coordenadas geográficas, demanda, coeficiente de emisor, altura piezométrica, presión; y en las tuberías: diámetro, rugosidad, caudal, longitud y pérdidas de carga por fricción.

En la presente tesis, los dígrafos son representados en el lenguaje de programación R y con ayuda de la interfaz Rstudio. El Epanet nos permite exportar una red como un archivo INP que contiene dos tablas: la primera tabla contendrá los datos de los nodos (coordenadas geográficas, cota, demanda) y la segunda contendrá los datos de las tuberías (nodo inicial y final, longitud, diámetro, caudal)

En el desarrollo de la creación del grafo, se utilizan los IDs de la primera y segunda tabla para construir una matriz de adyacencia. A partir de ambas se crea el grafo mediante el paquete Igraph para poder visualizar la red. Véase el Anexo I) En el programa Epanet se grafica la RDAP y se simula para encontrar el valor de las demandas en las tuberías. (Ver Figura 18)

Figura N° 18. Cotas y Diámetros del Sector 253

(Fuente: Elaboración propia)

Capitulo III: Diagnóstico

3.1 Metodología de investigación

La metodología usada para la presente tesis se divide en 4 etapas principales, las cuales se muestra en la Figura N° 19

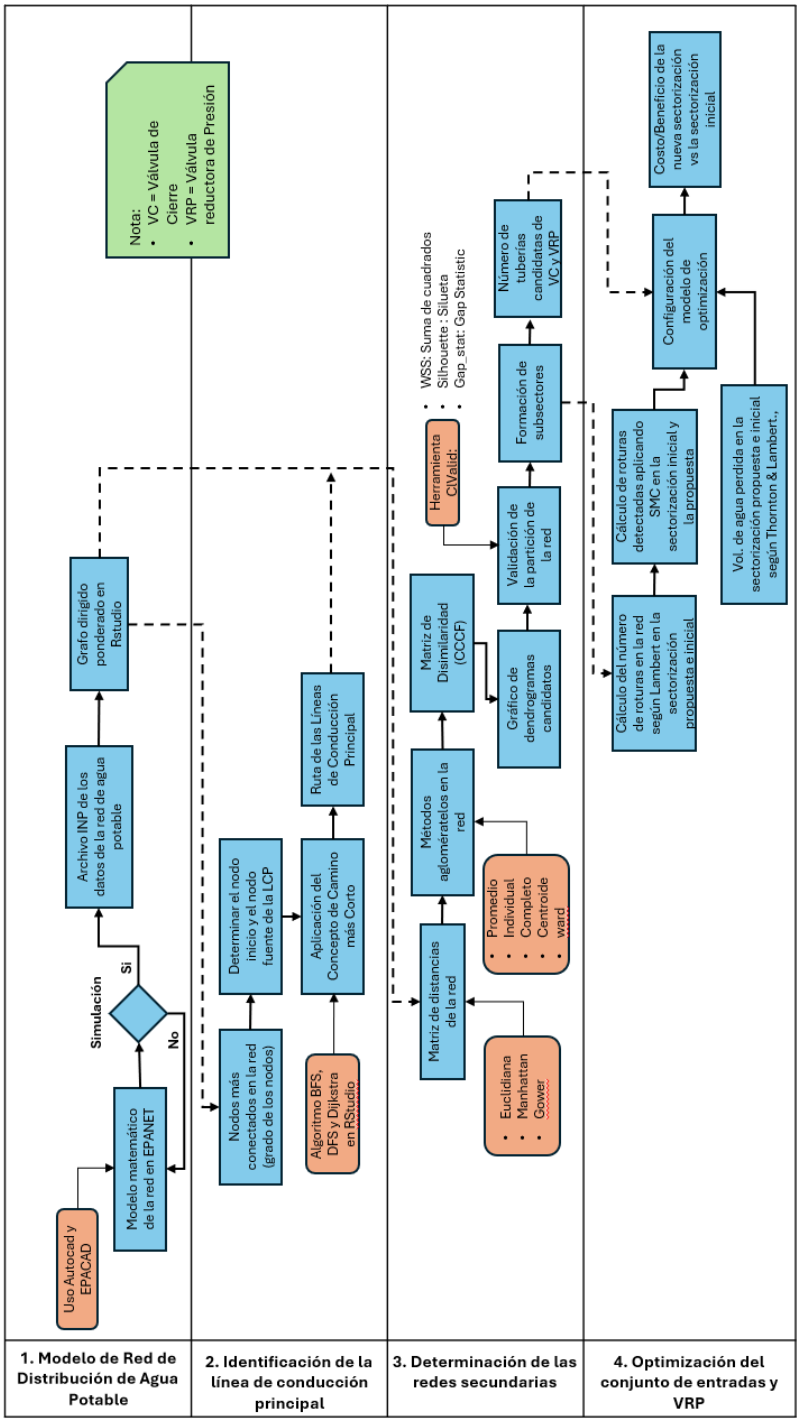

Figura N° 19. Esquema de la metodología aplicada en la tesis

(Fuente: Elaboración propia)

3.1.1 Zona de Estudio

La ciudad de Lima Metropolitana tiene 43 distritos y 7 distritos de la provincia constitucional del Callao, con un aproximado de 9 millones y medio de habitantes que son el 26.7% de la población del Perú. La provincia del Callao tiene casi 1 millón de habitantes, de los cuales 441 mil son parte del distrito de Ventanilla y el distrito de San Martin de Porres tiene poco más de 860 mil habitantes.

Sedapal como empresa gestora del agua potable en Lima y Callao lanzó en el 2018 un proyecto denominado “Víctor Raúl Haya de la Torre” para dotar y mejorar el servicio de agua potable a la zona de Márquez en Ventanilla, el AH. Nueva Jerusalén del Paraíso y varias Urbanizaciones de Villa Rica en San Martin de Porres, proyecto que consta de 5 sectores que son definidos como los sectores 253, 254, 255, 258 y 259.

El ámbito geográfico para el desarrollo de la presente tesis de investigación se realizará.

- **Departamento:** Lima
- **Provincia:** Lima y Callao
- **Distrito:** Callao, Ventanilla y San Martin de Porres
- **Red:** Victor Raúl Haya de la Torre
- **Sector:** 253, 254, 255, 258 y 259
- **Área:** 5.9 km2

Los planos de referencia que se utilizará para la tesis de investigación son los planos del Proyecto Víctor Raúl Haya de la Torre que beneficia a un promedio de 14000 habitantes con un total de 900 conexiones domiciliarias totales, los planos contienen las líneas de conducción principal y las redes secundarias que fueron proporcionados por Sedapal.

Los objetivos de esta investigación antes mencionados han sido elaborados considerando los datos actuales del año 2022 de la red “Víctor Raúl Haya de la Torre” recién construida. Las fuentes de agua que alimentan a cada sector se muestran en el Anexo 2. En la Figura N° 20 se muestra una imagen satelital de la ubicación de los sectores 253, 254, 255, 258 y 259 para una mejor apreciación.

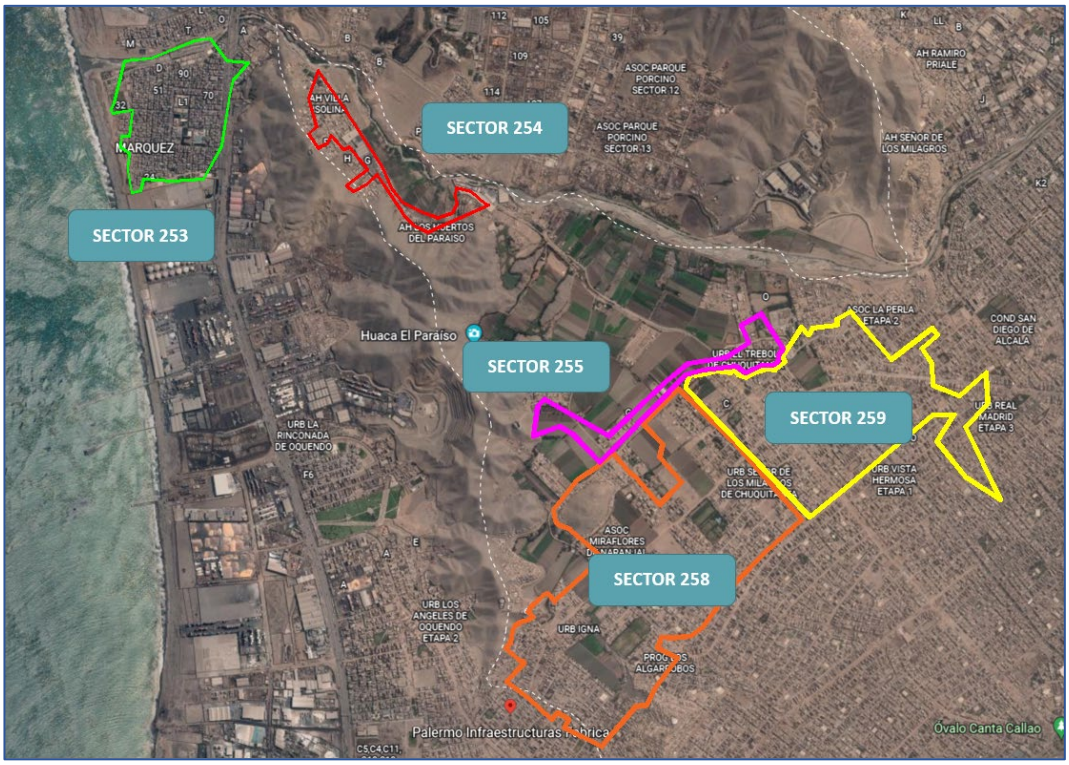

Figura N° 20. Vista satelital de la red "Víctor Raúl Haya de la Torre"
(Fuente: Google Earth)

En la Tabla N° 10 se muestra el área de los sectores 253, 254, 255, 258 y 259 respectivamente en metros cuadros y kilómetros cuadrados que resulta en un total de 5.9 kilómetros cuadrados del área de estudio.

Tabla N° 10. Área de los sectores de la RDAP

ID	N° de Sector	Área (m2)	Área (km2)
1	SECTOR 253	891,282.899	0.89
2	SECTOR 254	457,310.190	0.46
3	SECTOR 255	270,730.598	0.27
4	SECTOR 258	2,171,694.088	2.17
5	SECTOR 259	2,108,440.862	2.11
	Total	5,899,458.64	5.90

(Fuente: Elaboración propia)

3.2 Componentes de la red de distribución de agua potable “Víctor Raúl Haya de la Torre de los sectores 253-254-255-258-259”

La RDAP “Víctor Raúl Haya De La Torre” consiste en una red compuesta por tuberías 793 tuberías que tienen un estimado 110 mm a 450 mm de diámetro. La red cuenta con 5 fuentes de agua distribuidas en los 5 sectores iniciales del proyecto. Dicha red tendrá más de 10 mil beneficiarios en los distritos de San Martin de Porres, Callao y Ventanilla. A continuación, se hará una breve descripción de los cinco sectores iniciales del proyecto.

En la Figura N° 21 se puede apreciar al sector 253 que está conformado por 168 nodos, 6 cámaras reductoras de presión, 4 cámaras de válvulas, un reservorio de 200 m³ y en la Tabla N° 11 se aprecia la distribución de las CRP, CV y reservorios.

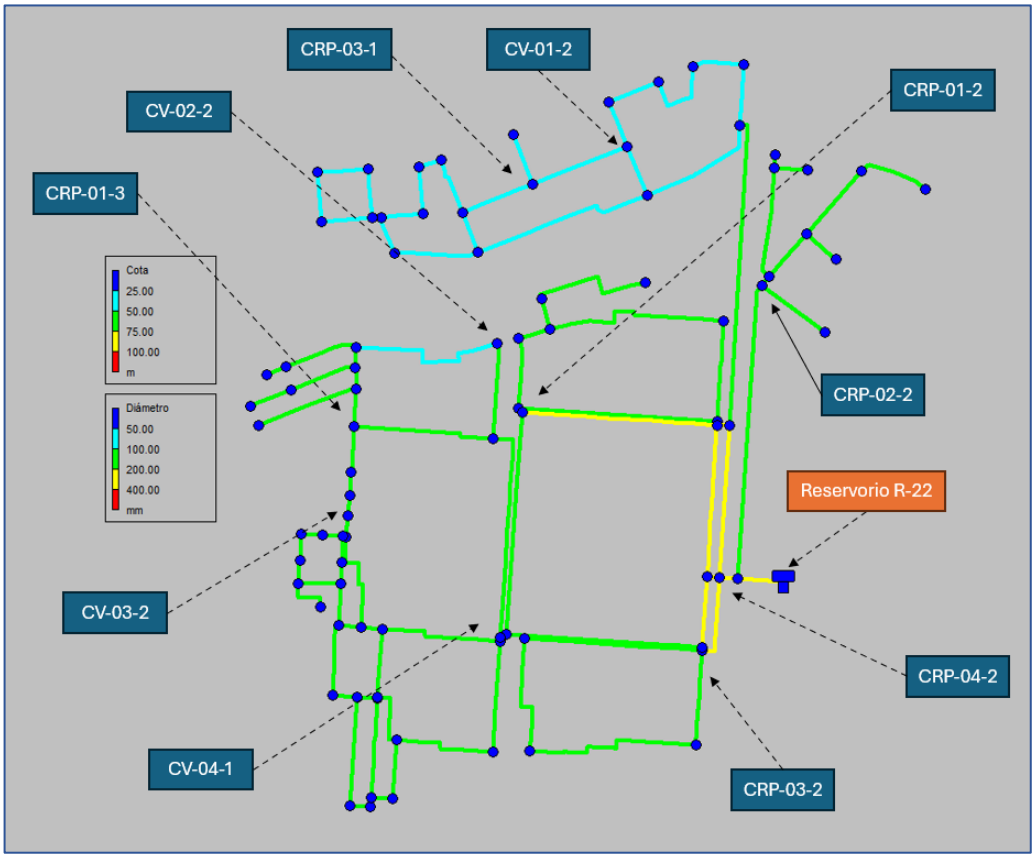

Figura N° 21. Cotas y Diámetros del Sector 253

(Fuente: Elaboración propia)

Tabla N° 11. Datos de nodos del Sector 253

Sector 253	
# de nodos	168
CRP	Cámara Reductora de Presión CRP-01-2
	Cámara Reductora de Presión CRP-02-2
	Cámara Reductora de Presión CRP-03-2
	Cámara Reductora de Presión CRP-04-2
	Cámara Reductora de Presión CRP-01-3
	Cámara Reductora de Presión CRP-03-3
	Cámara de Válvulas CV-01-2
	Cámara de Válvulas CV-02-2
	Cámara de Válvulas CV-03-2
	Cámara de Válvulas CV-04-1
Fuente	Reservorio R-22 200m3

(Fuente: Elaboración propia)

En la Tabla N° 12 se encuentra la distribución de tuberías del Sector 253, compuesta por tuberías de PVC y HD K-9.

Tabla N° 12. Distribución de diámetros de tuberías del Sector 253

Descripción	Número de Tuberías
90 mm PVC	24
100 mm PVC	82
110 mm PVC	63
150 mm HD K-9	13
160 mm PVC	7
200 mm PVC	12
250 mm PVC	2
Subtotal	203

(Fuente: Elaboración propia)

En la Figura N° 22 se aprecia al sector 254 y en la Tabla N° 13 se muestra la distribución de nodos conformado por 46 nodos, 4 cámaras reductoras de presión, 4 cámaras de válvulas, un reservorio RAP-03 de 300 m³.

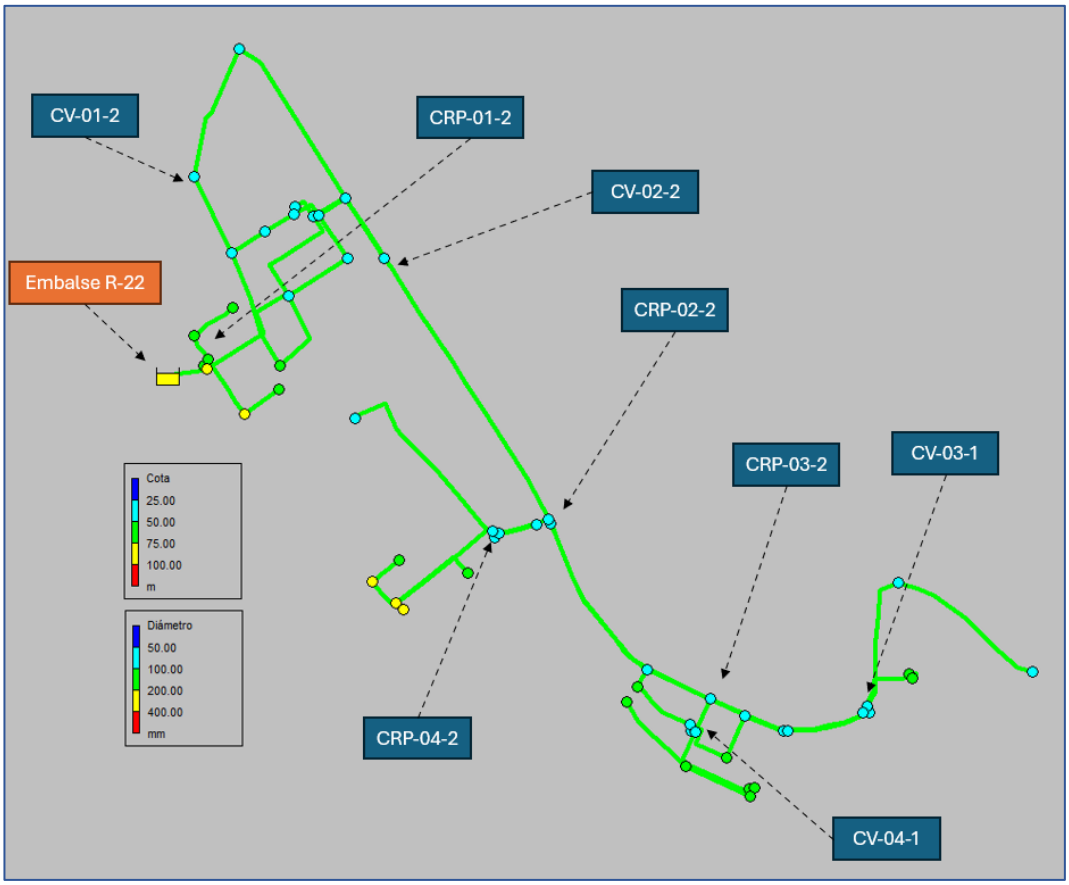

Figura N° 22. Cotas y Diámetros del Sector 254

(Fuente: Elaboración propia)

Tabla N° 13. Datos de nodos del Sector 254

Sector 254	
# de nodos	46
CRP	Cámara Reductora de Presión CRP-01-2
	Cámara Reductora de Presión CRP-02-2
	Cámara Reductora de Presión CRP-03-2
	Cámara Reductora de Presión CRP-04-2
	Cámara de Válvulas CV-01-2
	Cámara de Válvulas CV-02-2
	Cámara de Válvulas CV-03-2
	Cámara de Válvulas CV-04-1
Fuente	Reservorio Apoyado Proyectado RAP-03

(Fuente: Elaboración propia)

Tabla N° 14. Distribución de diámetros de tuberías del Sector 254

Descripción	Número de Tuberías
100 mm PVC	50
150 mm HD K-9	7
Subtotal	57

(Fuente: Elaboración propia)

El subsector 255 está conformado por 33 nodos, 2 cámaras reductoras de presión, un reservorio RAP-03 de 600 m³ (Ver Tabla N° 15) y 41 tuberías (Ver Figura N° 23).

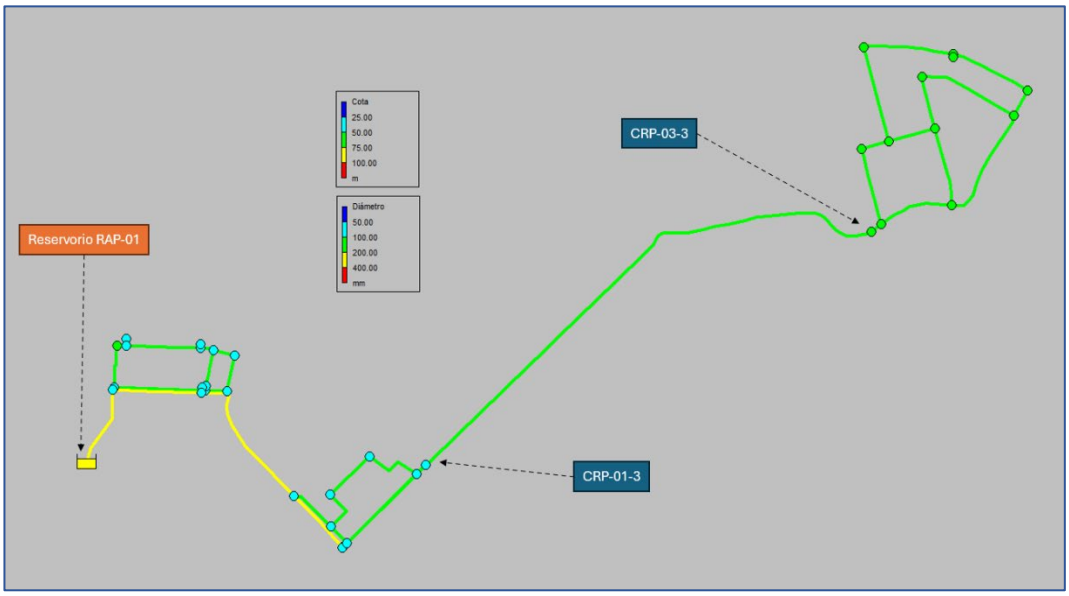

Figura N° 23. Cotas y Diámetros del Sector 255

(Fuente: Elaboración propia)

Tabla N° 15. Datos de nodos del Sector 255

Sector 255	
# de nodos	33
CRP	Cámara Reductora de Presión CRP-01-3
	Cámara Reductora de Presión CRP-03-3
Fuente	Reservorio Apoyado Proyectado RAP-01

(Fuente: Elaboración propia)

Tabla N° 16. Distribución de diámetros de tuberías del Sector 255

Descripción	Número de Tuberías
100 mm PVC	31
150 mm HD K-9	6
200 mm PVC	4
Subtotal	41

(Fuente: Elaboración propia)

El subsector 258 está conformado por 270 nodos, 6 cámaras reductoras de presión, 2 cámaras de válvulas, un reservorio RAP-02-4 de 2200 m³ (Ver Figura N° 24 y Tabla N° 17) y 31 tuberías (Ver Tabla N° 18).

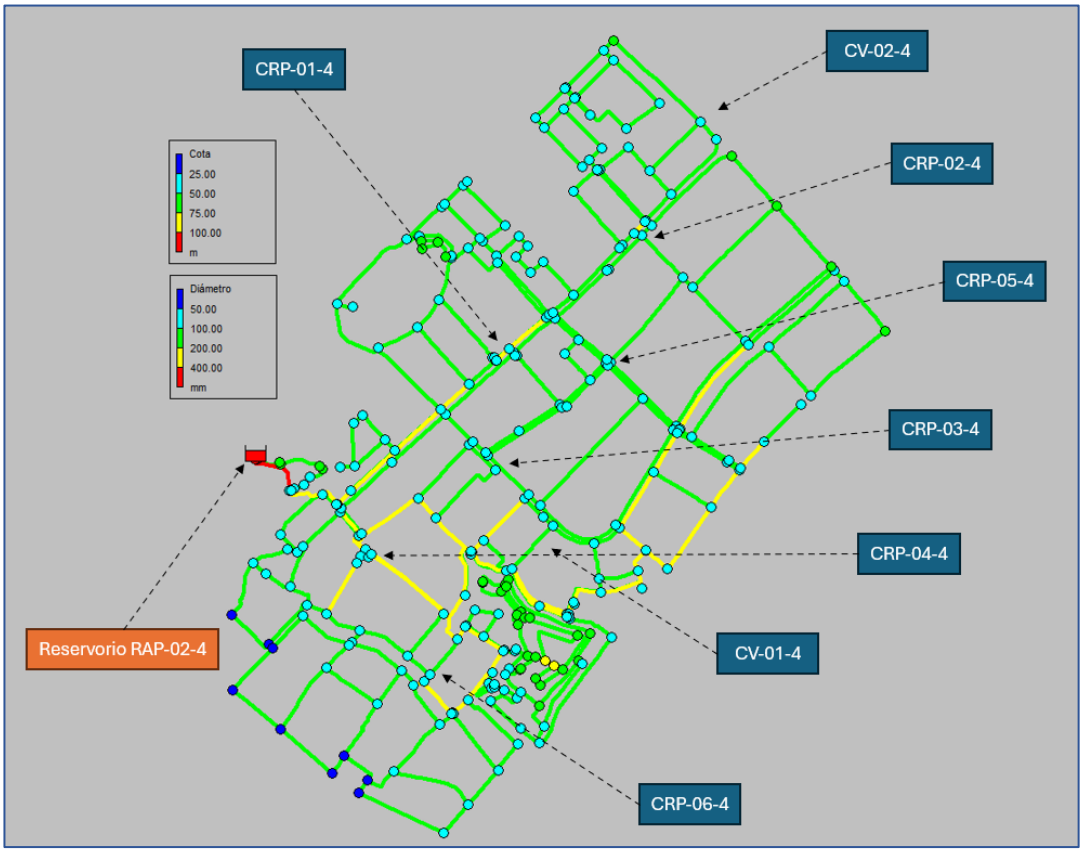

Figura N° 24. Cotas y Diámetros del Sector 258

(Fuente: Elaboración propia)

Tabla N° 17. Datos de nodos del Sector 258

Sector 258	
# de nodos	270
CRP	Cámara Reductora de Presión CRP-01-4
	Cámara Reductora de Presión CRP-02-4
	Cámara Reductora de Presión CRP-03-4
	Cámara Reductora de Presión CRP--04-4
	Cámara Reductora de Presión CRP-05-4
	Cámara Reductora de Presión CRP-06-4
	Cámara de Válvulas CV-01-4
	Cámara de Válvulas CV-02-4
Fuente	Reservorio Apoyado Proyectado RAP-02-4

(Fuente: Elaboración propia)

Tabla N° 18. Distribución de diámetros de tuberías del Sector 258

Descripción	**Número de Tuberías**
110 mm PVC	269
150 mm HD K-9	8
160 mm PVC	17
200 mm PVC	31
250 mm PVC	3
300 mm PVC	8
315 mm PVC	4
350 mm PVC	2
400 mm PVC	2
Subtotal	344

(Fuente: Elaboración propia)

En la Figura N° 25 se muestra el subsector 259 está conformado por 162 nodos, 8 cámaras de válvulas, un reservorio RAP-02-4 de 2000 m³ (Ver Tabla N° 19) y 209 tuberías (Ver Tabla N° 20).

