UNIVERSIDAD POLITECNICA DE VALENCIA

Título de Trabajo del Trabajo de Fin de Máster

Propuesta Para una Metodología de Sectorización de Redes de Abastecimiento de Agua Potable

Intensificación:
HIDRÁULICA URBANA

Autor:
ENRIQUE CAMPBELL GONZALEZ
Director/es:
DR. RAFAEL PÉREZ GARCÍA
DR. JOAQUÍN IZQUIERDO SEBASTIÁN
Fecha: **SEPTIEMBRE, 2013**

CONTENIDO

1 RESÚMENES ..... - 8 -

1.1 Resumen ..... - 8 -

1.2 Abstract ..... - 9 -

1.3 Resum ..... - 11 -

2 INTRODUCCION ..... - 14 -

3 OBJETIVOS DE LA TESIS ..... - 20 -

4 ORGANIZACION DEL DOCUMENTO ..... - 21 -

5 ANTECEDENTES ..... - 23 -

5.1 Conceptualizaciones sobre la sectorización y sus propósitos ..... - 23 -

5.2 Historia de la sectorización ..... - 25 -

5.3 Casos de implementación de la sectorización ..... - 27 -

5.3.1 Sectorización de la ciudad de Managua, capital de Nicaragua. ..... - 27 -

5.3.2 Dos casos de sectorización en Perú. ..... - 28 -

5.3.3 Sectorización de la RDAP de Tegucigalpa. Capital de Honduras. ..... - 28 -

5.3.4 Sectorización de San Luis de Rio Colorado, Sonora, México ..... - 29 -

5.4 Diseño de sectores ..... - 30 -

5.4.1 Factores a tener en cuenta en el diseño de sectores. ..... - 30 -

5.4.2 Reglas para establecer sectores ..... - 34 -

5.4.3 Tamaño óptimo de los sectores ..... - 35 -

5.5 Conceptualización de ML y Gestión de fugas ..... - 36 -

5.5.1 Teoría de Grafos ..... - 37 -

5.5.2 Teoría de Formación de Clústeres ..... - 44 -

5.5.3 Conceptualizaciones sobre Machine Learning (*ML*). ..... - 46 -

5.5.4 R studio ..... - 79 -

5.5.5 Aplicación de AHP para ponderar aspectos hidráulicos ..... - 80 -

5.5.6 Fugas en las RDAP ..... - 83 -

5.5.7 Energía en las RDAP ..... - 88 -

5.6 Investigaciones Previas ..... - 92 -

5.6.1 Verificación de planos de sectorización. ..... - 93 -

5.6.2 Herramientas de ayuda a la sectorización. ..... - 93 -

5.6.3 Sectorización de redes existentes. ..... - 94 -

5.6.4 Partición de grafo para sectorización automática. ..... - 95 -
5.6.5 Sectorización de sistemas de abastecimiento de agua. ..... - 95 -
5.6.6 Metodología heurística de diseño de distritos métricos. ..... - 95 -
5.6.7 Sectorización de redes de agua basadas en algoritmos genéticos. ..... - 96 -
5.6.8 Definición de clústeres en RDAP. ..... - 97 -
6 METODOLOGIA PROPUESTA..... - 98 -
6.1 Descripción de la Metodología ..... - 98 -
6.2 Ventajas comparativas de la metodología propuesta ..... - 100 -
6.3 Implementación de la Metodología ..... - 103 -
6.3.1 Selección del número de sectores ..... - 104 -
6.3.2 Subdivisión de la red en sectores ..... - 108 -
6.3.3 Ubicación de Unidades Operativas de Control (UOC). ..... - 113 -
6.3.4 Evaluación de propuesta de sectorización ..... - 118 -
6.3.5 Mejora de la eficiencia energética de la RDAP sectorizada ..... - 120 -
7 CONCLUSIONES Y LINEAS FUTURAS ..... - 122 -
7.1 Conclusiones ..... - 122 -
7.2 Líneas Futuras ..... - 123 -
8 REFERENCIAS..... - 126 -
8.1 Referencias Propias ..... - 126 -
8.2 Referencias ..... - 126 -
9 ANEXO I: Fracción del código Implementado..... - 135 -

INDICE DE TABLAS

Tabla 1: Matriz de adyacencia (W_{ij}) del grafo - 40 -
*Tabla 2: Matriz de afinidad (**Aij**) del grafo* - 42 -
*Tabla 3: Matriz laplaciana (**Lij**) del grafo* - 43 -
Tabla 4: Datos de ejemplo de clústering jerárquico - 65 -
Tabla 5: Matriz de disimilaridad de ejemplo de clústering jerárquico - 66 -
Tabla 6: Matriz de aglomeración actualizada - 66 -
Tabla 7: Matriz ultramétrica - 67 -
Tabla 8: Ventaja y desventaja de los indicadores para evaluación de particiones en un dendrograma. - 79 -
Tabla 9: Paquetes de R empleados - 80 -
Tabla 10: La escala absoluta de números absolutos (Saaty, 2008) - 82 -
Tabla 11: Evaluación del CPCC - 105 -
Tabla 12: Comparación de criterios a tomar en cuenta en la partición - 108 -
Tabla 13: Cálculo de CPP e I_r-Sector 1 - 115 -
Tabla 14: Cálculo de CPP e I_r-Sector 2 - 116 -
Tabla 15: Cálculo de CPP e I_r-Sector 3 - 117 -
Tabla 16: Valores de presión máximos y mínimos en los sectores antes y después de la sectorización - 120 -
Tabla 17: Índices de energía para dos propuestas de sectorización - 121 -

INDICE DE FIGURAS

Ilustración 1: Punto óptimo de control de fugas (Farley et al., 2008) (Pearson y Trow, 2008) - 31 -
Ilustración 2: Red ejemplo - 40 -
Ilustración 3: Grafo ponderado - 41 -
Ilustración 4: Clasificación de datos mediante SVM (Cristianini y Shawe-Taylor, 2000). - 50 -
Ilustración 5: Etapas de clústering espectral (Herrera, 2011a) - 56 -
Ilustración 6: Ejemplo de aplicación de clústering jerárquico - 65 -
Ilustración 7: Valores de altura en el dendrograma - 67 -
Ilustración 8: Optimización del número de clústeres con base en el índice de conectividad - 70 -
Ilustración 9: Optimización del número de clústeres con base en el ancho de silueta - 71 -
Ilustración 10: Ancho de silueta para una partición de dos clústeres - 72 -
Ilustración 11: Ancho de silueta para una partición de tres clústeres - 72 -
Ilustración 12: Ancho de silueta para una partición de cuatro clústeres - 72 -
Ilustración 13: Optimización del número de clústeres con base en el índice de Dunn - 73 -
Ilustración 14: Selección del número de clústeres con base en el criterio del codo - 74 -
Ilustración 15: Valores de inconsistencia correspondiente a cada clada del dendrograma - 76 -
Ilustración 16: Definición de sectores válidos mediante p-values - 77 -
Ilustración 17: Sectores válidos definidos mediante p-values - 78 -
Ilustración 18: Los cuatro pilares de la gestión de fugas (Pilcher et al., 2007) - 85 -
Ilustración 19: Red de Estudio - 104 -
Ilustración 20: Balance de caudales en red ejemplo - 104 -
Ilustración 21: CPCC obtenidos - 105 -

Ilustración 22: Dendrograma de la red obtenido _____ *- 105 -*
Ilustración 23: Curva de altura vs número de clústeres _____ *- 106 -*
Ilustración 24: Evaluación de medidas internas _____ *- 107 -*
Ilustración 25: Ancho de silueta para una partición en tres sectores _____ *- 107 -*
Ilustración 26: p-values AU/BP obtenidos por remuestreo multiescala a partir de los mismos datos _____ *- 108 -*
Ilustración 27: Gráfica para determinar el valor alfa de la suma de matrices kernel __ *- 109 -*
Ilustración 28: Clústeres en la red ejemplo _____ *- 110 -*
Ilustración 29: Partición de la red ejemplo _____ *- 110 -*
Ilustración 30: Red con un valor de coeficiente de emisor elevado en algunos nodos ___ *- 111 -*
Ilustración 31: p-values AU/BP obtenidos por remuestreo multiescala a partir de los mismos datos _____ *- 111 -*
Ilustración 32: Evaluación de medidas internas _____ *- 112 -*
Ilustración 33: Nueva partición de la red usando pesos del ejemplo anterior _____ *- 112 -*
Ilustración 34: Clústeres en la nueva partición _____ *- 113 -*
Ilustración 35: Nueva partición de la red dando importancia máxima al coeficiente de emisor _____ *- 113 -*
Ilustración 36: Esquema de sectorización antes de la selección de UOC _____ *- 114 -*
Ilustración 37: Líneas candidatas para UOC del sector 1 _____ *- 114 -*
Ilustración 38: Líneas candidatas para UOC del sector 2 _____ *- 115 -*
Ilustración 39: Líneas candidatas para UOC del sector 3 _____ *- 116 -*
Ilustración 40: Propuesta de sectorización final _____ *- 117 -*
Ilustración 41: MDT de presiones para las 18:00 horas. Periodo de mayor consumo ___ *- 118 -*
Ilustración 42: MDT de presiones para las 00:00 horas. Periodo de menor consumo ___ *- 118 -*
Ilustración 43: Curva de presión pre y post sectorización en el nodo crítico del sector 1 - 119 -
Ilustración 44: Curva de presión pre y post sectorización en el nodo crítico del sector 2 - 119 -
Ilustración 45: Curva de presión pre y post sectorización en el nodo crítico del sector 3 - 119 -
Ilustración 46: Curvas de consumo y caudales de fugas antes y después de la sectorización _ *- 120 -*
Ilustración 47: Modificación de la primera propuesta de sectorización. Original (izquierda), modificada (derecha). _____ *- 121 -*

GLOSARIO DE TERMINOS

AHP:	*Analytic hierarchy process (proceso de análisis jerárquico)*
ANR:	*Agua no registrada*
ANC:	*Agua no contabilizada*
AU:	*Aproximadamente insesgado (approximately unbiased)*
AVSA:	*Aguas de Valencia S.A*
BA:	*Búsqueda en amplitud (Breadth first search)*
BP:	*Búsqueda en profundidad (Depht first search)*
BPr	*Probabilidad de remuestreo (Bootstrap probability)*
CAF:	*Control activo de fugas*
CI:	*Coste de inspección*
CMA:	*Coste medio del agua*
CPCC:	*Coeficiente de correlación cofenética (Cophenetic correlation coefficient)*
CPP:	*Cociente de pérdida de potencia*
CU:	*Coeficiente de uniformidad*
CVIA:	*Centro Virtual de Información del Agua*
DMA:	*District metering area (Distrito Métrico)*
EML:	*Máxima probabilidad de igual varianza (Equal-Variance Maximun Likehood)*
EPA:	*Agencia de Protección Ambiental (Environmental Protection Agency)*
FAVAD:	*Area fija y variable de descarga (Fixed and variable area discharged paths)*
FOI:	*Frecuencia óptima de inspección*
GIZ:	*Sociedad Alemana de Cooperación Internacional (Deutsche Gesellschaft für Internationale Zusammenarbeit GmbH*)
H:	*Carga hidráulica o altura piezométrica*
IA:	*Inconsistencia aleatoria*
IIAMA:	*Instituto del Agua y Medio Ambiente*
IIC:	*Índice de inconsistencia*
IMTA:	*Instituto Mexicano de Tecnología del Agua*
INAF:	*Índice natural de aumento de fugas*
IWA:	*Asociación Internacional de Agua (International Water Association)*
I_r:	*Índice de resiliencia*
I_{rd}:	*Desviación del índice de resiliencia*
m^3/h:	*metros cúbicos por hora*
MA:	*Multiagente (Multi-agent)*
MAS:	*Sistema multiagente (Multi-agent system)*

mca:	*metros de columna de agua*
ML:	*métodos automáticos de aprendizaje computacional (Machine Learning)*
MLRB:	*Bisección recursiva multinivel (Multilevel Recursive Bisection)*
P:	*presión*
PMA:	*Área de gestión de presión (Pressure management area)*
PNUMA:	*Programa de Naciones Unidas para Medio Ambiente*
RDAP*:*	*Red de abastecimiento de agua potable*
Q:	*Caudal*
RI:	*Ratio de inconsistencia*
SANAA:	*Servicio Autónomo Nacional de Abastecimiento, Acueductos y Alcantarillados*
SAWUADB:	*The Southeast Asian Water Utilities Network and Asian Developing Bank*
SIG:	*Sistema de información geográfica*
SVM:	*Máquina de vectores de soporte (Support vector machine)*
UMF:	*Umbral mínimo de fugas*
UNU:	*Universidad de las Naciones Unidas (United Nations University)*
UOC:	*Unidad operativa de control*
UNWDCP:	*Programa Década del Agua de las Naciones Unidas (United Nations Water Decade Program)*
WSN:	*Redes de abastecimiento de agua potable (Water supply networks)*
WWC:	*Consejo Mundial del Agua (World Water Council)*
XDAP:	*Redes de abastecimiento de agua potable (Xarxes de Distribució d'Aigua Potable)*

1 RESÚMENES

1.1 Resumen

El agua potable es un recurso indispensable para todo proceso relacionado con la vida. Los problemas de las Redes de Abastecimiento de Agua Potable (RDAP) pueden resumirse en cuatro aspectos generales: fugas y agua no contabilizada; integridad física de la red; calidad de agua a distribuir; fiabilidad y calidad de la base de datos de los sistemas de distribución de agua. En particular, las fugas pueden llegar a ser del orden del 50% del agua que es inyectada a una RDAP. Hoy por hoy, redes sin ningún tipo de pérdida se considerarían una utopía, tanto por las implicaciones técnicas como económicas que esto representaría; 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. Dentro de estas técnicas se destaca la sectorización, que es considerada como una opción estratégica que implica la subdivisión de las redes en pequeñas subredes mediante el cierre de válvulas e instalación de caudalímetros. 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 subred debido a la reducción de su tamaño. En la mayor parte de los casos en los que se ejecuta un proyecto de sectorización, no se suele seguir el proceso con un rigor científico-técnico y por el contrario, suele basarse en aproximaciones de prueba y error. En redes de menor tamaño, este tipo de aproximación, no necesariamente tiene que representar un gran problema. El problema es la definición apropiada de sectores en redes de mayor tamaño, en donde, dada la gran cantidad de información asociada a las mismas, sería inviable la ejecución de un proceso de esta naturaleza sin el apoyo de herramientas informáticas. El objetivo principal de la tesis es establecer un procedimiento informático para la obtención de un plano de red sectorizada, que divida una RDAP en una red de alta y una red de distribución y que, además, no sólo emplee como criterios las características hidráulicas tradicionales o geográficos de las RDAP, sino

que también tenga en cuenta las fugas en la red. Para esta tarea, se sigue un proceso que implica: (1). Definir dentro de la RDAP de estudio las líneas que constituyen la red primaria y las líneas que constituyen la red de distribución. Este paso se efectuará en función del diámetro de las conducciones, estableciendo un rango de diámetro mínimo para distinguir entre red de alta y red de distribución; (2). Estimar el número de sectores que debe tener la red de distribución en cuestión, mediante la aplicación de una técnica informática de análisis de clústeres; (3). Establecer la distribución del número de microsectores calculados en el paso anterior, mediante *clústering espectral*; (4). Definir las entradas a los sectores con base en criterios energéticos; (5). Validar el esquema de sectorización resultante mediante modelización matemática. Con la segregación de la red de distribución y la red de alta, se conserva la flexibilidad de la red en el caso de que en el futuro se requiera variar el esquema de sectorización seleccionado y también se ahorran costes por la instalación de válvulas y caudalímetros de gran diámetro. En este trabajo se demuestra la aplicabilidad de Aprendizaje Automático Computacional (ML, *machine learning*) para abordar la tarea propuesta. Mediante clústering jerárquico, se logró estimar un número de sectores en los que se conserve el mejor grado de homogeneidad posible de las características de los sectores, lo que luego facilita que la red presente un buen rendimiento energético. Mediante el proceso de clústering espectral se logran mejorar los resultados de clústering jerárquico, encontrando una partición que además de mantener la homogeneidad de las características de cada uno de los sectores, minimice el número de válvulas que se deben emplear para hacer la partición. El empleo de indicadores de disipación de energía a través de la red, ha permitido encontrar las entradas a cada sector de manera tal que se minimice la energía disipada por la red y se garantice la mayor presión posible en los nodos de consumo. Con esta propuesta, se logró obtener un plano de red sectorizada (100 km de red, división en tres sectores) que mantiene la presión dentro de los rangos establecidos como apropiados y que a su vez conduce a una disminución del nivel de fugas tan sólo por implementarla.

