UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

DEPARTAMENTO DE INGENIERÍA HIDRÁULICA Y MEDIO AMBIENTE

Sectorización de Redes de Abastecimiento de Agua Potable basada en detección de comunidades en redes sociales y optimización heurística

TESIS DOCTORAL

Autor:
Enrique O. Campbell González

Directores:
Dr. Joaquín Izquierdo Sebastián
Dr. Idel Montalvo Arango

Valencia, julio de 2017

AGRADECIMIENTOS

Agradecimientos a todos los compañeros/amigos de FluIng, en especial a los Profesores Rafael Pérez García (Q.E.P.D.) y Joaquín Izquierdo, al Dr. Idel Montalvo y al M.Sc. Bruno Brentan (Universidad Estatal de Campinas, Brasil). A todo el personal de Investigación y Desarrollo de las Empresas Operadoras de Agua de Berlín, Alemania, en especial a la Dipl. Fereshte Sedehizade y a la Dipl. Regina Gnirss. Todo este trabajo se ha hecho posible gracias a su apoyo constante y ejemplo.

A mi madre, hermana y resto de familiares y amigos, de Nicaragua, Tegucigalpa, Valencia, Bruselas, Berlín y Karlsruhe, en especial a la familia Hebrero Campbell por la financiación, acogida y amistad incondicional y desinteresada.

A la Agencia de Cooperación Española para el Desarrollo (AECID) por la beca para la primera fase del periodo de investigación.

.

ACKNOWLEDGEMENT

Thanks to all of the colleagues / friends from FluIng UPV, especially to Professors Rafael Pérez García (R.I.P.) and Joaquín Izquierdo, Dr. Idel Montalvo and M.Sc. Bruno Brentan (State University of Campinas, Brazil). To all the Research and Development personnel of the Water Operators of Berlin, in special to Dipl. Fereshte Sedehizade and Dipl. Regina Gnirss. All this work has been possible thanks to their constant support and inspiration.

To my mother, sister and rest of the family and friends, from Nicaragua, Tegucigalpa, Valencia, Brussels, Berlin and Karlsruhe, especially to the Hebrero Campbell family for the funding, hosting and unconditional and disinterested friendship.

To the Spanish Cooperation Agency for development (AECID) for the funding of the first phase of the research period.

DANKSAGUNGEN

Vielen Dank an alle Kollegen / Freunde von Fluing UPV, vor allem an die Professoren Rafael Pérez García (R.I.F) und Joaquín Izquierdo, Dr. Idel Montalvo und Msc. Bruno Brentan (Bundesuniversität von Campinas, Brasilien). Dank an alle Forschungs- und Entwicklungsmitarbeiter (FE) der Berliner Wasserbetreibe, Deutschland, vor allem an Dipl. -Ing. Fereshte Sedehizade und Dipl.-Ing. Regina Gnirss. Die ganze Arbeit ist dank ihrer ständigen Unterstützung und Inspiration möglich.

Dank an meine Mutter, Schwester und den Rest meiner Familie und Freunde, aus Nicaragua, Tegucigalpa, Valencia, Brüssel und Berlin. Vor allem an die Familie Hebrero Campbell für die Finanzierung, Hosting und bedingungslose und uneigennützige Freundschaft.

Dank an die spanischen Kooperationsagentur für Entwicklung (AECID) für die Finanzierung der ersten Phase des Forschungszeitraums.

Dedicado a:

Profesor Rafael Pérez García († Octubre, 2016)

Ron Rivera († Septiembre, 2008)

Davi Campbell († Diciembre, 1998)

Enrique Campbell Hooker († Diciembre, 1982)

A mis tías Ruth & Salvadora en donde estén

RESUMEN

La sectorización de las Redes de Abastecimiento de Agua Potable (RDAPs) se puede considerar como una estrategia de gestión que implica su subdivisión en subgrupos homogéneos. Dicha subdivisión tiene como fin poder gestionar de mejor manera en cada sub-área (sector) aspectos tales como: fugas, reparaciones, aspectos de calidad, entre otros, mediante el monitoreo permanente de los caudales que ingresan a cada sector.

Debido a la popularidad que ha ganado la técnica en los últimos 15 años, se han venido proponiendo múltiples metodologías para el diseño de sectores. Sin embargo, la mayoría de estas metodologías se orientan a redes con muchas fuentes y no consideran en toda su extensión los beneficios económicos generados, producto de la sectorización resultante.

En esta tesis se plantea una serie de metodologías de sectorización innovadoras en que primero se definen los sectores basados en algoritmos de detección de comunidades en grafos de redes sociales. Los algoritmos de detección de comunidades en redes sociales están concebidos para analizar redes de gran extensión, por lo que permiten trabajar con RDAPs también de gran extensión. En un segundo paso, se optimiza el conjunto de entradas y válvulas de cierre (CEVC) de cada sector utilizando técnicas heurísticas de optimización. En dicha optimización se incluyen los beneficios de la sectorización en términos de reducción de fugas producto de la reducción de presión y de la capacidad aumentada para detectar nuevos eventos de fugas. Para el abordaje del segundo aspecto se hace uso de la técnica de Monte Carlo para representar eventos de fugas en cada sector basados en una distribución de probabilidades dada. En tal sentido, es importante destacar que este es el primer trabajo en el que se hace un abordaje de esta naturaleza.

Las estrategias empleadas para subdividir RDAPs deben tener en cuenta la

topología de las mismas. En redes con suficientes fuentes dentro de su mallado, se puede proceder a través de una subdivisión por fuentes, de tal manera que cada sector cuente con una fuente, o un par de fuentes exclusivas. Este tipo de aproximación no sería factible para RDAPs con pocas fuentes o con fuentes que se encuentren fuera del mallado de la red. En este último caso, la red es dependiente de una red de conducción principal o red troncal (de este punto en adelante los términos son intercambiables), y cualquier estrategia de sectorización que se plantee deberá evitar cierres en la misma, a fin de preservar la fiabilidad del sistema. El intentar establecer sectores por fuentes (alrededor de las fuentes) en una red dependiente de una red de conducción principal, conlleva el problema de contar con sectores excesivamente grandes, en los cuales es difícil hacer un monitoreo efectivo de los caudales de consumo. Es por esta razón que dentro de las metodologías que se plantean en este trabajo, se lleva a cabo un proceso de identificación y segregación de la red de conducción principal. Es lógico pensar que no existe una definición universal del alcance de dicha red y, por el contrario, que la misma sea distinta para cada RDAP. El método de identificación de la red troncal propuesto en este trabajo se basa en el concepto de *Caminos más Cortos*, propio de la teoría de grafos, en combinación con un análisis de los caudales (y direcciones de los mismos) que circulan por la red en el escenario de mayor demanda. Como resultado, se obtiene un ranking de tuberías, a partir del cual se puede seleccionar el alcance de la red de conducción principal.

Una vez identificada la red troncal, la misma se aísla de la red distribución y, sobre esta última, se definen los sectores utilizando tres algoritmos de detección de comunidades en redes sociales: Clústering Jerárquico, Algoritmo de Detección Multinivel o Método Louvain y Detección de Comunidades a través de Caminos Aleatorios. Tras definir el área que corresponde a cada sector, se debe establecer el conjunto de válvulas cerradas y el punto de abastecimiento del sector. Para tal fin, se implementan procedimientos de optimización basados en los algoritmos de optimización heurística: Algoritmos Genéticos (Genetic Algorithms), Optimización de Enjambres de Partículas (Particle Swarm Optimization) y Optimización de

Enjambres de Agentes (Agent Swarm Optimization).

En el primer procedimiento, no sólo se toma en cuenta el beneficio de la sectorización en términos de reducción de caudales asociados a fugas de fondo, como consecuencia de reducir la presión, sino que también se tienen en cuenta otros efectos de gran relevancia, tales como: aumento de la capacidad para poder detectar y reparar, con mayor rapidez, nuevos eventos de fugas; reducción de la frecuencia de aparición de nuevas roturas; y reducción del caudal de consumo doméstico, tanto interno como externo. Esto permite que el análisis coste/beneficio de la sectorización sea más realista que el que se podría realizar si sólo se tuviera en cuenta la reducción de caudales de fugas de fondo. Tal y como se mencionó previamente, para poder predecir el efecto que tiene un CEVC sobre la detección de nuevos eventos de fugas, se hace uso de un modelo de simulación Monte Carlo, en el cual se toman como criterios tanto el tamaño de los sectores, como el número de entradas/válvulas cerradas de cada uno de ellos. Así, por ejemplo, en sectores más grandes, con mayor número de entradas, se puede preservar mejor la fiabilidad del servicio, aunque es más difícil poder detectar y reparar un nuevo evento (de fuga). Y, por el contrario, cuanto más pequeño sea un sector y menos entradas tenga, más fácil sería detectar un nuevo evento de fugas.

En el segundo método se emplea optimización multinivel para, además de optimizar el conjunto de válvulas cerradas/entrada de sectores, determinar el punto de ajuste de válvulas reductoras de presión en la entrada de los sectores.

En el tercer método de optimización sólo se optimiza el CEVC mediante un análisis económico que no tiene en cuenta el efecto sobre la aparición de nuevas fugas.

Para la aplicación de las metodologías propuestas es importante contar con un modelo hidráulico correctamente calibrado. Para ello, se desarrolló un método de calibración de RDAPs que no sólo tiene en cuenta la rugosidad de las tuberías, sino también los coeficientes de emisor en los nodos. En el método se implementan

algoritmos genéticos como técnica de optimización, estableciendo como variables de decisión: las rugosidades de las tuberías, los coeficientes de emisor en los nodos y los factores horarios de la curva de demanda. Previo a la optimización, los nodos y las tuberías se subdividen en clases mediante el uso de Mapas Auto-Organizados (Self-Organized Maps) y la técnica clústering Jerárquico. Las variables de decisión también se generan en función de las clases establecidas. Este método tiene la ventaja de mejorar la distribución de las fugas en una RDAP, respecto a lo que lo hacen otros métodos convencionales en los que se distribuyen los emisores únicamente en función de la media de la longitud ponderada de tubería que le corresponde a cada nodo.

Para fines de ejemplificación, las metodologías propuestas se implementan sobre una sección de la RDAP de la ciudad de Managua, Nicaragua. Esta red tiene 246 km de tubería, 3 fuentes de agua, y cuenta con 4126 nudos y 4231 tuberías. Como resultado de la implementación se reporta un beneficio neto de 104,764 $ (dólares americanos)/año.

ABSTRACT

The partition of Water Supply Networks (WSNs) into sectors can be considered as a management strategy that entails its subdivision into homogeneous subgroups. This subdivision aims to enhance such management aspects as leakage, repairs, and quality aspects, among others, in each sub-area (sector) carried out by permanently monitoring the inlet flows of each sector. Given the popularity that the technique has gained over the last 15 years, multiple automatic sectorization approaches have been put in place. However, most of them can be implemented only on networks with many sources and fail making a complete approach of the economic benefits generated by its implementation.

This thesis presents a series of innovative sectorization methodologies where the sectors are previously defined by means of social networks community detection algorithms. Such algorithms are meant to analyze large networks, which allow them to address the problem of sectorization in large RDAPs. In a second step, the set of boundary valves/sector entrance is optimized based upon optimization heuristic techniques. Such techniques include the benefits of sectorization in terms of both, leakage reduction, as a result of reducing pressure, and increasing the capacity to detect new leakage events. To tackle the later, the Monte Carlo technique is used to simulate the occurrence of new leakage events. In this sense, it is worth mentioning that this is the first work in which an approach of this nature is conducted.

WSNs subdivision strategies, must take into account their network topology. In networks with sufficient water sources (tanks, pumps) within their layout, a subdivision in which each sector relies on at least one individual and independent source could be feasible. Such approximation would not be suitable, however, for WSNs with only few sources or with sources located out of the layout area. In this case, it is said that the network is dependent on a main conduction network, also called trunk network (from this point onwards, both terms will be interchangeable), and to preserve the reliability of the system, any sectorization strategy should avoid

closure of its pipes. Attempting to establish sectors around sources in a trunk-network-dependent WSN would lead to extremely large sectors, where it is difficult to effectively monitor the inlet flows. Within the methodology presented in this work, a process of identification and segregation of the trunk network is carried out. As can be imagined, there is no universal definition of the scope of the trunk network and, on the contrary, it is expected to be different for each WSN. The herein proposed trunk network identification method, is based on the concept of Shortest Path from the graph theory, in combination with an analysis of the flows (and their directions) circulating through the network in the pick-demand scenario. As a result, the pipes are graded, and the range of pipes belonging to the trunk network can be selected.

Once the trunk network is identified, it is isolated from the distribution network and sectors are defined on the later, based on three social network based community detection algorithms, namely: Hierarchical Clustering, Multilevel Detection Algorithm or Louvaine Method and Random Walk community detection. After defining the area corresponding to each sector, the set of entrance / boundary valves must be established. To this end, heuristic-based optimization algorithms (Genetic Algorithms, Particle Swarm Optimization and Agent Swarm Optimization) are implemented.

The first procedure not only takes into account the benefit of sectorization in terms of reduction of flows associated with background leakage as a result of reducing pressure, but also considers other effects of great relevance, namely: increased capacity to faster detect and repair new leakage events; reduction of the frequency of appearance of new breaks; and reduction of domestic (internal and external) consumption flow. This leads to a more realistic cost-benefit analysis than the one that could be carried out if only the reduction of background leakage flows was considered. As previously mentioned, to predict the effect of a given arrangement of sector entrance / boundary valves on the detection of new leakage events, a Monte Carlo simulation model is implemented. Both, sector size and number of

entrance/boundary valves are considered as criteria. For example, in larger sectors with a larger number of entries, network resilience can be better preserved, but it is more difficult to detect and repair a new (leakage); on the contrary, in smaller sectors with a reduced number of entrance points, it would be easier to detect a new leakage event.

In the second method, multilevel optimization is implemented to optimize the set of boundary valves / sector entrance, in the first level, and to determine the set point of pressure reducing valves located at the entrance of each sector, in the second level.

In the third optimization method, only the boundary valves/sector entrance set is optimized based on an economic analysis that does not take into account the effect on the occurrence of new leakages.

For the application of the proposed methodologies, it is mandatory to count on an appropriately calibrated hydraulic model. Thus, a WSN calibration method which not only takes into account the roughness of the pipes but also the emitter coefficients at the nodes was developed. The optimization of the method is carried out by means of genetic algorithms, establishing as decision variables: the roughness of the pipes, the emitter coefficients in the nodes and the time factors of the demand pattern. Prior to optimization, nodes and pipelines are subdivided into classes by means of hierarchical clustering using Self-Organized Maps. The decision variables are also generated according to the established classes. This method has the advantage of improving the distribution of leakages in WSNs, compared to other conventional methods, in which emitters are distributed only according to the length portion of pipe that corresponds to each node.

For exemplification purposes, the proposed methodologies are implemented on a section of the WSN of Managua city, capital of Nicaragua. This network has 246 km of pipeline, 3 water sources, 4126 nodes and 4231 pipes. As a result of the implementation, a net profit of 104,764 $ (American dollars)/year is reported.

RESUM

La sectorització de les Xarxes d'Abastament d'Aigua Potable (XAAPs) es pot considerar com una estratègia de gestió que implica la seva subdivisió en subgrups homogenis. Aquesta subdivisió té com a finalitat poder gestionar de millor manera en cada subàrea (sector) aspectes com ara: fuites, reparacions, aspectes de qualitat, entre d'altres, mitjançant el monitoratge permanent dels cabals que ingressen a cada sector.

A causa de la popularitat que ha guanyat la tècnica en els últims 15 anys, s'han vingut proposant múltiples metodologies per al disseny de sectors. No obstant això, la majoria d'aquestes metodologies s'orienten a xarxes amb moltes fonts i no consideren en tota la seva extensió els beneficis econòmics generats producte de la sectorització resultant.

En aquesta tesi es planteja una sèrie de metodologies de sectorització innovadores en que primer es defineixen els sectors basats en algoritmes de detecció de comunitats en grafs de xarxes socials. Els algoritmes de detecció de comunitats en xarxes socials estan concebuts per analitzar xarxes de gran extensió, pel que permeten treballar amb XAAPs també de gran extensió. En un segon pas, s'optimitza el conjunt d'entrades i vàlvules de tancament (CEVT) de cada sector utilitzant tècniques heurístiques d'optimització. En aquesta optimització s'inclouen els beneficis de la sectorització en termes de reducció de fuites producte de la reducció de pressió i de la capacitat augmentada per detectar nous esdeveniments de fuites. Per l'abordatge del segon aspecte es fa ús de la tècnica de Monte Carlo per representar esdeveniments de fuites en cada sector basats en una distribució de probabilitats donada. En aquest sentit, és important destacar que aquest és el primer treball en què es fa un abordatge d'aquesta naturalesa.

Les estratègies emprades per subdividir XAAPs han de tenir en compte la topologia de les mateixes. En xarxes amb suficients fonts dins del seu mallat, es pot procedir a

través d'una subdivisió per fonts, de manera que cada sector compti amb una font, o un parell de fonts exclusives. Aquest tipus d'aproximació no seria factible per XAAPs amb poques fonts o amb fonts que es trobin fora del mallat de la xarxa. En aquest últim cas, la xarxa és depenent d'una xarxa de conducció principal o xarxa troncal (d'aquest punt en endavant els termes són intercanviables), i qualsevol estratègia de sectorització que es plantegi d'evitar tancaments en la mateixa, a fi de preservar la fiabilitat del sistema. El intentar establir sectors per fonts (al voltant de les fonts) en una xarxa depenent d'una xarxa de conducció principal, comporta el problema de comptar amb sectors excessivament grans, en els quals és difícil fer un monitoratge efectiu dels cabals de consum. És per aquesta raó que dins de les metodologies que es plantegen en aquest treball, es porta a terme un procés d'identificació i segregació de la xarxa de conducció principal. És lògic pensar que no hi ha una definició universal de l'abast d'aquesta xarxa i, per contra, que la mateixa sigui diferent per a cada XAAP. El mètode d'identificació de la xarxa troncal proposat en aquest treball es basa en el concepte de *camins més curts*, propi de la teoria de grafs, en combinació amb una anàlisi dels cabals (i direccions dels mateixos) que circulen per la xarxa en l'escenari de major demanda. Com a resultat, s'obté un rànquing de canonades, a partir del qual es pot seleccionar l'abast de la xarxa de conducció principal.

Una vegada identificada la xarxa troncal, la mateixa s'aïlla de la xarxa de distribució i, a aquesta última, es defineixen els sectors utilitzant tres algoritmes de detecció de comunitats en xarxes socials: Clustering jeràrquic, Algorisme de Detecció Multinivell o Mètode Louvain i Detecció de Comunitats a través de Camins Aleatoris. Després de definir l'àrea que correspon a cada sector, s'ha d'establir el conjunt de vàlvules tancades i el punt d'abastament del sector. Per a tal fi, s'implementen procediments d'optimització basats en els algoritmes d'optimització heurística: Algorismes Genètics (Genetic Algorithms), Optimització de Eixams de Partícules (Particle Swarm Optimization) i Optimització de Eixams d'Agents (Agent Swarm Optimization).

En el primer procediment, no només es té en compte el benefici de la sectorització en termes de reducció de cabals associats a fuites de fons, com a conseqüència de reduir la pressió, sinó que també es tenen en compte altres efectes de gran rellevància, com ara : augment de la capacitat per poder detectar i reparar, amb més rapidesa, nous esdeveniments de fuites; reducció de la freqüència d'aparició de noves ruptures; i reducció del cabal de consum domèstic, tant intern com extern. Això permet que l'anàlisi cost / benefici de la sectorització sigui més realista que el que es podria fer si només es tingués en compte la reducció de cabals de fuites de fons. Tal com es va esmentar prèviament, per poder predir l'efecte que té un CEVT sobre la detecció de nous esdeveniments de fuites, es fa ús d'un model de simulació Monte Carlo, en el qual es prenen com a criteris tant la mida dels sectors, com el nombre d'entrades / vàlvules tancades de cada un d'ells. Així, per exemple, en sectors més grans, amb major nombre d'entrades, es pot preservar millor la fiabilitat del servei, encara que és més difícil poder detectar i reparar un nou esdeveniment (de fuita). I, per contra, com més petit sigui un sector i menys entrades tingui, més fàcil seria detectar un nou esdeveniment de fuites.

En el segon mètode s'empra optimització multinivell per, a més d'optimitzar el conjunt de vàlvules tancades / entrada de sectors, determinar el punt d'ajust de vàlvules reductores de pressió a l'entrada dels sectors.

En el tercer mètode d'optimització només s'optimitza el CEVT mitjançant una anàlisi econòmica que no té en compte l'efecte sobre l'aparició de noves fuites.

Per a l'aplicació de les metodologies proposades és important comptar amb un model hidràulic correctament calibrat. Per a això, es va desenvolupar un mètode de calibratge de XAAPs que no només té en compte la rugositat de les canonades, sinó també els coeficients d'emissor en els nodes. En el mètode s'implementen algoritmes genètics com a tècnica d'optimització, establint com a variables de decisió: les rugositats de les canonades, els coeficients d'emissor en els nodes i els factors horaris de la corba de demanda. Previ a l'optimització, els nodes i les

canonades es subdivideixen en classes mitjançant l'ús de Mapes Auto-Organitzats (Self-Organized Maps) i la tècnica Clustering Jeràrquic. Les variables de decisió també es generen en funció de les classes establertes. Aquest mètode té l'avantatge de millorar la distribució de les fuites en una XAAP, respecte al que ho fan altres mètodes convencionals en els quals es distribueixen els emissors únicament en funció de la mitjana de la longitud ponderada de canonada que li correspon a cada node .

Per a fins d'exemplificació, les metodologies proposades s'implementen sobre una secció de la XAAP de la ciutat de Managua, Nicaragua. Aquesta xarxa té 246 km de canonada, 3 fonts d'aigua, i compta amb 4126 nodes i 4231 canonades. Com a resultat de la implementació es reporta un benefici net de 104,764 $ (dòlars americans)/any.

TABLA DE CONTENIDO

INDICE DE ILUSTRACIONES

INDICE DE ILUSTRACIONES APENDICE I

INDICE DE TABLAS

INDICE DE TABLAS APENDICE I

ACRÓNIMOS

ACO: Ant Colony Optimization (Optimización de Colonia de Hormigas)
AG: Algoritmos Genéticos (Genetic Algorithms)
AD: Average Distance (Distancia Promedio)
ADM: Average Distance Between Means (Distancia Promedio entre Medias)
APN: Average Portion of non-Overlap (Porción Promedio de no Traslape)
AS: Ancho de Silueta
ANR: Agua no Registrada
ASO: Agent Swarm Optimization (Optimización de Enjambre de Agentes)
AU: Approximately Unbiased (Sesgo Aproximado)
BABE: Burst and Background Estimates Methodology (Metodología de Estimación de Roturas y Fugas de Fondo)
BFS: Breadth First Search (Búsqueda en Anchura)
BMU: Best Matching Unit (Mejor unidad de Coincidencia)
BP: Bootstrap Probability (Probabilidad de Remuestreo)
CMC: Camino más Corto
CCCF: Cophenetic Correlation Coefficient (Coeficiente de Correlación Cofenética)
CPP: Coeficiente de Pérdida de Potencia
CEVC: Conjunto de entradas y válvulas de cierre
DFS: Depth First Search (Búsqueda en Profundidad)
DEC: Desviación Estándar de Características
EPA: Environmental Protection Agency (Agencia Protección Ambiental)
FAVAD: Fixed and Variable Areas Discharge (Áreas de Descargas Fijas y Variables)
FOM: Figure of Merits (Figura de Mérito)
FOI: Frecuencia Óptima de Inspección
FDN: Factor Día-Noche
IWA: International Water Association (Asociación Internacional del Agua)
INAF: Índice Natural de Aumento de Fugas
MAS: Multi-Agent Systems (Sistemas Multiagente)
MC: Monte Carlo
MMC: Método de Muestreo Monte Carlo
MHL: Método Hipercubo Latino
HL: Hipercubo Latino
PNUMA: Programa de Naciones Unidas para el Medio Ambiente
PSO: Particle Swarm Optimization (Optimización de Enjambre de Partículas)
RDAP: Red de Abastecimiento de Agua Potable
SANAA: Sistema Nacional de Acueductos y Alcantarillados de Honduras
SMC: Simulación Monte Carlo
SOM: Self-Organized Maps (Mapas Auto-Organizados)
VCMC: Valor de Camino más Corto
VCMCA: Valor de Camino más Corto Acumulado
VRP: Válvulas Reguladoras de Presión
UOC: Unidad Operativa de Control
UMF: Umbral Mínimo de Fugas

NEFCP: Nivel Económico de Fugas a Corto Plazo
NEFNR: Nivel Económico de Fugas no Reportado
UKWIR: UK Water Industry Association (Asociación Inglesa de la Industria del Agua)
WLTF: Water Loss Task Force (Grupo de trabajo sobre Pérdidas de Agua)

I. INTRODUCCIÓN

I.1 Gestión del Abastecimiento de Agua Potable y Sectorización: Generalidades, Ventajas y Desventajas

El agua potable es un recurso indispensable para todo proceso relacionado con la vida; es un producto primario tanto para las actividades domésticas, como para las actividades urbanas y agrícolas. La disponibilidad de este recurso está totalmente ligada al bienestar y prosperidad de cualquier sociedad. De ahí la importancia que cobra la buena gestión de las Redes de Abastecimiento de Agua Potable (RDAPs) que, conceptualmente, se pueden definir como las infraestructuras que permiten transportar el recurso en cuestión desde las fuentes hasta los consumidores.

Desde un punto de vista técnico, y dando por supuesto que se hace una apropiada gestión administrativa, los problemas de las RDAPs pueden resumirse en cuatro aspectos generales: fugas y agua no contabilizada; integridad física de la red; calidad del agua a distribuir; y fiabilidad y calidad de la base de datos de los sistemas de distribución de agua. Con relación al primero de ellos, el control de las pérdidas se ha planteado como una preocupación desde la construcción de las primeras RDAPs. Incluso, en la antigua Roma, ya se implementaban acciones para reducir el volumen de pérdidas en los acueductos (Pilcher *et al*., 2007).

En países en vías de desarrollo, las fugas pueden llegar a representar hasta el 50% del agua inyectada a la red (Kingdom *et al*., 2006; ADB, 2007). Es decir, en una RDAP con tal índice de pérdida, se tienen que producir 2 m^3 de agua para que llegue 1 m^3 a los usuarios. Se ha estimado que, en estos países, el volumen anual de pérdidas de agua alcanza 26.7 miles de millones de m^3, lo que representa 5.9 miles de millones de dólares norteamericanos. Esto resulta peor si se tiene en cuenta que a la hora de valorar el coste del agua perdida, sólo consideran sus costos marginales, es decir, los costos asociados al proceso operativo que implica llevar el agua desde las fuentes hasta los usuarios; no obstante, existen otros costos que no son tan

fáciles de cuantificar, que se agrupan bajo una categoría denominada externalidades, y que representan los costos no asociados a la producción, que no están reflejados en el precio de mercado (Delgado-Galván *et al*., 2010; Delgado-Galván, 2011). Un ejemplo de externalidad es el costo por reparación de daños que puede causar una fuga en una tubería de gran diámetro sometida a 40 mca (metros de columna de agua) a los edificios residenciales o a una calzada. Con tan sólo reducir el valor anteriormente citado a la mitad, se podría abastecer hasta 90 millones de personas (WWC, 2009). En esta misma línea, IWA (2000) estima que reducir las pérdidas en países de renta baja y media a la mitad del nivel actual, representaría 11 mil millones de m^3/año lo que permitiría el acceso a agua potable a 130 millones de personas, implicando, además, un ahorro de 4 mil millones de dólares norteamericanos para los operadores de agua. Cifras de esta naturaleza resaltan la necesidad de equidad y gestión óptima y sostenible del agua con el fin de hacer frente a la creciente demanda del recurso a nivel mundial.

Pese a que en países desarrollados la situación es distinta, ya que el porcentaje de pérdida no suele superar el 15% (Kingdom *et al*., 2006), las previsiones, a nivel global son desalentadoras. Por ejemplo, un estudio conducido por el Programa de Naciones Unidas para el Medio Ambiente (PNUMA) estima que, para el año 2025, dos tercios de la población mundial serán objeto de "*estrés*" de agua ya sea moderado o alto. Este mismo estudio, estima que en EEUU las extracciones de agua pasarán del 10-20% (cifra de 1995) del agua disponible, al 20-40% en los próximos 15 años (Thornton *et al*., 2008). En países en vías de desarrollo es muy común que las empresas encargadas del suministro de agua potable busquen, muchas veces con carácter de urgencia, fondos para financiar la expansión del suministro de agua, ya que se estima que la mitad de los consumidores sufren un servicio intermitente y/o de baja calidad (Kingdom *et al*., 2006). Sin embargo, el llevar a cabo tareas de ampliación de RDAPs sin mejorar su eficiencia, crea un círculo vicioso de sobre-inversión que impide darle una solución a largo plazo al problema de desabastecimiento que se plantea.

Hoy en día, RDAPs con nivel cero de pérdidas se consideran una utopía, tanto por las implicaciones técnicas como económicas que esto representa; no obstante, ha habido un gran avance en el conocimiento y desarrollo de equipos y técnicas, que permiten hacer un seguimiento más exhaustivo de las fugas en ellas. A continuación, se hace mención de las más importantes (Pilcher *et al*., 2007).

- Subdivisión de las redes en pequeñas subredes mediante el cierre temporal de válvulas e instalación de caudalímetros que permitan un monitoreo permanente de los caudales de ingreso
- Métodos tradicionales de cierres controlados de válvulas (o variaciones de estas técnicas)
- Uso de grabadores acústicos como herramientas de búsqueda (también conocidos como grabadores de ruido)
- Búsquedas sonoras

La sectorización de RDAPs puede ser considerada como una estrategia que pueden seguir los operadores de agua para mejorar la eficiencia de las mismas. Esta implica la subdivisión de las redes en subredes con una o varias entradas permanentemente controladas a través de caudalímetros. En cada segmento de subdivisión se puede manejar un valor máximo de demanda o longitud de tuberías, y dentro de ellos se trata de mantener una homogeneidad en lo que a elevación de terreno y tipo de usuarios se refiere. Como uno de los grandes beneficios de su implementación, se destaca el aumento de la facilidad con la que se detecta cualquier anormalidad dentro de la red debido a la reducción espacial implícita en la definición de los sectores. Contar con una red sectorizada permite no sólo aplicar técnicas particulares de control de fugas sino, además, permite implementar modelos de gestión diversos (DVGW, 2003; GIZ, 2011).

Este concepto fue introducido a principios de la década de los 80 en el reporte 26 de Control de Pérdidas y Prácticas de las Asociación de la Industria del Agua de

Inglaterra (UKWIR por *UK Water Industry Association*), tras la implementación de sectores en dos ciudades británicas, Plymouth y Rickmansworth (Farley, 2010). De hecho, esta implementación inspiró la aparición de la guía de sectorización propuesta por Morrison *et al*. (2007), que actualmente es uno de los pocos documentos técnicos sobre la temática y que se encuentra ampliamente extendido.

En la actualidad, la sectorización de RDAPs es empleada en muchos países del mundo, siendo más popular en Europa y América Latina. Lamentablemente, en la mayor parte de los casos en los que se ejecutan proyectos de sectorización, no se suele seguir el proceso con un rigor científico-técnico y, por el contrario, se suelen seguir aproximaciones de prueba y error. Es decir, basándose en mapas de la red (y en algunos casos con modelos hidráulicos), se seleccionan zonas de tal manera que se minimice el número de válvulas a instalar y, luego, se van probando esquemas de cierres y entradas de sectores, hasta llegar a un esquema (o varios) que reporte niveles de presión aceptables. Este tipo de aproximación puede ser válida para redes de menor extensión, ya que la magnitud de la inversión que se requeriría no debería ser tan elevada y el número de posibles esquemas no debería ser tan extenso. En este caso, con poco esfuerzo se pueden evaluar todas las alternativas de manera manual. Sin embargo, en redes de mayor extensión, y con mayor grado de mallado, el número de posibles esquemas puede ser muy grande, siendo totalmente adecuado el uso de técnicas de análisis computacional que permitan obtener soluciones factibles y óptimas. Lo anterior se torna aún más importante en la medida en que al problema se van agregando más objetivos. En el párrafo anterior se hizo únicamente mención a la presión como único objetivo; no obstante, tal y como se explicará posteriormente, existen otros objetivos que son también de gran importancia, tales como: el nivel de fiabilidad; la uniformidad de las presiones; el impacto sobre el consumo energético; el caudal de fugas; y la frecuencia de aparición de roturas de tuberías, entre otros.

En términos topológicos, la sectorización de las RDAPs puede ser abordada desde dos perspectivas generales, las cuales dependen de su topología. Si la red cuenta

con muchas fuentes de abastecimiento y estas, a su vez, se ubican dentro de las mallas de la misma, se puede pensar en realizar una subdivisión en función de la zona de influencia de cada fuente. En este caso cada sector contaría con una o dos fuentes exclusivas (ver Ilustración 1). Este tipo de esquemas tiene como ventaja el hecho de que no es necesario instalar caudalímetros (por ende, se evita la construcción de una arqueta para ubicar dicho caudalímetro) ya que, por norma general, las fuentes cuentan con un medidor que facilita efectuar el control de caudales. Sin embargo, hay dos desventajas que se asocian con este tipo de aproximación. Por un lado, se destaca el tamaño de los sectores, ya que, en redes con un número no muy grande de fuentes de abastecimiento, es de esperar que los sectores tiendan a tener una longitud de tubería considerablemente grande, lo que luego reduce el nivel de precisión en el análisis de caudales y la efectividad en la detección de pérdidas físicas. Adicionalmente, al tener zonas tan grandes, el aislamiento de zonas puntuales (e.g. para efectuar la reparación de una tubería), puede afectar a un mayor número de usuarios. Finalmente, se debe mencionar el efecto negativo sobre la resiliencia de la red, ya que, bajo este esquema operativo, un fallo de la única fuente de abastecimiento no es automáticamente compensado por otra (ver Ilustración 2).

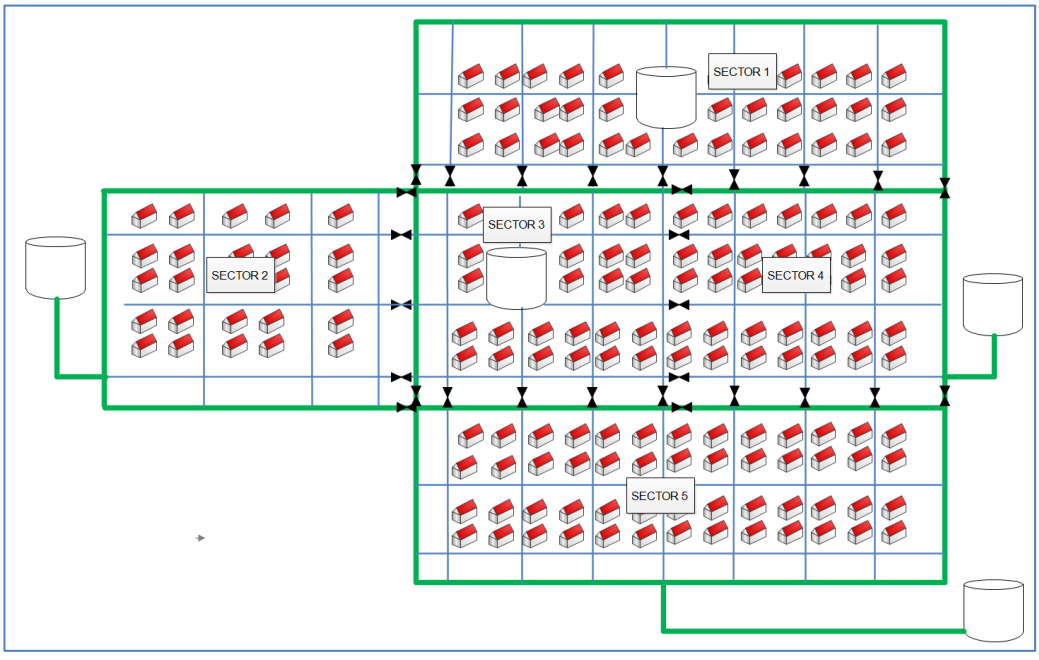


Ilustración 1: Esquema de sectorización por fuentes

En muchas redes, la implementación del tipo de esquema arriba descrito es inviable, dado que las fuentes se ubican fuera del mallado de la misma y/o constituyen un número reducido. Lógicamente, en este caso, no es posible establecer fuentes exclusivas por sector, ya que se tendrían sectores extremadamente extensos y poco fiables. En algunos casos podría ser imposible abastecer algunas zonas sin introducir rebombeos. En este tipo de redes, el abastecimiento se suele realizar mediante una red de conducción principal (o red troncal), a través de la cual se transporta el caudal desde las fuentes hasta las líneas de distribución (ver Ilustración 2). Esta red de conducción principal, por lo general, cuenta con un número relativamente reducido de conexiones, en su mayoría de diámetros medios. Dichas conexiones suelen corresponder a tuberías de abastecimiento a zonas residenciales, unidades de almacenamiento o bombeos intermedios, excluyéndose casi por completo las conexiones de tipo domiciliar.

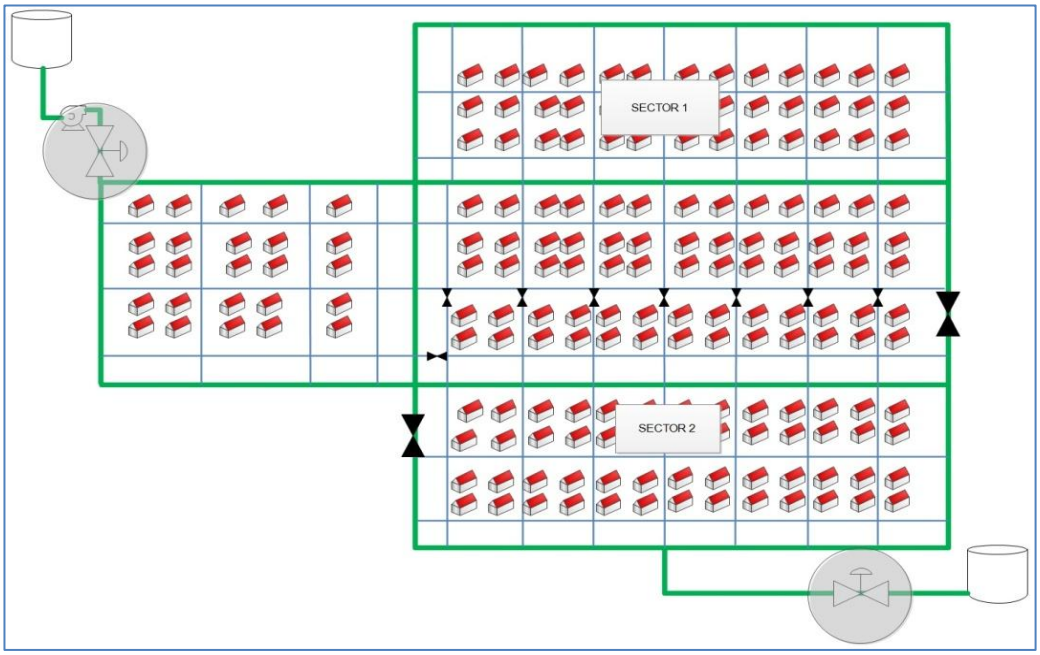


Ilustración 2: Esquema de sectorización sin segregar la red troncal

A diferencia del primer tipo de red, al sectorizar este segundo tipo de red (ver Ilustración 3), el número de sectores es incierto. Por un lado, se pueden establecer sectores más grandes, que ciertamente reducen la cantidad de válvulas que deben ser instaladas, aunque, tal y como se describió anteriormente, reducen la precisión

sobre el control permanente de caudal y la efectividad para aislar zonas puntuales. Por otro lado, se pueden diseñar sectores de menor extensión; sin embargo, esto implica un mayor nivel de inversión en la compra de válvulas de aislamiento, aunque siguiendo la línea de lo dicho hasta ahora, implican: mayor eficacia en lo que a detección de fugas se refiere; mayor precisión para el control permanente de caudal; y ventajas en términos de capacidad de aislamiento de zonas puntuales.

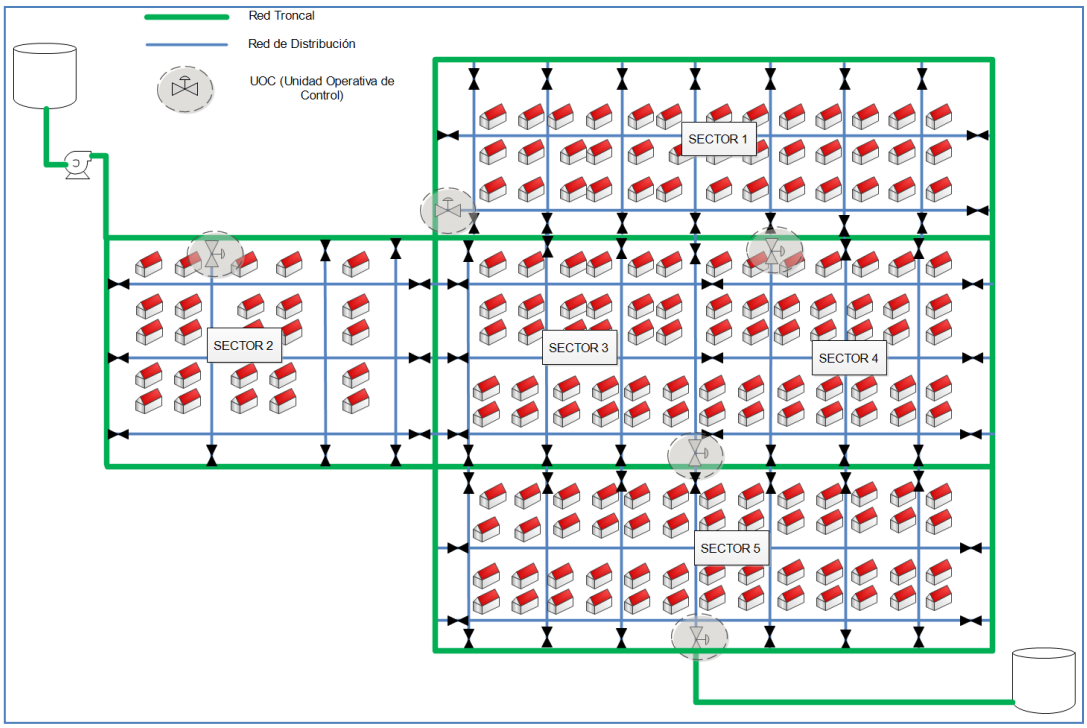


Ilustración 3: Esquema de sectorización segregando la red de conducción principal

Pese a los beneficios que conlleva la operación de una RDAP bajo un esquema de sectorización, hay ciertas desventajas que se deben tener en cuenta (ver resumen de ventajas/desventajas en la Tabla 1). En primer lugar, se puede destacar el efecto sobre la presión de entrega a los usuarios. Al tener que circular el caudal para abastecer a cada sector por una tubería (o un par), aumenta la fricción, la pérdida de carga y, por ende, disminuye la presión disponible. Pese a que una reducción de la presión puede generar un efecto positivo, en la medida que representa una reducción del caudal asociado a las pérdidas físicas (fugas de fondo, fugas no detectables y fugas detectables), también puede implicar desabastecimiento para

algunas zonas de la red en ciertos periodos del día (Thornton & Lambert, 2005; 2006; 2007). En algunos casos, el impacto puede no llegar a ser tan extremo como para generar desabastecimiento; sin embargo, el cierre de al menos una tubería, siempre reduce la fiabilidad del abastecimiento en cierto grado.

Con relación a los aspectos urbanísticos, no se puede soslayar el hecho de que la sectorización de una red implica llevar a cabo obras constructivas que pueden interrumpir de manera significativa las actividades normales de una ciudad. En ese sentido, es muy importante ubicar las válvulas de control de sectores y los medidores de caudal en lugares donde su acceso no sea muy complejo. Esto es particularmente importante para el caso de los medidores de caudal, los cuales tienen que ser ubicados en arquetas (normalmente hechas de hormigón) con un espacio suficientemente amplio como para poder colocar al menos cuatro válvulas de control (un par para el medidor principal y otro par para un bypass alternativo que sirva para suplir agua en caso de que la línea principal falle) y dar acceso a una persona para efectuar un registro de datos o para poder desarrollar actividades de mantenimiento (ver Ilustración 4). Desde este punto en adelante, tal estructura será definida como Unidad Operativa de Control (UOC).

Ilustración 4: Construcción de una UOC en un sector de la RDAP de la ciudad de Tegucigalpa, Honduras (Proyecto implementado por consultora española *Wasser* para el Sistema Nacional de Acueducto y Alcantarillados –SANAA– de Honduras)

Tras la implementación física de un esquema de sectorización, es de esperar que la velocidad con la que circula el agua en ciertas tuberías se reduzca de manera

drástica llegando a formar puntos muertos, propicios para el desarrollo de microorganismos responsables de enfermedades asociadas al consumo de agua contaminada. Esto es particularmente importante en los puntos que se encuentran más alejados de las entradas de los sectores, en donde se forman callejones sin salidas (tapones) en los que la velocidad del agua depende del consumo de agua por parte de un número reducido de usuarios.

Basándose en todos los aspectos mencionados anteriormente, se puede aseverar que un esquema de sectorización óptimo tiene que representar un compromiso entre una serie de aspectos positivos y negativos. El número de posibilidades puede ser considerablemente grande, dependiendo de la extensión de la red en la que se esté trabajando y su grado de mallado.

Tabla 1: Resumen de ventajas y desventajas de la sectorización

Ventajas	Desventajas
Seguimiento de evolución del comportamiento de las fugas	Inversión en compra e instalación de válvulas para cierre de sectores y UOCs
Aumento de la capacidad para aislar zonas en respuesta a anomalías	Aumento de la pérdida de carga debido a la reducción del área de tubería
Reducción de fugas de fondo como efecto de reducir la presión en la red de distribución	Disminución de la resiliencia de la red
Aumento de capacidad para monitorear de manera más focalizada parámetros de calidad de abastecimiento	Impacto negativo sobre la calidad del agua debido a la reducción de la velocidad de la misma
Ejecución de actividades de mantenimiento siguiendo un calendario que minimice las molestias a los clientes	Inconvenientes urbanísticos por la instalación de válvulas de cierre y UOCs
Facilidad para instalar dispositivos reguladores de presión	Posible aumento de consumo energético

El diseño de sectores en redes dependientes de una red troncal plantea una serie de aspectos que tienen que ser abordados de una manera distinta a como se abordarían en una red no dependiente de una red troncal, tales como: la definición de la red troncal; el número de sectores a establecer; y la ubicación de los mismos. Una vez establecidos los sectores, se hace necesario determinar las tuberías en las fronteras

de cada uno de los sectores obtenidos, que deben ser establecidas como límites de sector, y las líneas (o par de líneas) que deben ser establecidas como entrada(s) de sector.

Con relación al primero de los aspectos mencionados en el párrafo anterior, en algunos casos, las empresas operadoras de RDAPs cuentan con una definición previa de la red de conducción principal (la red de mayor diámetro sin conexiones menores). No obstante, siempre es conveniente llevar a cabo un análisis del comportamiento hidráulico de la red, para analizar el rol de las tuberías en el abastecimiento. Es una práctica relativamente común basar este análisis en lo diámetros de las tuberías; sin embargo, en algunos contextos, las redes se pueden encontrar sobredimensionadas o subdimensionadas, lo cual puede conducir a sobreestimaciones o subestimaciones de algunas de las tuberías. Por otro lado, basar el análisis sólo en la magnitud de los caudales puede generar el problema de dar excesiva importancia a tuberías dirigidas a zonas particulares, por ejemplo, una tubería que se dirige a un complejo industrial puede conducir caudales importantes; sin embargo, no es tan relevante para la hidráulica del resto del sistema. En relación a esto, Ferrari *et al*. (2014) señalan que la distinción entre tuberías principales y líneas de distribución es una función del tamaño de la RDAP; por lo tanto, tal distinción no es universal y varía de sistema a sistema. En el mismo artículo se menciona que es usual definir la red conducción principal a partir de las tuberías con un diámetro igual o mayor a aproximadamente 300 - 350 mm (12 - 14 pulgadas).

En este trabajo, la definición de la red de conducción principal parte de la simulación de la red en el escenario de máxima demanda; en este escenario se extraen los caudales y direcciones de caudales de cada tubería, y a continuación, se plantean una serie de cálculos basados en el concepto de “Detección de Caminos más Cortos” de la teoría de grafos. A través de estos cálculos, se establece una jerarquía de tuberías que refleja la importancia de cada tubería para el abastecimiento de toda la red. Finalmente, se establecen, mediante un análisis

estadístico, las tuberías que forman parte de la red de conducción principal y las que forman parte de la red de distribución.

Para abordar el segundo y el tercer problema mencionado, correspondiente al número y ubicación de sectores, se puede partir del hecho de que, en términos generales, la sectorización de una RDAP constituye un problema de agrupación de elementos (conexiones/nodos y usuarios) que permite tener mejor control sobre las mismas. Dicha agrupación puede ser abordada desde el ámbito de las técnicas de clústering en grafos, partiendo de la premisa de que las redes pueden ser representadas como grafos. Un grafo, es una estructura ampliamente utilizada en el campo de la ciencia de la computación y la matemática, que permite expresar la relación (a través de aristas o enlaces) entre un conjunto de elementos o vectores (nodos). Es importante hacer notar el paralelismo existente entre el concepto de grafo y el concepto de RDAPs, lo que permite la representación de las mismas como grafos. De manera similar al caso de la sectorización, los algoritmos para detectar clústeres en grafos, pueden ser subdivididos en dos tipos, por un lado, se encuentra el conjunto de algoritmos en donde el número de subdivisiones es *a priori* conocido, y, por otra parte, se ubican los algoritmos en los cuales, el número de subdivisiones no se conoce y por el contrario debe ser obtenido basándose en la estructura del conjunto de datos que se desean subdividir, o la topología de la red de interés. El segundo tipo es conocido como detección de comunidades en redes sociales, siendo una red social un conjunto de elementos interconectados por uno o más enlaces, que pueden ser subjetivos (por ejemplo, la amistad entre dos personas en *facebook*) o materiales, tal como sería la conexión entre dos o más nudos de una red mediante tuberías.

Es importante hacer notar la diferencia que existe entre los conjuntos de datos en forma de redes sociales y los conjuntos de datos sin ningún tipo de interconexión. En el primer caso, los elementos suelen ser vectores de una única dimensión (el nombre de una persona, el nombre de una ciudad, el ID de un nodo de una RDAP), y en el segundo caso estos suelen ser multidimensionales (las características de las

tuberías de una RDAP: diámetro, rugosidad, material, etc.).

I.2 Técnicas de Sectorización: Estado del Arte

En la presente sección se hace un análisis del estado del arte de la temática. Frente a cada uno de los documentos/trabajos/artículos citados se coloca una numeración que permite su identificación posterior en una tabla resumen (ver Tabla 2).

El documento más ampliamente conocido en torno a la técnica de sectorización corresponde a **(1)** las Directrices de Sectorización de Distritos Métricos (*District Metered Areas: Guidance Notes*) desarrollado por algunos de los miembros de Grupo de Trabajo de Pérdidas de Agua (WLTF por las siglas de "*Water Loss Task Force*") de la Asociación Internacional del Agua (IWA por las siglas de *International Water Association*) (Morrison *et al*. 2007). En dicho documento se plantean una serie de conceptualizaciones en torno a la técnica y, adicionalmente, se describen algunos aspectos a tener en cuenta para gestionar redes bajo un esquema de sectorización. Con respecto a la definición de sectores, sólo se plantean algunas pautas generales a tener en cuenta, tales como:

- Nivel económico de fugas requerido
- Tamaño (área geográfica y número de conexiones)
- Tipo de hogares
- Variación topográfica
- Consideraciones de calidad del agua
- Requerimientos de presión
- Capacidad de hidrantes
- Nivel de fugas objetivo
- Número de caudalímetros
- Condiciones de infraestructura

Desde la publicación del documento previamente descrito, hasta la fecha, varias metodologías han sido propuestas a fin de hacer frente al problema de la sectorización automática de RDAPs. En esencia, la mayor parte de estos estudios han abordado el problema mediante la teoría de grafos en combinación: con Sistemas Multi-agente (o MAS por las siglas en inglés de *Multi-agent System*), enfoques de optimización, criterios energéticos y, muy recientemente, mediante algunos conceptos derivados de la teoría de redes sociales.

En **(2)**, Tzatchkov *et al*. (2008), se describe la aplicación de dos algoritmos de exploración de grafos: Búsqueda en Amplitud (BFS por *Breadth First Search*) y Búsqueda en Profundidad (DFS por *Depth First Search*) (ver detalles de algoritmos en el Capítulo II), junto con la capacidad de simulación de la calidad del agua en el software de simulación hidráulica EPANET (Rossman, 2000), a fin de obtener el número de subredes independientes en una RDAP. Pese a ser el primer trabajo en que se aborda el problema de la sectorización mediante el uso de grafos, no establece ningún método para definir sectores. **(3)** Di Nardo & Di Natale (2011) presentan una metodología de soporte heurístico basado en el concepto Caminos más Cortos (CMCs) en grafos (ver detalles sobre el concepto de CMC en el Capítulo III), que permite la construcción de un grafo principal. Sobre este grafo principal, el número de sectores se define basándose en el índice de resiliencia propuesto por Todini (2000). Si bien se presenta una metodología que genera un esquema de sectorización que minimiza tanto el impacto sobre la resiliencia de la red así como el coste de implementación, no evalúa otros aspectos económicos relacionados con la sectorización, tales como los beneficios en términos de reducción de fugas, reducción/aumento del consumo energético o aspectos de calidad.

El concepto de CMC también se ha utilizado en la metodología propuesta por **(4)**, Alvisi & Franchini (2014), en combinación con el algoritmo DFS. En este método, la exploración con DFS se inicializa desde cada uno de los nodos de la red a fin de encontrar áreas que cumplan con un criterio de número de usuarios. Luego,

mediante el concepto de CMC y basándose en algunas características hidráulicas de las tuberías, se determinan las entradas a cada sector. El resultado es un conjunto de soluciones, la cuales son comparadas utilizando el índice de resiliencia previamente mencionado. Al igual que en el trabajo descrito en el párrafo anterior, este no incluye ningún tipo de abordaje económico.

En **(5)**, Perelman & Ostfeld (2011), se describe una metodología para encontrar grupos de nodos fuertemente conectados. Utilizan el algoritmo DFS. En este trabajo se hace un abordaje más general de la sectorización. En concreto, sólo se hace la identificación de los grupos, pero no se aborda ni el aislamiento de los mismos, ni el análisis de ningún criterio hidráulico o económico. **(6)**, Gomes (2012), presenta una metodología algo distinta que las anteriores. En la misma se parte de un esquema de sectorización ya diseñado y se optimiza: la ubicación de las entradas; el punto de funcionamiento de Válvulas Reguladoras de Presión (VRPs) que se colocan en dichas entradas; el coste de la implementación; y los beneficios económicos derivados de la gestión de la presión. El modelo se resuelve mediante el uso del algoritmo de optimización *Recocido Simulado* (*Simulated Annealing*). Si bien la metodología hace un análisis de los beneficios en términos de reducción de fugas de fondo, no toma en cuenta el efecto sobre fugas futuras ni sobre la resiliencia de la red. Además, no contempla la definición de los sectores. En **(7)**, Di Nardo *et al*. (2013), se aborda el problema utilizando el concepto de CMC a fin de hacer una definición inicial del diseño de sectores. Luego, los límites entre sectores se negocian a través de optimización mediante Algoritmos Genéticos (AG). Por otro lado, **(8)**, Di Nardo *et al*. (2014), propone una metodología basada en el algoritmo DFS y el índice de resiliencia anteriormente mencionado, a fin de encontrar un diseño de sectorización en el que se minimice el aumento de pérdida de carga. Nuevamente, al igual que en los tres trabajos anteriormente citados, las metodologías carecen de un análisis económico.

En un enfoque relativamente diferente, **(9)**, Izquierdo *et al*. (2009), y **(10)**, Herrera *et al*. (2012), proponen el uso de una combinación entre clústering espectral y MAS

para encontrar la zona de influencia de cada fuente de agua en una determinada red. Las metodologías parten de un número inicial de sectores y contemplan la generación de sectores alrededor de las fuentes de abastecimiento. Lo último les hace inaplicables a RDAP dependientes de una red troncal y, por otra parte, no toman en cuenta el efecto de la sectorización sobre la resiliencia de la red ni los beneficios económicos de su implementación.

(11), Hajebi *et al*. (2013), también implementa MAS para llevar a cabo un proceso de negociación entre los nodos situados en los límites de los sectores, previamente definidos, por medio de la técnica de detección de clústeres, *k-means*. Al igual que el caso anterior, no toma en cuenta ni los aspectos económicos de la sectorización, ni la resiliencia de la red.

(12), Hajebi *et al*. (2014), presenta una metodología que parte de la definición de la red troncal mediante la técnica de detección CMC en grafos. Los sectores luego son definidos a partir de la misma, y a continuación, mediante AGs se detectan las líneas que deben ser seleccionadas como entradas de sector y las que deben ser definidas como límites entre sectores. La optimización del Conjunto de Entradas y Válvulas de Cierre (CEVC) de sectores se centra en parámetros hidráulicos y no tiene en cuenta ningún aspecto económico.

(13), Diao *et al*. (2013), describe un enfoque basado en el concepto de detección de comunidades, propio de la teoría de redes sociales. En concreto, aplica el algoritmo de detección de comunidades planteado por Newman & Girvan (2004), y luego, mediante un modelo iterativo, define las líneas de entradas y de límite de cada sector. Este fue el primer trabajo en plantear el uso de la técnica de detección de comunidades en redes sociales para abordar el problema de sectorización, y a pesar de minimizar el efecto negativo de la sectorización sobre la resiliencia de la red, al igual que en los trabajos anteriormente citados, el mismo carece de un abordaje económico adecuado. En una línea relativamente similar, **(14)**, Campbell *et al*. (2014a), compara el uso de la técnica de clústering jerárquico y un algoritmo de

detección de comunidades en redes sociales para sectorizar RDAPs dependientes de una red troncal. El trabajo aporta el abordaje de las redes dependientes de redes de conducción principal y minimiza el efecto de la sectorización sobre la pérdida de resiliencia; no obstante, también carece de un adecuado abordaje económico. **(15)**, Giustolisi & Ridolfi (2014 a,b) y Boano & Berardi (2014), muestran algunas limitaciones en el uso del llamado Índice de Modularidad anteriormente mencionado para detectar comunidades en RDAPs. Para tales fines, dichos autores proponen un ajuste del mismo indicador. **(16)**, De Paola *et al*. (2014), presentan un modelo de optimización resuelto a través de un algoritmo bioinspirado. El modelo incluye, además de restricciones hidráulicas, el coste asociado a la compra de medidores de caudal; el coste asociado a la compra de válvulas nuevas; y los costes asociados a la reducción de caudal de fugas derivado de la reducción de la presión de la red; además tiene en cuenta un análisis de la inversión a largo plazo. Pese a ser la metodología más completa en términos de análisis económicos, sólo toma en cuenta la reducción de fugas producto de la reducción de presión; sin embargo, no considera la reducción de fugas futuras. En Ferrari *et al*. (2014) y Savic & Ferrari (2014), se describe un método, **(17)**, en el cual, basándose en los diámetros de las tuberías, se detecta, inicialmente, la red troncal y a partir de dicha línea, se detectan sectores a través de una exploración con el algoritmo BFS. Del conjunto de sectores encontrados se descartan aquellos que no satisfacen un número mínimo de conexiones. Por otro lado, los sectores que sí cumplen con un número mínimo de conexiones son recursivamente subdivididos, hasta obtener una configuración de sectores que cumplen con un rango de número máximo de conexiones. Si bien en la metodología se optimiza la resiliencia de la red, no plantea ningún tipo de análisis económico.

En el marco del presente trabajo se realizaron una serie de 10 publicaciones, entre artículos de congreso y revistas de alto impacto. Dos artículos de gran importancia corresponden a **(18)**, Campbell *et al*. (2016a) y **(19)** Campbell *et al*. (2016b). En el primero se propone un método de sectorización basado en detección de comunidades mediante caminos aleatorios y optimización mediante *Optimización*

de Enjambre de Agentes (ASO, por las siglas en ingles de *Agent Swarm Optimization*). En este trabajo se presenta la primera metodología basada en detección de comunidades en redes sociales que toma en cuenta aspectos económicos dentro de la optimización del CEVC; sin embargo, la misma no tiene en cuenta el efecto sobre fugas futuras. En el segundo artículo sí se aborda el último aspecto, siendo de hecho el primer trabajo en que se lleva a cabo tal tipo de análisis. A continuación, se presenta un listado que permite categorizar los trabajos de sectorización anteriormente citados. Las letras que se encuentran al final de cada una de las categorías, corresponde a la codificación que se emplea en la Tabla 2.

- Sectores alrededor de fuentes de abastecimiento (A)
- Tratamiento especial de la red troncal (B)
- Análisis de resiliencia (C)
- Coste de compra e instalación de válvulas de cierre y UOCs (D)
- Efecto de reducción de fugas por efecto de reducción de presión (E)
- Efecto sobre consumo energético (F)
- Análisis de inversión a largo plazo (G)
- Análisis de fugas futuras (H)
- Otros aspectos (I)

En la Tabla 2 se presenta un resumen de todos los trabajos citados.

Tabla 2: Resumen del estado del arte

			Aspecto abordado								
Trabajo	Objetivo	Algoritmo/Técnica	A	B	C	D	E	F	G	H	I
1	Establecer pautas generales sobre la implementación de sectorización	Ninguno									
2	Encontrar área de influencia de las fuentes	BFS y DFS									
3	Encontrar esquema de sectorización donde se minimice el impacto negativo sobre la fiabilidad del sistema	Optimización y CMC									
4	Encontrar conjunto de alternativas factibles, desde donde se puede seleccionar alguna con base en criterios de fiabilidad	DFS y CMC									
5	Encontrar grupos de nodos fuertemente conectados	DFS									
6	Optimizar colocación de VRPs en entrada de sectores	SA									
7	Encontrar sectores fiables y uniformes	CMC y AG									
8	Encontrar esquemas de sectorización en los cuales se minimice la pérdida de carga	DFS									

			Aspecto abordado								
Trabajo	Objetivo	Algoritmo/Técnica	A	B	C	D	E	F	G	H	I
9	Encontrar zonas con uniformidad de características	MAS									
10	Encontrar zonas con uniformidad de características	MAS + Clústering Espectral									
11	Encontrar zonas uniformes	*k-means* + MAS									
12	Encontrar sectores dependientes de una red de conducción principal	CMC + AG									
13	Sectorización de redes dependientes de una red de conducción principal	Detección de comunidades									
14	Comparación de técnicas de detección de comunidades	Detección de comunidades									
15	Adaptar índice de Modularidad a fin de poder abordar la sectorización de RDAPs mediante algoritmos de detección de comunidades en redes sociales	Detección de comunidades									
16	Optimizar la inversión de la sectorización tomando en cuenta el beneficio en términos de reducción de fugas de fondo	AG									
17	Definición de sectores en redes dependientes de una red troncal	CMC + BFS									

			Aspecto abordado								
Trabajo	Objetivo	Algoritmo/Técnica	A	B	C	D	E	F	G	H	I
18	Definición de sectores en redes dependientes de una red de conducción principal	Algoritmo de detección de comunidades + ASO									
19	Definición de sectores en redes dependientes de conducción principal teniendo en cuenta beneficios en término de detección de fugas futuras	Algoritmo de detección de comunidades + AG + SMC									

I.3 Objetivos de la Tesis

Objetivo General

La presente tesis persigue como objetivo general plantear una metodología para sectorizar RDAPs basada en detección de comunidades en redes sociales y en técnicas heurísticas de optimización. Esta metodología deberá equilibrar la calidad del suministro con aspectos económicos propios de la implementación de los esquemas de sectorización resultantes.

Objetivos Específicos

1. Definir los puntos de la red a partir de donde se establecen los sectores. Esto se realiza mediante un método basado en el concepto de caminos más cortos propio de la teoría de grafos.

2. Comparar la aplicación de tres algoritmos (clústering jerárquico, método Louvain y caminos aleatorios) de detección de comunidades en redes sociales para definir sectores en la red de distribución.

3. Establecer un marco metodológico para relacionar el diseño de sectores con la capacidad de detectar nuevos eventos de fugas.

4. Optimizar, mediante algoritmos evolutivos, el coste/beneficio de la sectorización. Para ello, se establecen como variables de decisión los estados (abierto/cerrado) del grupo de líneas que forman parte del CEVC.

I.4 Desarrollo de los Objetivos

De manera general, el trabajo partió de un análisis de las metodologías de sectorización que se han planteado hasta la actualidad. Se hizo énfasis en las metodologías en las que se generan sectores que no cuentan con su propia fuente de abastecimiento, así como de los factores económicos a tener en cuenta al momento

de implementar proyectos de sectorización. Se realizó un trabajo de programación en varios lenguajes/herramientas de programación R, C#, Visual Basic y Matlab. Para fines de evaluación, el método propuesto ha sido implementado en una sección de la RDAP de la ciudad de Managua, Nicaragua, en la cual se ejecutó, por parte de la consultora española WASSER S.A., un proyecto de optimización del suministro en el año 2009. El autor formó parte del departamento encargado del proyecto de sectorización.

Objetivo específico 1:

Establecer un mecanismo para definir la red troncal en una RDAP, que es la red que conecta la(s) fuente de abastecimiento con la red de distribución. A partir de esta red se establecen los sectores. Para la definición en cuestión se implementa, sobre el grafo de la RDAPs, el concepto de Caminos más Cortos, propio de la teoría de grafos.

Objetivo específico 2:

Para abordar este objetivo se programaron, en el lenguaje de programación estadístico *R*, procesos de definición de sectores basados en algoritmos de detección de comunidades en redes sociales: clústering Jerárquico, método Louvain y Caminos Aleatorios. Mediante criterios topológicos/operativos, se forman los sectores a partir de las comunidades generadas por los algoritmos de detección de comunidades en redes sociales. También se hizo una comparación entre todos los métodos, exponiendo sus ventajas y desventajas.

Objetivo específico 3:

Para abordar este objetivo se hizo un análisis de los parámetros de control de fugas descritos en la Metodología de Estimación de Roturas y Fugas de Fondo (BABE por "*Burst and Background Estimate*") de la Asociación Inglesa de la Industria del Agua (UKWIR, por las siglas en inglés de UK Water Industry) (UKWIR, 1994) y se

definió un mecanismo para relacionar tales parámetros con el diseño de sectores. En particular, se prestó especial atención a la relación entre el diseño de sectores y la detección de fugas futuras. Para el abordaje de este último aspecto se hizo uso de simulación Monte Carlo (SMC).

Objetivo específico 4:

En los sectores generados mediante los métodos basados en algoritmos de detección de comunidades se tiene que definir el CEVC. Esto se relaciona con el coste de implementación de los sectores, ya que cada configuración además de modificar el coste de implementación, modifica los costes de operación de la red. La optimización se realiza mediante tres métodos estocásticos: AG, ASO y un esquema multinivel basado en PSO. Para tal fin se realizó un trabajo de programación en los leguajes, Visual Basic, C# y Matlab.

I.5 Organización del documento

El presente trabajo está organizado de la siguiente manera:

Capítulo 2: *Metodología de Sectorización basada en la Detección de Comunidades en Redes Sociales*

En este capítulo se describe la metodología para diseñar sectores en RDAPs basada en algoritmos de detección de comunidades en redes sociales. En la primera parte del capítulo, se hace una descripción de aspectos de la teoría de grafos que son relevantes para la descripción del método de sectorización propuesto. En la segunda parte se describe el proceso de identificación de la red troncal mediante la implementación del concepto de CMC de la teoría de grafos. En la tercera parte, se abordan algunos aspectos teóricos sobre la teoría de redes sociales y detección de comunidades en las mismas, incluyendo una amplia descripción sobre el concepto de modularidad, que es el criterio más empleado por los algoritmos de detección de comunidades para evaluar la calidad de las particiones. En la cuarta parte, se describe la manera en que se genera un grafo a partir de un modelo matemático de

la red. En las partes 5, 6 y 7 se abordan la generación de sectores a partir de los algoritmos de detección de comunidades: clústering jerárquico, método Louvain y caminos aleatorios. En cada una de estas partes se describe el algoritmo en cuestión, la implementación del mismo, y se ejemplifica mediante su implementación en la RDAP de Managua. En la parte final, se hace una comparación de los tres métodos y se analizan las ventajas y desventajas que cada uno de ellos ofrece.

Capítulo 3: *Gestión de Pérdidas en Redes de Abastecimiento de Agua Potable Mediante Sectorización: Optimización del Conjunto de Válvulas de Cierre/Entradas de Sectores*

Este capítulo se centra en la optimización del CEVC, es decir, de las líneas que conectan a la red troncal con los sectores. Para ello se empieza, en la parte 1, con una descripción sobre aspectos relacionados con la gestión sostenible de pérdidas en RDAPs, haciendo especial hincapié en la metodología BABE propuesta por UKWIR. En la segunda parte se hace una descripción de criterios para evaluar el efecto de los CEVCs generados en la calidad del suministro (edad del agua), la topología (uniformidad de demandas y de elevaciones) y la fiabilidad del sistema (resiliencia). En la tercera parte se aborda la optimización del CEVC, primero mediante AGs, luego mediante ASO y, finalmente, mediante optimización multinivel basada en PSO. En la optimización con AGs se llevan a cabo SMCs a fin de predecir la ocurrencia y reparación de nuevos eventos de fugas, por lo cual en esta parte también se hace una descripción amplia de este método de simulación. En el segundo método se plantea la colocación de VRPs y en el tercero sólo se optimiza el CEVC. En el segundo de los casos la optimización se lleva a cabo en dos niveles. En el primer nivel se optimiza el CEVC y en el segundo nivel se optimiza el punto de funcionamiento de las VRPs. Finalmente, en la última parte se hace una comparación de los tres métodos/abordajes de optimización, describiendo las ventajas y desventajas de cada uno de ellos.

Capítulo 4: *Conclusiones y Líneas Futuras*

En este capítulo se presentan las conclusiones de todo el trabajo, incluyendo las conclusiones sobre las ventajas/desventajas de los métodos de detección de comunidades empleados, el resultado de la metodología propuesta y los beneficios de los esquemas de optimización presentados. Igualmente, se presenta una lista de futuras líneas de trabajo, haciendo énfasis en el método para evaluar la aparición de fugas futuras en función del esquema de sectorización a implementar.

Apéndice I: *Método de Calibración Mediante Mapas Auto-Organizados y Algoritmos Genéticos*

En esta sección se describe el método de calibración de RDAPs propuesto en la tesis. El método persigue la calibración no sólo teniendo en cuenta la rugosidad en las tuberías, que es el parámetro que se suele tener en cuenta en la mayor parte de los métodos de calibración de RDAPs, sino también la distribución de las fugas a través de la red. Se hace una descripción de un método de clasificación de nodos y tuberías en RDAPs mediante SOMs y clústering jerárquico. Con este método se divide la red en sub-áreas. Los parámetros rugosidad y coeficientes de emisor en cada sub-área son a continuación ajustados mediante el uso de AGs. El método de calibración se implementa sobre el modelo de la red de Managua, que es la red utilizada en los ejemplos de implementación a lo largo de la tesis.