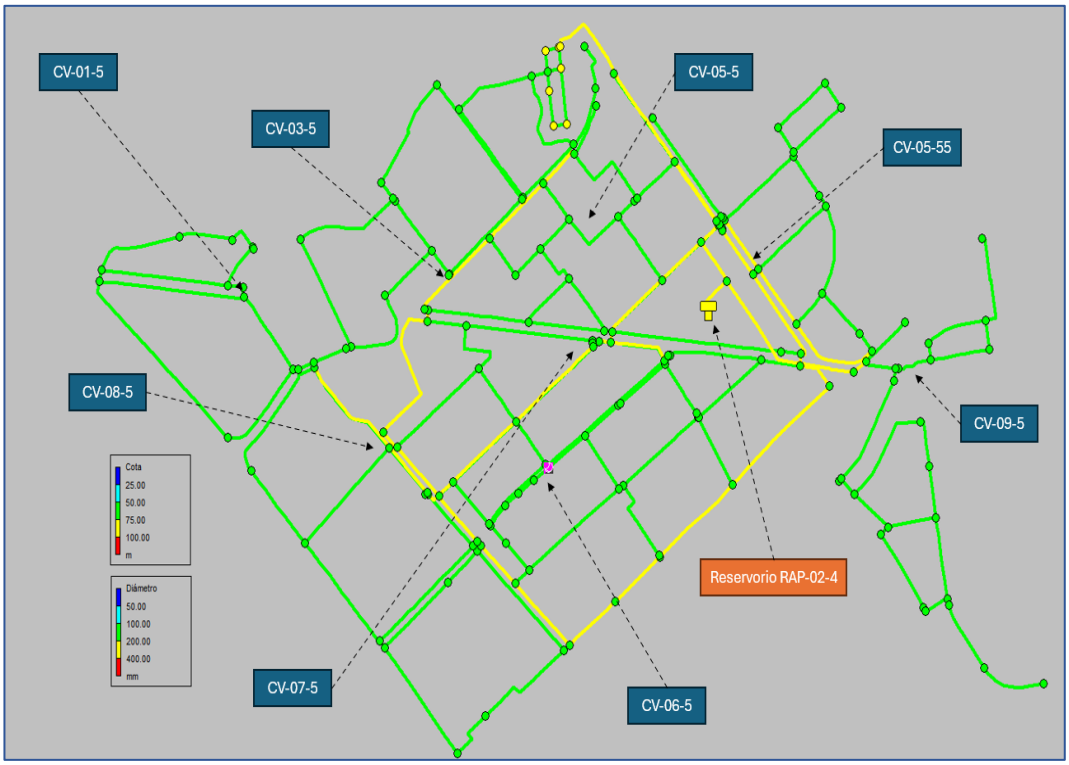

Figura N° 25. Cotas y Diámetros del Sector 259

(Fuente: Elaboración propia)

Tabla N° 19. Datos de nodos del Sector 259

Sector 259	
# de nodos	162
CV	Cámara de Válvulas CV-01-5
	Cámara de Válvulas CV-03-5
	Cámara de Válvulas CV-05-5
	Cámara de Válvulas CV-05-55
	Cámara de Válvulas CV-06-5
	Cámara de Válvulas CV-07-5
	Cámara de Válvulas CV-08-5
	Cámara de Válvulas CV-09-5
Fuente	Reservorio Proyectado REP-01

(Fuente: Elaboración propia)

Tabla N° 19-A. Datos de tuberías del Sector 259

Descripción	Número de Tuberías
110 mm PVC	129
150 mm HD K-9	2
160 mm PVC	31
200 mm PVC	39
250 mm PVC	6
300 mm PVC	2
Subtotal	209

(Fuente: Elaboración propia)

Capítulo IV: Proceso de sectorización

De acuerdo con lo indicado en ítem 3.2 se aplicó el software Epanet para evaluar las redes de estudio. Los resultados se muestran a continuación.

4.1 Identificación de las líneas de conducción principal mediante el Concepto de Caminos Más Cortos

Para identificar las líneas de conducción principal en cada sector se necesita aplicar algoritmos de exploración, los cuales tienen que hacer un mapeo y reconocimiento de que nodos y tuberías están más conectadas a los demás elementos de la red. También se debe tomar en consideración cual es la ruta más corta entre la fuente y el punto de llegada de la red para minimizar la perdida de energía del agua potable. Con este fin se usa el paquete shortest.paths que hace uso de algoritmos de exploración en el lenguaje de programación R. Para el correcto uso del paquete shortest.paths se ingresa el grafo a analizar, el punto de llegada y la fuente de agua en la red. (Ver Figura 26)

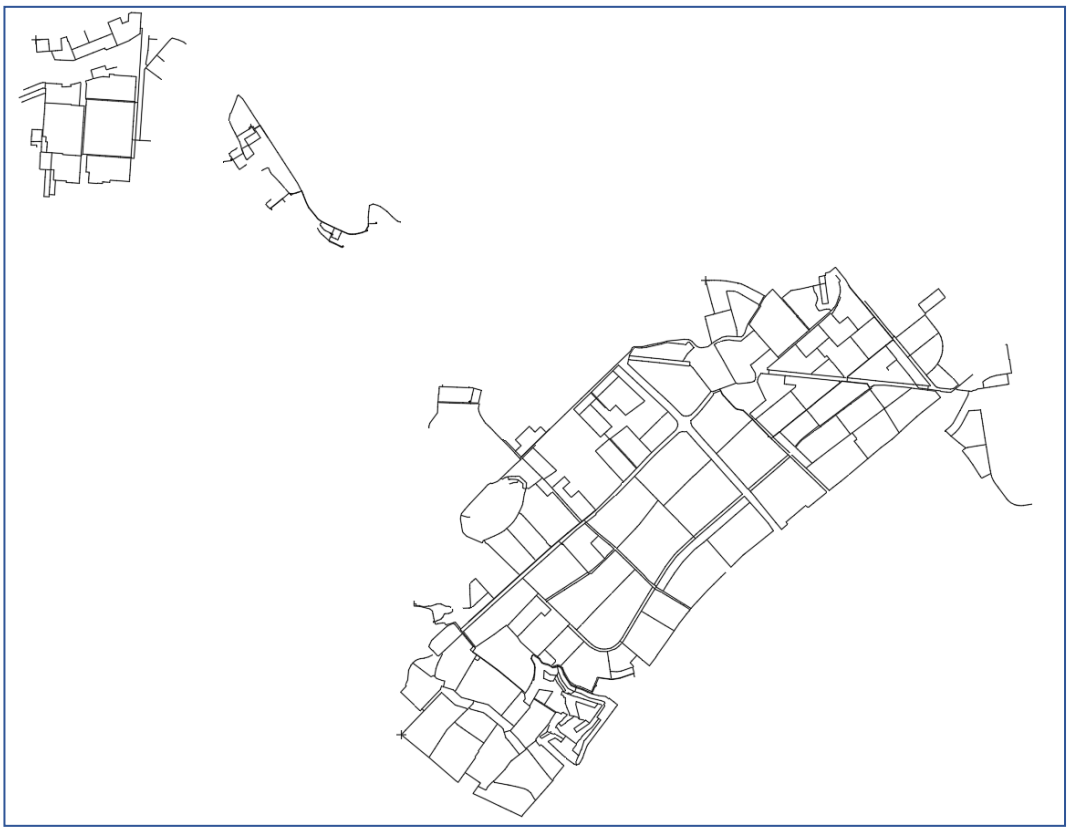

Figura N° 26. Grafo de la red de distribución en R

(Fuente: Elaboración propia)

En el sector 253 se pudo determinar dos rutas para la línea de conducción principal con el uso del concepto de Camino más Corto. En las tablas N° 20 y N° 21 muestran las tuberías y nodos que conforman la línea de conducción principal de la red hidráulica.

Tabla N° 20. Ruta del camino más corto entre la fuente R-522 y J-57

ID Tubería	Nodo 1	Nodo 2	Longitud (m)
P-35-1	R-522	J-48	64.671
P-34-1	J-48	J-74	26.07
P-33-1	J-74	J-75	17.82
P-57-1	J-75	J-53	216.3
P-64-1	J-53	J-57	277.3
Total (m)			602.2

(Fuente: Elaboración propia)

Tabla N° 21. Ruta del camino más corto entre la fuente R-522 y J-77

ID Tubería	Nodo 1	Nodo 2	Longitud (m)
P-35-1	R-522	J-48	64.671
P-34-1	J-48	J-74	26.07
P-62-1	J-74	J-73	216.3
P-22-1	J-73	J-72	438.8
P-14-1	J-72	J-93	181.1
P-75-1	J-93	J-86	261.7
P-76-1	J-86	J-81	120.1
P-77-1	J-81	J-77	54.27
Total (m)			1363.0

(Fuente: Elaboración propia)

La línea de conducción principal tiene una longitud de 602.2 metros en el primer tramo y 1363.0 metros en el segundo.

En la Figura N° 27. muestra las dos rutas que conforman la línea de conducción principal bajo el principio de Camino más Corto.

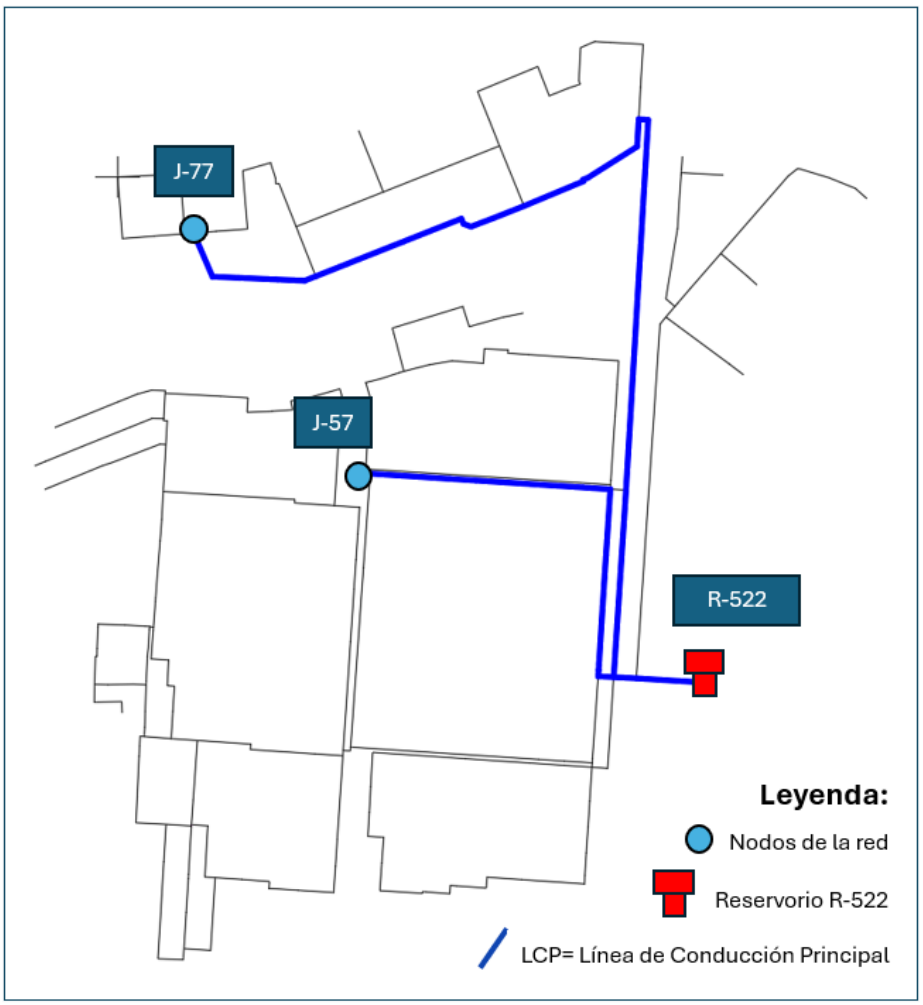

Figura N° 27. Líneas de conducción principal en el sector 253
(Fuente: Elaboración propia)

En el sector 254 se tiene una sola fuente de agua que alimenta a la red hidráulica. A continuación, se muestra la Tabla N° 22 en donde se describe las tuberías y nodos que conforman la línea de conducción principal.

Tabla N° 22. Ruta del camino más corto entre la fuente RAP-03-2 al J-8-2

ID Tubería	Nodo 1	Nodo 2	Longitud (m)
P-10-2	RAP-03-2	J-51-2	52.297
P-11-2	J-51-2	J-54-2	285.2
P-41-2	J-54-2	J-57-2	491.7
P-33-2	J-57-2	J-55-2	297.9
P-31-2	J-55-2	J-53-2	190.7
P-30-2	J-53-2	CV-04-1	88.65
P-35-2	CV-04-1	J-8-2	16.45
	Total (m)		1422.9

(Fuente: Elaboración propia)

En la Figura N° 28. muestra las dos rutas que conforman la línea de conducción principal bajo el principio de Camino más Corto con una longitud total de 1422.9m.

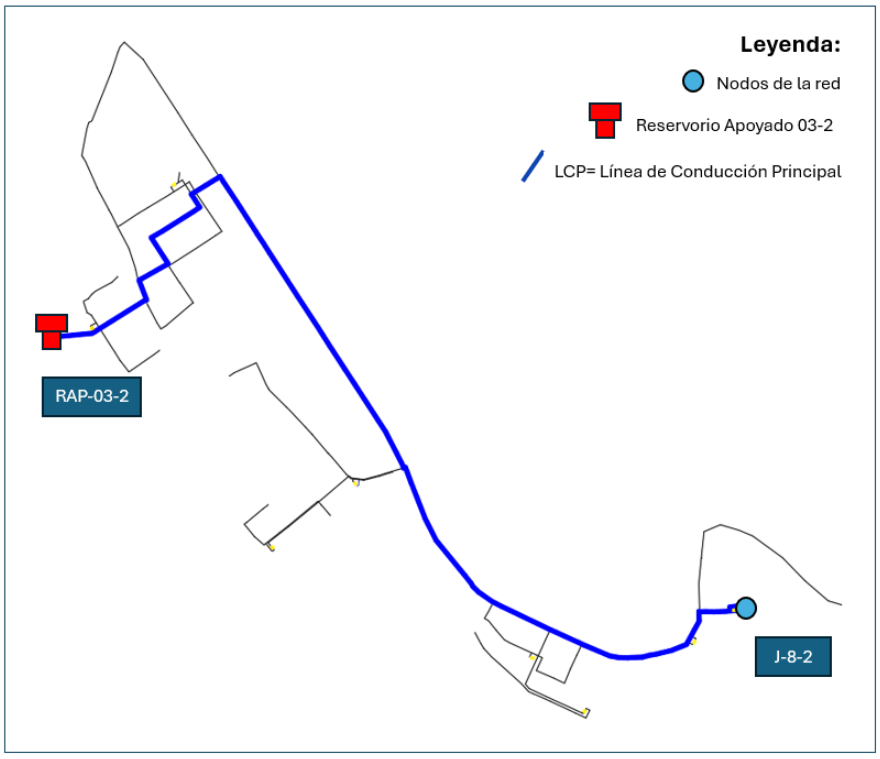

Figura N° 28. Línea de conducción principal en el sector 254

(Fuente: Elaboración propia)

En el sector 255 se tiene una sola fuente de agua que alimenta a la red hidráulica. A continuación, en la Tabla N° 23 se muestra las tuberías y nodos que conforman a su línea de conducción principal.

Tabla N° 23. Ruta del camino más corto entre la fuente RAP-01-3 al J-28-3

ID Tubería	Node1	Node2	Longitud (m)
P-4-3	RAP-01-3	J-36-3	166.4
P-2-3	J-36-3	J-39-3	181.6
P-1-3	J-39-3	J-38-3	318.3
P-32-3	J-38-3	J-34-3	145
P-33-3	J-34-3	CRP-03-3	1371
P-14-3	CRP-03-3	J-29-3	25.08
P-36-3	J-29-3	J-28-3	155.4
	Total (m)		2362.9

(Fuente: Elaboración propia)

En la Figura N° 29. muestra las dos rutas que conforman la línea de conducción principal bajo el principio de Camino más Corto con una longitud total de 2362.9m.

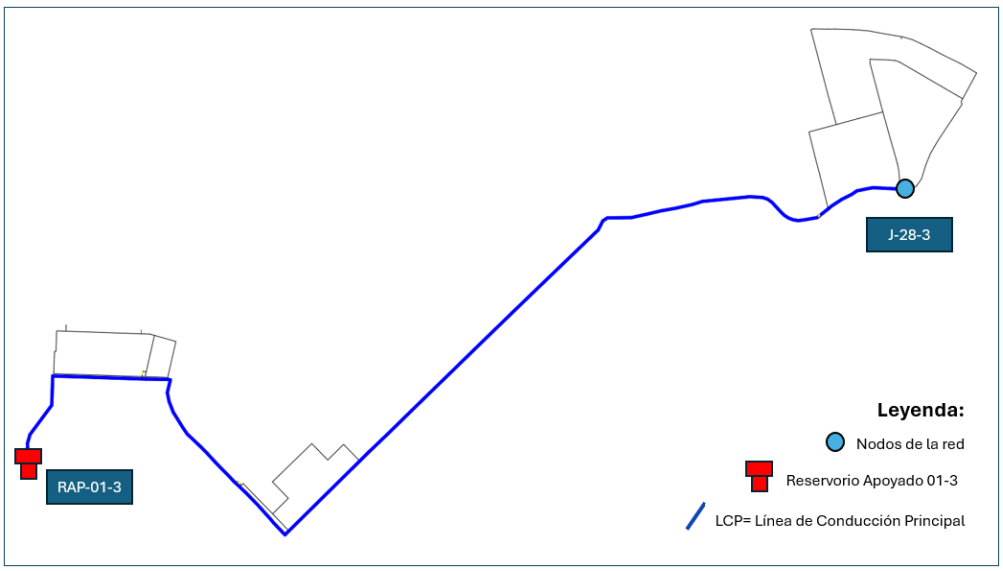

Figura N° 29. Línea de conducción principal en el sector 255

(Fuente: Elaboración propia)

En el sector 258 se tiene una sola fuente de agua que alimenta a la red hidráulica. A continuación, en la Tabla N° 24 y N° 25 se muestra las tuberías y nodos que conforman a su línea de conducción principal.

Tabla N° 24. Ruta del camino más corto entre la fuente RAP-03-4 al J-116-4

ID Tubería	Nodo 1	Nodo 2	Longitud (m)
P-90-4	RAP-03-4	J-272-4	14.7
P-91-4	J-272-4	J-270-4	184.3
P-79-4	J-270-4	J-278-4	192.8
P-328-4	J-278-4	J-279-4	4.1
P-320-4	J-277-4	J-279-4	218.9
P-156-4	J-277-4	J-275-4	910.4
P-157-4	J-275-4	J-271-4	95.0
P-156-4	PRV-8-4	J-271-4	56.9
P-330-4	PRV-8-4	J-182-4	9.5
P-309-4	J-182-4	J-116-4	292.9
	Total (m)		1979.5

(Fuente: Elaboración propia)

Tabla N° 25. Ruta del camino más corto entre la fuente RAP-03-4 al CRP-05-4

ID Tubería	Nodo 1	Nodo 2	Longitud (m)
P-90-4	RAP-03-4	J-272-4	14.66
P-91-4	J-272-4	J-270-4	184.3
P-79-4	J-270-4	J-278-4	192.8
P-80-4	J-278-4	J-273-4	702.1
P-81-4	J-273-4	J-274-4	55.46
P-82-4	J-274-4	J-276-4	188.7
P-83-4	J-276-4	CRP-05-4	305.2
	Total (m)		1643.2

(Fuente: Elaboración propia)

La línea de conducción principal tiene una longitud de 1979.5 metros en el primer tramo y 1643.2 metros en el segundo. En la Figura N° 30 muestra las dos rutas que conforman la línea de conducción principal bajo el principio de Camino más Corto.

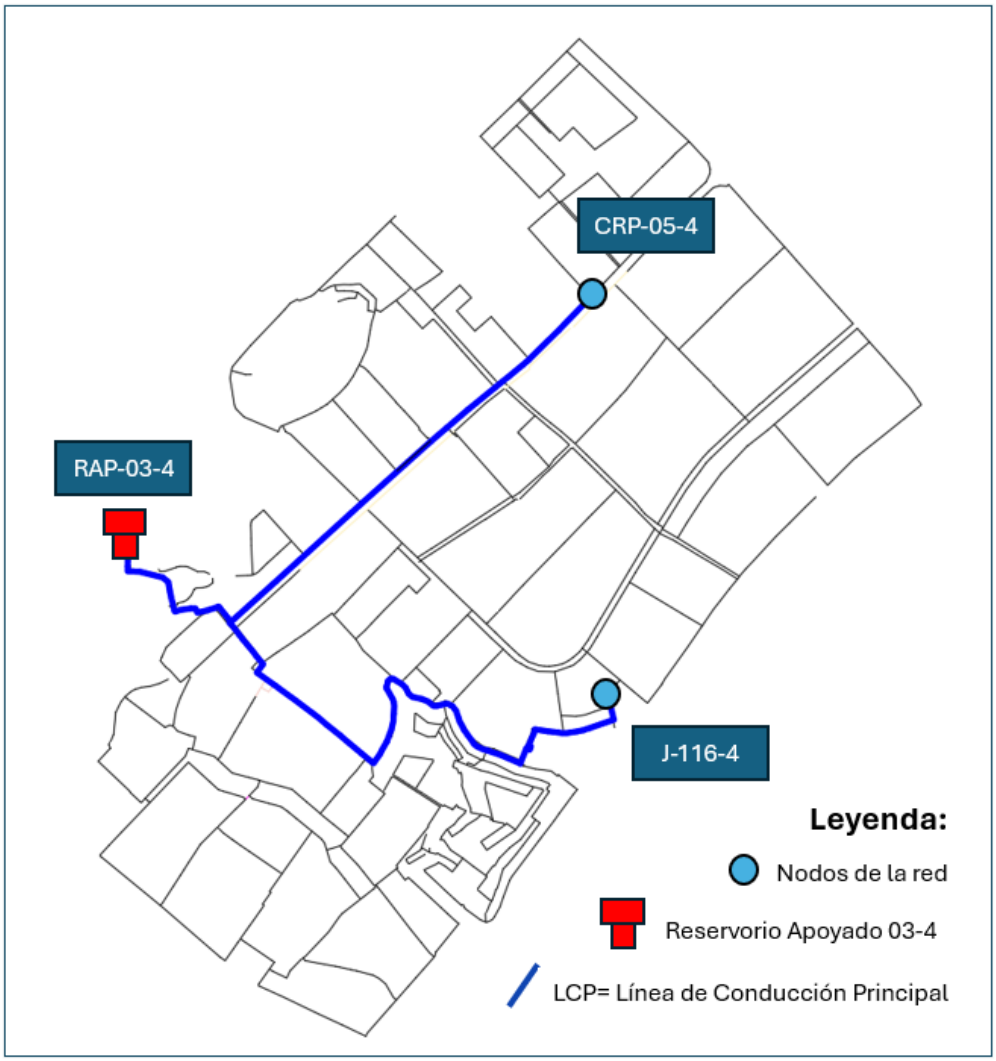

Figura N° 30. Líneas de conducción principal en el sector 258

(Fuente: Elaboración propia)

En el sector 259 se tiene una sola fuente de agua que alimenta a la red hidráulica. A continuación, en la Tabla N° 26 se muestra las tuberías y nodos que conforman a su línea de conducción principal de dicho sector.