1.2 Abstract

Drinking water is an indispensable resource for all processes involved in life. The problems of water supply networks (WSN) can be summarized in four broad areas: leakage and non-revenue water; physical integrity of the network; water quality and reliability, and quality of the databases associated to water distribution systems. In particular, leakage may become of the order of 50% of water injected in a WSN. Today, networks without any water losses would be considered utopian, given the technical and economic implications that this would imply. However, there have been important advances in the understanding and development of equipment and techniques that allow better comprehension and management of leakages. Among these techniques, it is important to highlight the segmentation, which is considered a strategic option which involves the subdivision of small sub networks by closing valves and installing flow meters. One of the major benefits of its implementation is the increased ease at which any abnormality is detected within the subsector due to the dimensional reduction. In most cases, sectorization processes are implemented with great lack of scientific rigor. On the contrary, sectorization is usually based on trial-and-error procedures. In smaller networks, this type of approach is usually straightforward. The problem is the proper definition of sectors in larger networks, where, given the large amount of information associated with the WSN, the implementation of a process of this nature, without the support of tools, would not be feasible. The main objective of the thesis is to establish a computational procedure for obtaining a layout of sectorization of a WSN. In this layout, the network is split into trunk and distribution network. The process not only uses traditional hydraulic or geographical features of RDAP as criteria, but also takes into consideration the leaks into the network. The process involves the next steps: (1). Define within the WSN the trunk and the lines that form the distribution network. This step is completed based on the diameter of the pipes, setting a minimum diameter range to distinguish between trunk and distribution network; (2). Estimate the number of sectors that should have the distribution network in question, by means of a computer technique of cluster analysis; (3). Define the number of microsectors calculated in the previous step, using spectral clustering; (4). Define the entrance (of water) to the sectors based on energy criteria; (5). Validation of the sectorization obtained (through mathematical modeling). Segregation of the distribution network and the trunk helps to conserve the flexibility of the network in

the event that, in the future, the sectoring scheme changes, and also to save costs by installing valves and flow meters of large diameter. This work demonstrates the applicability of Machine Learning to tackle the proposed task. By using hierarchical clustering a number of sectors with the highest degree of homogeneity regarding their characteristics are found. Then spectral clustering is able to improve the results of hierarchical clustering, finding a partition while maintaining consistency of the characteristics of each of the sectors, and minimizing the number of valves to be used to make the partition. The use of indicators of energy dissipation through the network enables to find water entrances to each sector so as to minimize the energy dissipated by the network, and ensures the highest possible pressure at the consumers' nodes. Following this proposal, It was managed to get an example network sectioned (100 km network divided into three sectors) that maintains the pressure within the ranges established as appropriated, and decreases the level of leakage just by its implementation.

1.3 Resum

L'aigua potable és un recurs indispensable per a tot procés relacionat amb la vida. Els problemes de les XDAP poden resumir-se en quatre aspectes generals: fugues i aigua no comptabilitzada; integritat física de la xarxa; qualitat d'aigua a distribuir; fiabilitat i qualitat de la base de dades dels sistemes de distribució d'aigua. En particular, les fugues poden arribar a ser de l'orde del 50% de l'aigua que és injectada a una XDAP. Ara per ara, les xarxes sense cap tipus de pèrdua es considerarien una utopia, tant per les implicacions tècniques com econòmiques que açò representaria; no obstant això, hi ha hagut un gran avanç en el coneixement i desenvolupament d'equips i tècniques, que

permeten fer un seguiment més exhaustiu de les fugues en elles. Dins d'estes tècniques es destaca la sectorització, que és considerada com una opció estratègica que implica la subdivisió de les xarxes en xicotetes subxarxes mitjançant el tancament de vàlvules i instal·lació de cabalímetres. Com un dels grans beneficis de la seua implementació es destaca l'augment de la facilitat amb què es detecta qualsevol anormalitat dins de la subxarxa a causa de la reducció de la seua dimensió. En la major part dels casos en què s'executa un projecte de sectorització, no se sol seguir el procés amb un rigor científic–tècnic i al contrari, sol basar-se en aproximacions de prova i error. En xarxes més xicotetes, aquest tipus d'aproximació, no necessàriament ha de representar un gran problema. El problema és la definició apropiada de sectors en xarxes més grans, on, donada la gran quantitat d'informació associada a les mateixes, seria inviable l'execució d'un procés d'esta naturalesa sense el suport de ferramentes informàtiques. L'objectiu principal de la tesis és establir un procediment informàtic per a l'obtenció d'un pla de xarxa sectoritzada, que dividisca una XDAP en una xarxa d'alta i una xarxa de distribució i que a més, no sols empre com a criteris les característiques hidràuliques tradicionals o geogràfiques de les XDAP, sinó que també tinga en compte les fugues en la xarxa. Per a esta tasca, se seguix un procés que implica: (1). Definir dins de la XDAP d'estudi les línies que constituïxen la xarxa primària i les línies que constituïxen la xarxa de distribució. Este pas s'efectuarà en funció del diàmetre de les conduccions, establint un rang de diàmetre mínim per a distingir entre xarxa d'alta i xarxa de distribució; (2). Estimar el nombre de sectors que ha de tindre la xarxa de distribució en qüestió, mitjançant l'aplicació d'una tècnica informàtica d'anàlisi de clusters; (3). Establir la distribució del nombre de microsectors calculats en el pas anterior, mitjançant *clústerin espectral*; (4). Definir les entrades als sectors amb base en criteris energètics; (5). Validar l'esquema de sectorització resultant mitjançant la modelització matemàtica. Amb la segregació de la xarxa de distribució i la xarxa d'alta, es conserva la flexibilitat de la xarxa en cas que en el futur es requerisca variar l'esquema de sectorització seleccionat i també s'estalvien costos per la instal·lació de vàlvules i cabalímetres de gran diàmetre. En este treball, es demostra l'aplicabilitat de *ML* per a abordar la tasca proposada. Mitjançant el clústerin jeràrquic es va aconseguir estimar un nombre de sectors en què es conserve el millor grau d'homogeneïtat possible de les característiques dels sectors, la qual cosa després facilita que la xarxa presente un bon rendiment energètic. Mitjançant el procés de clústerin espectral s'aconseguixen millorar els

resultats de clústerin jeràrquic, trobant una partició que a més de mantindre l'homogeneïtat de les característiques de cada un dels sectors, minimitze el nombre de vàlvules que s'han d'emprar per a fer la partició. L'ocupació d'indicadors de dissipació d'energia a través de la xarxa, ha permés trobar les entrades a cada sector de tal manera que es minimitze l'energia dissipada per la xarxa i sí garantisca la major pressió possible en els nodes de consum. Amb esta proposta, es va aconseguir obtindre un pla de xarxa sectoritzada (100 km de xarxa, divisió en tres sectors) que manté la pressió dins dels rangs establits com apropiats i que al seu torn comporta a una disminució del nivell de fugues tan sols per implementar-la.

2 INTRODUCCION

El agua potable es un recurso indispensable para todo proceso relacionado con la vida, es un producto primario tanto para la actividad doméstica, así 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 (RDAP). Estas son las infraestructuras que permiten transportar el recurso en cuestión desde las fuentes hasta los consumidores; es decir, a través de ellas se da el proceso de abastecimiento de agua potable. En tal sentido, es importante hacer notar la relación directa que existe entre la calidad del servicio de abastecimiento de agua potable de la que dispone cualquier ciudad (o localidad) y su grado de desarrollo y modernidad.

Desde un punto de vista técnico y dando por supuesto que se hace una apropiada gestión administrativa, los problemas de las RDAP pueden resumirse en cuatro aspectos generales: fugas y agua no contabilizada; integridad física de la red; calidad de agua a distribuir; 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 de agua ha sido una actividad asociada a los sistemas de distribución de agua desde que se construyeron las primeras RDAP. Incluso, en la antigua Roma ya existía conciencia de que una buena parte del agua que era inyectada a los sistemas de distribución no llegaba a los usuarios (Pilcher *et al*., 2007).

La causa principal para la existencia de agua no registrada (ANR) en una RDAP es el sub-registro de los contadores y las fugas en las tuberías, acometidas y unidades de almacenamiento (pérdidas reales). En países en vías de desarrollo, el segundo tipo de pérdidas puede representar más del 50% del agua inyectada a la red (Kingdom *et al*., 2006), (SAWUADB, 2007). En una RDAP con pérdidas del 50%, 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 m³, lo que representa 5.9 miles de millones de dólares norteamericanos. Con tan sólo reducir este valor por 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, lo que permitiría el acceso a agua potable a 130 millones de personas, implicando, en adición, un flujo de caja propio de 4 mil millones de dólares norteamericanos para los operadores de agua.

En países desarrollados la situación es distinta. El porcentaje de pérdida no suele superar el 15% (Kingdom *et al*., 2006); no obstante, las previsiones no indican una mejora. 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 será 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% (Thornton *et al*., 2008).

En países en vías de desarrollo es muy común que las empresas encargadas del suministro de agua potable busquen, 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 de baja calidad (Kingdom *et al*., 2006). Tal situación, se agrava más si se tiene en cuenta que a la hora de valorar el coste del agua perdida, sólo se tiene en cuenta 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. 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, a los edificios residenciales o a una calzada.

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.

Hoy por hoy, redes sin ningún tipo 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.
- 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, que es considerada como una opción estratégica, implica la subdivisión de la red en subredes con una entrada de agua controlada. En cada segmento de subdivisión se maneja un valor máximo de demanda y dentro de ellos se trata de mantener una homogeneidad en lo que a elevación de terreno 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 de su tamaño (Herrera, 2011a), (Morrison *et al*., 2007), (CVIA, 2010). Contar con una red sectorizada permite no sólo aplicar técnicas particulares de control de fugas, sino además permite implementar modelos de gestión diversos.

Este concepto fue introducido a principios de la década de 1980 en el reporte 26 de Control de Pérdidas y Prácticas de las Asociación de Autoridades del Agua de Inglaterra (Morrison *et al*., 2007). En la actualidad es una técnica empleada en muchos países alrededor del mundo, siendo más popular en Europa y América Latina.

En la mayor parte de los casos en los que se ejecuta un proyecto de sectorización, no se suele seguir el proceso con un rigor científico-técnico y por el contrario,

suele basarse en aproximaciones de prueba y error (Di Nardo *et al*., 2013b). En redes de menor tamaño, este tipo de aproximación no necesariamente tiene que representar un gran obstáculo. El problema es la definición de sectores en redes de mayor tamaño, en donde, dada la gran cantidad de información asociada a las mismas, sería inviable la ejecución de un proceso de esta naturaleza sin el apoyo de herramientas informáticas (Izquierdo *et al*., 2011). En la última década se han desarrollado una serie de trabajos científicos que le han dado a las técnicas de sectorización un nuevo enfoque metodológico. Estos han permitido incluir una visión de optimización dentro del proceso, ya no sólo para implementarla como herramienta de control de ANR, sino también como herramienta para una gestión sostenible de las RDAP. En ellos se ha unido la teoría de grafos con algoritmos de análisis de calidad de una sustancia no conservativa en RDAP para establecer subredes aisladas (Tzatchkov *et al*., 2008); también se ha planteado un algoritmo que divide la red en una red de alta y una red secundaria. Dentro de la red secundaria identifica macrosectores, y dentro de esos macrosectores hace la división de microsectores mediante un algoritmo de construcción del árbol dirigido de mínimo coste (Vegas, 2012). Se ha hecho uso de técnicas Multi-Agente (MA) para crear clústeres (o sectores) dentro de la red teniendo como criterio la homogeneidad de las cotas del terreno y una demanda máxima dentro de cada clúster (Izquierdo *et al*., 2011); también se han empleado técnicas de formación de clústeres para crear sectores-aislados (sectores con una fuente de agua exclusiva), mediante la implementación de Aprendizaje Automático Computacional (ML, *Machine Learning*), que permiten incluir múltiples criterios al momento de efectuar una partición de una RDAP en sectores (Herrera, 2011a). Igualmente se han empleado índices de eficiencia energética como criterio de definición de sectores (Di Nardo *et al*., 2013b), (Di Nardo *et al*., 2013a).

Con excepción del trabajo presentado por Vegas (2012), todos los trabajos anteriores plantean aislar zonas de la RDAP sin considerar una separación de la red primaria del resto de la red. La ventaja de no incluir las líneas de la red primaria (o red de conducción, red de alta o red principal) dentro de la sectorización radica en la conservación del nivel de flexibilidad de la misma, permitiendo modificar el esquema de sectorización si las circunstancias futuras o

circunstancias temporales lo ameritan. De igual manera, al no incluir la red primaria se reducen costes en la compra de caudalímetros o válvulas de gran diámetro (Morrison *et al*., 2007). También se debe tener en cuenta que en algunas RDAP, la partición en sectores aislados con una fuente exclusiva no es un concepto tan sencillo de aplicar, ya que las fuentes no se encuentran distribuidas dentro de la red misma, sino por el contrario en algunos casos se encuentran fuera de la ciudad.

A nivel normativo, no existen directrices claras respecto a los pasos a seguir para llevar a cabo un proceso de sectorización. Los criterios que se recomiendan en la actualidad se basan en un número de acometidas o longitud de red, sin tener en cuenta el contexto en el que se lleva a cabo el proceso. Esto trae como consecuencia el hecho de que en muchos casos su implementación se ejecute de manera excesivamente empírica, con las consecuencias negativas que eso puede acarrear, tales como: desabastecimiento, pérdidas de carga innecesarias y disminución de la calidad del agua. En este sentido, es importante destacar que la sectorización cambia el comportamiento hidráulico de cualquier RDAP, dado que en principio, está en conflicto con el criterio tradicional de diseño de redes malladas, que permite a las RDAP ser más confiables bajo condiciones de fallo mecánico e hidráulico (Mays, 2000), (Di Nardo y Di Natale, 2011a). Es por eso que su implementación debe apoyarse en un análisis apropiado.

De lo anteriormente expuesto nace la motivación de este trabajo. La idea que subyace, es proponer una metodología de sectorización con una base científico-técnica, para RDAP con las fuentes de abastecimiento no distribuidas dentro de ellas (con una red de alta de gran extensión), haciendo uso de *ML*.