Apéndice II: En este apéndice se describen los códigos preparados en *R*, *Visual Basic*, *C#* y *Matlab*.

II. METODOLOGÍA DE SECTORIZACIÓN BASADA EN LA DETECCIÓN DE COMUNIDADES EN REDES SOCIALES

II.1 Teoría de Grafos

La teoría de grafos es la rama de las matemáticas de conjuntos discretos (finitos e infinitos), o simplemente matemáticas discretas, que se encarga del análisis de este tipo de estructuras. Un grafo $G = \{V, E\}$ es un conjunto de objetos llamados vértices (V) o nodos (de ahora en adelante los términos son intercambiables) unidos por enlaces, aristas o arcos (de ahora en adelante los términos son intercambiables) (E), que permiten representar relaciones binarias entre los elementos de un conjunto (Thulasiraman & Swamy, 1992). El término grafo fue empleado por primera vez a mediados del siglo XVIII y se originó a partir de la "*notación gráfica*" que entonces, y aún en la actualidad, permite describir los enlaces moleculares en el ámbito de la química orgánica/bioquímica. El origen de la teoría de grafos como ciencia se da en el mismo siglo, ubicándose en la antigua Könisberg (actualmente Kaliningrado) en Rusia. Leonard Euler, publicó un documento en el que se le daba solución al llamado *Problema de Könisberg*, que perseguía encontrar una manera de recorrer los siete puentes de la ciudad pasando por cada uno de ellos una sola vez (Biggs *et al*., 1986; Goset, 2009). Posteriormente, en el año 1847, Gustav Kirchoff, produce una serie de publicaciones, en la que hace uso de la teoría de grafos para establecer las ampliamente conocidas leyes para cálculo de corriente, voltaje y resistencia en los circuitos eléctricos.

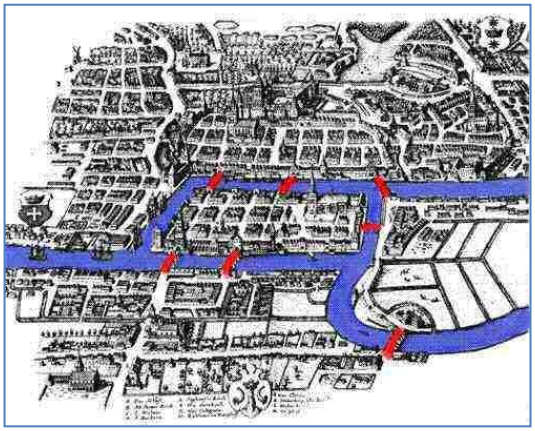


Ilustración 5: El problema de los puentes de Könisberg [Fuente: Johnson (2017)]

Pese a que los grafos pueden llegar a representar estructuras muy complejas, su estructura elemental es bastante simple. El grafo más simple es aquel que está compuesto únicamente por una arista conectando dos vértices (o una arista incidente a dos vértices).

Existen varios criterios topológicos que permiten clasificar los grafos y/o sus componentes. Por ejemplo, según de dónde parten y hacia dónde se dirigen los vértices de las aristas de un grafo, estas pueden ser clasificadas en: adyacentes, en el caso de que dos aristas tengan un vértice en común; paralelas o múltiples, en el caso de compartir los dos vértices; lazo, en el caso de que el vértice de partida sea el mismo vértice de llegada.

Dependiendo de si las aristas del grafo tienen un punto de partida y un punto de llegada definido o no, los grafos pueden ser clasificados como grafos dirigidos (dígrafos) o como grafos no dirigidos. En el caso de los grafos dirigidos, el vértice origen de cada arista es conocido como cola de la arista, y el vértice destino es conocido como cabeza. A su vez, los grafos dirigidos pueden ser clasificados en grafos simétricos, que corresponden a grafos dirigidos en los que para cada una de sus aristas direccionadas existe una correspondiente arista direccionada inversa; grafos dirigidos acíclicos, que corresponden a grafos dirigidos con ciclos no dirigidos; y grafos tipo torneo, que corresponden a grafos que se obtienen mediante la selección de una dirección para cada enlace en un grafo no dirigido completo.

Es común referirse a los grafos no dirigidos como grafos de aristas no ordenadas. En tal caso, los grafos son representados por puntos conectados por enlaces sin dirección. Por el contrario, es común referirse a los dígrafos como grafos de aristas ordenadas y son representados mediante puntos unidos a través de flechas.

Formalmente, un grafo no dirigido se puede formular de la siguiente manera:

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

Es decir, se asume que las aristas están formadas por un par no ordenado de nodos.

En función del número de aristas adyacentes a los vértices, los grafos se pueden clasificar como *k*-regulares, en el caso que todos los nodos cuenten con el mismo número de enlaces adyacentes y *k*-irregulares en caso contrario.

II.1.1 Caracterización de los Grafos

En las últimas dos décadas se han desarrollado una serie de criterios topológicos para caracterizar los grafos, lo que permite hacer comparaciones entre los mismos y medir la eficiencia de algoritmos de exploración de grafos (algunos de los cuales serán descritos más adelante). Algunos de estos criterios se implementan directamente sobre los componentes estructurales del grafo y algunos tienen una aplicación más global (se aplican sobre el grafo como un conjunto). Una de las propiedades más importantes es el grado de conectividad de los vértices (también conocido simplemente como grado o como valencia). Esta propiedad/criterio, corresponde al número de aristas que inciden sobre los vértices (los lazos o bucles se cuenta como dos aristas). Puede darse el caso de que ninguna arista incida en un vértice dado, en tal caso, el vértice se define como vértice aislado, o vértice de grado cero. Luego, en el caso de que sobre un vértice dado incida únicamente una arista, el mismo se define como vértice de grado 1 (también se denominan vértices hojas o vértices finales y sus aristas incidentes son llamadas artistas pendientes). En tanto, un vértice con conexión directa al resto de nodos del grafo (grado n-1) es

llamado vértice dominante. En esta misma línea, los grafos compuestos únicamente por dos vértices y una arista se definen como vértices pendientes (tal como se señaló arriba, este es el tipo de grafo más simple que puede existir). Por otro lado, cuando en un grafo todos los vértices tienen varias aristas y cada una de ellas tiene como punto de partida otro nodo del grafo, se dice que se trata de un grafo completo. En este caso, el número de aristas incidentes sobre cada uno de los vértices es *n*-1, donde *n* es el número de vértices que constituyen el grafo en cuestión. Este mismo número constituye el grado de cada uno de los vértices. Tal propiedad (grado) puede ser extrapolada al grafo, siendo el grado de un grafo regular igual al grado de todos sus vértices (ya que todos los nodos tienen el mismo grado); y en el caso de un grafo irregular, el grado del nodo con mayor número de aristas incidentes define el grado máximo del grafo, en tanto el nodo con menor número de aristas incidentes define el grado mínimo.

De acuerdo al popularmente conocido lema "*Handshaking*" (Euler, 1736), en cualquier grafo finito no dirigido, el número de vértices con grado impar es par. Es una consecuencia de la fórmula de suma de grados:

$$\sum_{v \in V} deg(v) = 2|E|. \qquad \text{(Ecuación 1)}$$

En esta ecuación V representa el conjunto de los vértices del grafo, la función $deg(\cdot)$ proporciona el grado de un vértice, y $|E|$es el número de aristas.

De acuerdo al peso (también conocido como coste) que se le asigne a las aristas, los grafos pueden ser clasificados como ponderados, en caso de que las aristas cuenten con un peso, o no ponderados en caso contrario. En el último caso también se puede decir que todas las aristas tienen peso 1.

Uno de los aspectos más estudiados dentro de la teoría de grafos es la accesibilidad entre los vértices, la cual se estudia a través de recorridos a lo largo de los grafos. Tales recorridos pueden formar cadenas, dado que son sucesiones finitas de aristas y vértices. Estas cadenas pueden ser abiertas o cerradas en función de si el vértice

inicial coincide o no con el vértice final. Las cadenas también pueden ser denominadas como caminos, en caso de que no se repita ninguno de los vértices ni aristas en el mismo. En el caso de que únicamente se repitan el nodo inicial y el final, se denominan ciclo.

Dos de los ciclos más conocidos y empleados dentro de la teoría de grafos son el ciclo de Hamilton y el ciclo de Euler. En el primero, se recorren todos los vértices sin repetir ninguno, salvo el primero y el último (que son el mismo). En el caso que el primer y último vértice sean distintos, pero se mantenga la regla de no repetición de vértices en el recorrido, el recorrido es denominado únicamente como camino Hamiltoniano. En cualquier caso, el número de aristas que forman parte del camino constituyen su longitud. Esto, a su vez, se relaciona con el concepto de diámetro del grafo, que es la longitud máxima que se requiere recorrer para llegar desde un nodo hasta otro.

Dentro de los grafos se pueden definir subgrafos, que corresponden a subconjuntos de aristas de un grafo (y sus respectivos vértices asociados). También puede haber componentes desconectados, si entre dos estructuras no existe un camino posible, siendo el grafo que contiene dichos componentes, un grafo desconectado.

Otras estructuras que permite caracterizar a los grafos son los árboles, que corresponde a grafos conexos simples y acíclicos. Un conjunto de árboles desconectados o disjunto corresponde a un bosque. El árbol que permite conectar a todos los vértices de un grafo corresponde al árbol de máxima expansión.

II.1.2 Representación Matricial de Grafos

Los grafos pueden ser representados mediante matrices cuadradas de tipo $n \times n$, siendo en *n* el número de vértices que constituyen el grafo.

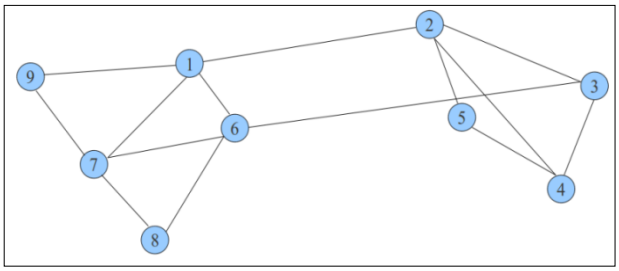


Ilustración 6: Grafo maqueta

Tomemos como ejemplo el grafo maqueta de la Ilustración 6. Esta misma red (grafo) podría ser representada mediante una matriz de adyacencia *W*, la cual es una matriz binaria $n \times n$ en la cual la adyacencia entre dos vértices se denota con el valor 1, y la no adyacencia con el valor 0 (ver Tabla 3).

W_{ij} = 1 si *i* y *j* son adyacentes (Ecuación 2)

W_{ij} = 0 si *i* y *j* no soy adyacentes

	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

Tabla 3: Matriz de adyacencia (W_{ij}) del grafo maqueta de la Ilustración 6

Como es de esperar, la diagonal de la matriz *W* sólo tomará el valor 0 en vista de que un vértice no puede estar conectado a sí mismo (salvo que existan bucles).

Por otro lado las matrices de disimilaridad representan las diferencias entre las propiedades que pueden tener los vértices de un grafo. Estas diferencias pueden ser medidas mediante distintas distancias métricas: Euclidiana, Manhattan, entre otras.

Esto será profundizado en la de clústering jerárquico (ver detalles en la Subsección II.6).

La matriz de grado, *D*, es una matriz diagonal que contiene información referida al grado de cada vértice tal como se muestra en la siguiente ecuación:

$$D_{ij} = \begin{cases} 0, si\ i \neq j \\ deg(V_i), si\ i = j \end{cases} \quad \text{(Ecuación 3)}$$

correspondiendo cada elemento de la diagonal al número de aristas incidentes en el nodo correspondiente.

La matriz de afinidad, *A*, por otro lado, contiene información de los pesos de las aristas incidentes en cada uno de los nodos. Así, si a las aristas del grafo anterior se les asignaran pesos (S_{ij}) (ver Ilustración 7), la matriz *A* vendría dada según:

$$A_{ij} = S_{ij} \text{ si } i \text{ y } j \text{ están conectados}$$
$$A_{ij} = 0 \text{ si } i \text{ y } j \text{ no están conectados} \quad \text{(Ecuación 4)}$$

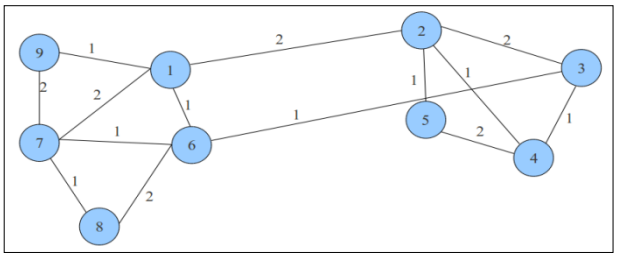


Ilustración 7: Grafo maqueta ponderado

La matriz *A* del grafo anterior sería la que se presenta en la Tabla 4. En la matriz *A*, la diagonal, en general, no toma valores diferentes a cero a la vista de que un nodo no suele estar conectado a sí mismo, salvo en el caso que existan bucles.

	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

Tabla 4: Matriz de afinidad $A = (A_{ij})$ del grafo de la Ilustración 7

La matriz Laplaciana, *L*, tal y como se muestra en Ecuación 5, es igual a la resta de la matriz *D* menos la matriz *A*.

$$L = D - A.$$

$$L_{ij} = \begin{cases} \deg(v_i) \text{ si } i = j \\ -1 \text{ si } i \neq j \text{ y } v_i \text{ es adyacente a } v_j \\ 0 \text{ en cualquier otro caso} \end{cases} \qquad \text{(Ecuación 5)}$$

Así, para el caso del grafo que se emplea de ejemplo, la matriz *L* no normalizada quedaría tal y como se presenta en la Tabla 5.

	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

Tabla 5: Matriz Laplaciana $L = (L_{ij})$ del grafo

Esta matriz se puede expresar de manera normalizada si se siguen las siguientes normas al momento de hacer su construcción:

$$L_{ij}^{*} = \begin{cases} 1 \text{ si } i = j \text{ y } L_{ii} \neq 0 \\ -\dfrac{1}{\sqrt{L_{ii} L_{jj}}} \; i \neq j \text{ y } v_i \text{ es adyacente a } v_j \\ 0 \text{ en cualquier otro caso} \end{cases} \qquad \text{(Ecuación 6)}$$

II.1.3 Caminos más Cortos en Grafos

El problema del camino más corto (CMC) en la teoría de grafos forma parte del problema de la teoría de flujo de coste mínimo. Este concepto tiene una serie de aplicaciones importantes, tales como la distribución de productos desde fábricas hasta sus respectivas casas comerciales, y el flujo de materias primas a través de líneas de producción, entre otros (Ravindrak *et al*. 1993; More, 1959).

Dado un grafo $G = (V, E)$ dirigido, definido por un conjunto de nodos V y un conjunto de enlaces dirigidos E_{ij} caracterizados por: un costo asociado C_{ij} que denota el coste por unidad de flujo en el enlace y que varía linealmente con la cantidad de caudal que pasa a través de E_{ij} ; una capacidad U_{ij} que denota la

máxima cantidad que puede fluir por un enlace; una cantidad mínima l_{ij} que debe fluir a través del enlace. A cada $v_i \in V$ se le asigna un valor entero $b(i)$ que permite reconocer el tipo de nodo: $b(i) > 0$ indica que el nodo es de abastecimiento; $b(i) < 0$ indica que el nodo es de demanda, y $b(i) = 0$ indica que el nodo es un nodo sólo de transporte. El flujo que circula por E_{ij} se denota como x_{ij}. Las variables de decisión en el problema son los caudales en los enlaces. El problema de CMC entre dos vértices dados se puede formular según la Ecuación 7, donde *P* es un camino arbitrario entre ambos vértices (Ahuja *et al*., 1993)

$$\min_{P} \sum_{(i,j)\in P} C_{ij}\, x_{ij} \qquad \text{(Ecuación 7)}$$

sujeto a:

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

$$I_{ij} \leq x_{ij} \leq u_{ij}, \forall i, j$$

Además, se tiene que $\sum_{i=1}^{n} b(i) = 0$.

La primera restricción es conocida como restricción de balance de masa, la cual establece que el flujo de salida menos el flujo de entrada debe igualar a la demanda del nodo. Si el nodo es un nodo de abastecimiento, su flujo de salida excede su flujo de entrada; si el nodo es un nodo de demanda, su flujo de entrada excede su flujo de salida; y si el nudo es un nodo de transporte, su flujo de salida es igual a su nudo de entrada. El flujo debe satisfacer la restricción (segunda restricción) de límite inferior y de capacidad (definida como restricción de límite de flujo). El límite de flujo se utiliza típicamente para modelar capacidades físicas o restricciones impuestas sobre los rangos operativos de caudal.

El problema del CMC es una simplificación del problema arriba descrito. En este se envía una unidad desde el nodo fuente hasta el nodo objetivo o sumidero. Dado uno nodo fuente, $b(i) = 1$, y un nodo sumidero, $b(j) = -1$, y $b(k) = 0$ para el resto de nodos del grafo, la solución del problema debe mandar una unidad desde el primero hasta el segundo a través del camino de coste mínimo. Para darle solución a este problema se han diseñado muchos algoritmos, destacándose los algoritmos BFS, DFS, PRIM y Dijkstra entre los más populares (Cormen *et al*., 2001). A continuación se hace una descripción del funcionamiento de cada uno de ellos.

II.1.3.1 Algoritmo de Búsqueda en Amplitud/Anchura

Este algoritmo fue creado por E. F. More en el año 1950 y posteriormente publicado por Lee (1961). Es uno de los algoritmos más simples para realizar exploraciones en grafos y sirve de engranaje para otros algoritmos de gran relevancia, siendo este el caso del algoritmo PRIM para la búsqueda del árbol de expansión mínima y del algoritmo Dijkstra para la búsqueda de caminos más cortos en grafos que cuentan con una única fuente. Dado un grafo $G = (V, E)$ y un nodo raíz (de ahora adelante denotado por R) el algoritmo BFS (*breath first search*) explora sistemáticamente todos los enlaces del grafo para "descubrir" cada uno de los nodos alcanzables desde R. El algoritmo computa las distancias (menor número de enlaces) desde el nodo R hasta cada uno de los nodos alcanzables. El resultado de este proceso es un árbol de amplitud (*breath-first tree*) que contiene todos los nodos alcanzables. El nombre del algoritmo se atribuye al hecho de que el mismo amplía la frontera entre nodos descubiertos y nodos no descubiertos de manera uniforme a través de la amplitud de la frontera (ver Ilustración 8). En otras palabras, el algoritmo descubre todos los vértices a distancia k desde el nodo R antes de descubrir cualquiera de los vértices a distancia $k - 1$.

En Pseudocódigo 1 describe el algoritmo BFS.

```
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 una fila Q
Introducir R en Q
Mientras Q no esté vacía {
Se extrae el primer nodo de la fila v' y se exploran todos
sus nodos adyacentes v''

Para cada v'' si estado == No visitado

{
Estado == Visitado
      Distancia = distancia v'  + 1
      Padre = v'
      Se introduce v'' en Q
}
}
Salida: Árbol de expansión del grafo
```

Pseudocódigo 1: Algoritmo BFS

El llenado y vaciado de Q en el pseudocódigo anterior sigue el principio FIFO (*first in first out*) lo que quiere decir que el primer nodo (R) que es introducido en la fila es el primero en salir.

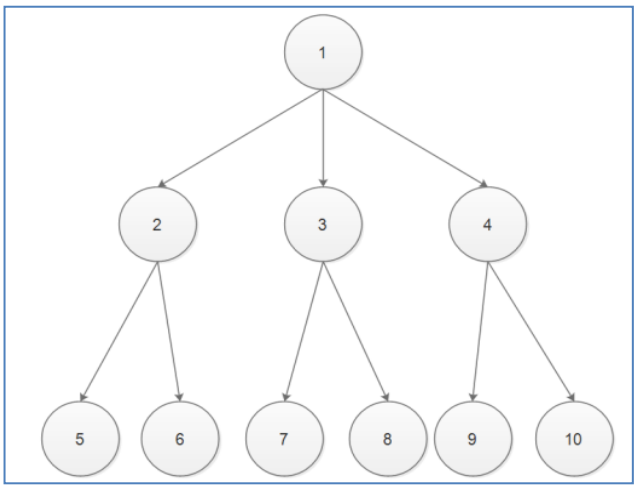


Ilustración 8: Orden de visita de los nodos mediante BFS en un grafo genérico

II.1.3.2 Algoritmo de Búsqueda en Profundidad

La estrategia que sigue el algoritmo DFS (*depth first search*), tal y como lo establece su nombre, consiste en realizar una búsqueda más profunda cada vez que sea posible (ver Ilustración 9). Este algoritmo explora los enlaces de los nudos más recientemente descubiertos que aún tienen enlaces no explorados. Una vez que todos los enlaces han sido explorados, la búsqueda retrocede para explorar los enlaces que se conectan al nodo desde donde v fue descubierto. Este proceso continua hasta que todos los vértices alcanzables desde el nodo R son descubiertos. Si algún nodo se mantiene sin ser descubierto, la búsqueda en profundidad lo selecciona como fuente R y repite el proceso de búsqueda. Este proceso continúa hasta que todos los vértices hayan sido descubiertos.

El Pseudocódigo 2 describe el algoritmo DFS.

```
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
```

Pseudocódigo 2: Algoritmo de búsqueda DFS

En este caso ya no se habla de Q si no de S. En este tipo de estructura la dinámica de llenado/vaciado es LIFO (*last in first out*), de manera que el último nodo en salir del saco es el primer nodo en entrar, que naturalmente corresponde a R.

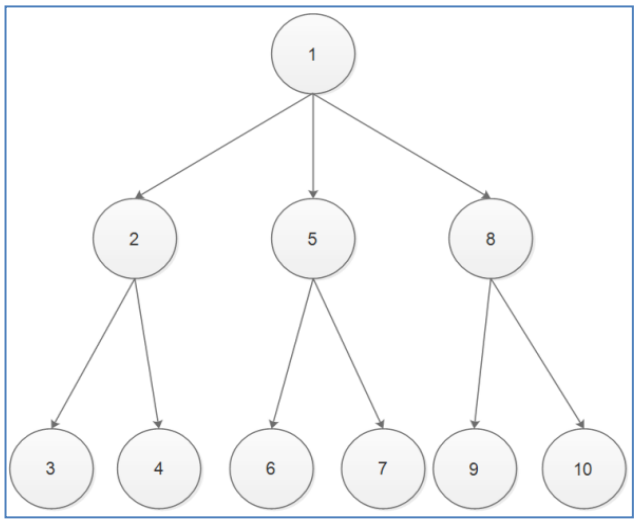


Ilustración 9: Orden de visita de los nodos mediante DFS sobre un grafo genérico

II.1.3.3 Algoritmo Dijkstra

El algoritmo Dijkstra (Dijkstra, 1959) está pensado para encontrar el CMC entre dos nodos por orden creciente de longitud o peso de los enlaces. En cada paso el algoritmo guarda los nodos para los que ya se sabe el camino mínimo y devuelve un vector indexado por vértice, de modo que para uno de estos vértices se puede determinar el coste de un camino más económico (de peso mínimo) desde el nodo inicial a tales vértices. Cada vez que se incorpora un nuevo nodo a la solución, se comprueba si los caminos, todavía no definitivos, se pueden acortar pasando por él. Se caracteriza por ser un algoritmo goloso (en inglés se usa el término *greedy*) y sólo funciona con pesos positivos. Nótese que este tipo de restricción es posible en muchas redes de contexto real (transporte por ejemplo), donde los enlaces representan distancias o tiempos medios (siempre con valores positivos).

```
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.
```

Pseudocódigo 3: Algoritmo Dijkstra

Nótese que en algunos casos se puede dar más de un CMC entre dos pares de nodos.

II.1.3.4 Algoritmo PRIM

Este algoritmo corresponde a un caso especial del método de árbol de mínima expansión genérico. Opera de una manera similar al algoritmo Dijkstra. Tiene como propiedad el hecho de que el conjunto de enlaces siempre forma un árbol individual. El árbol empieza desde un nodo R seleccionado al azar y se expande hasta que el árbol recubre todos los vértices en el conjunto V. En cada paso, adiciona al árbol E un enlace que conecta E a un nudo aislado – uno en el cual ningún enlace de E es incidente. Este algoritmo es de tipo *goloso*, dado que a cada paso adiciona al árbol un enlace que contribuye con la mínima cantidad posible al peso del árbol. A fin de implementar este algoritmo de manera eficiente, se necesita una manera rápida para seleccionar un nuevo enlace para adicionar al árbol conformado por los enlaces en E. Tal y como se muestra en el Pseudocódigo 4, el grafo conectado y el nodo R del árbol de mínima expansión son los inputs necesarios para hacer expandir el árbol. Durante la ejecución del algoritmo, todos los vértices que no están en el árbol se almacenan en Q, que se basa en un atributo clave. Para cada vértice v, el atributo clave es el peso mínimo de cualquier enlace que conecte con un vértice del árbol.

```
1.    Para cada u ϵ V [G]
2.     Valor [u] ← inf; Π[u] ← NULL
3.    Valor [R] ← 0; Q ← V[G];
Mientras !Q {
                u ← Extraer − MinQ
Para cada v ∈ Adj[u] {
Si v ∈ Q   y w(u, v) < valor[v]
            π[v] ← u; valor[v] ← w(u, v)
}
}
                    π[v] ← u
```

Pseudocódigo 4: Algoritmo PRIM

II.1.4 Teoría de Formación de Clústeres

Un clúster se define como un conglomerado de objetos que comparten características entre sí, o dicho de otra manera, los perfiles de los objetos en un mismo grupo son muy similares entre sí, pero son muy disimilares a los de los objetos pertenecientes a otros clústeres (Karlson, 2008; Kaufman & Rosseeuw, 1990; Mooi & Sardtedt, 2011; Romesburg, 2004). Un análisis de clústeres se define como la partición de las observaciones en grupos de manera que las disimilitudes por parejas (medida de cuán diferentes son dos elementos) entre los elementos asignados a un clúster sean menores con respecto a elementos pertenecientes a otros clústeres (Eisen *et al*., 1998; Hastie *et al*., 2009). Los seres humanos nacen con una capacidad innata para formar clústeres, estableciendo categorías para todos los elementos que les rodean y luego ubicando dentro de cada una de esas categorías cada nuevo elemento que se visualice (clasificación). La técnica de clústering es una de las técnicas más utilizadas para análisis y exploración de datos. Esta técnica tiene aplicaciones en estadística, ciencias de la computación, biología, ciencias sociales y psicología.

El análisis de clúster puede ser clasificado de acuerdo al resultado que el mismo genera. Una primera clasificación crea, por un lado, una jerarquía de clústeres (*hierarchical clustering*) y, por otro lado, una partición en clústeres (*partitioning clustering*) (Arabie & Hubert, 1996; Sander, 1999). En el primer caso, los clústeres son formados por anidación o por des-anidado de los elementos. En el caso de des-anidación, se parte de un clúster global que contiene todas las observaciones en un sólo clúster y en función del algoritmo de desagrupamiento que se emplea, se van formando subclústeres hasta llegar a un número de clústeres igual al número de elementos (cada elemento constituye un *singleton*). En el caso de anidación sucede exactamente lo contrario, es decir, desde los *singleton* se llega a un clúster global que agrupa todos los elementos. El segundo caso (partición de clústeres), es una partición simple del conjunto de datos en subconjuntos disjuntos, de tal manera que cada elemento se encuentre en uno u otro subconjunto.

Otra clasificación del análisis de clústeres genera como resultado clústeres exclusivos, o clústeres en los cuales cada clúster tiene su subconjunto exclusivo de elementos, que no se repiten en otro(s) clúster(es); clústeres no disjuntos, que corresponde al caso en que un mismo elemento (o varios) puede existir en diferentes clústeres al mismo tiempo; clúster borroso (*fuzzy cluster*) en el que los elementos no pertenecen *per se* a un determinado clúster, sino que su pertenencia al conjunto de clústeres está asociada a un peso, siendo 1 el peso que indica que el elemento pertenece completamente a un clúster dado y 0 el peso que indica que el elemento no tiene ninguna relación de pertenencia con un clúster dado (Abonyi & Balázs, 2007). La última clasificación de análisis de clúster de la que se hará mención puede generar clústeres completos o clústeres parciales. En el caso de los clústeres completos, todos los elementos se agrupan dentro de algún subgrupo, en tanto en el caso de la partición parcial, ciertos elementos no pueden ser ubicados en ninguno de los subconjuntos resultantes. Estos elementos corresponden generalmente al ruido (*outliers*) del conjunto de datos.

También existe una clasificación de los clústeres en función de las metas que persiga el análisis de clúster. Así, estos pueden ser: bien separados, basados en prototipo, basados en densidad, basados en grafos y en propiedades compartidas. Los clústeres bien separados siguen una conceptualización idealista del término clúster, es decir, los clústeres son formados por elementos con un grado de similitud muy fuerte entre sí y se encuentran lejos de otros elementos. En ciertos casos se emplean umbrales mínimos para definir si una similitud es suficiente para establecer el agrupamiento o no. Estos tipos de clústeres también son conocidos como clústeres naturales. Los clústeres basados en prototipos, o también conocidos como clústeres basados en el centro (*centered-based cluster*) corresponden a clústeres que se forman siguiendo un prototipo, como lo puede ser un centroide o un medoide, a partir del cual los elementos se agrupan (Ding *et al*., 2008). Los clústeres basados en grafos se forman a partir de nodos que se encuentran conectados entre sí a través de links (conexiones) o aristas. En este caso, el agrupamiento se centra en la conexión entre los elementos. Así, los elementos (o nodos) con un determinado grado de similaridad se encuentran conectados y los

elementos (o nodos) con un nivel de disimilaridad se encuentran desconectados. En el caso de los clústeres basados en densidad, el agrupamiento se da por regiones de mayores densidades de elementos, siendo las separaciones entre dos clústeres zonas de baja densidad de elementos (Sander, 1999). Los clústeres basados en propiedades compartidas van un paso más allá de todos los tipos de clústeres mencionados previamente, e incluyen clústeres que pueden estar enlazados entre sí mediante ciertos elementos en común.

Típicamente, en el caso del clústering jerárquico, las fusiones sucesivas se representan mediante un diagrama llamado dendograma (ver Ilustración 10). En el eje x se colocan los casos o elementos de estudio, en tanto que en el eje y se colocan los valores de disimilaridad calculados. La fusión o segregación de dos elementos se produce para un valor de disimilaridad dado que se representa mediante una línea horizontal, la cual recibe el nombre de hoja o clada. En la cima del dendograma se encuentra la clada que abarca todas las fusiones sucesivas que se produjeron durante la ejecución del algoritmo. Tal y cómo se abordará más adelante, en los algoritmos de detección de comunidades empleados en este trabajo se generan jerarquías de comunidades, las cuales son representadas a través de este tipo de estructura gráfica.

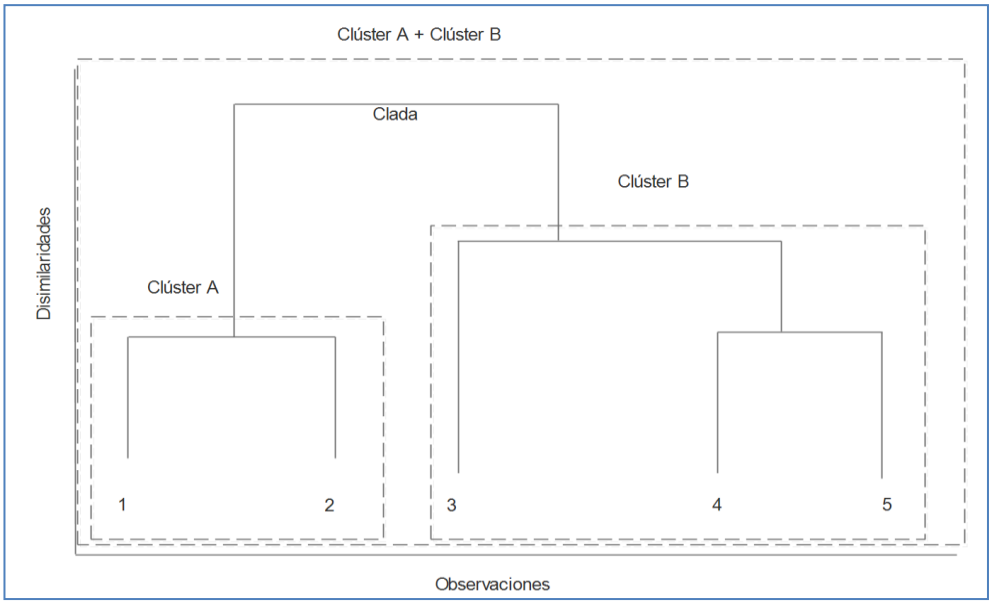


Ilustración 10: Estructura de un dendograma (clústering jerárquico)

II.1.4.1 Calidad de clústeres/Número de clústeres: Métodos de Evaluación

Uno de los principales problemas con respecto a la implementación de las técnicas de detección de clústeres es la determinación del número de clústeres que es más apropiado para un problema dado. Esto es particularmente importante para la técnica de clústering jerárquico, en el que no sólo se genera una sola partición sino una jerarquía de particiones. En el caso de las técnicas de clústering no jerárquico, la pregunta es en cuántos clústeres se debe hacer la partición, cuando este número no está a priori definido. Existe una serie de medidas que permiten abordar este problema. A continuación se hace una descripción de algunas de ellas. Primero se presentan un conjunto de medidas agrupadas en dos grupos: internas y de estabilidad. Luego se discute sobre un criterio gráfico conocido como es el criterio del codo, cuya estimación parte de graficar la escala de altura de un dendograma y el número de clústeres relativo a cada valor de dicha escala. A continuación se describe el cálculo del valor de inconsistencia, que es característico de cada clada en el dendograma, y que se puede combinar con el criterio del codo para establecer un número de clústeres distintivos. Se finaliza la discusión con un método más complejo propuesto por (Shimodaira, 2004), que contempla un remuestreo interno multiescala (*Multiscale Boostrap Resampling*), el cual arroja un indicador de Sesgo Aproximado (AU por *Approximately Unbiased*) mediante el cual también se puede estimar el número de clústeres distintivos dentro de un dendograma.

- **Medidas Internas**

Las medidas internas son una serie de parámetros (conectividad, Ancho de Silueta – *AS* - e índice de Dunn) que se calculan a partir de tres características generales del conjunto de clústeres en una partición dada: compactación, conectividad y separación. La compactación, por un lado, evalúa la homogeneidad de los clústeres, mientras la separación, por el contrario, como indica su nombre, evalúa el nivel de separación. Por otra parte, la conectividad se centra en la justificación del porqué algunos nodos pertenecen a un mismo clúster y otros no (Brock *et al*., 2008). Las características de compactación y separación son antagónicas, dado que cuantos más clústeres arroje una partición, mayor es el grado de compactación y menor la

separación; por otro lado, cuanto menos clústeres se produzcan mayor será la separación y menor la compactación. Es por esta razón que el cálculo de los parámetros que se utilizan para evaluar las particiones emplea una combinación de estas características.

La conectividad es un parámetro que mide la relación global de todos los casos pertenecientes a un clúster mediante comparación de los mismos con los casos más cercanos pertenecientes a otros clústeres (Handl *et al*., 2005). Se mide mediante la Ecuación 8; en esta expresión, cada comparación $\chi_{i,nn_{i(j)}}$ entre elementos (*i*-*j*) que están en el mismo clúster arrojará un valor igual a cero, en tanto que, la comparación entre un caso *i* y un vecino *j* seleccionado arrojará un valor igual a 1/*j*. La cantidad de vecinos *j* a emplear se selecciona mediante el parámetro *L*. El valor de la conectividad resultante puede tomar valores entre 0 e infinito, denotando los valores más cercanos a 0 un mayor nivel de conectividad.

$$Conn(\sigma) = \sum_{i=1}^{N} \sum_{j=1}^{L} \chi_{i,nn_{i(j)}} \qquad \text{(Ecuación 8)}$$

El parámetro *AS* fue propuesto por Rouseeuw (1987). Para el cálculo de este parámetro se combinan tanto la característica de cohesión como la de separación. El resultado es un valor final para cada clúster, que representa el promedio de valores *AS* de cada conjunto de casos dentro de un clúster. En el caso de los clústeres mejor conformados, el valor del *AS* será próximo o igual a 1 y en el caso de los peores conformados, el valor será próximo o igual a -1. Su cálculo se realiza mediante la Ecuación 9.

$$AS(i) = \frac{b_i - a_i}{\max(b_i, a_i)} \qquad \text{(Ecuación 9)}$$

En la Ilustración 11 se muestra la representación típica de la partición en clústeres de una base de datos. En este caso se trata de una partición en tres clústeres.

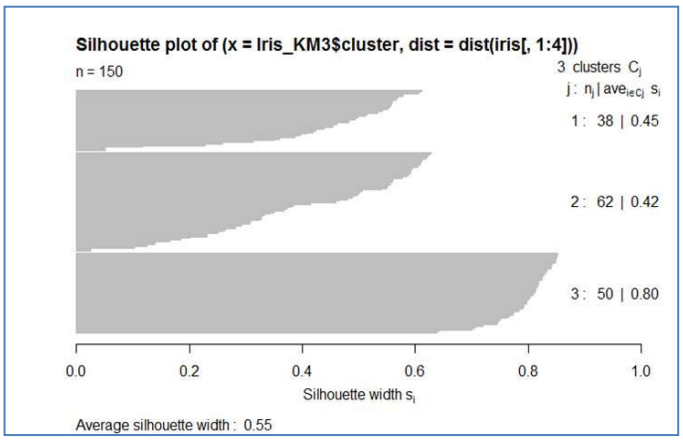


Ilustración 11: Representación típica del ancho de silueta de una partición de clústeres

Esta medida fue empleada por Herrera (2011) para evaluar sectores en RDAPs definidos mediante Clústering Espectral.

El índice de Dunn (Dunn, 1974) es el resultado que se obtiene al dividir la mínima distancia extra-clúster entre dos casos entre la máxima distancia intra-clúster, o lo que se llama diámetro de un clúster. Se calcula mediante la Ecuación 11.

$$D(C') = \frac{\min\limits_{C_k, C_l \in C', C_k \neq C_l} \left(\min\limits_{i \in C_k, j \in C_l} dist(i,j) \right)}{\max\limits_{C_m \in C'} diam\,(C_m)} \qquad \text{(Ecuación 11)}$$

Aquí $diam(C_m)$ es la distancia máxima entre observaciones en el clúster C_m. Dado que las distancias intra y extra clúster pueden ser infinitas, este índice puede tomar valores entre 0 y +∞, correspondiendo la mejor partición a la que se obtiene cuando este indicador se maximiza.

- **Medidas de Estabilidad**

Estas medidas comparan los resultados del clústering a partir de la base de datos completa menos un vector columna. Para ello se remueve una columna por vez. Esta categoría incluye las siguientes medidas: la Porción Promedio de no Traslape (*APN* por *Average Portion of non Overlap*), la Distancia Media (*AD* por *Average*

Distance), la Distancia Promedio entre Medias (*ADM* por *Average Distance Between Means*), o la Figura de Mérito (*FOM* por *Figure of Merit*). En cada uno de los casos, se persigue la minimización de los valores. El cálculo de las cuatro medidas sigue la misma estrategia, a saber, se hace el clústering en una cantidad dada de clústeres y luego se repite pero con una columna menos en la base de datos.

La primera de las medidas (*APN*), mide la porción de las observaciones posicionadas en el mismo clúster cuando este se hace con toda la base de datos o cuando el clúster se hace con una columna menos (Ver Ecuación 12).

$$APN\ (K) = \frac{1}{MN}\sum_{i=1}^{N}\sum_{l=1}^{M}\left(1 - \frac{n\left(C^{i,l} \cap C^{i,0}\right)}{n(C^{i,0})}\right) \qquad \text{(Ecuación 12)}$$

Aquí $C^{i,0}$ representa el conjunto de las observaciones en el clúster con toda la base de datos, $C^{i,l}$ las observaciones cuando el clúster se hace sin la columna l, M correspone al número total de columnas de la base datos, N corresponde al número total de observaciones, y K al número de clústeres en que se hace la partición.

La segunda medida (*AD*) compara la distancia entre elementos en el mismo clúster en los dos escenarios (ver Ecuación 13).

$$AD\ (K) = \frac{1}{MN}\sum_{i=1}^{N}\sum_{l=1}^{M}\frac{1}{n(C^{i,0})n(C^{i,l})}\left[\sum_{i \in C^{i,0}, j \in C^{i,l}} dist(i,j)\right] \qquad \text{(Ecuación 13)}$$

La tercera medida (*ADM*) compara la distancia de las observaciones al centro de su respectivo clúster (ver Ecuación 14).

$$ADM\ (K) = \frac{1}{MN}\sum_{i=1}^{N}\sum_{l=1}^{M} dist(\overline{x}_{C^{i,l}}, \overline{x}_{C^{i,0}}) \qquad \text{(Ecuación 14)}$$

Aquí $\overline{x}_{C^{i,0}}$ es el valor medio de la observación al centro del clúster cuando se tiene en cuenta toda la base de datos y $\overline{x}_{C^{i,l}}$ es la media de observación cuando el clústering se realiza sin una de las columnas (ver Ecuación 15).

$$DPM\ (l,K) = \sqrt{\frac{1}{N}\sum_{k=1}^{K}\sum_{i\in C_k(l)} dist(x_{i,l},\overline{x}_{C_{k(l)}})} \qquad \text{(Ecuación 15)}$$

- ***ClValid***

El lenguaje de programación estadística *R* (R Core Team, 2015) cuenta con un paquete que permite hacer evaluaciones del número de clústeres usando los dos tipos de medidas anteriormente descritas para distintas cantidades de clústeres. Una de las capacidades más interesantes de *ClValid* (Brock *et al*. 2008), es la posibilidad de emplear distintas técnicas de clústering: Jerárquico, *k-means*, DIANA, SOM, PAM, CLARA, AGNES, para evaluar en qué partición se obtiene el mejor valor de los parámetros anteriores. A continuación se presenta una descripción somera de cada uno de estos métodos, excepto de los métodos Jerárquico y SOM, dado que estos últimos son ampliamente descritos en secciones sucesivas (Subsección II.6 y Apéndice I, respectivamente):

- ***k-means*:** Esta es una técnica basada en prototipo. Primero se asigna un número *k* de centroides, donde *k* corresponde al número de clústeres deseados. Luego cada punto se asigna a un centroide, y la colección de puntos asignados a cada centroide es un clúster. A continuación se reasigna el centroide de cada clúster con base en la reasignación anterior. La reasignación se repite hasta que se alcance una estabilidad en la posición de los centroides.
- **DIANA:** Se construye una jerarquía de clústeres, empezando con un clúster que contiene las n observaciones. Los clústeres son entonces divididos hasta que cada uno de ellos contenga una única observación. En cada paso se selecciona el clúster con el diámetro más

grande (el diámetro de un clúster es la disimilaridad más grande entre todas las parejas de observaciones). Para el clúster seleccionado, el algoritmo empieza por buscar las observaciones más dispares (las que tiene el valor de disimilaridad promedio mayor con respecto a las otras observaciones del clúster seleccionado). Esta observación inicia el grupo separador o *splinter group*. En los pasos subsecuentes, el algoritmo reasigna las observaciones que están más cerca del grupo separador. El resultado es la división del clúster seleccionado en dos nuevos clústeres. Luego, el proceso se repite para cada nuevo clúster.

- **PAM:** Muy similar a *k-means,* este algoritmo divide la base de datos en grupos (clústeres), y ambos funcionan tratando de minimizar la disimilaridad entre elementos en un mismo clúster, pero PAM funciona con medoides, que son identidades del conjunto de datos que representan el grupo en que esta es insertada, mientras *k-means* opera con centroides. El algoritmo funciona en dos fases: en la primera fase, se construye una colección de objetos que es definida como un conjunto inicial, y en la segunda fase se trata de mejorar la calidad de los clústeres mediante el intercambio entre objetos seleccionados con objetos no seleccionados.

CLARA: Este algoritmo, propuesto por Kaufman & Rousseeuw (1990) fue diseñado para abordar aplicaciones con grandes cantidades de datos. CLARA extiende la aproximación propuesta por *k*-medoides a conjuntos conformados por un número considerablemente grande de objetos. Funciona mediante la clusterización de una muestra de la base de datos y luego asignando los elementos restantes a los clústeres creados.

- **Otras Medidas**

Una de las maneras más simples para poder estimar un número de clústeres que represente una partición óptima de una red consiste en hacer una gráfica en la que uno de los ejes (por lo general el eje-*x*) contenga una escala en la que estén representados todos los posibles valores de altura (*S´*) que puedan tener todas las *cladas* del *dendograma* (o porcentaje acumulado de este valor) y el otro eje

contenga el número de clústeres posibles. En esta gráfica se suele producir un codo (cambio de pendiente). Más allá de este codo, cualquier nueva fusión se produce a una distancia mucho más pequeña, de manera que el número de clústeres antes de esta fusión es la solución más probable (Mooi & Sardtedt, 2011); dicho de otra manera, se debe seleccionar el número de clústeres a partir del cual, agregar un nuevo clúster no provea una mejor modelación de los datos. Pese a la facilidad que representa este método, presenta como desventaja el hecho de que no en todos los casos aparece un codo.

En la Ilustración 12 se puede ver un ejemplo de gráfica que permite este análisis. La misma muestra el porcentaje acumulado del total del valor de altura, asociado a cada una de las particiones en la medida en que el número de clústeres va disminuyendo (hasta llegar a un único clúster en donde se acumula el 100% de la altura). También se ve la variación de esta curva (línea azul), llegando un punto en que se forma un codo, para una partición de tres clústeres.

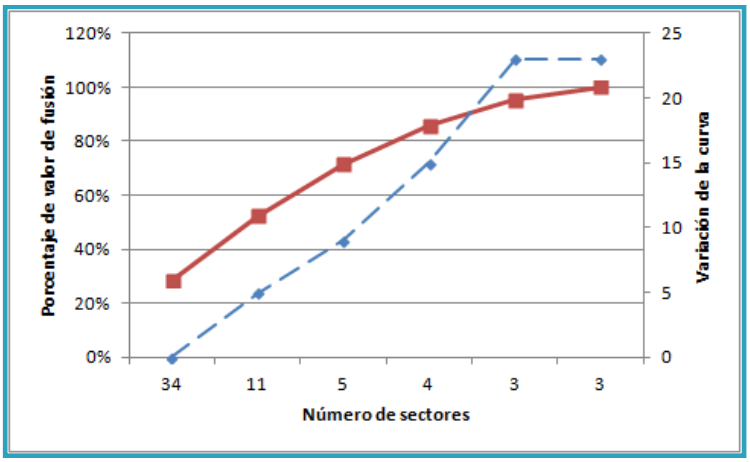


Ilustración 12: Selección del número de clústeres con base en el criterio del codo [Fuente: (Campbell, 2013a)]

Otra alternativa la representa el criterio de inconsistencia, el cual parte de etiquetar cada una de las cladas del dendograma con un valor denominado de inconsistencia, que sirve como indicador de similitud entre clústeres. Cuanta más similaridad tengan dos clústeres entre sí, menor será el valor de inconsistencia de la clada que los une. Fijando un umbral de tolerancia de inconsistencia, se puede definir que los

clústeres más representativos de la partición son aquellos que se ubican por debajo del umbral definido (Ghahramani, 2004). La aplicación de este método implica un esfuerzo de cálculo un poco mayor que el método anterior; no obstante, tiene la ventaja de aportar un poco más de precisión. La desventaja radica en la subjetividad implícita en la selección del umbral apropiado. Una manera para hacerlo es combinar este método con el método anterior, creando una gráfica en la que en lugar de la escala de alturas de enlace, se coloque la escala de valores de inconsistencia, y definir como número de clústeres apropiado aquel indicado por el codo de la gráfica.

Para calcular la inconsistencia de un clúster $\varphi_{k'}$, se parte de una matriz ultramétrica (ver ejemplo en Tabla 10, más adelante). De ella se extrae el valor de altura con el que se formó cada nuevo clúster S'.

A continuación se estima la media $\overline{S'_{(k')}}$ y la desviación estándar $\mu_{k'}$ de todos los valores de altura $S'_{(k')}$ de las cladas que constituyen cada clúster conformado al menos por tres *singleton.*

$\varphi_{k'}$ es la porción que representa la resta de cada valor de altura menos la media de los valores de las cladas que se encuentran bajo esta (incluyéndola) sobre la desviación estándar correspondiente. En la Ilustración 13 se muestra un ejemplo de dendograma con los valores de inconsistencia calculados para cada uno de sus clústeres.

$$\varphi_{k'} = \frac{S'_{(k')} - \overline{S'_{(k')}}}{\mu_{k'}} \qquad \text{(Ecuación 16)}$$

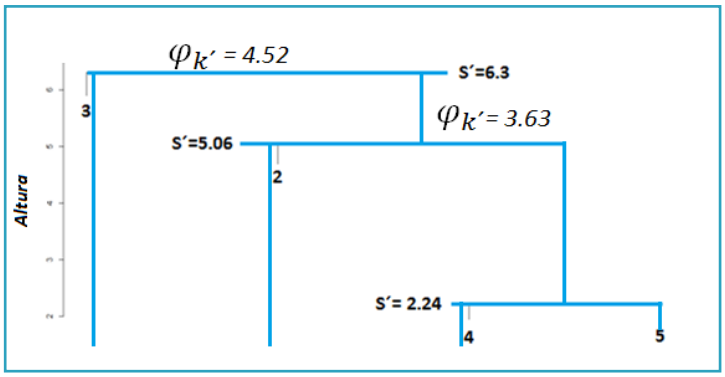


Ilustración 13: Valores de inconsistencia correspondiente a cada clada del dendograma

Finalmente, describimos la herramienta *pv-clust* para re-muestreo interno multi-escala. *pv-clust* es un paquete, creado para el programa *R*, que permite validar particiones hechas mediante clústering jerárquico representadas en un dendograma (Suzuki & Shimodaira, 2006). La validación se hace mediante dos valores probabilísticos o*p-values*, *AU* (*Aproximattely Unbiased*) y *BP* (*Bootstrap Probability*). Ambos se calculan mediante un proceso iterativo que implica un remuestreo que puede ir de 1000 hasta 10,000 (o las que se deseen) repeticiones, con la diferencia de que en el primero se varía la escala (tamaño) de la muestra (remuestreo multiescala), en tanto que en el otro caso, el tamaño de esta se mantiene constante. Ambos *p-values* terminan siendo aproximaciones; no obstante, la variación de la escala implicada en el cálculo del *p-value AU*, hace que corresponda a una aproximación menos sesgada que el *p-value BP*. El uso de ambos valores ayuda a estimar el nivel en que los clústeres están respaldados por los datos (Shimodaira, 2004). Estos pueden tomar valores entre 0 y 1.

La idea del método del remuestreo multiescala fue abordada inicialmente por Efron *et al*. (1996). En este caso, se toman muchas muestras de un conjunto de datos (remuestreo); se forma un dendograma; y se hace el análisis de clúster en cada una de estas réplicas. De esta manera se puede calcular la probabilidad de aparición de un clúster que será expresada mediante el *p-value BP*. Luego, para calcular el valor del *p-value AU*, se utiliza el mismo proceso, pero ahora se altera el tamaño de la muestra. El *p-value AU* se calcula mediante la observación en el cambio en la frecuencia a lo largo del cambio del tamaño de la muestra. Gráficamente, la validación se observa mediante rectángulos sobre el dendograma de la partición. Un

clúster es considerado cómo válido cuando está encerrado en un rectángulo, y a su vez está determinado por el *p-value AU* de la subclada superior. Las cladas que tengan mayor altura y cuyo *p-value AU* sea superior a 0.95, corresponden a las cladas por debajo de las cuales se encuentran los clústeres válidos. En la Ilustración 14, se muestra cómo se pueden identificar cuatro clústeres en una red con una sola fuente de abastecimiento, usando la herramienta *pv-clust*.

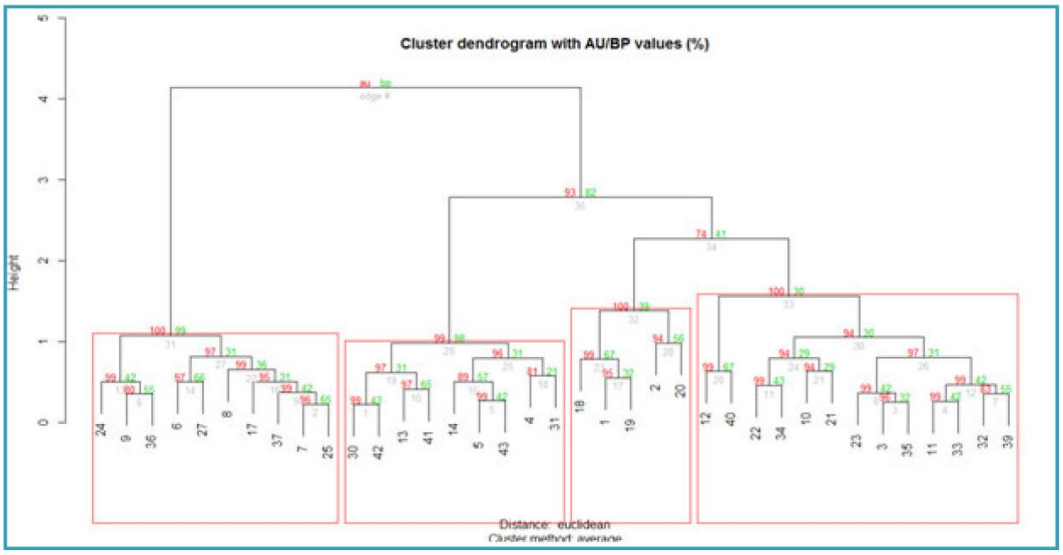


Ilustración 14: Identificación de clústeres "válidos" basada en los *p-values* en un dendograma [Fuente: (Campbell, 2013a)]

- **Conclusiones sobre los métodos de evaluación**

La Tabla 6 presenta una comparación de todos los métodos de evaluación de la calidad de particiones anteriormente expuestos. Las medidas internas y de estabilidad permiten realizar una evaluación más completa y profunda de la calidad de las particiones, en comparación a medidas tales como el índice de inconsistencia o el criterio del codo, las cuales, adicionalmente, implican un esfuerzo de cálculo mayor que las medidas anteriores sin garantizar al final un buen resultado. Los *p-values*, combinan, de alguna manera, las ventajas implícitas en las medidas internas y las medidas de estabilidad; sin embargo, la resolución del método en una red de pocos nodos implica una gran cantidad de tiempo. Teniendo esto en cuenta, en este

trabajo se procede a seleccionar las medidas internas y de estabilidad para definir el número de clústeres en que se hace la partición.

Vale la pena destacar que todas las medidas presentadas se centran directamente en la estructura del grafo y no en las implicaciones hidráulicas de la red, con lo cual no se puede esperar que las mismas tengan la capacidad para establecer un buen número de sectores en los cuales se debe subdividir la misma y por el contrario, la selección del número de clústeres debe ir acompañada de un buen análisis ingenieril.

Tabla 6: Ventajas y desventajas de la sectorización

Medida	Ventaja	Desventaja
Medidas Internas		
Conectividad	Rápido y sencillo	Medidas de evaluación interna que pueden ser distintas entre sí
Ancho de silueta	Rápido y sencillo	
Índice de Dunn	Rápido y sencillo	
Medidas de Estabilidad		
Porción promedio de no traslape PPNT	Rápido y sencillo	Medidas de evaluación interna que pueden ser distintas entre sí
Distancia media (DM)	Rápido y sencillo	
Promedio de Distancia entre medias (DPM)	Rápido y sencillo	
Figura de Merito (FM)	Rápido y sencillo	
Otras Medidas		
Criterio de codo	Rápido y sencillo	Algunas veces no se aprecia un codo claramente en la gráfica.
Inconsistencia	Rápido y sencillo	No hay reglas que establezcan que valor de inconsistencia es el correcto. Si se combina con el criterio del codo a veces pasa que en algunas gráficas no se aprecia el codo.
p-values	Alta precisión	Tiempo de procesamiento extenso

Vale la pena destacar que el índice de modularidad propuesto por Newman *et al.* (2006) es, quizá, el indicador más importante cuando se aborda el problema de

clústering en redes. Sin embargo, dada la relación que tiene el mismo con la teoría de redes sociales, su descripción se realiza en la siguiente subsección (Subsección II.2 Grafo de Redes Sociales y Detección de Comunidades).

II.2 Grafos de Redes Sociales y Detección de Comunidades

Las redes sociales virtuales se pueden describir como grafos especializados en describir la interacción de elementos que forman parte de una sociedad y que sostienen algún grado de interdependencia entre sí. En este sentido, un individuo es una entidad que genera un aporte a la sociedad: una persona que se comunica en un grupo, un espécimen o una especie en un ecosistema o un nudo en una red de agua potable. En este caso el aporte puede ser una demanda (ya sea positiva o negativa), o un valor de carga hidráulica. Los sistemas sociales, tecnológicos y de información pueden ser descritos en términos de redes sociales complejas con una topología de nodos interconectados (Easley & Kleinberg, 2010; Fortunato, 2007, 2010; Reichardt & Bornholdt, 2006; Wasserman *et al*., 2005).

Aunque muchos de los conceptos centrales relacionados con las redes sociales ya eran conocidos hace décadas, la rapidez con que los mismos se han ido integrando en los distintos campos académicos ha sido moderada. Es a partir de finales de la década de los 90s, con los trabajos de Barabasi & Albert (1999) y Watts & Strogatz (1998), que se despierta un gran interés en la aplicación de los mismos en diversos campos de estudio. Fue a partir de ese momento que el área de investigación logró una gran expansión debida, en gran parte, al impulso generado por el desarrollo de las nuevas técnicas/herramientas computacionales (Palla *et al*., 2005). En la actualidad existen herramientas de análisis y visualización de redes sociales bastante avanzadas: *Gephi* (Bastian *et al.*, 2009); *Pajek* (Bataglj & Mrvar, 2002); *Graphviz* (Bilgin *et al*., 2017); la librería *Igraph* en *R* (Csardi & Nepuz, 2006), por citar sólo algunas. No obstante, no sólo el desarrollo de herramientas es el responsable del aumento del interés en la temática de redes sociales, sino que gran parte de la responsabilidad recae en la adaptabilidad de la misma a la actual coyuntura social y tecnológica del mundo. Esto ha permitido emplear sus

algoritmos para, entre otras cosas, modelar interacción con el fin de entender procesos evolutivos o antropológicos ya acaecidos o predecir resultados de fenómenos que se viven en la actualidad. Dentro del campo de la ingeniería han tenido menos utilización; sin embargo, esto no quiere decir que su aplicabilidad no tenga validez.

Uno de los aspectos de mayor interés dentro del área de investigación de las redes sociales es el estudio de su topología, lo que permite comprender la organización y la función de los individuos que forman parte de las mismas. Concretamente, uno de los conceptos de mayor interés es el estudio de la detección de comunidades de individuos o clústering en redes sociales. La identificación de estas comunidades es de crucial importancia ya que permite revelar módulos funcionales a priori desconocidos. En las redes sociales algunos individuos pueden ser parte de un grupo altamente conectado o una élite social cerrada, otros pueden estar completamente aislados, mientras que algunos otros pueden actuar como puentes entre grupos. Los subgrupos en las redes sociales muchas veces tienen sus propias normas, orientaciones y subculturas, y el hecho de estar en una comunidad, les brinda una identidad. La idea de comunidades en redes sociales fue inicialmente formalizada y planteada en términos matemáticos en el campo de las ciencias sociales, sin embargo su expansión a otras áreas no tardó mucho en llegar. Así, en la actualidad, es muy frecuente encontrar análisis de comunidades en estudios de diversas temáticas: biología, internet, marketing, etc.

Una comunidad, clúster o subgrupo cohesivo es un subconjunto de individuos entre los cuales hay enlaces relativamente fuertes y, además, la proporción de los enlaces dentro de la comunidad (enlaces internos) es alta en relación a la proporción de enlaces entre distintas comunidades (Steinhaeuser & Chawla, 2010). En términos más simples, dentro de las comunidades la densidad de enlaces es alta y entre ellas baja. La idea de comunidades en una red social representada en un grafo compatibiliza con el concepto de subgrafo anteriormente expuesto (ver Subsección II.1.1). Es decir, una comunidad es un subgrafo; sin embargo, se tiene que tener cuidado con esta analogía, ya que no todos los subgrafos en un grafo de red social

tiene que ser representativos de comunidades con significado estructural. En la literatura de comunidades o subgrupos cohesivos se establecen varias maneras de conceptualizar la idea de subgrupos en redes sociales. En particular, existen cuatros criterios para describir la idea de estructura comunitaria: mutualidad de enlaces, cercanía (*closeness*) o alcanzabilidad (*reachability*), frecuencia de enlaces, frecuencia relativa de enlaces (Hamber *et al*., 2009).

- **Mutualidad de enlaces:** este criterio requiere que todos los pares de miembros de cada subgrupo se escojan entre ellos. Esta idea es formalizada mediante el concepto de cliques, que es un subgrafo máximo completo de tres o más nodos.
- **Cercanía o accesibilidad:** los subgrupos cohesivos basados en accesibilidad requieren que todos los miembros sean accesibles desde los demás. Nótese en este punto la importancia del concepto de componente, que es el subgrafo conectado máximo, por ejemplo, un subgrafo en el cual hay un camino entre cada par de nodos.
- **Frecuencia de enlaces:** referido al subgrafo máximo que contiene n nodos y en el cual cada nodo es adyacente a no menos de n-k nodos en el subgrafo.
- **Frecuencia relativa de enlaces:** esta idea es distinta de las tres anteriormente planteadas porque está basada en la comparación de los enlaces dentro de los subgrupos y los enlaces entre los subgrupos. De esta manera, subgrupos cohesivos se ven como áreas relativamente densas del grafo.

A diferencia del clústering tradicional, la detección de comunidades no se centra únicamente en los individuos (o nodos de la red) y sus características, sino que también tiene en cuenta los enlaces de los individuos con otros individuos, de manera tal, que las comunidades resultantes, no sólo están formadas por individuos sin más, sino por individuos que cuentan con una interacción (conexión) entre sí. Así, en una partición de nodos en un grafo mediante detección de comunidades no se encuentran comunidades subdividas, lo que sí ocurre cuando se aplican otras técnicas de detección de clústeres más centradas en los nodos.

Existen tres tipos diferentes de algoritmos de detección de comunidades (Maimon & Rokach, 2010):

- **Algoritmos divisivos:** detectan links que conectan comunidades, los remueve de la red y evalúan la estructura comunitaria.

- **Algoritmos aglomerativos:** fusionan nodos similares o comunidades, de manera recursiva.

- **Métodos de optimización**: emplean métodos de optimización para la maximización de una función objetivo.

Como es de esperar, cuanto más grande sea una red social, mayor es el número de posibles combinaciones entre individuos que pueden formar una comunidad. Esto hace muy compleja la definición de lo que es una buena partición en comunidades. En el pasado, esto representaba un reto para los investigadores centrados en esta área. Sin embargo, en la actualidad, tras la definición del indicador Modularidad, (Newman, 2006), el problema se considera solucionado. La Modularidad, hoy en día, es el criterio más aceptado para evaluar la calidad de las particiones en las redes sociales.

No obstante, aún existe mucho campo de desarrollo para los algoritmos de detección de comunidades, en especial en lo que a velocidad de procesamiento se refiere, cuando se trata de analizar redes de dimensiones astronómicas tales como *Facebook*, por ejemplo, que cuenta con más de 60 millones de usuarios o Google, que cuenta con varios miles de millones de sitios webs indexados.

II.2.1 Concepto de Modularidad

La Modularidad de una partición es un valor escalar entre -1 y 1 que cuantifica la cantidad de enlaces que caen dentro de una comunidad en comparación a la cantidad de enlace que caerían dentro de la comunidad si los mismos fueran reubicados aleatoriamente. En caso de que el número de enlaces en la comunidad supere el número de enlaces dentro de la comunidad dispuestos aleatoriamente, entonces el valor es positivo; caso contrario, el valor es negativo. La reorganización

aleatoria del grafo se realiza mediante el conocido modelo de configuración (Newman, 2003), de tal manera que los nodos preserven el mismo grado. Para ello, cada enlace en el grafo se parte por la mitad y cada media punta (también conocida como talón) se reconecta con otra media punta o talón de manera aleatoria. Definamos las medias puntas como l; así, para un grafo con una cantidad de enlaces m, la cantidad de talones viene dado por $l_n = \sum_i^n k_i = 2m$. El número de posibles enlaces es igual a la multiplicación de los grados de dos nodos conectados k_i y k_j dividido entre $2m$. Definiendo la adyacencia A_{ij} entre dos nodos como un elemento de la matriz de adyacencia previamente descrita (ver Subsección II.1.2), el índice de modularidad se calcula mediante la Ecuación 17.

$$Q = \sum_{ij} \frac{A_{ij}}{2m} - \left[\frac{k_i * k_j}{(2m)(2m)}\right] \delta\,(c_i, c_i) \qquad \text{(Ecuación 17)}$$

Aquí,$\delta\,(c_i, c_i)$ es, simplemente, un operador que asume el valor de 1 cuando dos nodos se encuentran dentro de la misma comunidad y 0 en caso contrario. Así, el valor de modularidad se computa únicamente entre nodos que pertenecen a la misma comunidad.

Una generalización común de la ecuación anterior viene dada por la Ecuación 18.

$$\sum_{i=1}^{c} (e_{ii} - {a_i}^2), \qquad \text{(Ecuación 18)}$$

en donde c denota comunidad, e_{ii} corresponde a la fracción de enlances que caen dentro de la misma comunidad y se calcula con el término $\sum_j \frac{A_{ij}}{2m}\ \delta\,(c_i, c_i)$, y a_i corresponde a la fracción de enlaces esperados y es calculado por el término $\frac{k_i}{2m}$.

La modularidad ha sido utilizada para comparar la calidad de las particiones obtenidas por diferentes métodos, pero también como función objetivo a optimizar. Desafortunadamente, la optimización exacta de la modularidad es un problema computacionalmente complejo, de tal manera que se hace necesario la utilización de

algoritmos de aproximación cuando se lidia con redes de gran extensión (de millones de nodos). Uno de los primeros algoritmos para tal fin fue el propuesto por Clauset *et al*. (2004), en el que se optimiza el valor de modularidad mediante una función recurrente de comunidades. Pese a la aceptación y el uso masivo de este indicador, varios estudios han destacado un problema de resolución en el mismo. El origen de este problema proviene del hecho de que la modularidad es la suma de términos, donde cada término corresponde a una comunidad. Por ende, encontrar el máximo valor de modularidad equivale a un compromiso entre el número de términos en la suma y el valor de cada término. Aumentar el número de módulos no necesariamente se traduce en un aumento de la modularidad, ya que esto implica una reducción del tamaño y, por ende, del valor de cada término. El problema es que esta partición "óptima" desde el punto de vista matemático, no necesariamente captura la estructura comunitaria de las redes, donde las comunidades pueden ser muy heterogéneas en tamaño, en especial en el caso de redes de gran tamaño. El problema tiende a centrarse en comunidades de menor tamaño, que debido al compromiso arriba mencionado, tienden a quedar fusionadas en comunidades de mayor tamaño (Giustolisi & Ridolfi, 2014 a, b; Lancichinetti & Fortunato, 2011).

II.3 Representacion de Redes de Abastecimiento de Agua Potable como Grafos de Redes Sociales

A lo largo de las últimas dos décadas, los modelos matemáticos de RDAPs se han vuelto una herramienta necesaria/indispensable para la correcta gestión de las mismas. Permiten a las empresas gestoras de acueductos, entre otras cosas, evaluar alternativas de gestión, gestionar situaciones de crisis de naturalezas diversas y definir tareas de mejora. Es posible que la herramienta de modelización de RDAPs más extendida en la actualidad sea EPANET, creada por el departamento de seguridad de la Agencia de Protección Ambiental de EEUU EPA (*Environmental Protection Agency*). En este software se pueden representar casi todos los elementos propios de una RDAP, siguiendo una topología básica de nodos (extremos de tuberías, tanques, reservorios) y enlaces (tuberías, válvulas, bombas). En EPANET un enlace requiere dos nodos extremos, lo que hace que la topología de las RDAPs

coincida con la topología de los grafos y, por ende, su transformación es una tarea sencilla. Para ello se establecen los nodos de consumo y abastecimiento (tanques, reservorios) como los vértices del grafo, y las tuberías, válvulas y bombas como las aristas. La definición de un dígrafo de RDAPs puede ser un poco compleja ya que la dirección de las tuberías en un modelo en EPANET es asignada arbitrariamente. Una alternativa es emplear la dirección del caudal que fluye por las tuberías para establecer la dirección de las aristas del grafo. En caso de querer convertir el dígrafo no ponderado en un dígrafo ponderado, las características que se pueden agregar a los vértices pueden ser las contenidas en los nodos del modelo matemático: coordenadas geográficas, cota, demanda, coeficiente de emisor, presión, altura piezométrica; y en las tuberías: diámetro, longitud, rugosidad, caudal, pérdidas de carga por fricción. También se pueden crear y/o agregar nuevas propiedades. Por ejemplo, en el caso de los vértices, se podrían agregar características sociales propias de los sitios de estudio, o la potencia hidráulica en los nodos, expresadas en $m^3/h * mc$. En esta tesis, los grafos son representados en el lenguaje de programación *R*. Para ello se exportan a partir de un archivo INP del modelo matemático dos tablas (con las características que se deseen agregar al grafo): primero la tabla de nodos y luego la tabla de tuberías. En el proceso de creación del grafo, se extraen los IDs de los nodos a partir de la primera tabla y con dichos IDs en conjunto con la segunda tabla se construye una lista de adyacencia. A partir de ambas se crea el grafo mediante el paquete *Igraph*. Para poder visualizar la red con su forma original, es importante establecer las coordenadas geográficas de cada uno de los nodos, que se obtienen igualmente a partir del archivo INP. Todo este proceso se encuentra incluido en la función `crear.grafo` creada para este trabajo y que se puede encontrar en Anexo II.