Tabla N° 26. Ruta del camino más corto entre la fuente REP-01-5 al J-167-5

ID Tubería	Nodo 1	Nodo 2	Longitud (m)
P-129-5	REP-01-5	J-173-5	92.28
P-130-5	J-173-5	J-174-5	105.3
P-131-5	J-174-5	J-172-5	311.2
P-79-5	J-169-5	J-172-5	47.09
P-154-5	J-169-5	J-168-5	527.6
P-74-5	J-168-5	J-167-5	394.9
	Total (m)		1478.4

(Fuente: Elaboración propia)

La línea de conducción principal tiene una longitud de 1478.4 metros en el primer tramo. En la Figura N° 31 muestra la ruta que conforma la línea de conducción principal bajo el principio de Camino en el sector 259.

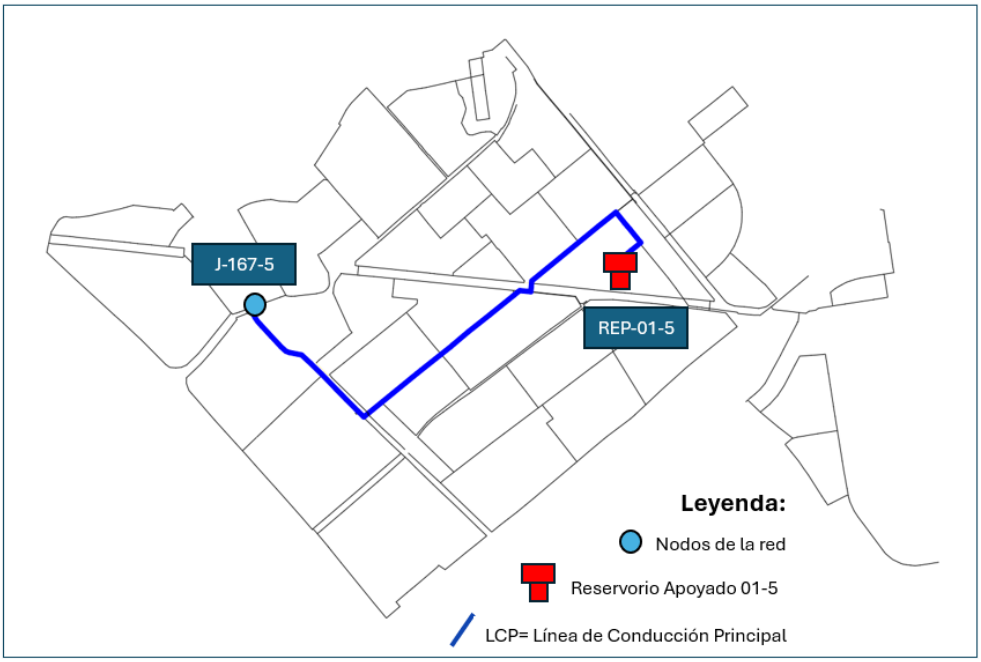

Figura N° 31. Líneas de conducción principal en el sector 259

(Fuente: Elaboración propia)

4.2 Sectorización de la red hidráulica basada en el método de Clustering Jerárquico

El primer paso para implementar la sectorización basado en el método de Clustering Jerárquico consiste en definir las líneas de conducción principal utilizando los criterios de la menor perdida de energía por fricción y los nodos más interconectados entre sí.

La red ejemplo utilizada en el presente trabajo cuenta con cinco sectores bien definidos por una fuente de agua independiente para cada uno de ellos. Para resumir la aplicación del método de Clustering se procederá a analizar los datos de los primeros tres sectores más pequeños (sector 253, 254 y 255). En la Figura N° 32 se muestra las líneas de conducción principal en los primeros tres sectores de la red ejemplo encontrados en la sección anterior.

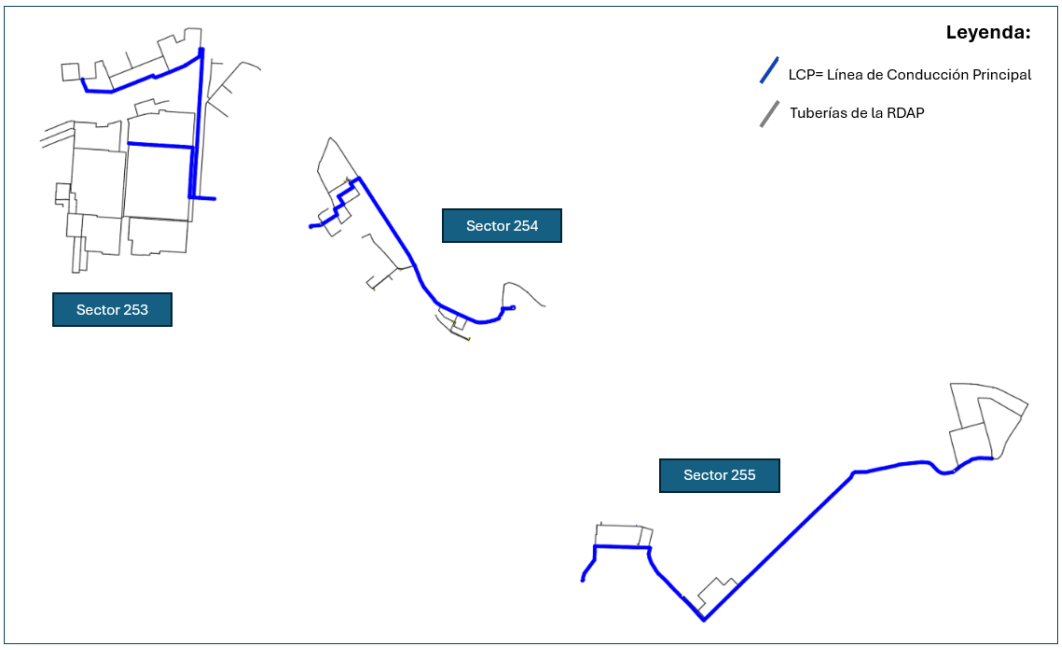

Figura N° 32. Líneas de conducción principal del sector 253, 254 y 255

(Fuente: Elaboración propia)

Las características utilizadas para la definición de la matriz de disimilaridad necesaria para implementar el Clustering Jerárquico en los tres sectores escogidos son las cotas y las coordenadas geográficas de los nodos de la red.

Si bien se puede utilizar las demandas en los nodos, esta característica no es recomendada debido a que hay más probabilidad que aparezcan valores atípicos relacionadas a demandas con un alto valor (hospitales e industria) en contraste con demandas de bajo valor (residenciales y domesticas) en el mismo sector. Como se explicó con detalle en el capítulo III, la técnica de Clustering Jerárquico ofrece alternativas de tipo métricas en la construcción de la matriz de disimilaridad y en los métodos de agrupación en el proceso de aglomeración de clústeres.

El uso del Coeficiente de Correlación Cofenética (CCCF) es una medida bastante aceptada para evaluar a las mejores alternativas. En la Tabla N° 27 se muestran resaltados en amarillo los valores más altos del CCCF para la sectorización de los datos mencionados líneas arriba.

Tabla N° 27. Valores del CCCF para distintas combinaciones de métrica y método Aglomerativo

	Euclidiana	Manhattan	Gower
promedio	0.8505921	0.8383431	0.8647751
individual	0.8608677	0.8137162	0.8453224
completo	0.8460288	0.8568064	0.8597181
centroide	0.8486717	0.8365721	0.8640229
ward	0.8302827	0.8228474	0.8086757

(Fuente: Elaboración propia)

La Figura N° 33, la Figura N° 34 y la Figura N° 35 muestran los dendrogramas de las tres combinaciones seleccionadas de los sectores 253, 254 y 255.

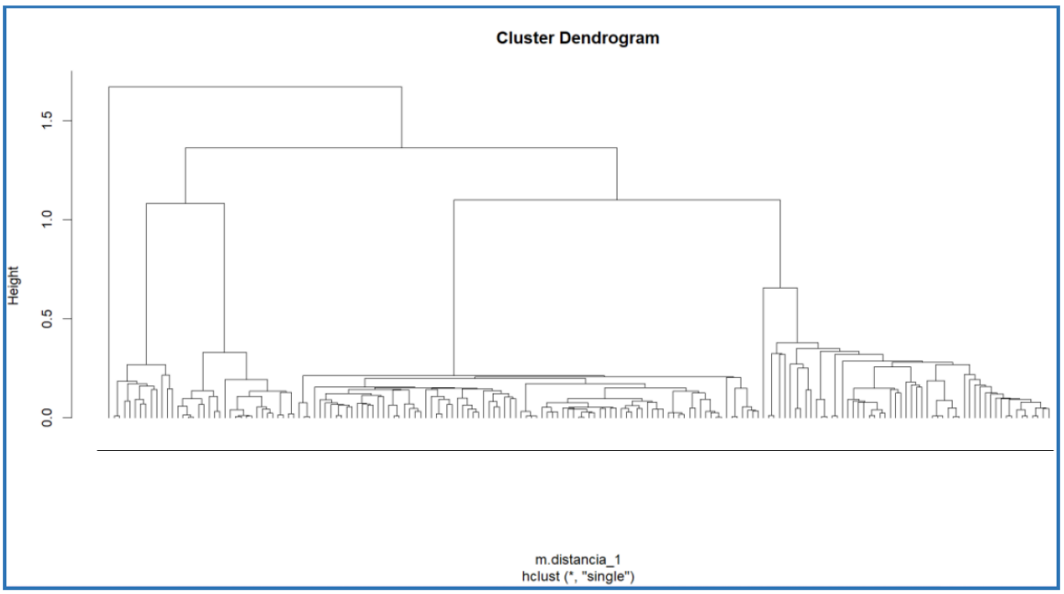

Figura N° 33. Dendrograma generado por la combinación distancia Euclidiana con método individual

(Fuente: Elaboración propia)

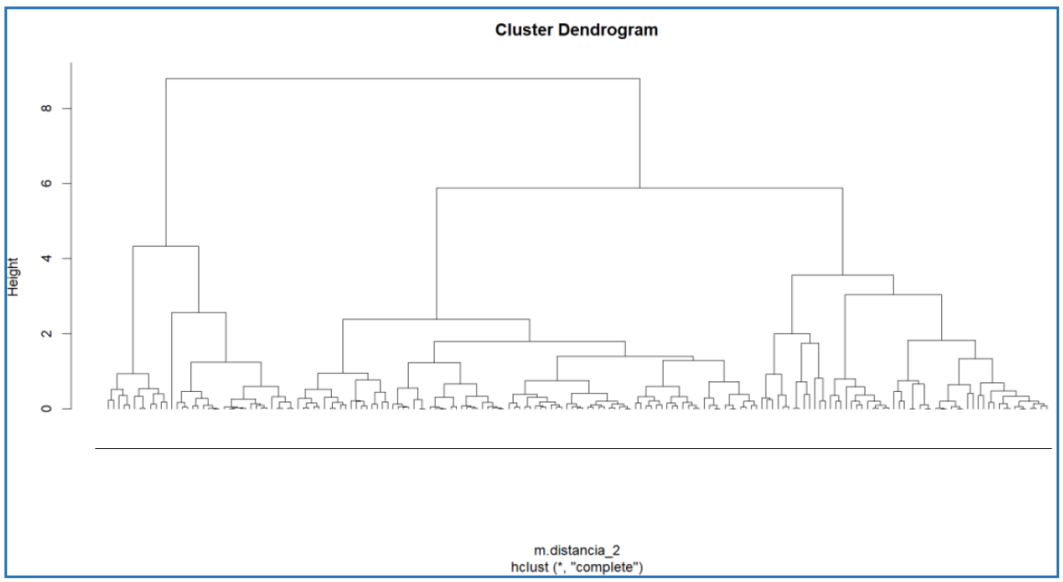

Figura N° 34. Dendrograma generado por la combinación distancia Manhattan con método completo

(Fuente: Elaboración propia)

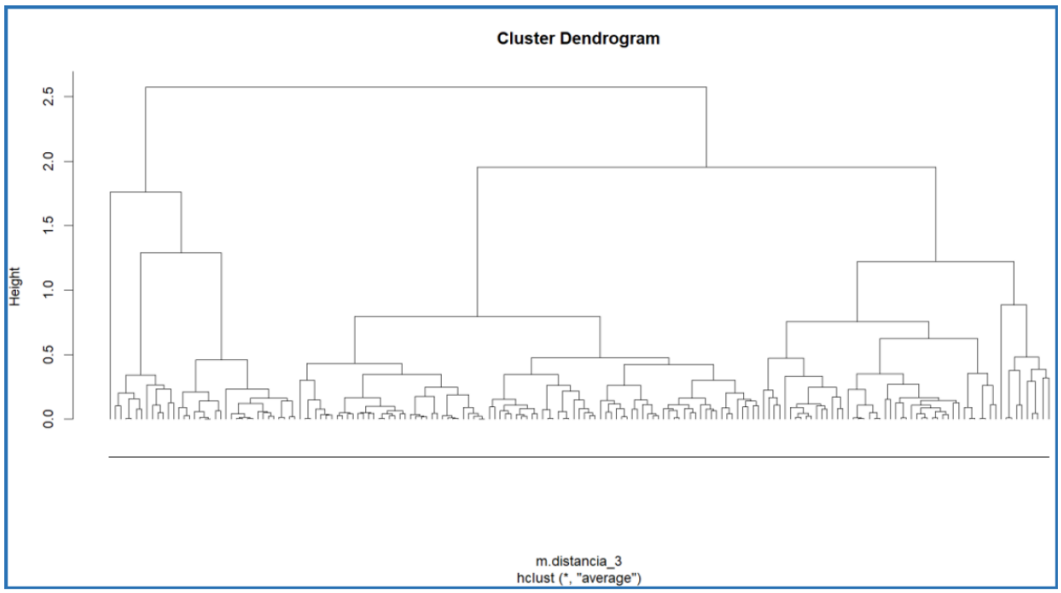

Figura N° 35. Dendrograma generado por la combinación distancia Gower con método promedio
(Fuente: Elaboración propia)

Como se puede apreciar el dendrograma resultante de la distancia Gower con el método promedio es el más homogéneo en la formación de clústeres. Para estimar el número de clústeres más eficiente en que se debe hacer la partición se aplicara la herramienta ClValid. Una de las capacidades más interesantes de dicha herramienta es el hecho que emplea diferentes métodos de evaluación del número óptimo de particiones tales como el WSS, Silhouette o gap_stat. En la Tabla N° 28 se muestra el resultado de aplicar ClValid y se puede concluir que el número de particiones optimo es tres.

Tabla N° 28. Valores más eficientes del número de particiones.

	kmeans	pam	clara
WSS	3	3	3
SILHOUETTE	3	3	3
GAP_STAT	3	3	4

(Fuente: Elaboración propia)

En la Figura N° 36 aplicando el criterio del codo (método grafico para identificar el cambio de concavidad de la gráfica) en la gráfica del método WSS resulta que el número óptimo de particiones es tres.

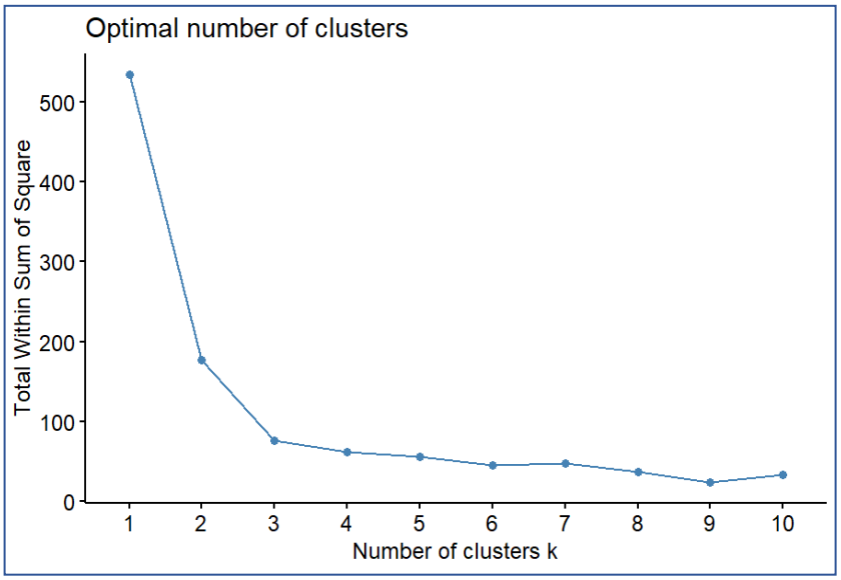

Figura N° 36. Grafica del método WSS de los sectores 253, 254 y 255

(Fuente: Elaboración propia)

En la Figura N° 37 aplicando el criterio del codo en la gráfica del método Silhouette resulta que el número óptimo de particiones es tres.

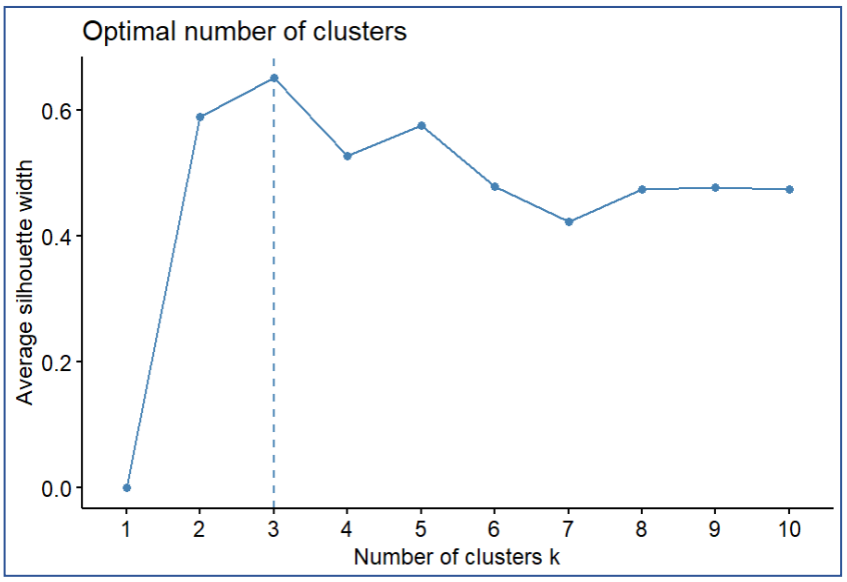

Figura N° 37. Grafica del método silhouette de los sectores 253, 254 y 255

(Fuente: Elaboración propia)

En la Figura N° 38 aplicando el criterio del codo en la gráfica del método gap_stat resulta que el número óptimo de particiones es cuatro.

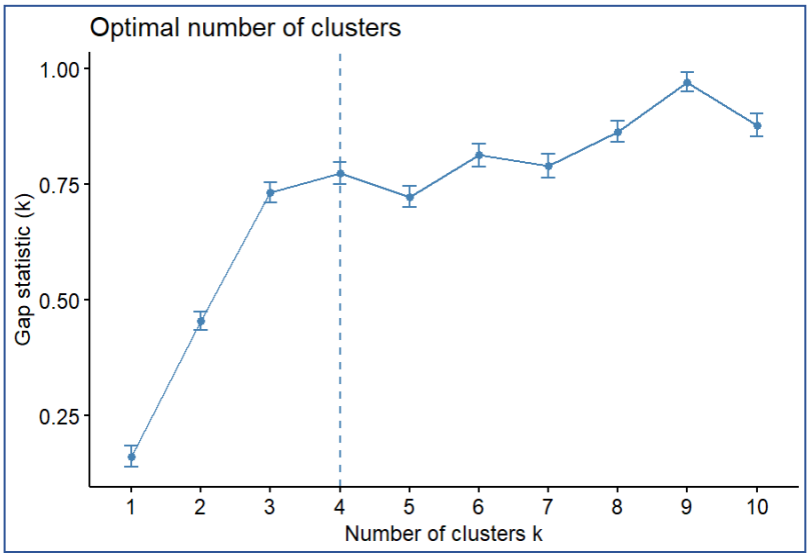

Figura N° 38. Grafica de barras del método gap_stat de los sectores 253, 254 y 255

(Fuente: Elaboración propia)

En la Figura N° 39 aplicando la herramienta resnumclust resultara que el número óptimo de particiones es tres, corroborando el resultado de la herramienta ClValid.

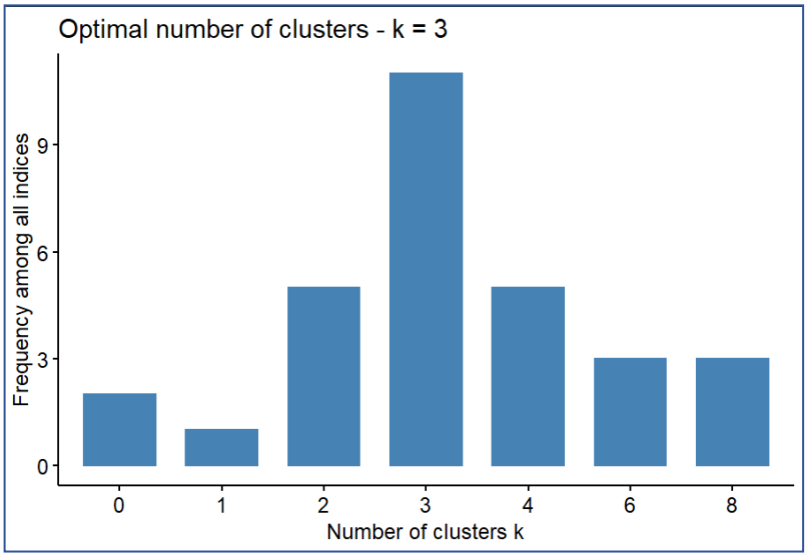

Figura N° 39. Grafica de barras aplicando la herramienta resnumclust de los sectores 253, 254 y 255

(Fuente: Elaboración propia)

El dendrograma resultante de aplicar el Clustering Jerárquico en los datos de los sectores 253, 254 y 255 al mismo tiempo corrobora la sectorización inicial del proyecto, teniendo a los nodos y tuberías de cada sector unidos a sí mismos después de la sectorización. (Ver Figura N° 40)

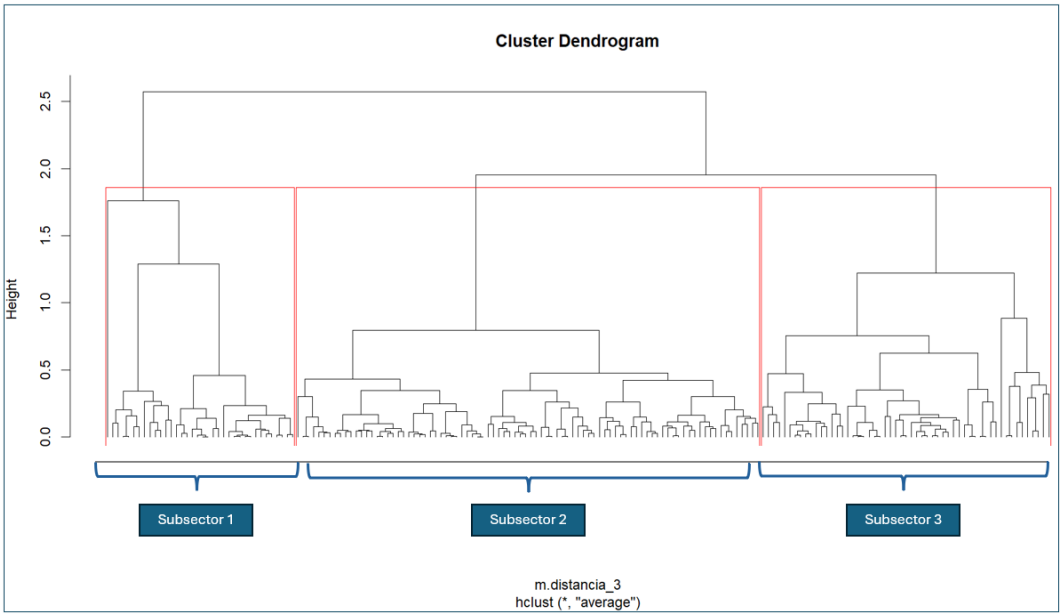

Figura N° 40. Dendrograma de los sectores 253, 254 y 255

(Fuente: Elaboración propia)

El resultado de la sectorización se muestra en la Figura N° 41 en donde se puede comprobar que es igual a la sectorización inicial del proyecto. El subsector 1 de color verde es el sector 253, el subsector 2 de color rojo es el sector 254 y el subsector 3 de color magenta es el sector 255 en el proyecto inicial.