El procedimiento propuesto se puede resumir en los siguientes pasos:

- ***Discriminación de la red primaria***: En primer lugar se identificará la red primaria dentro de la RDAP, empleando como criterio el diámetro de las tuberías, y la demanda. Una vez identificada la red de alta, esta se eliminará, dejando únicamente la red de distribución.

- ***Determinación del número de sectores***: Teniendo en cuenta las características de la RDAP, se procede a realizar una análisis de los datos mediante la técnica de formación de clústeres jerárquicos, y en función de algunos indicadores, se seleccionará el número de sectores en los que se hará la subdivisión de la red. Lo que se persigue con esta técnica es obtener sectores de modo que, sus características internas sean lo más homogéneas posible.

- ***Subdivisión de la RDAP en subsectores:*** A continuación se procederá a la aplicación de *clústering espectral* para definir el área de cobertura de cada uno de los sectores tal como plantea Herrera (2011a). A esta metodología se agrega un componente de fugas dentro de lo criterios efectuados para hacer la partición.

- ***Selección de las entradas de cada sector:*** Para seleccionar la entrada que abastecerá cada sector, se comparan los aspectos energéticos implicados en cada alternativa.

- ***Validación de la propuesta:*** Finalmente, las alternativas de sectorización planteadas se validan hidráulicamente, comprobando que los parámetros hidráulicos (presión y velocidad) se encuentre dentro de rangos aceptables.

Este trabajo persigue ser un aporte al desarrollo de técnicas de sectorización, que pese a estar siendo implementada en muchos países alrededor del mundo, sigue careciendo de criterios que ayuden a evitar las consecuencias excesivamente negativas que suelen darse como producto de su implementación y que además permitan optimizar los beneficios que puede ofrecer la misma en relación a la gestión de las RDAP.

3 OBJETIVOS DE LA TESIS

El objetivo principal de la tesis es establecer un procedimiento informático para la obtención de un plano de red sectorizada, que divida una red de distribución de agua potable (RDAP) en una red de alta y una red de distribución y que además, no sólo emplee como criterios los hidráulicos o geográficos, sino que también tenga en cuenta las fugas en la red. A fin alcanzar tal objetivo, se plantean la siguiente serie de objetivos específicos:

- Definir dentro de la RDAP de estudio las líneas que constituyen la red primaria y las líneas que constituyen la red de distribución. Este paso se efectuará en función del diámetro de las conducciones, estableciendo un rango de diámetro mínimo para distinguir entre red de alta y red de distribución.
- Estimar el número de sectores que debe tener la red de distribución en cuestión, mediante la aplicación de una técnica informática de análisis de clústeres.
- Establecer la distribución del número de microsectores calculados en el paso anterior, mediante *clústering espectral*, siguiendo la metodología propuesta por Herrera (2011a), pero adicionando como criterio, la distribución de fugas en la RDAP.
- Definir las entradas a los sectores con base a criterios energéticos.
  Validar el esquema de sectorización resultante mediante modelización matemática.

4 ORGANIZACION DEL DOCUMENTO

El presente trabajo está organizado de la siguiente manera:

Capítulo 5: *Antecedentes*

Se realiza un análisis del concepto sectorización y sus implicaciones. Se describen algunos ejemplos de sectorización en algunas ciudades y, además, se hace una descripción del estado del arte de la temática. Se incluyen la descripción de algunas guías metodológicas para el establecimiento y manejo de sectores. A continuación se presenta una descripción detallada de conceptos relativos a aprendizaje automático computacional (*ML, machine learning)*, tales como: clústering jerárquico, métodos de aprendizaje: *supervisados*, *no supervisados* y *semisupervisado*, y métodos *kernel* para el análisis de patrones. También se hace una explicación de algunos aspectos de la teoría de grafos. Posteriormente, se hace una breve descripción del Proceso de Análisis Jerárquico o Analytic Hierarchy Process (AHP), y finalmente se termina haciendo mención del lenguaje de programación R, y sus capacidades para abordar el problema aquí plantado.

Finalmente se describen algunos métodos ya existentes para subdividir RDAP utilizando procedimientos informáticos.

Capítulo 6: *El método propuesto*

En este capítulo se explica con detalle el método propuesto para generar sectores: se aborda en primer lugar el proceso de identificación de la red de alta. A continuación se detallan los pasos de la utilización de un método de aprendizaje automático *no supervisado* para seleccionar el número de sectores en el que será subdivida la red. Definido el número de sectores, se aplica un método de aprendizaje *semisupervisado* para formar clústeres aplicando métodos *kernel* para análisis de datos, y métodos espectrales de formación de clústeres. Se seleccionan

las entradas a cada uno de los subsectores resultantes mediante una comparación entre las suma de potencias en sus nodos una vez hecha la sectorización con la potencia antes de hacer la sectorización. También se hace una comparación de las potencias operativas de la red antes y después de la sectorización.

Se implementa el método en una red con 100 km de longitud de tubería. En este ejemplo se puede apreciar el efecto de incluir criterios de fugas dentro de la partición.

Capítulo 7: *Conclusiones y Líneas Futuras*

Se plantean las conclusiones de la aplicación del método y se hacen las recomendaciones pertinentes.

Capítulo 9: *Referencias bibliográficas*

Se presentan las referencias bibliográficas empleadas en esta investigación y las referencias propias de este documento

Capítulo 10: *Anexo*

Se presentan partes del código en R que se ha implementado en el trabajo.

5 ANTECEDENTES

5.1 Conceptualizaciones sobre la sectorización y sus propósitos

En términos generales, la sectorización de las redes de abastecimiento de agua potable (RDAP) puede ser considerada como el procedimiento encaminado a establecer dentro de las mismas, *subáreas* con una alimentación controlada (que puede ser exclusiva del sector o compartida por varios sectores al mismo tiempo). Tal procedimiento puede perseguir objetivos que van desde el Control Activo de Fugas (CAF) hasta el control de la calidad del agua (Farley *et al*., 2008). En cualquier caso, el contar con una red sectorizada, permite detectar con mayor facilidad cualquier anomalía que ocurra en un punto de la red, debido a la reducción dimensional implícita en la sectorización misma. En AVSA (2013), se establece que el objetivo principal para la creación de sectores es obtener la información necesaria distribuida y manejablemente escalada para llevar acciones claves en cada sector, tales como:

- Realizar auditorías para conocer el rendimiento hidráulico o el Agua No Contabilizada (ANC).
- Caracterizar la curva de demanda, especialmente el caudal nocturno.
- Detectar de la manera más rápida posibles fugas mediante el análisis de la evolución de los caudales mínimos nocturnos.
- Comprobar rápidamente los resultados de campañas rápidas de detección y reparación de fugas.
- Detectar el fraude, sub-registro, y diversos errores de medición.
- Disminuir los costes de mantenimiento.
- Establecer plan de inversiones para abastecer sectores con mayor índice de ANC.

La sectorización podría ser considerada también un primer paso para contrarrestar las situaciones de suministro intermitente, en vista que facilitan la detección y reparación de las fugas más importantes (GIZ *et al*., 2011).

En la literatura relativa al tema se encuentran varias definiciones para el término "sector". GIZ *et al*. (2011) define un sector como un DMA, que es un área discreta de una red de abastecimiento la cual se puede aislar del resto de la red ya sea mediante válvulas seccionadoras o mediante tuberías cortadas. Respecto a este concepto, es interesante destacar que para aislar un sector, además de colocar válvulas seccionadoras y realizar cortes de tuberías, se pueden emplear tuberías nuevas que permitan redistribuir el caudal (Izquierdo *et al*., 2011). Herrera (2011b), en lugar del término sector o DMA, emplea el concepto de clúster, siendo la creación de clústeres el proceso de agrupación o segmentación de objetos de una red en *subgrupos,* de manera tal que los objetos dentro de un clúster estén más cercanamente relacionados que los objetos que se encuentran fuera de él. En el caso de una RDAP, un clúster sería una pequeña red abastecida por al menos una o a lo sumo dos fuentes de abastecimiento, en donde la demanda tendría un valor máximo en función de la capacidad de la(s) fuente(s), y la elevación del terreno se encontraría dentro de un rango de similitud (Herrera, 2011a).

Con base en su funcionalidad, los sectores han sido clasificados de dos maneras: DMA y PMA (área de gestión de presión o *pressure management areas*). En los DMA, el objetivo principal es la estimación de demanda. En estos se contabiliza tanto el caudal de entrada como el de salida (consumo); sin embargo, no se efectúa ninguna acción directa encaminada a gestionar la presión. Por otro lado, en los PMA también se gestiona la presión además de los caudales (GIZ *et al*., 2011)

Morrison *et al*. (2007), establece una clasificación de tipo topológico de los DMA (que también es aplicable a los PMA) en función del tipo de alimentación de las que dispongan. Así, estos pueden ser clasificados en: DMA con una sola entrada; DMA con múltiples entradas; y “cascadas” de sectores. Las “cascadas” de

sectores corresponden a DMA dentro de otros DMA.

En función del número de fuentes disponibles en la RDAP, la división de la misma en sectores puede clasificarse en "Partición" y/o en "Sectorización". El primer término se aplica a RDAP que cuentan con un número reducido de fuentes; siendo los DMA separados, ya sea mediante la instalación de válvulas seccionadoras o mediante cortes de tuberías (Di Nardo *et al*., 2013a). El segundo término se emplea en redes con un mayor número de fuentes de abastecimiento, de manera tal que se puede asignar una fuente independiente a cada DMA. En el primer caso, las áreas resultantes son denominadas DMA y en el segundo caso son denominadas DMA-aislados. La manera de proceder para dividir RDAP en DMA o DMA-aislados depende de las características topológicas de la red y las características topográficas del sitio en la que la misma se ubica. En redes que han evolucionado por problemas coyunturales, en lugar de responder a una planificación previa, es muy posible que no se pueda proceder por un único método, sino por una combinación de ambos.

El presente trabajo está orientado exclusivamente al estudio de sectores en los que únicamente se controla el caudal y que, además, son alimentados por una red de alta; así, a partir de este punto y para fines de simplificación, el término sector, definirá *subareas* de una RDAP con una única entrada que no son alimentadas por una fuente exclusiva, sino a través de una red primaria; el término sectores-aislado definirá *subzonas* de abastecimiento con una o más fuentes exclusivas, y el término sectorización será empleado para definir la "partición" o subdivisión de una RDAP en sectores.

5.2 Historia de la sectorización

El concepto de sectorización de RDAP que se maneja en la actualidad se da a conocer en la década de 1980 en Inglaterra por parte la *Asociación de Autoridades de Agua*. Desde entonces, hasta la actualidad, el avance que ha tenido el estudio de la misma ha sido relativamente limitado (Herrera, 2011a).

Walski *et al.* (2001) propone el establecimiento de sistemas con medición para apoyar la implementación de la sectorización de redes de agua potable. Tzatchkov *et al*. (2008) aplica la teoría de grafos para identificar zonas hidráulicas independientes. El trabajo se lleva a cabo mediante el software SCARED[2] del Instituto Mexicano del Agua (IMTA). SCARED contiene una aplicación de sectorización que se basa en la zona de cobertura de las fuentes de abastecimiento para establecer sectores aislados. Posteriormente, Di Nardo *et al.* (2013a), teniendo como referencia índices de eficiencia energética, propone un método para reevaluar las fronteras de cada sector generado por el método arriba descrito. En el año 2007, Morrison *et al.* (2007), presenta la guía práctica de la IWA[3] para manejo de sectores. Esta guía establece conceptos en torno a la sectorización y brinda pautas generales para llevar a cabo un procedimiento de esta naturaleza. Hunaidi y Brothers (2007) se concentra en la búsqueda del tamaño óptimo de los sectores basados en diferentes criterios y tomando en cuenta sus costes económicos. En Izquierdo *et al*. (2008) se estudia la importancia relativa de las tuberías en las redes de abastecimiento de agua potable. Con base en este trabajo es posible establecer criterios para la división de las redes por zonas. Izquierdo *et al*. (2011) desarrolla un software para llevar a cabo sectorizaciones de RDAP basado en la aplicación de técnicas Multi-agente (MA). Saldarriga *et al*. (2008) presenta una metodología para evaluar el efecto de la sectorización de redes de abastecimiento empleando como medida de impacto de la misma el índice de resiliencia antes y después de la sectorización. Más recientemente, Herrera (2011a) presenta la tesis doctoral: "*Mejora de la gestión de las redes de agua potable mediante la división eficiente en clústeres de abastecimiento*". En el mismo trabajo se muestra un método de división de RDAP en clústeres mediante aprendizaje *semisupervisado*. También, para el mismo propósito se emplea la técnica MA. El primero de los métodos presenta dificultades al aplicarse a redes reales (de gran extensión), con lo cual también se presenta un método de *remuestreo* de los datos que, mediante aprendizaje *semisupervisado* y la técnica MA, permite resolver el problema para redes de mayor extensión. Vegas (2012) presenta en una de tesis de máster, una herramienta para efectuar sectorización de

[2] SCARED: Software de gestión de redes de abastecimiento del IMTA
[3] International Water Association o Asociación Internacional del Agua

redes de manera automática. La misma se basa en la teoría de grafos y está implementada en el software GISRED. A manera de descripción general, la misma funciona identificando inicialmente la red de alta, y luego encontrando en el resto de la red (de distribución) los sectores, que no necesariamente deben encontrarse aislados; es decir, a través de esta herramienta se pueden encontrar sectores separados entre sí mediante válvulas y compartiendo una misma fuente de abastecimiento.

Di Nardo y Di Natale (2011a) proponen una técnica heurística para encontrar sectores a través del camino de mínima disipación de energía. Con esta técnica, los operadores pueden escoger la partición que optimice el consumo energético en la red. Di Nardo *et al*. (2011b) aplica una la técnica MLRB[4] para subdividir redes a múltiples niveles, seleccionando la mejor partición con base a índices de eficiencia energética. También en Di Nardo *et al*. (2013b) se desarrolla una metodología para identificar DMA-aislados mediante combinación de la técnica de grafos con algoritmos genéticos.

5.3 Casos de implementación de la sectorización

5.3.1 Sectorización de la ciudad de Managua, capital de Nicaragua.

La ciudad de Managua es la capital de la república de Nicaragua. En esta ciudad se realizó un proyecto de optimización en una parte de la RDAP. El área de actuación cubrió 1340 km de extensión y entre sus fuentes de abastecimiento de agua potable, cuenta con un lago y 138 pozos distribuidos por todo el territorio, algunos de ellos agrupados a manera de campos de pozos. Una parte de la red fue subdivida en 65 sectores, todos alimentados desde la red de alta de la ciudad. El proceso de división se realizó mediante un proceso empírico, teniendo en cuenta

[4] Multilevel Recursive Bisection o Bisección Recursiva Multinivel

un modelo matemático de la red de alta de la RDAP y un modelo matemático conjunto de la red de alta y la red secundaria, además de las características topológicas de la red y otros aspectos de carácter urbanístico. Dentro del diseño de los mismos se contempló que cada uno de ellos contara con una sola entrada, eliminando otras posibles entradas, ya sea mediante el cierre de válvulas o mediante el corte de tuberías. Al final, los 65 sectores fueron conformados con un tamaño aproximado de 20 km de longitud de tubería. En cada una de las entradas se colocó una unidad operativa de control (UOC) con un macromedidor y una toma de presión. En 64 de los 65 sectores se comprobó el aislamiento de los mismos evaluando las caídas de presión al cierre de la línea de alimentación. En uno de ellos no se logró hacer dicha comprobación dado que la antigüedad de la red y las carencias catastrales impidieron detectar todos los puntos de inyección de agua al mismo.