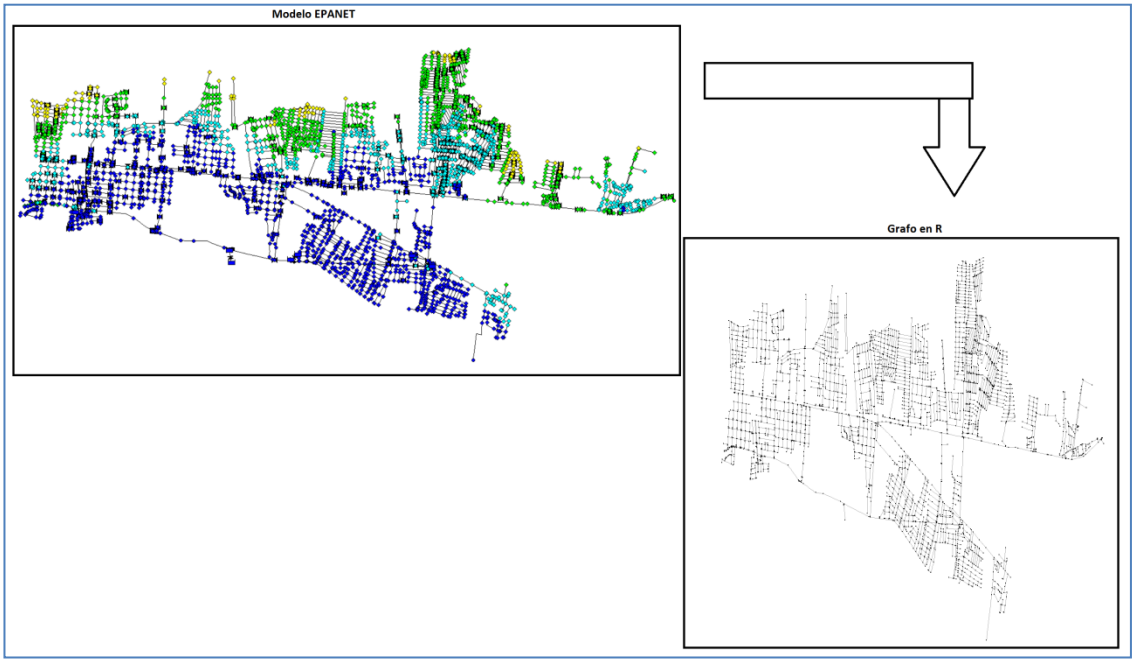


Ilustración 15: Modelo matemático en EPANET (izquierda) y grafo (derecha) de porción de la red de Managua

II.4 Identificación de Red de Conducción Principal Mediante Concepto de Caminos más Cortos

Tal y como se señaló en la Subsección II.1.3 Caminos más Cortos en Grafos, el CMC entre los nodos de una red determinada, o el camino de mínimo número de vértices, junto con la detección de comunidades, han sido dos de los conceptos más ampliamente explorados dentro del área de investigación tanto de la teoría de grafos como del dominio de la teoría de redes sociales. Mediante la aplicación del concepto de CMC sobre un dígrafo es posible encontrar el camino de flujo máximo entre cualquier par de nodos. Tal y como se estableció en la subsección anterior (II.3 Representación de RDAPs como Grafos de Redes Sociales), la información contenida en las tuberías de una RDAP dada se puede incorporar en los enlaces de su respectivo grafo, incluyendo la dirección del flujo que viaja a través de las mismas, lo cual permite definir las direcciones de los enlaces del grafo en cuestión. Evidentemente, para conocer la dirección del flujo, se requiere una simulación hidráulica. Cuando dicha simulación se lleva a cabo en el escenario de máxima

demanda (el punto más alto del patrón de demanda), la dirección del flujo en cada tubería se corresponde con la dirección en el escenario más crítico.

Una vez que se definen las direcciones de los enlaces del grafo, es posible realizar el cálculo de un Valor de Camino más Corto (VCMC) entre cada par de nodos. Esto se lleva a cabo a través de los algoritmos BFS, DFS y Dijkstra, también descritos en la Sección II.1.3. Caminos más Cortos en Grafos.

El VCMC para cada par de nodos puede ser almacenado en una matriz cuadrada (Ilustración 16), en la cual, las filas etiquetan a los nodos de partida y las columnas a los nodos de llegada. Para la construcción de dicha matriz se sigue el siguiente conjunto de reglas, donde NI significa Nodo Inicial y NF significa Nodo Final.

Si $NI = NF \rightarrow 0$
Si desde NI no se puede alcanzar $NF \rightarrow \infty$
Para cualquier otro caso → Número de nodos en el CMC

$$\begin{pmatrix} & Nodo_1 & Nodo_2 & \dots & Nodo_n \\ Nodo_1 & 0 & & & \\ Nodo_2 & & 0 & & \\ \vdots & & & 0 & \\ Nodo_n & & & & 0 \end{pmatrix}$$

Ilustración 16: Matriz modelo para los VCMCs

Es importante destacar dos aspectos: en primer lugar, en el caso de grafos no dirigidos (grafos en los que no se establece la dirección de los enlaces), cada nodo es accesible desde cualquier otro nodo en el grafo, por lo tanto, ∞ no es una posibilidad en ese caso. En segundo lugar, en el caso de los grafos no ponderados, se asume que todos los enlaces tienen el mismo peso (1), y por lo tanto, el CMC entre cualquier par de nodos se evalúa basándose únicamente en el número de nodos.

Mediante la matriz anterior se puede realizar una clasificación de los nodos en función de su importancia dentro de la estructura del grafo. Un nodo dado se considera más importante cuanto más grande sea la cantidad de otros nodos que es

capaz de alcanzar; lo cual se corresponde con la suma de cada fila en la matriz anterior. En este trabajo, tal suma se define como Valor de Camino más Corto Acumulado (VCMCA).

El VCMCA puede ser transferido a los enlaces del grafo, lo cual se lleva a cabo asignando a cada enlace un valor que es igual al VCMCA de su correspondiente nodo aguas abajo + 1. Obsérvese que 1 se añade con el fin de incluir el nudo final de cada camino, que siempre tiene 0 de VCMCA. De esta manera, cuanto mayor es el VCMCA de una tubería dada, mayor es su importancia dentro del grafo y, por lo tanto, con el fin de preservar al máximo la fiabilidad de la RDAP, esta misma no debería incluirse entre los sectores. Dado que el VCMCA de algunas de las tuberías puede ser considerable mayor que de otras, el valor en cuestión se puede reescalar por medio de la división del valor de cada una de ellas entre el máximo VCMCA de toda la red. De esta manera, el VCMCA de las tuberías conectadas a la fuente principal tendría un valor igual a 1 y los tubos conectados a los nodos extremos tendrían valores cercanos a 0.

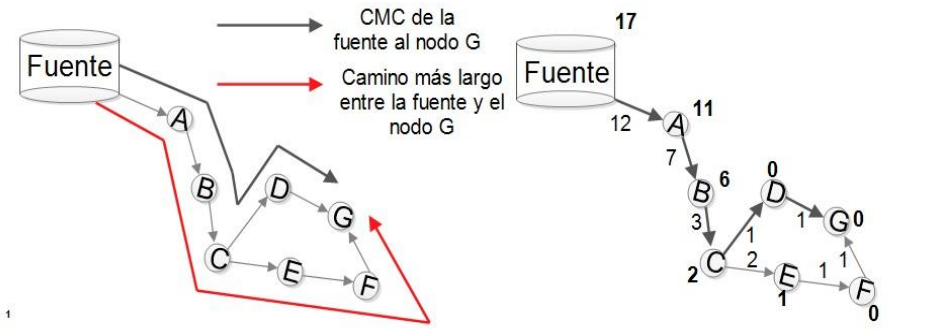
Es importante tener en cuenta que si el VCMCA se utiliza solo, algunas tuberías con pequeños valores de diámetro y caudal pueden ser incluidas dentro del conjunto de tuberías de la red de conducción principal, siempre y cuando los mismos sean capaces de introducir caudal a las tuberías de diámetro mayor. Sin embargo, si el VCMCA se utiliza como un factor multiplicador del caudal que se transporta a través de cada tubería, únicamente las tuberías con valores de diámetro y caudal grandes, se incluyen dentro del conjunto de la red de conducción principal. Esta multiplicación se denota con el símbolo VCMCA*. Como resultado final, se forma una jerarquía (o un árbol) de tuberías y nodos en el cual las fuentes de abastecimiento conforman las raíces.

Las RDAPs de la mayoría de las grandes ciudades están formadas por bloques de mallas conectados a otros bloques y a una red de conducción principal mediante un número reducido de tuberías. Esto hace que el número de tuberías que conforman la red de conducción principal sea significativamente reducido y que estas, a su vez,

reporten valores más altos de VCMCA*. De manera opuesta, es de esperar que las tuberías de la red de distribución sean muy frecuentes y que se caractericen por tener valores de VCMCA* significativamente menores (cercanos a 0 en el caso de tuberías que están conectadas a un nodo que está completamente aislado o que solamente recibe agua). Con base en esta idea, en este trabajo, se propone el análisis de frecuencia para determinar qué líneas son parte de la red de conducción principal y cuáles son parte de la red de distribución. Para ello, se construye una tabla de frecuencia con un número muy grande de intervalos de VCMCA* (lo que permite obtener intervalos muy pequeños). A continuación se calcula la frecuencia relativa y el porcentaje de frecuencia acumulada. Cuando la frecuencia relativa y la frecuencia acumulada se representan gráficamente en un diagrama de Pareto, se hace evidente que hay un gran porcentaje de tuberías con bajo VCMCA* (aproximadamente 80 % de las tuberías) y que, por ende, las tuberías con altos VCMCA* son significativamente menos frecuentes (el 20 % restante). También entre los intervalos de frecuencia se hacen notables algunos cambios abruptos. Estos puntos de cambio pueden ser utilizados como criterio de segregación de los dos tipos de red. En vista de que se espera que haya varios puntos alternativos, el personal técnico de la empresa de agua debe escoger, de entre los posibles, el que resulte más conveniente. Por ejemplo, el alcance puede ser limitado al conjunto de tuberías cuyo cierre conllevaría al desabastecimiento ya sea parcial o general. De una manera más general, el alcance puede incluir todas las tuberías con valores de VCMCA* significativamente menos frecuentes.

La idea de marcar, distinguir y/o desacoplar la red de conducción principal de la red de distribución ya ha sido previamente planteada en algunos trabajos de sectorización de RDAPs (ver Tabla 2 en la Introducción). Algunos de estos estudios han sugerido el uso de algunos parámetros asociados al comportamiento hidráulico de la red (caudal, pérdida de carga, diámetros), o el uso de una medida conocida como *Centralidad de Enlaces* (*Edge Betweenness*) (Campbell *et al*., 2014b), mediante la cual se puede caracterizar la centralidad de cada una de las tuberías en función de su rol en el funcionamiento de la red (Freeman, 1979). La utilización única del parámetro *Centralidad de Enlaces* tiene como inconveniente el hecho de

que las tuberías conectadas a los tanques tendrían menores valores de este parámetro, haciendo el proceso de distinción de la red de conducción principal más complejo. El uso de caudal puede fallar en el caso de que las fuentes de agua se encuentren delimitadas para zonas reducidas. El uso de diámetros puede funcionar bien en un diseño lógico de RDAPs. Sin embargo, en algunas situaciones reales, es posible encontrar ya sea sobre-dimensionamiento o sub-dimensionamiento de tuberías. Con base en esto, el indicador VCMCA aquí propuesto se puede considerar un parámetro más robusto y preciso para distinguir la red de conducción principal de la red de distribución. La Ilustración 17 muestra, gráficamente, el concepto de VCMCA aplicado sobre una red conceptual de un solo tanque.



	S	A	B	C	D	E	F	G	CMC
S	0	0	1	2	3	3	4	4	17
A	0	0	0	1	2	2	3	3	11
B	0	0	0	0	1	1	2	2	6
C	0	0	0	0	0	0	1	1	2
D	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	1	1
F	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0

Ilustración 17: Concepto de VCMCA

Para ejemplificar el concepto descrito anteriormente, el mismo se implementa sobre una pequeña sección de la red de Managua, Nicaragua. Como puede verse en la

Ilustración 18, los valores de intervalo cambian drásticamente en la zona donde el porcentaje de frecuencia acumulada no varía mucho.

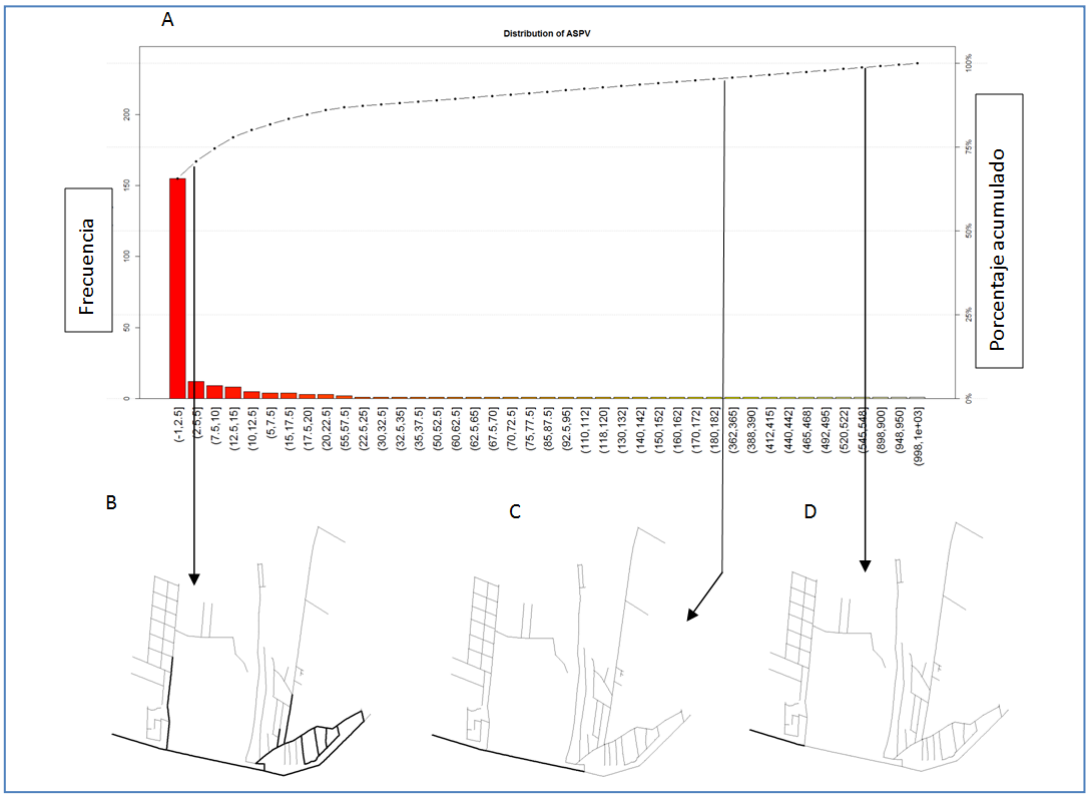


Ilustración 18: (A) Grafo de frecuencia de VCMCA* (diagrama de Pareto) y (B-D) alcances alternativos de la red de conducción principal

Las ilustraciones 18(B) - 18(D) muestran tres alcances alternativos de la red de conducción principal. La primera (B) incluye todas las tuberías con valores de frecuencia acumulada por debajo del 85%. Este corresponde al escenario menos estricto. La alternativa (C) corresponde al primer cambio en los valores de intervalo. El cierre de cualquiera de las tuberías de este subconjunto conllevaría al desabastecimiento en al menos una parte de la red. Finalmente, la alternativa (D) representa el segundo cambio de los valores de intervalo, siendo este el escenario más estricto. La eliminación de cualquiera de las tuberías en este subgrupo implicaría el corte de suministro en toda la red. Las Ilustración 19 y la Ilustración 20 muestran el gráfico de Pareto y el histograma, respectivamente, de los valores de VCMCA* calculados para este ejemplo.

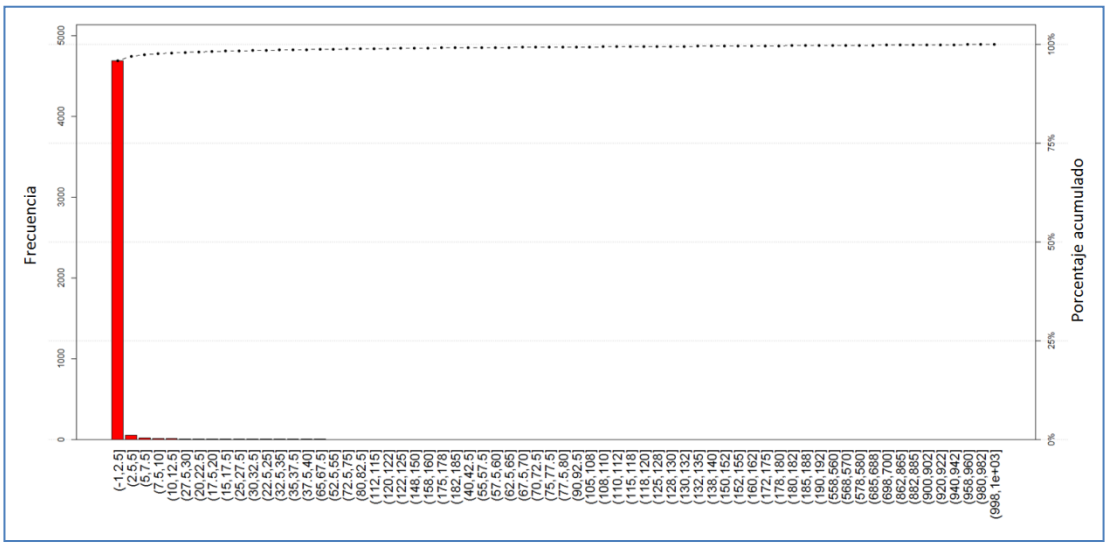


Ilustración 19: Diagrama de Pareto de los valores de VCMCA*

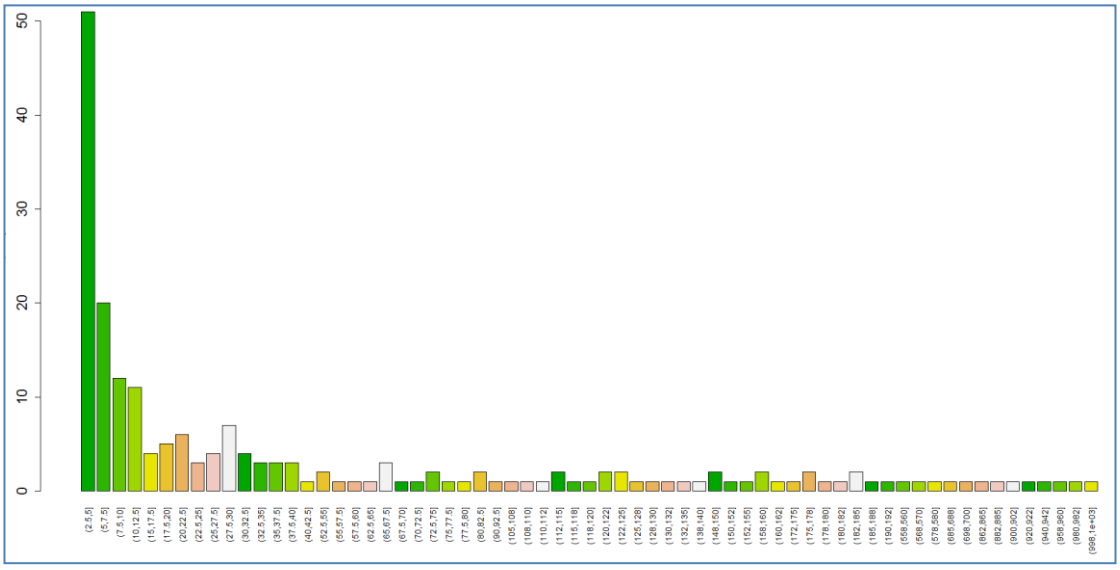


Ilustración 20: Histograma de valores de VCMCA* obtenidos en la red ejemplo

II.5 Aglomeración de Comunidades

Tal y como fue anteriormente establecido, al implementar los algoritmos de detección de comunidades, las posibilidades de particiones son muy amplias; las mismas pueden ser representadas en un dendrograma y, a partir de ahí, se puede seleccionar una u otra de acuerdo a un interés en particular. Una opción es extraer la partición con mayor valor de modularidad; no obstante, las particiones con mayor modularidad no necesariamente tienen que responder a criterios hidráulicos o

topológicos. Es decir, no implican una longitud de tubería definida, ni uniformidad de cotas. Para ello se propone un proceso de re-fusión recursiva (ver Pseudocódigo 5) de las comunidades detectadas por la partición de máxima Modularidad (o por una partición dada).

Para ello se establece TL (como conjunto de tuberías límite) y TC (como conjunto de tuberías candidatas); el índex m representa un tubería; los nodos extremos de m se representan con NI_m, para el nodo inicial, y NF_m, como nodo final; C_i representa una comunidad o sector; L representa la característica utilizada como criterio (longitud total de sector, demanda total de sector, máxima cota de sector, etc.,); y ∇ representa una operación cuyos argumentos son dos valores de L; la operación (suma, máximo, etc.,) depende del significado de L.

Input: una partición con el valor máximo de Modularidad

1. Se calcula un valor de L para cada sector (comunidad) en la partición
2. Para cada m, si $IN_m \in C_i$ y $FN_m \in C_j$, $i \neq j \rightarrow m \in BP$
3. De TL, seleccionar las tuberías cuyos nodos pertenecen a comunidades que cumplen una restricción específica para una característica L; contruir CP con dichas tuberías.
4. Para cada $m \in TC$, sea i y j valores tales que $NI_m \in C_i$ y $NF_m \in C_j$;
   Si $L_i \nabla L_j$ cumple con la restricción $L_{\text{límite}} \rightarrow C_i \cup C_j$
   reemplazar C_i y C_j en la partición
5. La característica de cada C en la nueva partición es recalculada

Pseudocódigo 5: Proceso de re-fusión de comunidades

La esencia del proceso está en el paso 4, en el cual se establece que dos sectores C_i y C_j se funsionan sólo si la unión $C_i \cup C_j$ produce un nuevo sector que cumple

con la restricción de interés ($L_{límite}$). Nótese que sólo una característica se puede utilizar como criterio de fusión.

Es también importante fijar un valor de límite menor para la característica empleada para definir el tamaño de los sectores. Así, si al final del proceso hay algunos sectores con un valor más pequeño que el límite, estos se establecen como minisectores (sin válvulas o UOCs). Esto sólo aplica a minisectores que no pueden ser fusionados con sectores más grandes. En el caso de que un minisector comparta al menos una conexión con un sector que ha alcanzado el valor máximo para la característica establecida, el límite máximo es levemente relajado para permitir la fusión. Por ejemplo, si la característica usada como criterio es la longitud de tubería (e.g. 30 km como máxima restricción), la longitud máxima final que un sector dado puede eventualmente tener será igual a la longitud máxima (30 km) más un valor que es menor al valor mínimo de longitud de tubería de cualquier otro sector (e.g. 4 km); en otras palabras, un valor entre 30 km y 34 km también puede ser aceptado.

II.6 Sectorización basada en el Clústering Jerárquico

II.6.1 Descripción de Definición de Comunidades Mediante Clústering Jerárquico

El clústering jerárquico es una herramienta de análisis de datos que se basa en la construcción de clústeres bajo un orden de jerarquía. Se agrupa dentro de los métodos de aprendizaje automático no supervisado, y tienen como objetivo global hacer una exploración de los datos. La idea tras el método es construir árboles binarios que, sucesivamente, se fusionen en grupos en función de su similaridad (Han *et al*., 2006). A partir del estudio del árbol que se genere, se puede extraer información útil que ayude a comprender la estructura de los datos. Se diferencia de otros métodos de formación de clústeres en el hecho de que no se requiere introducir *a priori* un número de clústeres en el que se hará la partición (Manning *et al.*, 2008), por el contrario, tal y como se explicará a continuación, el número de

clústeres óptimo en el que se deben partir los datos puede ser estimado a través del árbol de jerarquía resultante. En el resto de los métodos de partición de datos en clústeres se asigna una "tarea" inicial (un centroide por ejemplo) a los clústeres, con respecto a la cual estos se terminan de conformar.

II.6.1.1 Pasos de Clústering Jerárquico Aglomerativo

- **Paso 1: Matriz de Disimilaridad**

Partiendo de un conjunto de datos finitos dado (casos u observaciones), y un conjunto de propiedades (las mismas para todos los casos) descritas por variables aleatorias ya sean continuas o discretas, el proceso de clústering jerárquico comienza estableciendo cada uno de los casos como un clúster individual (o *singleton*). A continuación, se forma una matriz $n \times n$ en donde n es el número de casos (que pueden ser nodos de una red). Esta matriz sirve para evaluar las disimilaridades en parejas del conjunto de variables en cada uno de los casos, lo cual se puede hacer usando diferentes medidas métricas (métrica Euclidiana, métrica Manhattan y métrica Gower), que se describen a continuación. Si cada caso cuenta con más de una variable, la comparación se hace entre los valores que toma cada variable individual para cada caso. En caso de emplear como métrica la distancia Euclidiana, el valor de disimilaridad es igual a la raíz cuadrada de la suma de los cuadrados de cada una de las disimilaridades obtenidas. Si se emplea como métrica la distancia Manhattan, la distancia corresponde a aquella que existe entre cada par de casos siguiendo un camino tipo malla. Finalmente, también se puede emplear la distancia Gower, apropiada cuando se tiene un conjunto de datos mixtos, tal como podría ser una mezcla de variables cuantitativas y cualitativas (o categóricas).

Distancia euclidiana

$$d_E(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \qquad \text{(Ecuación 19)}$$

$$d_E(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad \text{(Ecuación 20)}$$

Distancia Manhattan

$$d = \sum_{i=1}^{n} |X_i - Y_i| \qquad \text{(Ecuación 21)}$$

Distancia Gower

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

$$s_{ij} = \frac{\sum_{h=1}^{p1} \frac{1-|x_{jh}-x_{jh}|}{G_h} + a + \alpha}{p_1 + (p_2 - d) + p_3} \qquad \text{(Ecuación 22)}$$

donde: p_1 es el número de variables cuantitativas continuas; p_2 es el número de variables binarias; p_3 es el número de variables cualitativas (no binarias); *a* es el número de coincidencias (1,1) en las variables binarias; *d* es el número de coincidencia (0,0) en las variables binarias; α es el número de coincidencias en las variables cualitativas (no binarias); y G_h es el rango (o recorrido) de la (*h-ésima*) variable cualitativa.

- **Paso 2: Aglomeraciones de Casos en Clústeres**

En el proceso de aglomeración de casos en clústeres, en un primer momento, se producen los clústeres en función de la comparación entre las variables en cada caso individual y, a continuación, de uno u otro elemento de los clústeres conformados

(una vez que los clústeres dejen de ser *singleton*). El elemento característico de cada clúster que se emplee para hacer la comparación depende del método de aglomeración que se seleccione. Los métodos más comúnmente empleados son: agrupación por promedio, agrupación completa (Defays, 1977), agrupación individual (Sibson, 1973), y finalmente, agrupación basada en un centroide (Everitt *et al.*, 2011). Existen otros métodos que se pueden emplear, tal como el conocido método de la mínima varianza (de Ward) (Ward Jr., 1963; Murtagh & Legendre 2011, 2014); el método de la mediana; el método de máxima probabilidad de igual varianza (o EML); el método de McQuitty (McQuitty, 1966), y el método flexible-beta. A continuación, se hará una explicación de los cinco métodos más reconocidos, que serán los empleados en el caso de estudio que se presenta en este documento.

Partiendo de una matriz $n \times n$ que contiene los distintos valores de disimilaridades entre pares de casos, el proceso consiste en formar un primer clúster con el primer par de casos, que tenga el menor valor de distancia entre sí, según el método de aglomeración que sea seleccionado. Este valor mínimo que se emplea como criterio de agrupación se denomina *altura de enlace*. Formado este primer clúster, se actualiza la matriz $n \times n$, en donde el clúster formado anteriormente pasa a ser un nuevo caso, repitiendo nuevamente el paso anterior hasta que en la matriz exista un único caso, que agrupe a todos los elementos del conjunto.

Método de Aglomeración Promedio

En este método, las agrupaciones se van dando según la distancia entre el promedio entre los elementos de un clúster con respecto al promedio de los elementos de otro clúster. Este, junto con el método basado en centroides, está pensado para datos espaciales consistentes en medidas escaladas en intervalos (Chen & Xu, 2003).

$$D_{KL} = \frac{1}{n_k n_l} \sum_{ie\, C_K} \sum_{ie\, C_l} d(x_i x_j) \qquad \text{(Ecuación 22)}$$

Aquí k y l representan dos clústeres distintos y $d(x_i x_j)$ es la distancia entre dos elementos que pertenecen al mismo clúster.

Método de Aglomeración Centroide

En este caso se define un centroide entre los clústeres y las agrupaciones se dan entre los pares de clústeres con mayor similitud en el valor de sus respectivos centroides. Los centroides de cada nodo corresponden a la distancia Euclidiana entre los nodos de cada sector.

$$D_{KL} = \left|\left|\text{prom}X_k - \text{prom}X_l\right|\right|^2 \qquad \text{(Ecuación 23)}$$

Este método tiende a formar clústeres pequeños, con poca varianza intra-clúster. Dado que en el mismo se comparan los centroides de los clústeres, los valores atípicos del conjunto de datos tienden a distorsionar el resultado; no obstante, la distorsión que esto causa en el método es menor que la que se genera si se emplea el método completo (Mooi & Sardtedt, 2011).

Método de Aglomeración Completa o de Vecinos más Lejanos

En este caso, la comparación de similitudes en parejas se limita a los dos casos más alejados entre cada clúster.

$$D_{KL} = \text{Máx}_{I \in C_k J \in C_L} d\left(x_i, x_j\right) \qquad \text{(Ecuación 24)}$$

La desventaja del mismo radica en su tendencia a formar clústeres compactos con diámetro aproximadamente similar y, además, el resultado, en algunos casos, se ve distorsionado por la presencia de valores atípicos (*outliers*). Esto último se debe a que el criterio de unión no es local, sino global; así, la estructura entera de clústering puede influenciar la decisión de fusión (Manning *et al*., 2008).

Método de Aglomeración Individual

Es, en cierto sentido, opuesto al método centroide. En este caso, las comparaciones en parejas de casos en distintos clústeres se hacen empleando los casos con valores de disimilaridad más cerca o los *vecinos más próximos*.

$$D_{KL} = \text{Mín}_{I \in C_k J \in C_L} d(x_i, x_j) \qquad \text{(Ecuación 24)}$$

Pese a tener propiedades teóricas atractivas que lo hacen muy popular, este método tiene una tendencia a formar clústeres muy dispersos, que lo hace insostenible para aislar clústeres esféricos o clústeres que se encuentren pobremente separados (Amar *et al*., 1997).

Método Ward o de Mínimo Incremento de Suma de Cuadrados

$$E = \sum_{k=1}^{h} E_k \qquad \text{(Ecuación 25)}$$

donde $E_k = \sum_{i=1}^{n_k} \sum_{j=1}^{n} \left(x_{ij}^k - m^k\right)^2$; x_{ij}^k es la variable i del individuo j del clúster k suponiendo que el cluster tiene n_k individuos ; m^k es el centroide del clúster k; E_k es la suma de cuadrados de los errores del clúster k, o sea, la distancia Euclidiana al cuadrado entre cada individuo del clúster k a su centroide.

- **Paso 3: Representación del Clúster Jerárquico Aglomerativo**

Para hacer una representación del proceso de formación de clústeres siguiendo un proceso jerárquico, la herramienta utilizada más habitualmente es el dendrograma, el cual fue descrito en la Subsección II.1.4.

- **Paso 4: Selección de Métodos a Emplear**

De todo lo expuesto previamente se puede hacer notar que existen muchos posibles

caminos para llevar a cabo el proceso de clústering jerárquico. Cada camino está definido por una métrica para medir la disimilaridades entre casos (métricas Euclidiana, Manhattan, Gower) y un método seleccionado para definir la unión entre dos clústeres (promedio, completo, individual, centroide, Ward). Para un mismo conjunto de datos, cada camino arroja un resultado distinto, lo que hace surgir la pregunta sobre qué camino seguir. Al respecto, Everitt *et al*. (2011) establece que no hay un método particular de clústering jerárquico que se pueda recomendar. Una medida bastante aceptada para evaluar el mejor camino a seguir es el Coeficiente de Correlación Cofenética (CCCF por C*ophenetic Correlation Coefficient*), que ha sido ampliamente utilizado en estudios de clasificación fenética (Farris, 1969; Gonçalves *et al*., 2008; Podani & Dénes, 2006). Este es muy reconocido dentro de la estadística para medir el grado de fiabilidad con que se puede decir que un dendrograma conserva las distancias en parejas entre los datos originales que no han sido modelados. Se emplea para evaluar el grado de ajuste de una clasificación a un conjunto de datos y como criterio para evaluar la eficiencia de varias técnicas para obtención de clústeres. Más concretamente, el CCCF en un conjunto de datos que ha sido dividido en clústeres se define como el valor de correlación de la matriz de disimilaridad inicial del conjunto de casos en los que se hace la partición y una matriz conocida como matriz ultramétrica, la que contiene para cada par de casos, el valor de altura con el cual tales casos se agruparon por primera vez (Sokal & Michener, 1958; Sokal & Rohlf, 1962).

A continuación se presenta la ecuación para la obtención del índice CCCF.

$$r_{xy} = \frac{\sum xy - \left({}^{1}/_{n}\right)(\sum x)(\sum y)}{\{[\sum x^2 - \left({}^{1}/_{n}\right)(\sum x)^2][\sum y^2 - \left(\frac{1}{n}\right)(\sum y)^2]\}^{{}^{1}/_{2}}} \qquad \text{(Ecuación 26)}$$

Aquí x representa los valores de disimilaridades en pares S_{ij} de la matriz de disimilaridad; y representa los valores de altura de la matriz ultramétrica y n el número de casos en estudio.

El coeficiente resultante puede tomar valores entre 0 y 1 y, a través de él, se puede evaluar el nivel de distorsión que genera el método empleado sobre el conjunto de datos. Pese a que no existe una directriz clara respecto al nivel que es tolerable, en la mayor parte de los estudios en que se ha empleado la técnica de clústering jerárquico, se acepta un valor superior a 0.8 como indicador de una distorsión no excesiva (Romesburg, 2004).

II.6.2 Ejemplo de Clústering Jerárquico

Dado un conjunto de datos *x* que contiene un número *n* de casos, el proceso se inicia creando la matriz de disimilaridad entre los pares de casos S_{ij}. Esta es una matriz $n \times n$, en la que los valores cumplen con la siguiente característica

$$D = (S_{ij}),\ 1 \leq S_{ij} \leq n.$$

Inicialmente se hace una primera partición: $x = \{1\}+\{2\}+.....+\{n\}$. De manera que cada caso es un clúster individual o *singleton.*

Por comparación entre los distintos pares se obtendrá un par (i, j) que tiene menor disimilitud y, por ende, mayor cercanía entre todo el conjunto de datos. Este par se une para constituir el primer conglomerado.

$$\{i\} \cup \{j\} \rightarrow \{i, j\}$$

El valor de disimilaridad entre ambos elementos representa la altura de enlace S'_{ij}, y, por ende, el primer valor de la matriz ultramétrica.

A continuación se compara el nuevo clúster con el resto de elementos.

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

donde S' representa el nuevo valor que tendrá la matriz $n \times n$ para la comparación

de cada elemento (k) con el clúster $\{i, j\}$. Dicho valor se obtiene a partir de una función seleccionada (ƒ) a partir del algoritmo que se emplee para la función (promedio, centroide, individual, completa).

Esto implica una nueva partición, en donde $\{i, j\}$ ya están constituidos como un clúster.

$$\sigma = \{1\}+.....\{i, j\}+.....+\{n\}.$$

Posteriormente se repite el proceso desde la selección del menor valor en la matriz de disimilaridad. El proceso termina cuando todos los casos queden agrupados en un único clúster.

$$x = \{1, 2 \dots, n\}.$$

Esta unión tendrá el mayor valor de altura ($S'_{ijk\,..n}$) en la matriz utramétrica.

Ahora, se puede representar el proceso anterior empleando como ejemplo una zona del grafo presentando previamente (Ver Ilustración 6 e Ilustración 21). Tomando los nodos 2, 3, 4 y 5 como casos, y asignándole a los mismos dos características, tal y como se puede ver en la Tabla 4, se puede iniciar el proceso construyendo una matriz de disimilaridad entre todos los casos.

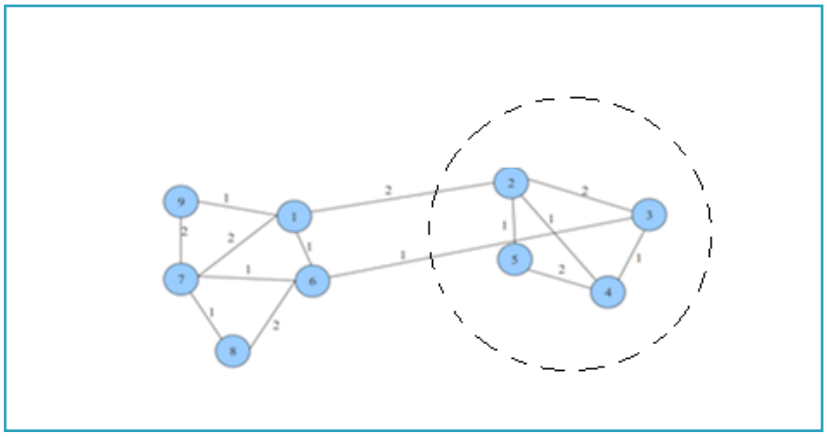


Ilustración 21: Ejemplo de aplicación de clústering jerárquico

caso	Característica 1	Característica 2
2	1.00	3.00
3	3.00	9.00
4	5.00	4.00
5	7.00	3.00

Tabla 7: Datos de ejemplo de clústering jerárquico

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

$$x_1 = \{2\}+\{3\}+\{4\}+\{5\}.$$

Para efectuar el cálculo de disimilaridad por pares entre los casos, se emplea como métrica la distancia Euclidiana.

	2	**3**	**4**	**5**
2	0.00			
3	6.32	0.00		
4	4.12	5.39	0.00	
5	6.00	7.21	**2.24**	0.00

Tabla 8: Matriz de disimilaridad del ejemplo de clústering jerárquico

En la matriz de disimilaridad obtenida (Tabla 8), se puede ver que los casos 4 y 5 son los que menor valor de disimilaridad tienen (con un valor de altura S'_{ij} = 2.24), con lo cual constituyen el primer clúster que se formará $\{4, 5\}$. Por tanto, se compara este nuevo clúster con el resto de los casos ($k = 3, \ldots, n$) a través de una función promedio.

{4,5},2 = [4-2], [5-2] = Promedio (4.12, 6) = 5.06

{4,5},3 = [4-3], [5-3] = Promedio (5.39, 7.21) = 6.3

La nueva matriz $n \times n$ de disimilaridad queda de la siguiente forma:

	2	3	4-5
2	0.00		
3	6.32	0.00	
4-5	**5.06**	6.3	0.00

Tabla 9: Matriz de aglomeración actualizada

El par de casos con la mayor similaridad lo constituyen el conjunto $\{\{4,5\},2\}$. Así, se constituye el clúster {4,5,2}. Para este enlace, la altura correspondiente $S'_{ij\ k}$ = 5.06. Ahora se procede a hacer la comparación entre este nuevo clúster con el único caso restante k_n.

$\{\{4,5,2\},3\}$= [4,5-3], [2-3] = Promedio (6.30, 6.32) = 6.31

La matriz ultramétrica resultante es la correspondiente a la Tabla 10.

	2	3	4	5
2	0.00			
3	6.3	0.00		
4	5.06	6.3	0.00	
5	5.06	6.3	2.24	0.00

Tabla 10: Matriz *ultramétrica*

El CCCF entre esta matriz y la matriz de disimilaridad inicial es igual a 0.88, lo que valida el camino seleccionado para la obtención jerárquica de los clústeres.

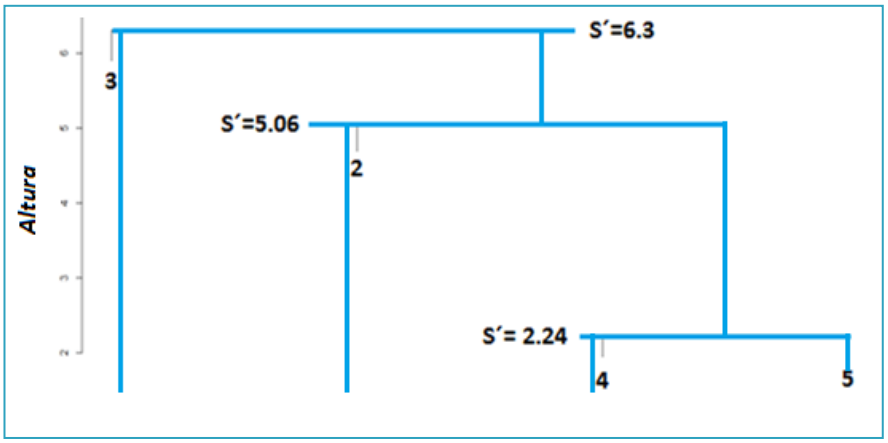


Ilustración 22: Valores de altura en el dendrograma (Fuente: Campbell, 2013a)

II.6.3 Ejemplo de Implementación

El primer paso para la ejecución del ejemplo de implementación consiste en la definición de la red de conducción principal sobre la red ejemplo (ver Ilustración 23). En la Ilustración 24 y en la 25 se muestra el diagrama de Pareto de los valores de VCMCA* y el histograma de VCMCA*, respectivamente, obtenidos en la red ejemplo. En la misma se puede ver que más del 80% de las tuberías presentan valores de VCMCA* inferiores a 10. Al ver con más detalle la zona con valores superiores a 10 (en la Ilustración 25), se puede ver que dentro de la misma se pueden establecer más categorías de tuberías (por ejemplo, las que tienen valores menores a 70 y las que tienen valores mayores a 70). En la Ilustración 27 y la Ilustración 28, se muestran distintas alternativas de alcance de red troncal mediante el establecimiento de estos tres criterios de definición. Para el ejemplo de implementación se emplea la alternativa 1 (Ilustración 26), dado que cubre todas las fuentes de abastecimiento de la red, evitando de esta manera que alguna de ellas quede ubicada dentro de alguno de los sectores.

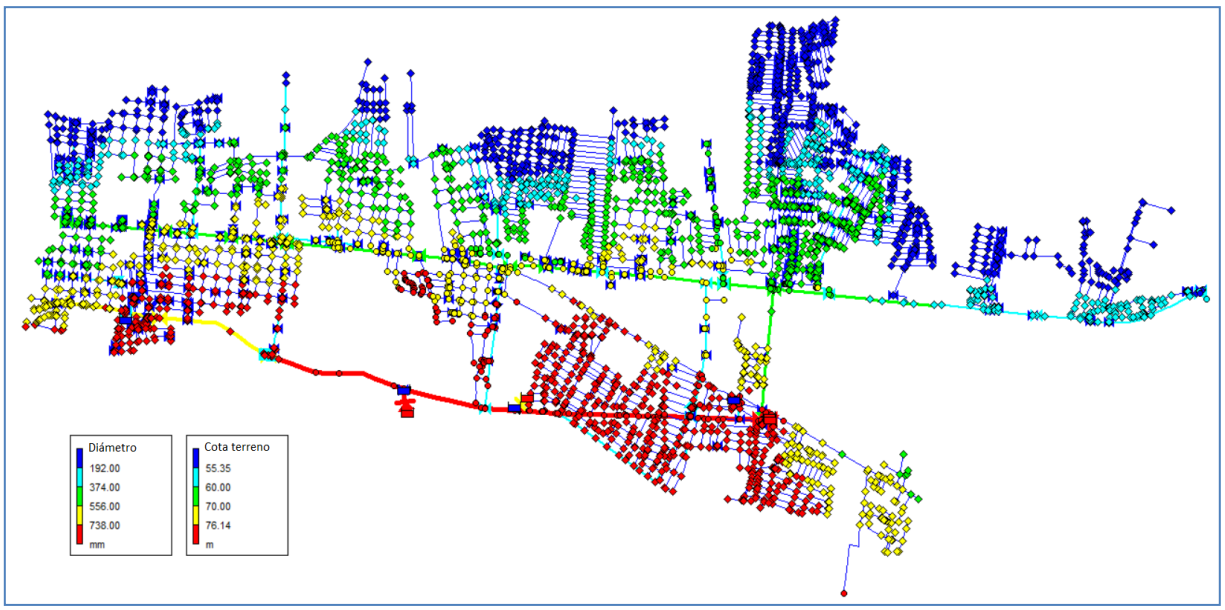


Ilustración 23: Diámetros y cotas de la red ejemplo

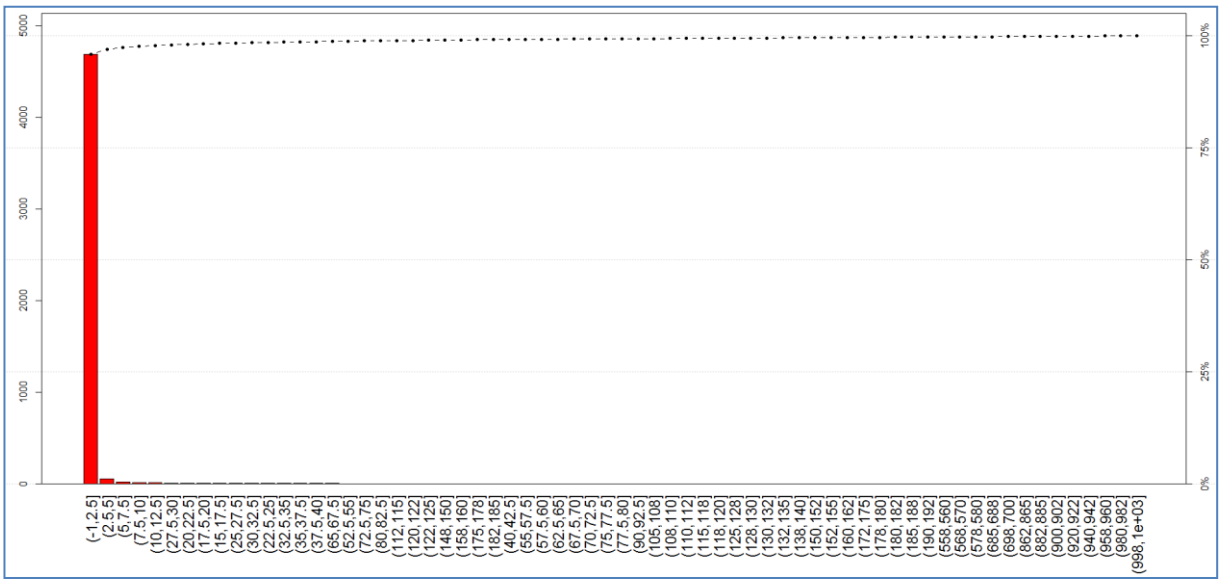


Ilustración 24: Diagrama de Pareto de los valores de VCMCA* en la red ejemplo

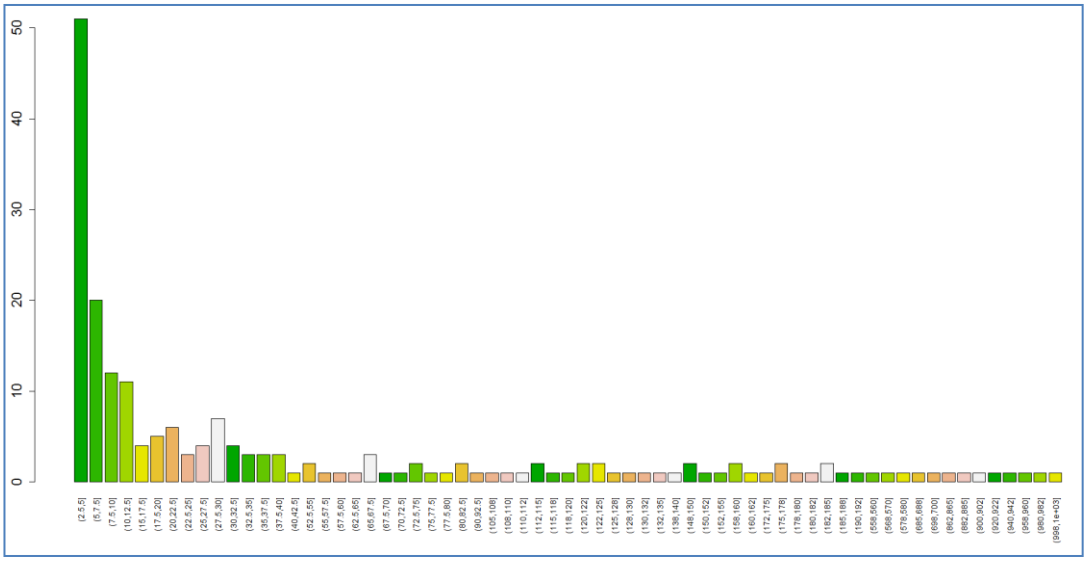


Ilustración 25: Histograma de los valores de VCMCA* de la red ejemplo

Ilustración 26: Red troncal utilizando VCMCA* = 10 como criterio de definición

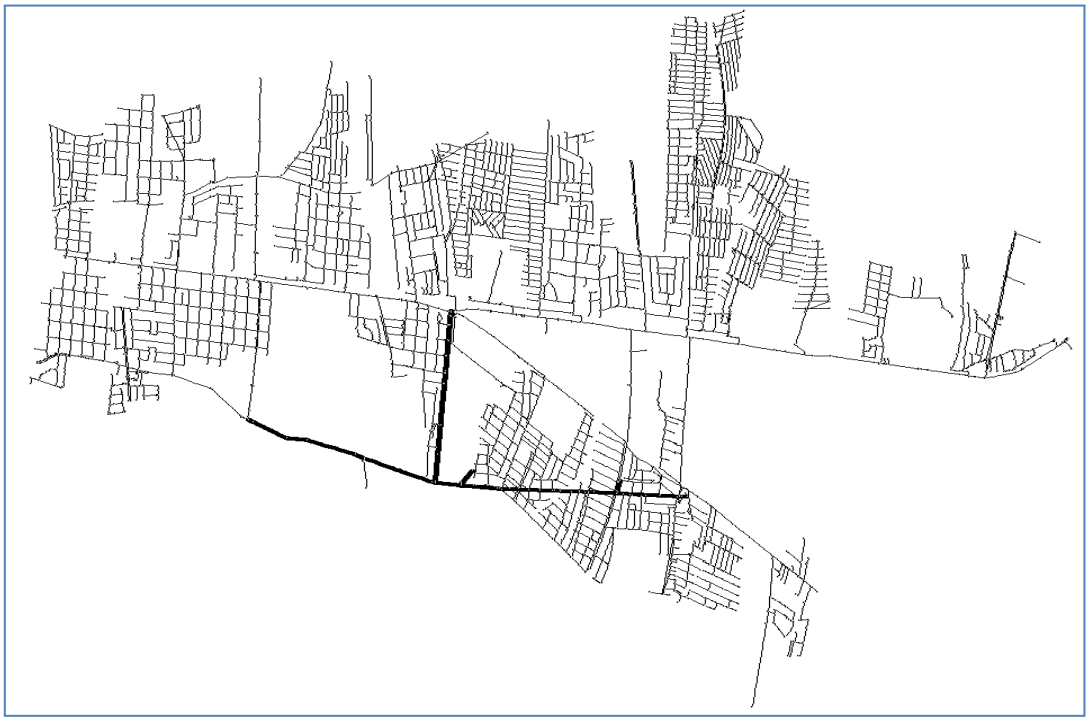


Ilustración 27: Red troncal utilizando VCMCA* = 90 como criterio de definición

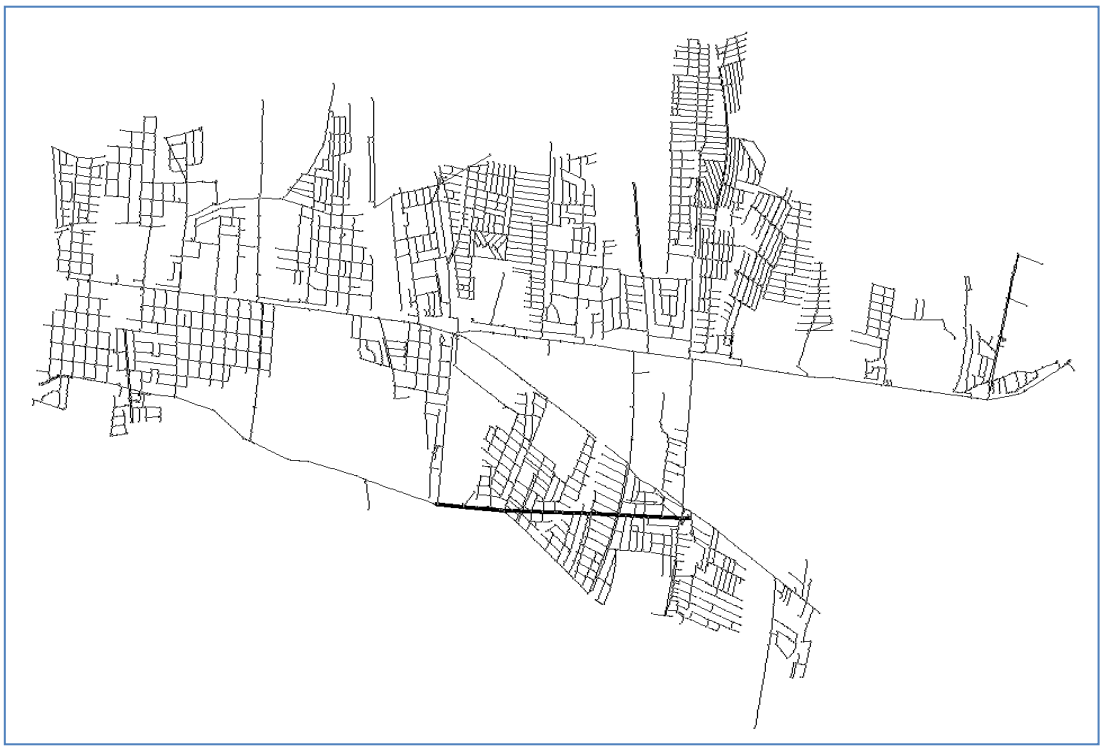


Ilustración 28: Red troncal utilizando VCMCA* = 600 como criterio de definición de red troncal

En la Ilustración 29 se muestra la red troncal cuando para la definición de la misma se emplea como valor 0.1. En este caso se obtiene una red troncal mucho más extendida, que no define claramente áreas de abastecimiento.

Ilustración 29: Red troncal extendida. VCMCA* igual a 0.1

La características empleadas para la construcción de la matriz de disimilaridad sobre la que se implementa el clústering jerárquico corresponden con los vectores de coordenadas geográficas x-y, y con el vector de cota. Alternativamente, se pudo haber incluido el vector de demandas. Sin embargo, la distribución de la demanda entre los nodos de una red es muy subjetiva y no necesariamente tiene que ser muy representativa de la realidad. Es una práctica relativamente común distribuir las demandas de manera uniforme entre el número de nodos, sin tener en cuenta la cantidad de consumidores que se encuentran en la cercanía de cada uno de los nodos. Es por esta razón que esta variable se establece únicamente como criterio en el proceso de re-fusión de sectores y no en el proceso de construcción inicial de clústeres.

Tal como se explicó anteriormente, la técnica de clústering jerárquico ofrece múltiple alternativas de tipo de métricas y métodos de agrupación, para la construcción de la matriz de disimilaridad y para el proceso de aglomeración, respectivamente. Como se describió anteriormente, la métrica Gower ofrece la ventaja de poder lidiar con datos de distinta naturaleza (ver detalles en Subsección II.6.1), tal como sería el caso de las coordenadas geográficas y las cotas en una RDAP, y además, al emplear la misma, se tiene la capacidad de poder adicionar pesos a las variables, de tal manera, que tal y como se verá más adelante, se puede favorecer uno u otro aspecto dentro del resultado de la sectorización. Pese a las mencionadas ventajas, se hace un análisis de la combinación más apropiada entre las distintas métricas de la matriz de disimilaridad y los distintos métodos de aglomeración. Dicho análisis se lleva a cabo mediante el CCCF previamente descrito. De acuerdo a los resultados mostrados en la Tabla 11, la distancia más versátil corresponde a la distancia Euclidiana, ya que produce buenos valores del CCCF con cualquiera de las funciones de agrupación, menos con la función individual. A continuación le sigue la distancia Manhattan, que presenta buenos valores para todas las funciones de agrupación con excepción de la función mediana y la función individual. Por último se ubica la distancia Gower, que produce buenos valores únicamente con las funciones promedio y centroide. De todas las combinaciones en la Tabla 11, la correspondiente a la distancia Euclidiana con el

método promedio presentó el valor de CCCF más alto; sin embargo, vale la pena destacar que la combinación distancia Euclidiana con método centroide y la combinación distancia Gower con método centroide producen valores de CCCF que apenas son ligeramente más bajos que el primero. Además, la última combinación presenta como ventaja la capacidad de asignar pesos a los vectores que forman la matriz de disimilaridad. De ahí que se opte por ejecutar el ejemplo de implementación con estas tres combinaciones.

Tabla 11: Valores del CCCF para distintas combinaciones de métrica y método de aglomeración

	Euclidiana	Manhattan	Gower
mediana	0.7297111	0.5458488	0.5545863
promedio	0.7792184	0.7124197	0.7111165
completo	0.7030722	0.7493743	0.5976114
individual	0.3517099	0.3607142	0.2553433
Ward	0.7604539	0.7461218	0.6423487
centroide	0.7666416	0.7529143	0.7686893

Valores de CCCF más altos

La Ilustración 30, la Ilustración 31 y la Ilustración 32 muestran los dendogramas respectivos de las tres combinaciones seleccionadas. Cómo se puede ver, tanto la primera como la segunda combinación generan clústeres más uniformes y mejor distribuidos, destacándose la primera combinación por formar clústeres compactos en pocos pasos (pocas fusiones). Caso contrario es la tercera combinación, en donde se forman clústeres más grandes en paralelo a clústeres más pequeños, es decir, clústeres con peor distribución. Este comportamiento es propio de bases de datos en donde hay una característica dominante. En el caso de esta red, la característica más dominante corresponde a la cota, ya que hay una zona con una gran cantidad de nodos que tienen valores similares de cota. Si se adiciona un peso a la característica cota, a fin de minimizar su influencia, se puede observar un cambio drástico en el dendrograma (ver Ilustración 33). En este caso, los clústeres se distribuyen de una manera más uniforme.

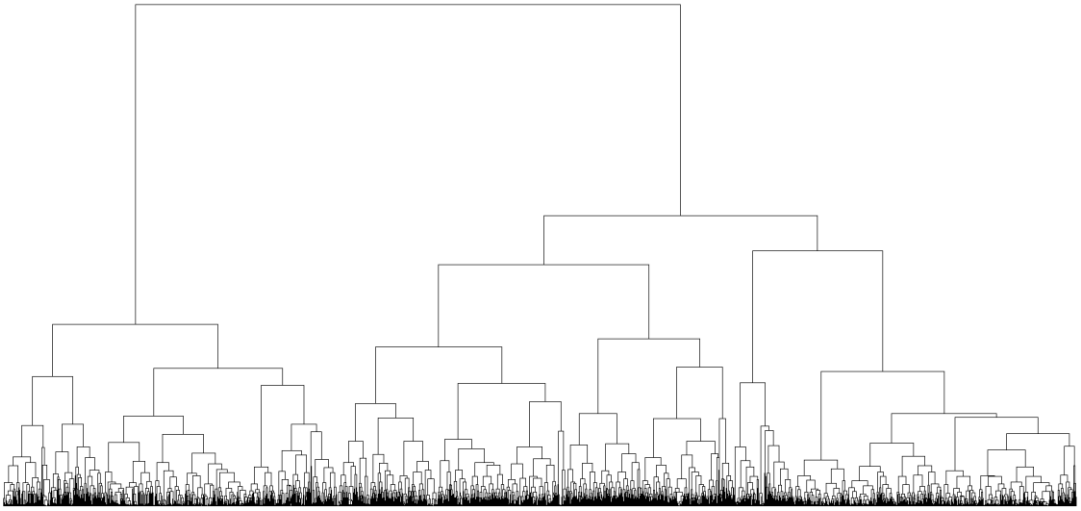


Ilustración 30: Dendrograma generado por la combinación distancia Euclidiana con método promedio

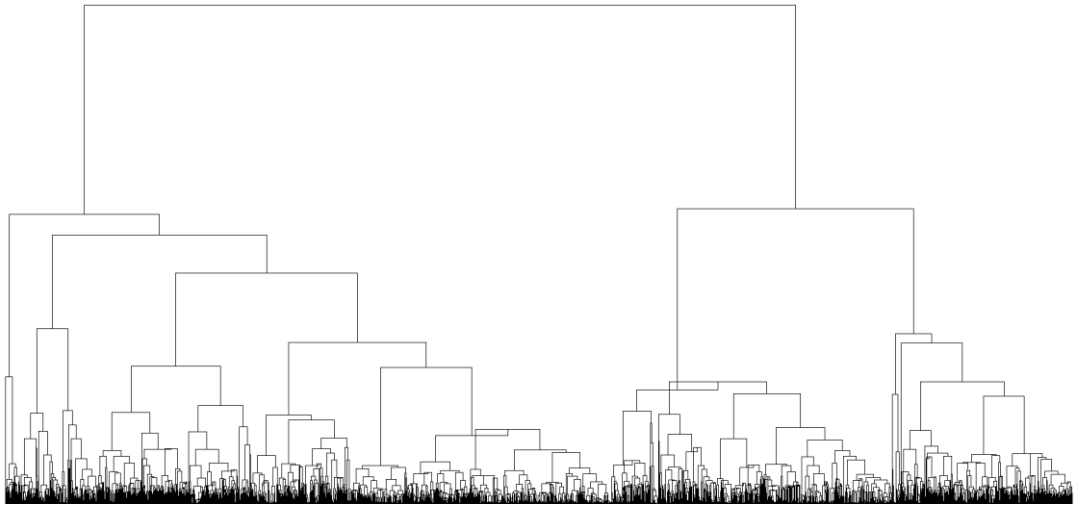


Ilustración 31: Dendrograma generado por la combinación distancia Euclidiana con método centroide

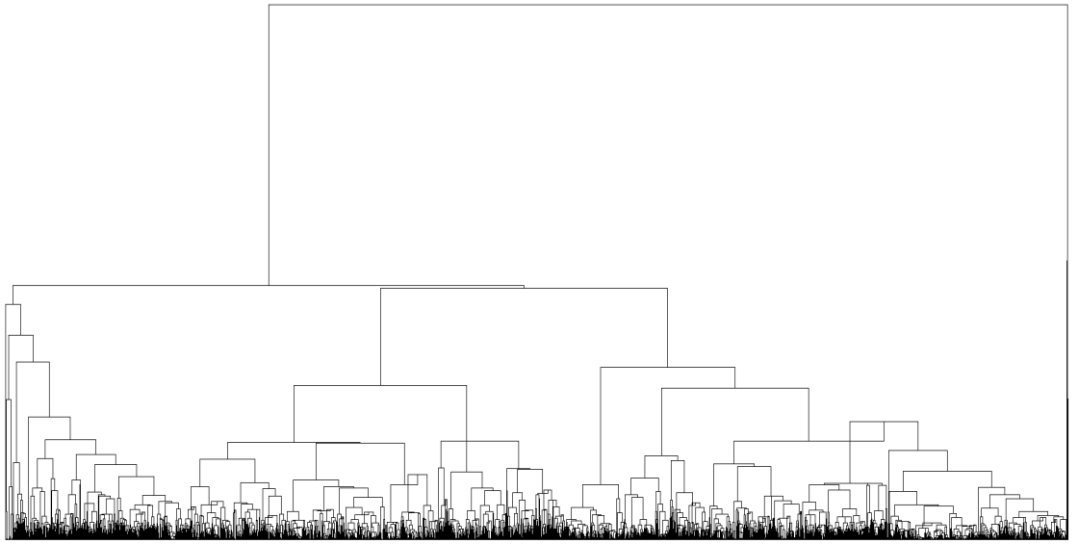


Ilustración 32: Dendrograma generado por la combinación distancia Gower con método centroide

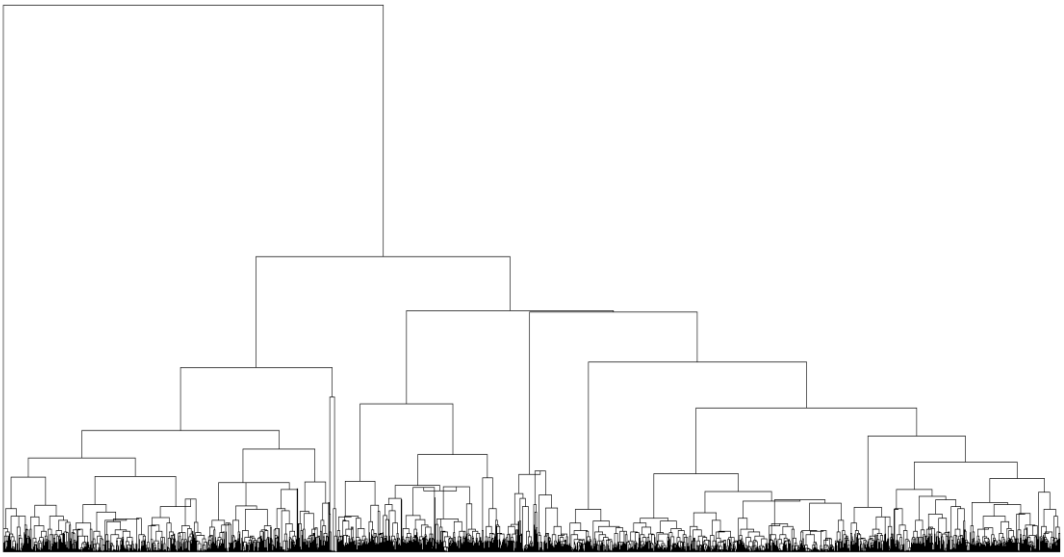


Ilustración 33: Dendrograma generado por la combinación distancia Gower con método centroide al minimizar la influencia de la cota

Se procedió a realizar un análisis del número de comunidades con el que se inicializaría el proceso de re-fusión para cada uno de los casos. Para tal fin se utilizó la herramienta *ClValid*, empleando las técnicas de clústering: jerárquico, DIANA, *k-means*, PAM y SOM, tanto para las medidas de validación internas (ver Ilustración 34) como para las medidas de validación de estabilidad (ver Ilustración 35).

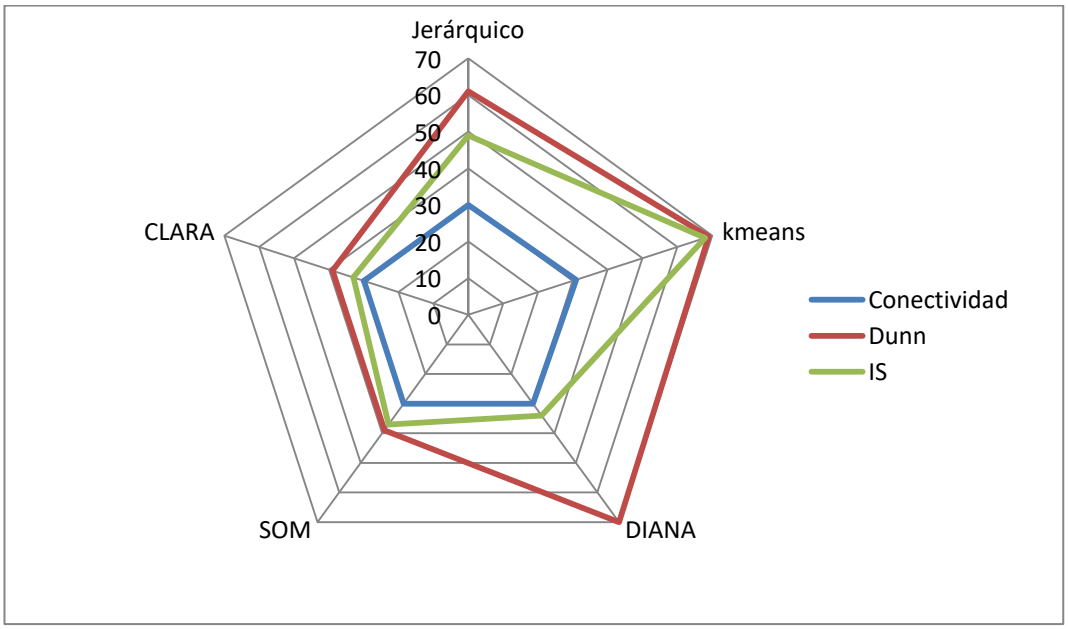


Ilustración 34: Número de clústeres obtenidos por *ClValid*. Medidas de validación interna

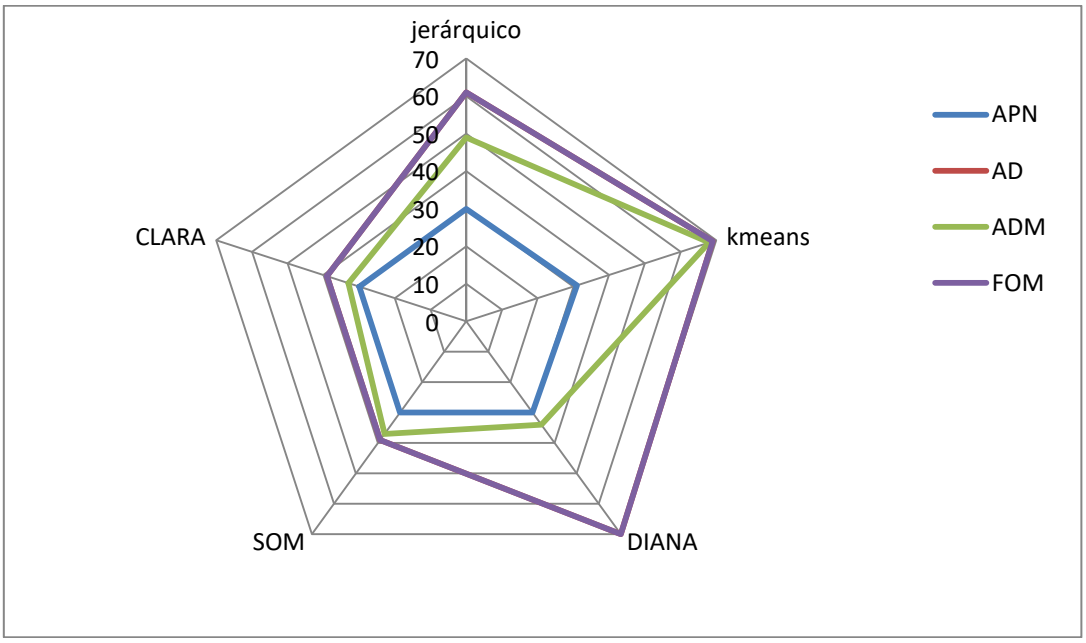


Ilustración 35: Número de clústeres obtenidos por *ClValid*. Medidas de validación de estabilidad

Para ambos casos se observa un número de clústeres entre 30 y 70. De hecho, para ambos tipos de medidas se obtienen, en general, dos franjas de resultados, 30-37 y 60-70.

Por términos de practicidad se hace la partición en 70 clústeres para las combinaciones anteriormente descritas, ya que así se reduce la probabilidad de obtener sectores no conectados, tal como se muestra en el doble recuadro de la Ilustración 36, en donde aparece una comunidad con elementos desconectados entre sí. Vale la pena señalar que este problema se puede solventar mediante un proceso de separación y re-enumeración de todos los componentes del grafo. En este proceso, se eliminan temporalmente todas conexiones existentes entre nodos que pertenecen a distintos sectores y se asigna un nuevo id a cada uno de los componentes desconectados. Dicha asignación se realiza mediante la ejecución del algoritmo BFS sobre cada uno de los nodos, así se encuentran todos los elementos conectados y a cada uno de los nodos en cada grupo se le reasigna el mismo id (ver Ilustración 37).