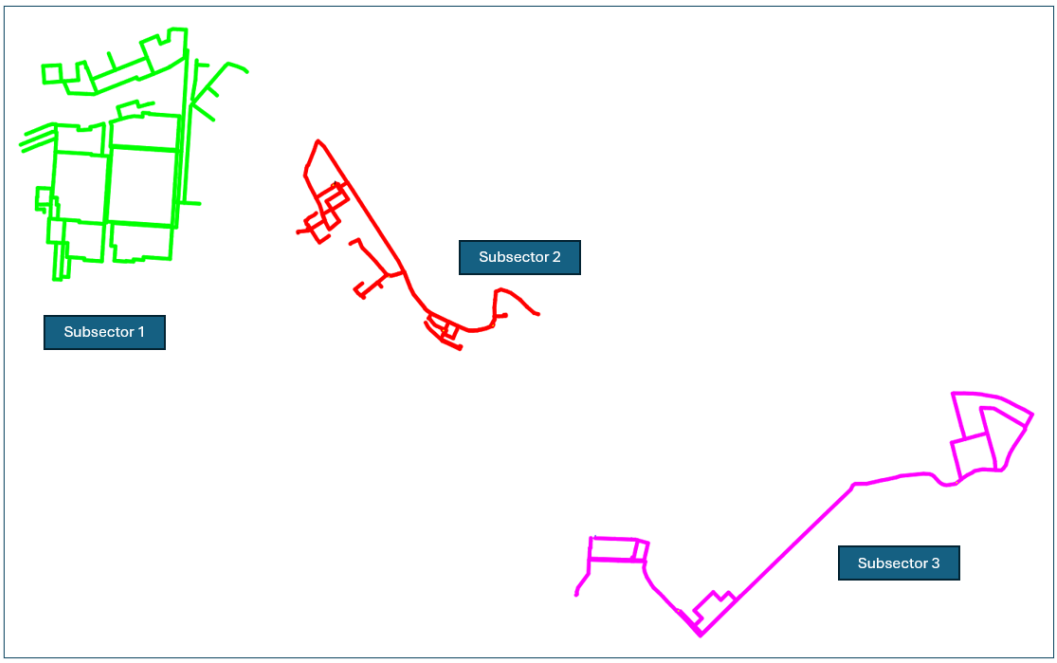

Figura N° 41. Resultados de la sectorización empleando Clustering Jerárquico

(Fuente: Elaboración propia)

En la Tabla N° 29 se tiene el resumen de la longitud total de los sectores definidos por el método del Clustering Jerárquico y los límites de las cotas de cada sector.

Tabla N° 29. Resumen de la longitud y límites de cotas de la sectorización

Número de sector	Long de tubería (m)	Cota máx. (m.s.n.m.)	Cota min (m.s.n.m.)
Subsector 1	12282.2	15.22	3.46
Subsector 2	3778.3	76.13	39.07
Subsector 3	5070.6	56.25	44.5

(Fuente: Elaboración propia)

Para el sector 258 alimentado por una fuente independiente de agua se repetirá el mismo procedimiento líneas arriba. En la tabla N° 30 se muestran los valores más altos del CCCF para las medidas métricas promedio, individual, completo, centroide y Ward para el sector 258.

Tabla N° 30. Valores más eficientes del número de particiones.

	Euclidiana	Manhattan	Gower
promedio	0.7671526	0.7379168	0.7839376
individual	0.5877672	0.5334133	0.6278521
completo	0.6538616	0.6251528	0.6386347
centroide	0.7253485	0.6951242	0.7291346
ward	0.578165	0.5633132	0.5282035

(Fuente: Elaboración propia)

La Figura N° 42, la Figura N° 43 y la Figura N° 44 muestran los dendrogramas respectivos para las tres combinaciones seleccionadas del sector 258 respectivamente.

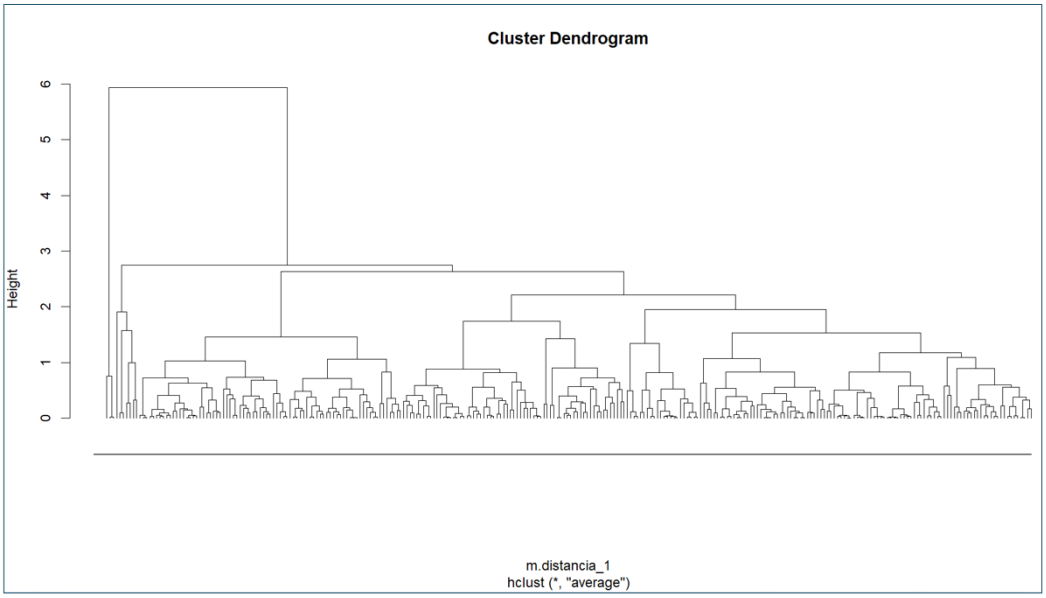

Figura N° 42. Dendrograma generado por la combinación distancia Euclidiana con método promedio
(Fuente: Elaboración propia)

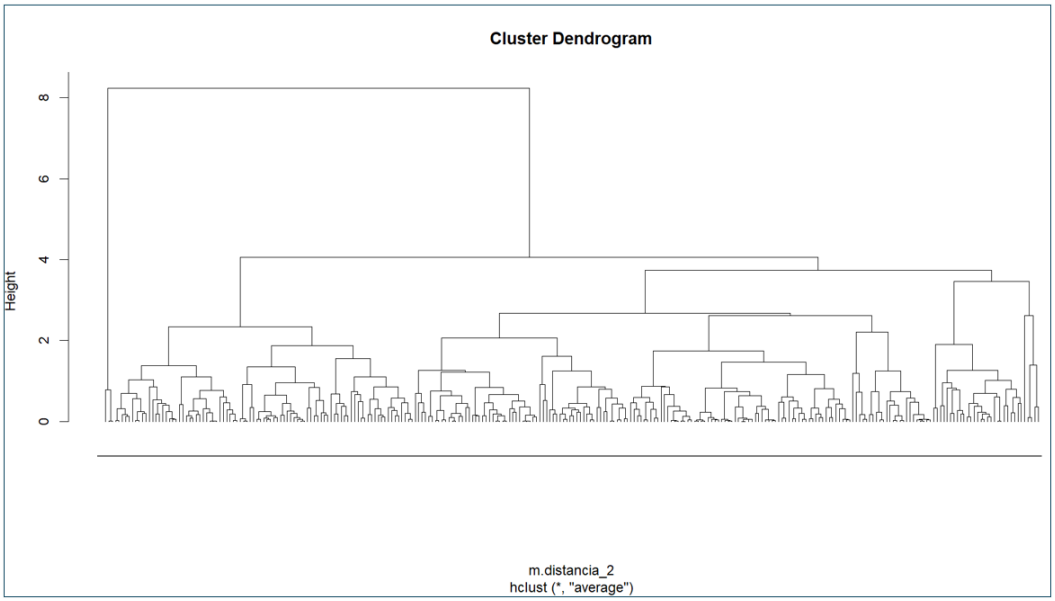

Figura N° 43. Dendrograma generado por la combinación distancia Manhattan con método promedio
(Fuente: Elaboración propia)

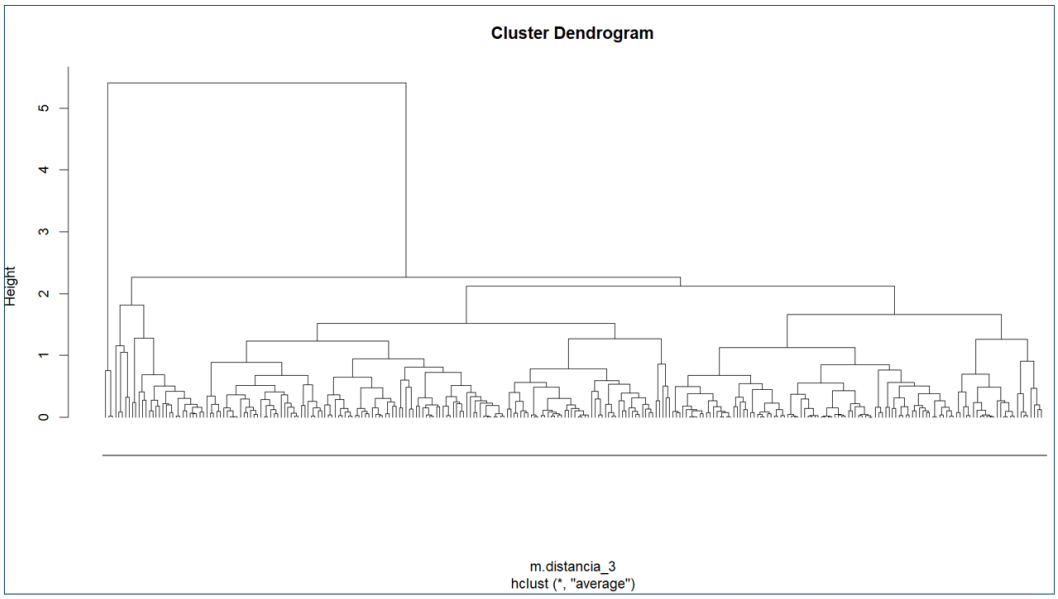

Figura N° 44. Dendrograma generado por la combinación distancia Gower con método promedio
(Fuente: Elaboración propia)

Como se puede apreciar el dendrograma resultante de la distancia Gower con el método promedio es el más homogéneo en la formación de clústeres. En la Tabla N° 31 se muestra el resultado de aplicar ClValid y se puede concluir que el número de particiones optimo es seis.

Tabla N° 31. Valores más eficientes del número de particiones.

	kmeans	pam	clara
WSS	7	6	6
SILHOUETTE	6	6	6
GAP_STAT	6	6	5

(Fuente: Elaboración propia)

En la Figura N° 45 aplicando el criterio del codo en la gráfica del método WSS resulta que el número óptimo de particiones es seis.

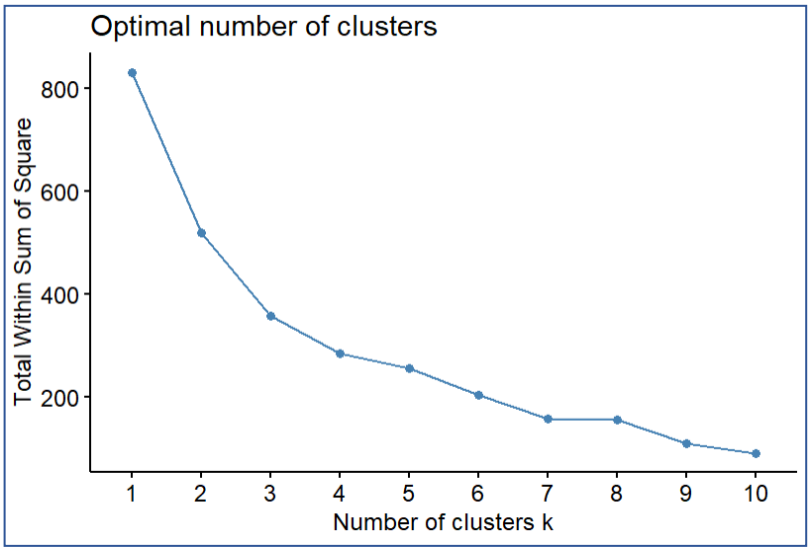

Figura N° 45. Grafica del método WSS usando el criterio del codo del sector 258

(Fuente: Elaboración propia)

En la Figura N° 46 la línea punteada muestra en la gráfica del método Silhouette que el número óptimo de particiones es seis.

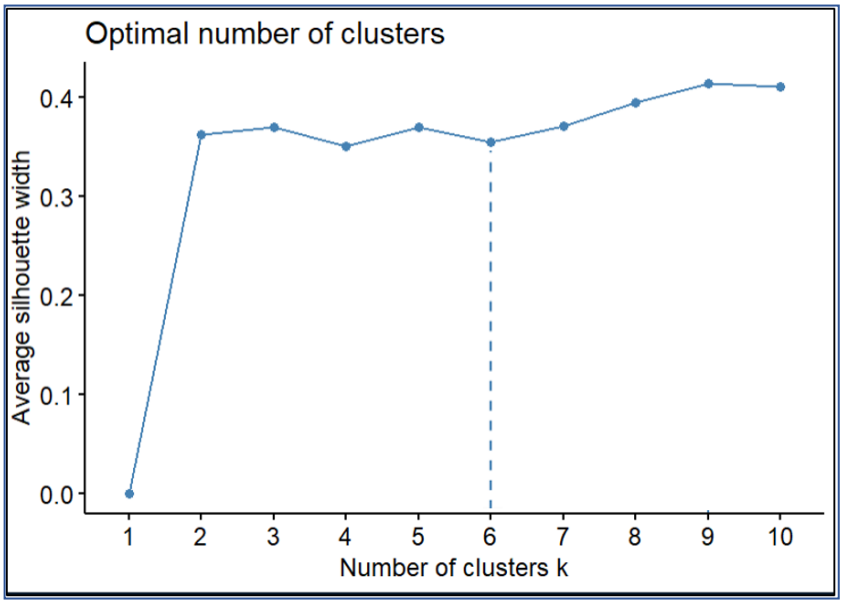

Figura N° 46. Grafica del método silhouette del 258

(Fuente: Elaboración propia)

En la Figura N° 47 la línea punteada muestra en la gráfica del método gap_stat que el número óptimo de particiones es seis.

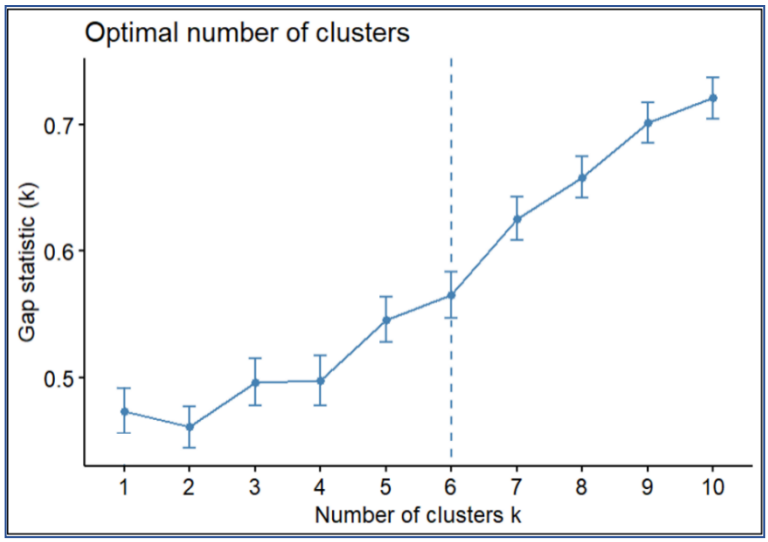

Figura N° 47. Grafica de barras del método gap_stat del sector 258
(Fuente: Elaboración propia)

El dendrograma resultante de aplicar el Clustering Jerárquico en el sector 258 muestra que se debe particionar en seis subsectores. (Ver Figura N° 48)

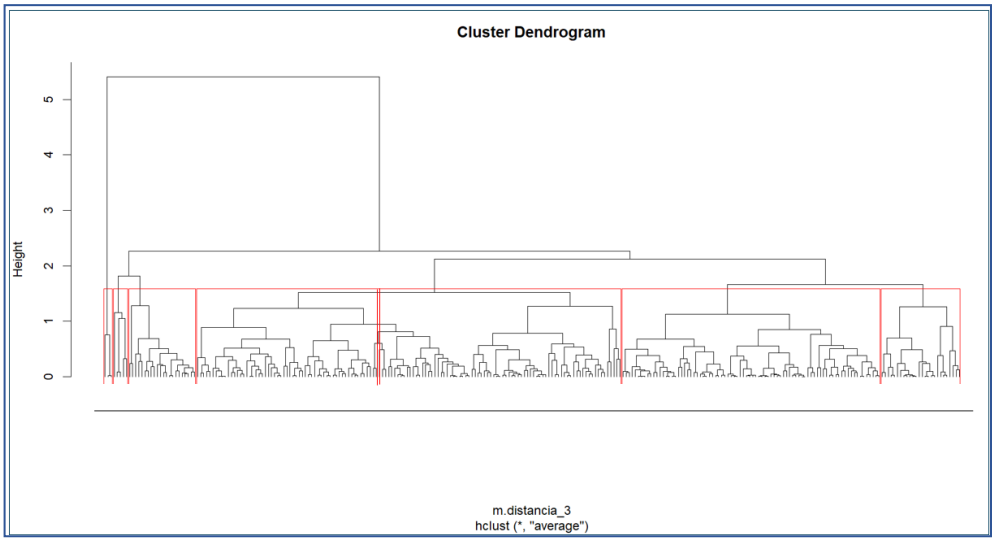

Figura N° 48. Dendrograma generado por la combinación distancia Manhattan con método promedio
(Fuente: Elaboración propia)

El resultado de la sectorización se muestra en la Tabla N° 32 donde se encuentran los límites de las cotas de los nuevos sectores formados. La sectorización del sector 258 en seis particiones se encuentra en la Figura N° 49.

Tabla N° 32. Resumen de la longitud y límites de cotas de la sectorización

Número de sector	Long de tubería (m)	Cota máx. (m.s.n.m.)	Cota min (m.s.n.m.)
Subsector 4	6245.14	79.15	26.54
Subsector 5	6515.45	54.5	45.58
Subsector 6	6448.56	72.33	33.78
Subsector 7	7891.45	52.51	41.87
Subsector 8	6945.62	32.09	22.43
Subsector 9	7656.14	54.5	29.01

(Fuente: Elaboración propia)

Figura N° 49. Resultados de la sectorización empleando Clustering Jerárquico

(Fuente: Elaboración propia)

Para el sector 259 alimentado por una fuente independiente de agua se repetirá el mismo procedimiento líneas arriba. En la tabla N° 33 se muestran los valores más altos del CCCF para las medidas métricas promedio, individual, completo, centroide y Ward para el sector 259.

Tabla N° 33. Valores más eficientes del número de particiones.

	Euclidiana	Manhattan	Gower
promedio	0.680916	0.6670687	0.7116518
individual	0.5413326	0.4320098	0.5600745
completo	0.5870436	0.6664152	0.6382151
centroide	0.6526081	0.6397309	0.6603967
ward	0.5903308	0.6254227	0.5393107

(Fuente: Elaboración propia)

La Figura N° 50, la Figura N° 51 y la Figura N° 52 muestran los dendrogramas respectivos para las tres combinaciones seleccionadas del sector 259 respectivamente.

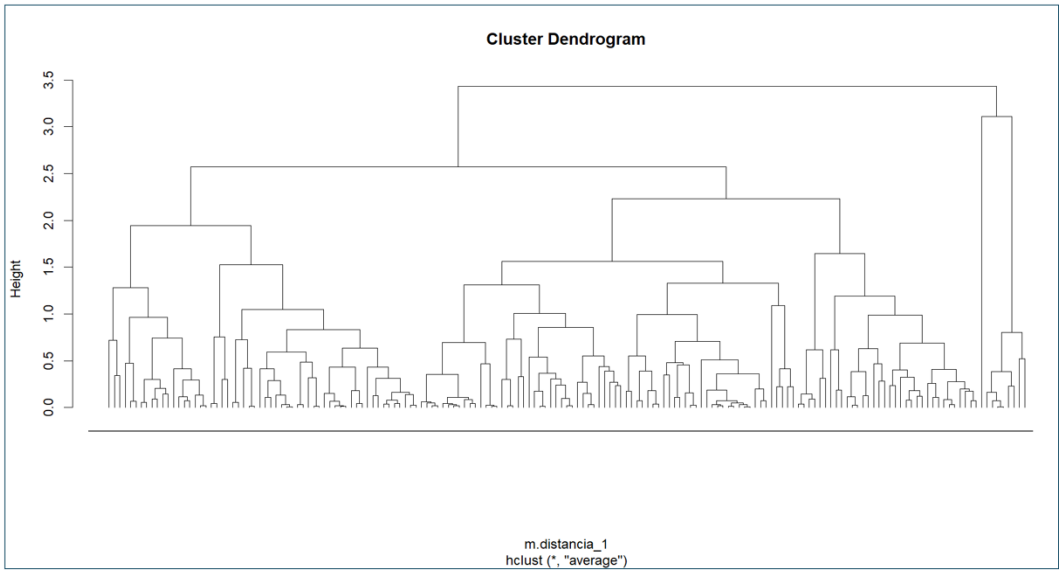

Figura N° 50. Dendrograma generado por la combinación distancia Euclidiana con método promedio

(Fuente: Elaboración propia)

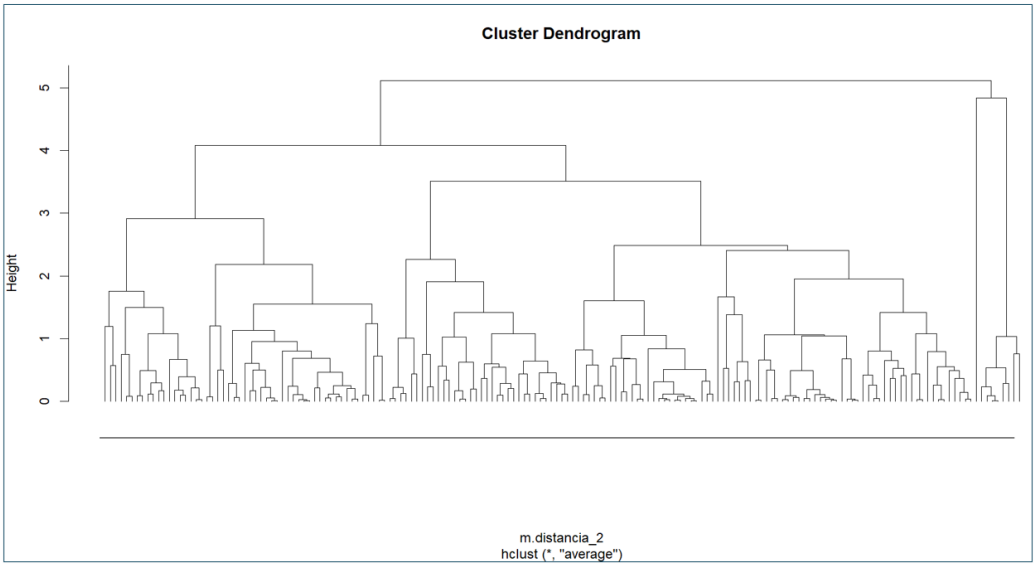

Figura N° 51. Dendrograma generado por la combinación distancia Manhattan con método promedio
(Fuente: Elaboración propia)

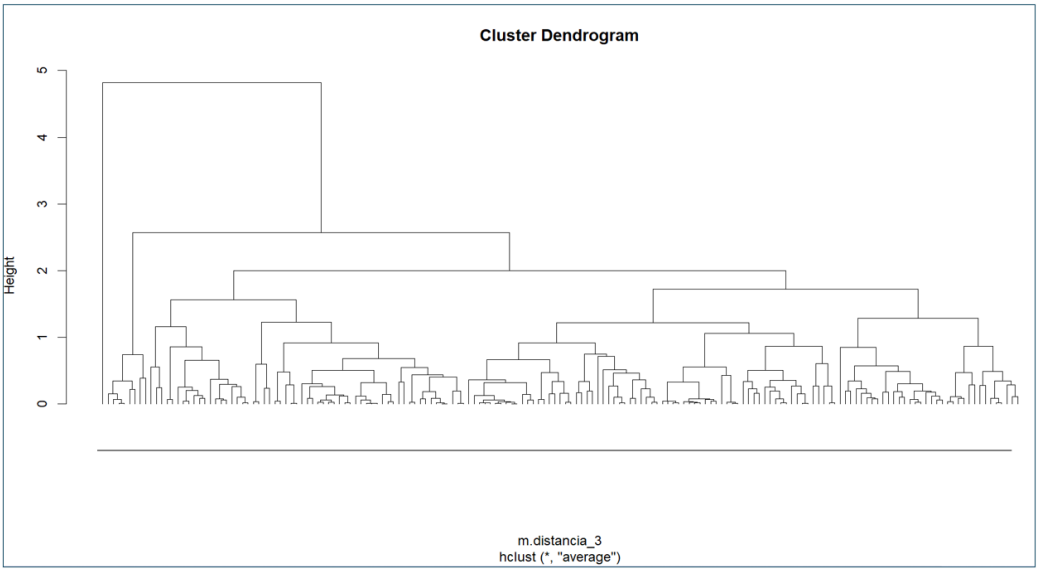

Figura N° 52. Dendrograma generado por la combinación distancia Gower con método promedio
(Fuente: Elaboración propia)

Como se puede apreciar el dendrograma resultante de la distancia Gower con el método promedio es el más homogéneo en la formación de clústeres. En la Tabla N° 34 se muestra el resultado de aplicar ClValid y se puede concluir que el número de particiones optimo es seis.