5.3.2 Dos casos de sectorización en Perú.

Vegas (2012) menciona dos casos de sectorización en Perú. Uno en la ciudad de Lima y Callo y el otro en la ciudad de Huacho. En el primero de los casos se estableció como criterio el área de los sectores. Estos no debían tener un área superior a las 300 ha, con lo que se esperaba tener sectores en los que se alimentara 400 ~ 4000 usuarios. En el otro caso se establecieron sectores con áreas en torno 3 km^2, los cuales debían tener un único punto de entrada.

5.3.3 Sectorización de la RDAP de Tegucigalpa. Capital de Honduras.

La ciudad de Tegucigalpa es la capital de Honduras. Su RDAP cuenta con 1800 km de tubería, 50 tanques de distribución y dos plantas potabilizadoras, más un campo de pozos. Esta ciudad destaca por una topografía bastante irregular, de manera que, en general, el sistema se abastece mediante tanques ubicados en las partes altas que son alimentados a través de bombeos desde las plantas

potabilizadoras. En el periodo 2010-2012, la empresa estatal SANAA[5] llevó a cabo un proyecto de optimización de la red de agua potable. El proyecto, además de incluir el levantamiento catastral de una parte de la RDAP, también incluyó una campaña de detección de fugas y la ejecución de la sectorización. El diseño de la sectorización se apoyó en un modelo matemático de la red de alta, cuya calibración no se pudo llevar a cabo debido a la intermitencia característica del servicio (las distintas zonas residenciales en general son abastecidas aproximadamente 10 horas por semana). Cada tanque tiene varias salidas (en algunos casos hasta seis) independientes para diferentes zonas, de tal manera que la red ya se encontraba algo sectorizada, por lo que a la hora de plantear el esquema de sectorización, se trató de mantener el esquema de sectorización ya existente, haciendo divisiones en la zonas en que la presión pudiese superar el límite máximo de presión establecido por la autoridad local. Estas divisiones se hicieron mediante cortes de tuberías, instalación de válvulas nuevas o cierre de válvulas existentes, aunque la autoridad local manifestó preferencia por los cortes de tuberías. Al final la red quedó subdividida en 75 sectores, alimentados bien por líneas salientes de tanques, o mediante inyecciones salientes de la red de alta.

5.3.4 Sectorización de San Luis de Rio Colorado, Sonora, México

En el informe de acciones IMTA-2008 se describe, el trabajo de reducción de fugas realizado en un microsector de una RDAP. Con la actividad de reducción de pérdida se logró alcanzar una eficiencia física de 77% (IMTA, 2008).

La red de la ciudad abastece a una población de 18,000 habitantes. Para ejecutar el proceso de sectorización se creó un modelo matemático apropiadamente calibrado con mediciones de caudal y presión, además de niveles de pozos. El mismo fue implementado en SCARED. Para definir lo sectores hidrométricos (10 en total) se implementaron en SCARED una serie de algoritmos que permiten encontrar la zona de influencia de cada una de las fuentes de abastecimiento. Dichos algoritmos se describen ampliamente en Tzachkov *et al.* (2008).

[5] Sistema Nacional Autónomo de Acueductos y Alcantarillados

5.4 Diseño de sectores

Para abordar el tema relativo al diseño de sectores en las RDAP, es de suma importancia tener en cuenta la naturaleza "caótica" de las mismas, dado que en muchos casos su crecimiento, más que responder a un proceso de planificación, es un reflejo de la subsanación de problemas locales y en el corto plazo, asociados al crecimiento de las poblaciones, o visto desde el punto de vista de la demanda, son la consecuencia de años de respuestas anárquicas a un aumento de la demanda. De ahí que sea lógico esperar que las mismas carezcan de una clara estructura desde el punto de vista topológico y luego sean complicadas para ser comprendidas, controladas y gestionadas (Izquierdo *et al*., 2011). En tal sentido, el diseño de sectores en una RDAP puede llegar a ser una tarea relativamente sencilla en caso de redes de pequeña extensión, siendo posible el empleo de mapas básicos de las RDAP para generar una definición de los mismos que mantenga los parámetros hidráulicos dentro de rangos aceptables; no obstante, en redes de gran extensión, el problema se torna complejo porque, más allá de la necesidad de mantener la homogeneidad general de la red dentro de cada sector, se deben tener en cuenta criterios de optimización, es decir, generar sectores homogéneos con el menor coste posible (reduciendo el número de válvulas a instalar) que además mantengan los parámetros hidráulicos dentro de un rango de aceptabilidad, teniendo siempre presente el efecto que esto puede generar sobre el resto de la red. Con respecto a esto, Saldarriaga *et al*. (2008) y Todini (2000), establecen que la sectorización reduce la respuesta de una red ante la falta de uno o varios de sus elementos, haciéndola más vulnerable. De ahí la necesidad de buscar metodologías que permitan brindar soluciones a este tipo de problemas de manera eficiente.

5.4.1 Factores a tener en cuenta en el diseño de sectores.

De manera general, los factores que se deben tener en cuenta al momento de diseñar sectores son los siguientes (Morrison *et al*., 2007).

- ***El nivel económico de fugas requerido***

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 que existe un umbral mínimo de fugas (UMF) que no puede ser evitado. En la Ilustración 1 se puede apreciar de manera más clara este concepto. La curva color negro representa el comportamiento del gasto por reparación y mantenimiento, que es mayor conforme menor es el volumen de agua que se fuga, tal a como se puede intuir. La curva color azul representa el volumen de agua incontrolada, cuyo crecimiento es directamente proporcional a su coste. La curva color rojo es una 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.

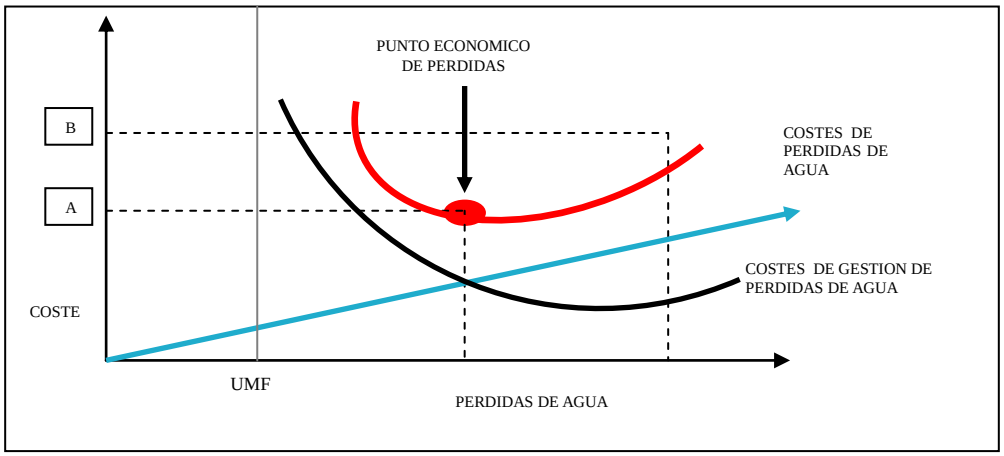


Ilustración 1: Punto óptimo de control de fugas (Farley *et al.*, 2008) (Pearson y Trow, 2008)

Por otro lado, también existen métodos para calcular el tiempo óptimo de inspección de fugas (FOI), tal como el que plantea la siguiente ecuación.

$$FOI = \sqrt{\frac{CI}{CMA * INAF * 0.5}}$$

En donde *CMA* representa el coste medio del agua, *INAF* representa el índice natural de aumento de fugas (es decir, la tasa de aumento de fugas de

una RDAP cuando se sigue una política pasiva para el control de fugas) y *CI*, el coste que representa cada una de las inspecciones. Si se aplica esto como criterio al momento de diseñar sectores en una RDAP, se hace evidente que al tener sectores de menor tamaño el coste de inspección es lógicamente menor y luego la *FOI* disminuye.

- ***Tamaño***

El tamaño de los sectores se relaciona con el nivel de precisión con que se deseen ejecutar las campañas de identificación de fugas (DVGW, 2003; Morrison *et al*., 2007; GIZ *et al*., 2011). También depende de las características urbanas de la red. Cuando una red ya se encuentra naturalmente aislada, es lógico que las subredes existentes se configuren como sectores.

- ***Tipo de viviendas o tipo de consumidor***

Es de gran importancia caracterizar cada área conformada como sector, de manera tal que se pueda mantener la homogeneidad dentro de la misma. En tal sentido, se debe identificar cualquier gran consumidor o algún consumo especial, cualquier edificio que requiera un abastecimiento con una presión superior a la norma establecida para el área, contornos de nivel de terreno, etc. (Morrison *et al*., 2007).

- ***Variación en el nivel del terreno***

Conservando el criterio de homogeneidad, se trata de mantener la red dentro de un rango de elevación de terreno para así lograr mantener las presiones en un rango de similitud.

- ***Consideraciones de calidad de agua***

Se debe evitar que la sectorización implique la aparición de zonas muertas (zonas de estanqueidad de agua). En caso necesario, se debe considerar la colocación de desagües.

- ***Requerimiento de presión***

El valor de presión en el sector se debe mantener en el rango establecido en las normativas locales. Si el problema es un exceso de presión, se puede considerar realizar una gestión de la presión dentro de los sectores, lo que se hace mediante la implementación y seguimiento de una curva de consigna. También puede ser necesario el aumento del diámetro de algunas líneas para reducir la pérdida de potencia de entrada al sector.

- ***Número de válvulas que se deben cerrar***

En general, se debe minimizar; no obstante, se debe compatibilizar con el CAF, de tal manera que la minimización del número de válvulas debe tener también en cuenta las implicaciones que tienen sectores más grandes sobre el coste del CAF.

- ***Capacidad del sistema contra incendios***

Pese a que la sectorización conlleva una pérdida de capacidad, en la medida de lo posible se deben tratar de mantener los rangos de presión en los valores adecuados para poder reaccionar ante una situación de emergencia.

- ***Número de caudalímentros (idealmente optimizado)***

Este aspecto se relaciona con el número de sectores. En la medida de lo posible se trata de que cada uno de los sectores cuente con un sólo caudalímetro, ya que esto reduce los costes asociados al mantenimiento posterior y reduce la complejidad para hacer el balance de caudales del sector; no obstante, dependiendo de las características de la red, puede

haber algunos sectores en los que sea necesario colocar más de uno. También se tiene que tener en cuenta la colocación de caudalímetros especiales para los grandes consumidores, en donde el comportamiento y la magnitud del consumo no sean comparables al consumo residencial.

5.4.2 Reglas para establecer sectores

Antes de arrancar un proceso de sectorización de una ciudad, es importante tener en cuenta los factores tales como las condiciones de operación, diseños del sistema de distribución y factores ambientales, así como la economía de las empresas encargadas de las distribución del agua (Kleppen, 2011). El primer paso para ejecutar la sectorización de una RDAP es contar con un buen conocimiento de la estructura y dinámica de funcionamiento de la misma. Este aspecto es fundamental para, además de optimizar el diseño de los sectores, evitar que los mismos acarreen consecuencias negativas, tales como problemas de desabastecimiento o incluso de calidad de agua. El diseño de los sectores debe comenzar por la línea de alta y a partir de ahí, dirigirse a la red de distribución (Morrison *et al*., 2007). El objetivo es separar lo máximo posible la red de alta de manera tal que siga siendo flexible, por si en el futuro se desea realizar alguna modificación particular. En el caso de redes de abastecimiento con problemas de baja presión, reviste vital importancia apoyar el proceso de toma de decisión en modelos matemáticos que permitan validar la idoneidad de cualquier decisión.

Se han establecido algunas reglas generales para el diseño de sectores, tales como las que a continuación se detallan (GIZ *et al*., 2011):

- El diseño de un sector no debe incluir bucles o tanques de almacenamiento. Si esto es inevitable se deben instalar caudalímetros que controlen el caudal de entrada y de salida.
- Cada sector debe ser alimentado por un único punto, el cual debe contar con un contador. Dado que en algunos casos se hace necesario contabilizar caudales muy bajos, por lo general es necesario que el contador en la entrada tenga un diámetro menor que el del tubo de

alimentación en el cual está instalado.

- Los límites del sector deben ser definidos mediante válvulas cerradas. De ser posible, estos límites deben ser naturales (ríos, carreteras, líneas de metro) para así minimizar la utilización de válvulas.
- Se debe hacer una evaluación de los distintos tipos de consumidores y el suministro que se le brinda a estos mismos.
- Se deben respetar las regulaciones locales.
- El cierre de válvulas para aislar sectores puede provocar la aparición de zonas muertas; por ende, en el diseño de los sectores se debe tener en cuenta los problemas relacionados con la estanqueidad del agua.
- Dado que la presión juega un rol muy importante en la gestión de las fugas, en la medida de lo posible, la gestión se debe incorporar en el proceso de reconfiguración del sistema cuando se diseñan sectores

5.4.3 Tamaño óptimo de los sectores

Tal y como se estableció al principio de esta sección, el componente de optimización económica cobra vital importancia al momento de establecer el tamaño de los sectores. El coste por sector en una RDAP es más alto en la medida en que su tamaño sea menor y así puedan representar un mayor número total. Esto se debe a que, para su conformación, se debe instalar un mayor número de válvulas o cortar un mayor número de líneas. Como se puede intuir, lo contrario pasa con sectores de mayor extensión; no obstante, los sectores de menor tamaño presentan varias ventajas, tales como: mayor agilidad para detección temprana de nuevas fugas, mayor capacidad para distinguir entre pequeñas fugas, caudal mínimo nocturno y fugas de fondo y luego una localización más rápida de las mismas. En este sentido, se ha establecido que existen tres aspectos claves que definen la cantidad de agua que se pierde por una fuga individual: tiempo de conocimiento, tiempo de detección y tiempo de reparación (Pilcher *et al*., 2007). En RDAP gestionadas con sectores, el tiempo de cada uno de estos factores se puede reducir de manera significativa y por ende reducir el caudal de fugas.

De igual manera, cuanto menor es el tamaño de sectores, más sencillo resulta

priorizar zonas para ejecutar inspecciones. La IWA recomienda definir el tamaño de los sectores en función del número de acometidas (conexiones) dentro de ellos, estableciendo un rango que va de 500 a 3000 conexiones como el rango aceptable (Morrison *et al*., 2007). Por otra parte en GIZ *et al*. (2011) se establece como criterio la longitud de tubería dentro del sector, definiendo un rango que va de 4 km a 30 km como un rango aceptable dentro del cual se debe encontrar el tamaño de los sectores. Algunos análisis económicos han demostrado que en la mayor parte de los casos, el tamaño de los sectores debe estar entre 3000 y 5000 conexiones (Thornton *et al*., 2008).

Tal como se explica más adelante, un criterio que se podría emplear para definir el tamaño apropiado de los sectores es la eficiencia energética de la RDAP. Tomando como punto de partida la energía aportada por cada una de las fuentes antes de hacer la sectorización, se podría buscar el esquema de sectorización que implique menor porcentaje de energía disipada a través de la red.