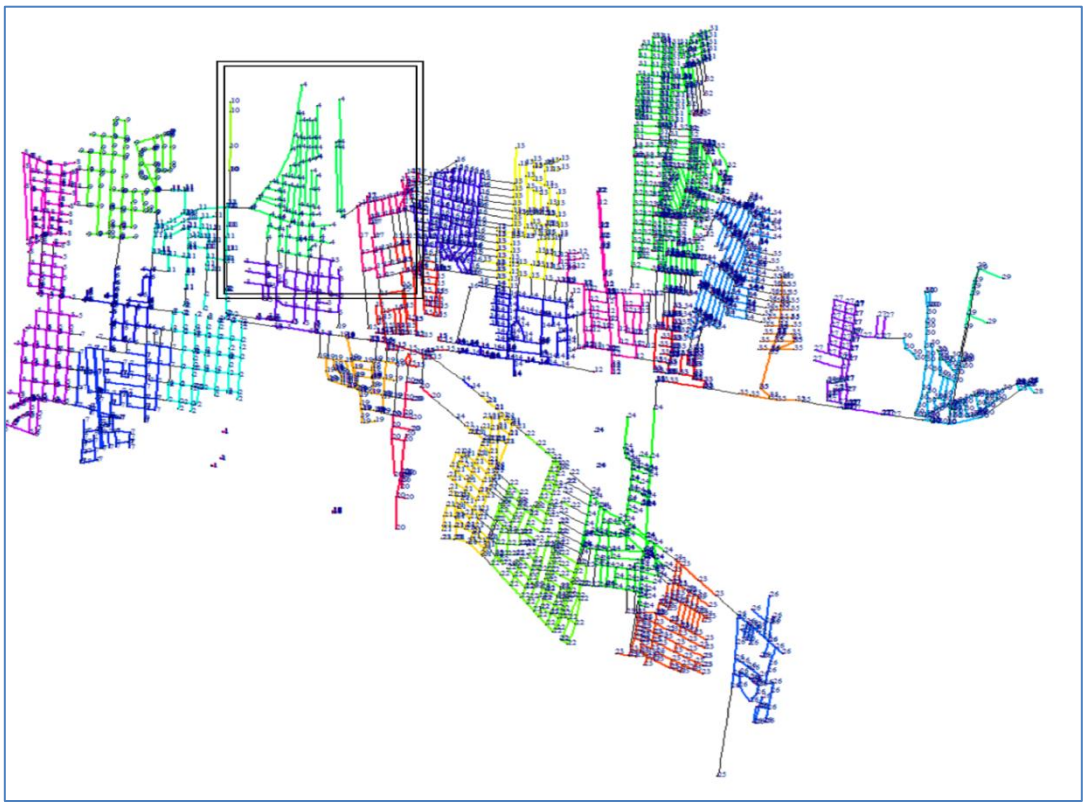


Ilustración 36: Partición en 30 clústeres (comunidades desconectadas)

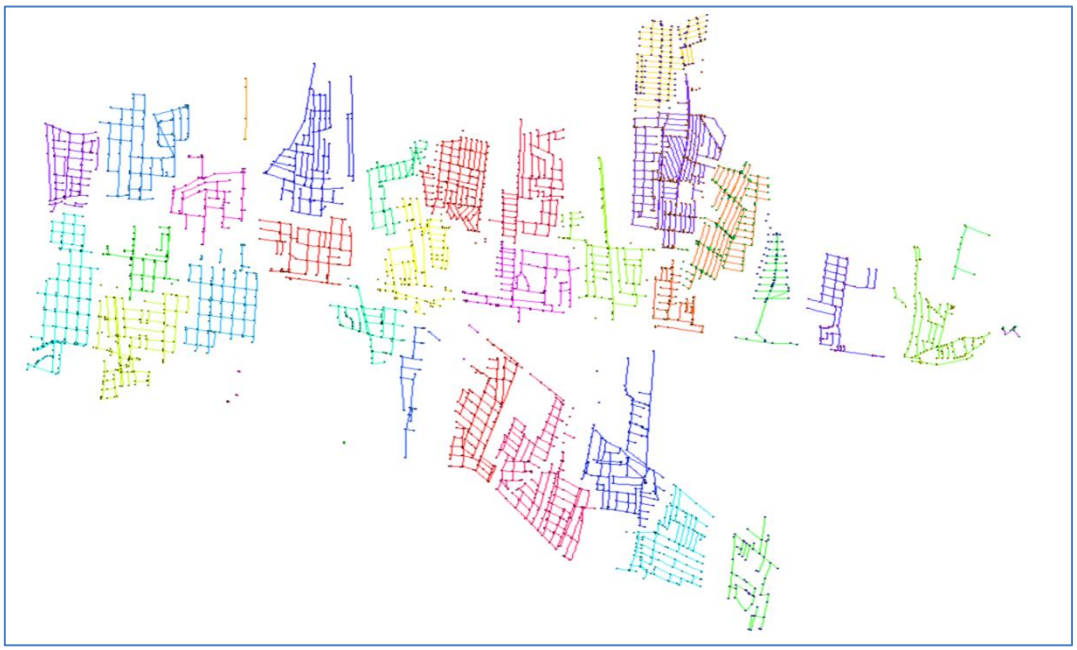


Ilustración 37: Componentes desconectados de la partición en 30 clústeres

La Ilustración 38 muestra el resultado obtenido al realizar la partición en 70 clústeres sobre el dendograma generado a partir de la combinación métrica Euclidiana con método promedio.

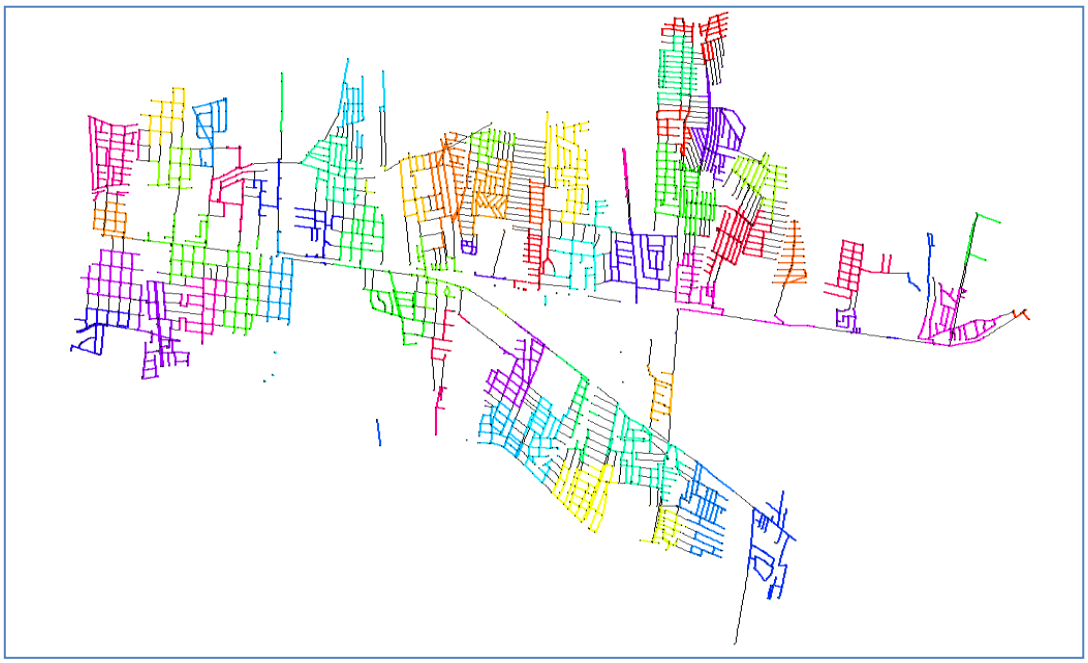


Ilustración 38: Partición en 70 clústeres obtenida mediante la métrica Euclidiana y el método promedio

La Ilustración 39 muestra el resultado obtenido al realizar la partición en 70 clústeres sobre el dendograma generado a partir de la combinación métrica Euclidiana con método centroide.

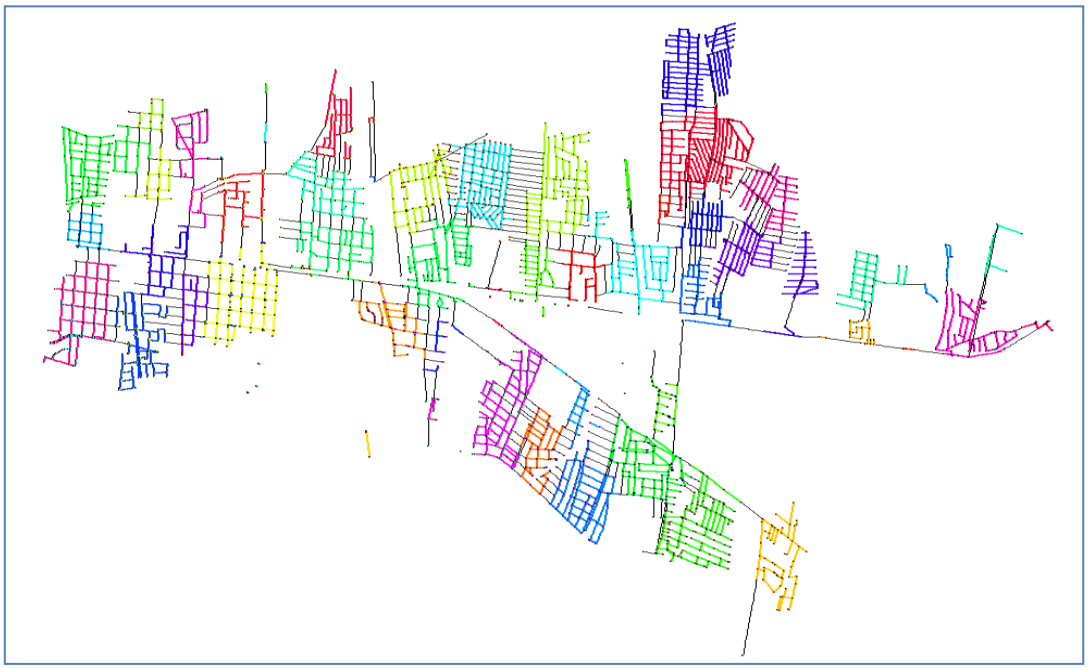


Ilustración 39: Partición en 70 clústeres métrica Euclidiana y método centroide

La Ilustración 40 muestra el resultado obtenido al realizar la partición en 10 clústeres sobre el dendograma generado a partir de la combinación métrica Gower con método centroide. En este caso, no se establecieron pesos para las variables al momento de calcular la matriz de disimilaridad. Como se puede ver, lo sectores resultantes cubren zonas con distintas cotas.

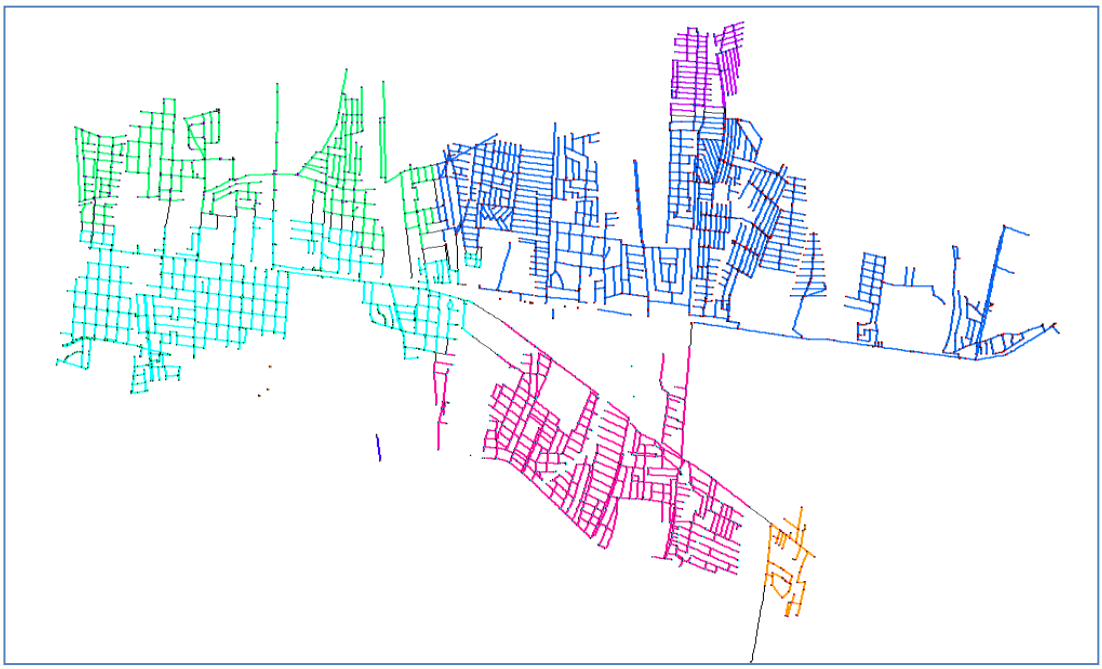


Ilustración 40: Partición en 70 clústeres métrica Gower (pesos iguales) y método centroide

El resultado varía significativamente cuando se da mayor peso al vector de cotas (ver Ilustración 41), se puede notar como las zonas que son definidas como un único sector en el caso previo, ahora se dividen en subsectores.

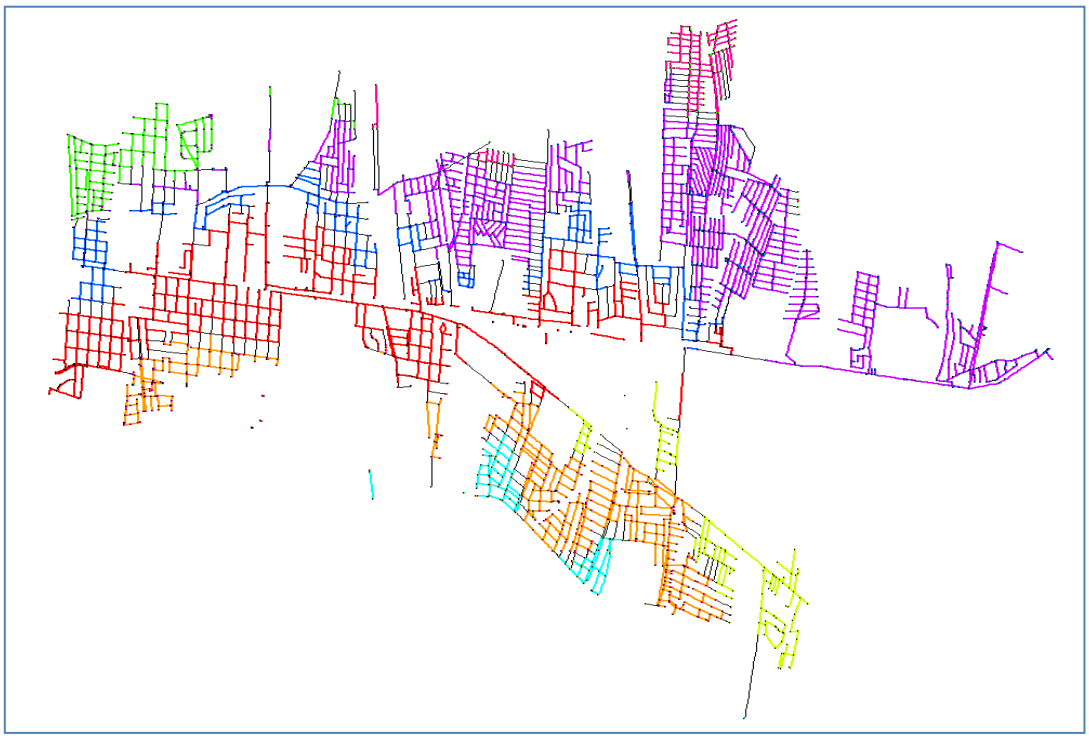


Ilustración 41: Partición en cuatro clústeres asignando el máximo peso a la cota

Para ejemplificar el proceso de re-fusión se empleó la partición en 70 sectores que generó el valor CCCF más alto (combinación distancia Euclidiana con método promedio). Al establecer como criterio de re-fusión la longitud de tubería (entre 30 y 3 km de longitud de tubería), se obtuvo una configuración de 12 sectores tal como se muestra en la Ilustración 42. En la misma figura, las líneas marcadas como "Tuberías Candidatas" constituyen el conjunto de tuberías que pueden ser definidas ya sea como límites o como entradas de sector, en el proceso de optimización que se explica más adelante. Al reducir el criterio de longitud mínima de 3 km a 1.5 km, el número de sectores aumentó de 12 a 16 (ver Ilustración 43).

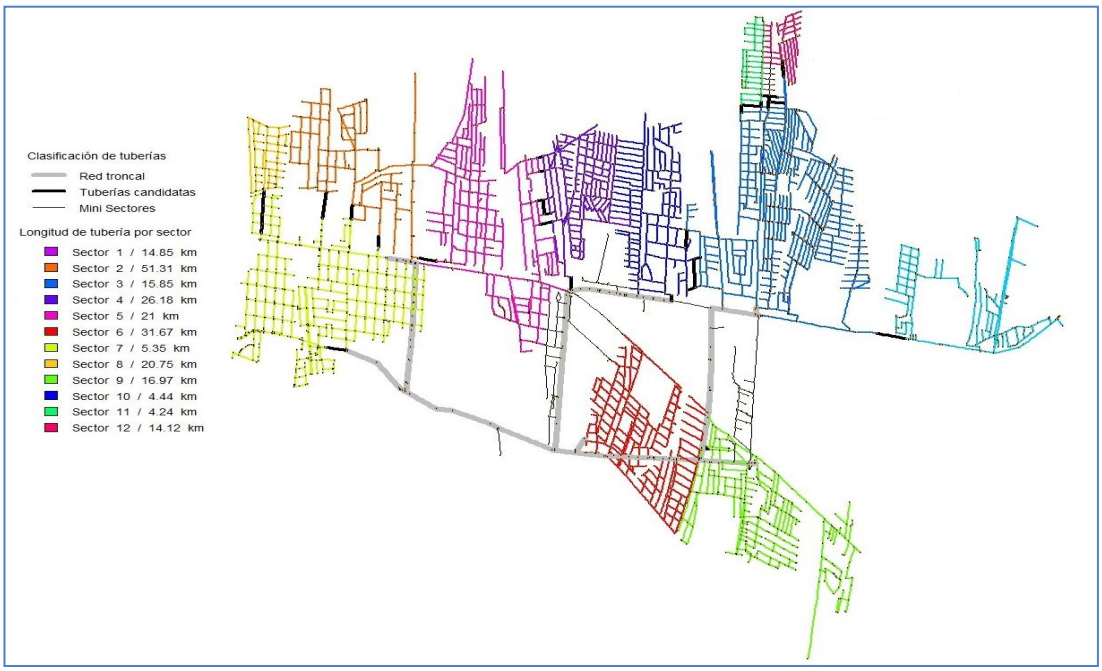


Ilustración 42: Configuración final de sectorización estableciendo como restricción la longitud de tubería (30 km – 3 km) en el proceso de re-fusión

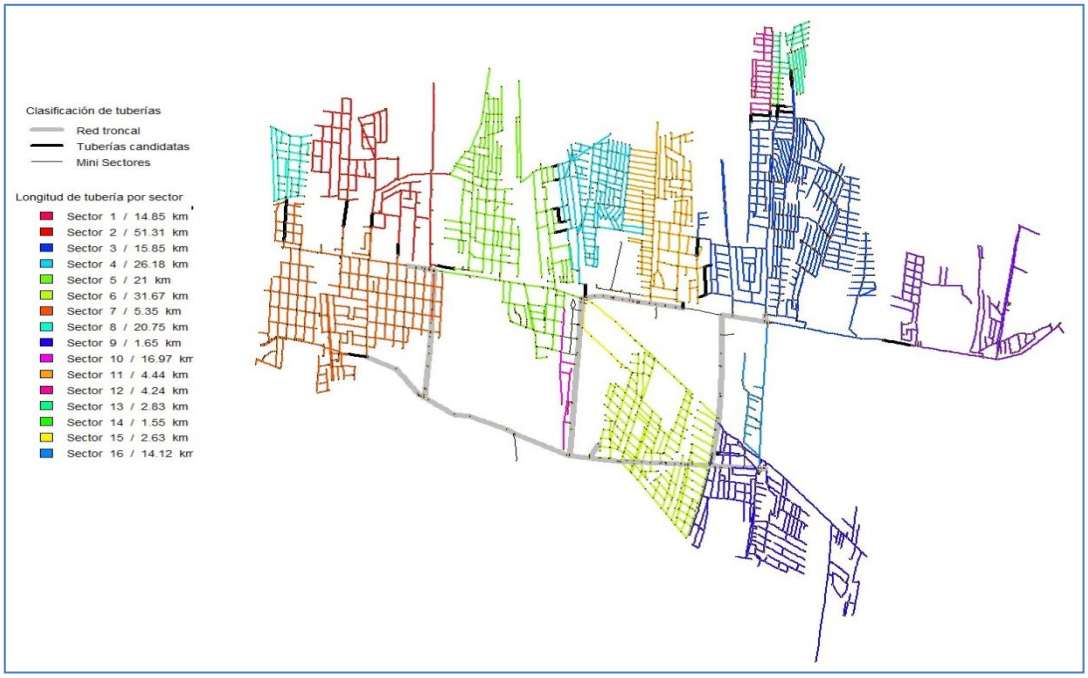


Ilustración 43: Configuración final de sectorización estableciendo como restricción la longitud de tubería (30 km – 1.5 km) en el proceso de re-fusión

Al establecer como criterio de re-fusión la cota (10 m), se obtuvo una configuración de 10 sectores tal como se muestra en la Ilustración 44, y al reducir el criterio de cota de 10 m a 5 m, el número de sectores aumentó de 10 a 15 (ver Ilustración 45).

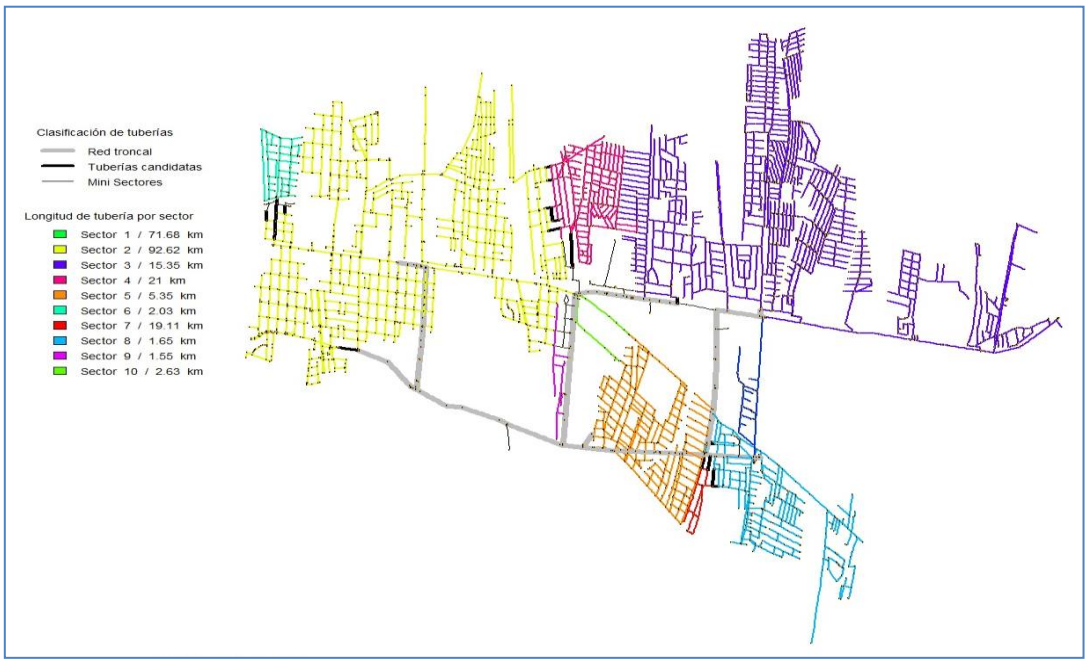


Ilustración 44: Configuración final de sectorización estableciendo como restricción la cota (10 m) en el proceso de re-fusión

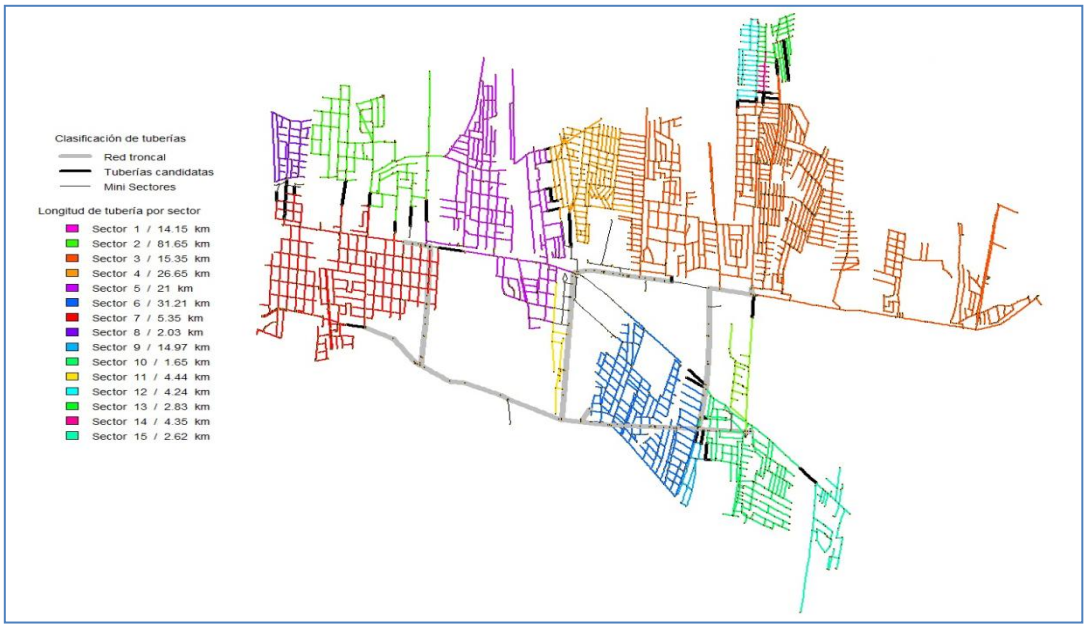


Ilustración 45: Configuración final de sectorización estableciendo como restricción la cota (5 m) en el proceso de re-fusión

II.6.4 Conclusiones sobre la Definición de Sectores mediante Clústering Jerárquico

Tal y como fue descrito en la presente subsección, la técnica clústering jerárquico destaca por su sencillez; sin embargo, cuando de generar sectores en RDAPs se trata, los resultados pueden no ser necesariamente adecuados, ya que al basarse únicamente en información almacenada en los nodos, se pierde el sentido de conectividad de la red. Lo anterior se traduce, en algunos casos, en particiones con nodos desconectados, lo cual en una RDAP no es factible. Si bien, este problema se puede resolver con los procesos de detección de todos los subgrupos y posterior proceso de re-fusión acá propuesto, esto representa un esfuerzo de cálculo extra que no es necesario en otros métodos que se abordarán más adelante. Por otro lado, tal y como se ha visto, la obtención de una buena partición, está sujeta a encontrar la correcta combinación de métrica con método de agrupación, lo cual también representa un trabajo adicional tampoco requerido en otros métodos de detección de comunidades.

Pese a lo anterior, el hecho de que existan tantas posibles combinaciones entre métrica y método de agrupación brinda un marco de flexibilidad para lidiar con datos de distinta naturaleza, dependiendo de la información que se puede utilizar a la hora de sectorizar RDAPs (e.g., el tipo de datos que corresponde a las coordenadas geográficas es muy distinto al tipo de datos que corresponde ya sea a la demanda o la cota en los nodos). Especialmente ventajosa es la métrica Gower, que está pensada para lidiar con este tipo de problema (variables de distinta naturaleza). Tal como se puede ver en los resultados, la inclusión de pesos puede conllevar a modificaciones importantes del resultado final. Pese a esto, también se tiene que tener en consideración que en las RDAPs no existen muchas variables que se pueden emplear como criterios de clústering. De hecho, las variables que se pueden utilizar son: coordenadas geográficas, cota y demandas. De estas cuatro variables, únicamente las coordenadas geográficas y la cota son buenas variables de clusterización, ya que la asignación de la demanda es bastante subjetiva. De manera tal, que la versatilidad que ofrece la técnica se podría aprovechar de mejor manera

si se incluyeran más variables, lo cual se puede establecer como una recomendación para trabajos futuros.

II.7 Método de Sectorización basado en Detección Multinivel de Comunidades en Redes Sociales

El algoritmo de detección de comunidades Multinivel (Blondel *et al*., 2008), también conocido como método Louvain, es un algoritmo no supervisado desarrollado en el Departamento de Ingeniería Matemática de la Universidad Católica de Lovaina (Louvain), Bélgica, en el año 2008. El mismo funciona en dos fases, las cuales, se repiten iterativamente (ver Ilustración 43). En un principio, a cada nodo se le asigna una comunidad individual, de manera tal que en la inicialización del algoritmo hay tantas comunidades como número de nodos. A continuación se evalúa la ganancia en el valor de Modularidad, que se obtiene mediante el traslado del nodo i a la comunidad del nodo j. Dicha ganancia es computada mediante la Ecuación 27.

$$\Delta Q = \left[\frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_{tot} + k_i}{2m}\right)^2\right] - \left[\left(\frac{\sum_{in}}{2m}\right) - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \quad \text{(Ecuación 27)}$$

En esta ecuación, para una comunidad *C*, $\sum_{in}$ es la suma de los pesos de los enlaces dentro de *C*, $\sum_{tot}$ es la suma de los pesos de los enlaces incidentes en los nodos en la comunidad *C*; k_i es la suma de los pesos de los enlaces incidentes al nodo i; $k_{i,in}$ es la suma de los pesos de los enlaces i a los nodos en *C* y m es la suma de los pesos de todos los enlaces en la red.

En caso de que no se logre ninguna ganancia sobre el valor de Modularidad, el nodo i retorna a su posición original. Este proceso se aplica de manera repetida y secuencial para todos los nodos, hasta que ya no se logre más ganancia en el valor del parámetro en cuestión. Nótese que, en el proceso, cada nodo puede ser seleccionado más de una vez.

En la segunda parte del algoritmo, se crea una nueva red cuyos nodos corresponden a las comunidades encontradas en la primera fase. El valor de peso en los nuevos enlaces es igual a la suma en todos los enlaces que conectan a dos comunidades. Los enlaces que parten de y llegan a la misma comunidad son establecidos como auto-bucles. Nuevamente, el proceso se repite hasta que no se logre más ganancia en el valor del índice de Modularidad.

Entre las ventajas de este algoritmo se destacan, la sencillez de su implementación, la facilidad con que se puede calcular la ganancia en la Modularidad (Ecuación 27), y el hecho de que el número de comunidades decrece en cada paso, lo que hace que a cada paso, su velocidad aumente. La última ventaja también implica que a cada a paso, el algoritmo, se aleja más del problema de resolución detallado en la Subsección II.2.1.

En términos comparativos, los resultados encontrados por Aynaud *et al*. (2013), destacan las ventajas del método sobre otros métodos de detección de comunidades también basados en la maximización del índice de Modularidad, tal como es el caso del método propuesto por Clauset *et al*. (2004), el método propuesto por Pons & Lapatty (2005) (que se aborda en la siguiente subsección), y el método propuesto por Wakita & Tsurumi (2007). Los criterios empleados para llevar a cabo esta comparación correspondieron a la velocidad y al valor de modularidad alcanzado. Una de las redes que se empleó correspondió a la ampliamente conocida Red del Club de Karate Zachary (ver Ilustración 46).

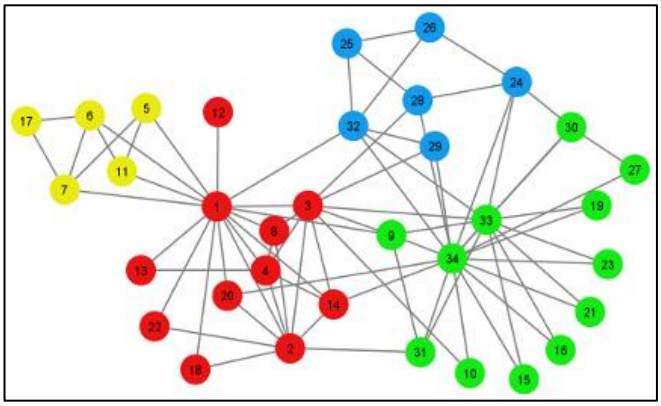


Ilustración 46: Red club de Karate Zachary [Fuente: (Zachary, 1977)]

Desde la publicación del algoritmo en 2008, se han publicado algunas variaciones, por ejemplo, *El Algoritmo de Detección de Comunidades Multinivel refinado Louvain* propuesto por (Rotta & Noack, 2011). Este extiende el algoritmo multinivel mediante un procedimiento de refinamiento en la primera fase del algoritmo.

II.7.1 Ejemplo de Implementación de Sectorización con base en el Método de Detección de Comunidades Multinivel

Para el ejemplo de implementación se parte de la red mostrada en la Ilustración 26, en la que para la selección de la red troncal se utilizó el valor de 10 como criterio de selección. En la Ilustración 47 se muestra el resultado generado por el algoritmo. En este caso se generaron 71 comunidades completamente conectadas (no se encontraron comunidades desconectadas). Una de las desventajas que conlleva la implementación del algoritmo Multinivel en *R* es el hecho de que no se puede extraer la jerarquía de comunidades, y a partir de ahí, seleccionar una partición en comunidades u otra. En este caso, se tiene que partir de la partición en comunidades en la que es máximo el índice de Modularidad.

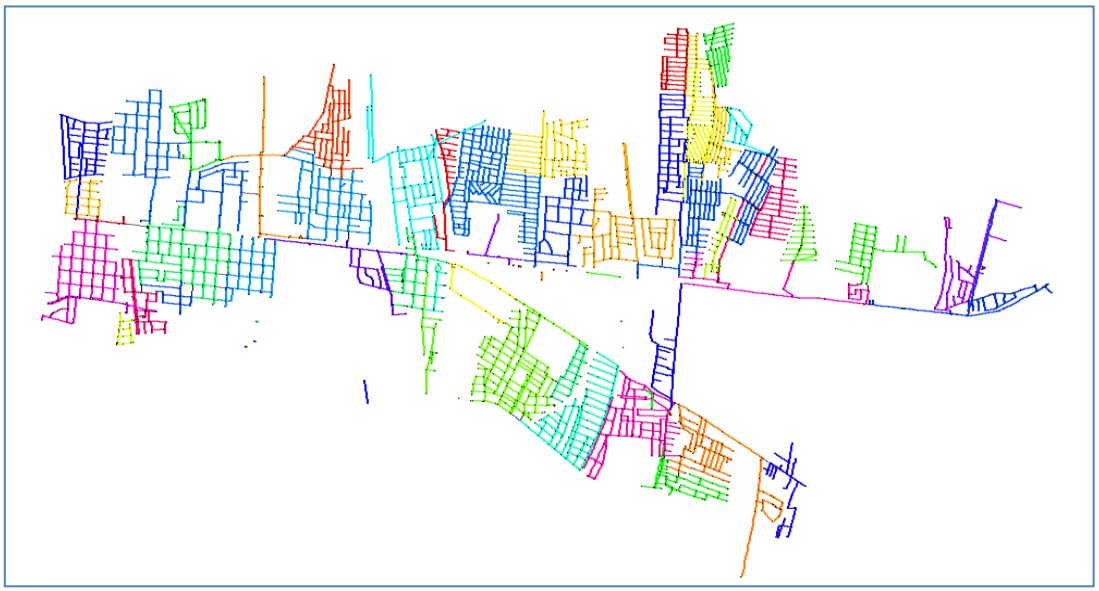


Ilustración 47: Comunidades (71) generadas por el algoritmo Multinivel

El proceso de re-fusión se llevó a cabo estableciendo como restricciones la longitud de tubería y la cota en los nodos. En el primero de los casos se definió una longitud de tubería admisible en el rango 30 km – 3 km y el rango 30 km – 1.5 km. En el segundo de los casos, se establecieron dos distintos valores admisibles de diferencia de cota, 10 y 5 m.

Tal y como se muestra en la Ilustración 48, cuando se establece la longitud de tubería como criterio estableciendo el rango 30 km – 3 km, se obtuvieron 9 sectores.

Ilustración 48: Sectores definidos tras el proceso de re-fusión estableciendo la longitud de tubería como criterio (30 km - 3 km)

Tal como se muestra en la Ilustración 49, al reducir la horquilla de longitud de tubería admisible (estableciendo un valor de longitud mínima de tubería igual a 1.5 km), el número de sectores generados aumenta de 9 a 12, lo que quiere decir que tres de los sectores que en el caso anterior eran establecidos como mini-sectores pasan a ser establecidos como sectores.

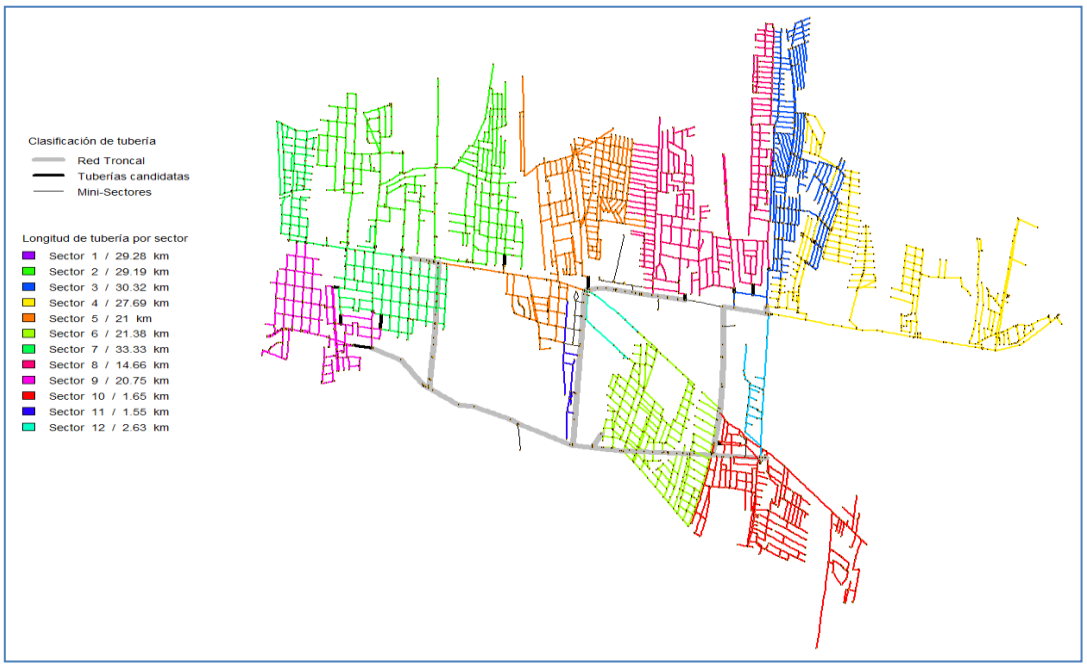


Ilustración 49: Sectores definidos tras el proceso de re-fusión al reducir el valor de longitud de tubería mínima

Al cambiar el criterio de fusión de sectores de longitud de tubería a cota en los nodos (10 m como límite de cota) (ver Ilustración 50), se obtiene un número de sectores igual a 11, lo cual indica que la topografía del terreno es bastante uniforme. Cuanto más irregular sea la topografía del terreno, mayor es el número de sectores a esperar cuando se emplea este criterio.

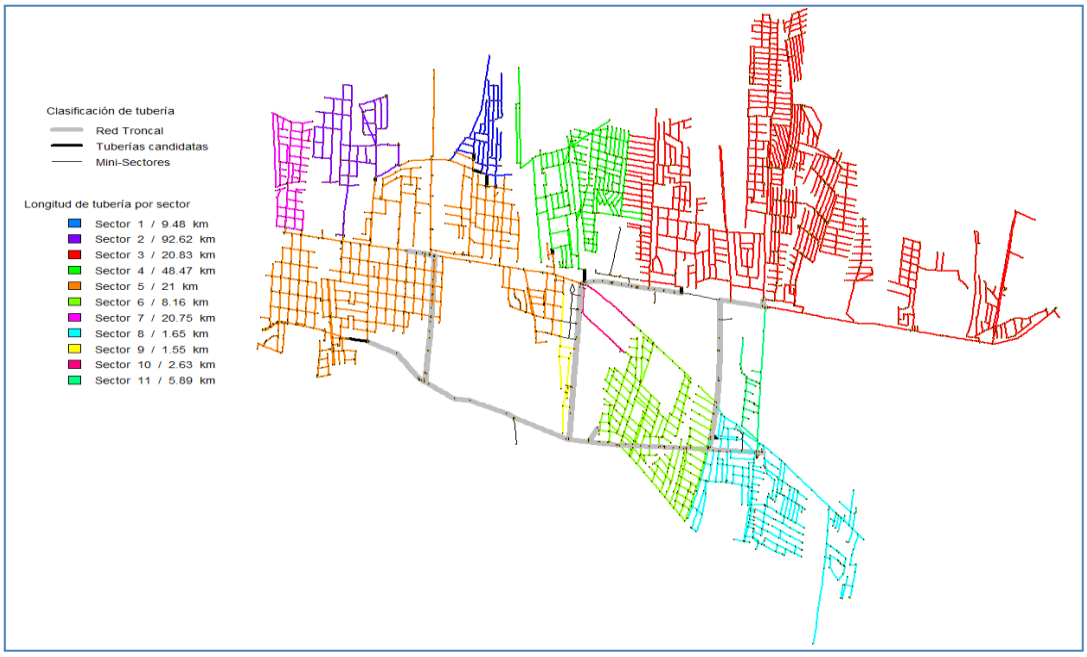


Ilustración 50: Sectores generados estableciendo como criterio de fusión la cota (10 m)

En la Ilustración 51 se observa el resultado final al reducir el criterio cota de 10 m a 5 m. En este caso se aumentó el número de sectores de 11 a 22.

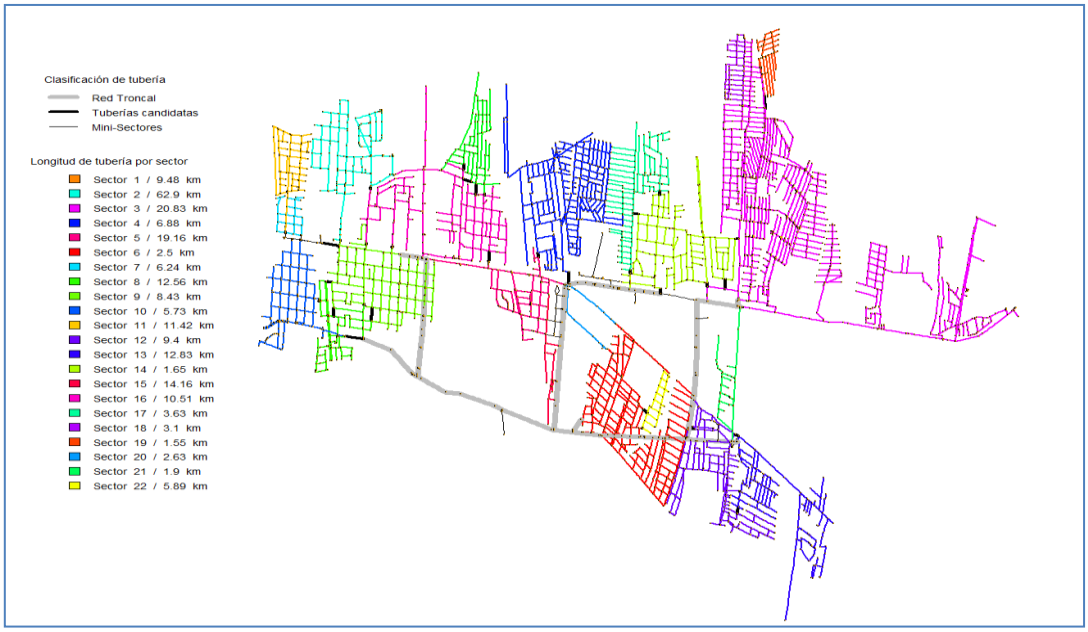


Ilustración 51: Sectores finales generados al reducir el criterio de fusión de 10 m a 5 m

II.7.2 Conclusiones sobre la Definición de Sectores con base en el Algoritmo de Detección de Comunidades Multinivel

En la presente subsección se ha visto la factibilidad del algoritmo Multinivel para definir esquemas de sectorización de RDAPs. La primera ventaja se relaciona con el hecho de que en términos de eficiencia y eficacia se trata de uno de los mejores algoritmos para detectar comunidades en redes sociales. Tal como se ha visto, su implementación tiene como ventaja la practicidad, ya que para la misma no se tienen que definir ningún tipo de parámetro; no obstante, su implementación en *R* tiene la desventaja, con respecto a otros algoritmos de detección de comunidades, de no permitir extraer la jerarquía de comunidades que se genera. La única alternativa es obtener la partición con máxima modularidad, perdiéndose flexibilidad en la implementación, de ahí que sería interesante en el futuro poder abordar este aspecto. Otra de las tareas que sería interesante llevar a cabo es evaluar el efecto que tiene el definir pesos en las tuberías al momento de implementar el algoritmo. Una ventaja adicional del algoritmo es que, a diferencia del clústering

jerárquico, siempre se encuentra comunidades conectadas, con lo cual no se hace necesario implementar la reasignación de *id* que se tiene que hacer en el caso del clústering jerárquico. Los resultados muestran que la técnica *per se* logra reconocer las características propias de cada zona de la red, estableciendo separaciones adecuadas sin necesidad de establecer una serie de características/criterios para tal fin. No obstante, sería interesante evaluar las mejoras que incluiría la inclusión de características dentro del proceso de clústering. Esto, naturalmente, conllevaría un proceso de reedición del algoritmo, lo cual se excluyó del alcance de este trabajo. Sin embargo, se puede establecer como recomendación para futuros trabajos. En la implementación del algoritmo, se hizo notoria la mejoría, en términos de límite de resolución, que tiene el algoritmo con respecto a otros algoritmos de detección de comunidades. En concreto, esto se puede visualizar por la presencia de comunidades de gran extensión conectando comunidades de menor extensión. De acuerdo al principio del problema de resolución, no se encontrarían comunidades pequeñas. Nótese que la partición en la que se maximiza el índice de Modularidad, corresponde a una partición en 71 sectores, la cual se encuentra dentro del rango de número de sectores encontrados por las medidas internas y de estabilidad por medio de *Clvalid* en el ejemplo de implementación de clústering jerárquico.

II.8 Método de Sectorización basada en la Detección de Comunidades Mediante Caminos Aleatorios

II.8.1 Nociones Básicas de Caminos Aleatorios

El concepto de caminos aleatorios (*Random Walks*) fue introducido a principios del siglo 20 por el matemático Carl Pearson (Pearson, 1905). En el campo de las matemáticas y la teoría de probabilidad, los caminos aleatorios son definidos como procesos estocásticos en donde la posición de una partícula dada en un cierto instante depende sólo de su posición en un instante previo y de una variable aleatoria, la cual determina su dirección subsecuente y la longitud de paso. Un camino aleatorio puede ser matemáticamente modelado mediante la Ecuación 28.

$$X(t + \tau) = X(t) + \Phi(\tau) \qquad \text{(Ecuación 28)}$$

En esta ecuación, Φ representa una variable aleatoria que describe la ley de probabilidad para tomar el siguiente paso y τ es el intervalo de tiempo entre pasos subsecuentes. Dado que la longitud y dirección de un paso dado dependen sólo de la posición $X(t)$ y no de alguna posición previa, se dice que el camino aleatorio posee la propiedad de *Markov* también conocida como la propiedad de la *Cadena de Markov* (*Markov Chain*). Una cadena de *Markov* es un sistema en un tiempo discreto, que en cada instante cambia de estado, entre un conjunto finito de estados, y en cada instante el cambio de un estado a otro depende de un conjunto de probabilidades dadas.

La matriz de transición es una matriz cuadrada cuyos elementos son las probabilidades de transición de un estado a otro:

$$P_{x,y} = [X_1 = y \mid X_0 = x] \qquad \text{(Ecuación 29)}$$

Esta matriz está caracterizada por el hecho de que todas las filas de la misma suman 1. La misma se rellena siguiendo las reglas que se presentan en la Ecuación 30.

$$P_{ij} = P(X_{k+1} = j \mid X_k = i) = \begin{cases} \dfrac{1}{d(i)}, & \text{si } ij \in E \\ O, \text{en cualquier otro caso} \end{cases} \qquad \text{(Ecuación 30)}$$

donde $d(i)$ corresponde al grado del nodo i.

II.8.2 Algoritmo Walktrap

El algoritmo Walktrap propuesto por Pons & Lapaty (2005) se basa en la idea de que si se ejecutan caminos aleatorios en un grafo, estos tienden a quedar atrapados en partes densamente conectadas, las cuales corresponden a comunidades. Si sobre un grafo dado se ejecutan caminos aleatorios, la información en la matriz de probabilidad o matriz *P* relativa a las probabilidades P_{ij}^{k} de ir desde un nodo i a un

nodo j en K pasos, es lo suficientemente grande como para obtener suficiente información sobre la topología de la red.

En este algoritmo se computa una medida de distancia (ver Ecuaciones 30-31) entre nodos y comunidades basada en la matriz P anteriormente descrita.

$$r_{ij}(t) = \sqrt{\frac{\sum_{k=1}^{n} (P_{ik}{}^{t} - P_{jk}{}^{t})^2}{d(k)}} \qquad \text{(Ecuación 30)}$$

$$r_{c_1 c_2} = \sqrt{\sum_{k=1}^{n} \frac{(P_{c_1 k}{}^{t} - P_{jk}^{t})^2}{d(k)}} \qquad \text{(Ecuación 31)}$$

En las Ecuaciones 30-31, $d(k)$ representa el grado de k.

Esta distancia es empleada para obtener la matriz de disimilaridad de todo el grafo. Con base en la comparación por parejas sobre la matriz de disimilaridad, se construye la primera comunidad. A fin de determinar qué comunidades fusionar, se emplea el previamente descrito método Ward (ver Subsección II.6.1.1 Pasos de Clústering Jerárquico Aglomerativo), tal y como lo describe la Ecuación 32. En este método las medias de las distancias al cuadrado entre nodos en cada sector es definida como función objetivo y la idea que se persigue es su minimización.

$$\sigma_k = \frac{1}{n} \sum_{C \in \mathcal{P}_k} \sum_{i \in C} r_{iC}^2 \qquad \text{(Ecuación 32)}$$

En la Ecuación 32, $\mathcal{P}_k$ corresponde a una partición k y r_i corresponde a la distancia arriba descrita.

Vale la pena destacar dos aspectos: en primer lugar, esta función sólo depende de cada clúster y su minimización no requiere información sobre otros sectores. En segundo lugar, el método sigue una estrategia *golosa*; por lo tanto, cada par de sectores adyacentes es fusionado y la variación $\Delta\sigma$ es calculada. La fusión que produce el menor valor de $\Delta\sigma$ es seleccionada como nueva partición.

A cada paso, el algoritmo forma una nueva partición que minimiza $\Delta\sigma$, de manera tal que al final no hay una única partición, sino una secuencia de particiones a diferentes niveles (los niveles del dendogramas). De acuerdo a Pons & Lapaty (2005), a partir de este conjunto de particiones, la mejor ajustada a los requerimientos puede ser seleccionada. Para este fin, los mismos autores proponen la llamada *Tasa de Incremento* (η_c, *Increase Ratio*) (ver Ecuación 33), la cual calcula la fracción que representa el $\Delta\sigma_k$ de una partición dada en relación a la partición previa $\Delta\sigma_{k-1}$. El mismo es evaluado para todos los niveles del dendrograma. El nivel con el valor más alto de η_c, es definido como la partición que mejor captura la estructura de comunidad.

$$\eta_c = \frac{\Delta\sigma_k}{\Delta\sigma_{k-1}} \qquad \text{(Ecuación 33)}$$

El algoritmo Walktrap tiene una serie de ventajas sobre otros algoritmos de detección de comunidades, estando la primera relacionada con la medida de distancia descrita arriba, la cual es fácilmente computable.

La estructura comunitaria se puede calcular en un tiempo $\mathcal{O}(mnH)$, donde m es el número de enlances, n es el número de nodos y H es la altura del dendrograma. En término comparativos respecto a otros algoritmos de detección de comunidades, las pruebas realizadas por Kesiban Orman *et al*. (2011) lo clasifican como el segundo mejor algoritmo de detección de comunidades, justo después del algoritmo *Infomap* (Rosvall & Bergstrom, 2008). Por otro lado, el resultado obtenido por Savić *et al.* (2012), al compararlo con otros tres algoritmos de detección de comunidades, a saber: *Newman-Girvan* (Newman & Girvan, 2004); *Label Propagation* (Raghavan,

2007) y *Greedy Modularity Optimization* (Clauset *et al.*, 2004), muestra que tiene ventajas significativas en términos de calidad y de estabilidad evolutiva.

II.8.3 Ejemplo de Implementación

A diferencia del algoritmo Multinivel, previamente discutido, en la implementación del presente algoritmo se debe definir un número de pasos que el algoritmo debe dar en cada camino. De manera general, es de esperar mayor calidad en los resultados en la medida que se incrementa el número de pasos en la implementación. Para comprobar esto, se registró el número de particiones en la que se maximiza el valor de modularidad para distintos números de iteraciones. Los resultados se presentan en la Ilustración 52. En la misma se aprecia que a partir de un número de iteraciones, el número de comunidades en que se maximiza la Modularidad se mantiene en el mismo rango (60-70). No obstante, la cantidad de sectores no garantiza igualdad topológica, aunque sí bastante similitud. Lo anterior se aprecia más claramente al comparar la Ilustración 53 y la Ilustración 54. En ella se muestran dos particiones en 65 comunidades, empleando en uno de los casos un número bajo de iteraciones (200) (ver Ilustración 53) y en otro de los casos, un número más alto (1500) (ver Ilustración 54). Como resultado se observa que para la partición obtenida con 1500 iteraciones, las comunidades cuentan con mejor definición (se encuentran bien diferenciadas unas respecto a las otras). En el primero de los casos, se aprecia cómo algunos de los nodos de algunas comunidades quedan rodeados por nodos pertenecientes a otras comunidades. Basado en ello, se opta por continuar el ejemplo con una partición en 65 comunidades empleando 1500 iteraciones. Lo anterior también puede ser respaldado mediante el valor de modularidad obtenido para ambas particiones: en el primero de los casos, se obtuvo un valor de 0.85, en tanto que en el segundo caso se obtuvo un valor de 0.97.

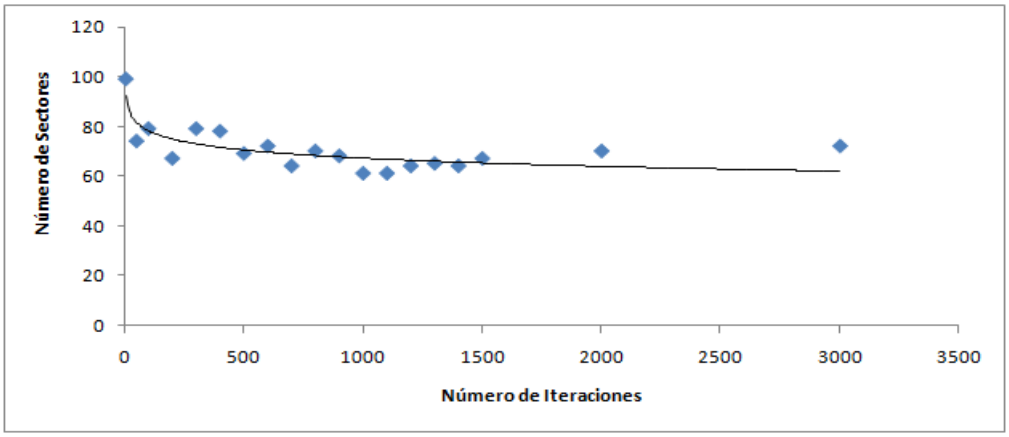


Ilustración 52: Número de sectores generados por el algoritmo Walktrap mediante distintos números de iteraciones

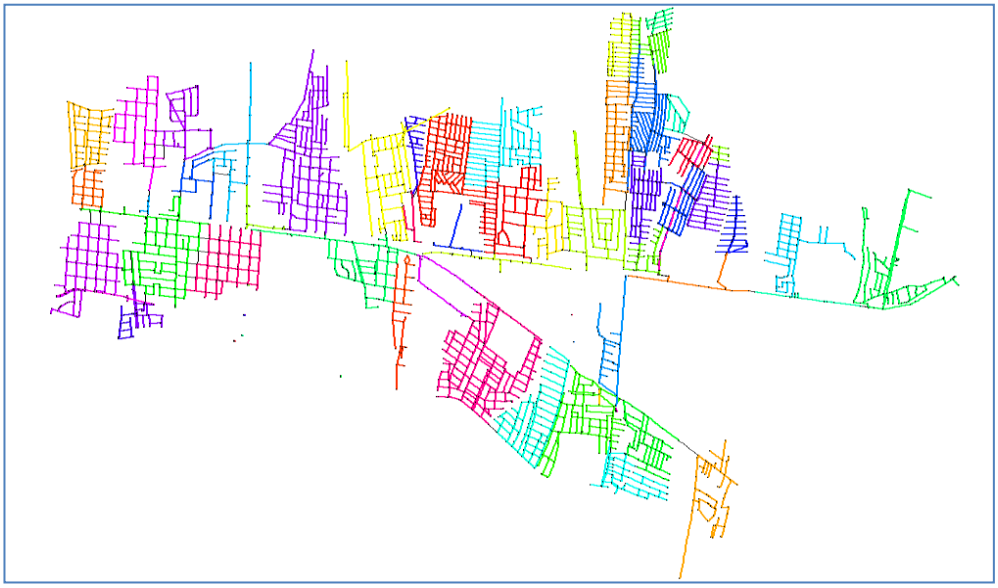


Ilustración 53: Partición obtenida usando un número de iteraciones igual a 200

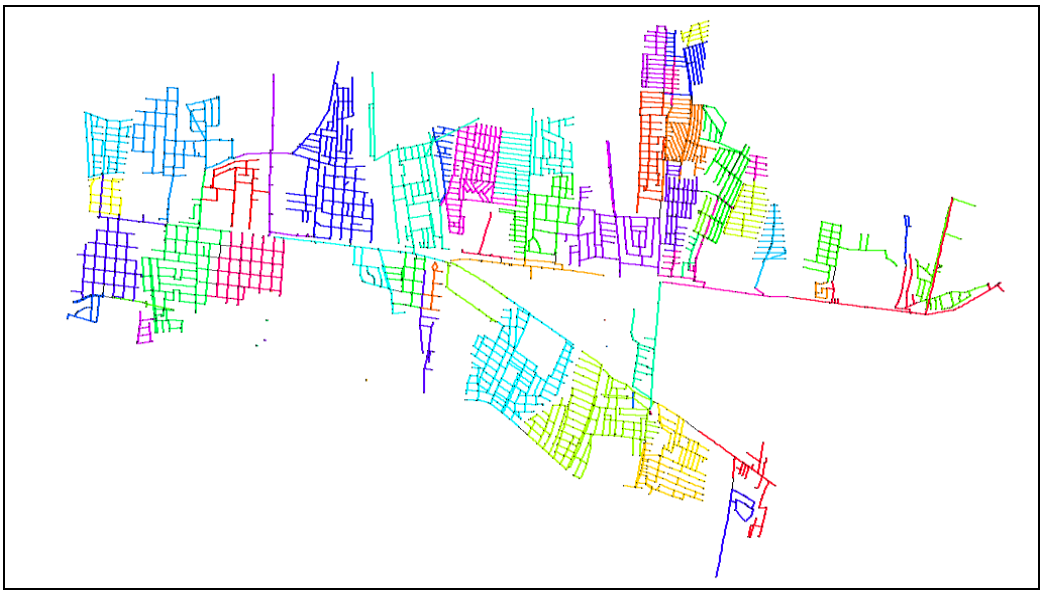


Ilustración 54: Partición obtenida utilizando un número de iteraciones igual a 1500

En el proceso de re-fusión empleando la longitud de tubería de sectores como criterio (30 km – 3 km), se obtuvieron 10 sectores en total (ver Ilustración 55). Luego, al reducir la restricción como longitud mínima de sectores a 1.5 km, se aumentó el número de sectores a 13 sectores (ver Ilustración 56)

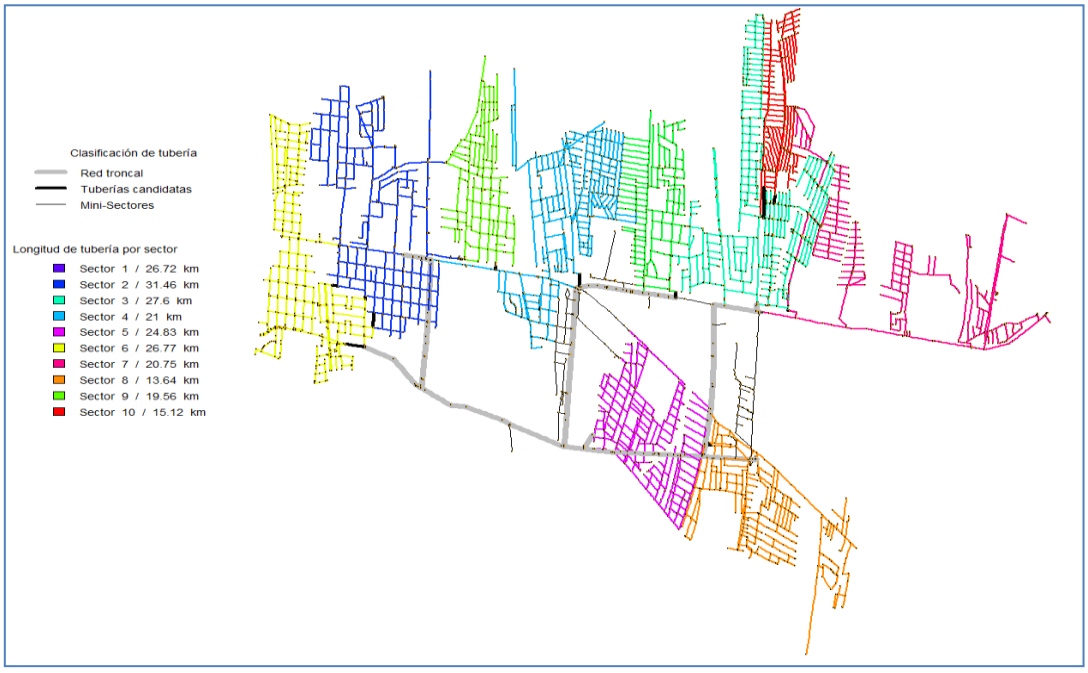


Ilustración 55: Sectores generados tras el proceso de re-fusión utilizando la longitud de tubería como criterio

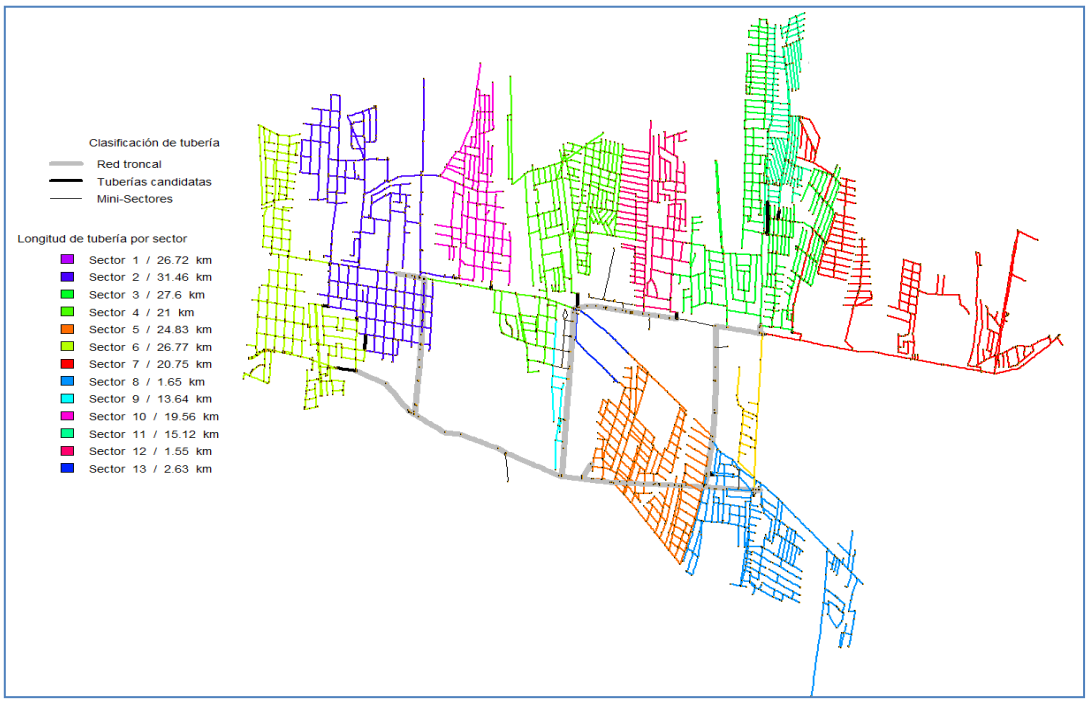


Ilustración 56: Sectores generados tras el proceso de re-fusión reduciendo el criterio de longitud mínima

Al cambiar el criterio de re-fusión de comunidades de longitud de tubería a cota, se redujo el número de sectores a 10 (ver Ilustración 57).

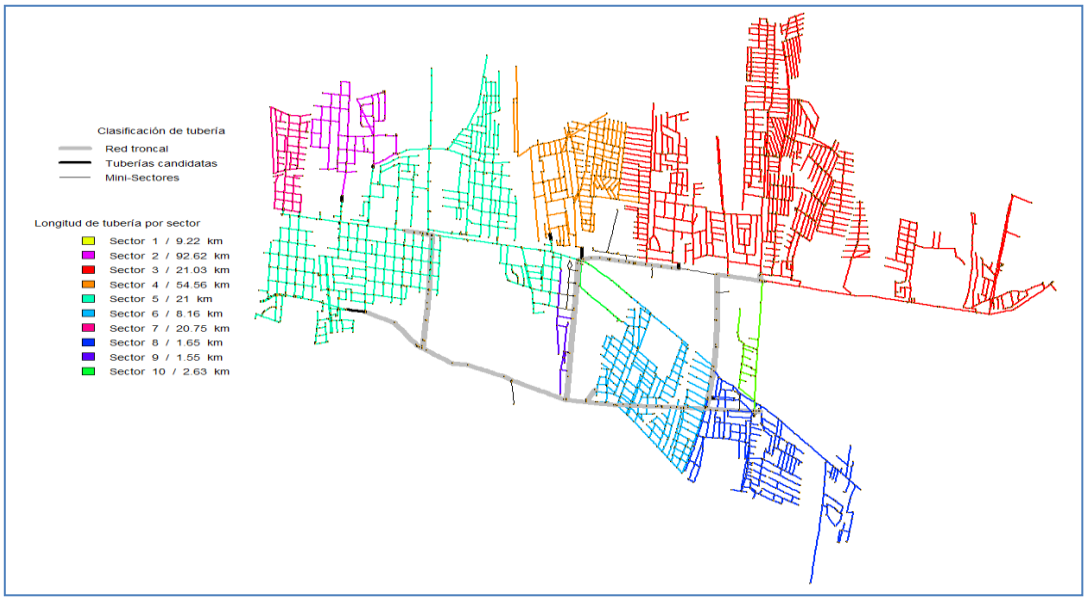


Ilustración 57: Sectores generados después del proceso de re-fusión empleando la cota como criterio

Al restringir el criterio de cota, el número de sectores aumentó de 10 a 22, lo que se debe al hecho de que el rango de cotas en el que se encuentran los nodos de la red es superior a 5 m (ver Ilustración 58).

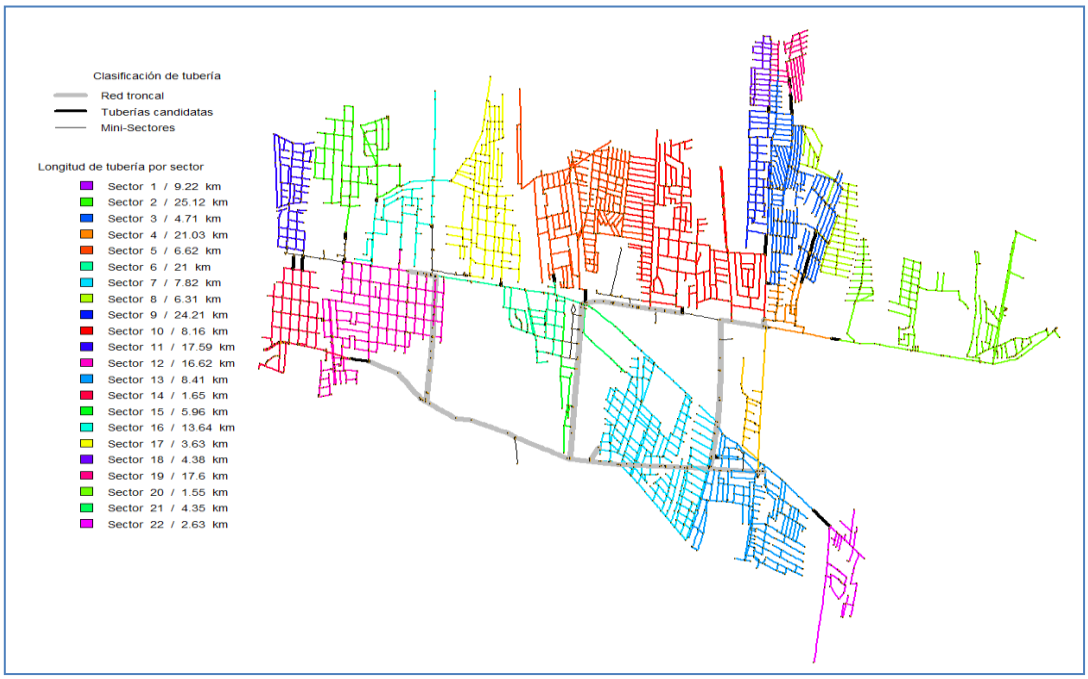


Ilustración 58: Sectores generados tras el proceso de re-fusión al reducir el criterio de elevación

II.8.4 Conclusiones Sobre la Definición de Sectores basada en el Algoritmo de Detección de Comunidades Walktrap

El método de definición de sectores basado en el algoritmo Walktrap ha demostrado ser completamente aplicable al problema de sectorización que se aborda en este trabajo. El mismo ofrece, entre sus ventajas, la capacidad de definir el número de iteraciones, lo cual permite mejorar el resultado final de la definición de sectores. Tal como se mostró en el ejemplo de implementación, si bien, después de cierto rango, el número de iteraciones deja de tener mucho efecto sobre el número de sectores en el que se maximiza el valor de modularidad, este sí puede tener efecto sobre la topología de las comunidades obtenidas. En el ejemplo de implementación, se mira que cuando se generan comunidades con un menor número de iteraciones, se encuentran tuberías pertenecientes a algunas comunidades dentro del área de otras comunidades.

La otra ventaja que tiene la implementación del algoritmo en *R* está relacionada con la capacidad de poder extraer una partición u otra del dendograma, lo cual puede

ayudar a reducir el tiempo del proceso de re-fusión de sectores; no obstante, siempre queda pendiente la pregunta sobre por dónde debe empezar el proceso de fusión.

El resultado en que se maximiza la modularidad correspondió a 65 sectores, que es muy similar al resultado obtenido en las dos técnicas previamente descritas. Al igual que en el caso del algoritmo Multinivel, el algoritmo logra encontrar las zonas de elevación de la red sin necesidad de definir características. No obstante, al igual que el caso anterior, sería interesante evaluar los beneficios que tendría la inclusión de pesos en el proceso de definición de comunidades.