Tabla N° 34. Valores más eficientes del número de particiones.

	kmeans	pam	clara
WSS	7	6	6
SILHOUETTE	6	6	6
GAP_STAT	6	6	5

(Fuente: Elaboración propia)

En la Figura N° 53 aplicando el criterio del codo en la gráfica del método WSS resulta que el número óptimo de particiones es seis.

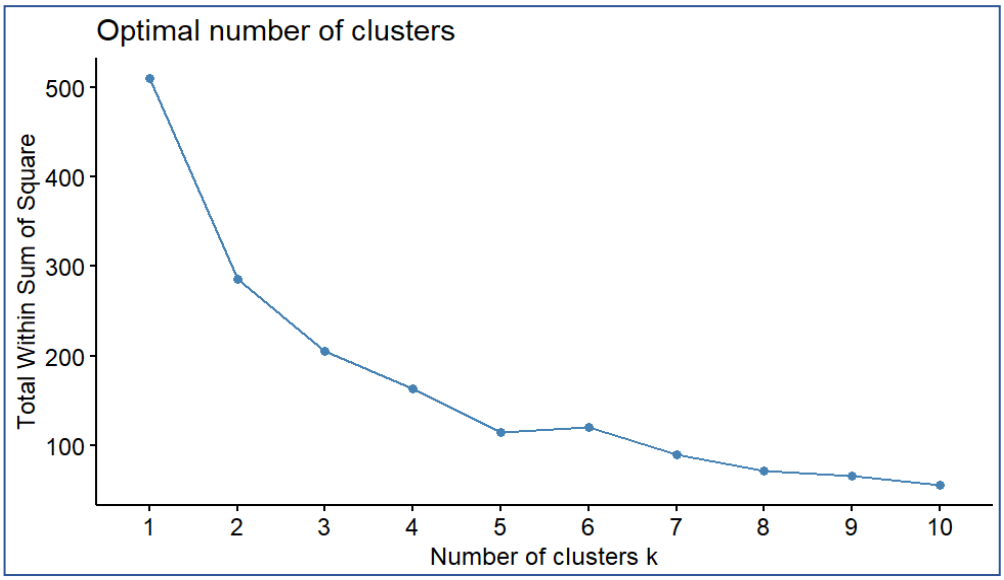

Figura N° 53. Grafica del método WSS usando el criterio del codo del sector 259

(Fuente: Elaboración propia)

En la Figura N° 54 la línea punteada muestra en la gráfica del método Silhouette que el número óptimo de particiones es seis.

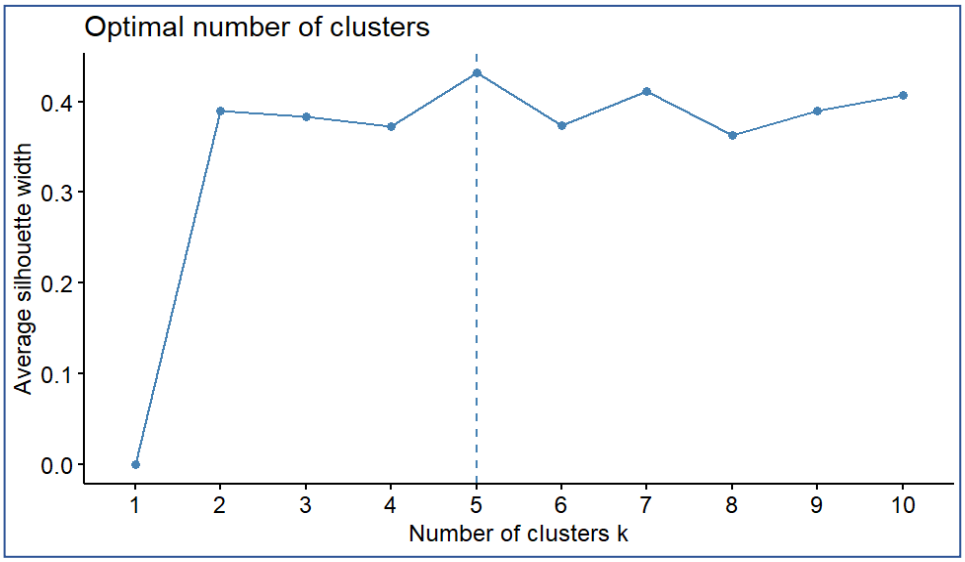

Figura N° 54. Grafica del método silhouette del 259

(Fuente: Elaboración propia)

En la Figura N° 55 la línea punteada de la gráfica del método gap_stat muestra que el número óptimo de particiones es seis.

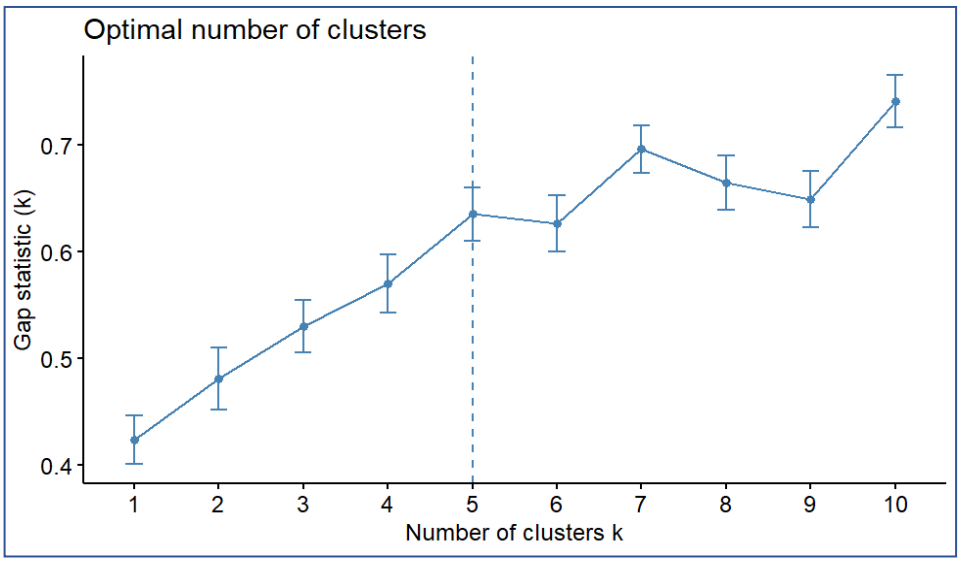

Figura N° 55. Grafica de barras del método gap_stat del sector 258

(Fuente: Elaboración propia)

El dendrograma resultante de aplicar el Clustering Jerárquico en el sector 259 muestra que se debe particionar en seis subsectores. (Ver Figura N° 56)

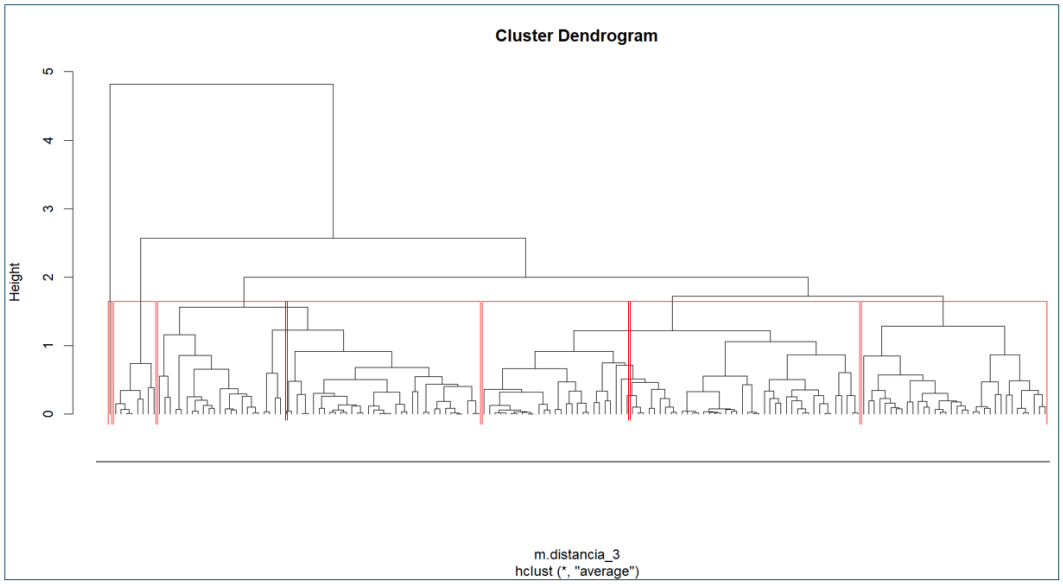

Figura N° 56. Dendrograma generado por la combinación distancia Manhattan con método promedio
(Fuente: Elaboración propia)

El resultado de la sectorización se muestra en la Tabla N° 35 donde se encuentran los límites de las cotas de los nuevos sectores formados. La sectorización del sector 259 en seis particiones se encuentra en la Figura N° 57.

Tabla N° 35. Resumen de la longitud y límites de cotas de la sectorización

Número de sector	Long de tubería (m)	Cota máx. (m.s.n.m.)	Cota min (m.s.n.m.)
Subsector 10	3097.60	72.83	66.08
Subsector 11	4355.26	81.86	57.19
Subsector 12	3184.94	55.38	50.45
Subsector 13	3510.12	62.07	54.93
Subsector 14	4243.21	61.17	50.01
Subsector 15	7095.624	65.43	55.03

(Fuente: Elaboración propia)

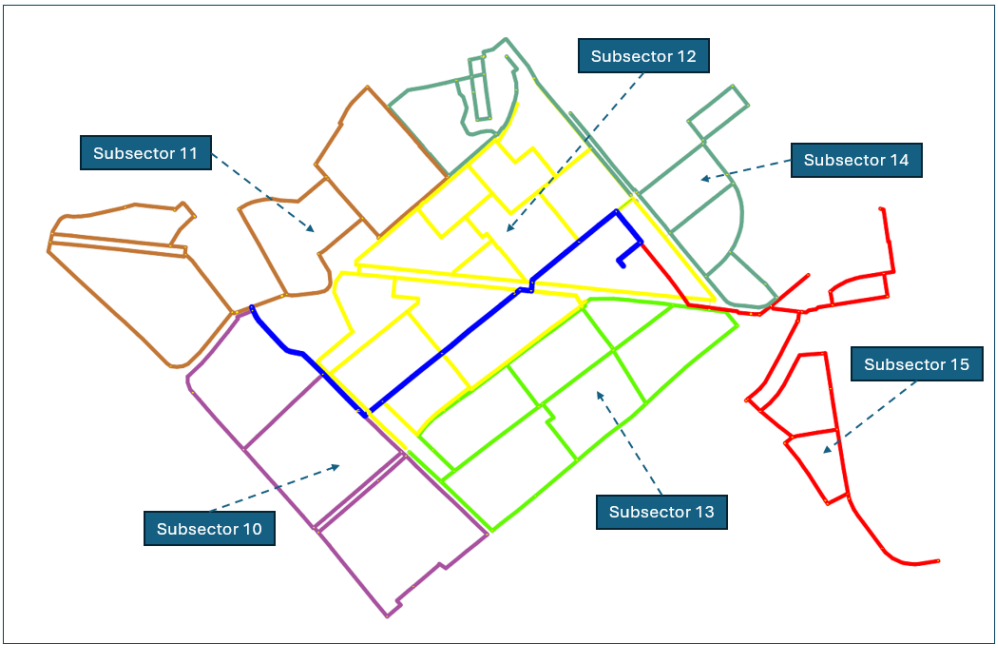

Figura N° 57. Resultados de la sectorización empleando Clustering Jerárquico
(Fuente: Elaboración propia)

Si bien la aplicación de clustering jerárquico deja algunos tramos y nodos separados de algunos sectores, estos deben unirse al sector más cerca posible debido a que en redes de agua potable no existe el caso de tuberías separadas de la fuente de agua.

La sectorización propuesta para la red distribución de agua potable (RDAP) “Víctor Raúl Haya de la Torre” considerara todos los subsectores generados después de aplicar la técnica Clustering Jerárquico. Al superponer todos los resultados se obtiene la Figura N° 58.

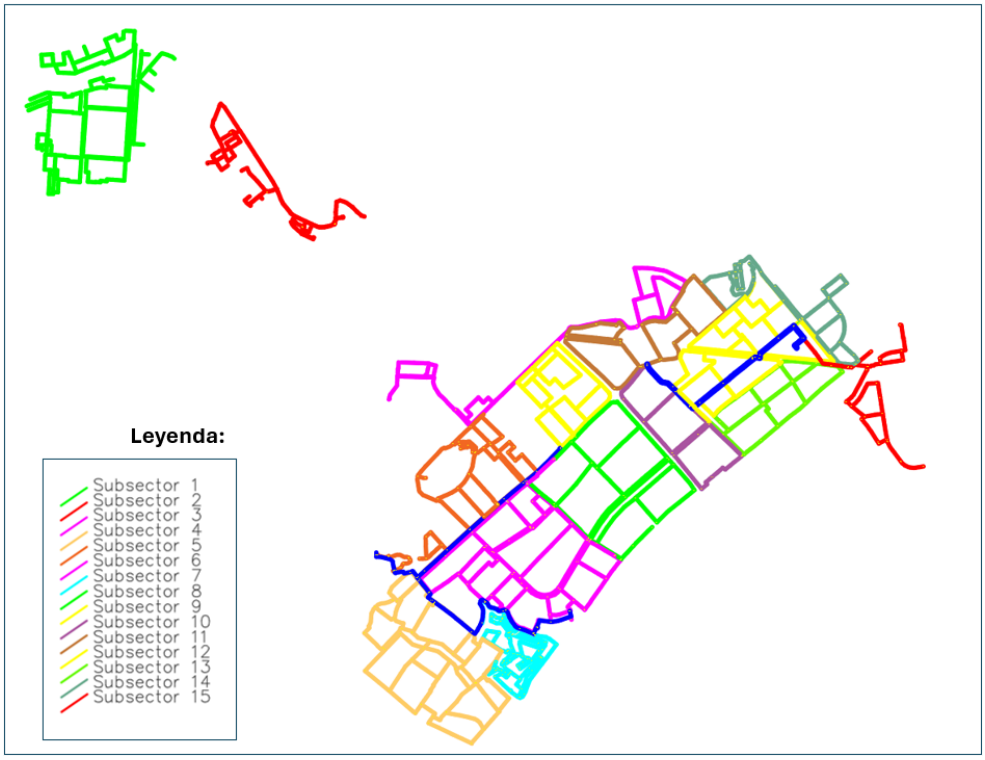

Figura N° 58. Resultados de la sectorización empleando Clustering Jerárquico promedio
(Fuente: Elaboración propia)

4.3 Implementación de optimización del conjunto de UOC en sectores mediante Algoritmos Genéticos y simulación Monte Carlo

Para la implementación se empleará el resultado obtenido mediante la definición de sectores usando el método de clustering jerárquico (ver Figura N° 59), en el cual se obtuvieron 15 subsectores con longitud de tuberías entre los 3 y 7 km. Los tres primeros sectores de la red analizada son independientes y cuentan con sus respectivas fuentes de agua; es así como se consideró una UOCs en la entrada a dichos sectores respectivamente. Para los últimos dos sectores que cuentan con una fuente de agua independiente se dividieron en seis subsectores cada uno. La red en cuestión se caracteriza por contar con sobrepresiones en puntos extremos, con valores entre los 65 mca en los puntos más bajos y 22 mca en los puntos más altos antes de la implementación de la sectorización aquí propuesta. Si bien implementar un número de UOC extra a la red incrementa el monto del proyecto, este se compara con el ahorro de agua a larga plazo y se le considerara como ganancia. Los costos de operación de la red se encuentran en la Tabla N° 36.

Tabla N° 36. Costos operativos de la red

Costo	Valor
Costo de agua (S/m3)	2.35
Costo de reparación (S/rotura)	774.28
Costo de transporte por cisterna (S/m3)	7.74
Costo de inspecciones (S/inspección)	57428.32

(Fuente: (SUNASS, 2021))

Si bien el costo por reparación depende del diámetro de la tubería, se consideró un valor promedio para ser conservadores en el análisis. Los costes por UOCs y válvulas de cierre se encuentran en la Tabla N° 37.

Tabla N° 37. Costos de UOC y válvulas

Diámetro (mm)	UOC (S/unidad)	Válvulas (S/unidad)
25	8702	1459
50	17408	2918
75	26110	4376
100	34816	5835
150	52223	8753
200	69635	11670
250	87043	14588
300	104450	17505
350	121858	20423
400	139270	23340
450	156678	26258
500	165379	27716

(Fuente: (SUNASS, 2021))

Según Lambert et al., 1999, para calcular el número de roturas futuras de la sectorización inicial y la sectorización propuesta se requiere la longitud de los subsectores y las conexiones en las redes secundarias respectivamente (Ver Tabla N°38 y 39).

Tabla N° 38. Número de roturas esperadas para LC principal y redes secundarias

Roturas	Según Lambert	
13	100km	principal
3	1000 conexiones	secundaria

(Fuente: Elaboración propia)

Tabla N° 39. Número de conexiones de la sectorización inicial

Descripción	Conexiones
Sector 1	782
Sector 2	474
Sector 3	490
Sector 4	3324
Sector 5	2040

(Fuente: Elaboración propia)

En la sectorización propuesta los últimos dos sectores (sector 258 y 259) se dividieron en 6 subsectores de los cuales se tuvieron que contabilizar la cantidad de conexiones y la longitud de las líneas de conducción principal resultantes de aplicar CMC. Cabe mencionar que para los primeros tres sectores la cantidad de conexiones se mantienen porque el método de clustering jerárquico los dejo como en la sectorización inicial. (Ver Tabla N° 40 y N° 41)

Tabla N° 40. Número de conexiones de la sectorización propuesta

Descripción	Conexiones
Subsector 1	782
Subsector 2	474
Subsector 3	490
Subsector 4	501
Subsector 5	560
Subsector 6	430
Subsector 7	654
Subsector 8	612
Subsector 9	567
Subsector 10	321
Subsector 11	350
Subsector 12	319
Subsector 13	336
Subsector 14	304
Subsector 15	410

(Fuente: Elaboración propia)

Tabla N° 41. Longitud de las líneas de conducción de la sectorización propuesta

Descripción	metros
Línea de conducción principal 1	4569.1
Línea de conducción principal 2	3622.7
Línea de conducción principal 3	1478.4

(Fuente: Elaboración propia)

Según Campbell (2017) para simular la ocurrencia y detección de roturas se establecieron distribuciones triangulares en función de la longitud de tuberías. En la Tabla N° 42 se muestra los porcentajes de detección basándose en el juicio de expertos debido a que no se cuenta con datos históricos de redes similares; dicha tabla está conformada por un valor máximo, más probable y mínimo. En el presente trabajo se considera el mismo porcentaje de detección de fugas reportadas y no reportadas para ser conservadores en la simulación, pero luego

tienen que ser geográficamente detectadas por medios acústicos. (Ver Tabla N° 42 y N° 43)

Tabla N° 42. Porcentaje de detección de la sectorización inicial

Porcentaje de detección		
Máximo	**Probable**	**Mínimo**
0.7	0.6	0.4
0.65	0.57	0.4
0.47	0.45	0.35
0.6	0.5	0.3
0.6	0.5	0.3

(Fuente: Elaboración propia)

Tabla N° 43. Porcentaje de detección de la sectorización propuesta

Porcentaje de detección		
Máximo	**Probable**	**Mínimo**
0.7	0.6	0.4
0.65	0.57	0.4
0.47	0.45	0.35
0.44	0.42	0.34
0.37	0.35	0.33
0.6	0.58	0.5

(Fuente: Elaboración propia)

Se definirá la distribución triangular utilizando la máxima, más probable y mínima cantidad de posibles roturas detectadas en un periodo de un año en la vida útil de la red. Este proceso se repite en cada sector y se procede a configurar la simulación a una SMC. (Ver Figura N° 59)

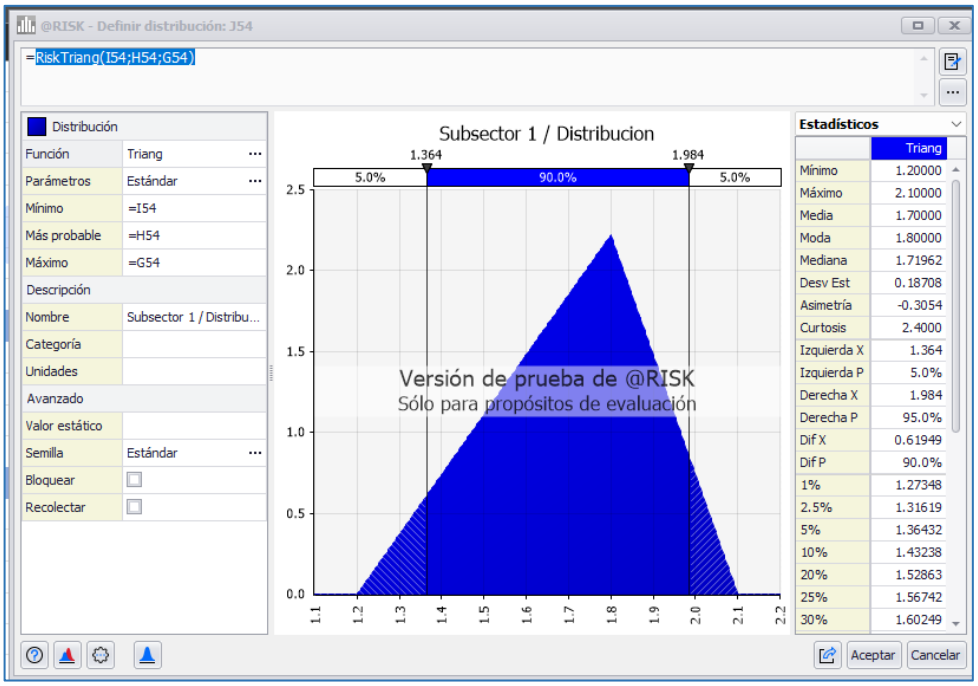

Figura N° 59. Valores máximos, más probable y mínimos para la simulación

(Fuente: Elaboración propia)

Para la sectorización inicial se aplicará la SMC en los sectores en función de la cantidad de conexiones, lo cual dio como resultado a 4 roturas en los primeros tres sectores y 8 roturas en los últimos dos sectores con una probabilidad de ocurrencia del 95%. (Ver Tabla N° 44 y N° 45)

Tabla N° 44. Número de roturas detectadas en los tres primeros sectores 253 (Sector 1), 254 (Sector 2) y 255 (Sector 3)

		Según Lambert		# roturas detectadas			SMC	
Descripción	conexión	roturas	# roturas	Max	Probable	Min	Distribución	
Sector 1	782	2.346	3	2.100	1.800	1.200	1.700	
Sector 2	474	1.422	2	1.300	1.140	0.800	1.080	
Sector 3	490	1.47	2	0.940	0.900	0.700	0.847	
	1746		7	4.340	3.840	2.700	3.627	4

(Fuente: Elaboración propia)

Tabla N° 45. Número de roturas detectadas en los tres primeros sectores 258 (Sector 4) y 259 (Sector 5)

		Según Lambert		# roturas detectadas			SMC	
Descripción	conexión	roturas	# roturas	Max	probable	min	Distribución	
Sector 4	3324	9.972	10	6.000	5.000	3.000	4.667	
Sector 5	2040	6.12	7	4.200	3.500	2.100	3.267	
	5364		17	10.200	8.500	5.100	7.933	8

(Fuente: Elaboración propia)

En la Figura N° 60 se muestra el resultado de la simulación Montecarlo de la sectorización inicial.