5.5 Conceptualización de ML y Gestión de fugas

En la presente sección se hace una revisión de conceptos claves para la compresión de las técnicas aplicadas. Se inicia con un repaso de los aspectos más relevantes de la teoría de grafos, que tal y como se verá más adelante, corresponde a la herramienta empleada para lograr transformar una red de abastecimiento en un conjunto de clústeres. A continuación se aborda la teoría de clúster, en vista de que cada sector se trata como tal. A continuación se aborda la teoría de *ML*, describiendo la idea básica de esta, se presta especial atención al algoritmo máquina de soporte de vectores (SVM, *Support Vector Machine*); a los métodos *kernel* como herramienta para efectuar clasificaciones no lineales; al clústering espectral y al clústering jerárquico. Este último se emplea para estimar el número de sectores en que puede ser subdividida una RDAP. Luego se hace mención del software-lenguaje de programación R, mediante el cual se lleva a cabo el procedimiento planteado, y se finaliza con una descripción de aspectos relacionados con las fugas en las RDAP.

5.5.1 Teoría de Grafos

La teoría de grafos es uno de los campos de estudio de las matemáticas de conjuntos discretos (finitos e infinitos) o matemáticas discretas. Un grafo es un conjunto de objetos llamados vértices o nodos unidos por enlaces, llamados aristas o arcos, que permiten representar relaciones binarias entre los elementos de un conjunto (Thulasiraman y Swamy, 1992); (Xu, 2003). El término grafo, es empleado por primera vez a mediados del siglo XVIII y se origina a partir de la "notación gráfica" para describir los enlaces moleculares en las ciencias químicas. El origen de esta rama de la ciencia se remonta al siglo XVIII, ubicándose en la ciudad rusa que actualmente tiene el nombre de Kalingrado (antiguamente Konisberg). Leonard Euler, publicó un documento que le daba solución al llamado "*problema de Konisberg"*, que perseguía plantear una manera de recorrer siete puentes de la ciudad pasando por cada uno de ellos una sola vez (Biggs *et al*., 1986); (Goset, 2009). Este trabajo es considerado el primero publicado sobre la teoría en cuestión. Con posterioridad hay otros trabajos destacables, como los de Gustav Kirchoff, publicados en 1847, que empleó la teoría de grafos para establecer las muy conocidas leyes para cálculo de corriente, voltaje y resistencia en los circuitos eléctricos.

De una manera más técnica, se puede decir que un grafo (G) es un conjunto finito de objetos (VE) llamados vértices (V) o nodos, unidos por enlaces llamados aristas o arcos (V,V) o (E), siendo J la relación de incidencia de dos elementos tipo (V) que se asignan a un enlace (E); así, se puede denotar un grafo como $G = \{V, E_J\}$. De lo anterior se puede ver cómo estas estructuras permiten representar relaciones binarias entre los elementos de un conjunto, o dicho de otra manera, permiten estudiar las interrelaciones de unidades que se encuentran en interacción.

Las aristas, o elementos de enlace de los grafos, según los vértices inicial y final desde donde parten y donde terminan, pueden ser clasificadas en: adyacentes, en caso de dos aristas que tienen un sólo vértice en común; aristas paralelas o múltiples, en caso de compartir los dos vértices; lazo, en caso de que el vértice de

partida sea el mismo vértice de llegada. En cualquiera de los casos, si una arista termina en un vértice dado, se dice que la arista es incidente a tal vértice. Cuando en un grafo existe una sola arista entre dos vértices y además no existen lazos o aristas paralelas, se dice que se trata de un grafo simple o sencillo.

Respecto a los vértices, por un lado, existe una propiedad que se asocia a estos mismos, que tiene el nombre de grado de conectividad G (también es conocida como valencia). Esta, corresponde al número de aristas que inciden sobre los mismos, y por otro lado, también se pueden clasificar de acuerdo a la situación de las aristas que inciden o no en ellos. Así, un vértice sobre el cual no incide ni una sola arista se conoce como vértice aislado, o vértice de grado cero y un vértice sobre el que incide una sola arista, se conoce como vértice de grado 1. En esta misma línea, los grafos compuestos únicamente por dos vértices y una arista se definen como vértices pendientes. 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 valor del grado de cada uno de los vértices. Cuando se da la situación en que las aristas de un grafo son paralelas (o bien forman lazos), se dice que el grafo es no simple.

Retomando la idea de grafos simples, si el grado de los vértices (*k*-aristas), es igual para todos los vértices, es decir, *k*-aristas es una constante, se dice que el grafo es *k* regular.

A las aristas de un grafo se le puede asociar un valor al que se llama peso o coste. En este caso se dice que el grafo es un grafo ponderado. Por lo general los pesos de un grafo se representan mediante matrices cuadras $n \times n$.

Las aristas de un grafo pueden tener o no dirección, es decir, los grafos pueden ser dirigidos o no dirigidos. En caso de que las aristas de un grafo tengan bien definido un vértice de origen y el vértice de fin, o dicho en otras palabras, que las

aristas tengan dirección, se dice que el grafo es dirigido (o simplemente digrafo). En caso contrario el mismo es no dirigido. En el caso de los grafos dirigidos, el vértice de origen de cada arista es conocida como cola de la misma, y el vértice de destino es conocido como cabeza.

De una manera más general, en los grafos no dirigidos, se dice que las aristas no están ordenadas y el grafo se representa con puntos (correspondiente a los vértices) y líneas (correspondientes a las aristas). Por el contrario, en los digrafos, se establece que las aristas si están ordenadas, el término aristas es reemplazado por arcos y el digrafo se representa por puntos (correspondiente a los vértices) y flechas (correspondientes a los arcos).

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 (o arcos en caso de digrafos) y vértices. Estas mismas (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 que no se repita ninguno de los vértices ni aristas en el mismo. En 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 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*.

5.5.1.1 Representación matricial de los grafos

5.5.1.1.1 Matrices de afinidad, de grado y de adyacencia

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

Si se toma como ejemplo el siguiente grafo (Ilustración 2).

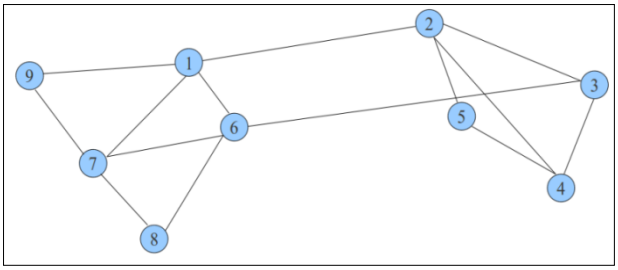


Ilustración 2: Red ejemplo

Esta misma red 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 1).

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

(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 1: Matriz de adyacencia (W_{ij}) del grafo

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

Por otro lado las matrices de *disimilaridad* representan las diferencias entre las

propiedades que pueden tener las aristas. Estas diferencias pueden ser medidas mediante métrica *euclidiana*, *manhattan*, entre otras. Esto será profundizado en la sección de clústering jerárquico.

$$(W_{ij}) = \textit{disimilaridad} \text{ entre } i \text{ y } j$$

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

$$D_{ij} = 0 \text{ si } i \neq j$$
$$D_{ij} = d_{vi} \text{ si } i = j$$
$$D = \text{diag}\ (d_1 ..... d_n)$$

Siendo cada elemento de la diagonal el 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 3), la matriz ***A*** vendría dada según:

$$A_{ij} = S_{ij} \text{ si } i \text{ y } j \text{ están conectados}$$
$$A_{ij} = 0 \text{ si } i \text{ y } j \text{ no están conectados.}$$

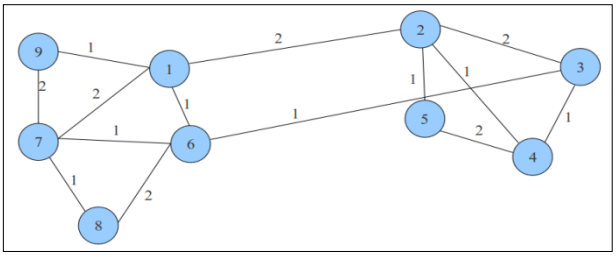


Ilustración 3: Grafo ponderado

La matriz ***A*** de afinidad del grafo anterior sería la que se presenta en la Tabla 2. En la matriz ***A*** la diagonal, en general, no toma valores diferentes a cero en vista

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 2: Matriz de afinidad (A_{ij}) del grafo

5.5.1.1.2 Construcción de la matriz Laplaciana

La matriz laplaciana (***L***), tal y como se muestra en la siguiente ecuación, es igual a la resta de ***D*** menos ***A***.

$$L = D - A$$

L es la matriz *laplaciana* no normalizada

D es la matriz de grado

A es la matriz de afinidad

Para la construcción de la matriz ***L*** no normalizada se establecen tres reglas:

$$L_{ij} = \begin{cases} L_{ij} = grad_{vi} si\ i = j \\ -1\ si\ i = j\ y\ vi\ es\ adyacente\ a\ j \\ 0\ en\ cualquier\ otro\ caso \end{cases}$$

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 3.

	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 3: Matriz *laplaciana* (L_{ij}) del grafo

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

$$L_{ij}^{*} = \begin{cases} 1 \; si \; i = j \; y \; k_i \neq 0 \\ -\dfrac{1}{\sqrt{k_i k_j}} \;\; i \neq j \; y \; n_i \text{es adyacente a } n_j \\ 0 \; en \; cualquier \; otro \; caso \end{cases}$$

5.5.1.2 Grafos de las redes de abastecimiento

Las RDAP representadas en EPANET 2.0[6] pueden ser concebidas como un grafo, en donde los nodos de consumo y abastecimiento (tanques), son los vértices, y las tuberías y válvulas, las aristas. En los digrafos, la dirección del caudal de las tuberías representaría la dirección de la misma. Las características que se pueden agregar a los vértices de un grafo de RDAP son las mismas que tienen los nodos en su modelo matemático (*coordenadas geográficas, elevación, demanda, coeficiente de emisor, presión, altura piezométrica*). En el caso de las tuberías, las características que se pueden emplear son: *diámetro, longitud, rugosidad y pérdidas de carga por fricción*. Tanto para las aristas, como para los vértices, se pueden crear y/o agregar nuevas propiedades. Por ejemplo, en el caso de los vértices, se podría agregar características sociales propias de los sitios de estudio,

[6] Software de simulación hidráulica cuasi-estacionaria desarrollado por EPA (Rossman, 2008)

o la potencia hidráulica en los nodos, expresadas en $m^3/h * mca$. También se podría agregar las pérdidas en las tuberías, la cual se puede expresar en $m^3/h * mca$, en donde los mca, es la pérdidas de carga (H) que se da en estas.

5.5.2 Teoría de Formación de Clústeres

Un clúster se define como un conglomerado de objetos que comparten muchas características entre sí, o dicho de otra manera, aquel en que los perfiles de los objetos en un mismo grupo sean muy similares entre sí, pero son muy disimilares a aquellas en los objetos que pertenecen a otros clústeres (Karlson, 2008; Romesburg, 2004; Mooi y Sardtedt, 2011). Un análisis de clústeres se define como la partición de las observaciones en grupos de manera que las disimilitudes por parejas[7] (medida de cuán diferentes son dos elementos) entre los elementos asignados a un clúster sean menores con respecto a elementos pertenecientes a otros clústeres (Hastie *et al*., 2009). Los seres humanos nacen una capacidad nata para formar clústeres, estableciendo categorías para todos los elementos que le 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 (Von-Luxburg, 2007).

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 (clústering jerárquico) y por otro lado una partición en clústeres (*partitioning clustering*), (Arabie y 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 clúster o *"singleton"*).

[7] Pairwise dissimilarities

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, clústeres no disjuntos, que corresponde al caso en que un mismo elemento(s) 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 asociado 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 (Abonyo y 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 del conjunto de datos.

También existe una clasificación de los clúster generados en función de las metas que persiga el análisis de clúster. Así, estos pueden ser bien separados, cuando se basan 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. Este tipo 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, siguiendo la tipología propia de un grafo de acuerdo a la teoría de grafos, que será abordada en la siguiente sección. 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 la 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, estos incluyen clústeres que puedan estar enlazados entre sí mediante ciertos elementos en común.

En relación con las RDAP, Herrera (2011a) define a los clústeres como sectores hidráulicos producidos con el uso de *ML*. En este sentido, un elemento agrupado en un clúster corresponde a un nodo de la RDAP perteneciente a un sector. Dentro de ellos existen una serie de características hidráulicas tales como: *elevación*, *demanda*, *coordenadas geográficas*, *coeficiente de emisores*, y otras que pueden ser creadas, que se emplean como criterios para llevar a cabo la formación clústeres o sectores.

5.5.3 Conceptualizaciones sobre Machine Learning (*ML*).

ML es la rama de la inteligencia artificial que se dedica al estudio y diseño de sistemas que son capaces de aprender reglas por sí mismo, en función de la información que se le provea para su entrenamiento. Pero más allá de aprendizaje de reglas, también tienen la capacidad de adaptarse a cambios y mejorar su despeño (Alpaydin, 2004; Koronacki *et al*., 2010). Es decir, en lugar de programar los ordenadores para efectuar una tarea, se enseña al ordenador a auto-programarse a partir de los datos de entrenamiento. En este caso, emplea un mecanismo inductivo de aprendizaje, descubriendo leyes generales o conceptos a

partir de un número limitado de ejemplos, basando su funcionamiento en la observación de similaridades entre los ejemplos (Sammut y Webb, 2010; Koronacki *et al*., 2010). El origen de este término se remonta al año 1959, cuando Arthur Samuel definió el aprendizaje automático como "el campo de estudio que enseña a los ordenadores a aprender sin necesidad de ser explícitamente programado para esta tareas". Esta rama de la ciencia tiene un amplio número de aplicaciones prácticas, incluyendo la clasificación, la detección de patrones, predicción, detección de valores atípicos, extracción de conocimiento, entre otros.

El aprendizaje automático está ampliamente relacionado con otra rama de la ciencia, conocida como minería de datos (*data mining* en inglés). Algunos autores establecen que la minería de datos es el paso del aprendizaje automático a grandes bases de datos; no obstante, algunos son un poco más conservadores, al establecer un grado de diferencias entre ambas, definiendo a la minería de datos como la herramienta que identifica o descubre propiedades desconocidas en los datos, en tanto define aprendizaje automático como la tarea final para hacer predicciones a partir de propiedades ya conocidas que se descubren mediante entrenamiento con datos (Han *et al*., 2012; Nagabhushana, 2006). Una de las propiedades más importantes dentro del campo del aprendizaje automático es la generalización. Una vez que el sistema construye su algoritmo en función de lo datos de entrenamiento que se han brindado, éste tiene que tener la capacidad de ejecutar la misma tarea sobre datos nuevos que no han sido vistos previamente (Sammut y Webb, 2010; Han *et al*., 2012; Witten y Frank., 2005). Esto se relaciona con la capacidad del algoritmo de ejecutar de manera eficiente la tarea para la cual fue programado, aun en presencia de ruido.

5.5.3.1 Clasificaciones de ML

Los algoritmos que se generan por *ML* se clasifican taxonómicamente en función de la salida deseada. De acuerdo a Taiwo (2010), los tres algoritmos más conocidos son: *aprendizaje supervisado*, *aprendizaje no supervisado* y *aprendizaje semisupervisado*. Otros tres algoritmos, un poco menos conocidos son: *aprendizaje por refuerzo, transducción* y *aprendiendo a aprender*.