II.9 Conclusiones sobre Métodos de Sectorización con base en Detección de Comunidades en Redes Sociales

En la Tabla 12 se muestra la comparación entre el número de comunidades generados por cada uno de los métodos de generación de sectores, estableciendo distintos criterios en el proceso de re-fusión de sectores. Tal como se puede ver, todos los métodos generan un número similar de sectores. A pesar de que el criterio de elevación sólo fue incluido de manera directa en el clústering jerárquico, el resultado encontrado por las otras dos técnicas de detección de comunidades establece como comunidades las mismas zonas. La diferencia estriba en que en la primera se tienen que establecer más pasos durante la ejecución para el análisis de número de sectores en que se tiene que hacer la partición, sin poder obtener un valor claro al final. Esto, sin tener en cuenta que el primer método no garantiza la continuidad de las comunidades detectadas (se pueden encontrar clústeres desconectados).

Tabla 12: Resumen de número de sectores obtenidos con cada uno de los métodos

Característica empleada en en el proceso de re-fusión	Clústering jerárquico	Algoritmo multinivel	Caminos aleatorios
30000 – 3000 (km de longitud de tubería)	14	9	10
30000 – 1500 (km de longitud de tubería)	16	12	13
10 (m de diferencia de cota)	10	11	10
5 (m de diferencia de cota)	19	22	22

Basados en lo anterior se puede establecer que, en términos de practicidad y velocidad, la definición de sectores basados en los métodos Multinivel y Caminos Aleatorios resulta más apropiada que la del método basado en clústering jerárquico. Respecto a la implementación en R, el algoritmo Walktrap presenta la ventaja de generar el dendrograma de todas las particiones y, además, de poder definir el número de iteraciones para la definición de comunidades. Esto permite reducir el número de pasos para la definición de sectores, ya que se puede seleccionar una partición con un menor número de sectores y/o se puede definir un número más bajo de iteraciones, lo que hace que el proceso de re-fusión de sectores se inicialice con un menor número de sectores y, por ende, que tome menor tiempo.

Respecto a las diferencias entre los criterios que se emplean para el proceso de re-fusión, como era de esperar, al emplear la cota, las particiones generadas resultan con mayor uniformidad para esa variable. Cuanto más estricto es el valor de cota empleado como criterio, mayor es el número de sectores y, a su vez, mayor la uniformidad. Esta variación de tamaño resulta muy interesante al momento de gestionar fugas en la redes; por ejemplo, redes con menores volúmenes de fugas puede ser gestionadas con sectores más grandes. En este caso el criterio de cota podría ser menos estricto. Por otro lado, en sectores con mayores niveles de fugas, sectores más pequeños permitirían mayor control sobre las fugas. En dicho caso, el criterio cota debería ser más estricto.

En el proceso de re-fusión presentado sólo se puede emplear un único criterio. Sin embargo, sería interesante que al final se pudiera obtener una solución de

compromiso entre dos o más criterios. De igual manera, sería interesante poder incluir los aspectos de fugas dentro de la definición de los sectores.

III. GESTIÓN DE PÉRDIDAS EN REDES DE ABASTECIMIENTO DE AGUA POTABLE MEDIANTE SECTORIZACIÓN: OPTIMIZACIÓN DEL CONJUNTO DE VÁLVULAS DE CIERRE/ENTRADAS DE SECTORES

III.1 Gestión Sostenible de Pérdidas en Redes de Abastecimiento de Agua Potable

Tal como ya ha sido mencionado en la Introducción, las pérdidas en las RDAPs constituyen uno de los aspectos más serios con el que tienen que lidiar las empresas gestoras de tales redes. Las pérdidas apuntan a mala gestión y a deterioro de la infraestructura, y, por tanto, implican la necesidad de incorporar un programa de gestión de las mismas ajustado a las características del sistema de abastecimiento en el que se trabaja. A pesar de esto, la eliminación completa de las pérdidas en una RDAP es completamente utópica y no factible. En cualquier sistema de abastecimiento de agua, siempre prevalecerá un porcentaje de pérdida, salvo que se inviertan cantidades extremadamente altas de recursos técnicos y financieros para evitarlas, lo cual haría que la operación del sistema fuera ineficiente.

Los problemas de las pérdidas de agua son de tipo técnico o de tipo financiero/económico. Los problemas de tipo técnico abarcan la imposibilidad de poder abastecer a todos los clientes, en tanto, los problemas financieros se refieren al hecho que no toda el agua que es consumida es facturada.

En 1991 fue establecida, por parte de la UKWIR, la Iniciativa Nacional de Fugas, con el fin de actualizar y revisar las directrices sobre el control de fugas que se habían utilizado hasta el año 1980. Esta iniciativa puso sobre la mesa una serie de técnicas encaminadas a que todas las empresas operadoras de RDAPs en Reino Unido pudiesen abordar el problema de las pérdidas de agua de una manera pragmática y directa. A este conjunto de técnicas se le asignó el nombre de BABE (*Burst and Background Estimates Methodology* o Metodología de Estimación de

Roturas y Fugas de Fondo) y fueron plasmadas en un conjunto de nueve reportes (UKWIR, 1994, SAWRC, 1999), a saber:

- Reporte A- Reporte Resumen
- Reporte B- Reporte de Comparación de Desempeño en Términos de Pérdidas
- Reporte C- Establecimiento de Metas Económicas de Fugas
- Reporte D- Estimación de Agua Inyectada y no Medida
- Reporte E- Interpretación de Medición de Mínimos Nocturnos
- Reporte F- Uso de Datos de Medición de Caudales Mínimos Nocturnos
- Reporte G- Gestión de Presiones
- Reporte H- Lidiando con Fugas en Puntos de Consumo
- Reporte J- Técnicas de Gestión de Fugas, Tecnología y Entrenamiento

En el marco BABE, se establece que la gestión de fugas en las RDAPs debe ser abordada mediante cuatro componentes principales: (1) aspectos económicos de control de fugas; (2) interpretación y uso de datos de caudales nocturnos; (3) gestión de presiones; (4) componentes anuales de las pérdidas y balance hídrico.

III.1.1 Balance Hídrico de acuerdo al Marco BABE

Un primer paso para una gestión correcta de cualquier sistema, incluyendo RDAPs, es el estructurar, lo más precisamente posible, un balance de sus entradas y salidas. En el caso de las RDAPs, este es conocido como balance hídrico y debe contemplar los volúmenes de agua inyectados en el sistema (incluyendo agua importada) y las distintas salidas, tales como: consumo, fugas, exportaciones, etc.

En la actualidad, el balance hídrico más ampliamente conocido es el que se presenta en el marco BABE, y que ha sido ya adoptado como un estándar internacional. Tal y como se puede ver en la Tabla 12, el mismo consiste en tres componentes

principales: consumo medido autorizado; consumo autorizado no medido y el resto de los consumos, que representa el volumen no facturado. Tal y como se mencionará a continuación, este volumen restante se subdivide en dos categorías: pérdidas reales (o pérdidas físicas) y pérdidas aparentes. Vale la pena destacar que en comparación con los balances hídricos tradicionales, este nuevo balance tiene como ventaja el hecho de subdividir los ítems en componentes más pequeños que pueden ser medidos o estimados.

Tabla 13: Modelo de balance hídrico establecido en el marco BABE

<table>
<tr><td rowspan="10">Volumen de entrada en el sistema (corregido en función de los errores conocidos)</td><td rowspan="5">Consumo autorizado</td><td rowspan="2">Consumo autorizado y facturado</td><td>Consumo facturado y medido (incluyendo agua exportada)</td><td rowspan="2">Agua con retorno financiero</td></tr>
<tr><td>Consumo facturado no medido</td></tr>
<tr><td rowspan="3">Consumo autorizado no facturado</td><td>Consumo no facturado medido</td><td rowspan="8">Agua sin retorno financiero</td></tr>
<tr><td>Consumo no medido ni facturado</td></tr>
<tr><td rowspan="5">Pérdidas de agua</td><td rowspan="2">Pérdidas aparentes</td><td>Consumo no autorizado</td></tr>
<tr><td>Imprecisiones en medidores de usuarios</td></tr>
<tr><td rowspan="3">Pérdidas reales</td><td>Fugas en líneas de conducción principal o líneas de distribución</td></tr>
<tr><td>Fugas o sobrellenando en tanques de almacenamiento</td></tr>
<tr><td>Fugas en conexiones de servicio hasta los contadores domésticos</td></tr>
</table>

Volumen de entrada: la entrada anual para una parte definida del sistema de abastecimiento de agua.

Consumo autorizado: el volumen anual medido y no medido que es extraído por los usuarios registrados, los abastecedores de agua y otros, implícita o

explícitamente autorizados para poder hacerlo. Incluye agua exportada, fugas y sobrellenado después de los contadores domiciliares, caudal de hidrantes, limpieza de calles, irrigación de jardines municipales, etc.

Pérdidas de agua: la diferencia entre el volumen de agua ingresado en el sistema y el consumo autorizado, conformado por pérdidas aparentes y pérdidas reales.

> **Pérdidas aparentes:** consumos no autorizados y todos los tipos de imprecisiones en las mediciones.
>
> **Pérdidas reales:** volúmenes anuales de pérdidas a través de todos los tipos de fugas, roturas y sobrellenados en tuberías de conducción principal, unidades de almacenamiento y red de distribución, hasta los medidores domésticos.

Agua sin retorno económico: La diferencia entre el ingreso de agua al sistema y el consumo facturado autorizado. Consiste en consumos autorizados no facturados (normalmente un componente muy pequeño del balance hidráulico) y pérdidas de agua.

La Ilustración 59, también ha sido planteada dentro del marco BABE y es ampliamente utilizada para describir el abordaje de las pérdidas reales en las RDAPs. En ella, se propone la gestión de pérdidas reales a través de cuatro componentes: (1) velocidad y calidad de reparaciones; (2) gestión de presiones; (3) control activo de fugas y (4) catastro y reemplazo de tuberías.

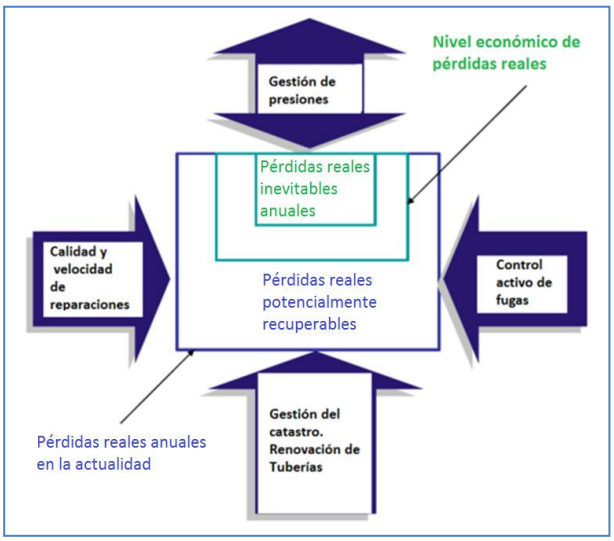


Ilustración 59: Frentes de acción para la gestión económica de las pérdidas reales en RDAPs [Fuente: basado en Lambert (2003)]

Entre el volumen inevitable de fugas (fugas de fondo) y el volumen total de fugas, existe un nivel económico de gestión, que corresponde al nivel en el que el coste de una nueva reparación excede al beneficio derivado de los ahorros. Este aspecto se refiere al punto de equilibrio entre el coste económico que representan las pérdidas por fugas y el coste de inversión necesario para su reparación, partiendo del hecho de que existe un Umbral Mínimo de Fugas (UMF) que no puede ser evitado. En la Ilustración 60 se puede apreciar de manera más clara este concepto. La curva de color negro representa el comportamiento del gasto por reparación y mantenimiento, que es inversamente proporcional al volumen de agua que se fuga, tal como se puede intuir. La curva de color azul representa el volumen de agua incontrolada, cuyo crecimiento es directamente proporcional a su coste. La curva de color rojo es la suma de las dos curvas anteriores, representado una curva de coste total. El punto más bajo de esta curva representa el nivel económico de fugas o el punto óptimo para la gestión del sistema. La línea gris vertical representa el ya mencionado UMF.

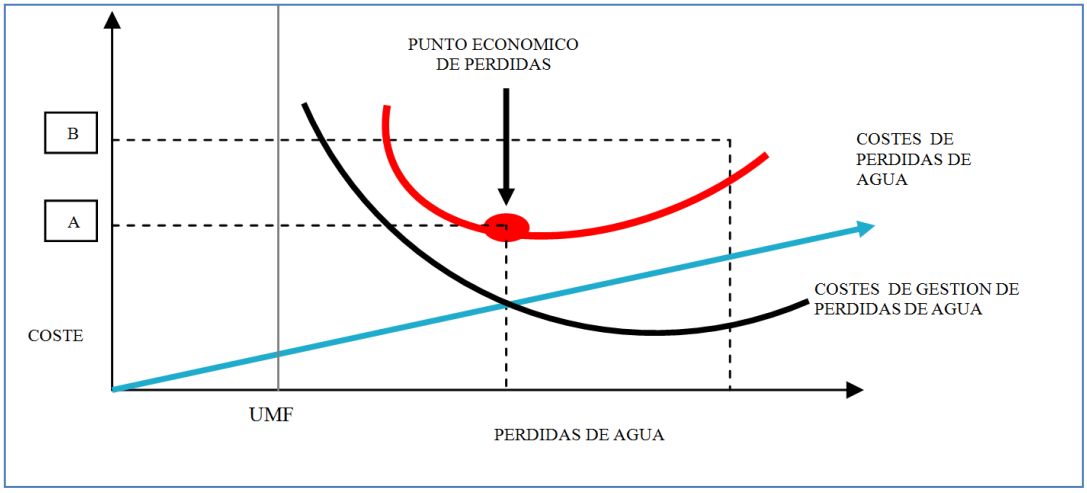


Ilustración 60: Punto económico de fugas [Fuente: basado en Morrison *et al.* (2007)]

III.1.2 Determinación de Caudales de Pérdidas Reales Siguiendo el Marco BABE

Con el fin de hacer un reconocimiento del problema de fugas, la técnica más eficaz consiste en realizar mediciones del caudal mínimo nocturno, que ocurre cuando el consumo es mínimo y las presión es más elevada, por lo que se espera que la mayor parte del caudal que esté entrando al sistema se deba a fugas. Por lo general, estas mediciones se realizan entre las 0 y las 4 horas. Una vez realizadas las mediciones, se debe distinguir el porcentaje del caudal que se debe a roturas, el porcentaje que se debe a fugas de fondo y el que se debe a consumo de los usuarios. Para ello, se debe partir asumiendo que el caudal mínimo nocturno está conformado por tres componentes: consumo nocturno legítimo, fugas de fondo y roturas en las tuberías.

En relación a los consumos legítimos, estos pueden ser de tres tipos: doméstico normal (piscinas, cisternas de inodoros, aspersores, etc.), no domésticos-bajos (bares, cafeterías, garajes), y grandes usuarios (fábricas, hospitales, aeropuertos, hoteles, etc.). Dada la gran cantidad de equipos que demandan agua por las horas de la noche, sería inviable intentar discriminar este consumo mediante mediciones. Por lo cual, el mismo se estima a través del uso de valores típicos tabulados en tablas.

En relación a las fugas de fondo, estas pueden ser divididas en tres categorías: fugas de fondo en las líneas de conducción principal; fugas de fondos en las conexiones y fugas de fondo en las instalaciones. El primer tipo de fugas se debe a fallos en las juntas de las tuberías, pequeñas roturas u hoyos en las tuberías y tal como se explicará más adelante, su magnitud depende de las condiciones en las que opera el sistema. El segundo tipo se ubica en las líneas que conectan las líneas de conducción principal con los usuarios, es decir, las tuberías entre las líneas de conducción principal y los contadores domésticos. Por lo general estas se originan por malas prácticas al momento de realizar las uniones entre las tuberías (fallos en las juntas). Vale la pena destacar que en muchos acueductos, esta es la principal fuente de fugas. El tercer tipo, se relaciona con los fallos (fallos en justas, desacoples) en las instalaciones de equipos ubicados aguas abajo de los contadores domésticos. Por lo general se espera que representen un volumen muy pequeño con respecto al volumen que pueden representar las fugas de fondo en las líneas de la conducción principal y en las líneas de distribución.

Dado que no todos los operadores de RDAPs cuentan con la capacidad técnica y económica para poder llevar a cabo las campañas de medición necesarias para poder contabilizar todos los ítems previamente descritos, se han tabulado algunos valores generales que se puede adaptar a las condiciones de cada RDAP. Por ejemplo, para el caso de las fugas de fondo en las líneas de conducción se establece un valor de 40 l/km de tuberías de conexión/h con un rango de variación del 50%; en el caso de las líneas de distribución, se recomienda un valor de 3 l/propiedad/hora, con un rango de variación también del 50%; finalmente, en el caso de las fugas en instalaciones se establece un valor de 1 l/propiedad/hora, siempre con el mismo rango de variación del 50% (Lambert *et al*. 1999; Lambert & McKenzie, 2002).

Una vez completado el cálculo de los consumos legítimos domésticos y no domésticos, y habiendo medido o estimado el volumen de fugas de fondo total, se puede discriminar el caudal que se fuga a través de las roturas de las tuberías, que equivale al volumen total mínimo nocturno, menos el resultado de sumar los primeros dos tipos de fugas.

III.1.3 Teoría FAVAD

Tradicionalmente se asumía que el caudal que salía por una rotura, variaba con el cuadrado de la variación de la presión; sin embargo, en muchas implementaciones, se hizo notorio que las variaciones de caudal respondían a exponentes más elevados. En torno a esto, May (1994) sugirió la posibilidad de existencia de Áreas de Descargas Fijas y Variables (o FAVAD por la siglas en inglés de *Fixed and Variable Area Discharge*). De acuerdo a esta teoría, los sistemas reaccionan de manera diferente a la presión dependiendo del tipo de fugas considerada. Si la fuga se debe a un hoyo producido por corrosión (e.g. en una tubería metálica), el tamaño de la abertura permanecerá fijo ante los ciclos diarios de variación de la presión. En tal caso, el agua perdida a través del hoyo seguirá el principio de la raíz cuadrada. Este tipo de fugas es conocido como *Fuga en Área Fija*. Por otra parte, si la fuga se produce en una junta, el tamaño de la apertura aumenta en la medida que la presión aumenta. En dicho caso, el caudal de fuga aumentaría más que en el caso de un área fija y por ende esta variación es representada con exponente superior a 0.5. En el caso de las fugas longitudinales, el área de fuga puede aumentar tanto en anchura así como en longitud, tal como sucede en las tuberías plásticas. En este caso el exponente tendría que ser aún mayor que en el caso anterior.

Dada la diversidad de posibles tipos de roturas, se ha adoptado, como esquema general, utilizar el valor 0.5 en el caso de fugas asociadas a roturas (ya que se espera que la mayor parte de estas ocurran en áreas fijas). En el caso de las fugas de fondo, se espera que las fugas se den a través de áreas variables, por lo cual se requeriría valores de exponentes más elevados. En este caso se suele utilizar un valor de 1.5, el cual se considera representativo de la variedad de fugas con exponentes entre 0.5 y 2.5. En caso de que la mayoría de las tuberías sean plásticas, se recomienda el uso de valores más altos y, por el contrario, en caso de que el material predominante sea rígido (por ejemplo hierro fundido) se recomienda el uso de valores más bajos.

La Ilustración 61 muestra gráficamente las variaciones de caudal de fugas en función de las variaciones de presión para diferentes valores del exponente N1 (tipo de fuga).

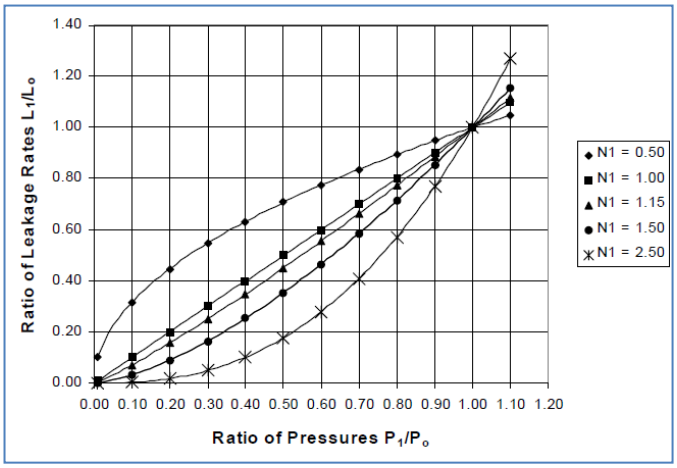


Ilustración 61: Variación de los caudales de fugas en función de la variación de presiones para distintos exponentes N1 [Fuente: Thornton & Lambert (2005)]

III.1.4 Estimación del Nivel Económico de Fugas a Corto Plazo

Hace más de una década, (Benvenuti *et al*., 2007; Fantozzi & Lambert, 2005, 2007; Lambert & Lalonde, 2005) propusieron un método para estimar el Nivel Económico de Fugas a Corto Plazo (NEFCP) con base en algunos conceptos de las metodologías BABE y la teoría FAVAD. Un resultado muy importante que se deriva de esta estimación es el Nivel Económico de Fugas no Reportadas (NEFNR), el cual corresponde al volumen de fugas no reportadas que es viable dejar ocurrir en un lapso de tiempo.

El método para estimar el NEFCP parte de hacer una división de los caudales de fugas de acuerdo a lo establecido en el marco BABE, tal y como se muestra en la Ilustración 62. En la misma ilustración se puede observar que, iniciando en un tiempo cero, las fugas de fondo se mantienen constante a lo largo del tiempo; las fugas reportadas se van presentando y reparando (barras verticales); en tanto, las fugas no reportadas, van aumentando paulatinamente hasta que llegan a un punto en que su coste se iguala al coste de inspeccionar toda la red. La pendiente de este

aumento se conoce como Índice Natural de Aumento de Fugas (INAF o RR por las siglas de *Rate of Rise*). El promedio de estos tres componentes corresponde al NEFCP y se representa con la línea punteada (horizontal) en la misma ilustración.

Es importante tener en cuenta que las fugas reportadas tienen un componente de coste asociado al caudal que se pierde a través de ellas, y un componente de coste asociado a la reparación de la tubería. Cuanto más extenso sea el período de notificación y reparación, mayor será el componte de coste asociado al caudal que se fuga, mientras que el componente asociado a la reparación se mantiene invariable. Es de esperar que en RDAPs con un bajo nivel de control de pérdidas, el primero de los componentes sea mayor que en redes con mayor control de pérdidas (e.g. redes sectorizadas).

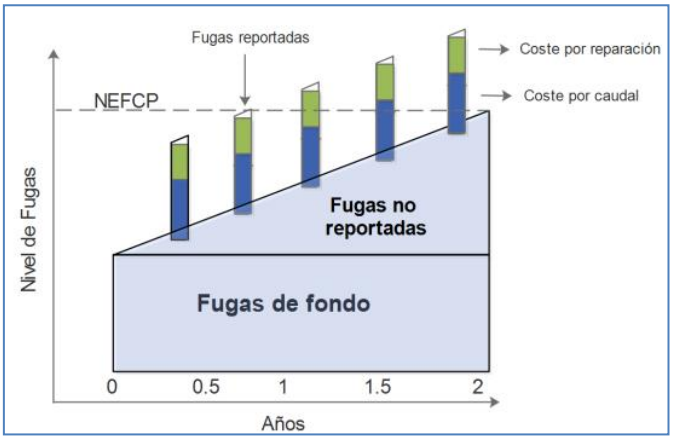


Ilustración 62: División de caudales de fugas establecido en el marco BABE [Fuente: Basado en Fantozzi & Lambert (2007)]

Vale la pena mencionar que el umbral de distinción entre roturas y fugas de fondo no es algo estático, y por el contrario, varía en cada sitio. De manera general, se establece que las fugas de fondo corresponden a aquellas que son muy difíciles de detectar con las tecnologías disponibles.

III.1.5 Formulación del Cálculo de Nivel Económico de Fugas no Reportadas

La Ilustración 63 conceptualiza la ecuación para calcular la Frecuencia Óptima de Inspección (FOI) en RDAPs. En ella, se puede ver que a lo largo de un tiempo T, expresado en días, el caudal nocturno aumenta desde Q_0 a Q_1, es decir, tiene un aumento Q_u. Si la presión nocturna promedio es P_0 cuando se mide Q_0, y P_1cuando Q_1 se mide, y si P_0 y P_1son significativamente diferentes, entonces sería necesario corregir Q_0 (multiplicando Q_0 por P_1/P_0) antes de proceder con el cálculo. También, el caudal de fugas nocturno en m^3/hora se debe multiplicar por un factor día-noche (*FDN*) apropiado, el cual, relaciona los caudales de fugas en la noche con los caudales promedios de fugas diurnos.

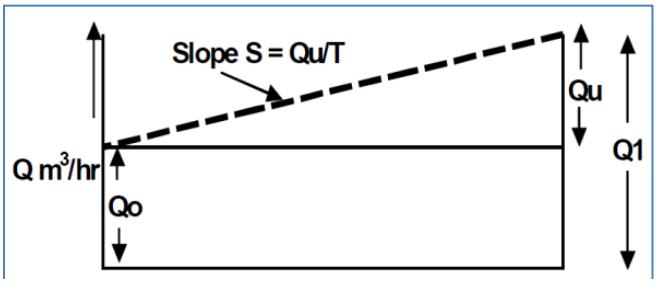


Ilustración 63: Ilustración para el cálculo de la frecuencia óptima de inspección [Fantozzi & Lambert (2005)]

Una vez que se ha corregido Q_0 en función de las presiones (en caso necesario) y que Q_u se convierte a m^3/día mediante la aplicación del FDN, se calcula la tasa de aumento de fugas no reportadas (Ecuación 33).

$$INAF(m^3/day/day) = Q_u \times \frac{FDN}{T} \qquad \text{(Ecuación 33)}$$

En términos más simples, el valor de *INAF* se obtiene mediante la comparación del caudal mínimo nocturno, justo antes de efectuar la reparación de todas las fugas no reportadas y reportadas, con el mismo parámetro al menos 1 año más tarde. Luego se divide la variación del caudal mínimo nocturno expresada en m^3/día, por la cantidad de años en que se realizó la evaluación, obteniendo un valor expresado en m^3/día/año.

Después del tiempo *T*, el volumen expresado por el triángulo (en la Ilustración 63) que representa las fugas no reportadas viene dado por la Ecuación 34.

$$V(m^3) = 0.5 \times INAF \times T^2 \qquad \text{(Ecuación 34)}$$

Si el costo marginal del agua es *CW* (*$/m^3*),* el valor del volumen V de fugas no reportadas en el tiempo T viene dado por la Ecuación 35.

$$CW * V = CW * 0.5 * INAF * T^2 \qquad \text{(Ecuación 35)}$$

Usando suposiciones similares a las que se hacen cuando se implementa la *Teoría de Control Económico de Inventario*, se puede demostrar que la *FOI* ocurre cuando el coste de intervención sobre todo el sistema (excluyendo los costos de reparación) iguala al valor del volumen de fugas no reportadas $CW \times V$, por lo tanto, el CI viene dado por la Ecuación 36. Sustituyendo términos se puede obtener la *FOI* (en días) tal y como se muestra en la Ecuación 37.

$$CI = CW \times V = CW \qquad \text{(Ecuación 36)}$$

$$FOI(días) = \left(\frac{CI}{CW \times 0.5 \times INAF}\right)^{0.5} \qquad \text{(Ecuación 37)}$$

Si el INAF se expresa en $m^3/día/año$, en lugar de $m^3/día/día$, entonces, la Ecuación 37 puede ser expresada en meses, tal y como se muestra en la Ecuación 38.

$$FOI(meses) = \sqrt{\frac{0.789 \times CI}{CW \times INAF}} \qquad \text{(Ecuación 38)}$$

De esta ecuación se obtiene un valor en meses, a partir del cual se puede calcular el porcentaje de la red que debe ser inspeccionado anualmente (PI) (Ecuación 39).

$$PI(\%/año) = 100 \times \frac{12}{FOI} \quad \text{(Ecuación 39)}$$

Y a partir de ahí, también se puede calcular el presupuesto del que se debe disponer cada año para tal fin (Ecuación 40).

$$PAI(\$/año) = PI \times CI \quad \text{(Ecuación 40)}$$

Donde *PAI* corresponde al presupuesto anual de inspección.

Al dividir el *PAI* entre el *CW*, se obtiene el *NEFNR* (Ecuación 41).

$$NEFNR(m^3/año) = \frac{PAI}{CW} \quad \text{(Ecuación 41)}$$

El método de cálculo anteriormente presentado se puede emplear para predecir los beneficios de la gestión de la presión en una RDAP. Siguiendo el principio de la teoría FAVAD, al reducir la presión de la red, es de esperar una reducción del caudal de fugas (de fondo, no reportadas detectables y reportadas), tal y como se observa en la Ilustración 64. En la misma Ilustración se puede observar que al reducir la presión, se disminuye el *INAF*, se extiende el período necesario para realizar una inspección completa (que implica una reducción del *PI* que se debe ejecutar) de la red y por tanto se reduce el *PAI*. Como consecuencia de todo ello, se reduce el *NEFCP* y el *NEFNR*.

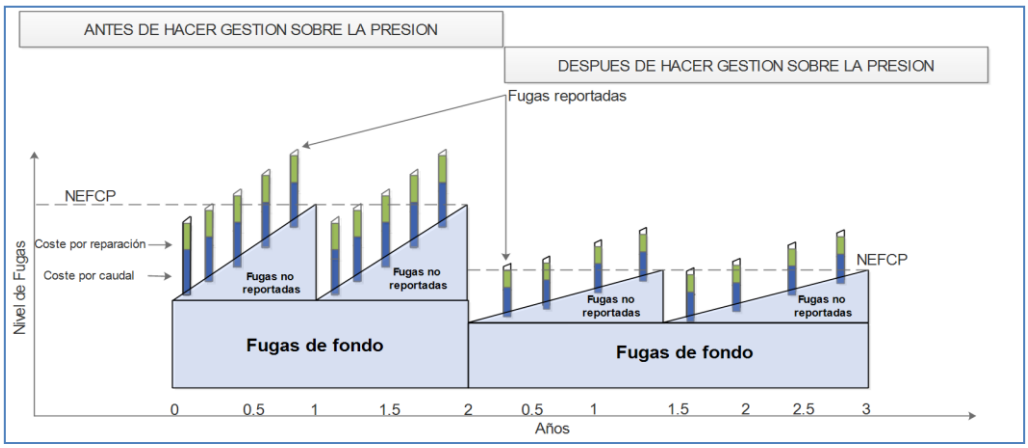


Ilustración 64: Efecto de la reducción de presión sobre el *NEFCP* [Fuente: Basado en Fantozzi & Lambert (2007)]

III.1.6 Beneficio de la Reducción de Presión sobre la Aparición de Nuevas Roturas y Sobre el Caudal de Consumo Doméstico

Además del beneficio de reducción de fugas, existen otros beneficios que se pueden obtener de gestionar (reducir) la presión en una RDAP: reducción de aparición de nuevas roturas, reducción del consumo doméstico interno y externo.

En el año 2006, se planteó el enfoque con el que actualmente se aborda la predicción en la reducción de roturas (Thornton & Lambert, 2007). Dicho enfoque se basa en datos (relación entre variación de presión y reducción de roturas) obtenidos en 110 sectores de 10 países distintos (Ver Ilustración 65).

Percentage reductions in new break numbers, before and after pressure management.

Country	Water Utility or System	Number of Pressure Managed Sectors in study	Assessed initial maximum pressure (metres)	Average % reduction in maximum pressure	Average % reduction in new breaks	Mains (M) or Services (S)
Australia	Brisbane	1	100	35%	28%	M,S
	Gold Coast	10	60-90	50%	60%	M
					70%	S
	Yarra Valley	4	100	30%	28%	M
Bahamas	New Providence	7	39	34%	40%	M,S
Bosnia Herzegovina	Gracanica	3	50	20%	59%	M
					72%	S
Brazil	Caesb	2	70	33%	58%	M
					24%	S
	Sabesp ROP	1	40	30%	38%	M
	Sabesp MO	1	58	65%	80%	M
					29%	S
	Sabesp MS	1	23	30%	64%	M
					64%	S
	SANASA	1	50	70%	50%	M
					50%	S
	Sanepar	7	45	30%	30%	M
					70%	S
Canada	Halifax	1	56	18%	23%	M
					23%	S
Colombia	Armenia	25	100	33%	50%	M
					50%	S
	Palmira	5	80	75%	94%	M,S
	Bogotá	2	55	30%	31%	S
Cyprus	Lemesos	7	52.5	32%	45%	M
					40%	S
England	Bristol Water	19	62	40%	40%	M
					55%	S
	United Utilities	10	47.6	32%	72%	M
					75%	S
Italy	Torino	1	69	10%	45%	M,S
	Umbra	1	130	39%	71%	M,S
USA	American Water	1	199	36%	50%	M
Total or Average		**110**		**37%**	**51%**	

Ilustración 65: Reducción de roturas como consecuencia de reducción de presión (Thornton & Lambert, 2006)

A partir de los datos en la Ilustración 65 se establecieron las curvas de relación que se presentan en la Ilustración 66, y a partir de ellas se establecen tres constantes que se multiplican por la variación máxima de la presión, para obtener *Factores de Reducción de Roturas*, tal y como se plantea en la Ecuación 42 (Pearson *et al.*, 2005).

$$r' = r * FR \qquad \text{(Ecuación 42)}$$

Aquí r' corresponde a la cantidad de roturas una vez reducida la presión; r corresponde a la cantidad de roturas iniciales y FR al factor de reducción medio, máximo o mínimo, los cuales se calculan multiplicando el porcentaje de variación de la presión media, máxima, o mínima por 1.4; 2.8 o 0.7, respectivamente.

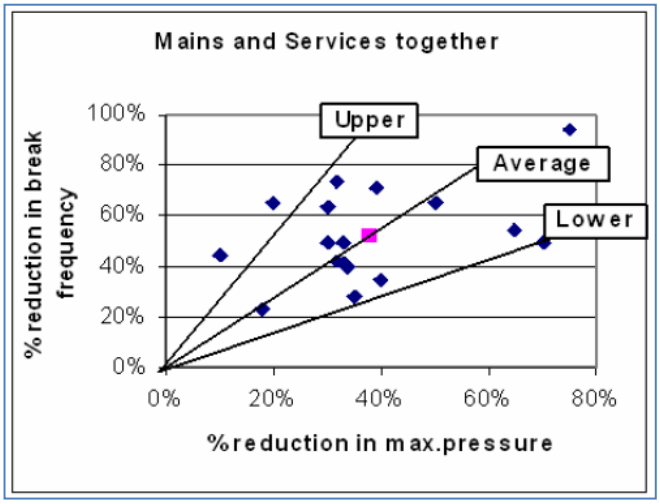


Ilustración 66: Curva de factores de reducción de roturas por efecto de reducción de la presión [FUENTE: Thornton & Lambert (2007)]

En las primeras implementaciones de esta ecuación se hizo notorio que en algunos casos se producían reducciones importantes en la frecuencia de roturas en las líneas de conducción principal, pero no así en las líneas de distribución, y viceversa. En la búsqueda de una explicación para este hecho, se desarrolló lo que actualmente se conoce como el *Modelo Conceptual* propuesto por Thornton & Lambert (2006, 2007). De acuerdo a este modelo, sólo se pueden esperar reducciones significativas de la frecuencia de roturas si la frecuencia de roturas actual es significativamente superior a un valor de frecuencia inicial de roturas que se asumió para establecer la ecuación *Pérdidas Reales Inevitables* (o *UARL* por la siglas en inglés de *Unavoidable Annual Real Losses*) descrita por Lambert *et al*. (1999),

$$UARL(l/día) = \left(1.8 * L_m + \ 0.8 * N_c + \ 25 * L_p\right) * P \qquad \text{(Ecuación 43)}$$

donde L_m indica longitud de tuberías de conducción; N_c indica número de conexiones; L_p indica longitud de tuberías de transporte y P indica presión.

Para la definición de ecuación del *UARL* se asumió que en una red con alto nivel de mantenimiento, el nivel de roturas esperado en la red de conducción principal es 13/100 km/año (sin incluir las reparaciones en las válvulas y los hidrantes); en tanto,

en la red de distribución, el valor esperado es 3/1000 conexiones/año, sin incluir pequeñas fugas en los contadores y en grifos (Lambert *et al*., 1999).

Cuando las tuberías son nuevas y los factores externos (bajas temperaturas, tracciones, corrosión) aún no han causado daños importantes en las mismas, es de esperar que estas puedan soportar sobrepresiones muy superiores a las presiones de diseño. Esto es particularmente importante en RDAPs que son abastecidas por bombeos, en las cuales ocurren transitorios en cada ciclo de apagado y encendido de las bombas. La Ilustración 67 y la Ilustración 68 muestran la conceptualización del comportamiento de la frecuencia de aparición de roturas en redes que son abastecidas por gravedad y por bombeo, respectivamente. En el segundo de los casos, la variación de la presión genera un mayor impacto sobre la frecuencia de aparición de nuevas roturas.

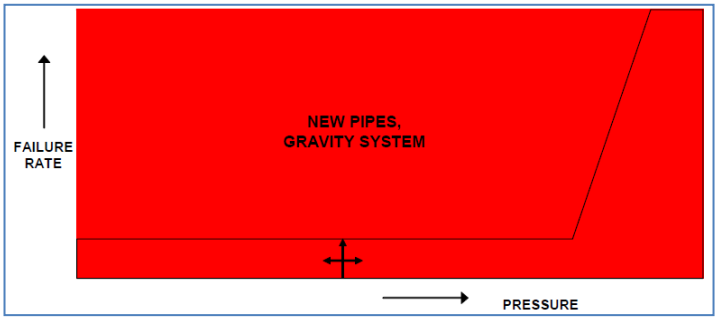


Ilustración 67: Aumento del porcentaje de roturas en proporción al comportamiento de la presión de acuerdo al *Modelo Conceptual* (abastecimiento por gravedad) [Fuente: Thornton & Lambert (2007)]

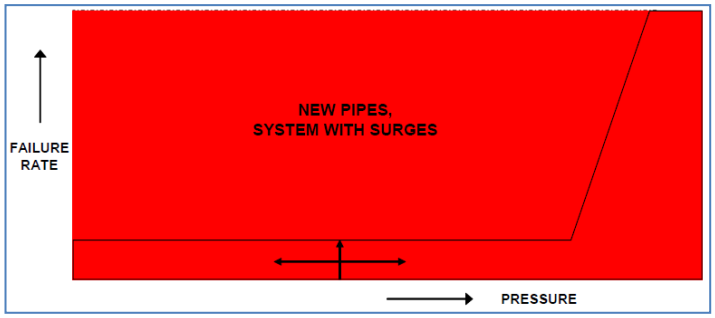


Ilustración 68: Aumento del porcentaje de roturas en proporción al comportamiento de la presión de acuerdo al *Modelo Conceptual* (abastecimiento por bombeo) [Fuente: Thornton & Lambert (2007)]

En la medida en que los factores externos empiezan a deteriorar las tuberías, las mismas se van haciendo más susceptibles a excesivas variaciones de la presión,

aumentado la frecuencia de roturas (ver Ilustración 69 e Ilustración 70). En ese punto, una reducción de la presión podría conllevar una reducción importante en la frecuencia de roturas, siempre y cuando el efecto de los factores externos no sea tan grande como para que una reducción conlleve desabastecimiento. En dicho caso, se haría necesaria la implementación de un programa de renovación de tuberías.

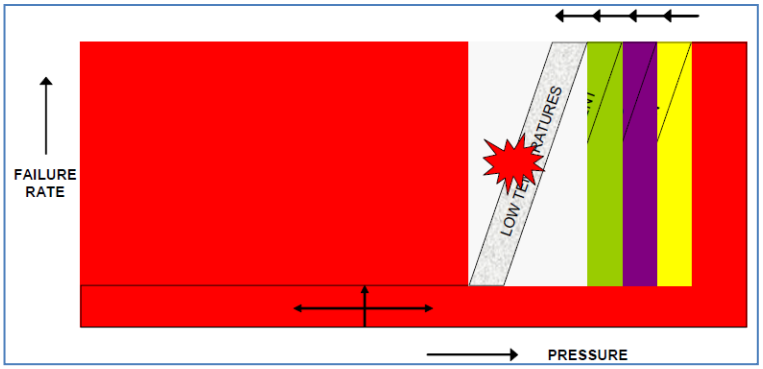


Ilustración 69: Efecto de la variación de presión sobre la frecuencia de roturas de acuerdo al *Modelo Conceptual* (sistema nuevo) [Fuente: Thornton & Lambert (2007)]

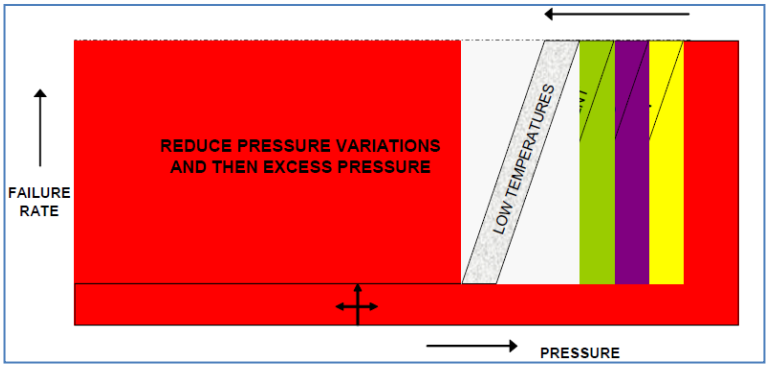


Ilustración 70: Efecto de la variación de presión sobre la frecuencia de roturas de acuerdo al *Modelo Conceptual* (sistema antiguo) [Fuente: Thornton & Lambert (2007)]

III.1.7 Reducción del Caudal de Consumo Doméstico

El consumo doméstico, tanto interno (para labores domésticas) como externo (e.g. irrigación de jardines), también se ve afectado por la variación de la presión. Esta variación se representa mediante una modificación de la ecuación FAVAD (ver Ecuación 44), en la que los términos de caudal ahora se refieren a consumos domésticos (internos y externos) y el exponente se denota con N_3. Para consumo externo, un valor de N_3 igual a 0.5 es usualmente apropiado, en tanto que, para consumo residencial interno, se recomienda un valor de 0.1, a menos que la mayor

parte de las conexiones cuenten con cisterna (tanque o aljibe) privada, en cuyo caso el valor 0 sería más apropiado (Fantozzi & Lambert, 2007),

$$\frac{Q^0}{Q^1} = \left(\frac{P^0}{P^1}\right)^{N_3} \qquad \text{(Ecuación 44)}$$

donde las variables Q indican caudales de consumo doméstico (interno y externo) inicial y posterior a la variación de la presión P.

III.1.8 Asociación de la Gestión de Fugas con la Sectorización

De lo expuesto hasta este punto, se puede concluir que una reducción de la presión se traduce en un beneficio económico debido a:

- Una disminución del caudal de fugas de fondo
- Una disminución del caudal de fugas reportadas
- Una disminución del caudal de fugas detectables no reportadas
- Una disminución del número de roturas reportas y no reportadas
- Una disminución del consumo doméstico (tanto interno como externo)
- Una disminución de la frecuencia de inspección de la red

La Ilustración 71 muestra gráficamente estos puntos. Según la configuración final de la sectorización, el consumo energético podría aumentar o disminuir. Si el aumento de pérdidas de carga que genera la sectorización no es excesivo, es de esperar que el consumo energético disminuya (debido a la reducción del caudal de fugas).

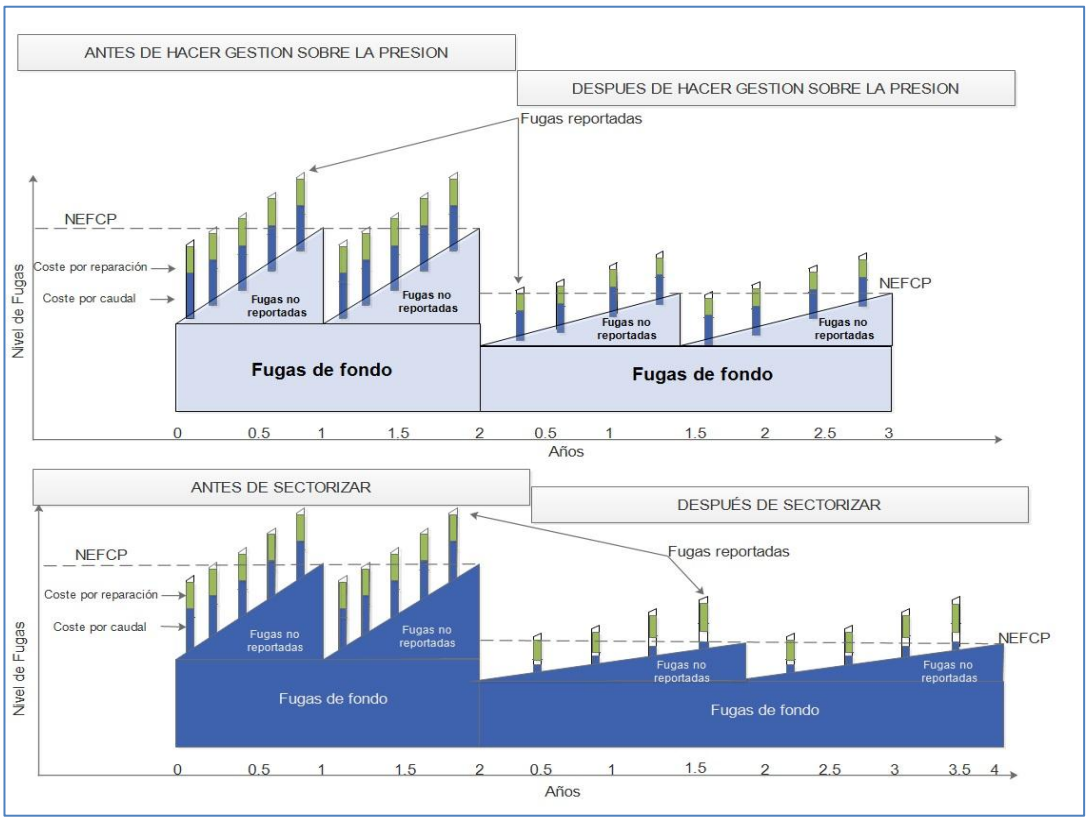


Ilustración 71: Efecto de gestión de RDAPs mediante sectorización

A fin de disminuir la presión en las RDAP, se pueden emplear VRPs, implementar sectores, o emplear VRPs en sectores. El emplear VRPs en combinación con sectores resultará, evidentemente, una opción más cara.

En términos generales, la sectorización no está concebida para reducir presiones, sino, para generar un mejor control de las fugas; no obstante, al sectorizar, inevitablemente se reduce un porcentaje de la presión, al menos en la red de distribución. Esto es válido sólo en el caso en que se cierren tuberías. En caso de que todos los sectores se delimiten con caudalímetros, es de esperar que la presión no varíe. La idea aquí propuesta, es contabilizar el beneficio que genera el sectorizar una red como consecuencia de la reducción de la presión producida por el aislamiento parcial de los sectores (definición del CEVC) y del mayor control con que se cuenta al poder monitorizar, de manera permanente, los caudales que entran a cada uno de los sectores. Para ello, (1) primero se obtienen los valores de reducción de caudal derivados de la reducción de la presión (fugas de fondo, fugas reportadas, consumo doméstico) basados en la ecuación FAVAD y se calcula el

NEFNR; (2) a continuación, se distribuye el caudal asociado a fugas reportadas y el *NEFNR* entre cada uno de los sectores, con base en los criterios técnicos de los operarios de la red; (3) en función de la longitud de tubería y la cantidad de entradas de cada sector, se estima un porcentaje de detección de eventos de fugas (detección inmediata y reparación); (4) se actualiza el valor de caudal de fugas reportadas y el *NEFNR*. En la Ilustración 72 se muestra gráficamente el proceso descrito.

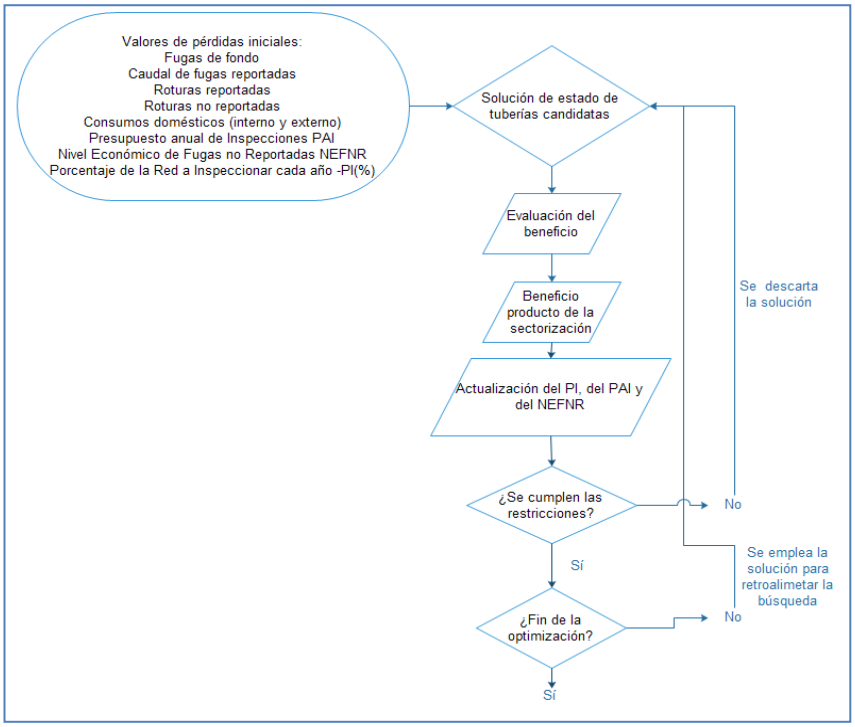


Ilustración 72: Propuesta de esquema de optimización del CEVC

Los beneficios se suman, para obtener el beneficio anual bruto. A este valor se le resta un coste de penalización asociado a los nudos en donde no se logra satisfacer el criterio de presión mínima, más un porcentaje del coste de inversión y un coste de mantenimiento anual de los equipos que sean instalados (UOCs y válvulas delimitadoras). Dado que los beneficios brutos se calculan en términos anuales el coste de inversión se multiplica por un factor de amortización (Ecuación 45), obteniendo, así, el porcentaje de la inversión que se debe cubrir cada año,

$$FA = \frac{(1+r)^T \times r}{(1+r)^T - r} \qquad \text{(Ecuación 45)}$$

donde FA es el Factor de Amortización; r es la tasa de descuento anual en % y T es el tiempo de vida de los activos (en este caso: caudalímetros y válvulas).

Para calcular el valor de la energía, se multiplica el consumo energético de cada hora (en *kWh*) en que se emplea bombeo, por el coste de la tarifa energética a esa hora ($/*kWh*). Luego se suma el valor de cada hora para obtener el consumo energético de cada día. Este valor diario se transforma a términos anuales y se agrega al beneficio bruto aunque, tal y como se explicará más adelante, también puede ser agregado a los gastos, dependiendo del esquema final de sectorización.

En la Ilustración 73 se puede ver, gráficamente, la relación de los beneficios por reducción de pérdidas físicas y la inversión en la compra de válvulas y caudalímetros. La línea más alta corresponde al beneficio neto anual (beneficio bruto anual - gasto anual), la cual no debe cruzar nunca el límite de estándar de abastecimiento establecido. También, es importante hacer notar, cómo la curva de ahorro cruza la línea de *NEFCP*, lo cual está asociado a los beneficios adicionales que tiene la gestión de presiones mediante la sectorización, más allá de únicamente reducir el caudal de fugas de fondo.

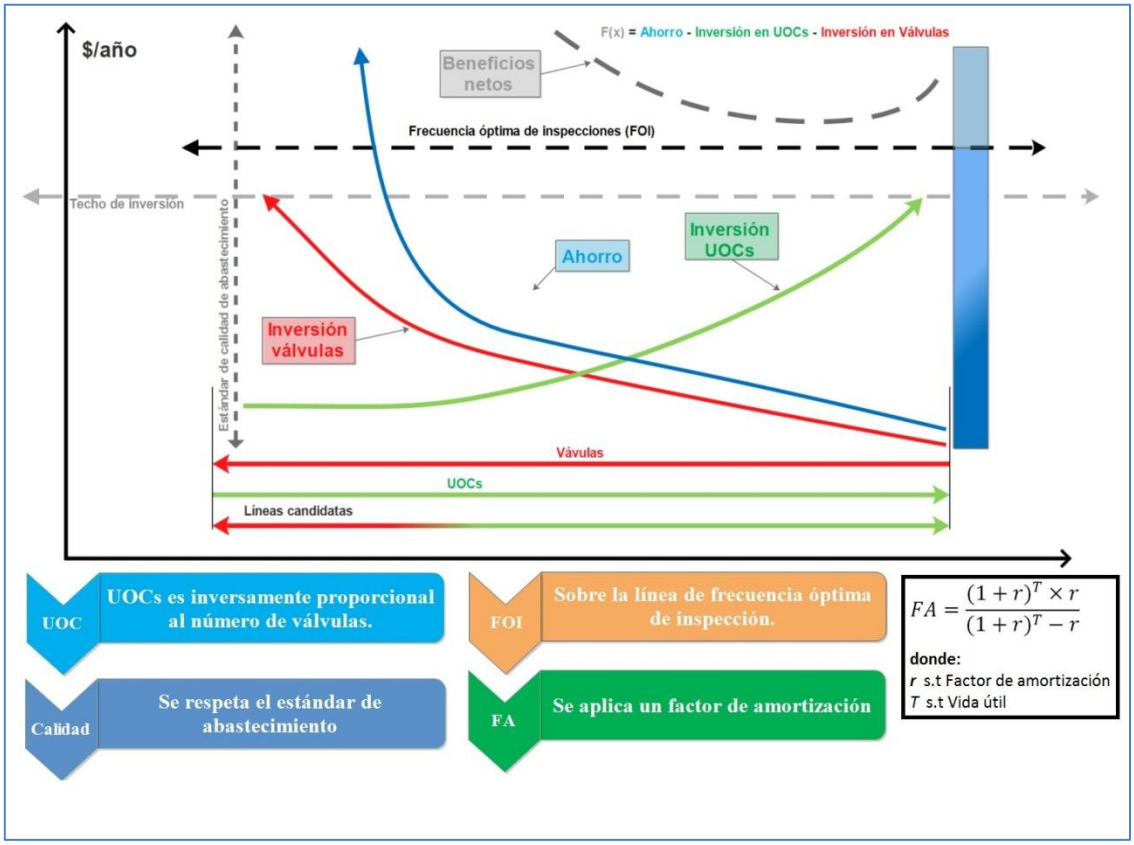


Ilustración 73: Relación coste/beneficio de la sectorización

La Ecuación 46 permite calcular el beneficio total de la sectorización, teniendo en cuenta todos los factores anteriormente descritos.

$$\text{Max} f(x) = A + B + C + D + E + F + G \pm H - A' - B' - C' \qquad \text{(Ecuación 46)}$$

Sujeto a:

$$\Delta I_r < \Delta I_r{}^{\text{máx}}; P_{\text{mín}} < P_{\text{mín}}{}^{req}; 0 < \text{A} + \text{B} < (A + B)^{\text{máx}}$$

ΔI_r y $\Delta I_r{}^{\text{máx}}$: Desviación del índice de resilencia con respecto al máximo valor permitido

$P_{\text{mín}}$ y $P_{\text{mín}}{}^{req}$: Presión máxima admitida y presión mínima admitida

$(A + B)$ y $(A + B)^{\text{máx}}$: Presupuesto para la compra de válvulas de aislamiento y UOCs y presupuesto límite

A: Ahorro por reducción de fugas de fondo (volumen) ($/año)

B: Ahorro por reducción de fugas reportadas (volumen) ($/año)

C: Ahorro en la reducción de fugas a reparar (fugas reportadas) ($/año)

D: Ahorro por reducción del número de tuberías a reparar (fugas no reportadas) ($/año)

E: Ahorro por reducción de consumo doméstico (volumen) ($/año)

F: Ahorro por reducción de consumo doméstico externo (volumen) ($/año)

G: Ahorro por reducción de fugas no reportadas (volumen) ($/año)

H: Ahorro/gastos por aumento/reducción de consumo energético ($/año)

A': Costo amortizado de válvulas y UOCs ($/año)

B': Costo de compensación por déficit de presión ($/año)

C': Costo de mantenimiento de válvulas y UOCs ($/año)

Las fugas de fondo representadas por el término A corresponden a aquellas que no se pueden detectar por medio de la tecnología disponible y que sólo se pueden disminuir mediante la reducción de la presión; las fugas reportadas, indicadas por el término B son las fugas que son visualmente detectables; los términos C y D indican reparación de grietas en las tuberías, ya sea para las fugas que son detectables visualmente o fugas que no son detectables visualmente, pero que pueden ser localizadas en una campaña de detección de fugas. Los términos E y F indican el consumo de agua dentro y fuera de los hogares; el término G indica fugas que pueden ser detectadas (localizadas) en una campaña de detección de fugas; H indica el consumo de energía debido al bombeo; A' indica el costo anual (coste amortizado) de la inversión para la adquisición de válvulas límites y UOCs; B' indica el coste de compensación por los nodos que no pueden cumplir con la presión de servicio; este se calcula multiplicando la demanda total de los nodos críticos por el costo del suministro de un metro cúbico de agua por medio de una cisterna. Finalmente, el costo de mantenimiento representado por C' se estima como un porcentaje del pago anual de la inversión.

El índice de resiliencia utilizado aquí corresponde al índice propuesto por Todini (2000), que se calcula restando de 1 la fracción que representa la pérdida de carga actual sobre la pérdida de carga total requerida para asegurar que todos los nodos tengan un límite de presión determinada. En la Subsección III.2 Criterios Hidráulicos para Sectorización, se aborda el indicador en cuestión con más nivel de detalle.

Es importante tener en cuenta la complejidad de la función objetivo, sobre todo si se compara con funciones objetivo utilizadas en investigaciones previas. De hecho, este enfoque tiene dos desventajas principales. La primera de ellas está relacionada con extensos tiempos de cálculos. Sin embargo, como se ha explicado anteriormente, la sectorización aporta beneficios muy importantes que deben tenerse en cuenta con el fin de hacer la optimización más realista. Una manera de resolver el problema del tiempo de cálculo podría ser la agrupación de algunos de los beneficios. Sin embargo, esto requeriría una nueva investigación a fin de conocer la forma en que se relacionan los mismos entre sí, lo que en este momento no está muy claro y de hecho, es uno de los temas sobre los cuales los autores están trabajando actualmente, con lo cual este aspecto se plantea como una recomendación para el futuro. El segundo inconveniente está relacionado con la información en sí misma, ya que no es fácil de obtener, y por el contrario, su generación requiere un gran esfuerzo por parte de las empresas operadoras de RDAPs.

III.2 Criterios Hidráulicos para Sectorización

En la presente sección se hace una descripción detallada de una serie de criterios hidráulicos y operacionales que se pueden tener en cuenta al momento de optimizar el CEVC en un esquema de sectorización.

III.2.1 Índice de Resiliencia

De acuerdo con Creaco *et al*. (2013), en una red mallada, un cambio en el caudal debido al fallo en una tubería o por un aumento de la demanda incrementaría las

pérdidas de energía. En caso de que los nodos sean abastecidos con la presión exacta necesaria, sería imposible suplir la demanda necesaria en el escenario anteriormente planteado, por lo que se hace necesario contar con una cantidad de energía adicional o *surplus* que permita hacer frente a este tipo de imprevistos, garantizando el abastecimiento en condiciones críticas.

La presión con la que el agua es entregada a un usuario es el resultado de la transferencia de energía desde la fuente (o fuentes) a través del circuito conductor conformado por las líneas de transporte. Como es lógico pensar, en este transporte, una parte de la energía se pierde tanto en las tuberías como en los accesorios. Todini (2000) planteó una serie de ecuaciones para describir esta transferencia. A través de estas, se establece que la potencia de entrada P_{inp} de la red es equivalente a la suma de las potencias entregadas a los usuarios P_{out} y la potencia de operación P_{int} (pérdidas por fricción y fugas) (Ecuación 47).

$$P_{inp} = P_{out} + P_{int} \qquad \text{(Ecuación 47)}$$

A su vez, P_{inp} también se puede expresar en función de la potencia que es suministrada por los bombeos ($j = 1, \dots, n_p$) y por embalses ($e = 1, \dots, n_e$):

$$P_{inp} = \sum_{e=1}^{n_e} Q_e \times H_e + \sum_{j=1}^{n_p} P_j \qquad \text{(Ecuación 48)}$$

La potencia de entrega a los consumidores ($j = 1, \dots, n_n$) se puede expresar como una magnitud real o como una magnitud máxima. En ambos casos el método de cálculo es el mismo, sólo que en uno se emplea la altura piezométrica real H_j (Ecuación 49) y en el otro se emplea la altura piezométrica máxima H_j^* requerida para satisfacer un requerimiento mínimo de presión (Ecuación 50).

$$P_{out}^{\text{real}} = \sum_{j=1}^{n_n} Q_j * H_j \qquad \text{(Ecuación 49)}$$

$$P_{out}^{\text{máx}} = \sum_{j=1}^{n_n} Q_j * H_j^* \qquad \text{(Ecuación 50)}$$

Teniendo en cuenta que la potencia de entrada es igual a la suma de la potencia operacional más la potencia entregada, se puede estimar la potencia operacional como la diferencia entre la potencia de entrada menos la potencia de entrega. Al poder expresar la potencia de entrega en dos términos (máxima y real) la potencia operacional también se puede expresar en los mismos dos términos.

$$P_{int} = P_{inp} - P_{out} \qquad \text{(Ecuación 51)}$$

$$P_{int}^{\text{real}} = P_{inp} - P_{out}^{\text{real}} \qquad \text{(Ecuación 52)}$$

$$P_{int}^{\text{máx}} = P_{inp} - P_{out}^{\text{máx}} \qquad \text{(Ecuación 53)}$$

Estos valores de potencia operacional real y requerida podrían ser pensados como la energía que se requiere disipar para llevar la presión correspondiente a un determinado nodo, que su vez depende de la configuración de la red (la rugosidad de la tubería y el camino que tiene que recorrer el agua).

El mismo autor también presenta un índice de resiliencia (Ir) para evaluar la eficiencia energética de las RDAPs. Dicho índice es descrito como la capacidad de un sistema de reaccionar y superar estados no normales, o el incremento de la redundancia energética y decrecimiento de la energía disipada internamente en una red. En términos generales, éste se calcula restando de la unidad el porcentaje que representa la potencia de operación real en la red con respecto a la potencia de operacional máxima requerida para satisfacer una presión mínima.

$$Ir = 1 - \frac{P_{int}^{\text{real}}}{P_{int}^{\text{máx}}} \qquad \text{(Ecuación 54)}$$

Luego, sustituyendo términos, se obtiene una expresión más específica para obtener el índice de resiliencia.

$$Ir = 1 - \frac{[\sum_{i=1}^{n_e}(Q_e * H_e)_i + \sum_{i=1}^{n_p} P_i] - \sum_{j=1}^{n_n} Q_J * H_J}{[\sum_{i=1}^{n_e}(Q_e * H_e)_i + \sum_{i=1}^{n_p} P_i] - \sum_{j=1}^{n_n} Q_J * H_j^*} \qquad \text{(Ecuación 55)}$$

$$Ir = 1 - \frac{\sum_{j=1}^{n_n} Q_J * (H_J - H_J^*)}{[\sum_{i=1}^{n_e}(Q_e * H_e)_i + \sum_{i=1}^{n_p} P_i] - \sum_{j=1}^{n_n} Q_J * H_j^*} \qquad \text{(Ecuación 56)}$$

Con este índice se puede estimar cómo de fiable es la red ante el fallo de uno de sus elementos. El mismo presenta su valor óptimo en 0.5. Un valor menor a este, sería propio de una red que no podría responder bien ante una falla de uno de sus elementos y, el caso contrario, se tendría una red sobredimensionada que, además de costosa, probablemente presentaría valores bajos de velocidad, que luego pueden desencadenar problemas de calidad de agua.

En esta línea, Di Nardo *et al.* (2013) presenta un índice de desviación del índice de resiliencia, que permite evaluar lo que disminuye la resiliencia de una RDAP debido a la implementación de un esquema de sectorización. Es decir, evalúa, cuánta energía extra se disipa como consecuencia de la implementación de la sectorización. El índice en cuestión se calcula mediante la Ecuación 57.

$$I_{rd} = \left(1 - \frac{I^*_r}{I_r}\right) \times 100 \qquad \text{(Ecuación 57)}$$

donde, I_{rd} representa la desviación del índice de resiliencia; I^*_r representa el índice de resiliencia de la RDAP con el nuevo esquema e I_r representa el índice de resiliencia de la RDAP con el esquema original.

Pese a que este índice de resiliencia no es el único disponible, es en la actualidad el

más ampliamente utilizado. Baños *et al*. (2011), comparan el citado índice de resilience con otros dos índices (Prasad & Park, 2004; Jayaram & Srinivasan, 2008) empleados en un marco de optimización heurístico, concluyendo que ninguno de estos es capaz de determinar con precisión la capacidad de la red para abastecer bajo condiciones inciertas.

III.2.2 Uniformidad de Presiones

De acuerdo con Araque & Saldarriaga (2005), al minimizar el valor de I_{rd}, que representa la relación entre la energía disipada por el sistema actual con una configuración dada respecto a la energía óptima disipada, se logra uniformizar el estado de las presiones. La definición de energía óptima disipada hace referencia a cuánta energía se espera que la RDAP potable disipe en cada una de las tuberías que la conforman. Para evaluar el grado de uniformidad de presiones de una RDAP, estos autores proponen el coeficiente que se presenta en la Ecuación 58.

$$CU = \frac{\sum_{j=1}^{n} P_j}{n * \text{máx}[P_j]} \qquad \text{(Ecuación 58)}$$

donde n representa el número de nodos de la red y P_j representa la presión en cada uno de los nodos de la red.

En algunas RDAPs antiguas con bajo nivel de inversión, o con escasez regular de agua, el concepto de presión mínima de servicio es un concepto muy difícil de implementar. En este tipo de situación, la aplicación de estas ecuaciones requeriría un valor de presión mínima muy baja que en algunos puntos podría representar desabastecimiento.

III.2.3 Coeficiente de Pérdida de Potencia

Otra manera de evaluar la eficiencia energética de las redes, sin tener que establecer un mínimo de presión, es hacer una comparación entre la potencia de los nodos de consumo para dos o más esquemas de red. Así, para una red abierta sin sectorizar (sin haber cerrado válvulas de sectorización), se obtiene la potencia de cada uno de sus nodos multiplicando su demanda (incluyendo caudal de fugas) por su presión. Posteriormente, se suman las potencias de todos los nodos, dando como resultado la potencia de entrega de la red abierta. Se repite el mismo proceso pero con otro esquema de la misma red (red ya sectorizada). A continuación se comparan ambos resultados. En este trabajo, está comparación es definida con el Coeficiente de Pérdida de Potencia (*CPP*) (ver Ecuación 59), que resulta mayor conforme mejor sea el esquema de sectorización:

$$CPP = \frac{\sum_{i=1}^{n} Q_{i_{df}}^{*} * P_i^{*}}{\sum_{i=1}^{n} Q_{i_{df}} * P_i} \qquad \text{(Ecuación 59)}$$

donde *n* representa el número de nodos de la red; $Q_{i_{df}}^{*}$ representa el caudal de demanda y fugas en cada nodo en el nuevo esquema de RDAP; P_i^{*} representa la presión en cada uno de los nodos de la red en el nuevo esquema de RDAP; $Q_{i_{df}}$ representa el caudal de demanda y fuga en cada nodo en el esquema original y P_i representa la presión en cada uno de los nodos de la red en el esquema original.

La comparación se puede hacer para cuantos esquemas de red se deseen probar (comparar un esquema sectorizado dado con respecto al esquema original). Posteriormente se puede establecer un *ranking* de mejores esquemas de red en función de su valor de CPP.

III.2.4 Uniformidad de Características

La uniformidad en ciertas características de los nodos de una RDAPs puede ser útil para evaluar el tamaño y la similaridad de los sectores que se establecen en la misma. En esta línea Alvisi & Franchini (2014) destacan el uso de datos de demanda para gestionar en tiempo real tareas que se pueden simplificar mediante el establecimiento de sectores con un alto grado de similitud. Este indicador se calcula mediante la desviación estándar de la demanda total o cota de cada sector, tal y como se muestra en la Ecuación 60. Como es de esperar, altos valores de este indicador reflejan baja calidad en la sectorización, dado que esto supone agrupación de nodos con valores de demanda o cota muy dispares. En el caso de la demanda, en realidad, no sería extraño encontrar algunos nodos con valores de demanda disimilares en un mismo sector. En efecto, tal como ya fue descrito en el capítulo II, por un lado, la asignación de demandas puede ser arbitraria y no necesariamente representar la realidad geográfica de las redes; y, por otro lado, aunque la demanda represente la realidad geográfica del sitio, puede darse el caso de que algunos nodos representen un gran consumidor (e.g. hospital, industria) y que el mismo se encuentre rodeado por nodos de baja demanda (zonas residenciales).

La utilización de la cota, en cambio, sí puede ser factible como indicador de uniformidad, ya que es de esperar que el valor de cota almacenado en los nodos sí refleje la realidad topográfica de la RDAP.

$$DEC = \frac{\sqrt{\frac{\sum_{i=1}^{N_n}(q_i - q_{av})^2}{N_n}}}{q_{av}} \qquad \text{(Ecuación 60)}$$

Aquí DEC (desviación estándar de característica) corresponde a la similaridad de demanda o cota y q_{av} es el promedio de la característica de referencia (demanda o cota) en la red.

III.2.5 Calidad del Agua

La calidad del agua se evalúa a través del análisis de la edad del agua. De acuerdo a EPA (2002), la reducción de la velocidad con que circula el agua a través de las tuberías puede conllevar problemas de tipo físico (congelamiento en zonas frías, color, olores, turbidez), químico (decaimiento de desinfectantes) o biológico (desarrollo bacteriano). Por ello, se han desarrollado indicadores para la evaluación de este parámetro, tal como el propuesto por Marchi *et al.* (2014):

$$I_{WA} = \frac{\sum_{i=1}^{N_n} \sum_{j=1}^{T} k_{i,j} \cdot \left(WA_{i,j} - WA_{lim\,i,j} \right)}{\sum_{i=1}^{N_n} \sum_{j=1}^{T} q_{i,j}} \qquad \text{(Ecuación 61)}$$

donde I_{WA} conrresponde al indicador de calidad de agua; $WA_{i,j}$ es la edad del agua en el nodo i, en el tiempo j; $WA_{lim\,i,j}$ es el límite de edad de agua en el nodo i, en el tiempo j y $k_{i,j}$ es un coeficiente igual a 1, si $WA_{i,j} > WA_{lim\,i,j}$ y 0 si $WA_{i,j} \leq WA_{lim\,i,j}$.