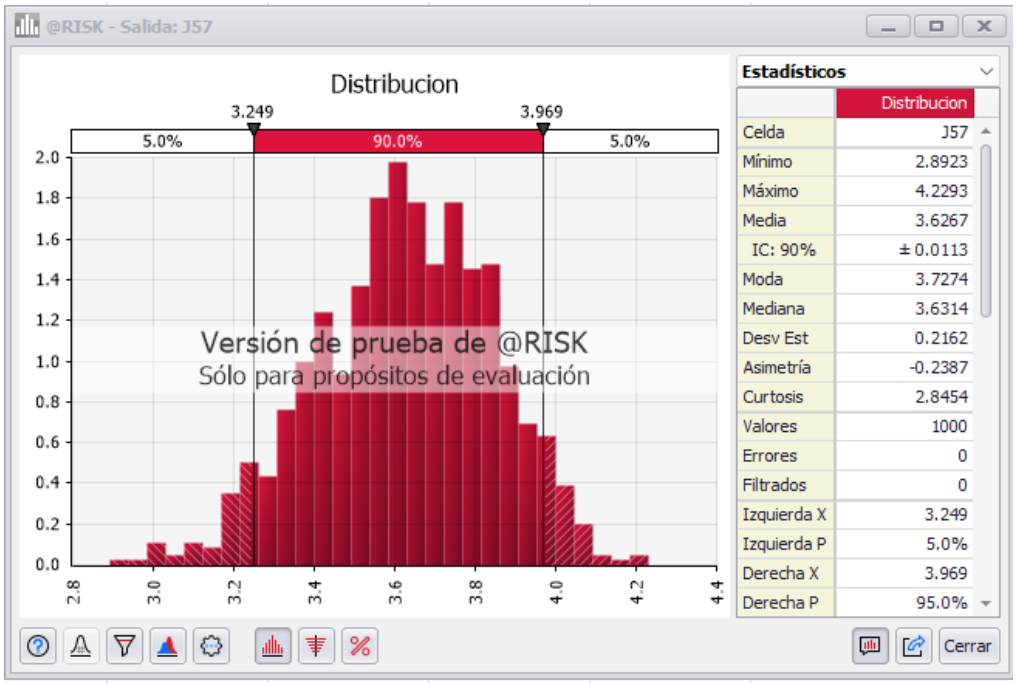

Figura N° 60. Simulación Monte Carlo de la sectorización inicial

(Fuente: Elaboración propia)

Para la sectorización propuesta, en los tres primeros sectores se predicen 4 roturas con una probabilidad del 95% de ocurrencia (Ver Tabla N° 46) y para los últimos dos sectores que se dividieron en 6 sectores cada uno se predicen 11 roturas. (Ver Tabla N°47 y N° 48)

Tabla N° 46. Número de roturas detectadas en los tres primeros sectores 253 (Subsector 1), 254 (Subsector 2) y 255 (Subsector 3)

		Según Lambert		# roturas detectadas			SMC	
Descripción	conexión	roturas	# roturas	Max	probable	min	Distribución	
Subsector 1	782	2.346	3	2.100	1.800	1.200	1.700	
Subsector 2	474	1.422	2	1.300	1.140	0.800	1.080	
Subsector 3	490	1.470	2	0.940	0.900	0.700	0.847	
	1746		7	4.340	3.840	2.700	3.627	4

(Fuente: Elaboración propia)

Tabla N° 47. Número de roturas del sector 258 aplicando la SMC

		Según Lambert		# roturas detectadas			SMC	
Descripción	und	roturas	# roturas	Max	probable	min	Distribución	
Subsector 4	501	1.503	2	1.400	1.200	0.800	1.133	
Subsector 5	560	1.68	2	1.300	1.140	0.800	1.080	
Subsector 6	430	1.29	2	0.940	0.900	0.700	0.847	
Subsector 7	654	1.962	2	0.880	0.840	0.680	0.800	
Subsector 8	612	1.836	2	0.740	0.700	0.660	0.700	
Subsector 9	567	1.701	2	1.200	1.160	1.000	1.120	
	3324		12	6.460	5.940	4.640	5.680	6

(Fuente: Elaboración propia)

Tabla N° 48. Número de roturas del sector 259 aplicando la SMC

		Según Lambert		# roturas detectadas			SMC	
Descripción	und	roturas	# roturas	Max	probable	min	Distribución	
Subsector 10	321	0.963	1	0.700	0.600	0.400	0.567	
Subsector 11	350	1.05	2	1.300	1.140	0.800	1.080	
Subsector 12	319	0.957	1	0.470	0.450	0.350	0.423	
Subsector 13	336	1.008	2	0.880	0.840	0.680	0.800	
Subsector 14	304	0.912	1	0.370	0.350	0.330	0.350	
Subsector 15	410	1.23	2	1.200	1.160	1.000	1.120	
	2040		9	4.920	4.540	3.560	4.340	5

(Fuente: Elaboración propia)

Tabla N° 49. Número de roturas de la sectorización de las LCP

		Según Lambert		# roturas detectadas			SMC	
Descripción	m	roturas	# roturas	Max	probable	min	Distribución	
LCP 1	4569.1	0.137	1	0.700	0.600	0.400	0.567	
LCP 2	3622.7	0.109	1	0.650	0.570	0.400	0.540	
LCP 3	1478.4	0.044	1	0.470	0.450	0.350	0.423	
	9670.2		3	1.820	1.620	1.150	1.530	2

(Fuente: Elaboración propia)

En la sectorización inicial se pueden estimar un total de 12 roturas detectadas y para la sectorización propuesta al dividir el área de los sectores se logran estimar un total de 17 roturas detectadas. Para calcular el beneficio total de la sectorización propuesta se deben considerar tres aspectos importantes; en primer lugar, se tiene la cantidad de agua que no se pierde por rotura reparada en un periodo de un año, en segundo lugar, el tiempo de respuesta entre la detección y reparación de la rotura y por último al coste de implementar UOC o válvulas de cierre para la entrada de los subsectores.

Como ya se ha visto capítulos anteriores, el implementar accesorios (en este caso serian UOC o válvulas de cierre) en una red de agua potable disminuye la energía transmitida a los nodos y, por ende, la presión disminuye ligeramente. Es por lo que se estableció como restricción el índice de resiliencia de la red con una variación máxima del 50% del índice de resiliencia de la sectorización inicial.

Para poder contabilizar la cantidad de agua que se pierde por una rotura en una red con un alto grado de mantenimiento o en su defecto en una red con poco tiempo de funcionamiento como es el caso de la presente tesis se aplicara el Modelo Conceptual propuesto por Thornton & Lambert.

Para el proceso de optimización se definió una población equivalente a 120 individuos que equivalen a 10 veces el número de variables de decisión (12 tuberías candidatas), una tasa de mutación del 0.5 % y una tasa de cruzamientos equivalente al 50 %. En la solución final, de las 12 tuberías candidatas, 10 se definieron como entradas de sector y el resto como límites de sector (válvulas de

cierre). La solución más optima se encontró en la iteración 80 como se puede apreciar en la Figura N° 61.

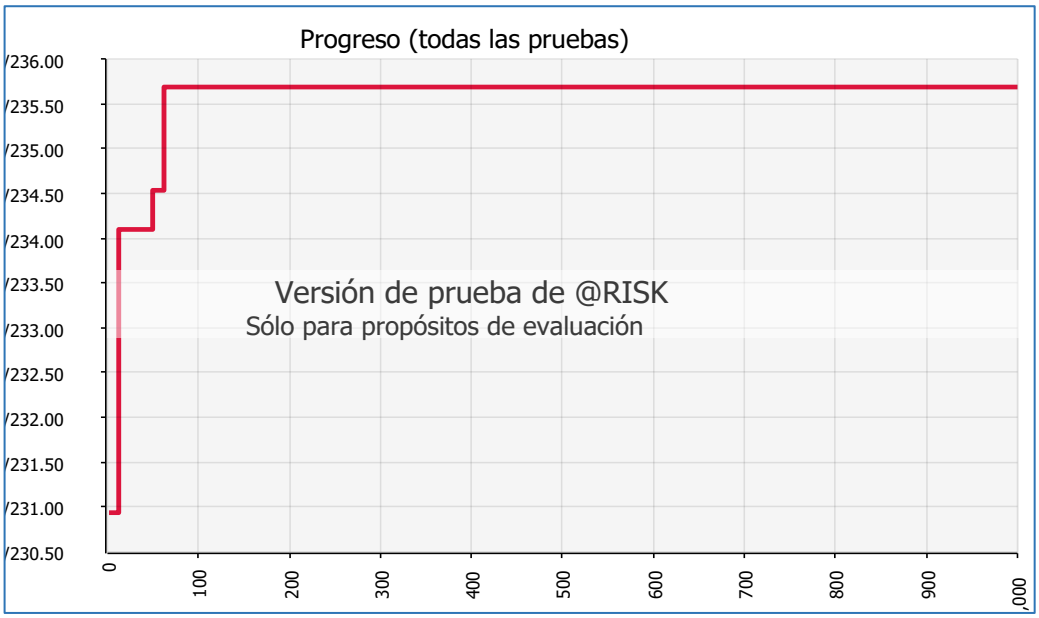

Figura N° 61. Comportamiento de la función objetivo

(Fuente: Elaboración propia)

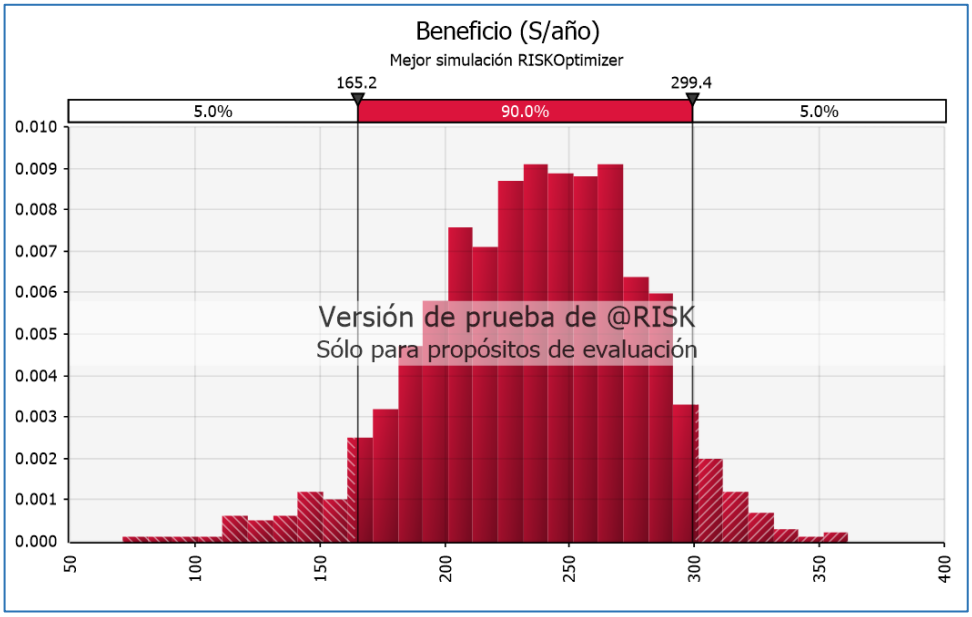

Figura N° 62. Variación de la simulación Monte Carlo en la Red Hidráulica

(Fuente: Elaboración propia)

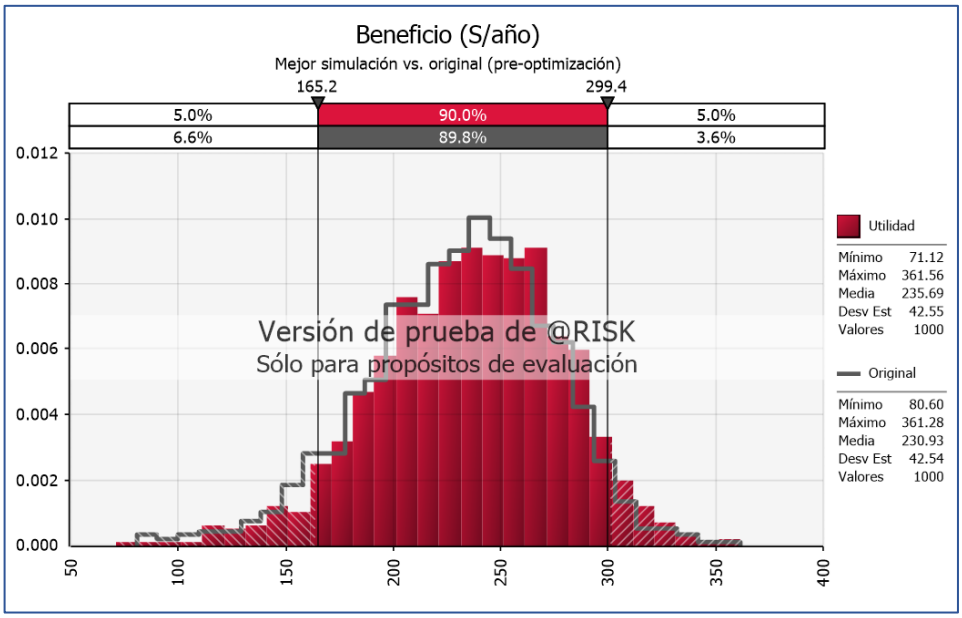

Figura N° 63. Comparación de la mejor simulación vs original

(Fuente: Elaboración propia)

Se obtuvo una media de S/. 235.69, con un mínimo de S/. 71.12 y un máximo de S/. 361.56 en beneficio por sobre los costos de inversión en UOC más la reparación de las roturas detectadas gracias a la implementación de la nueva sectorización. El eje horizontal indica el Beneficio en el primer año y el eje vertical muestra las probabilidades que se obtenga dicho resultado.

La tabla N° 45 muestra el Costo/Beneficio de la implementación de la nueva sectorización propuesta en la presente tesis, en donde se considera los ahorros por detección de roturas futuras, por inspección y el caudal que se ahorra al reparar las roturas futuras.

Capítulo V: Análisis y discusión de resultados

5.1 Fase I: Revisión de documentación del expediente técnico. Revisión de las líneas de conducción principal de la red hidráulica víctor Raúl Haya de la Torre

El primer punto importante de una sectorización de agua potable tratado es la definición de la línea de conducción principal, los criterios que se utilizaron para definirla se basan en minimizar la cantidad de energía perdida por fricción en las tuberías y el grado de conectividad de la de los nodos que pertenezcan a la LCP. El Concepto de Camino más Corto y el grado de conectividad fueron los criterios utilizados para definir la ruta de la LCP en cada sector (253, 254, 255, 258, 259) partiendo desde la fuente, no obstante, se requiere del criterio ingenieril para tener una mejor visión de la realidad y es por esta razón que fueron considerados las direcciones del caudal de agua en las tuberías.

Las líneas de conducción principal de la sectorización inicial cuentan con una UOC en la entrada de cada sector; para el sector 253 cuenta con una UOC en una tubería de DN 250 mm en la entrada del sector, en el sector 254 cuenta con una UOC en una tubería de DN 150 mm en la entrada del sector con cinco válvulas de cierre, en el sector 255 cuenta con una UOC en una tubería de DN 200 mm con dos válvulas de cierre, en el sector 258 cuenta con cuatro UOC en una tubería de DN 400 mm con nueve válvulas de cierre y para el sector 259 cuenta con cuatro UOC en tuberías de DN 300 mm con seis válvulas de cierre (Ver Figura N° 64).

En la sectorización inicial se cuenta con 5 líneas de conducción principal cada una repartida en los 5 sectores con una longitud entre los 600 y 2200 metros. Para el sector 253 la línea de conducción principal mide 602.2 metros, en el sector 254 la línea de conducción principal mide 1172.9 metros, en el sector 255 la línea de conducción principal mide 2112.9 metros, en el sector 258 la línea de conducción principal mide 1729.5 metros y en el sector 259 la línea de conducción principal mide 1478.4. (Ver Tabla N° 50)

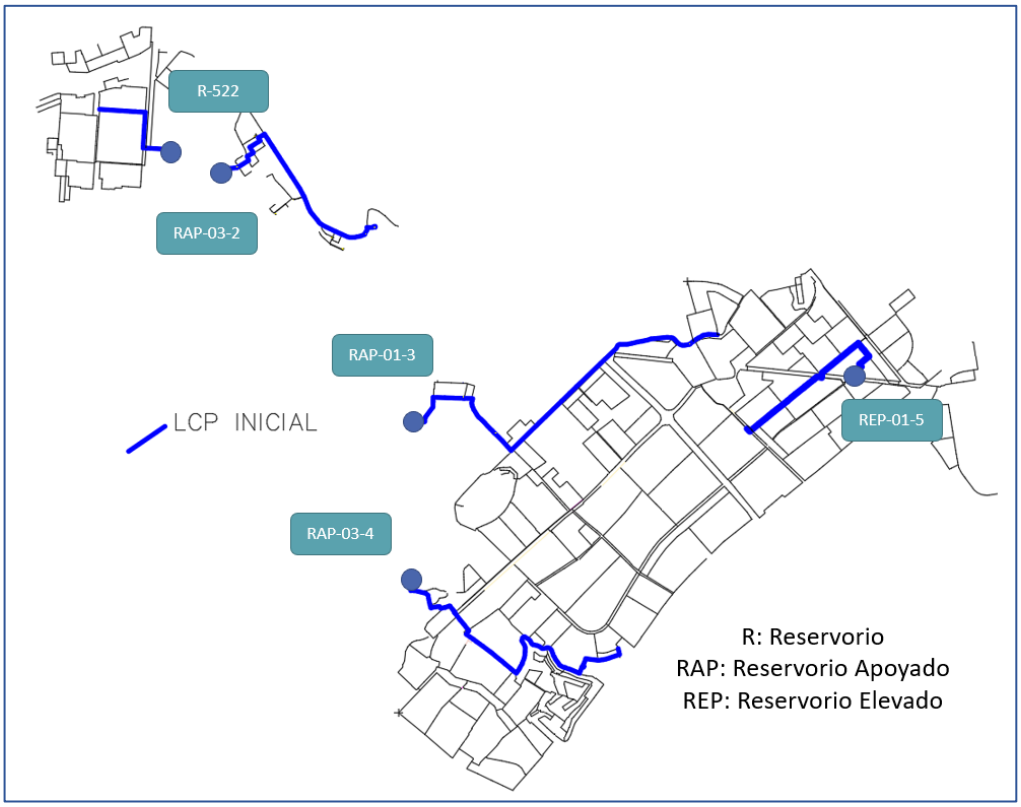

Figura N° 64. LCP de la sectorización inicial
(Fuente: Elaboración propia)

Para la sectorización propuesta, basándose en el concepto de caminos más cortos y el grado de conectividad de los nodos unos a otros, se llegó a la conclusión de que es más favorable para la red hidráulica ramificar algunas LCP para así llegar a más nodos y poder tener más control sobre los caudales. En el sector 253 se propone tener dos líneas de conducción principal en vez de una sola, la primera seria la misma LCP que en la sectorización inicial pero la segunda partiría del nodo J-74 con una longitud de 1363 metros. En el sector 254 se propone tener una LCP más larga con una longitud de 1422.9 metros y llega hasta el nodo J-78. En el sector 255 se propone tener una LCP más larga con una longitud de 2362.9 metros que llegue hasta el nodo J-28. En el sector 258 se propone tener dos LCP, en la primera ruta llegaría hasta el nodo J-116 con una longitud de 1979.5 metros y la segunda ruta llegaría hasta el nodo J-242 con una longitud de 1643.2 metros. En el sector 259 la LCP es la misma que en la sectorización inicial con una longitud de 1478.4 metros y llega hasta el nodo J-167. (Ver Tabla N° 50)

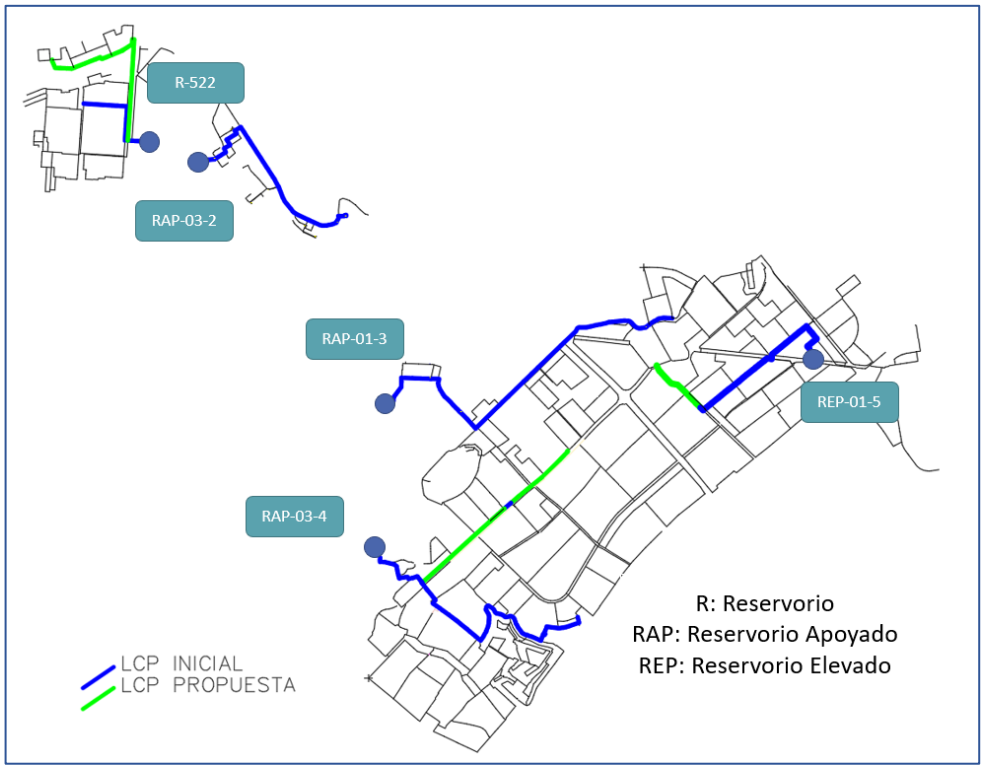

Figura N° 65. Comparación de la mejor simulación vs original

(Fuente: Elaboración propia)

Tabla N° 50. Longitud de las LCP de la sectorización propuesta e inicial

	Sectorización propuesta		Sectorización inicial	
Sector	**ID**	**Longitud de la LCP (m)**	**ID**	**Longitud de la LCP (m)**
253	LCP-1	602.2	LCP-1	602.2
	LCP-2	1363		
254	LCP-3	1422.9	LCP-2	1172.9
255	LCP-4	2362.9	LCP-3	2112.9
258	LCP-5	1979.5	LCP-4	1729.5
	LCP-6	1643.2		
259	LCP-7	1478.4	LCP-5	1478.4

(Fuente: Elaboración propia)

5.2 Revisión de los sectores de la red hidráulica víctor Raúl Haya de la Torre

En la sectorización inicial se cuenta con cinco sectores definidos por un área menor a 3 km2, en el sector 253 tiene un área de 0.98 km2, en el sector 254 tiene un área de 0.82 km2, en el sector 255 tiene un área de 0.83 km2, en el sector 258 tiene un área de 2.94 km2 y para el sector 259 tiene un área de 2.85 km2. Cabe mencionar que para la entrada en cada sector cuenta con una UOC y válvulas de cierre dentro de los sectores. (Ver Figura N° 66)

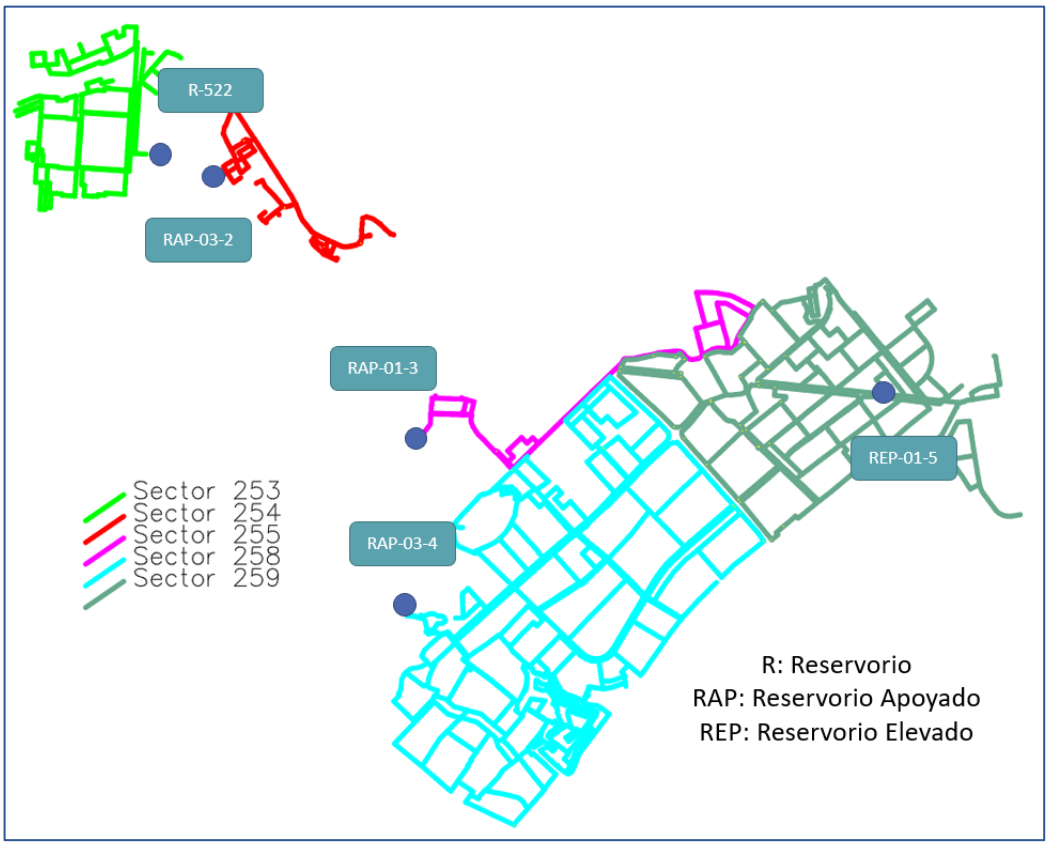

Figura N° 66. Sectorización inicial

(Fuente: Elaboración propia)