5.5.3.1.1 Algoritmos basados en métodos supervisados

En los métodos de *aprendizaje supervisado*, se trata de construir algoritmos que razonan a partir de instancias suplidas del exterior, a partir de las cuales se genera una hipótesis general que luego puede hacer predicciones sobre instancias futuras, o dicho de otra manera, se infiere una función para relacionar los datos de entrada con los datos de salida. En este caso, cada una de las instancias le corresponde una etiqueta, es decir un valor de salida establecido *a priori* (Cord y Cunningham, 2008). Dentro de los algoritmos más conocidos que funcionan con este método de aprendizaje, destacan: las redes neuronales, el árbol de decisión y las SVM. En general las SVM fueron concebidas para ejecutar particiones binaria de datos; no obstante, tal y como se explica más adelante, cuando se hace una apropiada conjugación de las mismas con los métodos *kernel* para la detección de patrones, se pueden emplear para resolver problemas de tipo *multiclase* (en que los datos se agrupan en más de una clase).

5.5.3.1.2 Algoritmos basados en métodos no supervisados

Estos métodos se basan en el aprendizaje de los sistemas a partir de grupos particulares de datos de entrada, tales que reflejen la estructura estadística de todo el conjunto de entrada. En este tipo de algoritmo no hay datos etiquetados, es decir, no existen datos de salida asociados *a priori* a una entrada. Los únicos datos con que se trabaja en este método son patrones de entradas que se suponen independientes, provenientes de una distribución probabilística desconocida (Daya *et al*., 1999; Ghahramani Z., 2004).

Este tipo de métodos puede ser dividido en dos tipos de problemas: la formación de clústeres a partir del conjunto de datos y la extracción de características del conjunto de datos (Shental, 2004). Tal y como se planteó previamente, la formación de clústeres o análisis exploratorio de los datos, persigue revelar estructuras inherentes al conjunto de datos, en tanto la extracción de características persigue reducir las dimensiones de los datos para poder hacer una representación más compacta del conjunto.

5.5.3.1.3 Algoritmos basados en métodos semisupervisados

Los métodos de aprendizaje *semisupervisado*, son un caso intermedio de los dos métodos descritos anteriormente. En este caso, también se infiere una hipótesis que relacione los datos de entradas con la salida; no obstante, en este caso, a diferencia de los *métodos supervisados*, sólo una parte de los datos de entrada corresponden a datos etiquetados, en tanto la otra parte corresponde a datos sin etiquetar. Dado los altos costes implicados en la obtención de etiquetas para los datos de entrada (costes de experimentación, o de personal con alto nivel de experiencia), es de esperar que la proporción de datos de entrada etiquetados sea significativamente inferior que la proporción de datos de entrada no etiquetados (Wang *et al*., 2009). La idea del aprendizaje *semisupervisado* es clasificar los datos no etiquetados con base en su distribución y con base en los datos etiquetados proveídos (Albalate y Minker, 2011). En este sentido, es importante mencionar que algunas investigaciones recientes han demostrado el beneficio sobre la precisión que tiene el empleo conjugado de datos de entrada con y sin etiquetas (Ando y Zhang, 2005).

5.5.3.2 Algunos Algoritmos de ML

5.5.3.2.1 Support Vector Machines (SVM)

Las SVM son algoritmos de tipo de aprendizaje automático *supervisado* que permiten hacer clasificaciones de patrones en conjuntos de datos *n*-dimensionales en espacios alta o infinitamente dimensionales (Boswell, 2002; Shawe-Taylor y Cristiani, 2004). De manera general, se caracterizan por establecer un *hiperplano* óptimo que separe un conjunto de datos en dos categorías (en caso de que se trate de un problema binario o los datos sean linealmente separables), o se pueden conjugar con métodos *kernel*, para lograr lidiar con problemas de tipo *multiclase*, en donde los datos se agrupan en más de dos categorías o cuando los problemas no son linealmente separables (Scholkopf y Smola, 2001; Soman *et al*., 2009). Cabe destacar que en estos métodos, los datos se manejan como productos escalares (producto propio o productor interno).

Se denomina *hiperplano* óptimo aquel que, además de separar los datos en las respectivas clases, maximice la distancia entre él y los puntos, dentro del conjunto de datos que se encuentren más cercanos al mismo. Dicho de otra manera, es aquel que tiene el mayor margen con respecto a las dos categorías de datos.

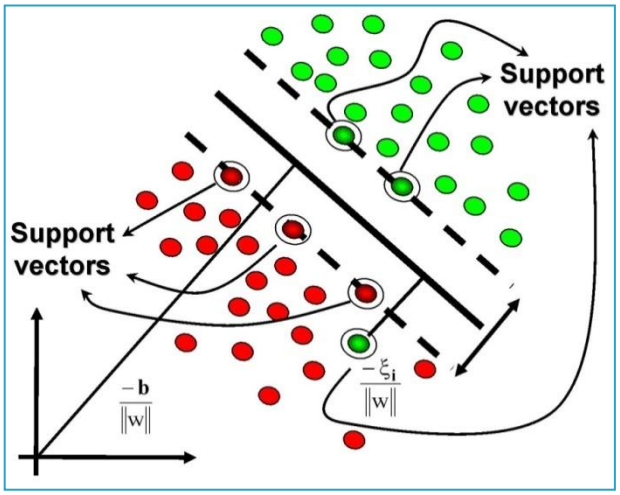


Ilustración 4: Clasificación de datos mediante SVM (Cristianini y Shawe-Taylor, 2000).

La motivación para combinar las SVM con los métodos *kernel* reside en el hecho de que la mayor parte del tiempo, los conjuntos de datos no son linealmente separables, por lo cual el *hiperplano* no puede ser construido en un espacio bidimensional, sino por el contrario, se construye en un espacio altamente dimensional.

Entre las características exclusivas de estos algoritmos se puede destacar que están explícitamente basados en modelos teóricos de aprendizaje. Tienen una garantía teórica respecto a su desempeño, tienen un diseño modular que les permite diseñar e implementar de manera separada sus componentes, y no se ven afectados por mínimos locales.

En este caso, los datos de entrenamiento conforman vectores de soporte que definen los márgenes del *hiperplano* óptimo de clasificación de los nuevos datos (datos no vistos). Estos datos de entrenamiento (elementos críticos) son los que especifican completamente la función de decisión del problema. Encontrar un

hiperplano que separe los datos termina siendo un problema de optimización.

5.5.3.2.2 Métodos *kernel*

En el campo de la ciencia de la computación, los métodos *kernel* son una serie de algoritmos empleados para hacer análisis de patrones. Tal análisis lidia con la detección automática de patrones en los datos, y juega un papel importante y central en la inteligencia artificial moderna y los problemas de las ciencia de la computación (Shawe-Taylor y Cristiani, 2004). Por patrón se entiende cualquier relación, regularidad o estructura inherente en una fuente de datos. El objetivo principal del análisis de patrones es predecir características de un dato en función de un valor característico.

El análisis de patrones basado en métodos *kernel* inserta los datos en un espacio característico adecuado y luego se emplean algoritmos basados en álgebra lineal, geometría y estadística para descubrir patrones en los datos incrustados (Shawe-Taylor y Cristiani, 2004). Este espacio corresponde al espacio Hilbert **H**, que es un espacio infinito dimensional con producto interior (Akheizer y Glazman, 1993). Un *kernel* se puede definir como una medida de similitud que puede ser pensada como un producto interno en el área de un espacio característico. También se puede definir como un producto interno en un espacio **H** por vía de una aplicación o mapa **Φ** (Scholkopf y Smola, 2001).

$$K(x,x') = <\Phi(x),\Phi(x')>.$$

Esta técnica recibe el nombre de "*Truco kernel*" (o "*kernel trick")*. Se basa en el teorema de *Mercer*, que establece que cualquier función *kernel* $K(x,y)$ continua, simétrica, semidefinida positiva, puede expresarse con un producto escalar en un espacio altamente dimensional, si los argumentos para el *kernel* están en un espacio medible *X*, y si el *kernel* es semidefinido positivo.

A su vez, el uso de *kernel* se basa en el teorema de *Cover*, que establece que un problema complejo de clasificación de patrones, emitido no linealmente en un espacio altamente dimensional, es más probable que sea linealmente separable que en un espacio de baja dimensionalidad siempre que el espacio no esté densamente poblado (Cover, 1965).

Una tal matriz $K = (K_{ij})$ es llamada matriz *kernel*. Con esta herramienta, se pueden buscar relaciones lineales en espacios altamente dimensionales a un costo computacional muy bajo. Esta contiene la función *kernel* en todos los pares de puntos y el conjunto de datos de entrenamiento, y debe ser semidefinida positiva. Sus valores propios deben ser no-negativos, para la convergencia de los métodos utilizados en los problemas de optimización.

En el método propuesto por Herrera (2011a), se emplean los métodos *kernel* para trabajar con la variabilidad natural presente entre toda la información que existe en una RDAP y que podría ser empleada para sectorizar la misma. Por otra parte, el uso de métodos *kernel*, permite aplicar algoritmos de formación de clústeres sobre datos que tiene conectividad entre sí (nodos unidos a través de tuberías). En este método, se transforma la RDAP en un grafo no dirigido. A partir de la matriz de afinidad *A* y la matriz de grado *G* de este grafo, se obtiene una matriz laplaciana *L* normalizada y se transforma en matriz *kernel*. Las matrices del resto de las características de la red (*elevación*, *demanda*, *coordenada-x*, *coordenada*-y y *coeficiente de emisor*) se transforman en matrices *kernel*. Una de las características de los métodos *kernel*, es que la suma de dos o más matrices *kernel* da como resultado una nueva matriz *kernel*; así, en este caso, las matrices *kernel* obtenidas se suman de acuerdo a la siguiente expresión.

$$K = \lambda_A k_A + (1 - \lambda_A) \sum \omega_i K_i$$

En donde

K: Matriz kernel sobre la que se realizará el proceso de formación de clústeres

k_A: *Matriz kernel asociada a la matriz de afinidad del grafo*

K_i: *es la matriz asociada con los inputs de interés*

λ_A: *Representa la importancia del grafo*

ω_i: *Son los pesos de la combinación lineal*

El valor de λ_A se calcula mediante un análisis de coste por número de válvulas necesarias para la sectorización. Para cada valor posible de λ_A (entre 0 y 1), se estima el número de válvulas que se requeriría instalar. Se genera una curva de comportamiento de ambas variables, y se selecciona el valor λ_A correspondiente al número de válvulas más bajo.

En el caso de los valores ω_i, que corresponden a los pesos de las distintas características hidráulicas: demanda, coordenadas geográficas (*x*,*y*), elevación y emisores, estos se evalúan mediante la aplicación del proceso de AHP[8], el cual será abordado más adelante.

5.5.3.2.3 Algoritmo *k*-means

Tal y como se ha venido introduciendo en secciones anteriores, la idea detrás de la formación de clústeres es introducir una medida de similitud en las entidades bajo consideración y combinar entidades similares en los mismos clústeres, mientras se mantienen entidades disimilares en diferentes clústeres (Mirkin, 2013; Karlson, 2008; Romesburg, 2004; Mooi y Sardtedt, 2011). Uno de los algoritmos más conocidos para la formación de clústeres es el algoritmo *k*-means. Este es un algoritmo basado en prototipo, es simple, "particional" y tiene como objetivo encontrar un número (*K*) de clústeres disjuntos. Es a su vez el algoritmo particional más popular (Jain, 2010). Los clústeres generados por él están representados por sus *centroides*.

Al ejecutar *k*-means, el primer paso es escoger el número deseado de centros de clústeres, *K*, y el procedimiento de *k*-means mueve iterativamente el centro para minimizar la varianza total dentro del clúster (Hastie *et al*., 2009). El objetivo del análisis *k*-means es llegar a una partición estable de *K* clústeres en la cual cada

[8] Analytic Hierarchy Process of Proceso de Análisis Jerárquico

caso esté lo más cercano a la media del clúster al que fue asignado (Wishart, 2001; Kanungo *et al*., 2002)

Wishart (2001), describe el procedimiento de *clústering* por *k*-means de la siguiente manera:

- *Se selecciona una partición inicial de los casos en K clústeres.*
- *Se computan las distancias desde cada caso a la media de cada clúster y se asignan los casos al clúster más cercano.*
- *Se re-computan las medias de los clústeres según cualquier cambio de membrecía de clúster del paso 2.*
- *Se repite el paso 2 y el 3 hasta que no ocurren cambios en la membrecía del clúster en una iteración completa. El procedimiento converge en una partición estable de K clústeres.*

La dificultad de *k*-means radica en encontrar clústeres "naturales", cuando el conjunto de datos tiene forma no *esférica* o cuando éstos presentan amplias diferencias de tamaño o densidad (Herrera, 2011b). En adición, el algoritmo *k*-means puede caer fácilmente en óptimos locales y a través de él, difícilmente se obtienen óptimos globales (Liu *et al*., 2011; Selim e Ismael, 1984). Otro problema que presenta, es la necesidad de elegir a priori un número *K* para realizar las particiones, cuando no existe alguna función eficiente y universal que permita elegir un número *K* más óptimo (Hamfelt *et al*., 2011).

5.5.3.2.4 Clústering Espectral

El clústering espectral representa una alternativa atractiva a las técnicas de agrupación controladas tradicionales, dado que son aplicables a situaciones en las cuales los objetos no están naturalmente representados en términos de vectores característicos y evitan asumir que todos los ejemplos en un clúster deben estar cerca de un prototipo. Esto quiere decir que están disponibles para objetos de forma irregular (Zelnik-Manor y Perona, 2004; Divakaran, 2009; Aggarwal, 2011). Su éxito se basa principalmente en el hecho que no hace fuertes asunciones

de la forma del clúster. Puede ser implementado eficientemente aun para extensas series de datos, con tal que se asegure que la matriz de similitud del grafo sea dispersa (Von-Luxburg, 2007). El algoritmo espectral de formación de clústeres se deriva de la teoría de división de espectro, y su esencia radica en convertir el problema de definición de clúster en un problema de partición óptima de un grafo. El algoritmo puede formar clústeres en un espacio de muestra de cualquier forma, es fácil de entender e implementar y no cae en óptimos locales.

Los pasos para realizar el *clústering* espectral son (Von-Luxburg, 2007):

Input: Matriz de similaridad S $\in R_{n*n}$, número *K* de clústeres a construir:

- *Construir un grafo de similaridad. Sea W su matriz de adyacencia ponderada.*
- *Computar la matriz L normalizada.*
- *Computar el primer K vector propio* U_1*.....* U_k *de la matriz anterior.*
- *Sea* $U \in R_n * K$ *la matriz que contenga los vectores* $U_1 \ldots\ldots, U_k$ *como columnas.*
- *Formar la matriz* $T \in R_n * K$ *a partir de U mediante la normalización de las filas siguiendo de manera que tenga 1 unidad de longitud. Es decir* $t_{ij} = u_{ij}/(k\, u^2)1/2$.
- *Para i=1,....,n, sea* $y_i \in R_k$ *el vector correspondiente a la n-ésima fila de T.*
- *Formar clústeres de los puntos* $(y_i)\ i = 1, \ldots, n(y_i)$ *en clústeres* $C_1, \ldots, C_k$ *con el algoritmo k-means.*

Output: Clústers $A_1, \ldots\ldots, A_k$ con $A_i = \{j \mid y_j \in C_i\}$.