III.3 Generalidades sobre Optimización

En el campo de las matemáticas y la ciencia de la computación, los problemas de optimización matemática pueden ser descritos como la búsqueda de la mejor solución cuando el número de buenas alternativas es muy grande. Es decir, la selección de los mejores elementos dentro de un conjunto de alternativas en un dominio dado que se define por un conjunto de criterios o restricciones. Pese a que, tal y cómo se mencionará más adelante, existe una diversidad de métodos de optimización, en general, todos ellos están orientados a la misma tarea, a saber, la maximización de un problema, representado a través de una o varias funciones objetivo. Tal como lo establecen Eiben & Smith (2003), en un problema de optimización se conoce el modelo, junto con el resultado deseado (o por lo menos

una descripción del resultado deseado) y la tarea es encontrar la salida o salidas que conlleven a dicha esperada salida (ver Ilustración 74).

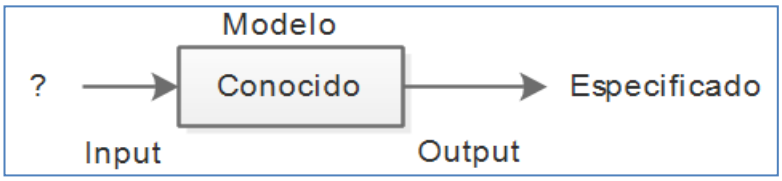


Ilustración 74: Gráfico conceptual de los problemas de optimización [Fuente: Eiben & Smith (2003)]

La búsqueda de soluciones óptimas en problemas de optimización puede ser realmente compleja, sobre todo cuando se abordan problemas con muchos objetivos y/o muchas restricciones. La selección de una solución óptima se relaciona con el concepto de *optimalidad*, mediante el cual se establece que una solución es mejor que otra siempre y cuando en la misma todos los criterios sean mejores que en la primera, y además, en la segunda no se pueda dar la mejora de uno de los objetivos sin detrimento de otro. Todos los elementos que cumplen con esta última característica son catalogados como dominantes y dentro del plano *x*-*y* se agrupan en una línea denominada curva (frente o frontera, los términos son intercambiables) de Pareto (ver Ilustración 75). Vale la pena destacar la diferencia entre la curva de Pareto acá abordada y el diagrama de Pareto que se emplea para el análisis de la red troncal que se presenta en la Subsección II.4 (ver Ilustración 19). El último no describe ningún tipo de optimalidad.

El concepto en cuestión se puede formular, matemáticamente, de la siguiente manera:

Dado un vector $u = (u_1, \dots, u_k)$ se dice que domina a otro vector $v = (v_1, \dots, v_k)$ si y sólo si $v_i \leq u_1 \ \ \forall\, v_i$ y $\exists i \in \{1, \dots, k\} / v_i < u_i$. Una solución $\boldsymbol{x}^*$ se dice que es Pareto-óptima si y sólo si no existe otro vector $\boldsymbol{x}$ tal que $f(\boldsymbol{x})$ domine a $f(\boldsymbol{x}^*)$, donde f es la función vectorial de todos los objetivos considerados.

Así, a la largo de la frontera de Pareto no existe una única solución, sino por el contrario, un conjunto de vectores de solución que caen dentro de la categoría de soluciones no dominadas.

La Ilustración 75 muestra la frontera de Pareto de un problema genérico. Los puntos sobre la línea corresponden a soluciones no dominadas.

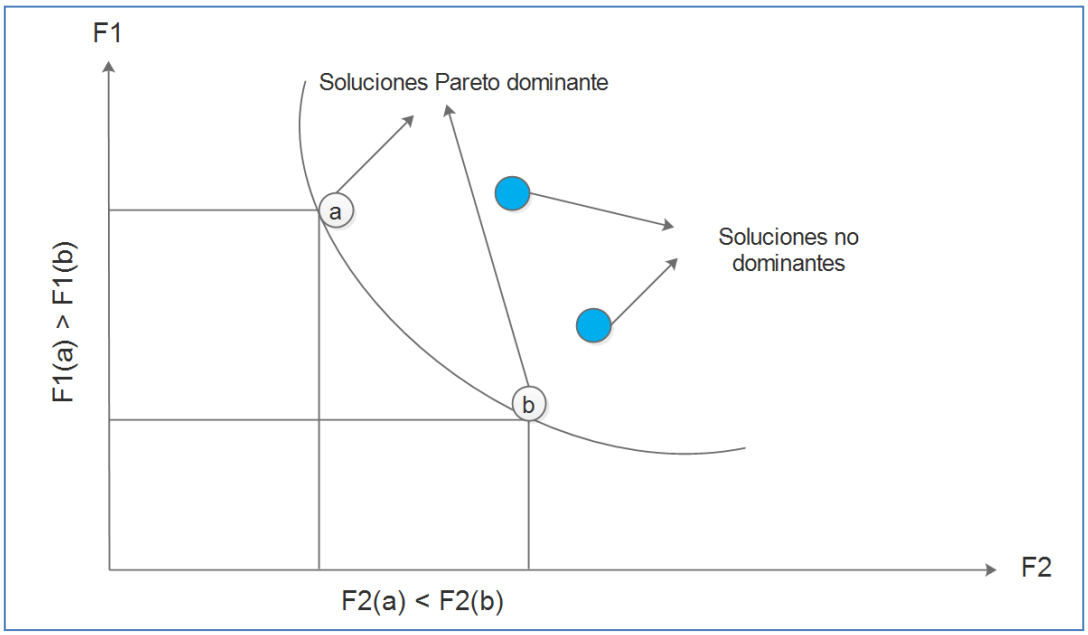


Ilustración 75: Ejemplo de frontera de Pareto

Las soluciones en un espacio de búsqueda unidimensional pueden ser clasificadas en dos tipos: de tipo local o de tipo global (ver Ilustración 76). Se dice que una solución es óptima global si a través de esta, se logra alcanzar un mínimo valor para la función $f_0(x)$ que no logre ser superado por ninguna otra solución dentro del espacio de búsqueda. En tanto, una solución es óptima local, si su optimalidad depende de las otras soluciones presentes en el vecindario en donde se localiza la misma. En otras palabras, la optimalidad ya no es relativa a un espacio de búsqueda, sino a una esfera suficientemente pequeña alrededor de la solución en cuestión.

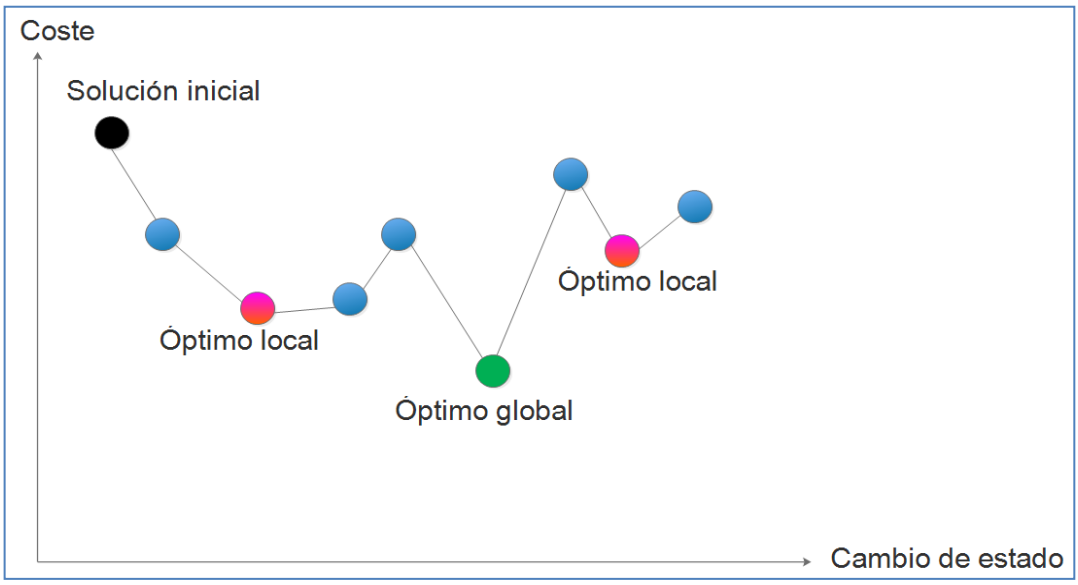


Ilustración 76: Tipos de óptimos en un espacio de búsqueda

Cuatro son los componentes que conforman el abordaje de un problema de optimización: (1) una o varias funciones objetivo; (2) una o más restricciones/penalización; (3) un conjunto (universo) de soluciones factibles, y (4) un método de solución. Describimos someramente cada uno de estos componentes.

Función objetivo: función objetivo (también conocida como función de costo, función de utilidad directa o función de energía) es una función que representa (parcialmente) el problema mediante el establecimiento de las relaciones existentes entre (todas) las variables de decisión. Las relaciones generan como resultado un costo/beneficio que se trata de minimizar/maximizar. Con frecuencia, la optimización se describe en términos de minimización de funciones objetivo; no obstante la maximización de una función objetivo puede ser concebida como el negativo de su minimización. Este principio de dualidad permite que los algoritmos de optimización puedan ser empleados para las dos tareas con sólo hacer cambios menores en los objetivos y sin necesidad de tener que modificar el algoritmo mismo. En la mayoría de los casos, las relaciones que se representan en la función objetivo son de tipo numéricas; sin embargo, en algunos casos, una representación cualitativa también es posible. En dicho caso se suele emplear una codificación numérica que permita incluir el criterio cualitativo dentro de la función objetivo. La optimización se puede llevar a cabo con un único objetivo (optimización mono-

objetivo), o con varios objetivos (habitualmente) enfrontados entre sí (optimización multi-objetivo). Ciertos algoritmos están únicamente concebidos o funcionan mejor con problemas mono-objetivos. No obstante, esto no quiere decir que con los mismos no se puedan abordar problemas multiobjetivo. Una estrategia relativamente común para tal fin, consiste en establecer el objetivo más importante como función objetivo y el resto de los objetivos como restricciones.

La transformación de un problema multi-objetivo en un problema mono-objetivo, puede generar problemas cuando la definición de la importancia de un objetivo sobre otro no es tan evidente (no es tan fácil definir qué objetivo es el más importante). Adicionalmente, en algunos casos puede ser bastante difícil establecer a qué rango de valores se deben restringir el resto de los objetivos, a fin de dejarlos definidos como restricciones. En este caso, la optimización multi-objetivo presenta la ventaja de no tener que seleccionar uno entre todos los objetivos. Sin embargo, dependiendo de la complejidad del problema a tratar, incluso las aproximaciones multi-objetivo, pueden no tener la capacidad para abordarlo. Surge, pues, la necesidad de desarrollar aproximaciones metaheurísticas en las que intervienen más de un algoritmo.

Restricciones y/o penalizaciones y espacio de búsqueda: El espacio de búsqueda se refiere a una colección de vectores candidatos a ser solución de un problema dado (Mitchell, 1998). Las restricciones representan algunas relaciones funcionales entre las variables de decisión con otros parámetros de diseño que satisfacen ciertos fenómenos físicos y ciertas limitaciones de recursos. Su función es reducir la cantidad de alternativas posible, definiendo un subconjunto llamado de soluciones factibles (Baquela & Redchuk, 2013). La naturaleza y el número de restricciones a incluir en la formulación dependen del usuario o del problema que se quiera abordar. Estas pueden tener una formulación matemática o no y, así mismo, pueden ser de tipo desigualdad o de tipo igualdad. En el primer caso, se establece que los valores de ciertas relaciones entre las variables de decisión tienen que ser menores o iguales a un valor umbral establecido, en tanto en el caso de las restricciones tipo igualdad, tiene que igualar el valor de la restricción.

Métodos de Optimización:

Dada la diversidad de problemas de optimización existentes, es de esperar que los mismos cuenten con una cantidad, igualmente diversa, de métodos de resolución, que se pueden clasificar/caracterizar de acuerdo a distintos tipos de criterios, que tal y como se verá a continuación, en algunos casos se traslapan entre sí. Por ejemplo, en función del tipo de variable contenida en el espacio de búsqueda, los métodos de optimización pueden ser de tipo continuo (optimización continua), dentro de los cuales se encuentra la optimización convexa lineal o programación lineal (e.g. el ampliamente conocido algoritmo *Simplex*), o de tipo combinatorio, en caso de que las variables sean de tipo discreto. En caso de tener una combinación de ambos tipos de variables, se trata de un problema de optimización mixta, cayendo dentro de esta categoría la optimización binaria, que se suele emplear en problemas de categorización (Baquela & Redchuk, 2013).

Una segunda categoría se basa en la tipología de la función objetivo y de las restricciones. Así, cuando tanto la función objetivo como las restricciones son lineales, se emplea programación lineal, y en caso de que las restricciones sean lineales pero la función objetivo sea cuadrática, se emplea programación cóncava o cuadrática.

Por otro lado, en función de la naturaleza probabilística de las variables, los métodos pueden ser determinísticos o estocásticos. Los enfoques determinísticos (e.g. programación lineal, programación no lineal y programación no lineal mixta-integral) aprovechan las propiedades analíticas del problema para generar una secuencia de puntos que convergen hacia una solución óptima global, en tanto los métodos estocásticos se emplean cuando se abordan problemas en donde uno o alguno de los datos incorporan incertidumbres. Se ha encontrado que, para muchos problemas reales, los enfoques heurísticos son más flexibles y eficientes que los enfoques deterministas; sin embargo, no se puede garantizar la calidad (proximidad al óptimo global) de la solución obtenida. Además, la probabilidad de encontrar la

solución global disminuye cuando aumenta el tamaño del problema (Lin *et al*, 2012).

Dependiendo de la estrategia de resolución, los métodos pueden ser divididos en técnicas de cálculo, técnicas de búsqueda y técnicas de convergencia de soluciones. De estos tres, el primero es el que menos se emplea en el campo de la ingeniería, dado que exige propiedades muy restrictivas a las funciones involucradas. Dentro de esta clasificación se podría incluir la programación dinámica, el método de búsqueda de gradiente y la optimización golosa. La primera corresponde al marco que se emplea cuando los problemas cuentan con estructura temporal (el problema se va subdividiendo iterativamente); la segunda se emplea cuando la función objetivo es diferenciable y estrictamente convexa y el óptimo puede ser encontrado con una condición de primer orden, y la tercera siempre prefiere ir al el siguiente paso que mejore la solución.

Dentro de las técnicas de búsqueda que no utilizan derivadas, se encuentran los métodos de optimización heurística, que siguen los siguientes pasos: empiezan con una solución más o menos arbitraria; iterativamente producen nuevas soluciones obtenidas utilizando ciertas reglas; evalúan estas nuevas soluciones, y eventualmente reportan la mejor solución encontrada durante el proceso de búsqueda (Maringer, 2005).

Dentro de las técnicas heurísticas se encuentran los algoritmos inspirados en la naturaleza que corresponden a técnicas computacionales basadas en la observación de los comportamientos en la naturaleza para resolver problemas complejos. Estos mismos se pueden dividir en cuatro categorías principales: de inteligencia colectiva (*swarm intelligence* o SI), bioinspirados (excluyendo los SI), basados en principios físicos/químicos, otras técnicas (Newman & Witt, 2010). En otra clasificación, los mismos se dividen en tipo Evolutivo (e.g. AG); SI (e.g. SI) o Ecológico [e.g. PS20 (Chen & Zhu, 2008)] (Binitha & Siva, 2012). También cabe mencionar los algoritmos basados en el comportamiento humano, tales como los algoritmos

meméticos (Baños *et al.* 2010 a-b) o los mecanismos de búsqueda de la harmonía (Geem *et al*. 2001), entre otros.

Un paso más avanzado lo constituyen las aproximaciones meta-heurísticas, que formalmente se definen como procesos iterativos de búsqueda que guían una heurística subordinada mediante la combinación de diferentes conceptos para explorar y explotar el espacio de búsqueda a fin de encontrar, de manera eficiente, soluciones ubicadas cerca del óptimo global (Osman & Laportem, 1996; Blum & Andrea, 2003).

III.4 Optimización Mediante Algoritmos Genéticos y Simulación Monte Carlo: Predicción de Nuevas Fugas Mediante Sectorización

III.4.1 Descripción de Algoritmos Genéticos

Los AG, planteados por John Holland en el año 1975 (Holland, 1975), pueden ser considerados como programas de computadora que mimetizan el proceso de evolución biológica a fin de resolver problemas de optimización. Contienen una solución potencial para un problema específico en una estructura de datos tipo cromosoma y aplican operadores de recombinación y mutación a estas estructuras a fin de preservar información crítica. Por lo general se les considera como optimizadores de funciones, aunque el rango de los problemas en que pueden ser implementados es muy amplio. Desde su creación en 1975, han aparecido una serie de variaciones adaptadas para resolver problemas de amplia índole. Gran parte del éxito de los mismos radica en su robustez para tratar problemas multi-objetivo o problemas con restricciones no lineales (Konak *et al*., 2006). Otras ventajas incluyen su capacidad para optimizar variables muy distintas, producidas por entradas de funciones analíticas, datos experimentales u outputs de modelos numéricos. También cuentan con la capacidad de optimizar ya sea valores reales, variables binarias o variables enteras. Pueden procesar un número muy grande de variables y pueden producir una lista de mejores variables, así como una mejor solución individual; en ocasiones, son buenos para encontrar un mínimo global en

lugar de un mínimo local; pueden muestrear simultáneamente varias porciones de una superficie de costo; se puede adaptar fácilmente a la computación paralela y se pueden emplear para resolver problemas de gran complejidad (Rojas, 1996).

El funcionamiento de estos algoritmos se basa en los siguientes principios (Palisade, 2010):

La evolución se desarrolla a nivel de cromosomas: los organismos no evolucionan. Sólo sirven de recipiente en el que se transportan los genes. Son los cromosomas los que cambian dinámicamente con cada re-arreglo de genes.

La naturaleza tiende a hacer más copias de cromosomas que están en organismos mejor ajustados: si un organismo sobrevive suficiente tiempo y es saludable, sus genes tienen más probabilidades de pasar a la siguiente generación mediante la reproducción. Este principio se conoce frecuentemente como *supervivencia del más apto*.

Se debe mantener la diversidad en la población: A fin de asegurar la variación entre los organismos se dan frecuentemente mutaciones. Estas mutaciones genéticas suelen resultar en características que son normalmente útiles para la supervivencia de los individuos. Con un espectro más amplio de posibles combinaciones, la población es menos susceptible a alguna debilidad común que la pueda destruir por completo.

La Ilustración 77 muestra el proceso seguido por un AG durante la optimización.

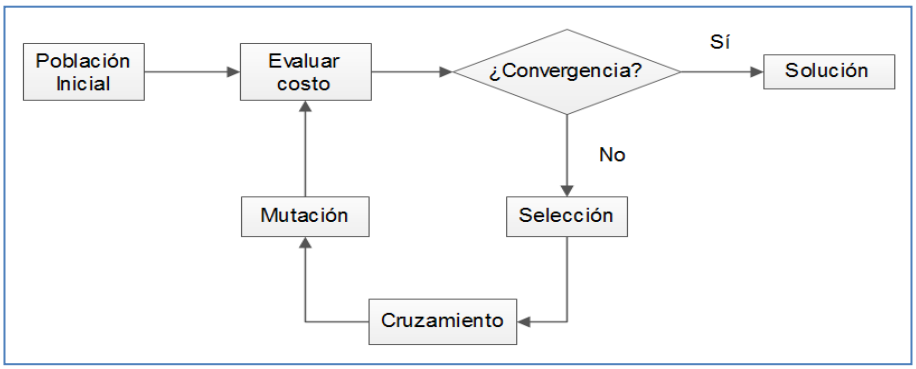


Ilustración 77: Proceso de funcionamiento de un AG

A continuación se hace una descripción de cada uno de los pasos establecidos en la Ilustración 77 (Mitchel, 1998; Coley, 1999).

Inicialización (población inicial): se crea una población inicial. Esta población se genera, en principio, aleatoriamente y puede ser del tamaño que se desee, desde un par de individuos hasta miles. Este es uno de los aspectos más críticos dentro de dominio de los AG. En general se establece que poblaciones muy pequeñas puede guiar el algoritmo a soluciones muy malas, en tanto, una población inicial muy grande, puede resultar en costes computacionales muy elevados. Sin embargo, una población inicial más grande, permite lograr contar con un nivel más alto de diversidad, que es otro aspecto clave para el funcionamiento adecuado de este método, tanto para evitar convergencia prematura, así como para ser establecido como criterio de parada. Por el contrario, si la población no es suficientemente diversa, entonces puede ser que no se alcance el óptimo global. Dado que la definición de una población inicial adecuada depende de la naturaleza del problema que se quiere abordar, la resolución de este problema aún sigue siendo un tema pendiente.

Evaluación de costes (aptitud): cada miembro de la población es evaluado y se calcula el ajuste o aptitud para ese individuo. El valor de ajuste se calcula en función de lo bien que el individuo se ajusta a los requerimientos establecidos. Estos requerimientos pueden ser simples o complejos. En la aplicación de AGs, la formulación de la función objetivo es de importancia crítica y determina la forma final de la superficie de búsqueda.

Selección: este procedimiento se ejecuta a fin de mejorar de manera constante la población. La selección descarta las soluciones malas de manera tal que se mantengan los mejores individuos en la población. En todos los métodos de selección la idea básica es la misma, hacer que los individuos mejor ajustados se preserven en las generaciones subsecuentes. Todas ellas se basan en un criterio de evolución que retorna una medida de calidad para cualquier cromosoma en el contexto del problema. Las técnicas de selección más populares son: la técnica de la ruleta, la selección por rango y la selección de estado estacionario.

Cruzamiento: durante el cruzamiento se crean nuevos individuos mediante la combinación de aspectos de los individuos seleccionados en la etapa anterior. Se puede pensar sobre esto como sobre la reproducción sexual en la naturaleza. Lo que se espera lograr es que mediante la combinación de ciertos rasgos de dos individuos se cree una población aún más apta que herede los mejores rasgos de ambos padres. En general, se identifica como la fuerza conductora de los AGs. El número de puntos de cruzamiento es decidido desde el inicio y se suele limitar a 1 o 2.

Mutación: se necesita agregar algo de aleatoriedad a la genética de la población, de otra manera, cada combinación de soluciones que se pueda crear, estaría en la población inicial, es decir, quedaría atrapada en óptimos locales. La mutación, típicamente, trabaja marcando cada pequeño cambio de manera aleatoria a los genomas individuales. Esto ayuda a mejorar la habilidad del AG para encontrar soluciones cerca del óptimo global para un problema a través de mantener un nivel suficiente de variedad genética en la población, lo que se requiere para asegurar que se haga una exploración a través de todo el espacio de búsqueda.

Iteración/convergencia: una vez que se tiene una nueva generación, se inicia el proceso desde el paso 2 hasta que se alcance una condición de finalización

III.4.2 Optimización Mediante Algoritmos Genéticos con Evolver

Evolver es una herramienta-complemento de optimización diseñada por la corporación Palisade® para el programa Excel (Palisade, 2017). A través de ella se pueden abordar diversos problemas de optimización, teniendo mejor capacidad que las utilidades de optimización que se encuentran instaladas por defecto en Excel (e.g. Solver). La herramienta funciona con AGs, y al instalarla en Excel (ver Ilustración 78), brinda una serie de posibilidades para hacer las adaptaciones necesarias del mencionado algoritmo al problema que se desee abordar. La idea es modelar el problema con todas las herramientas con las que cuenta Excel y luego establecer los requerimientos de la optimización (número de iteraciones, población inicial, restricciones, función objetivo, etc.) mediante la interfaz gráfica que provee Evolver.

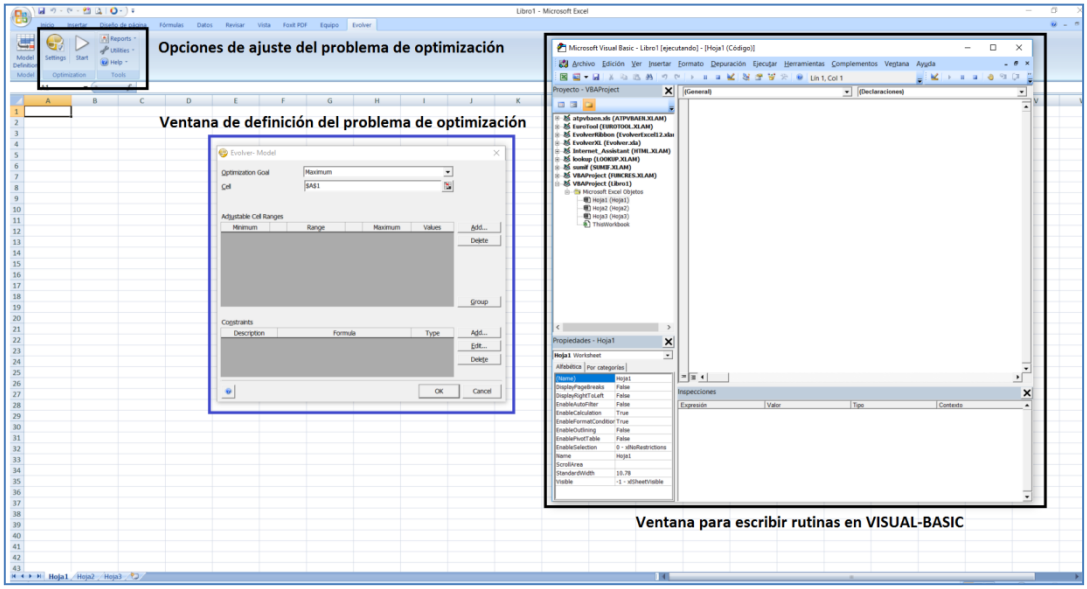


Ilustración 78: Herramienta Evolver anidada en Excel

Lamentablemente, la implementación de Evolver en Excel es únicamente mono-objetivo. En caso de querer abordar un problema con más de un objetivo, se debe seguir la estrategia de trasformación de problemas multi-objetivo en problemas mono-objetivo, arriba descrita. Luque (2013) y Millet García (2014), presentan ejemplos de implementación de la herramienta, en el campo de optimización de operación de RDAPs.

Antes de iniciar el proceso de optimización, una de las primeras tareas que hace el algoritmo, es un primer recorrido del espacio de búsqueda para encontrar una solución factible, en la cual se cumplan todas las restricciones, aunque el valor final adoptado por la función objetivo no se corresponda de ninguna manera con un óptimo. Esta es una alternativa que puede ser o no seleccionada por el usuario. En caso de que la misma no sea seleccionada, entonces el algoritmo procederá a la búsqueda desde un punto aleatorio.

Respecto a la visualización del proceso, esta se puede hacer a lo largo de toda la ejecución, de una manera completa o de una manera simplificada. Si se selecciona la manera corta, se puede ver el número de restricciones satisfechas, el número de restricciones no satisfechas, y el resultado de la función objetivo. En tanto, si se selecciona la opción reducida, se puede observar, además de la información anterior, un gráfico representativo del grado de diversidad existente entre las soluciones, un gráfico de evolución de la solución a lo largo de todas las iteraciones, los valores adoptados por todas las soluciones en una generación, y una barra desplegable que permite modificar la tasa de mutación o la tasa de cruzamiento.

La Ilustración 79 muestra el proceso de optimización que se sigue mediante Evolver en Excel.

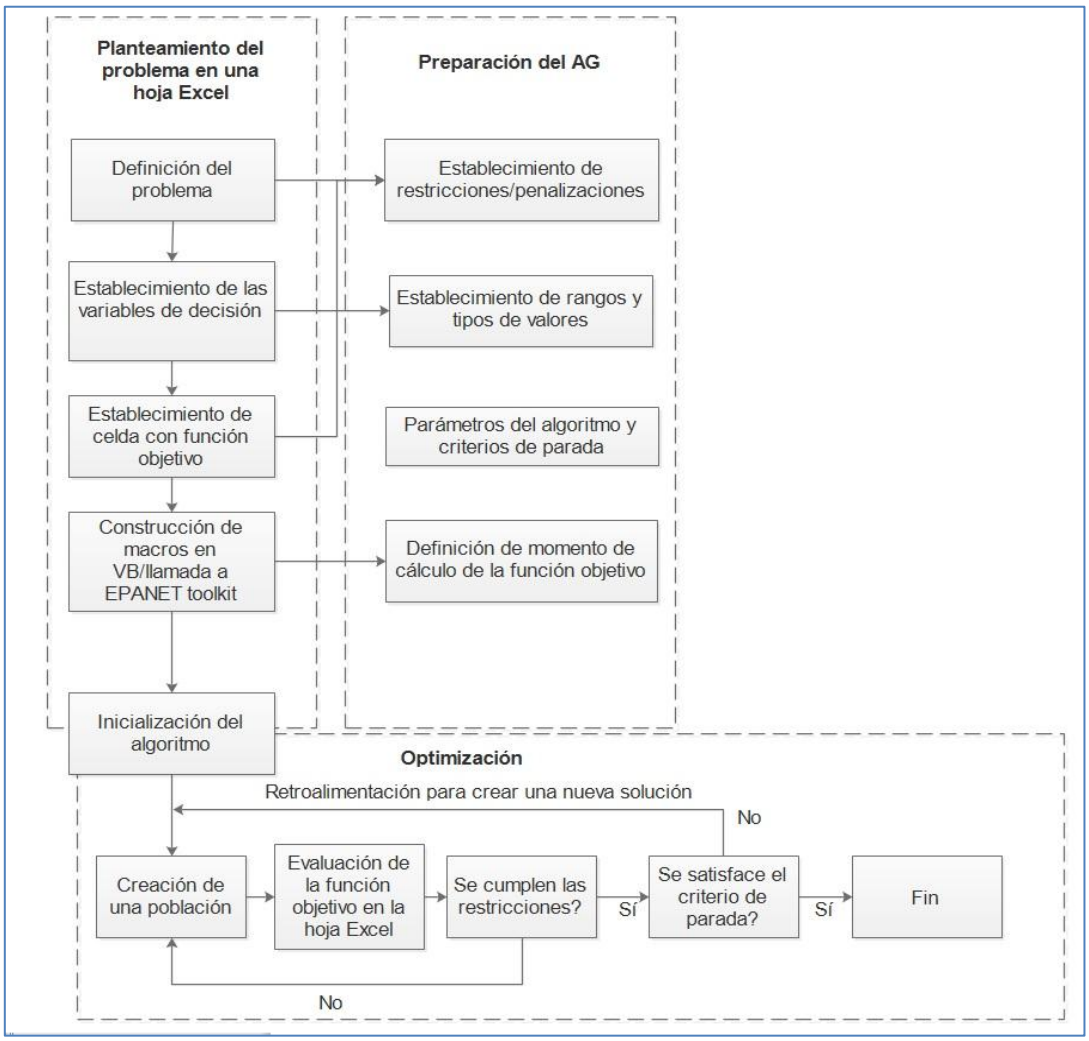


Ilustración 79: Proceso de optimización con Evolver en Excel

III.4.3 Optimización Mediante Algoritmos Genéticos y Simulación Monte Carlo

La aplicación RiskOptimizer de la ya mencionada compañía Palisade® es una extensión de la aplicación Evolver descrita en la subsección anterior. La misma está orientada a analizar/incluir factores que tienen incertidumbre/riesgo durante el proceso de optimización. Lo cual se logra a través de SMCs de las variables con incertidumbre/riesgo. En la SMC se calcula el resultado una y otra vez, utilizando un conjunto de valores aleatorios obtenidos a partir de una función de probabilidad. Dependiendo del número de incertidumbres y el rango especificado para los

valores, la simulación puede implicar miles o decenas de miles de re-cálculos antes de terminar, siendo el resultado una distribución de posibles valores de salida.

La aplicación RiskOptimizer en Excel es prácticamente igual que Evolver (ver Ilustración 80), con la salvedad de que cuenta con la probabilidad de seleccionar el método de muestreo de las variables aleatorias en la SMC y permite definir para cada una de las variables con incertidumbre el tipo de distribución de probabilidades a emplear (incluyendo el valor máximo, mínimo y medio). Adicionalmente, permite definir el número de iteraciones que se deben realizar en la SMC y el valor final de salida (media, máximo o mínimo).

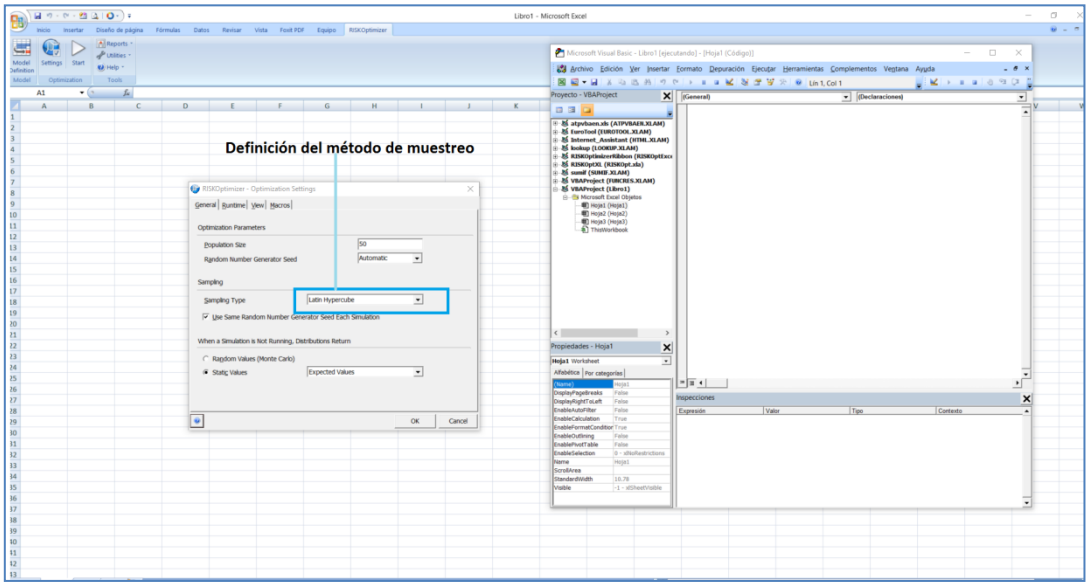


Ilustración 80: Aplicación RiskOptimizer anidada en Excel

III.4.3.1 Descripción del Método de Simulación Monte Carlo

De una manera muy general, la SMC puede describirse como un conjunto de métodos de simulación basados en la generación de valores aleatorios, que se incluyen como entradas en un problema dado. Los valores aleatorios son generados de manera artificial, emulando un proceso de muestreo realizado en la población actual. De esta manera, en el problema no sólo se evalúa el resultado con una suposición, sino como el resultado estadístico de repetir el experimento muchas veces con valores que son seleccionados al azar a partir de un rango probabilístico preestablecido. Cuando se implementa este método para resolver un problema dado,

la resolución del mismo se vuelve determinista-iterativa. La justificación del método SMC, se basa en la *Ley de los Grandes Números*, la cual describe el resultado de realizar el mismo experimento un número extenso de veces. Según la misma, el promedio de todos los resultados obtenidos a partir de un gran número de iteraciones debe ser próximo al valor esperado y, en la medida en que se aumente el número de iteraciones, la distancia entre el valor esperado y el promedio obtenido se va reduciendo cada vez más. A partir de esta ley se logra garantizar un grado de estabilidad a largo plazo en el resultado de eventos aleatorios. Tal como lo establece su mismo nombre, esta ley sólo es válida si el número de iteraciones es suficientemente grande.

Existen dos versiones de esta ley. Por un lado está la *Ley Fuerte de los Grandes Números* y, por otro lado, la *Ley Débil de los Grandes Números*. Ambas siguen el mismo concepto: el promedio de un número de iteraciones converge a un valor esperado,

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_1) \qquad \text{Ecuación 76}$$

$$\bar{X}_n \to \mu \; para \; n \to \infty$$

donde $\bar{X}_n$ corresponde al promedio de un número n de iteraciones realizadas y μ corresponde al valor esperado.

La *Ley Débil de los Grandes Números* establece que el promedio de la muestra de todos los resultados obtenidos mediante el conjunto de iteraciones, converge, en probabilidad al valor esperado, es decir $\bar{X}_n \xrightarrow{P} \mu$ para $n \to \infty$. Para cualquier número positivo ε, $\lim_{n\to\infty} Pr(|\bar{X}_n - \mu| > \varepsilon) = 0$. Según la misma, para cualquier margen diferente de cero, existirá una alta probabilidad de que el promedio de todas las observaciones esté cerca del valor esperado, siempre y cuando el número de iteraciones sea lo suficientemente grande.

Por el contrario, la *Ley fuerte de los Grandes Números* establece que el promedio de la muestra converge, de manera casi segura, al valor esperado, $\bar{X}_n \xrightarrow{a.s} \mu$ para $n \to \infty$, Lo que quiere decir que $\lim_{n\to\infty} Pr(|\bar{X}_n = \mu|) = 1$.

El método de SMC presenta muchas ventajas, tales como el hecho de que el resultado no sólo muestra qué puede pasar, sino la probabilidad de que pase cada resultado posible; en comparación con los análisis determinísticos, tiene la ventaja de poder definir qué variables tienen mayor impacto sobre el problema, lo que resulta de gran utilidad para llevar a cabo análisis adicionales; destaca por su simplicidad, flexibilidad y escalabilidad; es capaz de reducir sistemas muy complejos a un conjunto de acontecimientos básicos representados por una serie de iteraciones; puede ser paralelizable, lo que le permite abordar problemas muy complejos en tiempos considerablemente cortos; y al emplearlo en combinación con modelos estocásticos, ayuda a los algoritmos propios de estos modelos a escapar de óptimos locales, lo que permite mejorar los resultados de las búsquedas.

En el Pseudocódigo 6 se enumeran los pasos generales en la implementación la SMC.

1. Se crea un modelo paramétrico, $y = f(X_1, X_2, \dots, X_q)$.
2. Se genera una red de inputs aleatorios, $X_{i1}, X_{i2}, \dots, X_{iq}$.
3. Se evalúa el modelo y se almacena el resultado dentro del modelo paramétrico, como y_i.
4. Se repiten los pasos 2-3 para un número *n* de iteraciones o hasta un punto de convergencia.

Pseudocódigo 6: Funcionamiento de SMC

La distribución de probabilidades constituye la manera más realista de describir las incertidumbres en las variables en un análisis de riesgo. Algunas distribuciones de probabilidad comunes incluyen las siguientes.

Distribución Normal

También llamada *Curva de Campana Invertida*, en ella, se define la media o valor esperado y una desviación estándar para describir la varianza sobre la media. Los valores situados cerca de la media tienen más probabilidades de ocurrir. Es simétrica y describe muchos fenómenos naturales.

La *Distribución Normal* se define mediante la función de densidad de probabilidades presentada en la Ecuación 76,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \qquad \text{(Ecuación 76)}$$

donde μ es la media de la población y σ^2 es la variancia.

Cuando una variable aleatoria sigue una *Distribución Normal*, entonces $X \sim N(\mu, \sigma^2)$. En particular, la distribución normal con $\mu = 0$ y $\mu = 1$ es llamada la distribución estándar normal y se denota como $N(0, 1)$.

Este tipo de distribución tiene las siguientes características: es simétrica alrededor de la media; la media, la mediana y la moda son iguales; el área por debajo de la curva normal es igual a 1; es densa en el centro y menos densa en las colas; se caracteriza por la media y la desviación estándar; 68 % del área de la distribución se localiza dentro de una desviación estándar de la media, aproximadamente el 95 % del área de la distribución normal está dentro de dos desviaciones estándares de la media, y el 97 % de los datos cae dentro de tres desviaciones estándares de la media

Distribución Lognormal

Los valores están un poco sesgados positivamente y no son simétricos como en el caso de la *Distribución Normal*. Es utilizada para representar valores que no van por

debajo de cero y que tienen potencial positivo ilimitado. Una variable cuenta con una *Distribución Lognormal* sí $Y = ln(x)$ está distribuida normalmente. La Ecuación 77 presenta la fórmula general para su función de densidad de probabilidades

$$f(x) = \frac{e^{-((\ln(x-\theta)/m))^2/(2\sigma^2))}}{(x-\theta)\sigma\sqrt{2\pi}} \qquad \text{(Ecuación 77)}$$

Aquí $x > \theta; m;\ \sigma > 0; \sigma$ es el parámetro de forma (y es la desviación estandar de de la distribución; θ es el parámetro de localización y m es el parámetro de escalación (y es además la mediana de la distribución). Si $x = \theta$, entonces $f(x) = 0$. El caso donde $\theta = 0$ y $m = 1$ es llamado *Distribución Estándar Lognormal*.

Distribución Uniforme

En este tipo de distribución, todos los valores tienen igual probabilidad de ocurrir. Para su construcción sólo se necesita un valor mínimo y un valor máximo. El hecho de que todos los eventos tengan igual probabilidad de ocurrir, hace que la distribución no tenga moda y por ende, que su gráfica corresponda a un rectángulo. Por otro lado, al no haber sesgo (*skewness*) para la distribución uniforme, la media y la mediana coinciden.

La función de probabilidad de densidad en este caso viene dada por la Ecuación 78.

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a \le x < b \\ 0, & x < a \text{ o } x > b \end{cases} \qquad \text{(Ecuación 78)}$$

Distribución Triangular

En este tipo de distribución se define un valor mínimo (más probable) y un valor máximo. Los valores cercanos al más probable son los valores que tienen más probabilidades de ocurrir.

La función de densidad de esta distribución viene dada por la Ecuación 80.

$$p(x) = \begin{cases} \dfrac{2(x-a)}{(b-c)(c-a)}, & \forall\ a \leq x \leq c \\ \dfrac{2(b-x)}{(b-a)(b-c)}, & \forall\ c \leq x \leq b \end{cases} \qquad \text{(Ecuación 80)}$$

Y su función de distribución viene dada por

$$d(x) = \begin{cases} \dfrac{(x-a)^2}{(b-a)(c-a)}\ para\ a \leq x \leq c \\ 1 - \dfrac{(b-x)^2}{(b-a)(b-c)}\ para\ c \leq x \leq b \end{cases} \qquad \text{(Ecuación 81)}$$

donde $c \in [a, b]$ representa la moda.

<u>Función PERT</u>

Para la definición de este tipo de función de distribución, se tiene que definir un valor mínimo, un valor máximo y un valor más probable, tal y como ocurre en el caso de la *Distribución Triangular*. Nuevamente, los valores alrededor del valor más probable tienen más probabilidad de ocurrir; sin embargo, el resto de los valores tienen más oportunidades de ocurrir que en el caso de la *Distribución Triangular*. En otras palabras la "*Distribución PERT*" construye una curva suavizada que coloca progresivamente más énfasis en valores alrededor (cerca) del valor más probable, a favor de los valores que se encuentran alrededor de los límites y así se tiene la expectativa de que el valor resultante esté más cerca al valor estimado.

Tomando en cuenta que m es el valor más probable y que cumple con la siguiente restricción $a \leq m \leq b$, La función de densidad de probabilidades viene dada por la Ecuación 82,

$$f(x) = \frac{1}{B\ (\alpha_1, \alpha_2)} \frac{(x-a)^{\alpha_1 - 1}(b-x)^{\alpha_2 - 1}}{(b-a)^{\alpha_1 - \alpha_{2-1}}} \qquad \text{(Ecuación 82)}$$

donde $\alpha_1 = \frac{4m+b-5a}{b-a}$ y $\alpha_2 = \frac{5b-a-4\,m}{b-a}$.

III.4.3.2 Método de Muestreo en Simulación Monte Carlo

Al implementar la SMC, existen varias alternativas para llevar a cabo la toma de muestras que se evalúan en cada iteración. El método tradicional corresponde al *Método de Muestreo Monte Carlo* (MMC); sin embargo existen otras alternativas que destacan por tener mayor eficiencia, tal como es el caso del "*Método Hipercubo Latino*" (MHL) (Mackey *et al*., 1979). A continuación se hace una descripción de ambos métodos.

Método de Muestreo Monte Carlo

Este tipo de muestreo se refiere a la técnica tradicional para utilizar valores aleatorios o semi-aleatorios a fin de muestrear a partir de una distribución de probabilidad. En esta técnica, cada muestra dada puede caer en cualquier lugar dentro del rango de la distribución de entrada. Las muestras son más probables a salir de áreas de la distribución que tienen probabilidades más altas de ocurrencia. En general, con un número suficientemente grande de iteraciones, el MMC puede recrear las distribuciones de entradas a través de muestreo. Sin embargo, un problema relacionado con la clusterización ocurre cuando el número de iteraciones en muy bajo (Lee *et al*., 2006). Al existir salidas de baja probabilidades y un número muy reducido de iteraciones, al método le resulta difícil muestrear dentro de los clústeres con elementos de baja probabilidad, lo que hace que la salida resultante tenga cierto grado de sesgo.

Muestreo Hipercubo Latino

En el contexto de muestreo estadístico, un mallado cuadrado conteniendo posiciones de muestreo es un *Cuadrado Latino* (*Latin Square*) si (y sólo si) hay sólo una muestra en cada fila y cada columna. Un *Hipercubo Latino* (HL) es una generalización de este concepto a un número arbitrario de dimensiones, en las

cuales cada muestra es la única en cada hiperplano-eje alineado que la contiene. Cuando se muestrea una función de variables, el rango de cada variable es dividido en intervalos probables iguales. Entonces se colocan los puntos de muestreo para satisfacer los requerimientos del HL; nótese que esto fuerza el número de divisiones a ser igual para cada variable. También, nótese que este esquema de muestreo no requiere más muestras para más dimensiones (variables), siendo esta independencia una de las principales ventajas de este esquema de muestreo (Helton & Davis, 2003). Otra ventaja es que las muestras aleatorias pueden ser tomadas una a la vez, quedando registrado qué muestra ya ha sido previamente tomada. En el Muestreo Hipercubo Latino (MHL) se emplea muestreo sin reemplazo. El número de estratificaciones de la distribución acumulada (cumulativa) es igual al número de iteraciones llevadas a cabo. La idea es tomar una muestra desde cada una de las estratificaciones y, una vez que se toma una muestra de una estratificación, dicha estratificación no se vuelve a tomar en cuenta. Cuando el proceso de muestreo se lleva a cabo sobre varias variables a la vez, la muestra de cada una de las variables en un tiempo determinado es realizada en distintos estratos al mismo tiempo, evitando así correlaciones no deseadas.

La Ecuación 83 permite calcular el número máximo de combinaciones para un HL de M divisiones y N variables (dimensiones).

$$\left(\prod_{n=0}^{M-1}(M-n)\right)^{N-1} = (M!)^{N-1} \qquad \text{(Ecuación 83)}$$

La diferencia de este muestreo con respecto al muestreo aleatorio, radica en el hecho de que los puntos de muestreo deben ser decididos previamente y para cada muestra tomada se tiene que dejar registro de la fila y la columna de donde se tomó.

III.4.4 Simulación Monte Carlo en Redes de Abastecimiento de Agua Potable

En el desarrollo de modelos hidráulicos de RDAPs existe un alto nivel de incertidumbre asociado a parámetros cuyo valor es difícil de estimar, dado el

entorno de donde surgen los valores en cuestión. Para empezar, se puede hacer mención de dos parámetros críticos para el funcionamiento de un modelo matemático adecuado, tales como los valores de rugosidad en las tuberías y los valores de coeficiente de emisor en los nudos.

A pesar de que es posible estimar un valor de coeficiente de emisor global que permita balancear adecuadamente los caudales en las RDAPs, la distribución del mismo a través de los nodos de la red da lugar a un alto grado de incertidumbre. Puede existir un número considerablemente grande de alternativas de distribución de los valores de coeficiente de emisor, que aumenta en la medida que crece el número de tuberías en la red. A pesar de que más de una de las combinaciones puede hacer que el modelo funcione apropiadamente, no existe una garantía de que la distribución seleccionada se acerca a la realidad. Lo mismo es aplicable a las rugosidades en las tuberías, cuya apropiada distribución es importante debido al hecho de que las fugas dependen de los valores de presión. Al cambiar los valores de presión cambian también los caudales asociados a las fugas y, por ende, la asignación previa de coeficiente de emisor se ve afectada (ver más detalles sobre el tema en Apéndice I: Método de Calibración Mediante Mapas Auto Organizados y Algoritmos Genéticos).

No obstante, no sólo se puede hablar de incertidumbre asociada a valores que se generan a partir de elementos físicos en las RDAPs. También, las predicciones que se pueden hacer suelen contener un alto grado de incertidumbre; ejemplo de ello es la predicción de la ocurrencia de un evento de fuga en una zona dada de la red.

La probabilidad de que ocurra un evento de fuga en un tiempo T, puede ser expresada, matemáticamente, como una función del tamaño del sector (en términos de longitud de tubería); del número de entradas con las que cuenta (por ende, indirectamente; del número de tuberías cerradas); de la edad de las tuberías, y de las variaciones de presiones que se dan dentro del mismo.

$$|P| = f(L^t, \Delta P^n, E^A) \qquad \text{(Ecuación 84)}$$

$$\forall S\ 0 < |P| < 1\ , E^A < E^T y \sum_{a=1}^{i} E^A = E^T$$

Aquí L^t corresponde a longitud de tubería; ΔP^n corresponde a variación de presión y E^A corresponde la cantidad de entradas del sector.

Evidentemente, estos son parámetros difíciles de ser evaluados, por lo que una manera de tener una idea de ellos es realizar un análisis de registros históricos o analizar RDAPs que cuenten con características similares. También se puede asignar un número dado de roturas por año a cada sector, que sea igual a una fracción del número de roturas esperado a lo largo de toda la red. Para definir la fracción correspondiente a cada sector, se puede hacer uso del *Análisis Jerárquico de Prioridades* (AHP por *Analytic Hierarchical Process*) (Osirio & Orejuela, 2008) a partir del cual se obtiene un vector de pesos correspondiente a la fracción de roturas esperadas en cada sector cada año. En Herrera (2011); Delgado-Galván (2011); Delgado-Galván *et al* (2010) y Benítez *et al* (2012, 2014, 2015); Ilaya-Ayza *et al* (2016 a,b) se presentan amplias descripciones sobre la implementación de la técnica para abordar problemas propios de RDAPs. Una vez asignada la fracción de roturas que le corresponde a cada uno de los sectores, se realiza un análisis de la probabilidad de que esa cantidad de roturas ocurra dentro de cada sector en un tiempo dado, lo cual dependerá de los parámetros previamente mencionados.

El segundo aspecto a tratar, concierne a la probabilidad de que un evento de fuga dado sea detectado y reparado en un tiempo determinado. Esto es especialmente importante para la categoría de fugas no reportadas, en vista de que a lo que a roturas reportadas se refiere, el periodo de detección y reparación se espera que sea significativamente más corto (en vista de que las operadoras pueden ser notificadas por los mismos usuarios).

Por norma general, las roturas no reportadas sólo son detectadas cuando las empresas operadoras de RDAPs se plantean la realización de una campaña de detección de fugas. No obstante, al operar una red bajo un esquema de

sectorización, se espera que la situación sea distinta, ya que, al contar con capacidad de monitoreo permanente del caudal que entra a cada uno de los sectores, es posible reconocer eventos de fugas a partir de las variaciones en los comportamiento del caudal y, en el caso de que el caudal de pérdida sea categorizado como muy importante, el personal técnico puede proceder a lanzar una campaña rápida de localización y reparación de la fuga.

Tal y como se observa en la Ecuación 85, ahora la probabilidad de detección de un nuevo evento de fuga es una función del caudal que entra al mismo y de su longitud de tubería. En sectores con mayor longitud de tubería y mayor número de UOCs, la detección de la ubicación del punto en donde ocurre la fuga y, por ende, su reparación, toma más tiempo que en el caso de un sector más pequeño con un número reducido de UOCs. Los caudales que se miden en cada sector son también importantes, ya que cuanto más elevados sean los valores de caudal, tanto más difícil es que la alteración generada por el evento de fuga sea identificada en la curva de entrada de caudal del sector. Lo mismo ocurre en el caso que el sector cuente con muchas UOCs, ya que cuanto mayor sea el número de las mismas, más difícil resulta el reconocer una variación significativa de caudal.

$$|P| = f(n^v, L^t, Q^m) \qquad \text{(Ecuación 85)}$$

Otro aspecto clave, en lo que a detección de eventos se refiere, es la representación de la probabilidad de ocurrencia y de detección/reparación. Tales probabilidades pueden ser representadas de dos maneras distintas, ya sea como un valor fijo, o como un rango de valores probabilísticos muestreados a partir de una curva de distribución. Tal y como se anunció al principio de esta sección, es de esperar un alto grado de incertidumbre cuando se intenta realizar este tipo de predicciones en RDAPs. Con lo cual, el uso de un valor fijo puede conducir a predicciones muy engañosas, siendo totalmente factible el uso de SMC para abordar esta tarea.

En gran medida, la obtención de buenos resultados a partir del muestreo probabilístico depende de la curva de distribución que se emplee para cada uno de los sectores. Una posible solución ante la dificultad que en muchos casos implica la

obtención de datos históricos de las RDAPs para poder establecer curvas de distribución de ocurrencia y reparación de eventos de fugas, es el establecimiento de un sector piloto a partir del cual se puedan inferir patrones de comportamiento; sin embargo, no todas las empresas operadoras de RDAPs cuentan con los recursos económicos y la capacidad técnica para llevar a cabo esta tarea. En este último caso, se puede pensar en utilizar información obtenida a partir de otras RDAPs en donde este tipo de proyecto se hubiese realizado.

No sería correcto pensar que en todos los sectores se detectará el 100% de los eventos de fugas que se espera que ocurran y, adicionalmente, no sería lógico pensar que el tiempo desde el momento de la detección hasta el momento de la reparación pueda ser igual a cero. El considerar este tiempo es de suma importancia, ya que cuanto más amplio sea el mismo, mayor es el caudal que se pierde a través de las fugas, especialmente en las horas de la noche, cuando la presión alcanza su pico máximo.

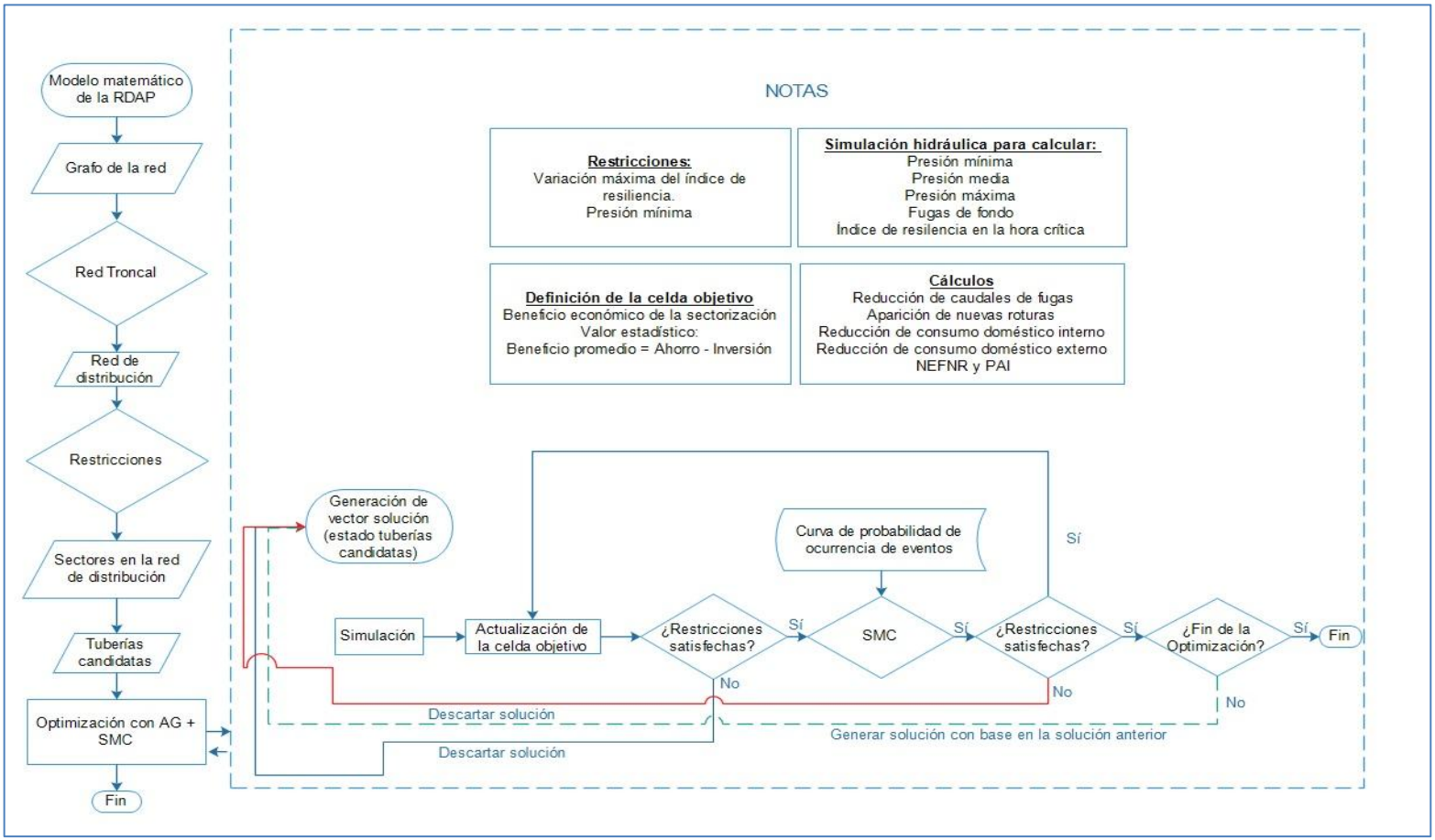


Ilustración 81: Método de optimización del CEVC con base en algoritmos genéticos y simulación Monte Carlo

III.4.5Ejemplo de Implementación de Optimización del Conjunto de Válvulas de Cierre/Entrada de Sectores Mediante Algoritmos Genéticos y Simulación Monte Carlo

Para el ejemplo de implementación se emplea el resultado obtenido mediante la definición de sectores obtenido a través del método basado en caminos aleatorios (ver Ilustración 55), en el cual se obtuvieron 10 sectores con longitud de tubería entre 13 y 32 km. La red en cuestión cuenta con un único patrón de demanda. Se caracteriza por tener sobrepresiones en puntos extremos, con valores de presión máxima que llegan a 60 m en el punto más bajo del patrón de demanda; 19 m en el punto más alto, y 46 m en el punto promedio. En la Tabla 14 se muestran los valores de fugas de la red antes de implementar la sectorización.

Tabla 14: Valores de fugas definidos para la red ejemplo

Parámetro	Valor
Fugas de fondo (m^3/año)	378233
Fugas reportadas (m^3/año)	264753
Fugas no reportadas (m^3/año)	6754
Roturas reportadas (roturas/año)	333
Roturas no reportadas (roturas/año)	200
Consumo interno (m^3/año)	7943600
Consumo externo (m^3/año)	794360

En la Tabla 15 se muestran los costes unitarios que fueron definidos.

Tabla 15: Costes de operación

Coste	Valor
Coste de agua ($/m^3)	0.6
Coste de reparación ($/rotura)	200
Coste de transporte por cisterna ($/m^3)	2
Coste de inspecciones ($/inspección)	20000

Pese a que el costo de reparación depende del diámetro de la tubería y de su ubicación, se utilizó el mismo coste para todos los diámetros, en vista de que no se contaba con información al respecto.

En la Tabla 16 se muestran los costes de compra e instalación de las válvulas de aislamiento y de las UOCs.

Tabla 16: Costos de UOCs y válvulas

Diámetro (mm)	UOC ($/unidad)	Válvulas ($/unidad)
25	2237	375
50	4475	750
75	6712	1125
100	8950	1500
150	13425	2250
200	17901	3000
250	22376	3750
300	26851	4500
350	31326	5250
400	35802	6000
450	40277	6750
500	42514	7125

Se definió una vida útil de 10 y 5 años para las UOCs y válvulas, respectivamente. En el caso de las válvulas, se definió una vida útil menor en vista de que al estar enterradas, su mantenimiento es más difícil y, por ende, menos regular. Para ambos caso se definió un interés anual equivalente al 20%. Respecto al coste de

mantenimiento, el mismo se definió como un porcentaje (10%) del descuento anual respectivo.

Se estableció como restricción el índice de resiliencia de la red y la presión mínima. En el segundo de los casos, no se estableció directamente el valor de presión mínima como restricción, sino que se estableció un coste para suplir la demanda de los nodos en donde no se logra cumplir el valor mínimo de presión. Dicho coste corresponde al ítem "coste de transporte por cisterna" que aparece en la Tabla 15. El valor máximo de variación del índice de resiliencia establecido como restricción correspondió a 50%.

Para el proceso de optimización se definió una población equivalente a 260 individuos (equivalente a 10 veces el número de variables de decisión – tuberías candidatas –), una tasa de mutación del 0.5 % y una tasa de cruzamiento equivalente al 50 %.

Para simular la ocurrencia y detección de eventos de fugas, se establecieron distribuciones de probabilidad triangulares en función de la longitud de tubería de cada sector y el número de entradas, tal como se muestra en la Tabla 17. En la misma, el valor en cada celda de la columna "porcentaje de detección" está conformado por un valor mínimo, un máximo y un valor más probable. En el presente ejemplo se empleó la misma tabla para el porcentaje de detección de fugas reportadas y de fugas no reportadas, aunque se espera que los porcentajes de detección de fugas reportadas sean mayores, en vista de que dichas fugas son visibles, en tanto, las otras se reconocen, inicialmente, por cambios en los valores de caudal que se registran en las UOCs, pero luego tienen que ser geográficamente detectadas por medios acústicos.

Tabla 17: Porcentajes de detección en función de la longitud de tubería de sector

Longitud de sector (km de tuberías)	Número de UOCs		Porcentaje de detección
10	0	1	0.5; 0.4;0.2
10	2	3	0.45;0.37;0.2
10	3	5	0.25;0.27;0.15
10	5	8	0.24;0.22;0.14
10	9	12	0.17; 0.15; 0.13
20	0	1	0.4;0.38;0.3
20	2	3	0.38;0.35;0.30
20	3	5	0.28;0.18;0.16
20	5	8	0.19;0.16;0.14
20	9	12	0.14;0.12;0.09

En la solución final, de las 26 tuberías candidatas, 14 se definieron como entradas de sectores y el resto como límites de sector (con válvulas cerradas), generando una reducción del índice de resiliencia equivalente a 43.8 % (ver Ilustración 83). La Tabla 18 muestra el resultado de uniformidad de presiones, presión promedio, presión mínima y presión máxima de la configuración final.

Tabla 18: Resultados de presión obtenidos mediante optimización con AG

Sector	1	2	3	4	5	6	7	8	9	10
Índice de uniformidad de presión	27.80	64.00	66.42	28.30	37.00	62.0	62.14	31.40	50.00	62.00
Presión promedio	24.32	28.85	30.10	30.10	27.00	31.59	26.44	31.36	25.65	31.08
Presión mínima	13.00	11.00	12.0	11.99	10.00	10.22	11.00	11.00	11.0	12.15
Presión máxima	57.50	57.30	53.30	52.60	52.73	54.56	52.45	54.28	57.15	59.35

La Tabla 19 muestra el balance coste beneficio de la optimización. Al resultar la presión en todos los nodos mayor al valor de presión mínimo establecido (mayor a 10 m), el ítem correspondiente a compensación por déficit de presión resultó 0.

Tabla 19: Balance coste/beneficio obtenido

Coste	Coste ($)
Inversión en válvulas y caudalímetros ($/año)	63405.5
Compensación por déficit	0.0
Coste total ($/año)	**63405**
Beneficio	**Beneficio ($)**
Ahorro en caudal fugas de fondo ($/año)	46519
Ahorro en roturas reportadas ($/año)	26640
Ahorro en roturas no reportadas ($/año)	16000
Ahorro en caudal por roturas reportadas ($/año)	20933
Ahorro por inspecciones ($/año)	2846
Variación de energía	0
Ahorro en consumo interno ($/año)	47661.6
Ahorro en consumo externo ($/año)	4766
Ahorro total ($/año)	**165368**
Beneficio Total ($/año)	**104764**

En la mejor solución generada se generó un beneficio neto anual de 104764 $/año (ver Ilustración 82 e Ilustración 83), que es astronómicamente mayor que el beneficio que se obtendría si no se redujera la presión (si se instalaran UOCs en cada una de las entradas de los sectores). En este último caso el beneficio neto sería sólo de 4134 $/año (ver Ilustración 82). Siguiendo el concepto de gestión sostenible, en la red inicial cada año se tiene que inspeccionar un 27%, en tanto, en el resultado final se debe inspeccionar solo 12%. Nótese que si se comparara el coste total 63405 $/año únicamente con el beneficio que se obtendría por el ahorro en reducción de fugas de fondo 46519 $/año, el proyecto no sería factible.

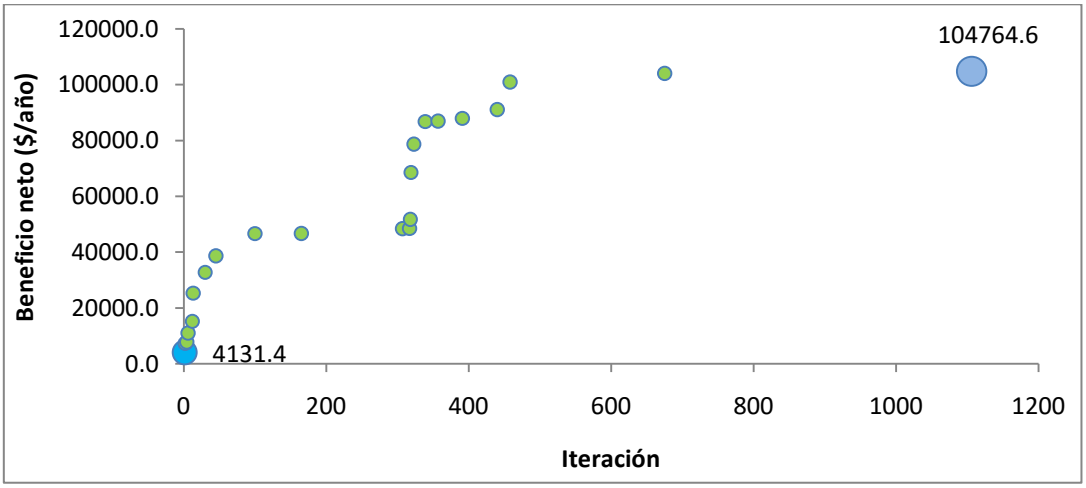


Ilustración 82: Progreso de optimización

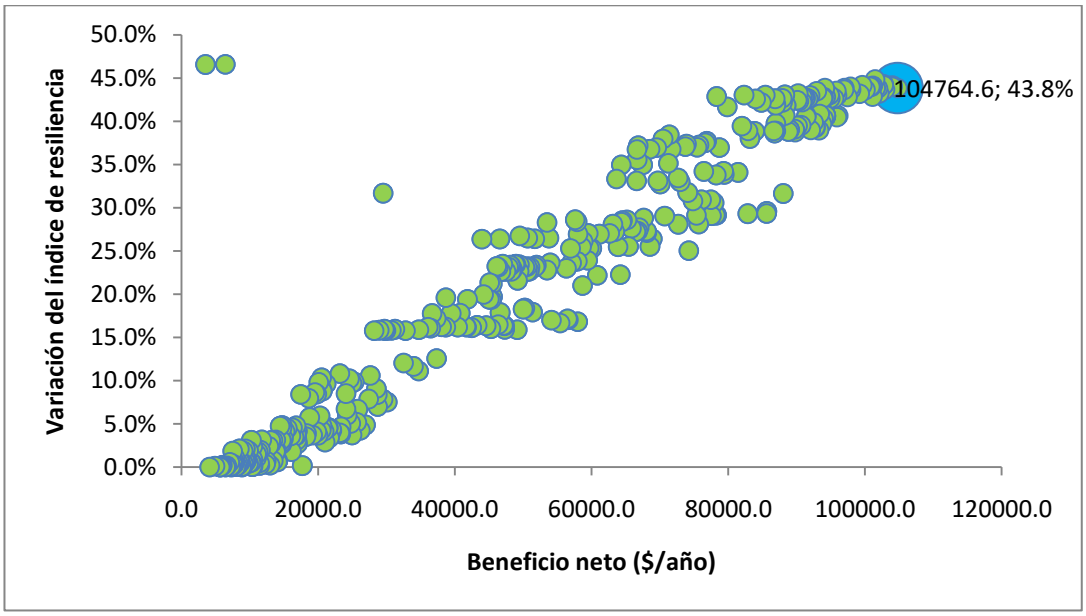


Ilustración 83: Beneficio neto vs variación del índice de resiliencia

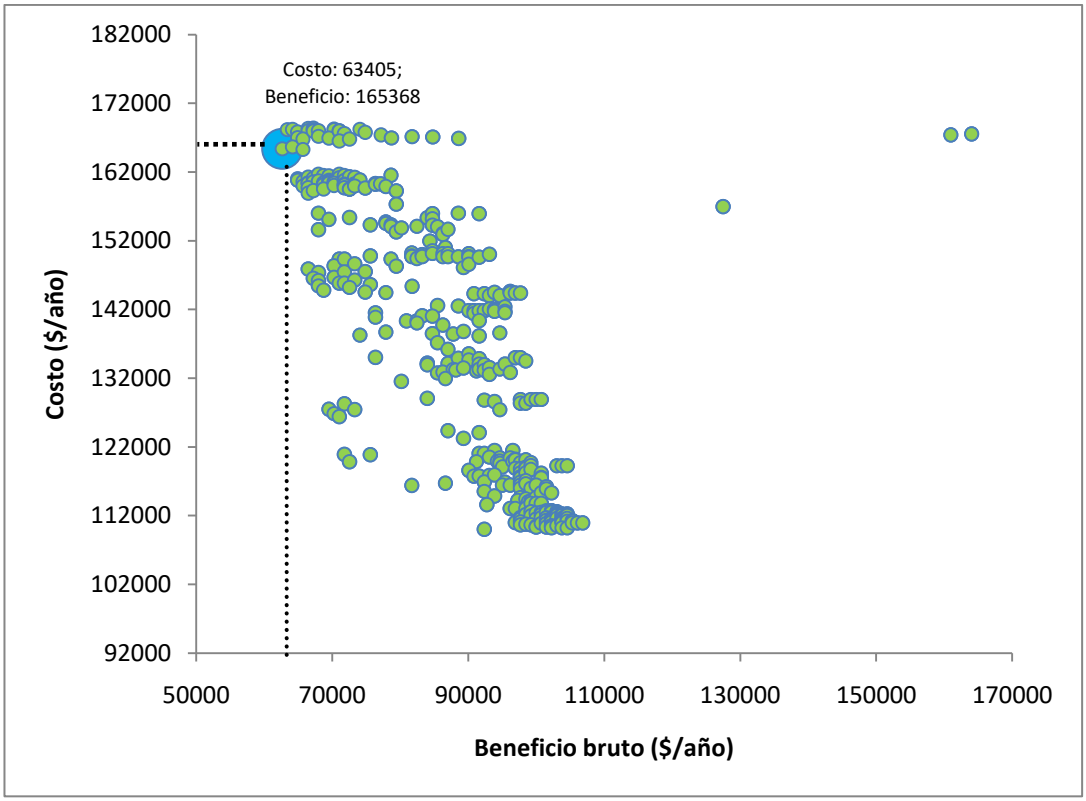


Ilustración 84: Coste vs beneficio bruto

Si se siguiera estrictamente lo establecido en el modelo conceptual de reducción de roturas presentado en la Subsección III.1.6, la reducción en la frecuencia de roturas sólo puede ser calculada mediante la reducción de la presión máxima de toda la red (comparando cada uno de los nodos antes y después del cierre de líneas candidatas).