La sectorización propuesta se basa en el agrupamiento de nodos y tuberías en función de la similitud de sus características (método de clustering jerárquico). Se representa a la red hidráulica como un grafo debido a que la topología de la red es similar a la topología de un grafo. Para analizar el grafo formado por las tuberías y nodos se aplicó el lenguaje de programación estadístico R. Cabe mencionar que analizar los grafos de toda la red hidráulica tomaría demasiado poder computacional debido a la cantidad de datos y variables. Para aligerar el análisis del nuevo grafo se optó por analizar el problema en tres partes; la primera parte

es elegir los grafos de los primeros 3 sectores (253, 254 y 255) que dio como resultado 3 grupos que coinciden con los sectores iniciales, la segunda parte es elegir al grafo del sector 258 que dio como resultado a 6 subsectores y en la tercera parte es elegir al sector 259 que dio como resultado a 6 subsectores. Para la nueva sectorización se tendrán 15 subsectores de los cuales los tres primeros serán los mismos que los sectores iniciales y los otros 12 serán de los sectores 258 y 259. (Ver Figura N° 67)

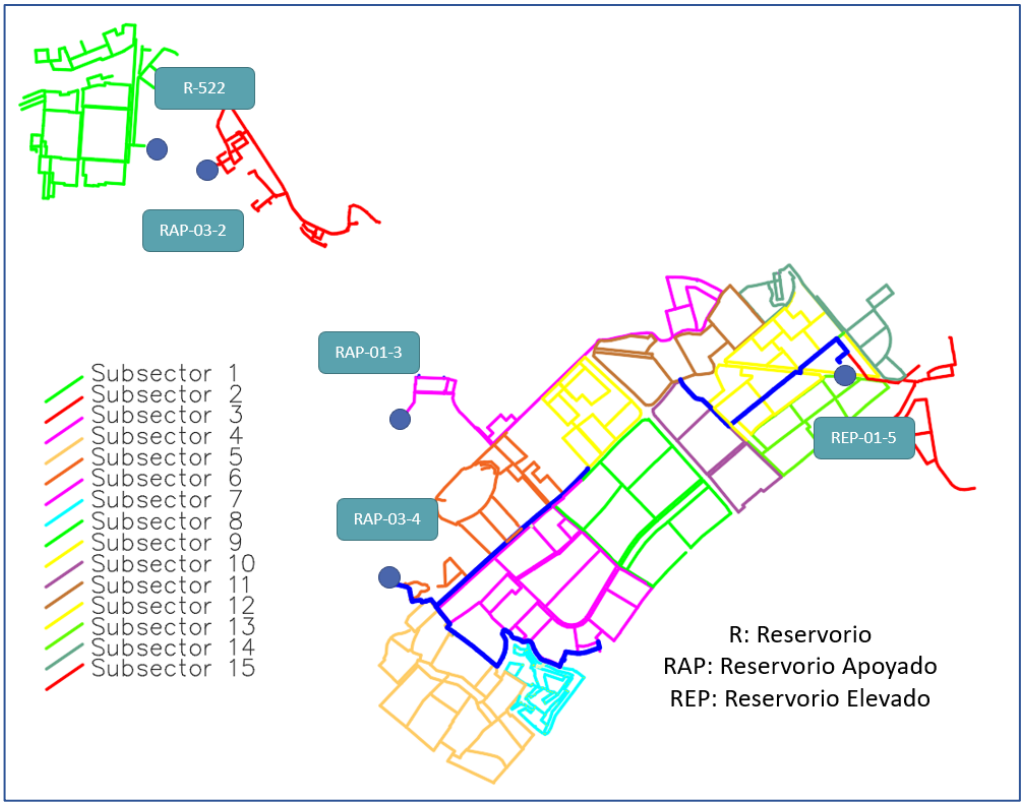

Figura N° 67. Sectorización propuesta
(Fuente: Elaboración propia)

Considerando la cantidad de subsectores se tomará en cuenta el número de válvulas de cierre para cada una de ellas. En la sectorización inicial se cuentan con 12 válvulas de cierre repartidas en los cinco subsectores, sin embargo, se necesitarán un mínimo de 3 válvulas de cierre más para la nueva sectorización. Como resultado en la tabla N° 51 se muestra el tamaño en términos de longitud de tuberías de cada sector y subsector de la sectorización inicial y propuesta respectivamente.

Tabla N° 51. Longitud de tuberías de la Sectorización propuesta e inicial

Sectorización propuesta		Sectorización inicial	
N° de subsector	**Long de tubería (m)**	**N° de Sector**	**Long de tubería (m)**
Subsector 1	12282.2	Sector 253	12282.2
Subsector 2	3778.3	Sector 254	3778.3
Subsector 3	5070.6	Sector 255	5070.6
Subsector 4	6245.14	Sector 258	41702.36
Subsector 5	6515.45		
Subsector 6	6448.56		
Subsector 7	7891.45		
Subsector 8	6945.62		
Subsector 9	7656.14		
Subsector 10	3097.6	Sector 259	25486.754
Subsector 11	4355.26		
Subsector 12	3184.94		
Subsector 13	3510.12		
Subsector 14	4243.21		
Subsector 15	7095.624		

(Fuente: Elaboración propia)

5.3 Optimización de conjunto de entradas y válvulas de cierre de la red hidráulica víctor Raúl Haya de la Torre

Considerando que la cantidad de roturas detectadas en una red es directamente proporcional al tamaño de la red (longitud de tuberías) y al detectar más roturas, el volumen de agua potable perdida por las fugas se reduce. No obstante, para implementar la nueva sectorización se requiere de un número de UOC y válvulas de cierre lo cual significa una inversión económica.

La sectorización inicial contiene un conjunto de UOC y válvulas de cierre, cuenta con 12 válvulas de cierre y 6 UOC distribuidas en los 5 sectores. Una vez definido los subsectores de la sectorización propuesta se procede a la refusión de nodos que quedaron fuera de los grupos formados luego de aplicar clustering jerárquico a la red. Como se mencionó líneas arriba la mínima cantidad de válvulas de cierre que se deben agregar son 3. La nueva sectorización dividió el tamaño de los

últimos dos sectores (258 y 259) en 6 subsectores cada uno, para calcular la posible cantidad de roturas detectadas extra se aplicará la Simulación Monte Carlo con ayuda del software RiskOptimizer. Para estimar el volumen de agua potable que no se pierde por detectar roturas extra en un periodo de un año se aplica la fórmula de Lampert. Para optimizar la función objetivo que calcula el número de UOC y válvulas de cierre se aplicaron algoritmos genéticos, como restricción se establece una variación máxima del 50 % del índice de resiliencia y un número de 1000 iteraciones. El software RiskOptimizer utiliza algoritmos genéticos y la simulación Montecarlo al mismo tiempo, también incluye gran versatilidad para definir las celdas ajustables e ingresar las restricciones de la función objetivo.

Antes de correr el modelo de optimización en RiskOptimizer se configura el número de núcleos que se utilizaran para el proceso. Para el presente trabajo se utilizó un procesador Core i7 9750H con 6 núcleos, 12 hilos más 16 GB RAM a 3200 Hz logrando resultados en un tiempo estimado de 55 minutos.

En la red hidráulica al agregar más accesorios (UOC y válvulas de cierre) se generan perdidas locales y, por ende, la energía que llega a los nodos disminuye. Simulando la red hidráulica con la nueva sectorización en EPANET se produce una variación en las presiones de los nodos (Ver Tabla N° 52) que, a su vez, produce una variación en el índice de resiliencia. (Ver Tabla N° 53)

Tabla N° 52. Presión de nodos con la Sectorización Propuesta

Subsector	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Índice de uniformidad de presión	29.15	65.35	67.77	29.65	38.35	63.35	63.49	32.75	51.35	28.8	65.35	67.42	29.3	38	63.35
Presión promedio (mca)	12.1	40.72	33.78	60.56	51.72	68.54	55.75	45.63	49.58	75.69	88.69	87.69	91.56	101.5	106.7
Presión mínima (mca)	10.67	10.21	14.93	10.75	11.2	10.29	10.78	11.21	11.02	65.58	68.69	71.54	79.62	81.56	76.32
presión máxima (mca)	14.12	55.28	43.66	70.2	66.5	78.25	82.4	81.6	87.5	89.96	96.56	98.68	96.65	113.6	121.7

(Fuente: Elaboración propia)

Tabla N° 53. Desviación del índice de resiliencia (Ir)

	Sectorización inicial	Sectorización propuesta	Desviación del Ir
	Ir	Ir	∆ Ir
Sector 259	0.631	0.541	14.26%
Sector 258	0.505	0.415	17.81%
Sector 253	0.675	0.585	13.33%
Sector 254	0.543	0.453	16.58%
Sector 255	0.675	0.585	13.33%

(Fuente: Elaboración propia)

El balance de costo beneficio de la sectorización inicial versus la sectorización propuesta en un año se muestra en la Tabla N° 54, en donde se tiene un costo total de inversión inicial de S/. 186,906.40 soles, pero con un beneficio de S/. 187,142.09 soles. En la inversión inicial comprende 5 válvulas de cierre de 150mm, 2 UOC de 200mm y 5 nuevas roturas detectadas. El beneficio comprende 4 roturas no detectadas, un caudal de 29,441.00 m3 ahorrado y 2 inspecciones por medios acústicos. El beneficio total para un año es de S/. 235.69 soles.

Tabla N° 54. Balance Costo/Beneficio obtenido

Costo	Costo (S/.)
Inversión en UOC y válvulas (S/año)	183,035.00
Reparación de roturas detectadas	3,871.40
Coste total (S/año)	186,906.40
Beneficio	**Beneficio (S/.)**
Roturas detectadas (S/año)	3,097.12
Caudal por roturas detectadas (S/año)	69,188.33
Inspecciones (S/año)	114,856.64
Beneficio Total (S/año)	187,142.09
Ahorro Total (S/año)	235.69

(Fuente: Elaboración propia)

Conclusiones

La aplicación de la teoría de grafos para la optimización de redes hidráulicas, empleando métodos de Clustering Jerárquico y algoritmos genéticos, ha demostrado ser una propuesta de sectorización efectiva. En un periodo de un año, este enfoque ha generado resultados positivos, con un beneficio que supera en un 0.12% los costos de implementación, sin comprometer la calidad y continuidad del servicio.

El trazo propuesto, basado en los caudales de mayor demanda y el concepto de Caminos más Cortos, ha revelado que los sectores 253, 254 y 255 comparten las mismas líneas de aducción que la sectorización inicial. Por otro lado, la ruta propuesta para los sectores 258 y 259 presenta dos ramificaciones adicionales en comparación con la sectorización inicial, lo que permite abarcar una mayor área de la red y mejorar la conectividad con los subsectores que se formaran con los pasos siguientes de la sectorización.

Se establecieron 15 subsectores en la red de distribución de agua potable utilizando la técnica de Clustering Jerárquico en la sectorización propuesta. En la sectorización inicial, los sectores 253, 254 y 255 conservaron la misma cantidad de subsectores. En contraste, los sectores 258 y 259 se dividieron en 6 subsectores cada uno, lo que permite un mayor control sobre aspectos importantes gracias a la reducción del área de cada subsector, en comparación con los 2 subsectores iniciales.

La optimización de la ubicación de las válvulas de cierre y puntos de abastecimiento se basó en el análisis costo/beneficio en donde el arreglo final debe cumplir con las mínimas presiones e índice de resiliencia, mediante la Simulación Montecarlo se obtuvo un ahorro de 29, 476.00 m3 de agua potable que no se pierde por fugas.

Los beneficios económicos de la sectorización propuesta se traducen en un ahorro anual de S/.187,142.09, conformado por S/.3,097.12 en roturas detectadas, S/.69,188.33 en caudal por roturas detectadas y S/.114,856.64 en inspecciones. Al comparar este ahorro con el costo de implementación, se obtiene un beneficio neto de S/.235.69 anuales, equivalente al 0.12% del costo total. Si bien este porcentaje puede parecer bajo, es importante tener en cuenta que estos resultados corresponden únicamente al primer año de implementación de la sectorización.

Recomendaciones

Se recomienda evaluar otras alternativas de sectorización aparte de la empleada en la presente investigación para comparar resultados como, por ejemplo: Método de sectorización basado en Detección Multinivel de Comunidades en Redes Sociales y Método de Sectorización basado en la Detección de Comunidades Mediante Caminos Aleatorios.

Se recomienda usar los métodos de inclusión de Válvulas Reguladoras de Presión en las Entradas de Sectores Mediante Optimización Multinivel y Optimización del Conjunto de Entrada de Sectores/Válvulas de Cierre Mediante Optimización de Enjambre de Agentes en reemplazo de los algoritmos genéticos.

Se recomienda investigar la Metodología BABE para redes de agua potable con varios años en funcionamiento y que cuenten con un registro de consumo para clasificar los tipos de perdidas en el análisis de costo/beneficio como nuevo tema de investigación.

Referencias Bibliográficas

Abonyi, J., & Balázs, F. (2007). Cluster Analysis for Data Mining and System Identification. Birkhäuser Verlag AG.

Ahuja, R., Magnanti, T., & Orlin, J. (1993). Network Flows: Theory, Algorithms, and Applications. Prentice Hall.

Alamo, M. A. (2015). USO DEL CRITERIO AHP PARA LA TOMA DE DECISIONES. UNIVERSIDAD NACIONAL AGRARIA LA MOLINA.

Álvarez, C. (2020). ¿Qué significa que el agua empiece a cotizar en el mercado de futuros de Wall Street? Obtenido de El País: www.elpais.com

Alvisi, S., & Franchini, M. (2014). A heuristic procedure for the automatic creation of district metered areas in water distribution systems. Urban Water Journal, 11, 37-159.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. Science, 286, 5439.

Bastian, M., Heyman, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating network. In Proc. International AAAI Conference on Weblogs and Social Media.

Bataglj, V., & Mrvar, A. (2002). Pajek – Analysis and Visualization of Large Network. Lecture Notes in Computer Science, Springer.

Bilgin, A., Ellson, J., Gansner, E., & North, Y. (2017). Graphviz. Obtenido de http://www.graphviz.org/users/arif-bilgin

Brock, G., Pihur, V., & Datta, S. (2008). Datta. ClValid. An R package for cluster validation. Journal of Statistical Software, 25, 1-22.

Campbell, E. O. (2017). Sectorización de Redes de Abastecimiento de Agua Potable basada en detección de comunidades en redes sociales y optimización heurística. Valencia.

Creaco, E., Fortunato, A., Franchini, M., & Mazzola, M. (2013). Comparison between entropy and resilience as indirect measures of reliability in the framework of water distribution network design. In Proc.12th International Conference on Computing and Control for the Water Industry, 70, 379-388.

Csardi, G., & Nepusz, T. (2017). The igraph software package for complex network research. Obtenido de Inter Journal: http://igraph.org

Cuadras, C. (2014). Introducci´on al An´alisis Cluster. Obtenido de https://www.ugr.es/~gallardo/pdf/cluster-g.pdf

Di Nardo, A., Di Natale, M., Santonastaso, G., Tzatchkov, V., & Alcocer Yamanaka, V. (2013). Water network sectorization based on a genetic algorithm and minimum dissipated power paths. Water Since and Technology: Water Supply, 13, 951-957.

Dunn, J. (1974). Well-separated clusters and optimal fuzzy partitions. Journal of Cybernetics, 4, 95-104.

Everitt, B., Landau, S., Leese, M., & Stahl, D. (2011). Cluster Analysis. John Wiley & Sons. Inc.

Fortunato, S., & Barthélemy, M. (2007). Resolution limit in community detection. In Proc. National Academy of Sciences, 104, 36-41.

Han, J., Kamber, M., & Pei, J. (2006). Data Mining: Concepts and Techniques. Morgan Kaufmann.

Handl, J., Knowles, J., & Kell, D. (2005). Computational cluster validation in post-genomic data analysis. Bioinformatics, 21, 3201-3212.

IWA (International Water Association). (2000). Performance Indicators for water Supply Services. London.

Kaufman, L., & Rousseeuw, P. (1990). Finding Groups in Data: An introduction to Cluster Analysis. Wiley.

Kingdom, B., Liemberger, R., & Philippe, M. (2006). The Challenge of Reducing Non-Revenue Water (NRW) in Developing Countries. The World Bank Group.

Lambert, A., Brown, T., Takizawa, M., & Weimer, D. (1999). A review of performance indicators for real losses from water supply systems. AQUA.

Leon Escobar, A., & Giraldo Suarez, E. (2005). Implementación de Algoritmos de recorrido de Grafos para el Calculo de la Regulación en Redes de Distribución Radiales.

Manning, C., Raghava, P., & Schutze, H. (2008). Introduction to Information Retrieval. Obtenido de http:http://nlp.stanford.edu/IR-book/html/htmledition/irbook.html

More, E. (1959). The shortest path through a maze. In Proc. International Symposium on the Theory of Switching.

Newman, M. (2003). The structure and function of complex networks. SIAM Review, 45, 167-256.

Palisade. (2010). Evolver. The Genetic Algorithm Solver for Microsoft Excel.

Pilcher, R., Stuart, H., Chapman, H., Field, D., Ristovski, B., & Stapely, S. (2007). Leak location and repair: guidance notes. Water Loss Task Force. International Water Association (IWA).

PNUMA. (2019). Programa de la ONU para el Medio Ambiente. Programa de las Naciones Unidas para el Medio Ambiente.

R Core, T. (2015). R: A Language and Environment for Statistical Computing. Obtenido de https://www.R-project.org/

Romesburg, C. (2004). Cluster analysis for researches. Lulu Press.

Sokal, R., & Michener, C. (1958). A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin, 28, 1409-1438.

Steinhaeuser, K., & Chawla, N. (2010). Identifying and evaluating community structure in complex networks. Pattern Recognition Letters, 31, 413-421.

SUNASS. (2021). Benchmarking Regulatorio de las Empresas Prestadoras.

SUNASS. (2023). BENCHMARKING REGULATORIO 2023 DE EMPRESAS PRESTADORAS.

Thornton, J., & Lambert., A. (2007). Pressure Management extends infrastructure life and reduces unnecessary energy cost. Water Loss 2007. Bucharest, Romania.

Thulasiraman, K., & Swamy, M. (1992). Graph Algorithms, in Graphs: Theory and Algorithms. John Wiley & Sons.

Todini, E. (2000). Looped water distribution networks design using a resilience index. Urban Water, 2, 115-122.

Vegas Niño, O. T. (2012). HERRAMIENTAS DE AYUDA A LA SECTORIZACIÓN DE REDES DE ABASTECIMIENTO DE AGUA BASADOS EN LA TEORÍA DE GRAFOS APLICANDO DISTINTOS CRITERIOS. Valencia.

Watts, D., & Strogatz, S. (1998). Collective dynamics of „small-world" networks. Nature, 40, 393.

WWC. (2009). Istambul Water Consensus for Local and REGIONAL Authorities. In Proc. 5th World Water Forum. Istanbul. Istanbul: World Water Council.

Anexos

Enlace:

https://github.com/Jhuramos2023/Sectorizacion_RDAP/blob/main/Simulacion%20red%20-%202.R

```r
#====================== Data Pre-procesamiento ========================#
#======================================================================#
# En esta seccion, recuperara un archivo de datos de red en formato csv  #
# al estudio R como un marco de datos y convertirlo en el igraph         #
# objeto.                                                                #
#======================================================================#

##########================== SECTOR 3 ====================##########

#1. Leer los archivos desde la ubicacion de una red
nodesdat<- read.table("C:\\Users\\User\\Desktop\\TESIS JAIME\\Modelamiento RStudio\\nodesdat2_sector3.csv", header=TRUE
pipesdat<- read.table("C:\\Users\\User\\Desktop\\TESIS JAIME\\Modelamiento RStudio\\pipesdat2_sector3.csv", header=TRUE

pipesdat<-as.data.frame(pipesdat)
DJ<-pipesdat[,c(2,3,4)]
nodesdat<-as.data.frame(nodesdat)
DJ_meta=nodesdat
DJ_meta$Cota_1=DJ_meta[,4]

#2. Administrar conjunto de datos (EN CASO DE DARLE PESO CON LA FRECUENCIA DE DATOS)
B<-as.data.frame(table(DJ))
B1<-subset(B,Freq>0)

#3.  Crear un objeto igraph a partir de los marcos de datos.
library(igraph)
# A menos que su lista de bordes en B1 se reconozca como 'factor', producira un error
Stucont_J<-graph_from_data_frame(DJ, directed = TRUE, vertices = DJ_meta)
E(Stucont_J)$weight<-E(Stucont_J)$Length  #Aqui se reconoce a la longitud como peso
Stucont_J #grafo construido

#======================================================================#
#================== Explore sus datos igraph ==================#
#======================================================================#
#======================================================================#

#1. igraph summary
Stucont_J
gsize(Stucont_J)    #numero de enlaces
gorder(Stucont_J)   #grado

#2. Nodelist
V(Stucont_J) #vertices

#3. Edgelist
E(Stucont_J) #enlaces

#4. Attributes  (revisar los atributos)
V(Stucont_J)$Cota_1
```

```
# # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #

as.numeric(DJ_meta$Cota_1) #por siacaso lo vuelves factor numerico

################ Para la particion por intervalos ################
x <- DJ_meta$Cota_1
categorias <- cut(x, breaks = c(-1, 60, 70, 80, Inf),
                  labels = c("SEC1", "SEC2", "SEC3", "SEC4"))
particion <- data.frame(x, categorias)

DJ_meta_cota=group_by(particion,
                      categorias) %>%
  summarise(n=n())

#Reemplazar la columna cota_1 por las categorias
DJ_meta$Cota_1=particion$categorias

########## Volver a la creacion de grafo ##########

Stucont_J<-graph_from_data_frame(DJ, directed = TRUE, vertices = DJ_meta)
E(Stucont_J)$weight<-E(Stucont_J)$Length
Stucont_J

#5. Adjacency matrix (Matriz de adyacencia)
Stucont_J[c(1:10),c(1:10)]

a=Stucont_J[c(1:171),c(1:171)]
jaimito=as.data.frame.matrix(a)

setwd("C:\\Users\\User\\Desktop")

install.packages("rio")
library(rio)
export(jaimito,"jaimito.xlsx")


#=================================================================#
#================== Midiendo la Centralidad ====================#
#=================================================================#
# En esta seccion, mediras la centralidad del igrafo             #
# objeto, "Stucont". Podras ver como funciona la teorica         #
# El concepto de cada centralidad, como el grado, el vector propio y  #
# la centralidad de intermediacion, se mide mediante el igraph.  #
#=================================================================   #

#1. Grado de centralidad
Stucont_deg_J<-degree(Stucont_J,mode=c("All"))
V(Stucont_J)$degree<-Stucont_deg_J
```

```
#=================== Midiendo la Centralidad ======================#
#=================================================================#
# En esta seccion, mediras la centralidad del igrafo              #
# objeto, "Stucont". Podras ver como funciona la teorica          #
# El concepto de cada centralidad, como el grado, el vector propio y #
# la centralidad de intermediacion, se mide mediante el igraph.   #
#=================================================================#

#1. Grado de centralidad
Stucont_deg_J<-degree(Stucont_J,mode=c("All"))
V(Stucont_J)$degree<-Stucont_deg_J
V(Stucont_J)$degree
which.max(Stucont_deg_J)


jaimito=as.data.frame.vector(V(Stucont_J)$degree)

setwd("C:\\Users\\User\\Desktop")
library(rio)
export(jaimito,"jaimito.xlsx")

#2. Centralidad del vector propio
Stucont_eig_J <- evcent(Stucont_J)$vector
V(Stucont_J)$Eigen<-Stucont_eig_J
V(Stucont_J)$Eigen
which.max(Stucont_eig_J)




#3. Centralidad de intermediaci?n
Stucont_bw_J<-betweenness(Stucont_J, directed = FALSE)
V(Stucont_J)$betweenness<-Stucont_bw_J
V(Stucont_J)$betweenness
which.max(Stucont_bw_J)

DF_J<-as_long_data_frame(Stucont_J)
Stucont_J

#=================================================================#
#============== Medicion de la estructura de la red ==============#
#=================================================================#
# En esta seccion mediras los indicadores de la red               #
# estructura como la densidad de la red, la variedad.             #
#=================================================================#

#1. Densidad de red
edge_density(Stucont_J) # Densidad global de toda la red
```

```
#=============================================================#
#============== Medicion de la estructura de la red ==========#
#=============================================================#
# En esta seccion mediras los indicadores de la red           #
# estructura como la densidad de la red, la variedad.         #
#=============================================================#

#1. Densidad de red
edge_density(Stucont_J) # Densidad global de toda la red
SEC1<-induced_subgraph(Stucont_J, V(Stucont_J)[Cota_1=="SEC1"], impl=c("auto"))
edge_density(SEC1)

#2. Assortativity ----> indica que tanto se unen a vertices con similares caracteristicas
values_J <- as.numeric(factor(V(Stucont_J)$Cota_1))
assortativity_nominal(Stucont_J, types=values_J)

#2.1. Calculate the observed assortativity (Calcular la variedad observada)
observed.assortativity_J <- assortativity_nominal(Stucont_J, types=values_J)
results_J <- vector('list', 1000)
for(i in 1:1000){results_J[[i]] <- assortativity_nominal(Stucont_J, sample(values_J))}  "sample para cre

#2.2.  Trace la distribucion de los valores de surtido y agregue una linea vertical roja en el valor ori
hist(unlist(results_J), xlim = c(0,1))
abline(v = observed.assortativity_J,col = "red", lty = 3, lwd=2)

#=============================================================#
#==================== Network Visualization ==================#
#=============================================================#

#1. Trazar una red con el grado de centralidad

set.seed(1001)
library(RColorBrewer) # Esta es la biblioteca de colores.
pal<-brewer.pal(length(unique(V(Stucont_J)$Cota_1)), "Set3") # Color de v?rtice asignado por cada numero
plot(Stucont_J,edge.color = 'black',vertex.label.cex =0.5,
     vertex.color=pal[as.numeric(as.factor(vertex_attr(Stucont_J, "Cota_1")))],
     vertex.size = sqrt(Stucont_deg_J)/3, edge.width=sqrt(E(Stucont_J)$weight/800),
     edge.arrow.width=0.5,
     edge.arrow.size=0.08,
     edge.curved = T,
     layout = layout.fruchterman.reingold)

#1. Trazar una red con la centralidad del vector propio

set.seed(1001)
plot(Stucont_J,edge.color = 'black',vertex.label.cex =0.5,
     vertex.color=pal[as.numeric(as.factor(vertex_attr(Stucont_J, "Cota_1")))],
     vertex.size = sqrt(Stucont_eig_J)*10, edge.width=sqrt(E(Stucont_J)$weight/800),
```

```
set.seed(1001)
plot(Stucont_J,edge.color = 'black',vertex.label.cex =0.5,
     vertex.color=pal[as.numeric(as.factor(vertex_attr(Stucont_J, "Cota_1")))],
     vertex.size = sqrt(Stucont_bw_J)/3, edge.width=sqrt(E(Stucont_J)$weight/800),
     edge.arrow.width=0.5,
     edge.arrow.size=0.08,
     edge.curved = T,
     layout = layout.fruchterman.reingold)

# Camino m?s corto entre 1 y 5
sp <- shortest.paths(Stucont_J, v = "REP-01-5", to = "J-167-5")
sp[]                                        # Distancia
gsp <- get.shortest.paths(Stucont_J, from = "REP-01-5", to = "J-167-5")
V(Stucont_J)[gsp$vpath[[1]]]                          # Secuencia de vertices

#########################################################################
####################     Clustering Jerarquico     ######################
#########################################################################
install.packages("factoextra")
library(cluster)
library(factoextra)

nodesdat<- read.table("C:\\Users\\User\\Desktop\\TESIS JAIME\\Modelamiento RStudio\\nodesdat2_se
DJ_meta=nodesdat
DJ_meta

df <- DJ_meta
df
#normalizar las puntuaciones
df <- scale(df)
head(df)

#calcular la matriz de distacias
m.distancia <- get_dist(df, method = "euclidean")
fviz_dist(m.distancia, gradient = list(low = "blue", mid = "white", high = "red"))

#estimar el n?mero de cl?sters
#Elbow, silhouette o gap_stat  method
fviz_nbclust(df, pam, method = "wss")
fviz_nbclust(df, pam, method = "silhouette")
fviz_nbclust(df, kmeans, method = "gap_stat")

#Aplicando la funcion resnumclust del Rstudio

resnumclust<-NbClust(df, distance = "euclidean", min.nc=2, max.nc=10, method = "kmeans", index =
fviz_nbclust(resnumclust)

#calculamos los dos clusters
```

```r
#calculamos los dos clusters
k2 <- kmeans(df, centers = 5, nstart = 25)
k2
str(k2)

#plotear los cluster
fviz_cluster(k2, data = df)
fviz_cluster(k2, data = df, ellipse.type = "euclid",repel = TRUE,star.plot = TRUE) #ellipse.type
fviz_cluster(k2, data = df, ellipse.type = "norm")
fviz_cluster(k2, data = df, ellipse.type = "norm",palette = "Set2", ggtheme = theme_minimal())

res2 <- hcut(df, k = 5, stand = TRUE)
fviz_dend(res2, rect = TRUE, cex = 0.5,
          k_colors = c("red","#2E9FDF","green","black","blue","coral","chocolate"))

#pasar los cluster a mi df inicial para trabajar con ellos

DJ_meta %>%
  mutate(Cluster = k2$cluster) %>%
  group_by(Cluster) %>%
  summarise_all("mean")

df <- DJ_meta
df
df$clus<-as.factor(k2$cluster)
df
```

```
279
280 ##########================== SECTOR 2 ======================##########
281
282 #1. Leer los archivos desde la ubicacion de una red
283 nodesdat<- read.table("C:\\Users\\User\\Desktop\\TESIS JAIME\\Modelamiento RStudio\\nodesdat2_sector2.csv",
284 pipesdat<- read.table("C:\\Users\\User\\Desktop\\TESIS JAIME\\Modelamiento RStudio\\pipesdat2_sector2.csv",
285
286 pipesdat<-as.data.frame(pipesdat)
287 DJ<-pipesdat[,c(2,3,4)]
288 nodesdat<-as.data.frame(nodesdat)
289 DJ_meta=nodesdat
290 DJ_meta$Cota_1=DJ_meta[,4]
291
292 #2. Administrar conjunto de datos (EN CASO DE DARLE PESO CON LA FRECUENCIA DE DATOS)
293 B<-as.data.frame(table(DJ))
294 B1<-subset(B,Freq>0)
295
296 #3.  Crear un objeto igraph a partir de los marcos de datos.
297 library(igraph)
298 # A menos que su lista de bordes en B1 se reconozca como 'factor', producira un error
299 Stucont_J<-graph_from_data_frame(DJ, directed = TRUE, vertices = DJ_meta)
300 E(Stucont_J)$weight<-E(Stucont_J)$Length
301 Stucont_J #grafo construido
302
303 #=============================================================#
304 #1. igraph summary
305 Stucont_J
306 gsize(Stucont_J)    #344
307 gorder(Stucont_J)   #278
308
309 #2. Nodelist
310 V(Stucont_J) #vertices
311
312 #3. Edgelist
313 E(Stucont_J) #enlaces
314
315 #4. Attributes  (revisar los atributos)
316 V(Stucont_J)$Cota_1    #para ver si valores atipicos se colan
317
318 # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
319 as.numeric(DJ_meta$Cota_1) #por siacaso lo vuelves factor numerico
320
321 ################ Para la particion por intervalos ################
322 x <- DJ_meta$Cota_1
323 categorias <- cut(x, breaks = c(-1, 60, 70, 80, Inf),
324                   labels = c("SEC1", "SEC2", "SEC3", "SEC4"))
325 particion <- data.frame(x, categorias)
```

```
326
327 DJ_meta_cota=group_by(particion,
328                        categorias) %>%
329   summarise(n=n())
330
331 #Reemplazar la columna cota_1 por las categorias
332 DJ_meta$Cota_1=particion$categorias
333
334 ########## Volver a la creacion de grafo
335 Stucont_J<-graph_from_data_frame(DJ, directed = TRUE, vertices = DJ_meta)
336 E(Stucont_J)$weight<-E(Stucont_J)$Length
337 Stucont_J
338
339 #5. Adjacency matrix (Matriz de adyacencia)
340 Stucont_J[c(1:10),c(1:10)]
341
342 #=================================================================#
343 #1. Grado de centralidad
344 Stucont_deg_J<-degree(Stucont_J,mode=c("All"))
345 V(Stucont_J)$degree<-Stucont_deg_J
346 V(Stucont_J)$degree
347 which.max(Stucont_deg_J)           #J-98-4  #96
348
349 #2. Centralidad del vector propio
350 Stucont_eig_J <- evcent(Stucont_J)$vector
351 V(Stucont_J)$Eigen<-Stucont_eig_J
352 V(Stucont_J)$Eigen
353 which.max(Stucont_eig_J)
354
355 #3. Centralidad de intermediacion
356 Stucont_bw_J<-betweenness(Stucont_J, directed = FALSE)
357 V(Stucont_J)$betweenness<-Stucont_bw_J
358 V(Stucont_J)$betweenness
359 which.max(Stucont_bw_J)
360
361 DF_J<-as_long_data_frame(Stucont_J)
362 Stucont_J
363
364 #=================================================================#
365 #1. Densidad de red
366 edge_density(Stucont_J) # Densidad global de toda la red
367 SEC1<-induced_subgraph(Stucont_J, V(Stucont_J)[Cota_1=="SEC1"], impl=c("auto"))
368 edge_density(SEC1) # Densidad de nivel de Cota
369
370 #2. Assortativity ----> indica que tanto se unen a vertices con similares caracteristicas
371 values_J <- as.numeric(factor(V(Stucont_J)$Cota_1))
372 assortativity_nominal(Stucont_J, types=values_J)
373
```

```
#2.1. Calculate the observed assortativity (Calcular la variedad observada)
observed.assortativity_J <- assortativity_nominal(Stucont_J, types=values_J)
results_J <- vector('list', 1000)
for(i in 1:1000){results_J[[i]] <- assortativity_nominal(Stucont_J, sample(values_J))}  "sample para crea

#2.2.  Trace la distribucion de los valores de surtido y agregue una linea vertical roja en el valor orig
hist(unlist(results_J), xlim = c(0,1))
abline(v = observed.assortativity_J,col = "red", lty = 3, lwd=2)

#=================================================================#

#1. Trazar una red con el grado de centralidad

set.seed(1001)
library(RColorBrewer) # Esta es la biblioteca de colores.
pal<-brewer.pal(length(unique(V(Stucont_J)$Cota_1)), "Set3") # Color de vertice asignado por cada n?mero
plot(Stucont_J,edge.color = 'black',vertex.label.cex =0.5,
     vertex.color=pal[as.numeric(as.factor(vertex_attr(Stucont_J, "Cota_1")))],
     vertex.size = sqrt(Stucont_deg_J)/3, edge.width=sqrt(E(Stucont_J)$weight/800),
     edge.arrow.width=0.5,
     edge.arrow.size=0.06,
     edge.curved = T,
     layout = layout.fruchterman.reingold)

#1. Trazar una red con la centralidad del vector propio

set.seed(1001)
plot(Stucont_J,edge.color = 'black',vertex.label.cex =0.5,
     vertex.color=pal[as.numeric(as.factor(vertex_attr(Stucont_J, "Cota_1")))],
     vertex.size = sqrt(Stucont_eig_J)*10, edge.width=sqrt(E(Stucont_J)$weight/800),
     edge.arrow.width=0.5,
     edge.arrow.size=0.06,
     edge.curved = T,
     layout = layout.fruchterman.reingold)

#2. Trazado de una red con la centralidad de intermediacion

set.seed(1001)
plot(Stucont_J,edge.color = 'black',vertex.label.cex =0.5,
     vertex.color=pal[as.numeric(as.factor(vertex_attr(Stucont_J, "Cota_1")))],
     vertex.size = sqrt(Stucont_bw_J)/3, edge.width=sqrt(E(Stucont_J)$weight/800),
     edge.arrow.width=0.5,
     edge.arrow.size=0.06,
     edge.curved = T,
     layout = layout.fruchterman.reingold)

# Camino m?s corto entre 1 y 5
sp <- shortest.paths(Stucont_J, v = "RAP-03-4", to = "J-271-4")
```

```
419
420 # Camino m?s corto entre 1 y 5
421 sp <- shortest.paths(Stucont_J, v = "RAP-03-4", to = "J-271-4")
422 sp[]                                    # Distancia
423 gsp <- get.shortest.paths(Stucont_J, from = "RAP-03-4", to = "J-271-4")
424 V(Stucont_J)[gsp$vpath[[1]]]                        # Secuencia de v?rtices
425
426 ######################################################################
427 ####################     Clustering Jerarquico     ####################
428 ######################################################################
429 nodesdat<- read.table("C:\\Users\\User\\Desktop\\TESIS JAIME\\Modelamiento RStudio\\nodesdat2_sector2.csv", heade
430 DJ_meta=nodesdat
431 DJ_meta
432
433 df <- DJ_meta
434 df
435 #normalizar las puntuaciones
436 df <- scale(df)
437 head(df)
438
439 #calcular la matriz de distacias
440 m.distancia <- get_dist(df, method = "euclidean") #el m?todo aceptado tambi?n puede ser: "maximum", "manhattan"
441 fviz_dist(m.distancia, gradient = list(low = "blue", mid = "white", high = "red"))
442
443 #estimar el n?mero de cl?sters
444 #Elbow, silhouette o gap_stat  method
445 fviz_nbclust(df, kmeans, method = "wss")
446 fviz_nbclust(df, kmeans, method = "silhouette")
447 fviz_nbclust(df, kmeans, method = "gap_stat")
448
449 #Aplicando la funcion resnumclust del Rstudio
450
451 resnumclust<-NbClust(df, distance = "euclidean", min.nc=2, max.nc=10, method = "kmeans", index = "alllong")
452 fviz_nbclust(resnumclust)
453
454 #calculamos los dos cl?sters
455 k2 <- kmeans(df, centers = 3, nstart = 25)
456 k2
457 str(k2)
458
459 #plotear los cluster
460 fviz_cluster(k2, data = df)
461 fviz_cluster(k2, data = df, ellipse.type = "euclid",repel = TRUE,star.plot = TRUE) #ellipse.type= "t", "norm", "e
462 fviz_cluster(k2, data = df, ellipse.type = "norm")
463 fviz_cluster(k2, data = df, ellipse.type = "norm",palette = "Set2", ggtheme = theme_minimal())
464
465 res2 <- hcut(df, k = 3, stand = TRUE)
466 fviz_dend(res2, rect = TRUE, cex = 0.5,
```

```
495
496
497 ##########================== SECTOR 1 ====================##########
498
499 #1. Leer los archivos desde la ubicacion de una red
500 nodesdat<- read.table("C:\\Users\\User\\Desktop\\TESIS JAIME\\Modelamiento RStudio\\nodesdat2_sector1.csv",
501 pipesdat<- read.table("C:\\Users\\User\\Desktop\\TESIS JAIME\\Modelamiento RStudio\\pipesdat2_sector1.csv",
502
503 pipesdat<-as.data.frame(pipesdat)
504 DJ<-pipesdat[,c(2,3,4)]
505 nodesdat<-as.data.frame(nodesdat)
506 DJ_meta=nodesdat
507 DJ_meta$Cota_1=DJ_meta[,4]
508
509 #2. Administrar conjunto de datos (EN CASO DE DARLE PESO CON LA FRECUENCIA DE DATOS)
510 B<-as.data.frame(table(DJ))
511 B1<-subset(B,Freq>0)
512
513 #3.  Crear un objeto igraph a partir de los marcos de datos.
514 # A menos que su lista de bordes en B1 se reconozca como 'factor', producira un error
515 Stucont_J<-graph_from_data_frame(DJ, directed = TRUE, vertices = DJ_meta)
516 E(Stucont_J)$weight<-E(Stucont_J)$Length
517 Stucont_J #grafo construido
518
519 #================================================================#
520 #1. igraph summary
521 Stucont_J
522 gsize(Stucont_J)    #244
523 gorder(Stucont_J)   #179
524
525 #2. Nodelist
526 V(Stucont_J) #vertices
527
528 #3. Edgelist
529 E(Stucont_J) #enlaces
530
531 #4. Attributes  (revisar los atributos)
532 V(Stucont_J)$Cota_1    #para ver si valores atipicos se colan
533
534 # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # # #
535 as.numeric(DJ_meta$Cota_1)
536
537 ################ Para la particion por intervalos ################
538 x <- DJ_meta$Cota_1
539 categorias <- cut(x, breaks = c(-1, 40, 60, 80, Inf),
540                   labels = c("SEC1", "SEC2", "SEC3", "SEC4"))
541 particion <- data.frame(x, categorias)
542
```

```
particion <- data.frame(x, categorias)

DJ_meta_cota=group_by(particion,
                      categorias) %>%
  summarise(n=n())

#Reemplazar la columna cota_1 por las categorias
DJ_meta$Cota_1=particion$categorias

########## Volver a la creacion de grafo
Stucont_J<-graph_from_data_frame(DJ, directed = TRUE, vertices = DJ_meta)
E(Stucont_J)$weight<-E(Stucont_J)$Length  #Aqui se reconoce a la longitud como peso
Stucont_J

#5. Adjacency matrix (Matriz de adyacencia)
Stucont_J[c(1:10),c(1:10)] #el numero solo es para ver una matrix de zxz

#=================================================================#
#1. Grado de centralidad
Stucont_deg_J<-degree(Stucont_J,mode=c("All"))
V(Stucont_J)$degree<-Stucont_deg_J
V(Stucont_J)$degree
which.max(Stucont_deg_J)                          #J-5-3  #144

#2. Centralidad del vector propio
Stucont_eig_J <- evcent(Stucont_J)$vector
V(Stucont_J)$Eigen<-Stucont_eig_J
V(Stucont_J)$Eigen
which.max(Stucont_eig_J)                      #J-34-3    #167

#3. Centralidad de intermediaci?n
Stucont_bw_J<-betweenness(Stucont_J, directed = FALSE)
V(Stucont_J)$betweenness<-Stucont_bw_J
V(Stucont_J)$betweenness
which.max(Stucont_bw_J)                      #J-58    #87

DF_J<-as_long_data_frame(Stucont_J)
Stucont_J

#=================================================================#
#1. Densidad de red
edge_density(Stucont_J) # 0.007658025
SEC1<-induced_subgraph(Stucont_J, V(Stucont_J)[Cota_1=="SEC1"], impl=c("auto"))
edge_density(SEC1) # 0.01306471

#2. Assortativity ----> indica que tanto se unen a vertices con similares caracteristicas
values_J <- as.numeric(factor(V(Stucont_J)$Cota_1))
assortativity_nominal(Stucont_J, types=values_J)   #0.890189
```

```
assortativity_nominal(Stucont_J, types=values_J)   #0.890189

#2.1. Calculate the observed assortativity (Calcular la variedad observada)
observed.assortativity_J <- assortativity_nominal(Stucont_J, types=values_J)
results_J <- vector('list', 1000)
for(i in 1:1000){results_J[[i]] <- assortativity_nominal(Stucont_J, sample(values_J))}  "sample para

#2.2.  Trace la distribucion de los valores de surtido y agregue una linea vertical roja en el valor
hist(unlist(results_J), xlim = c(0,1))
abline(v = observed.assortativity_J,col = "red", lty = 3, lwd=2)

#=============================================================#

#1. Trazar una red con el grado de centralidad

set.seed(1001)
library(RColorBrewer) # Esta es la biblioteca de colores.
pal<-brewer.pal(length(unique(V(Stucont_J)$Cota_1)), "Set3") # Color de v?rtice asignado por cada n?m
plot(Stucont_J,edge.color = 'black',vertex.label.cex =0.5,
     vertex.color=pal[as.numeric(as.factor(vertex_attr(Stucont_J, "Cota_1")))],
     vertex.size = sqrt(Stucont_deg_J)/3, edge.width=sqrt(E(Stucont_J)$weight/800),
     edge.arrow.width=0.5,
     edge.arrow.size=0.06,
     edge.curved = T,
     layout = layout.fruchterman.reingold)

#1. Trazar una red con la centralidad del vector propio

set.seed(1001)
plot(Stucont_J,edge.color = 'black',vertex.label.cex =0.5,
     vertex.color=pal[as.numeric(as.factor(vertex_attr(Stucont_J, "Cota_1")))],
     vertex.size = sqrt(Stucont_eig_J)*10, edge.width=sqrt(E(Stucont_J)$weight/800),
     edge.arrow.width=0.5,
     edge.arrow.size=0.06,
     edge.curved = T,
     layout = layout.fruchterman.reingold)

#2. Trazado de una red con la centralidad de intermediacion

set.seed(1001)
plot(Stucont_J,edge.color = 'black',vertex.label.cex =0.5,
     vertex.color=pal[as.numeric(as.factor(vertex_attr(Stucont_J, "Cota_1")))],
     vertex.size = sqrt(Stucont_bw_J)/3, edge.width=sqrt(E(Stucont_J)$weight/800),
     edge.arrow.width=0.5,
     edge.arrow.size=0.06,
     edge.curved = T,
     layout = layout.fruchterman.reingold)

```

```
     layout = layout.fruchterman.reingold)

# Camino m?s corto entre 1 y 5
sp1_1 <- shortest.paths(Stucont_J, v = "R-522", to = "J-57")
sp1_1[]                                          # Distancia
gsp1_1 <- get.shortest.paths(Stucont_J, from = "R-522", to = "J-57")
V(Stucont_J)[gsp1_1$vpath[[1]]]                          # Secuencia de vertices

sp1_2 <- shortest.paths(Stucont_J, v = "R-522", to = "J-81")
sp1_2[]                                          # Distancia
gsp1_2 <- get.shortest.paths(Stucont_J, from = "R-522", to = "J-81")
V(Stucont_J)[gsp1_2$vpath[[1]]]



sp2 <- shortest.paths(Stucont_J, v = "RAP-03-2", to = "J-8-2")
sp2[]                                        # Distancia
gsp2 <- get.shortest.paths(Stucont_J, from = "RAP-03-2", to = "J-8-2")
V(Stucont_J)[gsp2$vpath[[1]]]                          # Secuencia de vertices

sp3 <- shortest.paths(Stucont_J, v = "RAP-01-3", to = "J-28-3")
sp3[]                                        # Distancia
gsp3 <- get.shortest.paths(Stucont_J, from = "RAP-01-3", to = "J-28-3")
V(Stucont_J)[gsp3$vpath[[1]]]                          # Secuencia de vertices

#########################################################################
####################     Clustering Jerarquico     ####################
#########################################################################

DJ_meta=nodesdat
DJ_meta

df <- DJ_meta
df
#normalizar las puntuaciones
df <- scale(df)
head(df)

#calcular la matriz de distacias
m.distancia <- get_dist(df, method = "euclidean") #el metodo aceptado tambi?n puede ser:

#estimar el numero de clusters
#Elbow, silhouette o gap_stat  method
fviz_nbclust(df, kmeans, method = "wss")
fviz_nbclust(df, kmeans, method = "silhouette")
fviz_nbclust(df, kmeans, method = "gap_stat")

#Aplicando la funcion resnumclust del Rstudio:
```

```
680
681 #Aplicando la funcion resnumclust del Rstudio:
682
683 resnumclust<-NbClust(df, distance = "euclidean", min.nc=2, max.nc=10, method = "kmeans", index =
684 fviz_nbclust(resnumclust)
685
686 #calculamos los dos clusters
687 k2 <- kmeans(df, centers = 3, nstart = 25)
688 k2
689 str(k2)
690
691 #plotear los cluster
692 fviz_cluster(k2, data = df)
693 fviz_cluster(k2, data = df, ellipse.type = "euclid",repel = TRUE,star.plot = TRUE) #ellipse.type=
694 fviz_cluster(k2, data = df, ellipse.type = "norm")
695 fviz_cluster(k2, data = df, ellipse.type = "norm",palette = "Set2", ggtheme = theme_minimal())
696
697 res2 <- hcut(df, k = 3, stand = TRUE)
698 fviz_dend(res2, rect = TRUE, cex = 0.5,
699           k_colors = c("red","#2E9FDF","green","black","blue","coral","chocolate"))
700
701 res4 <- hcut(df, k = 4, stand = TRUE)
702 fviz_dend(res4, rect = TRUE, cex = 0.5,
703           k_colors = c("red","#2E9FDF","green","black"))
704
705 #Pasar los cluster a mi df inicial para trabajar con ellos
706
707 DJ_meta %>%
708   mutate(Cluster = k2$cluster) %>%
709   group_by(Cluster) %>%
710   summarise_all("mean")
711
712 df <- DJ_meta
713 df
714 df$clus<-as.factor(k2$cluster)
715 df
716
717 head(USArrests)
718
```