En el trabajo de Herrera (2011a), de la matriz *kernel* acumulativa que contiene las matrices de todas las características de la RDAP (incluyendo la de su grafo), se extraen los valores propios más altos (*eigenvalues*, en inglés), a partir de la cual se aplica, el previamente explicado algoritmo *k-means* para efectuar la partición en clústeres (o sectores) (ver Ilustración 5).

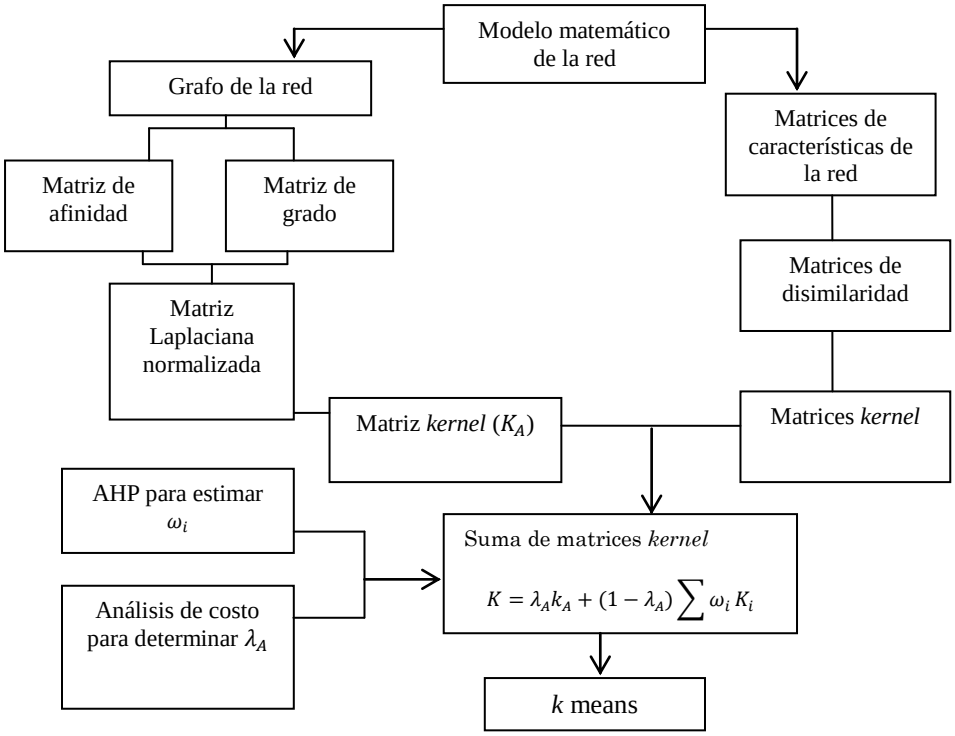


Ilustración 5: Etapas de clústering espectral (Herrera, 2011a)

En función de las restricciones con que se ejecute este algoritmo, este puede ser considerado parte de los métodos *semisupervisados* o parte de los métodos *no supervisados*. En el caso de Herrera (2011a), al fijar restricciones para establecer tanques que pertenecen a sectores separados, sin establecer *a priori* la pertenencia de los nodos a uno u otro sector, el algoritmo se emplea como un método *semisupervisado*; no obstante, tal y como se explicará más adelante, en el método propuesto en este trabajo no se aplica ningún tipo de restricción, con lo cual el algoritmo se trata como un método *no supervisado*.

5.5.3.2.5 Clústering Jerárquico

El clústering jerárquico es una herramienta de análisis de datos que se basa en la construcción de clústeres de los mismos 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 datos. La idea, tras este, es construir

árboles binarios que sucesivamente se fusionen en grupos en dependencia 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.

Existen dos tipos de procesos de clústering jerárquico, según el entorno desde el que se inicie éste. Si el proceso se inicia en un único clúster que contiene todos los casos agrupados como un sólo conjunto, se dice que se trata de un proceso de clústering divisivo (de arriba a abajo). En él, el clúster general se va subdividiendo hasta llegar a un nivel en donde cada una de los casos conforma un clúster (o *singleton*). El segundo procedimiento, conocido como clustrering de aglomeración en nido (*aglomerative nesting clustering)* sigue la ruta contraria; en él, a partir de los *singleton*, los casos se van agrupando, hasta llegar a un nivel en donde se forma un único clúster (de abajo hacia arriba) (Cimiano *et al*., 2004). De estos dos métodos, el método aglomerativo suele ser empleado más comúnmente debido a que tiene menor grado de complejidad. El clústering jerárquico divisivo tiene la desventaja de que cuando el número de datos en los que se hace la partición es muy grande, es computacionalmente costoso examinar todas las posibles particiones. Por esta razón, se suele recurrir a métodos heurísticos, que luego puede conducir a imprecisiones en los resultados (Han *et al*., 2012).

5.5.3.2.6 Pasos de clústering Jerárquico aglomerativo

5.5.3.2.6.1 Paso 1: Matriz de disimilaridad

Partiendo de un conjunto de datos finitos dados (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* × *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. En caso de que cada caso cuente con más de una variable, la comparación (*disimilaridad*) se hace entre los valores que toma cada variable individual para cada caso. Si se emplea 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 *disimilaridad* obtenidas. Si se empleara como métrica la distancia tipo *manhattan*, la distancia correspondería a aquella que existiría entre cada par de casos si siguiera 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 mezclas de variable cuantitativas y cualitativas (o categóricas).

- *Distancia euclidiana*

$$d_E(P,Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

- *Distancia Manhattan*

$$d = \sum_{i=1}^{n} |X_i - Y_i|$$

- *Distancia Gower*

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

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

donde:
P_1 es el conjunto de variables cuantitativas o cualitativas (o categóricas)
P_2 es el número de variables binarias
P_3 es el número de variables cualitativas (no binarias)
a es el número de coincidencia (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)
G_h es el rango (o recorrido) de la (h-ésima) variable cualitativa

5.5.3.2.6.2 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*, agrupación *individual*, y finalmente, agrupación basada en un *centroide* (Everitt *et al.*, 2011). A pesar de ello, existen otros métodos que se pueden emplear, tal como el conocido método de la mínima varianza (de *Ward)*, el método de la *mediana*, el método de *máxima probabilidad de igual varianza* (o *EML*), el método de *McQuitty* y el método *flexible-beta*. A continuación, se hará una explicación de los cuatro métodos más reconocidos, que serán los empleados en el ejemplo que se presenta al final de este documento.

Partiendo de una matriz *n* × *n* que contiene los distintos valores de *disimilaridad*es 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.

5.5.3.2.6.2.1 Método de aglomeración basado en el promedio

En este método, las agrupaciones se van dando, según la distancia de cada uno de los elementos dentro de un clúster con respecto al promedio de los elementos de otro clúster, o dicho de otra manera, la distancia promedio entre par de observaciones. Este, junto con el método basado en *centroide* está pensado para datos espaciales consistentes en medidas escaladas en intervalos (Chen y Xu, 2003).

$$D_{KL} = \frac{1}{n_k \ n_l} \sum_{ieC_K} \sum_{ieC_E} d(x_i \ x_j)$$

donde:
k y *l*: representan dos clústeres

El hecho de que tenga en cuenta todos los casos dentro de un clúster en lugar de uno sólo, hace que este método tienda a formar clústeres con un varianza pequeña, con un pequeño sesgo a formar clústeres con la misma varianza interna; no obstante, esto mismo hace que se vea menos influenciado por datos extremos en comparación a otros métodos.

5.5.3.2.6.2.2 Método de aglomeración basado en el 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*.

$$D_{KL} = \left|\left|prom\, X_k - \ prom\, X_l\right|\right|^2$$

Este método tiende a formar clústeres pequeños, con poca varianza intra-clúster.

Dado que en este método se comparan los *centroides* de los clústeres, los valores atípicos del conjunto de datos tienden a distorsionar el resultado; no obstante, la distorsión que estos causan en el método es menor que la que se genera si se emplea el método *completo* (Mooi y Sardtedt, 2011).

5.5.3.2.6.2.3 Método de aglomeración completo

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

$$D_{KL} = Máx_{I \in C_k \, J \in C_L} d(x_i, x_j)$$

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. 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).

5.5.3.2.6.2.4 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".

$$D_{KL} = Min_{I \in C_k \, J \in C_L} d(x_i, x_j)$$

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).

5.5.3.2.6.3 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*. Se trata de un diagrama sobre un plano X-Y, que muestra mediante líneas y columnas los enlaces jerárquicos que se van formando entre los casos. En el eje X se encuentran todos los casos y en el eje Y la altura de enlace. El objetivo de un *dendrograma* es representar las agrupaciones de los casos a manera de árbol, dónde el clúster más grande (en el que se agrupan todos los casos) es el tronco, y los rectángulos que unen dos o más clústeres de menor tamaño, son las hojas. Las líneas horizontales de estos rectángulos son denominados *cladas*, y se forman en la ordenada correspondiente a la altura de enlace con la que se aglomeran los clústeres por debajo de ella.

5.5.3.2.6.4 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 un método seleccionado para medir la *disimilaridades* entre casos (*euclidiana*, *manhattan*, *gower*) y un método seleccionado para definir la unión entre dos clústeres (*promedio*, *completo*, *individual*, *centroide*). 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 puede recomendar. Una medida bastante aceptada para evaluar el mejor camino a seguir es el CPCC[9], que ha sido ampliamente utilizada en estudios de clasificación fenética (Farris, 1969; Gonçalves *et al*., 2008; Podani y 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

[9] Cophenetic correlation coefficient o Coeficiente de correlación cofenética

CPCC en un conjunto de datos que han 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 y Rohlf, 1962).

A continuación se presenta la ecuación para la obtención del índice de CPCC.

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

donde:

x: son los valores de disimilaridades en pares S_{ij} de de la matriz de disimilaridad

y: son los valores de altura de la matriz ultramétrica

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 un lineamento en torno 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).

5.5.3.2.6.5 Ejemplo de clústering jerárquico

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

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

En un primer momento 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 obtendrán 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} , por ende el primer valor de la matriz *ultramétrica*.

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

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

donde:

S': representa el nuevo valor que tendrá la matriz n x *n para la comparación de cada elemento (k) con el clúster* $\{i, j\}$*. Dicho valor se obtiene a partir de una función seleccionada* (f) *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.

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

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). 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.

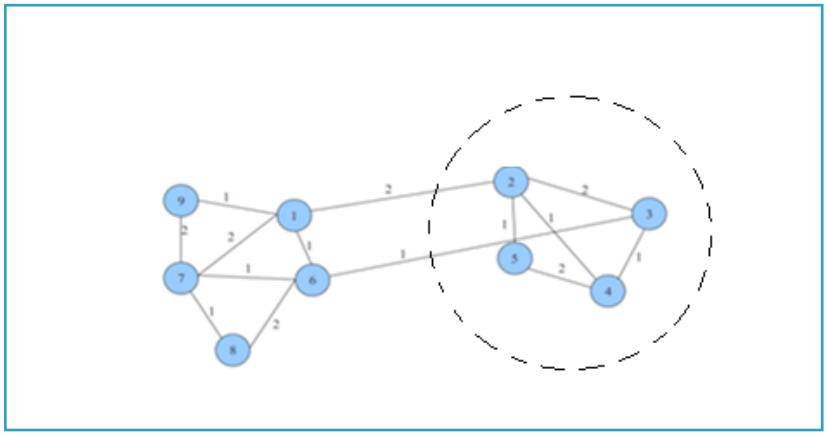


Ilustración 6: Ejemplo de aplicación de clústering jerárquico

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 4: 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 disimiliaridad por pares entre los casos, se emplea como métrica la distancia *euclidiana*.

	2	**3**	**4**	**5**
2	0.00			

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

Tabla 5: Matriz de *disimilaridad* de ejemplo de clústering jerárquico

En la matriz de *disimilaridad* obtenida (Tabla 5), 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.

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

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

La nueva matriz *n* × *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 6: 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

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

La matriz *ultramétrica* queda de la siguiente forma.

	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 7: Matriz *ultramétrica*

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

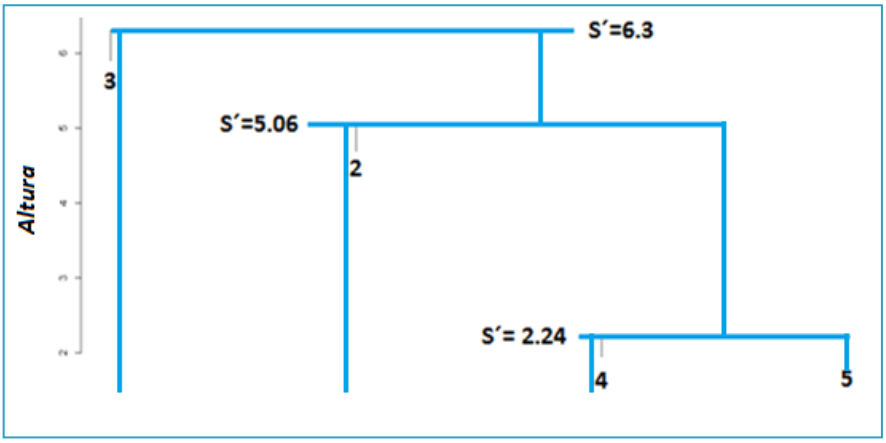


Ilustración 7: Valores de altura en el *dendrograma*

5.5.3.2.6.6 Selección del número de clústeres apropiado

La técnica descrita previamente, si bien puede mostrar de una manera visualmente clara la estructura de los datos, no indica cuál podría ser una partición que represente más apropiadamente las características de los datos y que a su vez preserve el concepto de clúster que ya se ha descrito previamente. Dicho de otra manera, un *dendrograma*, muestra los distintos pasos de la agrupación de casos hasta llegar a un único clúster en el que se agrupen todos ellos; no obstante, de ninguna manera indica qué nivel de altura es el apropiado para establecer una línea de corte por debajo de la cual se puedan identificar los clústeres que conserven de mejor manera el concepto de clúster. En realidad, este es un tema que hoy por hoy sigue siendo de amplia discusión. Pese a que se han propuesto muchos indicadores, hasta la actualidad no se ha logrado un consenso definitivo

en torno a cuál es la medida más exacta para la definición de una buena partición. Existen métodos que tienen un nivel de precisión bastante alto; sin embargo son computacionalmente muy costosos, en tanto los métodos más sencillos de aplicar, algunas veces, no arrojan resultados con el nivel de precisión necesario para tomarlos como concluyentes. En este sentido, uno de los más aceptados hoy por hoy para evaluar particiones en grafos es el índice de modularidad. Este ha sido definido por Newman y Girvan (2004), como el número de aristas de un grafo que caen dentro de un grupo, menos el número esperado en una red equivalente con las aristas colocadas aleatoriamente. Cuantifica la idea de que la verdadera estructura de la comunidad es un arreglo estadístico de aristas (Newman, 2006) y puede tomar valores negativos y/o positivos, siendo los valores positivos más altos, los característicos de las mejores particiones de una red; es decir, indica si una comunidad en un red es distintiva o no (Zhang y Wang, 2008).