Sin embargo, al ejecutar el ejemplo, se hizo notorio que en los nodos ubicados en las cotas más bajas, la presión máxima incluso se incrementaba levemente, con lo cual no era posible calcular reducción en la frecuencia de aparición de roturas. Dos alternativas se pueden emplear para resolver este problema. Por un lado se puede adoptar una estrategia más conservadora, incluyendo dentro de la optimización la colocación de VRPs para garantizar una reducción de la máxima presión en toda la red. Por otro lado, se puede hacer el cálculo de la reducción de frecuencia de roturas no con el valor máximo de presión, sino con el promedio de presión de todos los nodos en el instante en que la curva de demanda se encuentre en su punto más bajo. Para fines de ejemplificación se optó por la segunda de las alternativas, reduciéndose la presión máxima en la solución final, de 46 m a 30.46 m.

En el caso de la energía, no se reportó ningún cambio significativo en el costo de la energía, a pesar de que mediante la optimización se reduce el caudal de fugas, lo que implica una reducción del caudal que debe ser enviado por las bombas. Esto se debe al hecho de que las bombas operan con velocidad fija.

Tal y como se señaló previamente, una de las grandes desventajas de la metodología es la recolección de información. Los datos que fueron incluidos corresponden a estimaciones generales obtenidas mediante discusiones con el personal encargado de la operación de la RDAP de Managua. En particular, resulta especialmente arriesgada la definición del caudal que le corresponde a cada uno de los sectores y la predicción de detección de eventos de fugas dentro de cada uno de los sectores, aun cuando el error se trate de reducir mediante la utilización de la SMC. No obstante, tal y como se ha visto, la diferencia en el balance costo beneficio puede variar significativamente si los aspectos planteados son incluidos dentro de la optimización.

Finalmente, vale la pena destacar que para que la optimización acá planteada sea efectiva (realista) es necesario que la operadora de la RDAP cuente con un programa de gestión sostenible de fugas. En caso contrario, los únicos beneficios que se obtendrían son los que están exclusivamente ligados a la reducción de las presiones.

III.5 Inclusión de Válvulas Reguladoras de Presión en las Entradas de Sectores Mediante Optimización Multinivel

El *Algoritmo el Enjambre de Partículas* (PSO por las siglas en inglés de *Particle Swarm Optimization*) propuesto por Eberhart & Kennedy (1995) constituye una aproximación heurística que cae dentro de la clasificación de algoritmos evolutivos bio-inspirados y a su vez, sigue el paradigma de inteligencia colectiva. El mismo funciona mimetizando una bandada de aves en la naturaleza, constituyendo cada una de las aves (individuos) una potencial solución. En la búsqueda de alimentos, la bandada tiende a seguir al individuo o a los individuos que primero han localizado el alimento. En el contexto de optimización, la partículas se mueven por el espacio de búsqueda tratando de encontrar la solución que optimice (minimice o maximice) la función objetivo. Cada una de las partículas en el algoritmo, tiene dos vectores asociados: un vector de posición (x) y un vector de velocidad (v). Por lo general, cada vector comienza al azar dentro de un rango definido. El vector de posición se interpreta como una solución al problema y permite la evaluación de la función objetivo. La actualización de ambos vectores se lleva a cabo según lo que se establece en las Ecuaciones 86 - 87:

$$x_s^{t+1} = x_s^t + v_s^{t+1} \qquad \text{(Ecuación 86)}$$

$$v_s^{t+1} = w^t . v_s^t + c_1 . r_1 \left| l_b^t - x_s^t \right| + c_2 . r_2 \left| g_b^t - x_s^t \right| \qquad \text{(Ecuación 87)}$$

donde x_s^{t+1} es la posición de la partícula s en la iteración $t + 1$, actualizada mediante la utilización del vector de velocidad v_s^{t+1}.

La actualización de la velocidad es una combinación del último valor de velocidad, ponderado por el parámetro de inercia w^t, que permite evitar la excesiva itinerancia de partículas; la diferencia entre la mejor posición de la partícula s, l_b^t,, y la posición actual, ponderada por el parámetro cognitivo c_1; y, por último, la diferencia entre la posición actual de la partícula y la posición de la partícula líder del enjambre, g_b^t, ponderada por el parámetro social c_2. Los números aleatorios r_1 y r_2, trabajando como elementos dispersivos, evitan la convergencia prematura en los puntos óptimos locales. Los dos últimos sumandos en la Ecuación 87 son

responsables de la convergencia del método, ya que atraen a las partículas al mejor punto encontrado por el enjambre.

Tal como lo describe Montalvo (2008), el algoritmo PSO, al igual que otras técnicas de optimización, no garantizan la localización de un óptimo global y, de hecho, en función de la complejidad del problema, puede padecer de una tendencia a quedarse prematuramente atrapado en óptimos locales. Sin embargo, a través de él es muy probable encontrar al menos una buena solución. Lo cual, junto a la sencillez de su implementación, lo hacen una técnica muy atractiva.

Tal como se mencionó anteriormente, lo problemas de optimización pueden ser mono o multi-objetivo. Sin embargo, dependiendo de la complejidad del problema a tratar, las aproximaciones multi-objetivo pueden no tener la capacidad para abordarlos. En tal sentido, una posible solución podría incluir un abordaje multinivel, de tal manera que cada uno de los objetivos es optimizado en diferentes niveles, utilizando el aprendizaje obtenido en el primer nivel para agilizar la optimización de los objetivos en un nivel más elevado.

Un problema de optimización multinivel *P* con variables de decisión x_k, funciones objetivos f_k, y restricciones g_k, puede ser descrito tal como lo establece el siguiente conjunto de ecuaciones.

Encontrar la solución para el nivel superior del problema

$$P(1) = min_{x_1} f_1(x_1, x_2 \dots x_k), \qquad \text{(Ecuación 88)}$$

Sujeto a: $g_1(x_1, x_2 \dots x_k) \leq 0$,

donde x_1 es la solución para el segundo nivel del problema:

$$P(2) = min_{x_2} f_2(x_1, x_2 \dots x_k), \qquad \text{(Ecuación 89)}$$

Sujeto a: $g_2(x_1, x_2 \dots x_k) \leq 0$,

y así sucesivamente, hasta el nivel de problema k, con solución x_k.

Este marco de optimización ha sido planteado por estudiantes de doctorado en el Laboratorio de Hidráulica Computacional (LHC) en la Universidad Estatal de Campinas (UNIVCAMP), en Campinas, Brasil. Durante esta tesis, se presentó en colaboración con este grupo, una solución al problema planteando para la competición llamada Batalla de la Sectorización de Redes de Abastecimiento de Agua Potable (Brentan *et al.*, 2016), llevada a cabo en el marco del congreso WDSA en Cartagena de Indias, Colombia, en el año 2016.

III.5.1 Ejemplos de Implementación de Método de Optimización Multinivel para Colocación de Válvulas Reductoras de Presión en las Entradas de los Sectores

En este caso se combina el problema de optimización del CEVC de sectores con la optimización de la colocación de VRPs. Se asume que todas las entradas de los sectores cuentan con VRPs. En el primer nivel de la optimización se optimiza el CEVC, en tanto, en el segundo nivel se optimiza el punto de ajuste de las válvulas. Nótese que el primer grupo está relacionado con costos estructurales relacionados a la instalación de las válvulas, mientras que el segundo grupo está relacionado con aspectos hidráulicos, tales como la presión mínima o maximización de la resiliencia de la red.

Para mostrar la implementación del método, se empleó la red grande de Managua que se muestra en la Ilustración 55, la cual está subdividida en 10 sectores. El costo del CEVC resultó 259,437 $ en total. De las 26 tuberías candidatas, 11 fueron establecidas como entradas de sectores. La Tabla 20 muestra el resultado final en términos de índice de uniformidad, presiones promedio, mínima y máxima. Es interesante notar que en todos los casos la presión mínima se acerca al mínimo de presión establecida (10 m), en tanto que la presión final se redujo en comparación a la presión máxima final (61 m). La variación del índice de resiliencia correspondió a 45 %.

Tabla 20: Resultados de presión obtenido mediante el método de optimización multinivel

Sector	1	2	3	4	5	6	7	8	9	10
Índice de uniformidad de presión	26.80	61.17	69.42	27.29	35.48	63.90	65.14	35.40	52.17	65.74
Presión promedio	27.56	27.94	32.04	32.04	26.19	30.59	28.86	31.35	27.92	32.09
Presión mínima	10.01	10.00	10.30	12.99	10.00	10.19	10.00	10.00	10.16	10.15
Presión máxima	56.54	59.68	54.23	54.70	53.93	55.82	57.63	55.38	59.34	60.40

III.6 Optimización del Conjunto de Entrada de Sectores/Válvulas de Cierre Mediante Optimización de Enjambre de Agentes

La técnica/marco de optimización ASO constituye una técnica desarrollada por Montalvo *et al.* (2011), en el grupo de investigación FluIng de la Universidad de Valencia, España. Es un abordaje meta-heurístico mediante el cual se persigue mimetizar el juicio ingenieril (Montalvo *et al*., 2014). Para ello se combinan y ajustan/mejoran los principios de PSO y MAS (*por Multi-agent Systems*), dando como resultado un marco de optimización en donde varios algoritmos de optimización evolutiva, tales como AGs, PSO, Optimización de Colonia de Hormigas (o ACO por *Ant Colony Optimization*) cooperan entre sí. En este contexto, un agente es una entidad que representa una solución candidata. El concepto de agente es un principio de la teoría MAS. En esta, un agente es un elemento computacional capaz de llevar a cabo acciones en nombre de su usuario/dueño a fin de alcanzar un objetivo sin necesidad de que el usuario le diga que hacer. Tales agentes cuentan únicamente con una visión parcial del problema, lo que les imposibilita para resolver el problema de manera individual. Sin embargo, al cooperar en un grupo (enjambre de agentes) sí es capaz de abordar problemas muy complejos. De ahí el origen del término *Sistemas Multi-Agente*. Esta capacidad de interacción entre agentes se debe a la habilidad con la cual cuentan para cooperar, coordinar y negociar (Wooldridge, 2002). Así, en este marco, los agentes (que representan potenciales soluciones) viajan sobre el espacio de búsqueda, siguiendo los principios que se siguen en el algoritmo PSO, a saber, guiados por su resultado anterior y por el mejor resultado obtenido por el enjambre mismo. A su vez, dentro de cada agente, varios algoritmos de optimización

cooperan entre sí, de manera tal que el resultado generado por uno (algoritmo) puede servir para mejorar el desempeño de los otros. Esto es especialmente importante en lo que a definición de parámetros de algoritmos se define, que representa uno de los aspectos más difíciles de abordar en la implementación de algoritmos bio-inspirados.

Cuatro características/aportes importantes de este marco de optimización son: (1) la redefinición del concepto de liderazgo propio del algoritmo PSO. Lo cual se refiere al establecimiento de una partícula que sirve como guía para las otras partículas (en esta caso, agentes). Evidentemente, al inicio de la optimización se desconoce el punto *óptimo*, por lo cual, en cada iteración, se define como partícula líder (también conocida como *punto singular*) aquella que presente el mejor resultado dentro del conjunto de agentes. Sin embargo, para enriquecer la búsqueda y evitar la tendencia de las nuevas soluciones a dirigirse al *punto singular*, se fomenta la búsqueda en otras regiones mediante la creación de sub-enjambres con ciertas instrucciones específicas que se basan en el conocimiento del problema por parte del usuario; (2) la adopción de un procedimiento de normalización para uniformizar las unidades de los objetivos, de tal manera que, luego, la distancia entre dos objetivos pueda ser medida mediante la métrica Euclidiana; (3) el enriquecimiento de la frontera de Pareto, lo cual no sólo implica la adición de nuevos sub-enjambres, sino el aumento dinámico de las poblaciones a fin de aumentar la densidad de la frontera de Pareto y la interacción humana con el sistema computacional, a fin de completar áreas pobremente representadas de la frontera de Pareto; (4) la dotación de reglas a los agentes con reglas propias del problema que se quiere abordar.

III.6.1 Ejemplo de Implementación de Optimización del Conjunto de Válvulas de Cierre/Entradas de Sectores Mediante Optimización de Enjambre de Agentes

Una vez que se obtiene la lista de tuberías candidatas, se procede a la definición del CEVC. Como se ha explicado anteriormente, la disminución de la presión se debe al cierre de tuberías y en las tuberías que permanecen abiertas, se debe colocar un caudalímetro para controlar el caudal. Es preferible establecer más tuberías candidatas como tuberías cerradas que como entradas de sectores debido a dos

razones: en primer lugar, cuanto menor sea el número de entradas a cada sector, mayor será la precisión en el control del caudal de entrada y, por lo tanto, mayor será la probabilidad de detectar una nueva fuga; por otro lado, se espera que el costo de instalación de un caudalímetro sea más alto que el costo de instalación de una válvula. El problema se aborda como un problema de optimización multi-objetivo con cinco objetivos: (objetivo 1) la minimización de la desviación del índice de resiliencia propuesta por Todini (2000) (Ecuación 90); (Objetivo 2) la minimización del aumento de la potencia operativa (Ecuación 91); (Objetivo 3) la minimización de la variación de la desviación estándar de la presión entre los nodos de la red (Ecuación 92); (Objetivo 4) una penalización/costo por los nodos que no pueden cumplir con un mínimo valor de presión (Ecuación 93), y finalmente (objetivo 5) la minimización del costo asociado a la compra de nuevas válvulas y caudalímetros (Ecuación 94). Los dos primeros objetivos pueden ser considerados como criterios energéticos, los dos siguientes como criterios operativos y el quinto objetivo como un criterio económico.

$$f(x) = Min\ (\Delta I_r) \quad \text{(Ecuación 90)}$$

$$f(x) = Min\ (\Delta P_O) \quad \text{(Ecuación 91)}$$

$$f(x) = Min\ (\Delta \sigma_P) \quad \text{(Ecuación 92)}$$

$$f(x) = Min\ (P_{de}) \quad \text{(Ecuación 93)}$$

$$f(x) = Min\ (C_{cv} + C_{FM}) \quad \text{(Ecuación 94)}$$

No se define ninguna restricción, ya que el algoritmo ASO funciona con penalizaciones. Se agrega un costo de penalización al objetivo 4 (nodos con presión inferior a 15 m), por lo tanto, cuanto más grande sea el número de nodos con presión por debajo de ese valor, mayor será el costo del objetivo 4. También se asumió la ausencia de válvulas preexistentes, por lo que cualquier tubería candidata que resulte cerrada (válvula de aislamiento) en el proceso de optimización estará asociada a un coste simbólico (100 unidades monetarias / válvula), inferior al costo

asociado a las tuberías que deben permanecer abiertas (UOCs) (1000 Unidades monetarias). Estos costos son simbólicos, y tuvieron que ser asumidos ya que no se contaba con información sobre los costos reales.

Para mostrar la aplicación de este algoritmo, el mismo es implementado sobre la red que se muestra en la Ilustración 85.

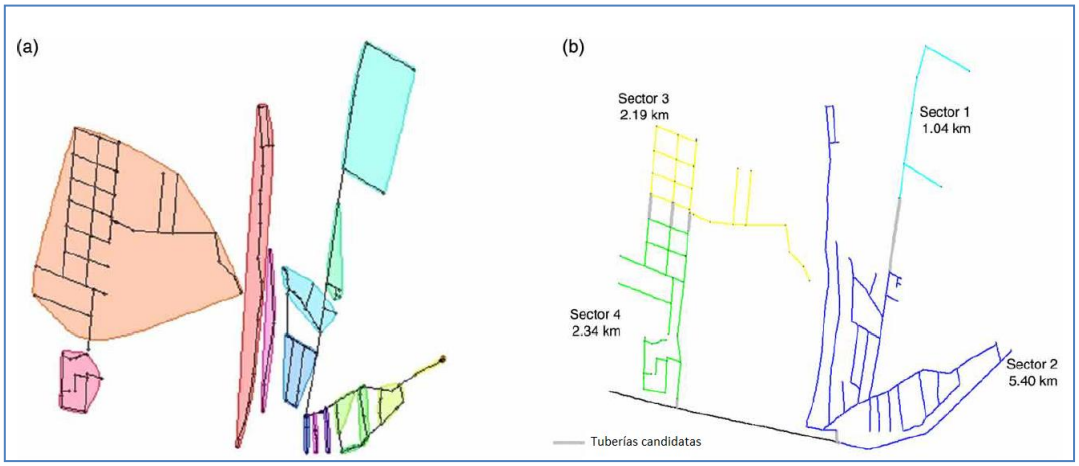


Ilustración 85: Red ejemplo pequeña

La Tabla 21 muestra una comparación entre los valores obtenidos para las dos soluciones factibles con los valores en la red original (red con todas las tuberías abiertas). La solución 1 es mejor en términos de control de caudal, ya que sólo requiere una caudalímetro para el sector de mayor tamaño. Sin embargo, también implica mayor impacto en términos de fiabilidad de la red, incremento en la potencia operativa, y adicionalmente, algunos nodos reportan presión inferior a 15 m (pero superior a 6 m). En este punto el personal técnico de la red debe decidir la configuración que es más apropiada. Es importante resaltar que los valores de presión inferiores a 15 m sólo se registran en un único período de tiempo (el peor escenario).

Tabla 21: Resultados de la optimización de la red pequeña

	Red original		Solución factible 1	Solución factible 2
Parámetro	**Valores**	**Función**	**Valores**	**Valores**
Índice de resiliencia	0.48	Variación del índice de resiliencia (%)	30.68 (valor 0.3)	0.65 (valor: 0.47)
Energía inicial	20524	Variación de la energía (kW)	22109 (valor: 22735)	86.73 (valor: 20610)
Desviación estándar de la presión (m)	6.98	Variación del desviación estándar de la presión (m)	3.30 (valor: 10.28)	0.029 (valor: 6.95)
Déficit de presión ($)	0	Déficit de presión ($)	259.59	0
Válvulas de cierre	0	Válvulas de cierre	2	2
UOC	0	UOC	5	5

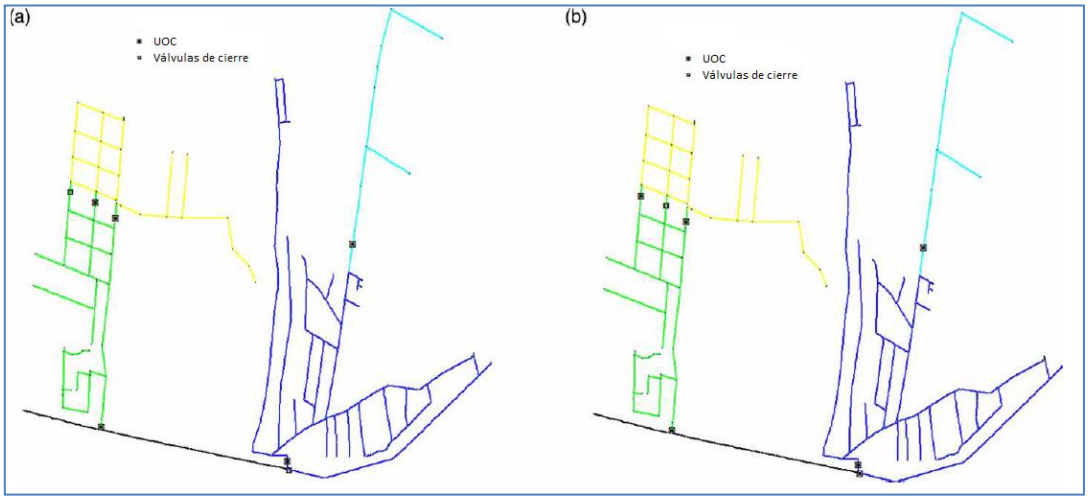


Ilustración 86: Ubicación de UOCs y válvulas de cierre (red pequeña)

Se realizó otro ejemplo con una parte de la RDAP de la ciudad de Managua. Esta nueva red tiene 1374 nodos y 1325 tuberías. Se suministra por una tubería principal que recibe agua de una serie de bombas y un lago. Dado que estas fuentes están lejos de la red estudiada, las mismas se representan mediante un depósito. La Ilustración 87 muestra la división de la red en ocho sectores, todos entre 3 y 30 km de longitud de tubería. En total, se incluyeron 21 tuberías en el conjunto de tuberías candidatas. La Tabla 22 muestra el valor de los objetivos en la red original y en las soluciones encontradas por el proceso de optimización. En este caso, se encontraron tres soluciones. Las tres implican la instalación de seis válvulas límite, con una reducción máxima en el índice resiliencia de 16.52%.

Tabla 22: Resultado de la optimización (red grande)

	Red original		Solución factible 1	Solución factible 2	Solución factible 3
Parámetro	**Valores**	**Funciones**	**Valores**	**Valores**	**Valores**
Índice de resiliencia	0.55	Variación del índice de resiliencia (%)	12.53	14.93	16.52
Energía inicial	24788.61	Variación de la energía (kW)	171.58	161.77	173.76
Desviación estándar de la presión (m)	5.47	Variación del desviación estándar de la presión (m)	0.511	1.20	0.51
Déficit de presión ($)	0	Déficit de presión ($)	0	0	0
Válvulas de cierre	0	Válvulas de cierre	6	6	6
UOC	0	UOC	15	15	15

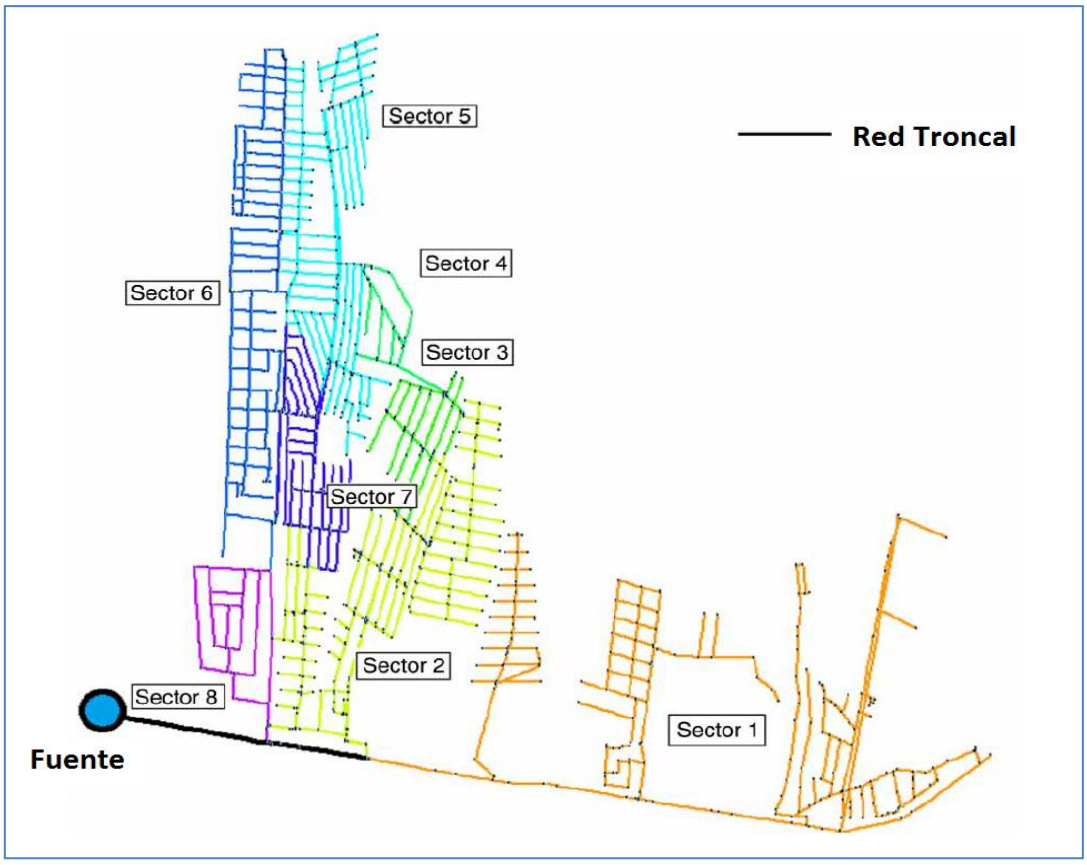


Ilustración 87: Resultado de la optimización (red grande)

III.7 Análisis Global de los Resultados Obtenidos en los Ejemplos de Implementación

Mediante los ejemplos de implementación presentados, se pudo evidenciar la aplicabilidad de algunas técnicas de optimización heurísticas para definir el CEVC de los sectores definidos mediante los métodos basados en detección de comunidades presentados en el Capítulo I. La tesis se centra principalmente en el método de optimización presentado en la Subsección III.4 (Optimización Mediante Algoritmos Genéticos y Simulación Monte Carlo: Predicción de Nuevas Fugas Mediante Sectorización) en la que se introducen una serie de conceptos que permiten valorar la implementación de los proyectos de sectorización más allá de sólo considerar la reducción de fugas de fondo, que es el único criterio que se ha empleado hasta ahora para dicha valoración. Tal como se observó en los resultados de esta subsección, el balance coste beneficio obtenido finalmente varía significativamente cuando se toman en consideración los criterios introducidos en esta tesis. Evidentemente, la magnitud de tal variación depende de los valores de fugas que existen en la red. En redes con menores índices de fugas, se espera que el beneficio no sea tan grande como en el caso de redes con alto nivel de deterioro. Al comparar los resultados obtenidos mediante la técnica de optimización basada en AGs, con los que se obtuvieron mediante la técnica de optimización basada en PSO, se puede destacar que en la primera se establecen tres UOCs más que en la última (ver Tabla 23); no obstante, la inversión total de la segunda (259,437 $) resultó menor que la inversión total resultante en el primer caso (304,552 $), lo que implica que la segunda de las técnicas es más eficaz para la definición del CEVC. Adicionalmente, tal y como se puede ver en la Ilustración 88, en la configuración generada mediante la técnica PSO, se obtuvieron mejores índices de uniformidad de presiones con respecto a la otra alternativa. Peso a ello, se debe tener en consideración que, en realidad, ambas técnicas no son comparables, ya que en la optimización mediante PSO, sólo se tiene en cuenta como beneficio la reducción del impacto sobre la presión y la uniformidad de las presiones dentro de los sectores, versus el costo de la inversión total, sin tener en cuenta la tasa de descuento anual; en tanto, en la optimización mediante AGs, se establece como beneficio la reducción de las fugas de fondo, la reducción de fugas producto de la probabilidad

de detección de eventos futuros, entre otros aspectos, y en adición, los costos se calculan sobre una base anual. La tercera técnica de optimización empleada, representa un paso más avanzado en la optimización, ya que no sólo hace uso de los dos algoritmos de optimización empleados en las primeras dos técnicas, sino que genera más de una solución Pareto dominante. Adicionalmente, en esta técnica no se tiene que establecer un único objetivo, sino que todos los objetivos se optimizan paralelamente pero de manera independiente.

Tabla 23: CEVC resultante mediante las técnica de optimización AG y PSO

Código de tubería	Diámetro (mm)	Estado de tuberías candidatas (PSO)	Estado de tuberías candidatas (AG)
TUB136078	304.8	0	1
TUB136094	406.4	1	0
111101	100.0	0	0
111102	100.0	0	0
111103	100.0	0	1
TUB84821	50.8	0	1
TUB84867	152.4	0	1
TTTUB1358340	203.2	1	0
11122	152.4	1	0
11123	152.4	0	0
11124	300.0	0	1
TUB547216	152.4	1	1
TUB2649	406.4	0	1
11141	300.0	1	1
TTTUB1347771	406.4	1	0
TUB547036	609.6	0	1
TTTUB5735510	152.4	0	1
TUB77984	406.4	1	0
TUB85129	101.6	1	1
11192	100.0	1	1
11191	200.0	0	0
TTTUB4271	457.2	0	0
TUB2565	152.4	1	1
TTTUB24870	203.2	0	0
11131	152.4	1	1
TTTUB4560351	101.6	0	0
	Total	11	14

De lo anterior, se puede concluir que todas las técnicas empleadas presentan una serie de ventajas importantes cuya combinación podría generar un esquema de

optimización del CEVC bastante interesante, por ejemplo, se podría pensar en un esquema de optimización mediante ASO en el cual se tengan en cuenta todos los aspectos de la detección de fugas presentando en el método de optimización mediante AG + SMC y que adicionalmente tome en cuenta en un segundo nivel, el punto de funcionamiento de las VRPs, con el fin de mejorar la uniformidad de las presiones en los sectores.

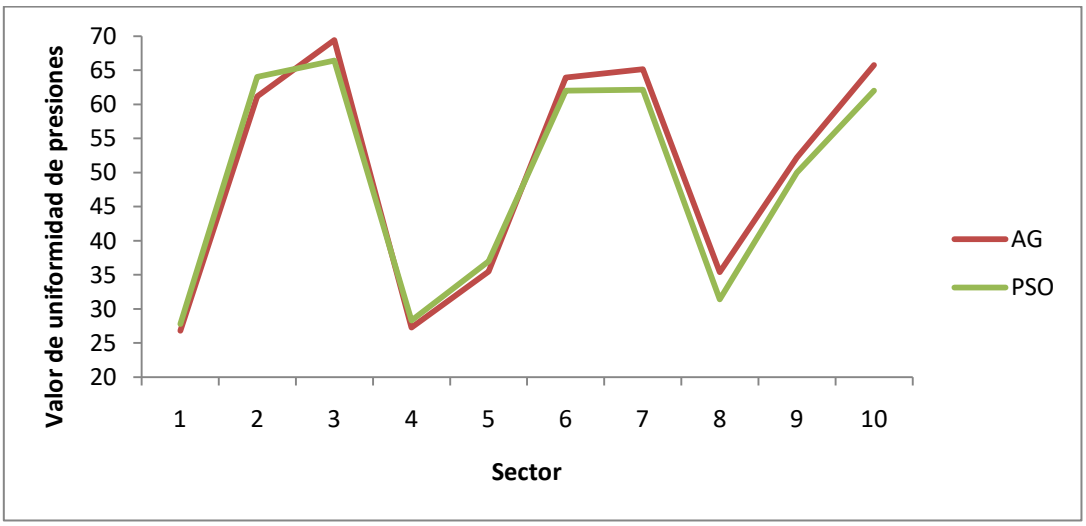


Ilustración 88: Comparación de valor de uniformidad de presiones para las técnicas AG y PSO

IV. CONCLUSIONES Y LÍNEAS FUTURAS

En esta tesis doctoral se aborda un tema de gran relevancia actual: métodos para el diseño de esquemas de sectorización en RDAPs aplicables a redes dependientes de una red de conducción principal, considerando una serie de beneficios económicos relacionados con la gestión sostenible de fugas. Estos beneficios se calculan siguiendo el marco metodológico propuesto en el esquema BABE propuesto por UKWIR. Este desarrollo permite considerar el beneficio de la sectorización, más allá de los ahorros generados sólo por la reducción de las fugas de fondo, que es el criterio económico más ampliamente considerado en las metodologías de sectorización que hasta ahora han sido propuestas.

Este trabajo integra nuevos conceptos en el campo de investigación de la sectorización de RDAPs, tales como algoritmos de detección de comunidades, propios de la Teoría de Redes Sociales, optimización mediante AGs ligada a SMC, método multinivel de optimización basado en PSO, y método ASO de optimización. Mediante los algoritmos de detección de comunidades, se detectan sectores basados en la topología de red, lo cual es clave cuando se abordan RDAPs que dependen de una red troncal.

Para la definición de la red troncal en una RDAP (donde la misma aún no está definida), se establece una jerarquía de tuberías, según su importancia en el suministro del resto de la red basada en el VCMCA aquí propuesto. Este criterio es más robusto que el uso de una sola característica de las tuberías, ya que considera la magnitud del caudal y la dirección del mismo en cada tubería en el escenario más crítico. Tal jerarquía permite a las empresas gestoras definir el alcance de la definición de la red troncal utilizando las características de cada contexto. Una vez definida, la red troncal se segrega de la red de distribución. El objetivo es no incluir a la primera dentro de la sectorización, reduciendo así el impacto negativo sobre la fiabilidad del sistema. A continuación, se detectan las comunidades mediante algoritmos de detección de comunidades, a saber: clústering jerárquico, algoritmo Multinivel, y algoritmo Walktrap.

- La primera técnica es bastante sencilla de implementar, y brinda bastante flexibilidad para poder mejorar la adaptación a datos de distinta naturaleza. Sin embargo, al basarse únicamente en información almacenada en los nodos, se pierde el sentido de conectividad de la red. Lo anterior se traduce, en algunos casos, en particiones con nodos desconectados, lo cual en una RDAP no es factible. Si bien, este problema se puede resolver con los procesos de detección de todos los subgrupos y posterior proceso de re-fusión aquí propuesto, esto representa un esfuerzo de cálculo extra que no es necesario en otros métodos que se han abordado. Una de las desventajas de este método es que en su implementación no existe una medida que permita de manera directa decidir la partición más adecuada, con lo cual, la implementación tiene que ser acompañada de medidas de validación externas.

- La segunda técnica de definición de sectores (basada en el algoritmo Multinivel) tiene como ventaja la practicidad, ya que, para su implementación, no se tiene que definir ningún tipo de parámetro; no obstante, su implementación en *R* tiene la desventaja, con respecto a otros algoritmos de detección de comunidades, de no permitir extraer la jerarquía de comunidades que se genera. La única alternativa es obtener la partición con máxima modularidad, perdiéndose flexibilidad en la implementación. Una ventaja adicional del algoritmo es que, a diferencia del clústering jerárquico, siempre encuentra comunidades conectadas, con lo cual no se hace necesario implementar la reasignación de *id* que en algunos casos es necesario hacer cuando se emplea la técnica de clústering jerárquico. Los resultados muestran que la técnica *per se* logra reconocer las características propias de cada zona de la red, estableciendo separaciones adecuadas sin necesidad de establecer una serie de características/criterios para tal fin.

- Por otro lado, el método de definición de sectores basado en el algoritmo Walktrap, ha demostrado ser completamente aplicable al problema de sectorización que se aborda en este trabajo. Este método ofrece, entre sus ventajas, la capacidad de definir el número de iteraciones, lo cual permite

mejorar el resultado final de la definición de sectores. Tal como se mostró en el ejemplo de implementación, si bien, después de cierto rango, el número de iteraciones deja de tener mucho efecto sobre el número de sectores en el que se maximiza el valor de modularidad, este sí puede tener efecto sobre la topología de las comunidades obtenidas. En el ejemplo de implementación se observa que, cuando se generan comunidades con un menor número de iteraciones, se encuentran tuberías pertenecientes a algunas comunidades dentro del área de otras comunidades. La implementación de este algoritmo en *R* tiene como ventaja la capacidad de poder extraer una partición u otra del dendograma, lo cual puede ayudar a reducir el tiempo del proceso de re-fusión de sectores; no obstante, siempre queda pendiente la pregunta sobre por dónde debe empezar el proceso de re-fusión.

Con respecto a la optimización del CEVC, el marco de optimización presentando, en el cual no sólo se tienen en cuenta la reducción de fugas de fondo, genera un cambio radical en el balance costo/beneficio de la sectorización. En este sentido, vale la pena destacar que este trabajo representa la primera metodología de sectorización en donde se tiene en cuenta el efecto de la sectorización sobre eventos de fugas futuros. En la actualidad, no existe ninguna metodología que conecte el aspecto topológico de la sectorización con la capacidad de detectar futuros eventos de fugas, radicando en este punto el componente innovador del presente trabajo. Por otra parte, a través de la SMC se espera obtener resultados más realistas, respecto a la capacidad de ocurrencia y detección de eventos de fugas futuros, que mediante el uso de valores estimados. Evidentemente, la validez y aplicabilidad de esta idea depende de la disponibilidad de las empresas operadoras para seguir un programa de gestión sostenible de fugas. En ese sentido, una de las principales desventajas de la metodología es la recolección de información, lo cual requiere de amplios periodos de tiempo (años).

Mediante la introducción de VRPs en el método optimización del conjunto CEVC basado en PSO, se logra mejorar la uniformidad de presiones dentro de los sectores y, además, se obtiene una solución del CEVC más económica (se requiere menor número de entradas de sectores, sin generar una reducción en el índice de

sectorización, sustancialmente mayor que el que se obtiene mediante AGs). No obstante, en este caso, no se tienen en cuenta los aspectos de reducción de fugas que se plantean en la técnica basada en AGs + SMC.

El tercer método de optimización, constituye un paso más avanzado en términos de optimización, ya que incluye los dos métodos de optimización anteriormente citados; sin embargo, su implementación no incluye ni los aspectos de reducción de fugas futuras, ni la inclusión de VRPs. Vale la pena destacar dos beneficios importantes de esta última técnica: por un lado no ofrece una única solución, sino un conjunto de soluciones Pareto dominantes; por otro lado, se pueden optimizar varios objetivos en paralelo. Así, para este problema en concreto, se podría maximizar el beneficio económico de la implementación y la uniformidad de las presiones, a la vez que se minimiza la reducción de la red.

Líneas Futuras de Investigación

- Tal como se destacó, sería interesante que en la implementación de los algoritmos de detección de comunidades Walktrap y multinivel se tomaran en cuenta pesos en las tuberías y en los nodos durante el proceso de definición de comunidades. En este trabajo, lo anterior se hace de manera indirecta mediante el proceso de re-fusión planteado, pero valdría la pena modificar los algoritmos de detección de comunidades para poder contemplar estos aspectos a nivel interno.

- Dados los resultados finales presentados en este trabajo, sería interesante plantear un método de optimización del CEVC, mediante ASO, incluyendo todos los aspectos de gestión de fugas presentados, la SMC y la optimización en un segundo nivel del punto de funcionamiento de las VRPs.

- También sería interesante poder incluir los eventos de fugas dentro del modelo matemático de la red al momento de ejecutar la optimización, para así poder considerar el impacto de la ocurrencia de los eventos de fugas sobre la resiliencia de la red.

- En este trabajo sólo se plantea la relación entre el CEVC y la probabilidad de ocurrencia y detección de nuevos eventos de fugas; sin embargo, tal y como se ha visto, la introducción de VRPs en las entradas de los sectores, mejora la uniformidad de las presiones dentro de los sectores. De esta forma, valdría la pena investigar el efecto de la colocación de VRPs sobre la probabilidad de ocurrencia y detección de futuros eventos de fugas.

- Sería recomendable profundizar en la relación entre la topología de sectores y la probabilidad de ocurrencia y detección de nuevos eventos de fugas. A su vez, valdría la pena estudiar el efecto que generaría incluir esta probabilidad no sólo dentro de la optimización del CEVC, sino dentro del proceso de definición de lo sectores.

CONCLUSIONS AND FUTURE LINES OF RESEARCH

The present doctoral dissertation addresses a topic of major and current relevance, namely, generation of automatic sectorization designs, for trunk network depending Water Supply Network (WSNs). The proposed methods consider a series of economic benefits related to sustainable leakage management. Such benefits are calculated according to the methodological framework in the BABE scheme proposed by UKWIR. The development here proposed allows considering the benefit of sectorization, beyond the savings generated only by the reduction of background leakage, which is the economical criterion most widely considered in the sectorization methodologies that have been proposed so far.

This work integrates new concepts in the WSN sectorization research field, such as community detection algorithms from the Social Network Theory, optimization by means of Genetics Algorithms (GAs) linked to Monte Carlo Simulation, a multi-level optimization method based on PSO, and optimization through ASO. The definition of sectors by means of community detection algorithms is based on the topology of the WSNs, which is a key aspect when dealing with water supply systems that depend on a trunk network.

For the definition of the trunk network in WSNs (where it is not yet defined) a hierarchy of pipes is established according to their importance for the supply of the rest of the network. Such hierarchy is based on an Accumulated Shortest Path Value here proposed. This criterion is more robust than the use of a single characteristic of the pipes, since it considers the magnitude of the flow and its direction in each pipe in the most critical scenario. Such hierarchy allows water authority to define one or more trunk network scope based on the characteristics of each context. Once the trunk network is defined, the same is decoupled from the distribution network. The objective is to exclude the first from the sectorization, in order to reduce the negative impact on the reliability of the system. After that, sectors are defined through community detection algorithms, namely: hierarchical clustering, multilevel algorithm, and Walktrap algorithm.

- The first technique is easy to implement, and provides enough flexibility to be suitable to tackle various data types. However, by relying only on nodes information, the sense of network connectivity is lost. This translates sometimes into partitions with disconnected nodes, which is not feasible in WSNs. Although this problem can be solved with the subgroup detection and merging process here proposed, this represents an extra computational effort which is not necessary in other methods that have been proposed. One of the disadvantages of this method is that in its implementation there is no a measure that directly allows deciding the most appropriate partition, therefore, the implementation has to be accompanied by external validation measures.

- The second sector definition technique (based on the Multilevel algorithm) has practicality as its main advantage, since, for its implementation, there is no need to define any parameter. However, its implementation in R has the disadvantage (in comparison to other community detection algorithms) that it does not allow to extract the hierarchy of communities that is generated, producing only the partition with maximum modularity available. This translates into flexibility reduction in the implementation. An additional advantage of the algorithm is that, unlike hierarchical clustering, it always finds connected communities, so it is not necessary to implement the subgroup merging process above mentioned. The results show that the technique *per se* is able to recognize the characteristics of each zone of the network, establishing adequate separations without having to establish a series of characteristics / criteria for this purpose.

- The sector definition method based on the Walktrap algorithm has proven to be fully applicable to the sectorization problem addressed in this work. It offers, among its advantages, the ability to define the number of iterations, which allows improving the final result of the sector definition. As shown in the implementation example, although, after a certain range, the number of iterations does not have much effect on the number of sectors in which the modularity value is maximized, it can have an effect on the topology of the obtained communities. In the example of implementation, it is observed that

when communities with a smaller number of iterations are generated, pipes belonging to some communities can be found inside the area of other communities. The implementation of this algorithm in R has as advantage the capacity to extract one or another partition from the dendogram, which helps reducing the time of the sector merging process. However, the question remains as to where the merging process should start.

Regarding the optimization of the group of entrances and closed valves (GECV), the optimization framework presented, which not only takes into account the reduction of background leakage, generates a radical change in the cost / benefit balance of sectorization. In this sense, it is worth noting that this work represents the first methodology of sectorization where the effect of sectorization on future leakage events is taken into account. Until now, there is no methodology that links the topological aspect of sectorization with the ability to detect future leakage events, which is an innovative component of this work. Moreover, through the Monte Carlo Simulation, it is expected to obtain more realistic results regarding the capacity of occurrence and detection of future leakage events, than by using estimated values. Evidently, the validity and applicability of this idea depends on the availability of the water companies to follow a sustainable leakage management program. In that sense, the collection of information, which requires extensive periods of time (years), represents one of the main disadvantages of the methodology.

The introduction of Pressure Regulating Valves (PRV) in the optimization method based on PSO, improves the uniformity of pressures within the sectors and, in addition, a more economical solution is found (a smaller number of sector entrances is required), without significantly reducing network resilience in comparison to the solution found by GAs. However, in this case, the leakage reduction aspects that are considered in the GAs + MCS-based technique are not taken into account, thus, both methodologies are not completely comparable.

The third optimization method is a further step in terms of optimization, since it includes the two above implemented optimization methods; however, its implementation does not include aspects of future leakage reduction or the

inclusion of PRVs. It is worth highlighting two important benefits of this technique: on the one hand it does not offer a single solution, but an entire set of Pareto dominant solutions; on the other hand, several objectives can be optimized in parallel. Thus, for this particular problem, the economic benefit of the implementation and the uniformity of the pressures could be maximized, while minimizing the reduction of the network.

Future Works

As pointed out, it would be interesting to include weights (in the nodes and pipes) in the community definition process by means of the Walktrap and Multilevel algorithms. In this work, weights are indirectly included through the proposed merging process, but it would be worthy it to modify the community detection algorithms in order to internally consider the weights (characteristics).

Based on the final results presented in this work, it would be interesting developing a method of optimization of the GECV, through ASO, including all aspects of leakage management presented, the MCS and the optimization at a second level of the operating point of the PRVs.

- It would also be interesting to include leakage events within the simulation of the network during the optimization, in order to be able to consider the impact of the occurrence of leakage events on the network resilience.

- In this work only the relationship between the GECV and the probability of occurrence and detection of new leakage events is discussed; however, as has been seen, the allocation of PRVs in the entrances, improves the uniformity of pressures within sectors. In this sense, it would be worth investigating the effect of the placement of PRVs on the probability of occurrence and detection of future leakage events.

- It would be advisable deepening in the relationship between the topology of sectors and the probability of occurrence and detection of new leakage events. At the same time, it would be worthy to study the effect that would generate

including this probability not only in the optimization of the GECV, but also in the sectors definition process.

V. REFERENCIAS BIBLIOGRÁFICAS

Publicaciones propias directamente asociadas al desarrollo de esta tesis

B.M. Brentan, E. Campbell & J. Izquierdo. Calibración conjunta de presiones y fugas en redes de abastecimiento de agua potable. *Aporte Santiaguino*, Vol. 8: 453-470. **2015**

B.M. Brentan, E. Campbell, G.L. Meirelles, E. Luvizotto & J. Izquierdo. Social network community detection for DMA creation: criteria analysis through multilevel optimization. *Mathematical Problems in Engineering*, Vol. 2017(Article ID 9053238). **2017**

E. Campbell. Propuesta para una Metodología de Sectorización de Redes de Abastecimiento de Agua potable. Tesina de Máster, *Universitat Politècnica de València*. **2013a**

E. Campbell, R. Pérez-García & J. Izquierdo. Propuesta de metodología para sectorización de redes de abastecimiento de agua potable. In *Proc*. XII Simposio Iberoamericano Sobre Sistemas de Abastecimiento de Agua y Drenaje Urbano, Buenos Aires, Argentina. **2013b**

E. Campbell, R. Pérez-García, J. Izquierdo & D. Ayala-Cabrera. Metodología para sectorización de redes de abastecimiento de agua potable. In *Proc*. III Jornadas de Ingeniería del Agua, Valencia, España. **2013c**

E. Campbell, D. Ayala-Cabrera, J. Izquierdo & R. Pérez-García. Label propagation algorithm based methodology for water supply networks sectorization. *International Journal of Complex Systems in Science*, Vol. 4 (1): 35-39. **2014a**

E. Campbell, D. Ayala-Cabrera, J. Izquierdo & R. Pérez-García. Graph clustering based on social network community detection algorithms. In *Proc*. VIII International Congress on Environmental Modelling and Software Society, San Diego, California, USA. **2014b**

E. Campbell, D. Ayala-Cabrera, J. Izquierdo, R. Pérez-García & M. Tavera. Water supply network sectorization based on social networks community detection algorithms. In *Proc*. Water Distribution Systems Analysis, Bari, Italy. **2014c**

E. Campbell, D. Ayala-Cabrera, J. Izquierdo, R. Pérez-García & M. Tavera. Water Supply Network Sectorization Based on Social Networks Community Detection Algorithms. *Procedia Engineering*, Vol. 89: 1208–1215. **2014d**

E. Campbell, J. Izquierdo, R. Pérez-García & D. Ayala-Cabrera. Unsupervised methodology for sectorization of trunk depending water supply networks. *In* J. C. Cortés López, L. A. Jordá Sánchez & R. J. Villanueva Micó (Edits), Mathematical Modelling in Social Sciences and Engineering, Nova Publishers, New York. **2014e**

E. Campbell, A. Ilaya-Ayza, J. Izquierdo & R. Pérez-García. Self-organized maps and clustering techniques to sectorize water supply networks based on energy criteria. In *Proc*. Mathematical Modelling in Engineering and Human Behaviour, Valencia, Spain. **2014f**

E. Campbell, R. Pérez-García, J. Izquierdo, D. Ayala-Cabrera, M. Tavera & J. Gutiérrez-Pérez. Sectorización multinivel de redes de abastecimiento de agua potable basada en métodos de detección de comunidades en redes sociales. In *Proc*. Congreso Latinoamericano de Hidrogeología e Hidráulica, Santiago de Chile, Chile. **2014g**

E. Campbell, D. Ayala-Cabrera, J. Izquierdo, R. Pérez-García & M. Tavera. A flexible methodology to sectorize water supply networks based on social network theory concepts and multi-objective optimization. *Journal of Hydroinformatics*, Vol. 18 (1): 62-76. **2016a**

E. Campbell, J. Izquierdo, R. Pérez-García & I. Montalvo. A novel water supply network sectorization methodology based on a complete economic analysis, including uncertainties. *Water*, Vol. 8 (5): 179. **2016b**

E. Campbell, A, Ilaya-Ayza, J. Izquierdo, R. Pérez-García & I. Montalvo. Social-network-based water supply network sectorization methodology using Montecarlo simulation to predict economical and operational benefits. *Acta Universitaria, Multidisciplinary Scientific Journal*, Vol. 26(NE-3): 44-53. **2016c**

J. Izquierdo, E. Campbell, I. Montalvo, R. Pérez-García & D. Ayala-Cabrera. Error analysis of some demand simplifications in hydraulic models of water supply networks. *Abstract and Applied Analysis*. Article ID 169670, 13 pages, **2013**

J. Izquierdo, I. Montalvo, R. Pérez-García, R & E. Campbell. Mining solution spaces for decision making in water distribution systems. *Procedia Engineering*, Vol. 70: 864-871. **2014a**

J. Izquierdo, E. Campbell, I. Montalvo, R. Pérez-García & D. Ayala-Cabrera. Piezometric error derived from some demand lumped models in water distribution. *International Journal of Complex Systems in Science*, Vol. 4 (1): 17-20. **2014b**

J. Izquierdo, I. Montalvo, E. Campbell, R. Pérez-García. A hybrid, auto-adaptive, and rule-based multi-agent approach using evolutionary algorithms for improved searching. *Engineering Optimization*, Vol. 48(8): 1365-1377. **2016a**

J. Izquierdo, E. Campbell, I. Montalvo & R. Pérez-García. Injecting problem-dependent knowledge to improve evolutionary optimization search ability. *Journal of Computational and Applied Mathematics*, Vol. 291: 281-292. **2016b**

I. Montalvo, J. Izquierdo, E. Campbell & R. Pérez-García. Cloud-based decision making in water distribution systems. *Procedia Engineering*, Vol. 89: 488–494. **2014**

L. Pastor, B.M. Brentan, E. Campbell, M. Nudelman, R. Pérez-García & J. Izquierdo. Estrategias para la modelación de una red de abastecimiento de agua a partir de información limitada. *Aporte Santiaguino*. Vol. 8: 327-346. **2015**

Otras referencias utilizadas en la tesis

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

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

S. Alvisi & M. Franchini. A heuristic procedure for the automatic creation of district metered areas in water distribution systems. *Urban Water Journal*, Vol.11 (2):137-159. **2014**

N. Amar, L.Tazi & A. Bensaid. Semi-supervised hierarchical clustering algorithms. In G. Grahne (Ed). In *Proc*. Sixth Scandinavian Conference on Artificial Intelligence, Vol.1:232-239. **1997**

P. Arabie & L.J. Hubert. An Overview of Combinatorial Data Analysis. In. Arabie, L. Hubert, G. De Soete, P. Arabie, L. Hubert & G. De Soete (Edits), Clustering and Classification. World Scientific Publishing. **1996**

D. Araque & J. G. Saldarriaga. Water distribution network operational optimization by maximizing the pressure uniformity at service nodes. In *Proc.* Water Distribution Systems Analysis, Bari, Italy. **2005**

(ADB) Asian Development Bank. Data Book of Southeast Asian water Utilities 2005. Asian Development Bank. **2007**

T. Aynaud, V. Blondel, J.L. Guillaume & R, Lambiotte. Multilevel Local Optimization of Modularity. Graph Partitioning. Wiley. **2013**

E. Baquela & A. Redchuk. Optimización Matemática con R. Introducción al Modelado y Resolución de Problemas. 1ª edición. Bubok Publishing S.L. España. **2013**.

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

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

V. Bataglj & A. Mrvar. Pajek – Analysis and Visualization of Large Network. In P. Mutzel, M. Jünger, S. Leipert (Eds). Graph Drawing Software. GD 2001. Lecture Notes in Computer Science, Springer. **2002**

R. Baños, C. Gil, J. Reca & F.G. Montoya. A memetic algorithm applied to the design of water distribution networks. *Applied Soft Computing*, Vol.10(1) 261-266. **2010a**

R. Baños, C. Gil, J. Reca & J. Ortega. A pareto-based memetic algorithm for optimization of looped water distribution systems. *Engineering Optimization*. Vol. 42 (3): 223-240. **2010b**

J. Baños, J. Reca, C. Martinez & G. Márquez. Resilience Indexes for Water Distribution Network Design: A Performance Analysis under Demand Uncertainty. *Water Resources Management*. Vol. 55 (10): 2351-2366. **2011**

D. Benvenuti, A. Lambert & M. Fantozzi. Practical experiences in applying advanced solutions for calculation of frequency of intervention with active leakage control: results obtained. In *Proc*. IWA International Specialised Conference, Bucharest, Romanian. **2007**

J. Benítez, X. Delgado-Galván, J. Izquierdo & R. Pérez-García. Improving consistency in AHP decision-making processes. *Applied Mathematics and Computation*, Vol.219: 2432-2441. **2012**

J. Benítez, L. Carrión, J. Izquierdo & R. Pérez-García. Characterization of consistent completion of reciprocal comparison matrices. *Abstract and Applied Analysis*. Article ID 349729, 12 pages. **2014**

J. Benítez, X. Delgado-Galván, J. Izquierdo & R. Pérez-García. Consistent completion of incomplete judgments in decision making using AHP. *Journal of Computational and Applied Mathematics*, Vol. 290: 412-422. **2015**

N. Biggs, E. Lloyd & R. Wilson. *Graph Theory*. Oxford University Press. **1986**

A. Bilgin, J. Ellson, E. Gansner & Y. Hu. North, *et al*. *Graphviz*. **2017**
URL {http://www.graphviz.org/users/arif-bilgin} (Última visita: Marzo 2017).

S. Binitha & S. Siva. A survey of bio inspired-optimization algorithms. *International Journal of Soft Computing and Engineering*. Vol.2(2): 2231-2307. **2012**

V. Blondel, J. Guillaume, R. Lambiotte & E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, Vol. 2008(10): P10008. **2008**

F. Boano & L. Berardi. Comparison of water distribution networks segmentation based on modularity indexes. In *Proc*. 16th Conference on Water Distribution System Analysis, Bari, Italy. **2014**

C. Blum & R. Andrea. Metaheuristic in combinatorial optimization: Overview and Conceptual comparison. *ACM Computing Surveys*, Vol.35(3): 268-308. **2003**

B.M. Brentan, E. Luvizotto Jr, I. Montalvo, J. Izquierdo & R Pérez-García. Water pump scheduling optimization using agent swarm optimization. *Modelling for Engineering & Human Behaviour 2015*, Valencia, Spain. **2015**

G. Brock, V. Pihur & S. Datta. ClValid. An R package for cluster validation. *Journal of Statistical Software*, Vol.25 (4): 1-22. **2008**

D. Coley. An Introduction to Genetic Algorithms for Scientists and Engineers. World Scientific Publishing. **1999**

L. Carrión. Aplicación del Método de las Jerarquías Analíticas para la Toma de Decisiones Participativa en la Gestión de Fugas en Redes de Abastecimiento de Agua. Tesina de Máster, *Universitat Politècnica de València*. **2013**

D. Chen & B. Xu. Geometric algorithms for agglomerative hierarchical clustering. In T. Warnow & B. Zhu (Eds). In *Proc. 9th Annual International Conference on Computing and Combinatorics* , Vol.9:30-39, Springer-Verlag. **2003**

H. Chen & Y. Zhu. Optimization based on symbiotic multi-species coevolution. *Journal on Applied Mathematics and Computation.* Vol.205. **2008**

A. Clauset, M. Newman & C. Moore. Finding community structure in very large networks. *Physical Review E*, Vol.70(6): 066111. **2004**

T. Cormen, C. Leiserson, R. Rivest & C. Stein. Introduction to Algorithms. MIT Press and McGraw-Hill. **2001**

E. Creaco, A. Fortunato, M. Franchini & M.R. Mazzola. 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, Vol.70: 379-388. **2013**

G. Csardi & T. Nepusz. The igraph software package for complex network research. *Inter Journal*, Vol. Complex Systems. **2016** URL {http://igraph.org} (Útima visita: Marzo 2017)

D. Defays. An eficient algorithm for a complete link method. *The Computer Journal (British Computer Society)*, Vol. 20(4): 364-366. **1977**

X. Delgado-Galván, R. Pérez-García, J. Izquierdo & M. Mora-Rodríguez. An analytic hierarchy process for assessing externalities in water leakage management. *Mathematical and Computer Modeling*, Vol. 52: 1194-1202. **2010**

X. Delgado-Galván. Aplicación del Método de Jerarquías Analíticas (AHP) a la Gestión de Pérdidas de Agua en Redes de Abastecimiento. Tesis Doctoral, *Universitat Politècnica de València*. **2011**

F. De Paola, N. Fontana, E. Galdiero, M. Giugni, D. Savic & G. Sorgenti. Automatic multi-objective sectorization of a water distribution network. In *Proc.* Conference on Water Distribution System Analysis, Bari, Italy. **2014**

A.E. Eiben & J.E. Smith. Introduction to Evolutionary Computing. Springer. **2003**

(DVGW) Deutsche Vereinigung des Gases und Wasser. Arbeitsblatt W 392 Rohrnetzinspektion und Wasserverluste Maßnahmen Verfahren. DVGW. **2003**

(GIZ) Deutsche Gesellschaf für International Zussamenarbeit, VAG (Armaturen GmbH), FHNW (Fachhochschule Nordwestschweiz), KIT (Karlsruhe Institute of Technology) & Institute for Ecopreneurship (IEC). Guidelines for Water Loss Reduction: a Focus on Pressure Management. GIZ. **2011**

K. Diao, Y. Zhou & W. Rauch. Automated creation of district metered area boundaries in water distribution systems. *Journal of Water Resources Planning and Management*, Vol. 139(2): 184-190. **2013**

A. Di Nardo & M. Di Natale. A heuristic design support methodology based on graph theory for district metering of water supply networks. *Engineering Optimization*, Vol. 43(2): 193-221. **2011**

A. Di Nardo, M. Di Natale, G. Santonastaso & S. Venticinque. Graph partitioning for automatic sectorization of a water distribution system. In *Proc.*11th International Conference on Computing and Control for Water Industry. Urban Water Management: Challenges and Opportunities, Vol. 3: 841-846, Exeter, UK. **2011**

A. Di Nardo, M. Di Natale, G. Santonastaso, V. Tzatchkov & V. Alcocer Yamanaka. Water network sectorization based on a genetic algorithm and minimum dissipated power paths. *Water Since and Technology: Water Supply*, Vol. 13(4): 951-957. **2013**

A. Di Nardo, M. Di Natale, G. Santonastaso, V. Tzatchkov & V. Alcocer Yamanaka. Water network sectorization based on graph theory and energy performance indices. *Journal of Water Resources Planning and Management*, Vol. 140(5): 620-629. **2014**

W. Ding, R. Jiamthapthaksin, R. Parmar, D. Jiang, Tomasz, F, Stepinski & F.E Christoph. Towards region discovery in spatial datasets. In *Proc.* 12th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, Osaka, Japan, **2008**

E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, Vol.1: 269-271. **1959**

J.C. Dunn. Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, Vol.4 (1): 95-104. **1974**

R. C. Eberhart & J. Kennedy. A new optimizer using particle swarm theory. In *Proc.* Sixth International Symposium on Micromachinery and Human Science, Nagoya, Japan. **1995**

D. Easley & J. Kleinberg. Networks, Crowds & Markets: Reasoning About a Highly Connected World. Cambridge University Press. 2010. URL {http://www.cs.cornell.edu/home/kleinber/networks-book/} (Útima visita: Marzo 2017)

B. Efron, E. Halloran & S. Holmes. Bootstrap confidence levels for phylogenetic trees. In *Proc.* National Academy of Science, Vol.93: 13429-13434. **1996**

M. Eisen, P. Spellman, P. Brown & D. Botstein. Cluster analysis and display of genome-wide expression patterns. In *Proc*. National Academy of Sciences, Vol. 95(25): 14863-14868. **1998**

L. Euler. Solutio problematis ad geometriam situs pertinentis, *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, Vol.8: 128–140. Reprinted and translated in N. L. Biggs, E. K. Lloyd, R. J. Wilson (1976), Graph Theory, Oxford University Press. **1736**

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

(EPA) Environmental Protection Agency. Effect of water age on distribution system water quality, Technical Report. **2002**

M. Fantozzi & A. Lambert. Recent advances in calculating economic intervention frequency for active leakage control, and implications for calculation of economic leakage levels. *Water Science and Technology: Water Supply*, Vol. 5(6):263-271. **2005**

M. Fantozzi & A. Lambert. Including the efects of pressure management in calculations of economic leakage level. In *Proc.* IWA Special Conference Water Loss 2007', Bucharest, Romania. **2007**

M. Farley. Are there alternatives to DMA? *Asian Water*, Vol. 26(10): 12-18. **2010**

J.S. Farris. On the cophenetic correlation coefficient. *Systematic Zoology*, Vol. 18(3): 279-285. **1969**

G. Ferrari, D. Savic & G. Becciu. Graph-theoretic approach and sound engineering principles for designing of district metered areas. *Journal of Water Resources Planning and Management*. Vol.140(12). **2014**

S. Fortunato. Community detection in graphs. *Physics Reports*, Vol.486: 75-174. **2010**

S. Fortunato & M. Barthélemy. Resolution limit in community detection. In *Proc.* National Academy of Sciences, Vol.104(1): 36-41, 2007

L. Freeman. Centrality in social networks: conceptual clarification. *Social Networks*, Vol.1(3): 215-239. **1979**

Z.W. Geem, J.H. Kim & G.V. Loganathan. A new heuristic optimization algorithm: harmony search, *Simulation*, 72(6): 60-68. **2001**

Z. Ghahramani. Unsupervised Learning. In O. Bousquet, G. Raestsch, U. Von-Luxburg (Eds). Advance Lectures on Machine Learning LNAI 3176. Springer-Verlag. **2004**.

O. Giustolisi & L. Ridolfi. Modularity index for hydraulic system segmentation. In *Proc*. 16th Conference on Water Distribution System Analysis, Bari, Italy, **2014a**

O. Giustolisi & L. Ridolfi. A New modularity-based approach to segmentation of water distribution networks. *Journal of Hydro Engineering*. Vol.140: 04014049. **2014b**

E. Goset. Discrete Mathematics with Proff. John Wiley & Sons. **2009**

R. Gomes, A. S. Marques & J. Sousa. Identification of the optimal entry points at district metered areas and implementation of pressure management. *Urban Water Journal*, Vol.9(6): 365-384. **2012**

L. Gonçalves, R. Rodriguez, A.J. Amaral, M. Karasawa & C. Sudré. Comparison of multivariate statistical algorithms to cluster tomato heriloon accessions. *Genetic and Molecular Research*, Vol.7 (4): 1289-1297. **2008**

S. Hajebi, S. Temate, S. Barret, A. Clarke & S. Clarke. Multi-agent simulation to automate water distribution network partitioning. In *Proc.* 27th European Simulation and Modelling Conference, Lancaster, UK. **2013**

S. Hajebi, S. Temate, S. Barret, A. Clarke & S. Clarke. Water distribution network sectorization using structural graph partitioning and multiobjective optimization. In *Proc*. 16th Conference on Water Distribution System Analysis, Bari, Italy, **2014**

K. Hamber, G. Low & G. Stephens. Investigating the Applicability of Structural Analysis Technique in Distributed Systems. In G. Papadopoulos, G. Wojtkowski, W. Wojtkowski, S. Wrycza & J Zuüacjc (Edits). Information Systems Development. Toward a Service Provision Society, Springer. **2009**

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

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

T. Hastie, R. Tibshirani & J. Friedman. *Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer-Verlag. **2009**

J.C. Helton & F.J. Davis. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering & System Safety*, Vol. 81(1): 23–69. **2003**

M. Herrera. Improving Water Networks Management by Eficient Division into Supply Clusters. PhD thesis, *Universitat Politècnica de València*. **2011**

M. Herrera, J. Izquierdo, R. Pérez-García & I. Montalvo. Multi-agent adaptive boosting on semi-supervised water supply clusters. *Environmental Modelling and Software*, Vol.50: 131-136. **2012**

J.H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press. **1975**

A. Ilaya-Ayza, E. Campbell, R. Pérez-García & J. Izquierdo. Network capacity assessment and increase in systems with intermittent water supply. *Water*, Vol.8(4): 126. **2016**

A. Ilaya-Ayza, J. Benítez, J. Izquierdo & R. Pérez-García. Multi-criteria optimization of supply schedules in intermittent water supply systems. *Journal of Computational and Applied Mathematics*, Vol.309: 695-703. **2017**

J. Izquierdo, M. Herrera, I. Montalvo & R. Pérez-García. Division of Water Supply Systems into District Metered Areas using a Multi-agent based Approach. In Software and Data Technologies (Communications in Computer and Information Science), Springer-Verlag Berlin Heidelberg. **2011**

N. Jayaram & K. Srinivasan. Performance-based optimal design and rehabilitation of water distribution networks using life cycle costing. *Water Resources Research*, Vol.440(1). **2008**

N. Johnson. Personal Website. Department of Mathematics. The Ohio State University, Newark. **2017**. URL {http://nilesjohnson.net/} (Útima visita: Marzo 2017)

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

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

G. Keziban Orman, V. Labatut & H. Cherifi. On accuracy of community structure discovery algorithms. *Journal of Convergence Information Technology*, Vol. 6(11): 283-292. **2011**

B. Kingdom, R. Liemberger & P. Marin. The Challenge of Reducing Non-Revenue Water (NRW) in Developing Countries. How the Private Sector can Help: A Look at Performance-Based Service Contracting. The World Bank Group. **2006**

J.H. Lee, Y. Ko & I. Yun. Comparison of latin hypercube sampling and simple random sampling applied to neural network modelling of HfO2 thin film fabrication. *Transactions on Electrical and Electronic Materials*, Vol. **7**(4): 210-214. **2006**

M. Lin, J. Tsai & C. Yu. A review of deterministic optimization models in engineering and management. *Mathematical Problem in Engineering*. **2012**. DOI: 10.1155/2012/756023

A. Lambert & A. Lalonde. Using practical prediction of economic intervention frequency to calculate short-run economic leakage level, with or without pressure management. In *Proc. IWA Specialised Conference 'Leakage 2005'*, Halifax, Nova Scotia, Canada. **2005**

A. Lambert, T. Brown, M. Takizawa & D. Weimer. A review of performance indicators for real losses from water supply systems. *AQUA*, Vol. 48(6). **1999**

A. Lambert. Assessing non-revenue water and its components: a practical approach. IWA publishing. **2003**

A. Lambert & R. McKenzie. Practical experience in using the infrastructure leakage index. In *Proc*. IWA Conference 'Leakage Management – A Practical Aproach', Cyprus. **2002**

A. Lancichinetti & S. Fortunato. Limits of modularity maximization in community detection. *Physical Review E*, Vol. 84(6): 066122. **2011**

C. Lee. An algorithm for path connection and its applications. *IRE Transactions on Electronic Computers*, Vol. 10(3): 346-365. **1961**

G. Luque. Diseño Multiobjetivo de un Sistema de Abastecimiento de Agua Incluyendo la Cosecha de Agua de LLuvia como Recurso Complementario. Tesina de Máster. *Universitat Politècnica de València*. **2013**

D. Maringer. Portafolio Management with Heuristic Optmization. Springer. **2005**

O. Maimon & L. Rokach. Data Mining and Knowledge Discovery Handbook. 2nd Edit. Springer Science + Bussiness Media. **2010**

M. Mackey, R. Beckman & W. Conover. A comparison of three methods for selecting values of input variables in analysis of output from a computer code. *Technometrics*, Vol.21(2): 239-245, **1979**

C. Manning, P. Raghava & H. Schutze. Introduction to Information Retrieval. **2008** URL {http:http://nlp.stanford.edu/IR-book/html/htmledition/irbook.html} (Última visita: Marzo 2015)

A. Marchi *et al*. (2014). Battle of the Networks II. *Journal of Water Resources Planning and Management*, Vol. 140(7): 1-14. **2014**

J. May. Leakage, pressure & control. In *Proc*. BICS International Conference on Leakage Control Investigation in Underground, London, UK. **1994**

L. McQuitty. Similarity analysis by reciprocal pairs for discrete and continuous data. *Educational and Psychological Measurement*, Vol.26: 825-831. **1966**

R.W. Meier & B.D Barkdoll. Sampling design for network model calibration using genetic algorithms. *Journal of Water Resources Planning and Management*, Vol. 126(4): 245–250. **2000**

M. Mitchell. An Introduction to Genetic Algorithms. MIT Press. **1998**

M. Millet García. Diseño Óptimo de una Red de Distribución de Agua con Objetivos Múltiples Utilizando Métodos Heurísticos (Algoritmos Genéticos). Tesina de máster. *Universitat Politècnica de València*. **2013**

E. Mooi & M. Sardtedt. Cluster Analysis. In E. Mooi & S. Marko (Edits.), A Concise Guide to Market Research. The Process, Data, and Methods Using IBM SPSS Statistics. Springer- Verlag. **2011**

I. Montalvo. Diseño Óptimo de Sistemas de Distribución de Agua Mediante Agent Swarm Optimization. Tesis de Doctorado. *Universitat Politècnica de València*. **2011**

I. Montalvo. Diseño Óptimo de Sistemas de Distribución de Agua Mediante Particle Swarm Optimization. Tesina de máster. *Universitat Politècnica de València*. **2008**

I. Montalvo, J. Izquierdo, R. Pérez-García & M. Herrera. Water distribution system computer-aided design by agent swarm optimization. *Computer-Aided Civil and Infrastructure Engineering.* Vol.29(6): 433-448. **2014**

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

J. Morrison, T. Stephen & D. Roger. District Metered Areas: Guidance Notes. Technical Report 1. Water Loss Task Force-International Water Association (IWA). **2007**

F. Murtagh & P. Legendre. Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm. **2011** URL {http://arxiv.org/pdf/1111.6285.pdf.} (Útima visita: Marzo 2017)

F. Murtagh & P. Legendre. Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion? *Journal of Classification*, Vol. 31(3):274-295. **2014**

M. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, Vol. 74(3):36104. **2006**

M. Newman & M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, Vol. 69(2):026113. **2004**

M.E.J. Newman. The structure and function of complex networks. *SIAM Review*, Vol. 45(2): 167-256. **2003**

F. Newman & C. Carsten. Bioinspired Computation in Combinatorial Optimization. Algorithms and Their Computational Complexity. Springer. **2010**

C. Karlsson. Handbook of Research on Cluster Analysis. Elgar Publishing Limited. **2008**

J. Kennedy & R. C. Eberhart. Particle swarm optimization. In *Proc.* of IEEE International Conference on Neural Networks, Piscataway, New Jersey. **1995**

A. Konak, D. Coit & A. Smith. Multi-objective optimization using genetic algorithms: A tutorial. *Reliability Engineergin & System Safety*. Vol. 91(9). **2006**

T. Koppel & A. Vassiljev. Calibration of a model of an operational water distribution system containing pipes of different age. *Advances in Engineering Software.* Vol. 40(8): 659-664, **2009**

J. Osorio & J. Orejuela. El proceso de análisis jerárquico (AHP) y la toma de decisiones multicriterio-ejemplo de aplicación. *Scientia et Technica*, Vol. 2(39): 247-252. **2008**

I.H. Osman & G. Laporte. Metaheuristic: a bibliography. *Ann. Oper. Res*. Vol. 63: 513-623. **1996**

Palisade. Evolver. The Genetic Algorithm Solver for Microsoft Excel. Version 5.7. **2010**

Palisade. The Decision Tools Suite. **2017** URL {https:\\www.palisade.com} (Útima visita: Marzo 2017)

G. Palla, I. Derényi, I. Farkas & T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, Vol. 435(7043):814-818. **2005**

L. Perelman & A. Ostfeld. Short communication: Topological clustering for water distribution systems analysis. *Environmental Modelling and Software*, Vol. 26(7):969-972. **2011**

C. Pearson. The random walk. *Nature*, Vol. **72**: 318-342. **1905**

D. Pearson, M. Fantozzi, D. Soares & T. Waldron. Searching for N2: How does pressure reduction reduce burst frequency? In *Proc*. IWA Special Conference 'Leakage 2005', Halifax, Canada. **2005**

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

J. Podani & S. Dénes. On dengrogram-based measures of funcional diversity. *OIKOS*, Vol. 115 (1): 179-185. **2006**

P. Pons & M. Lapaty. Computing communities in large networks using random walks. *Journal of Graph Algorithms and Application*, Vol. 10(2): 191-218. **2005**

DT. Prasad & N. Park. Multi-objective genetic algorithms for the design of pipe networks. *Journal of Resources Planning and Management*, Vol.130(1):73-84. **2004**

R. Puust & A. Vassiljev. Real water network comparative calibration studies considering the whole process from Engineer's perspective. In *Proc.* 16th Conference on Water Distribution System Analysis. **2014**

U. Raghavan, R. Albert & S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, Vol. 76(3): 036106. **2007**

R Core Team. R: A Language and Environment for Statistical Computing. **2015**. URL {https://www.R-project.org/} (Útima visita: Marzo 2017)

J. Reichardt & S. Bornholdt. Statistical mechanics of community detection. *Physical Review E,* Vol. 74: 016110. **2006**

R. Rojas. Genetic Algorithms. In. Neural Network: A Systematic Introduction. Springer. **1996**

C. Romesburg. Cluster analysis for researches. Lulu Press. **2004**

L. Rossman. EPANET2 Users Manual. United States Environmental Protection Agency (EPA). **2000**

M. Rosvall & C. Bergstrom. Maps of information flow reveal community structure in complex networks. In *Proc.* National Academy of Sciences USA. **2008** DOI: 10.1073/pnas.0706851105.

R. Rotta & A. Noack. Multilevel local search algorithms for modularity clustering. *Journal of Experimental Algorithmics,* Vol. 16. **2011**

P. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, Vol. 20: 53–65. **1987**

D. Savic & G. Ferrari. Design and performance of distric metering areas in water distribution systems. In *Proc.* 16th Conference on Water Distribution System Analysis, Bari, Italy. **2014**

M. Savić, M. Radovanović & M. Ivanović. Community detection and analysis of community evolution in apache ant class collaboration networks. In *Proc.* 5th Balkan Conference in Informatics, Novi Sad, Serbia. **2012**

J. Sander. Generalized Desity-Based Clustering for Spatial Data Mining, Herbert Utz Verlag. **1999**

H. Shimodaira. Technical Details of the Multistep-Multiscale Bootstraap Resampling. Research Reports on Mathematical and Computing Sciences. Department of Mathematical and Computing Sciences Tokyo Institute of Technology. **2004** URL {http://www.is.titech.ac.jp/~shimo/pub/B403.pdf} (Última visita: Marzo 2017)

R. Sibson. Slink: an optimal efficient algorithm for the single-link cluster method. *The Computer Journal (British Computer Society),* Vol. 16(1): 30–34. **1973**

(SAWRC) South African Water Research Commission. Development of a Standardised Approach to Evaluate Burst and Background Losses in Water Distribution Systems in South Africa. SANFLOW VER 1.5. User Guide. SAWRC**. 1999**

R. Sokal & C. Michener. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, Vol. 28(1): 1409–1438. **1958**

R, Sokal & F. Rohlf. The comparison of dendograms by objective methods. *Taxon*, Vol.11 (2): 33-40. **1962**

K. Steinhaeuser & N. Chawla. Identifying and evaluating community structure in complex networks. *Pattern Recognition Letters*, Vol. 31(5): 413–421. **2010**

C. R. Suribabu & T. R. Neelakantan. Design of water distribution networks using particle swarm optimization. *Urban Water Journal*, Vol. 3(2): 111-120. **2006**

R. Suzuki & H. Shimodaira. Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, Vol.22 (12): 1540-1542. **2006**

J. Thornton & A. Lambert. Progress in practical prediction of pressure: leakage, pressure: burst frequency and pressure: consumption relationships. In *Proc.* IWA Special Conference 'Leakage 2005', Halifax, Nova Scotia, Canada, **2005**

J. Thornton & A. Lambert. Managing pressure to reduce new break frequencies, and improve infrastructure management. *Water* 21. **2006**

J. Thornton & A. Lambert. Pressure Management extends infrastructure life and reduces unnecessary energy cost. In *Proc*. IWA special conference 'Water Loss 2007', Bucharest, Romania. **2007**

J. Thornton & A. Lambert. Managing pressure to reduce new breaks. *Water Supply Networks*. **2006**

J. Thorton & A. Lambert. Pressure management extends infrastructure life and reduces unnecessary energy costs. In *Proc.* IWA Conference Water Loss 2007, Bucharest, Rumania. **2007**

J. Thornton, R. Sturm & G. Kunkel. Water Loss Control, McGraw-Hill. **2008**

K. Thulasiraman & M.N.S. Swamy. Graph Algorithms, in Graphs: Theory and Algorithms, John Wiley & Sons. **1992**

E. Todini. Looped water distribution networks design using a resilience index. *Urban Water*, Vol.2 (1): 115-122. **2000**

V. Tzatchkov, V. Alcocer-Yamanaka & V. Bourguett-Ortiz. Sectorización de redes de distribución de agua potable a través de algoritmos basados en la teoría de grafos. *Tlaloc-AMH,* Vol. 40(Enero-Febrero 2008):14–22. **2008**

V. Tzatchkov, V. Alcocer-Yamanaka & V. Bourguett-Ortiz. Graph theory based algorithms for water distribution network sectorization projects. In *Proc.* 8Th Water Distribution Systems Analysis Symposium, Cincinnati, Ohio, USA. **2008a**

V. Tzatchkov & V. Alcocer-Yamanaka. Graph partitioning algorithms for water distribution network sectorization projects. In *Proc.* 10th International Conference on Hydroinformatics, Hamburg, Germany. **2012**

(UKWIR) UK Water Industry. The 'Managing Leakage' Series of Reports, UKWIR. **1994**.

K. Wakita & T. Tsurumi. Finding community structure in mega-scale social networks. In *Proc*. International conference on WWW/Internet, Alberta, Canada. **2007**

J. Ward, Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, Vol. 58(301): 236–244. **1963**

S. Wasserman, J. Scott & P. Carrington. Introduction. In P. Carrington, J. Scott, and S. Wasserman (Edits), Models and Methods in Social Network Analysis. Cambridge University Press, Cambridge. **2005**

D.J. Watts & S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, Vol. 40(393). **1998**

M. Wooldrige. An Introduction to MultiAgent Systems. John Wiley & Sons Ltd. **2002**

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

W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*. Vol.33(452).**1977**

APÉNDICE I: MÉTODO DE CALIBRACIÓN MEDIANTE MAPAS AUTO ORGANIZADOS Y ALGORITMOS GENÉTICOS

VI. DESCRIPCIÓN DEL PROBLEMA

Típicamente la calibración de las RDAPs se lleva a cabo mediante la modificación exclusiva de la rugosidad de las tuberías sin tener en cuenta ni los caudales de fugas, ni su distribución. Según Meier & Barckdoll (2000), el uso de modelos de calibración que toman en cuenta como variables de ajuste sólo la rugosidad, pueden tener graves problemas de convergencia cuando las pérdidas de carga son extremadamente bajas. Otro punto señalado por los mismos autores se refiere al tema de agrupación de tuberías. En concreto, se establece que el agrupamiento de tuberías puede dejar la calibración mal condicionada, si alguno(s) de los grupos tienen baja representatividad de caudal, afectando negativamente el resultado de asignación final de rugosidades.

Cuando dentro de los modelos de RDAPs se toman en cuentan las fugas, es una práctica común asignar los coeficientes de emisor en los nudos a través de un balance de caudales, para lo que, primero se tiene que encontrar un coeficiente de emisor global, que permita hacer cumplir el balance de caudales diarios, mismo cuando este no se cumpla hora a hora. A continuación, se modifican, de manera iterativa, los factores horarios del patrón de demanda hasta lograr que los balances de caudales sí se cumplan hora a hora. El inconveniente de este abordaje, está relacionado con el hecho de que las fugas en una RDAP son dependientes de las presiones, con lo cual, si primero se asignan las fugas y luego se calibran las rugosidades, las variaciones en las rugosidades generarían variaciones en las presiones que luego alterarían el balance de caudales llevado a cabo inicialmente. Por el contrario, si primero se calibran las rugosidades y luego se asignan las fugas, la distribución de coeficientes de emisores estaría completamente condicionada por la distribución de rugosidades obtenidas inicialmente. Con lo cual, salta a la vista que para abordar este problema apropiadamente, entran en juego simultáneamente dos aspectos fundamentales, a saber: la distribución tanto de los valores de

rugosidad a través de la red, como de los coeficientes de fugas. Por tanto, el ajuste de ambos parámetros debe ser conjunto.

En el proceso de modelación de RDAPs, se considera error a la diferencia entre los valores obtenidos mediante el modelo hidráulico y los valores reales. El objetivo de la calibración es ajustar ciertos parámetros para que la red modelada sea lo más próxima a la real. En ese sentido, el problema de optimización se centra en la minimización de los errores mediante la modificación iterativa de un conjunto de variables de decisión.

Tomando en cuenta las características de los procesos de optimización (ver Subsección III.3 Generalidades sobre Optimización), la definición de la función objetivo es fundamental para que el resultado sea satisfactorio. Para ello, en este trabajo se propone la evaluación de siete funciones objetivo. Algunas de ellas se basan en trabajos previos y otras son completamente nuevas. También se hace un análisis de la influencia de cada función objetivo sobre el resultado final.

La calibración mediante técnicas de optimización ya ha sido previamente planteada de diversas maneras, incluyendo el uso de métodos lineales, no lineales y más recientemente, mediante algoritmos bioinspirados. En la línea de los algoritmos bioinspirados, los AGs han sido ampliamente aplicados para calibración de modelos de RDAP (Puust & Vassilvej, 2014).

El método que aquí se propone implica la calibración conjunta de rugosidades en las tuberías y coeficiente de emisor en los nodos, mediante un método en que primero se subdividen los nodos y las tuberías en comunidades a través de SOMs y luego se ajustan los variables de decisión a través de AGs. En la segunda parte de este apéndice se describe el problema de optimización; en la tercera parte se hace una descripción de un método para subdividir las tuberías y los nodos en clústeres; en la cuarta parte se describe el planteamiento y solución del problema, en una hoja Excel, mediante la aplicación para AGs en Evolver. Esta última parte viene acompañada de una discusión sobre una serie de funciones objetivo planteadas. Finalmente, en la quinta parte se establecen las conclusiones del trabajo. El presente trabajo fue publicado en el año 2015 (Brentan *et al*. 2015)

VII. PLANTEAMIENTO DEL PROBLEMA DE OPTIMIZACIÓN

El problema que se plantea en este trabajo corresponde a la minimización de las diferencia entre un conjunto de datos de caudal (medidos en las líneas de alimentación de la red) y presión (medidos en un número limitado de nodos) vs el mismo conjunto, pero simulado por el modelo matemático. Las variables de decisión que se establecen son los factores horarios del patrón de demanda, las rugosidades en las tuberías y los coeficientes de emisor en los nudos.

VIII. CLÚSTERING DE NODOS Y TUBERÍAS

Pese la importancia que tiene la correcta definición de la función objetivo, no se puede soslayar la importancia que tiene el correcto abordaje de la selección de las variables de decisión. Esto es especialmente importante en RDAPs de gran extensión (con miles de tuberías) donde la definición de cada elemento (tubería/nodo) como función objetivo, haría el problema inabordable. De ahí que sea importante subdividir los nodos y las tuberías en subgrupos a los cuales se les asignan el mismo valor de rugosidad/coeficiente de emisor. Esto implica un ahorro importante en términos de coste computacional. Vale la pena destacar el hecho de que la distribución de los coeficientes de emisor y de los valores de rugosidad en las RDAPs, ha sido uno de los aspectos menos estudiados dentro del campo de estudio de los modelos hidráulicos de RDAPs.

La subdivisión de nodos y tuberías en subgrupos en las RDAPs se podría realizar únicamente en función de un único criterio, cómo podría ser la edad o los diámetros, para el caso de las tuberías, o las elevaciones o las demandas, para el caso de los nudos. Sin embargo, tal y como es de esperar, no es viable que las variaciones de la rugosidad en las tuberías o de las fugas en los nodos dependan únicamente de una única variable. Por el contrario, se espera que los valores que pueden asumir estos parámetros a lo largo del tiempo dependan de una combinación de varios parámetros. Por ejemplo, en tuberías con valores de diámetro y edades similares es de esperar que los valores de rugosidad sean muy similares (si no iguales). Por ello, se hace evidente la validez de la aplicación de una técnica de clústering para subdividir ambos tipos de elementos en subgrupos (clústeres).

Esta tarea puede llevarse a cabo mediante diversas técnicas de clústering (ver más detalles sobre clústering en Subsección II.1.4 Teoría de Formación de Clústeres).

VIII.1 Mapas Auto-organizados (SOMs)

Los SOMs pueden ser considerados como un tipo de red neuronal. Sin embargo, estos difieren de las anteriores, en el hecho de que los nodos (o neuronas) de un SOM no se encuentran interconectados entre sí. Por lo general, los SOMs se emplean para representar bases de datos conformadas por vectores multidimensionales, en una estructura bidimensional de fácil visualización. Sin embargo, tal como se describe a continuación, también pueden ser empleados como técnica de clústering. En el campo de la optimización en el contexto de las RDAPs han sido empleados por Izquierdo *et al*. (2014a; 2016 a-b) para inyectar reglas a fin de mejorar el proceso de optimización con ASO.

El primer paso para la construcción de un SOM es el establecimiento de su estructura topológica. Para ello, se tiene que definir el número de neuronas de que constará. La selección de éste número depende de la estructura implícita de la base de datos con la que se esté lidiando. Lógicamente, la estructura, en un principio sólo se puede intuir, por lo cual, una aproximación válida, sería probar distintas configuraciones de tamaño para el SOM y ver cuál es la que mejor se ajusta al conjunto de datos que se está estudiando. En un buen SOM, se encontraría una distribución bastante uniforme de los vectores multidimensionales entre todas las neuronas (no habría neuronas con cantidades de casos, significativamente distintas). Otra medida que se puede utilizar como referencia es la distancia entre los elementos que existen dentro de una misma neurona. De manera ideal, esta distancia tendría que ser la mínima posible.

Una vez se define la topología del SOM, se tienen que definir algunos parámetros con los que se entrenará a la red, tales como: número de veces que se presenta la base de datos al mapa; una tasa de aprendizaje; un radio de acción sobre las neuronas vecinas. Al ejecutar este paso, se crea, para cada neurona, un vector artificial, de ahora en adelante llamado *codebook*, mediante el cual se comparan los vectores multidimensionales que se desean mapear en el SOM.

En el proceso de entrenamiento, se le presenta al SOM el conjunto de datos, una cantidad *n* de veces. Para cada uno de los casos, se mide una distancia entre el mismo y todos los *codebooks* existentes en el SOM. La neurona cuyo *codebook* tenga la menor distancia al caso de comparación, se define como Unidad de Mejor Acople (o BMU, por las siglas en inglés de *Best Matching Unit*). El *codebook* original de la neurona en cuestión es, entonces, modificado para hacerlo más similar al vector de comparación y, a partir de ahí, se inicia un proceso de influencia del *codebook* de la BMU sobre los *codebooks* ubicados en las neuronas vecinas. Dicho proceso se lleva a cabo a través de un radio de acción temporal que se va contrayendo gradualmente. Este paso es el que le confiere la propiedad de vecindad a los SOMs, al tiempo que establece que los vectores con un significativo grado de similitud se agrupen en áreas cercanas del mapa; y es, bajo este principio, que se fundamenta la utilización de los SOMs para realizar clústering de bases de datos.

VIII.1.1 Características de Nodos y de Tuberías

Tal y como ya ha sido previamente anunciado, en este trabajo se lleva a cabo el clústering tanto de tuberías así como de nodos. Para ello, primero se tuvo que analizar qué características de los nodos y qué características de las tuberías podrían ser más adecuadas para revelar las clases subyacentes existentes entre el conjunto de elementos.

En un modelo hidráulico, los elementos se pueden caracterizar por variables dependientes y no dependientes de la simulación hidráulica. Si sólo se tienen en cuenta aspectos netamente hidráulicos, dentro de la primera categoría se pueden encontrar: la rugosidad, la longitud, el diámetro, para el caso de las tuberías, y cota, demanda base, coeficiente de emisor, coordenadas *x-y*, para el caso de los nodos (estas dos últimas variables están más relacionadas con el aspecto topológico que con el hidráulico). Tal y como ya ha sido mencionado anteriormente, la rugosidad y el coeficiente de emisor, son las variables incógnitas en el proceso de calibración, por lo que no se pueden considerar como criterios de clústering.

La longitud de tubería no es muy útil como criterio de clústering, ya que los valores asumidos por dicha variable dependen del nivel de detalle/simplificación con el que está definido el modelo. En general, se espera que tuberías con distintos valores de longitud, pero con valores similares de otras características, terminen asumiendo el mismo valor de rugosidad (o al menos, valores muy similares). En cambio, el diámetro, sí se puede considerar como una buena variable discriminatoria. En el caso de los nudos, la cota, junto con las coordenadas geográficas, son variables muy útiles para revelar clases entre el conjunto de nodos, ya que se espera que zonas con cotas similares tengan tanto valores de presión como ritmos de mantenimientos similares, y que, por ende, los caudales de fugas se asemejen mucho entre sí.

A continuación se listan las variables empleadas para llevar a cabo el proceso de clústering.

Características en las tuberías:

Edad: con el paso del tiempo, las tuberías van acumulando material, lo cual aumenta la rugosidad en las paredes de las tuberías.

Material: algunos materiales son más sensibles a la acumulación de materia o a la corrosión. En general, en materiales plásticos, la ocurrencia de incrustaciones es menos frecuentes que en materiales como el hierro y el cemento.

Características en los nodos:

Cota: La presión en las distintas zonas de la red, depende en gran medida, de su cota. En zonas más bajas, es de esperar presiones más elevadas.

Coordenadas geográficas: tal y como se señaló anteriormente, se espera que nodos ubicados en la misma zona tengan ritmos similares de mantenimiento, presiones similares y, por ende, caudales de fugas similares.

En la Ilustración 1 y la Ilustración 2 se muestran las características anteriormente descritas sobre la red de Managua.

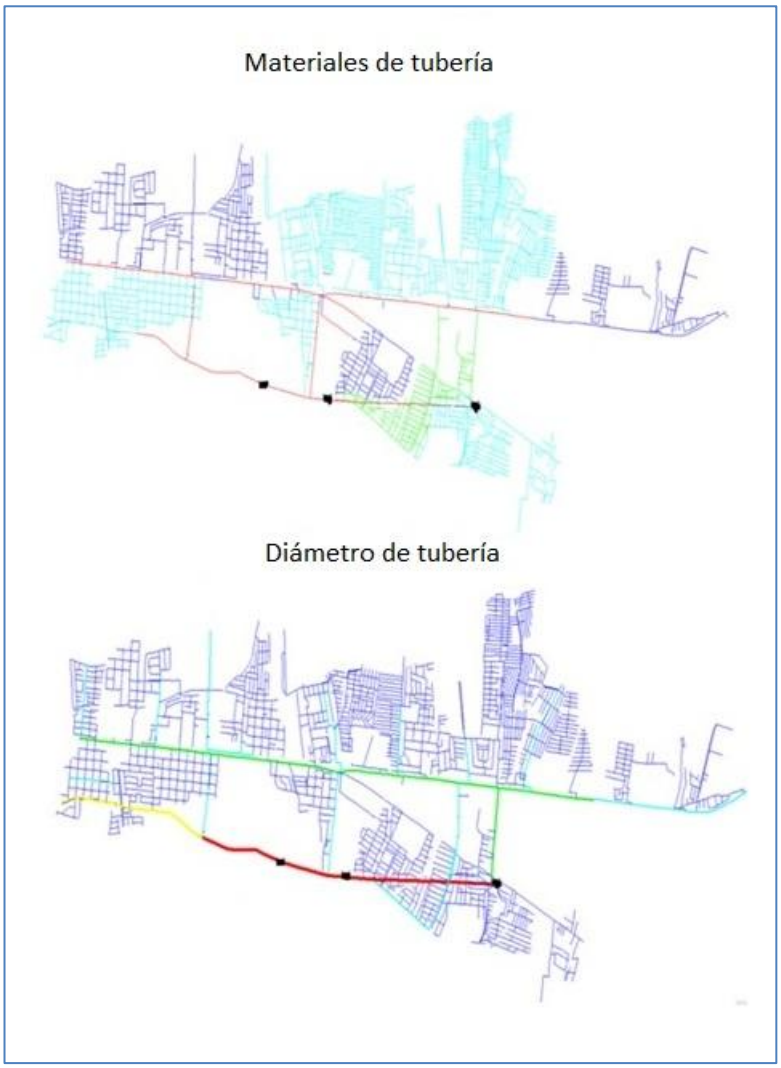


Ilustración 1: Información tuberías empleada para el clústering

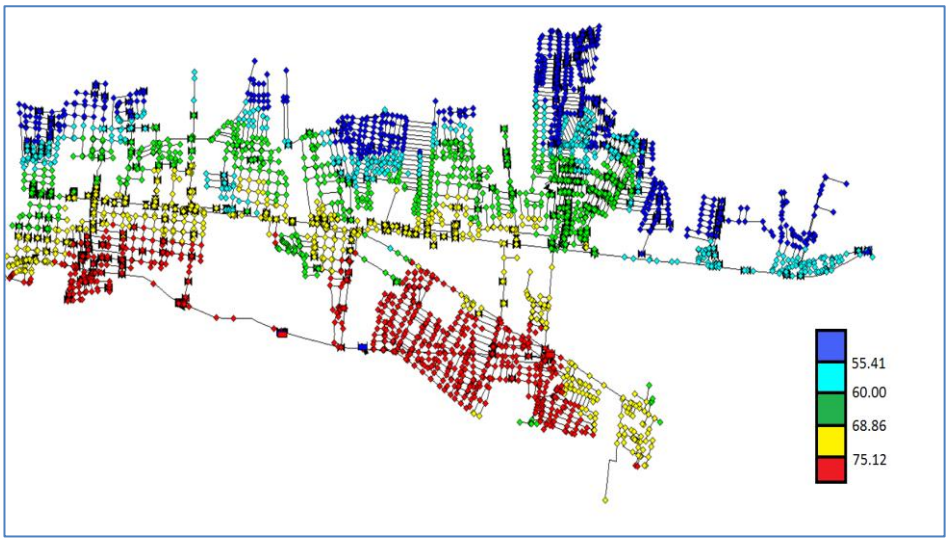


Ilustración 2: Información de nodos empleada para el clústering

En las siguientes Ilustraciones se muestran los SOMs obtenidos, tanto para los nodos (Ilustración 3) como para las tuberías (Ilustración 4). Nótese cómo para el caso de las tuberías, la cantidad de neuronas es significativamente diferente que en el caso de los nodos.

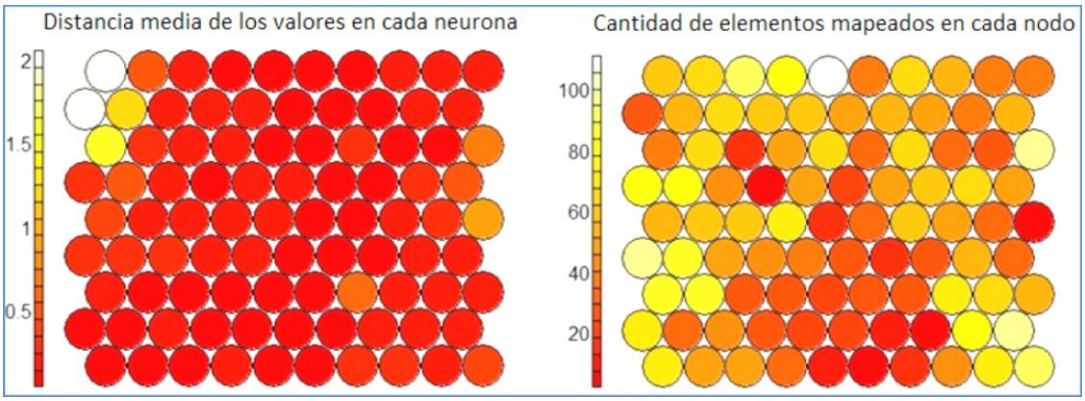


Ilustración 3: SOM generado para los nodos

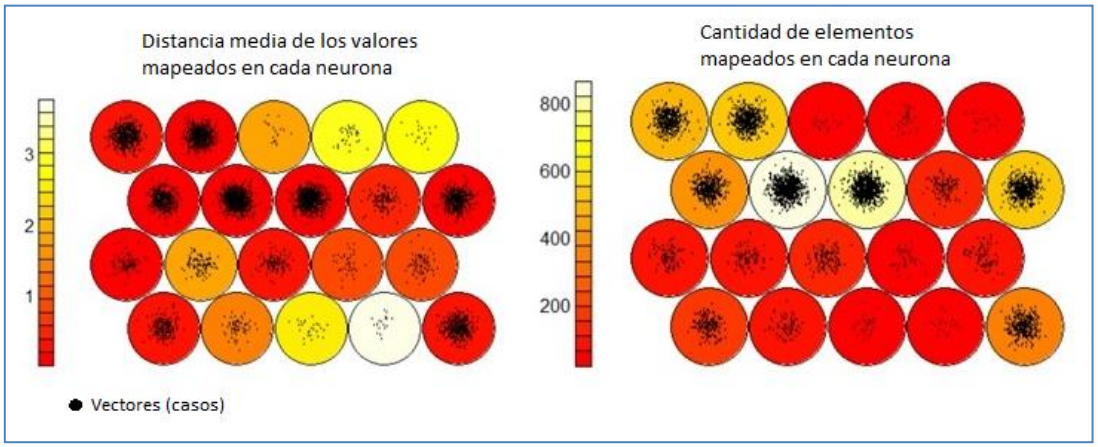


Ilustración 4: SOM generados para las tuberías

VIII.1.2 Clústering jerárquico sobre Mapas Auto-Organizados

Como ya ha sido ampliamente descrito, el clústering jerárquico es una herramienta de análisis de datos que se basa en la construcción de clústeres de los datos siguiendo un principio de jerarquía. Se agrupa dentro de los métodos de aprendizaje automático no supervisados y tiene como objetivo global hacer una exploración de los datos. Una vez generado un SOM, se puede implementar el clústering de los elementos en clases. Tal como ya se mencionó, en los SOM, las clases están representadas como *codebooks*, y es sobre ellos que se lleva a cabo el proceso de

clústering jerárquico. Dadas las particiones sobre el SOM, se procede a definir el clúster al que pertenece cada uno de los nodos y se evalúa el índice de silueta (IS) para cada una de las particiones. La idea es encontrar una combinación de medida métrica, método de fusión y número de partición que maximicen el índice en cuestión. En este trabajo se probaron sólo un número limitado de combinación de manera manual; sin embargo, sería interesante considerar el diseño de un método automático para el desarrollo de esta tarea.

En la Ilustración 5 (derecha) se muestran la partición, tanto para el SOM de nodos como para el SOM de tuberías.

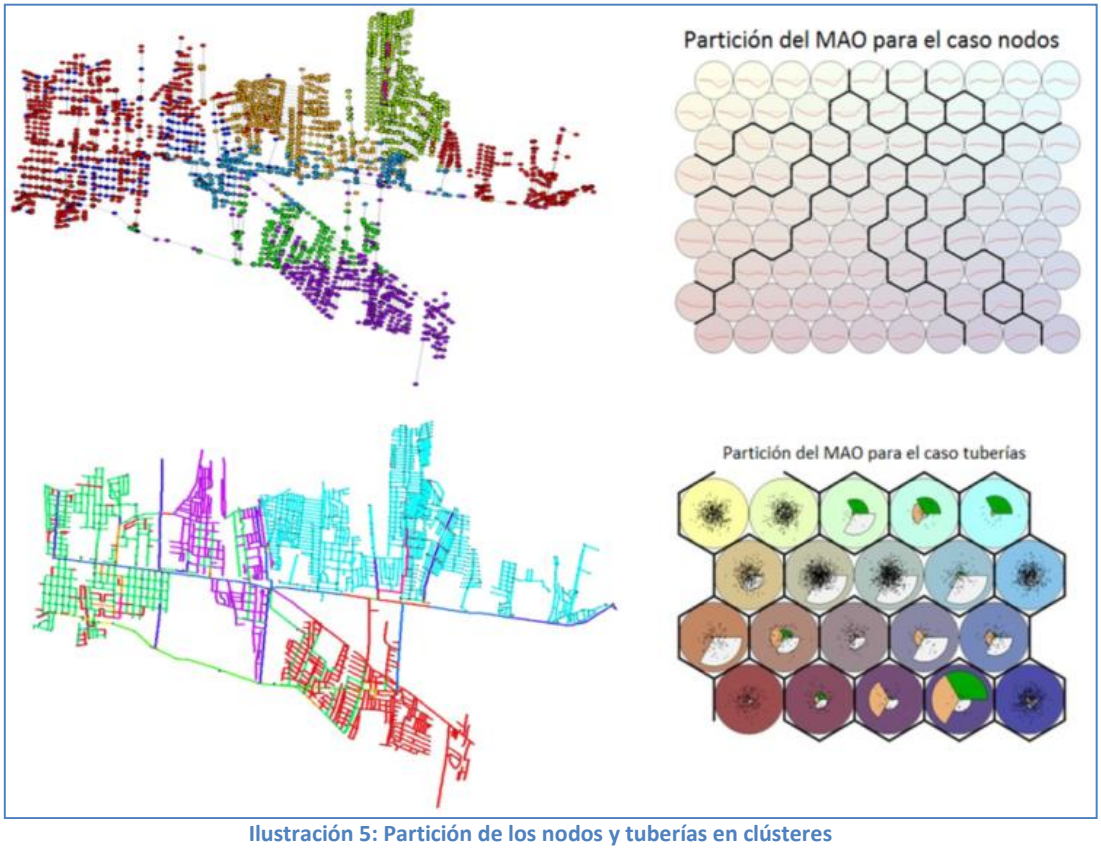


Ilustración 5: Partición de los nodos y tuberías en clústeres

Las variables de decisión corresponden a: los factores horarios de la curva de demanda, las rugosidades de las tuberías y los coeficientes de emisor de los nudos. Teniendo en cuenta que los coeficientes de emisor corresponden a 24, el total de variables de decisión sería 24 + el número de clústeres de nodos + el número de clústeres de tuberías.

En la hoja Excel, se establece un conjunto de celdas en donde se ubican todas las variables decisión. Al inicio se coloca un valor razonable, luego, mediante la macro escrita en Visual Basic, se leen los valores escritos en estas celdas y se ejecuta el modelo matemático. Los resultados arrojados por el mismo (caudales y presiones) en los puntos de control se exportan a la hoja Excel y son comparados con el conjunto de datos de referencia, se evalúa la función objetivo y luego se procede a la siguiente iteración.

Para el ejemplo de implementación, los parámetros del AG que se establecieron fueron: un conjunto de 60 individuos como población inicial, una tasa de cruzamiento equivalente a 0.1 y una tasa de mutación equivalente a 0.5.

Como ya fue planteando previamente, la obtención de un resultado satisfactorio mediante AGs está íntimamente vinculado a la determinación de una función objetivo apropiada. Todo el proceso de optimización pasa por la evaluación de esa función y en el caso del uso de AGs, una buena función objetivo puede hacer que la superficie de búsqueda sea más suave y que tenga menos regiones de óptimos locales. En ese sentido, el estudio más amplio de la función a ser aplicada en la calibración es un paso fundamental para la obtención de un modelo que pueda ser implementado por empresas operadoras de RDAPs.

La manera más sencilla de abordar la calibración de RDAPs es mediante una función que minimice la diferencia absoluta entre las medidas hechas en campo y los valores modelados, tal y como se muestra en la Ecuación 1.

$$\text{Mín } F(x) = \sum_{i=1}^{n} \left| x_i^0 - x_i^m \right| \qquad \text{(Ecuación 1)}$$

donde x_i^0 es el valor medido y x_i^m es el valor modelado en una red con n valores de control.

Dicha función objetivo fue usada por Vitkovsky *et al*. (2000) para la calibración y detección de fugas usando AGs a fin de minimizar los errores entre los valores de presión en los nodos.

Ormsbee (1989) presentan una variante de la Ecuación 1 en la que utiliza la diferencia relativa entre el modelo y la red real. En este caso se toma en cuenta la presión en los nudos como variable de medición (ver Ecuación 2).

$$\text{Mín } F(x) = \sum_{i=1}^{n} \frac{\left|x_i^0 - x_i^m\right|}{x_i^0} \qquad \text{(Ecuación 2)}$$

Otra variación de la Ecuación 1 es la presentada por Kopple & Vassiljev (2009). En esta (Ecuación 3) se plantea la minimización del error cuadrático entre los valores reales y los modelados.

$$\text{Mín } F(x) = \sum_{i=1}^{n} (x_i^0 - x_i^m)^2 \qquad \text{(Ecuación 3)}$$

Vale la pena destacar que en todos los trabajos previos, sólo se utilizan como referencia los valores de presión nodal. Dado que muchas veces, en función del tamaño de la red, la cantidad valores a determinar es bastante grande, un número reducido de puntos de presión puede que no sea suficiente para obtener un buen resultado. En este sentido, Lansey *et al.* (2001) presentan un modelo en el que no sólo se usa la presión como referencia, sino que también se toma en consideración los caudales en las tuberías y los niveles de los tanques (Ver ecuación 4).

$$\text{Mín } F(H, Q, L) = \sum_{i=1}^{j} \left|H_i^0 - H_i^m\right|^n + \sum_{k=1}^{p} \left|Q_k^0 - Q_k^m\right|^n + \sum_{t=1}^{T} |L_t^0 - L_t^m|^n \qquad \text{(Ecuación 4)}$$

Donde: H_i^0 es la carga total en el nudo i, con j nudos y H_i^m es la carga total modelada en el nudo i. Q_k^0 es el caudal medido en una tubería k de una red con p tuberías, Q_k^m es el caudal modelado en la tubería k y finalmente L_t^0 es el nivel del tanque t en la red con T tanques.

En el presente trabajo se presenta y evalúan el siguiente conjunto de ecuaciones.

$$\text{Min } F(Q,H) = \sum_{k=1}^{p} \frac{|Q_k - Q_k^m|}{\bar{Q}} + \sum_{i=1}^{p} \frac{|H_i - H_i^m|}{\bar{H}} \qquad \text{(Ecuación 5)}$$

$$\text{Min } F(Q,H) = \sum_{k=1}^{p} \frac{|Q_k - Q_k^m|}{Max\,(Q_k)} + \sum_{k=1}^{p} \frac{|H_i - H_i^m|}{Max\,(H_i)} \qquad \text{(Ecuación 6)}$$

$$\text{Min } F(Q,H) = \sum_{k=1}^{p} \frac{|Q_k - Q_k^m|}{Min\,(Q_k)} + \sum_{k=1}^{p} \frac{|H_i - H_i^m|}{Min\,(H_i)} \qquad \text{(Ecuación 7)}$$

$$\text{Min } F(Q,H) = \sum_{k=1}^{p} \frac{|Q_k - Q_k^m|}{\sigma_q} + \sum_{k=1}^{p} \frac{|H_i - H_i^m|}{\sigma_h} \qquad \text{(Ecuación 8)}$$

$$\text{Min } F(Q,H) = \sum_{k=1}^{p} \frac{|Q - Q_k^m|}{Q_k} + \sum_{k=1}^{p} \frac{|H_i - H_i^m|}{H_i} \qquad \text{(Ecuación 9)}$$

$$Min\ F(Q,H) = \sum_{k=1}^{p} |Q_k - Q_k^m| + \sum_{k=1}^{p} |H_i - H_i^m| \qquad \text{(Ecuación 10)}$$

$$\text{Min } F(Q,H) = \sum_{k=1}^{p} (Q_k - Q_k^m)^2 + \sum_{k=1}^{p} (H_i - H_i^m)^2 \qquad \text{(Ecuación 11)}$$

La Tabla 1 presenta los resultados obtenidos mediante cada una de las funciones objetivo en cada uno de los siete nudos de control (Puntos de Control: PC).

Tabla 1: Resultado de la evaluación de las funciones objetivos

Función Objetivo	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Ecuación 5	66,72%	75,66%	55,52%	88,08%	97,61%	97,61%	45,33%
Ecuación 6	92,83%	84,15%	91,24%	74,14%	92,83%	88,87%	84,15%
Ecuación 7	92,03%	57,55%	69,65%	86,50%	73,39%	95,22%	64,55%
Ecuación 8	84,15%	45,33%	79,49%	80,26%	97,61%	89,66%	89,66%
Ecuación 9	92,83%	54,85%	95,22%	80,26%	74,14%	99,20%	31,73%
Ecuación 10	54,85%	56,87%	77,95%	71,14%	65,27%	47,77%	0,96%
Ecuación 11	61,71%	31,25%	0,60%	68,18%	81,81%	81,03%	0,63%

Al comparar los valores anteriores, se puede determinar que la función representada por la Ecuación 11 es la que presenta peores resultados (ver Ilustración 6).

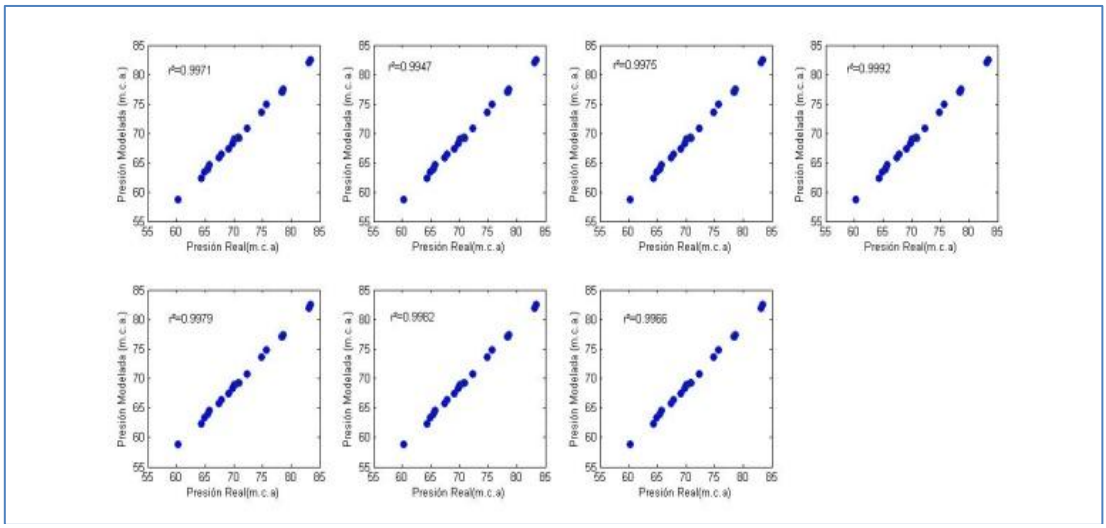


Ilustración 6: Correlación entre los datos medidos y los datos del modelo para cada PC

Para ampliar más el análisis de la calidad de las funciones objetivo, se tomó en consideración otro factor importante, como lo es, la convergencia del método en términos de velocidad. Es decir, la agilidad del método para converger una vez definidas sus variables. En ese trabajo, el Índice de Calidad (IC) desarrollado, tiene en cuenta el número de iteraciones con el que se llega al resultado esperado.

Para calcular IC, se hace una multiplicación entre todos los coeficientes de correlación obtenidos, y el resultado se divide entre el número de iteraciones llevadas a cabo. De esta manera, si el modelo tuviera un desempeño excelente, se obtendría como resultado un valor igual a 1, ya que el máximo valor asumido por r^2 es 1. Por otro lado, esa multiplicación se divide entre una fracción en la cual el

denominador es una división entre el número de iteraciones necesarias para la convergencia del modelo y el número mínimo de iteraciones en todos los test (Ecuación 12).

$$IC_0 = \frac{\prod_{e=1}^{E} r_e^2}{Iter_0 / \text{mín}(Iter)} \qquad \text{(Ecuación 12)}$$

donde r_e^2 es el coeficiente de correlación para un elemento en una red con E elementos, $Iter_0$ es el número de iteraciones necesarias para la convergencia del método e $Iter$ es el vector con todos los valores de iteracciones necesarias para la convergencia.

De acuerdo a los IC que se presentan en la Tabla 2, la función objetivo representada por la Ecuación 6, es la que tiene mayor eficiencia presenta, sin embargo, el resto de las funciones, con excepción de la representada por la Ecuación 11, presentan valores aceptables.

Tabla 2: IC para cada una de las funciones objetivo

Función objetivo	IC
Ecuación 5	0.76127
Ecuación 6	0.95826
Ecuación 7	0.88514
Ecuación 8	0.87499
Ecuación 9	0.82066
Ecuación 10	0.87516
Ecuación 11	0.57070

APÉNDICE II: PROGRAMACIÓN Y LIBRERÍAS CREADAS

LIBRERIAS CREADAS:

Crear grafos: crean un grafo a partir de archivo INP.

Shortest.path.tuberia: calcula el ranking de tuberías en función de los caudales.

Segregación de red troncal: dado cierto criterio segrega la red distribución de la red troncal.

Pipelenghtcalc: calcula la longitud de tubería de una partición en sectores.

Pipelenghtcalc: refusiona las comunidades obtenidas por otros algoritmos.

DEFINICION DE COMUNIDADES

```
# library:(cluster)
library(igraph)
library(reshape2)
library(Hmisc)
library(reshape)
library(qcc)
rm(list = ls())
load("remuestreo")
load("C:/Users/enrique/Documents/PHD/TESIS/Ejemplos/Edge
Betweenness/archivos secundarios/crear.grafo.R")
load("C:/Users/enrique/Documents/PHD/TESIS/Ejemplos/Edge
Betweenness/archivos secundarios/shortest.paths.tuberias.R")
load("C:/Users/enrique/Documents/PHD/TESIS/Ejemplos/Edge
Betweenness/archivos secundarios/segregacion.red.troncal.R")
load("C:/Users/enrique/Documents/PHD/TESIS/Ejemplos/Edge
Betweenness/archivos secundarios/pipelenghtcalc.R")

#nodesdat<-
read.table("C:/Users/enrique/Desktop/codigoMultinivelBatalla/RedBatalla/nod
esdat.csv", header=TRUE, sep=";")
#pipesdat<-
read.table("C:/Users/enrique/Desktop/codigoMultinivelBatalla/RedBatalla/pip
esdat.csv", header=TRUE, sep=";")

#nodesdat<-
read.table("C:/Users/enrique/Documents/PHD/TESIS/Ejemplos/ArchivosComunes/R
edManagua/nodesdat.csv", header=TRUE, sep=";")
#pipesdat<-
read.table("C:/Users/enrique/Documents/PHD/TESIS/Ejemplos/ArchivosComunes/R
edManagua/pipesdat.csv", header=TRUE, sep=";")

nodesdat<-      read.table("C:/Users/enrique/Desktop/Figuras      Tesis/Red
Grande/nodesdat.csv", header=TRUE, sep=";")
pipesdat<-      read.table("C:/Users/enrique/Desktop/Figuras      Tesis/Red
Grande/pipesdat.csv", header=TRUE, sep=";")
plotGrafoInit <- "si"

#CREACION DEL GRAFO INICIAL----------------------------
tipo<-1:nrow(pipesdat)
tipo<-1
pipesdat<-cbind(pipesdat, tipo)
```

```
ID<-paste(nodesdat[,1], "x", sep = "")
ID<-gsub(" ", "", ID, fixed = TRUE )
nodesdat<-cbind(ID, nodesdat[,-1])

namespipes1<-paste(pipesdat[,1], "x", sep = "")
namespipes1<-gsub(" ", "", namespipes1, fixed = TRUE )
ID<-namespipes1
namespipes2<-paste(pipesdat[,2], "x", sep = "")
namespipes2<-gsub(" ", "", namespipes2, fixed = TRUE )
Node1<-namespipes2
namespipes3<-paste(pipesdat[,3], "x", sep = "")
namespipes3<-gsub(" ", "", namespipes3, fixed = TRUE )
Node2<-namespipes3
pipesdat<-pipesdat[,-1]
pipesdat<-pipesdat[,-1]
pipesdat<-pipesdat[,-1]
pipesdat<-cbind(ID, Node1, Node2, pipesdat)

estado<-lapply(pipesdat,  function(x)  ifelse(pipesdat$Status  !=  "Open",
"Open", "Open"))
pipesdat$Status <- estado$Status

pipesdat<-pipesdat[!(pipesdat$Status == "Closed"),]

ra<-1:nrow(nodesdat)
nodesdat<-cbind(nodesdat, ra)
nodesdat$ra<-0

for(id in 1:nrow(pipesdat)){
  nodesdat$ra[nodesdat$ID            %in%            pipesdat$Node1[id]]<-
as.character(pipesdat$Node1)[id]}
for(id in 1:nrow(pipesdat)){
  nodesdat$ra[nodesdat$ID            %in%            pipesdat$Node2[id]]<-
as.character(pipesdat$Node2)[id]}

nodesdat<-nodesdat[!(nodesdat$ra == 0),]
nodesdat<-nodesdat[,1:5]

grafo1<-crear.grafo(nodesdat, pipesdat)

par(mar=c(0,0,0,0))

if(plotGrafoInit == "si"){

plot(grafo1[[1]], layout = grafo1[[2]], vertex.size = 0.0001, vertex.label
= NA, vertex.color = "black", edge.arrow.size = 0.0001, vertex.label.cex =
0, edge.label.cex = 0.6, vertex.label.dist = 0.3, edge.label = NA)

}
g<-grafo1[[1]]


#CÁLCULO DE SHORTEST PATHS-----------------------------------
#calculo de shortest paths
sp<-shortest.paths.tuberias.test(grafo1[[1]], nodesdat, pipesdat)
tabla<-sp[[2]]
#plot(grafo1[[1]],   edge.arrow.size   =   0.01,   layout   =   grafo1[[2]],
vertex.size   =   0.3,   vertex.color   =   "gray",   vertex.label   =   NA,
vertex.label.cex = 0.3, edge.label.cex = 0.5, edge.label = sp[[2]][,11])

#tabla de frecuencia
factorx <- factor(cut(tabla$peso, breaks=100))
xout <- as.data.frame(table(factorx))
xout   <-   transform(xout,   cumFreq   =   cumsum(Freq),   relative   =
prop.table(Freq))

xout1 <- xout[order(xout[,1]),]
```

```
#DEFINICION CRITERIO SELECCION DE RED TRONCAL---------------------

xout2<-xout$Freq
names(xout2)<-xout1$factorx
par(mar=c(6,2,2,2))
pareto.chart(xout2, main = "Distribution of ASPV")

xout3<-xout2[2:length(xout2)]
par(mar=c(9,4,2,2))
barplot(xout3, col = terrain.colors(10), cex.names = 0.9, las = 3, ylab =
"Frequency", cex.axis = 1)

criterio<- 10


#SEGREGACION DE LA RED TRONCAL----

tipo<-1:nrow(pipesdat)
tipo<-0
pipesdat1<-cbind(pipesdat, tipo)

vcmca<-sp[[1]]
pipesdat1<-cbind(pipesdat1,vcmca)

pipesdat1$tipo<-1
for(i in 1:nrow(pipesdat1)){
  if(pipesdat1$vcmca[i] >= criterio){
    pipesdat1$tipo[i] <-2
  }
}


#RED INICIAL-----------------------------
red_distribucion<-segregacion.red.troncal(criterio, nodesdat, sp[[2]])

#Dibujar grafo reducido
grafo_reducido<-crear.grafo(red_distribucion[[1]], red_distribucion[[2]])

plot(grafo_reducido[[1]], layout = grafo_reducido[[2]], vertex.size = 0.05,
vertex.color = "gray", edge.arrow.size = 0.01, vertex.label.cex = 0.5,
edge.label.cex = 0.6, vertex.label.dist = 0.3, vertex.label = NA)

vcmca<-sp[[1]]
pipesdat<-cbind(pipesdat,vcmca)
for(i in 1:nrow(pipesdat)){
  if(pipesdat$vcmca[i] >= criterio){
    pipesdat$tipo[i] <-2
  }
}

Elev<-red_distribucion[[1]][c("Elev")]
Demand<-red_distribucion[[1]][c("Demand")]

nodesdatax<-red_distribucion[[1]][c("ID", "Demand", "X.Coord", "Y.Coord")]
pipesdatax<-red_distribucion[[2]][c("Node1", "Node2", "Length", "peso",
"Diameter")]
grafox<-grafo_reducido[[1]]

#ALGORITMO COMUNIDAD-----------------

grafox1 <- as.undirected(grafox, "each")

coor2<-as.matrix(nodesdatax[c("X.Coord", "Y.Coord")])
wt<-walktrap.community(grafox, weights = NULL, steps = 3000, merges = TRUE,
modularity = TRUE, membership = TRUE)
#wt<-multilevel.community(grafox1)
value<-length(wt)
value
```

```
dd<-as.hclust(wt, hang=-1, use.modularity=FALSE)
plot(dd, cex = 0.1, hang = -0.1, xlab = NA, main = "Communities merges")
ns<-65
x<-rect.hclust(dd, k = ns, border = 1)#CON ESTO SE OBTIENE LA LISTA DE
SECTORES A CADA UNA DE LAS ALTURAS DEL DENGROGRAMA
ap<-melt(x)
ap <- ap[order(ap[,1]),]
pertenencia<-ap[,2]
valorTemp<-1:length(pertenencia)
valorTemp<-0
ap<-data.frame(pertenencia, nodesdatax, valorTemp)
rm(x)
rm(dd)

grafo1<-crear.grafo(nodesdat, pipesdat)
sp<-shortest.paths.tuberias.test(grafo1[[1]], nodesdat, pipesdat)
red_distribucion<-segregacion.red.troncal(criterio, nodesdat, sp[[2]])


#---------PLOT DE PARTICION ELEGIDA

pipesdatx1<-red_distribucion[[2]]

for(i in 1:nrow(comunidades)){
  pipesdatx1$PertNode1[pipesdatx1$Node1      %in%      comunidades$ID[i]]<-
comunidades$pertenencia[i]
}

pipesdatx1$PertNode1[pipesdatx1$Node1       %in%       comunidades$ID[i]]<-
comunidades$pertenencia[i]


for(i in 1:nrow(comunidades)){
  pipesdatx1$PertNode2[pipesdatx1$Node2      %in%      comunidades$ID[i]]<-
comunidades$pertenencia[i]
}

pipedattemp<-pipesdatx1

# cargar pertenencia en las tuberias
pipesPert<-1:nrow(pipedattemp)
pipesPert<-100
pipedattemp<-cbind(pipedattemp, pipesPert)

subset_tuberia<-subset(pipedattemp,        (pipedattemp$PertNode1        ==
pipedattemp$PertNode2))

for(i in 1:nrow(subset_tuberia)){
  pipedattemp$pipesPert[pipedattemp$ID     %in%     subset_tuberia$ID[i]]<-
subset_tuberia$PertNode2[i]
}

grafoxtemp<-grafox

colbar<-rainbow(ns)
l<-sample(1:ns, replace = F)
colbar2<-sort(colbar)[l]
colbar1<-colbar2

E(grafoxtemp)$color <- colbar1[factor(pipedattemp$pipesPert)]
V(grafoxtemp)$color <- colbar2[factor(ap$pertenencia)]


E(grafoxtemp)[(pipedattemp$pipesPert == 100)]$width<-0.1
E(grafoxtemp)[(pipedattemp$pipesPert == 100)]$color<-"black"
E(grafoxtemp)[(pipedattemp$pipesPert != 100)]$width<-2
E(grafoxtemp)[(pipedattemp$pipesPert == 0)]$width<-1
```

```
  E(grafoxtemp)[(pipedattemp$tipo == 2)]$width<-7
E(grafoxtemp)[(pipedattemp$tipo == 2)]$color<-"gray"
E(grafoxtemp)[(pipedattemp$Status == 1)]$color<-"black"
E(grafoxtemp)[(pipedattemp$Status == 1)]$width<-8

par(mar=c(0,0,0,0))
  plot(grafoxtemp, edge.arrow.size = 0, vertex.size = 0.00009,layout =
coor2, vertex.label = NA, vertex.label.dist = 0.1, vertex.label.cex = 0.5)


# Busqueda de comunidades desconectadas-----------------------------

ComponentesGrafox3<-clusters(grafox3)

col<-melt(ComponentesGrafox3$membership)
ap5<-cbind(nodesdatax, col)

factores<-unique(ap5$value)
nuevos_id_sectores<-1:length(factores)
df_nuevosID<-as.data.frame(cbind(factores, nuevos_id_sectores))

for(i in 1:nrow(df_nuevosID)){
  ap5$valorTemp[ap5$value          %in%          df_nuevosID$factores[i]]<-
df_nuevosID$nuevos_id_sectores[i]

}
ap$pertenencia<-ap5$valorTemp


#CRITERIO PARA FUSION DE COMUNIDADES----

criteriofusion <- "longitud"

#FUSION DE COMUNIDADES-------------------

NumeroSectoresInit<-0
MaxLongitud<-0

vabcontrol<-3

ns1<-1
ns2<-2

while(ns1 != ns2 ) {

  vabcontrol <-3

  long_tuberia <- c()

 ID<-unique(ap[,1])
  for(i in 1:length(ID)[1]){
  long_tuberia[i] <- (pipelenghtcalc(i, ap, pipesdatax))}

vector<-(long_tuberia>=0 & long_tuberia<=limite2)
table(vector)
long_tuberia<-as.data.frame(long_tuberia)
id<-1:nrow(long_tuberia)
long_tuberia<-cbind(id, long_tuberia)
MaxLongitud<-max(long_tuberia[,2]/1000)
MinLongitud<-min(long_tuberia[,2]/1000)
print(MaxLongitud)
print(MinLongitud)

comunidades<-ap
colnames(comunidades)<-c("pertenencia",    "ID",    "Demand",    "X.Coord",
"Y.Coord")
```

```
#------------------ELEVACION DE COMUNIDADES
nodesdat_elev<-cbind(ap[,1], Elev)
colnames(nodesdat_elev)<-c("L1", "Elev")

elevpromcalc <- function(x){

  member2<-subset(nodesdat_elev, L1 == x, select = c(Elev))

  promedio <- sum(member2)/length(member2$Elev)
  return(promedio)
}

elevacion_promedio <- c()
ID<-(unique(ap[,1]))
for(i in 1:(length(ID)[1])){
  elevacion_promedio[i] <- (elevpromcalc(i))
}

elevacion_promedio<-data.frame(round(elevacion_promedio, 2))
id<-1:nrow(elevacion_promedio)
elevaciones<-cbind(id, elevacion_promedio)
colnames(elevaciones)<-c("id", "elevprom")

for(i in 1:nrow(elevaciones)){
  pipesdatx1$ElevNode1[pipesdatx1$PertNode1    %in%    elevaciones$id[i]]<-
elevaciones$elevprom[i]
}
for(i in 1:nrow(elevaciones)){
  pipesdatx1$ElevNode2[pipesdatx1$PertNode2    %in%    elevaciones$id[i]]<-
elevaciones$elevprom[i]
}

#------------------DEMANDA DE LAS COMUNIDADES

nodesdat_demand<-cbind(ap[,1], Demand)
nodesdat_demand[is.na(nodesdat_demand)] <- 0

colnames(nodesdat_demand)<-c("L1", "Demand")

demandsector<- function(x){

  member2<-subset(nodesdat_demand, L1 == x, select = c(Demand))

  minima <- sum(member2)
  return(minima)
}

demanda_total_sector<-c()
ID<-(unique(ap[,1]))
for(i in 1:(length(ID)[1])){
  demanda_total_sector[i] <- (demandsector(i))
}

demanda_total_sector<-data.frame(round(demanda_total_sector, 2))
id<-1:nrow(demanda_total_sector)
demandasA<-cbind(id, demanda_total_sector)
colnames(demandasA)<-c("id", "demanda")
print(nrow(demandasA))

#caracteristicas_sectores<-data.frame(demandasA, long_tuberia)

for(i in 1:nrow(demandasA)){
 pipesdatx1$DemandNode1[pipesdatx1$PertNode1    %in%    demandasA[,1][i]]<-
demandasA[,2][i]
}
for(i in 1:nrow(demandasA)){
  pipesdatx1$DemandNode2[pipesdatx1$PertNode2    %in%    demandasA[,1][i]]<-
demandasA[,2][i]
```

```
}


#---------------------------Longitud de las comunidades

for(i in 1:nrow(long_tuberia)){
  pipesdatx1$LongNode1[pipesdatx1$PertNode1    %in%    long_tuberia$id[i]]<-
long_tuberia$long_tuberia[i]
}
for(i in 1:nrow(long_tuberia)){
  pipesdatx1$LongNode2[pipesdatx1$PertNode2    %in%    long_tuberia$id[i]]<-
long_tuberia$long_tuberia[i]
}



#-------------------------SEPARARCION DE COMUNIDADES y REFUSION

pipesdatx1$Status<-as.character(pipesdatx1$Status)
pipesdatx1[which(pipesdatx1$PertNode1 != pipesdatx1$PertNode2),"Status"] <-
"Closed"

subset_tuberia_close<-subset(pipesdatx1,      (pipesdatx1$PertNode1      !=
pipesdatx1$PertNode2))

vector_suma <- 1:nrow(subset_tuberia)

if (criteriofusion == "demanda"){
vector_suma<-(subset_tuberia_close$DemandNode1                              +
subset_tuberia_close$DemandNode2)
subset_tuberia_close_caract<-cbind(subset_tuberia_close, vector_suma)
subset_tuberia_close_caract_i<-subset(subset_tuberia_close_caract,
subset_tuberia_close_caract$vector_suma < LimiteDemanda)
nrowcontrol<-nrow(subset_tuberia_close_caract_i)
MaxLongitudCombinacion<-0
x<-1
i<-0
if(nrowcontrol==0)
{
  break
}

MinLongitudCombinacion <- min(subset_tuberia_close_caract_i$vector_suma)
subset_tuberi_x<-subset(subset_tuberia_close_caract_i,
subset_tuberia_close_caract_i$vector_suma == MinLongitudCombinacion)
subset_tuberi_final<-subset_tuberi_x[1,]
}


if (criteriofusion == "longitud"){
  vector_suma<-(subset_tuberia_close$LongNode1                              +
subset_tuberia_close$LongNode2)
  subset_tuberia_close_caract<-cbind(subset_tuberia_close, vector_suma)
  subset_tuberia_close_caract_i<-subset(subset_tuberia_close_caract,
subset_tuberia_close_caract$vector_suma < LimiteLongitud)
  nrowcontrol<-nrow(subset_tuberia_close_caract_i)
  MaxLongitudCombinacion<-0
  x<-1
  i<-0
  if(nrowcontrol==0)
  {
    break
  }

  MinLongitudCombinacion <- min(subset_tuberia_close_caract_i$vector_suma)
```

```
  subset_tuberi_x<-subset(subset_tuberia_close_caract_i,
subset_tuberia_close_caract_i$vector_suma == MinLongitudCombinacion)
  subset_tuberi_final<-subset_tuberi_x[1,]
}

if (criteriofusion == "elevacion"){
  vector_suma<-(abs(subset_tuberia_close$ElevNode1                          -
subset_tuberia_close$ElevNode2))
  subset_tuberia_close_caract<-cbind(subset_tuberia_close, vector_suma)
  subset_tuberia_close_caract_i<-subset(subset_tuberia_close_caract,
subset_tuberia_close_caract$vector_suma < LimiteElevacion)
  nrowcontrol<-nrow(subset_tuberia_close_caract_i)
  MaxLongitudCombinacion<-0
  x<-1
  i<-0
  if(nrowcontrol==0)
  {
    break
  }

  MinLongitudCombinacion <- min(subset_tuberia_close_caract_i$vector_suma)
  subset_tuberi_x<-subset(subset_tuberia_close_caract_i,
subset_tuberia_close_caract_i$vector_suma == MinLongitudCombinacion)
  subset_tuberi_final<-subset_tuberi_x[1,]
}
```

CÓDIGO DE OPTIMIZACIÓN EN VISUAL BASIC

```
Sub modelo_completo2()


Dim hora As Integer
Dim num_horas As Variant
Dim num_nudos As Variant
Dim num_tuberias As Variant
Dim nudo1 As Long
Dim nudo2 As Long
Dim nudo3 As Long
Dim nudo4 As Long
Dim presion1() As Single
Dim presion2() As Single
Dim presion3() As Single
Dim presion4() As Single
Dim rugosidad() As Single
Dim nueva_rugosidad() As Single
Dim nudo As Integer
Dim emitter() As Single
Dim coefi_demanda() As Single
Dim pressure() As Single
Dim tuberia As Integer
Dim qfugas() As Single
Dim linkindes As Long
Dim caudal() As Single
Dim demandab() As Single
Dim consumo() As Single
Dim caudalcal() As Single
Dim suma_presiones As Single


F1 = Cells(1, 2).Value
F2 = Cells(2, 2).Value
ENopen F1, F2, ""

num_horas = 24
```

```
presionreq = Cells(2, 3).Value

ENgetcount EN_NODECOUNT, num_nudos


ENgetcount EN_LINKCOUNT, num_tuberias
candidatas = Cells(11, 3).Value

ReDim emitter(1 To 311)
ReDim pressure(1 To 311)
ReDim coefi_demanda(1 To num_horas)
ReDim rugosidad(1 To 375)
ReDim nueva_rugosidad(1 To 375)
ReDim qfugas(1 To num_horas)
ReDim caudal(1 To num_horas)
ReDim presion1(1 To num_horas)
ReDim presion2(1 To num_horas)
ReDim presion3(1 To num_horas)
ReDim presion4(1 To num_horas)
ReDim demandab(1 To 311)
ReDim consumo(1 To num_horas)
ReDim caudalcal(1 To num_horas)
ReDim nombrescandidatas(1 To candidatas)
ReDim idcandidatas(1 To candidatas)
ReDim statuscandidatas(1 To candidatas)
ReDim presion(1 To num_nudos)
ReDim demand(1 To num_nudos)
ReDim elev(1 To num_nudos)
ReDim ir(1 To num_horas)
ReDim factori(1 To num_horas)
ReDim presionreg(1 To num_horas, 1 To num_nudos)
ReDim press_maxi(1 To num_horas)
ReDim demandabase(1 To num_horas, 1 To num_nudos)
ReDim dif_factores_med(1 To num_horas)
ReDim promedioi(1 To num_horas)
ReDim sd(1 To num_horas)
ReDim emitter(1 To num_horas, 1 To num_nudos)
ReDim presionminima(1 To num_horas)
ReDim costeenergia(1 To num_horas)


ENgetnodeindex "BNA568841", indexsource1

ENgetlinkindex "Pozo_B.Aires", indexpump1

ENgetnodeindex "546398", indexsource2

ENgetlinkindex "Mercedes", indexpump2

ENgetnodeindex "125053", indexsource3

ENgetlinkindex "Pozo_V.Fraterni", indexpump3

For hora = 1 To num_horas
costeenergia(hora) = Cells(40 + hora, 3)
Next

ENopenH
ENinitH 0


For tuberia = 1 To candidatas
nombrescandidatas(tuberia) = Cells(tuberia + 3, 17).Value
ENgetlinkindex nombrescandidatas(tuberia), idcandidatas(tuberia)
statuscandidatas(tuberia) = Cells(tuberia + 3, 19).Value
ENsetlinkvalue idcandidatas(tuberia), EN_STATUS, statuscandidatas(tuberia)
Next
```

```
For hora = 1 To num_horas
ENrunH T
energianudos = 0
energianudosreq = 0
eneraux = 0

ENgetlinkvalue indexpump1, EN_FLOW, flowsource1
ENgetnodevalue indexsource1, EN_HEAD, headsource1
ENgetlinkvalue indexpump2, EN_FLOW, flowsource2
ENgetnodevalue indexsource2, EN_HEAD, headsource2
ENgetlinkvalue indexpump3, EN_FLOW, flowsource3
ENgetnodevalue indexsource3, EN_HEAD, headsource3

energia_fuente = (flowsource1 * headsource1) + (flowsource2 * headsource2)
+ (flowsource3 * headsource3)
eneraux = energia_fuente


For nudo = 1 To num_nudos

ENgetnodevalue nudo, EN_PRESSURE, presionreg(hora, nudo)
ENgetnodevalue nudo, EN_BASEDEMAND, demandabase(hora, nudo)
ENgetnodevalue nudo, EN_EMITTER, emitter(hora, nudo)

If presionreg(hora, nudo) <= 0 Then
presionreg(hora, nudo) = 0
End If
ENgetnodetype nudo, tipoaux
If tipoaux = 1 Then
ENgetnodevalue nudo, EN_HEAD, headtanki
ENgetnodevalue nudo, EN_DEMAND, demandtanki
'eneraux = eneraux + (headtanki * Abs(demandtanki))
ElseIf tipoaux = 0 Then
ENgetnodevalue nudo, EN_DEMAND, basedemandaux
If basedemandaux >= 0 Then
ENgetnodevalue nudo, EN_HEAD, presionaux
ENgetnodevalue nudo, EN_DEMAND, demandaux
ENgetnodevalue nudo, EN_ELEVATION, elevaux
If presionaux < 0 Then
presionaux = 0.01
End If
energianudos = energianudos + (Round(presionaux, 2) * Round(demandaux, 2))
energianudosreq   =   energianudosreq   +   ((elevaux   +   presionreq)   *
Round(demandaux, 2))

End If
End If
Next

ir(hora) = 1 - ((eneraux - energianudos) / (eneraux - energianudosreq))


ENgetlinkvalue indexpump1, EN_ENERGY, energysource1
ENgetlinkvalue indexpump2, EN_ENERGY, energysource2
ENgetlinkvalue indexpump3, EN_ENERGY, energysource3

energy_pump   =   energy_pump   +   ((energysource1   +   energysource2   +
energysource3) * costeenergia(hora))

ENnextH 3600

Next


Cells(26, 13) = Round(energy_pump * 7 * 0.01 * (52 / 1.3), 1)

'punto mas alto de la curva
```

```
For hora = 1 To num_horas
ENgetpatternvalue 1, hora, factori(hora)
Next

factor_aux = factori(1)
For hora = 1 To num_horas
If factor_aux <= factori(hora) Then
factor_aux = factori(hora)
hora_puntoalto = hora
End If
Next

'extraer el punto mas bajo de la curva de demanda
For hora = 1 To num_horas
ENgetpatternvalue 1, hora, factori(hora)
Next
factor_aux = factori(1)
For hora = 1 To num_horas
If factor_aux >= factori(hora) Then
factor_aux = factori(hora)
hora_puntobajo = hora
End If
Next


'extraer el punto medio de la curva de demanda

For hora = 1 To num_horas
suma_factores = suma_factores + factori(hora)
promedio_factores = suma_factores / num_horas
dif_factores_med(hora) = Abs(factori(hora) - promedio_factores)
Next

dif_factores_med_aux = dif_factores_med(1)

For hora = 1 To num_horas
If dif_factores_med(hora) <= dif_factores_med_aux Then
dif_factores_med_aux = dif_factores_med(hora)
hora_puntomedio = hora
End If
Next


ENcloseH


'-------------------------IMPRIMIR EL INDICE DE RESILENCIA----------------

Cells(17, 7) = ir(hora_puntoalto)


'--------------------CALCULAR EL CAUDAL DE FUGAS-----------------------

For hora = 1 To num_horas
For nudo = 1 To num_nudos
caudal_fugas = caudal_fugas + (emitter(hora, nudo) * (presionreg(hora,
nudo)) ^ 0.5)
Next
Next
Cells(15, 3) = caudal_fugas


'--------------------CALCULAR LA DEMANDA A SATISFACER EN CASO DE
DESABASTECIMIENTO-----------------------------------------------

For hora = 1 To num_horas
For nudo = 1 To num_nudos
If presionreg(hora, nudo) < presionreq Then
```

```
ENgetnodevalue nudo, EN_BASEDEMAND, demanda_aux
ENgetpatternvalue 1, hora, coefi_aux
demanda_satis = demanda_satis + (demanda_aux * coefi_aux)
End If
Next
Next
Cells(16, 9) = demanda_satis

'Extraer el valor de presión mínima


presion_min_aux = presionreg(1, 1)
For hora = 1 To num_horas
For nudo = 1 To num_nudos
ENgetnodevalue nudo, EN_BASEDEMAND, demanda_aux
If demanda_aux > 0 Then
If presionreg(hora, nudo) < presion_min_aux Then
presion_min_aux = presionreg(hora, nudo)
End If
End If
Next
Next

Cells(14, 8) = presion_min_aux

'For hora = 1 To num_horas
'presion_min_aux = presionreg(hora, 1)

factor_aux_max = factori(1)
For hora = 1 To num_horas
If factor_aux_max <= factori(hora) Then
factor_aux_max = factori(hora)
hora_crit_max = hora
End If
Next

presion_min_aux = presionreg(hora_crit_max, 1)
For nudo = 1 To num_nudos
ENgetnodevalue nudo, EN_BASEDEMAND, demanda_aux
If demanda_aux > 0 Then
If presionreg(hora_crit_max, nudo) < presion_min_aux Then
presion_min_aux = presionreg(hora_crit_max, nudo)
End If
End If
Next


factor_aux_min = factori(1)
For hora = 1 To num_horas
If factor_aux_min >= factori(hora) Then
factor_aux_min = factori(hora)
hora_crit_min = hora
End If
Next

presion_max_aux = presionreg(hora_crit_max, 1)
For nudo = 1 To num_nudos
ENgetnodevalue nudo, EN_BASEDEMAND, demanda_aux
If demanda_aux > 0 Then
If presionreg(hora_crit_min, nudo) > presion_max_aux Then
presion_max_aux = presionreg(hora_crit_min, nudo)
End If
End If
Next
Cells(13, 8) = presion_max_aux


'------------------------------IMPRIMIR LA PRESION PROMEDIO---------------
```

```
For hora = 1 To num_horas
suma_presiones = 0
nudos = 0
For nudo = 1 To num_nudos
ENgetnodevalue nudo, EN_BASEDEMAND, demanda_aux
If demanda_aux > 0 Then
suma_presiones = suma_presiones + presionreg(hora, nudo)
nudos = nudos + 1
End If
Next
promedioi(hora) = suma_presiones / nudos
Next

'imprimir presion media final
Cells(12, 8) = promedioi(hora_puntomedio)
Cells(30, 13) = promedioi(hora_crit_min)

'----------------------------------IMPRIMIR LA PRESION MAXIMA--

'For hora = 1 To num_horas
'presion_aux = 0
'For nudo = 1 To num_nudos
'If demandabase(hora, nudo) > 0 Then
'If presionreg(hora, nudo) > presion_aux Then
'presion_aux = presionreg(hora, nudo)
'End If
'End If
'Next
'press_maxi(hora) = presion_aux
'Next


'imprimir la presion maxima final

ENclose
End Sub
```