A continuación se discute una serie de medidas que se pueden combinar con la técnica clústering jerárquico para determinar la mejor partición del conjunto de casos (vértices con las características relevantes para el proceso de sectorización). Primero se presenta una validación interna de clústeres, que incluyen tres parámetros cuya optimización puede servir de lineamiento para establecer una partición óptima. Luego se discute sobre un criterio gráfico conocido como criterio del codo, cuya estimación parte de graficar la escala de altura de un *dendrograma* 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 *dendrograma*, y que se puede combinar con el criterio del codo para establecer un número de clúster distintivos. Se finaliza la discusión con un método más complejo propuesto por (Shimodaira, 2004a), que contempla un remuestreo interno multiescala (*multiscale boostrap resampling*), el cual arroja un indicador de Sesgo Aproximado (AU, Approximately Umbiased) mediante el cual también se puede estimar el número de clústeres distintivos dentro de un *dendrograma*

5.5.3.2.6.6.1 Medidas Internas

Las medidas internas son una serie de parámetros (*conectividad*, *ancho de silueta* e *índice de Dunn*) que se calculan a partir de tres características generales del conjunto de clústeres en una partición dada, y se utilizan para evaluar la calidad de cada posible partición. Las características que se emplean son: *compactación*, *conectividad* y *separación*. Su objetivo final es poder determinar en qué estado de la partición (número de clústeres) estos parámetros alcanzan un valor óptimo. La *compactación*, por un lado, evalúa la homogeneidad de los clústeres, mientras la *separación*, por el contrario, a 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.

5.5.3.2.6.6.1.1 Conectividad

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 siguiente expresión; en esta expresión, cada comparación 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 y ∞, denotando los valores más cercanos a 0 un mayor nivel de *conectividad.*

$$\mathrm{Conn}(\sigma) = \sum_{i=1}^{N} \sum_{j=1}^{L} \chi_{i,nn_{i(j)}}$$

En la Ilustración 8 se puede ver el resultado del cálculo del parámetro para diferentes particiones en una red de 64 nodos. En ella se aprecia que este valor se optimiza (valor de *conectividad* cero) cuando la partición se hace en 2-3 clústeres.

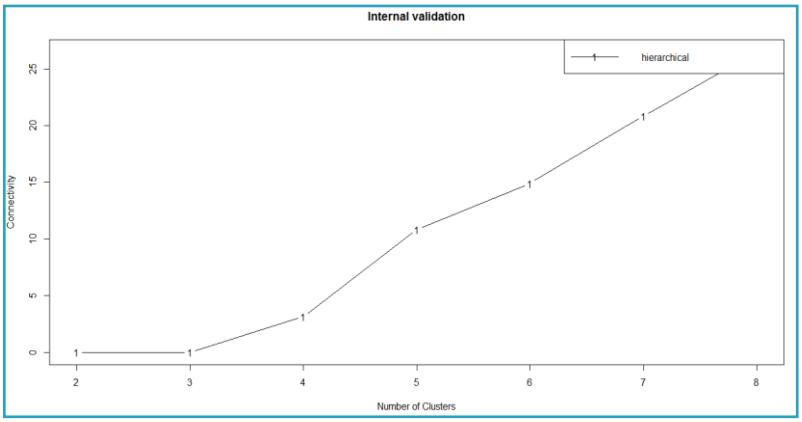


Ilustración 8: Optimización del número de clústeres con base en el índice de conectividad

5.5.3.2.6.6.1.2 Ancho de Silueta

Este método fue propuesto por Rouseeeuw (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 (de *ancho de silueta*) de cada conjunto de casos dentro de un clúster. En el caso de los clústeres mejor conformados, el valor del *ancho de silueta* será próximo o igual a 1 y en el caso de los peores conformados, el valor será próximo o igual a -1. Este valor se emplea como indicador de confiabilidad de una partición dada. Su cálculo se realiza mediante la siguiente expresión:

$$S(i) = \frac{b_{ij} - a_i}{\max(b_{ij}, a_i)}$$

En donde a_i es el promedio de las distancias entre todos los caso que pertenecen al mismo clúster, y b_{ij} también es un promedio de distancia, pero de los casos i en un clúster, con respecto a los casos j del clúster más cercano.

En la Ilustración 9 se puede apreciar que el valor óptimo de este parámetro se

alcanza para una partición compuesta por 3 clústeres.

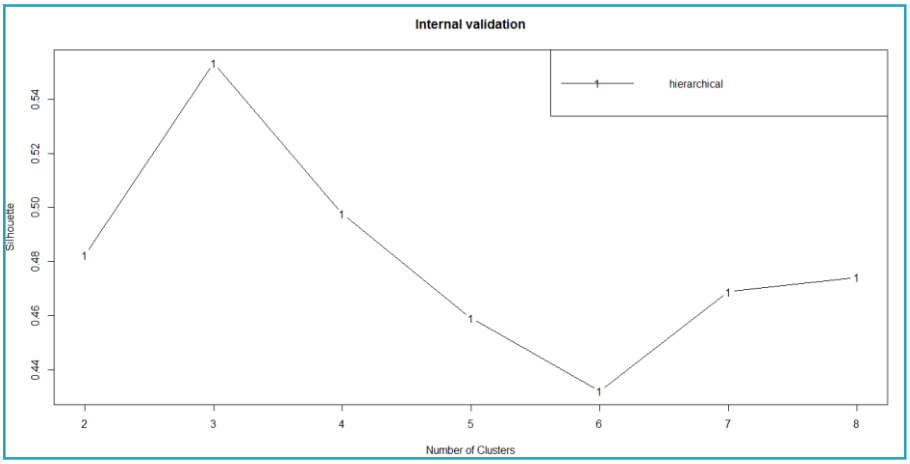


Ilustración 9: Optimización del número de clústeres con base en el *ancho de silueta*

La gráfica anterior se forma a partir del conjunto de resultados que produce el análisis del *ancho de siluetas* para cada partición $\sigma_{2\ldots n}$. A continuación, se muestra el gráfico de *ancho de silueta* para una partición de 2, 3 y 4 clústeres (ver Ilustración 10, 11 y 12). Cómo se puede ver en ellas, en el caso de la partición con tres clústeres, el promedio del valor del *ancho de silueta* es mayor. También, se hace notorio que en el caso de la partición en cuatro clústeres, se conforma un clúster con un único miembro, por lo que el valor de *ancho de silueta* para este caso es 0, siendo la partición no válida.

Esta medida, fue empleada por Herrera (2011a) para evaluar los clústeres de RDAP generados mediante *clústering espectral*.

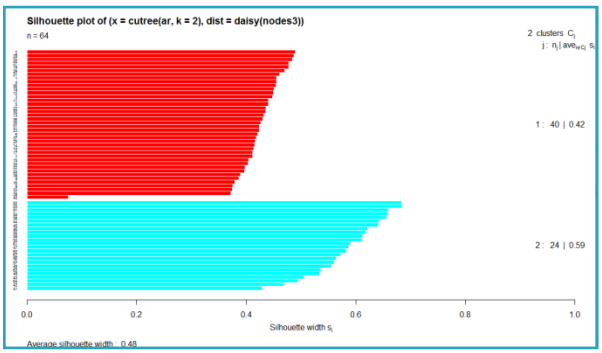


Ilustración 10: *Ancho de silueta* para una partición de dos clústeres

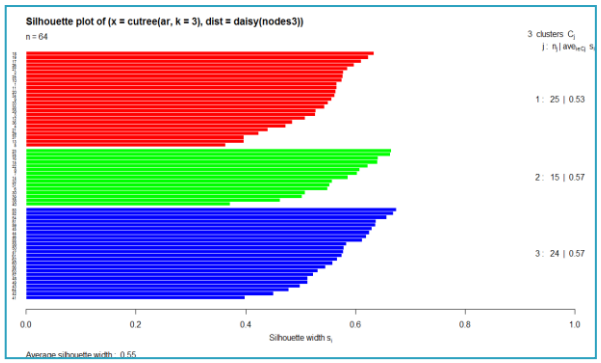


Ilustración 11: *Ancho de silueta* para una partición de tres clústeres

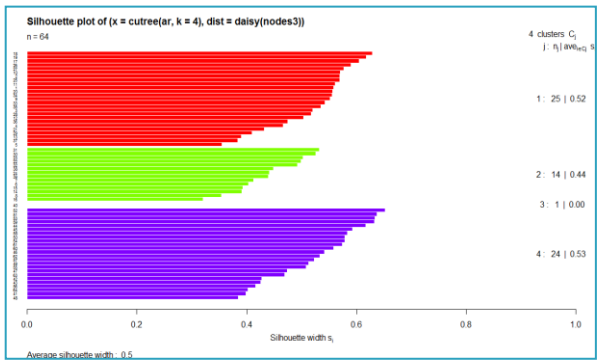


Ilustración 12: *Ancho de silueta* para una partición de cuatro clústeres

5.5.3.2.6.6.1.3 Índice de Dunn

Este índice, propuesto por 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 siguiente

expresión:

$$D(i) = \frac{\min\limits_{c_k c_i \in c_{k \neq} c_i} \left(\min\limits_{i \in c_k, j \in c_i} dist(i,j) \right)}{\max\limits_{c_m \in i} diam(C_m)}$$

donde:

$diam(C_m)$ 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 ∞, teniendo la mejor partición en el punto en que este se maximice.

En la Ilustración 13, se puede apreciar que el valor óptimo de este parámetro se alcanza para una partición compuesta por 3 clústeres.

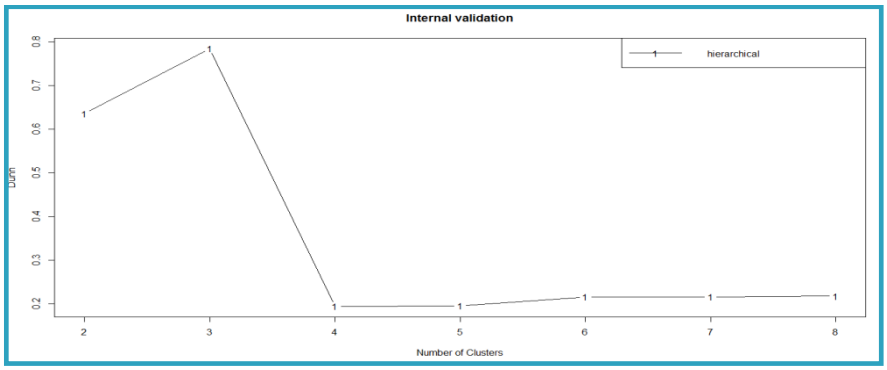


Ilustración 13: Optimización del número de clústeres con base en el *índice de Dunn*

5.5.3.2.6.6.1.4 Criterio del codo

Una de las maneras más simples de poder estimar un número de clústeres que represente una óptima partición 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 representado todos los posibles valores de altura (S´) que puedan tener todas las *cladas* del *dendrograma* (o porcentaje acumulado de este valor) y el otro eje contenga el número de clústeres posibles (1...$k_{2...n}$). En esta gráfica se suele

producir un *codo*. Más allá de este *codo*, cualquier nueva fusión se produce a una distancia mucho más pequeña, de manera que el número de clúster antes de esta fusión es la solución más probable (Mooi y 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 modelización de los datos. Pese a la facilidad que representa este método, el mismo presenta como desventaja el hecho de que no en todos los casos aparece un codo.

En la Ilustración 14 se muestra el porcentaje acumulado del total del valor de altura, asociado a cada una de las particiones, en la medida 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), llegado un punto en que se forma un codo, para una partición de tres clústeres.

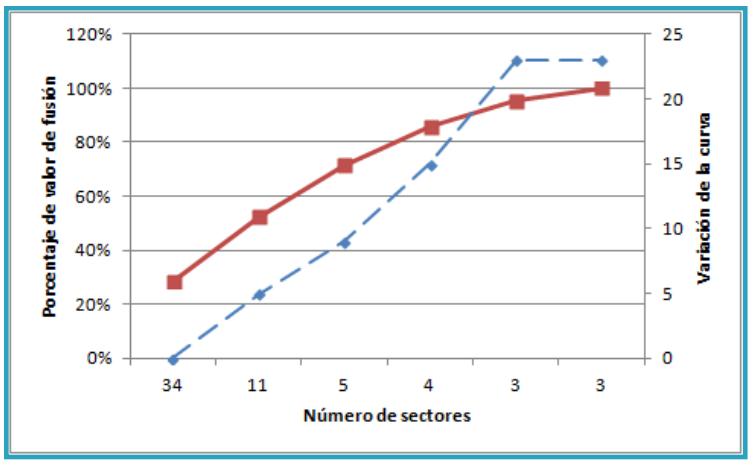


Ilustración 14: Selección del número de clústeres con base en el criterio del codo

5.5.3.2.6.6.1.5 Criterio de inconsistencia

Este método parte de etiquetar cada una de las *cladas* del *dendrograma* con un valor denominado *inconsistencia*, que sirve como medida que indica que tan parecido son dos 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 clúster apropiado aquel indicado por el codo de la gráfica.

Para calcular la *inconsistencia* de un clúster $\varphi_{k'}$, se parte de una matriz *ultramétrica* (la de la Tabla 7 por ejemplo). De ella se extrae el valor de altura con el que se formó cada nuevo clúster.

$$S'_{(5-4)} = 2.24$$

$$S'_{(5-4-2)} = 5.02$$

$$S'_{(5-4-2-3)} = 6.31$$

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.*

$$\overline{S'}_{5\text{-}4} = 3.63$$

$$\overline{S'}_{5\text{-}4\text{-}2} = 4.52$$

El valor de *inconsistencia* 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 15 se muestra el *dendrograma* del ejemplo anterior con los valores de *inconsistencia* correspondiente a cada uno de los clústeres.

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

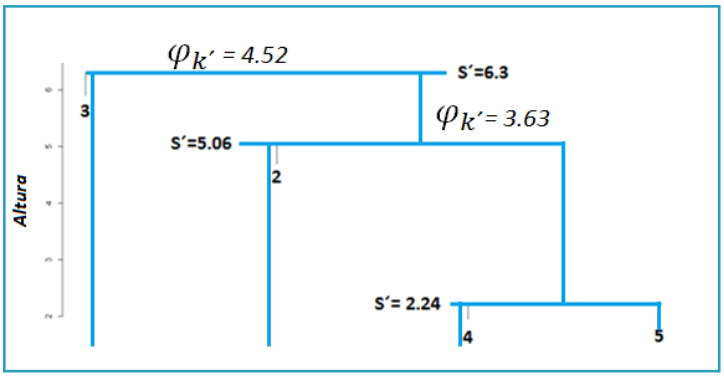


Ilustración 15: Valores de *inconsistencia* correspondiente a cada *clada* del *dendrograma*

5.5.3.2.6.6.1.6 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 *dendrograma* (Suzuki y Shimodaira, 2006). La validación la hace mediante dos valores probabilísticos o *p-values*, Aproximadamente Insesgado (AU, *Approximately Umbiased*) / Probabilidad de Remuestreo (BPr, *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 en que en el primero se varía la escala (tamaño) de la muestra (*remuestreo multiescala*), en tanto 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* BPr. El uso de ambos valores ayuda a estimar el nivel en que los clústeres están respaldados por los datos (Shimodaira, 2004b). Estos, pueden tomar valores entre 0 y 1. Estableciendo un grado de significancia – α –, se puede hacer una prueba de hipótesis, en la que se rechacen los clústeres con un valor inferior a este α.

La idea del método del *remuestreo multiescala* fue abordada inicialmente por

Efron *et al*. (1996). En este caso, se toman muchas muestras de conjunto de datos (*remuestreo*); se forma un *dendrograma*; 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* BPr. Luego, para calcular el valor *p-value* AU, se utiliza el mismo proceso, pero ahora se altera el tamaño de la muestra. El *p-value* AU, es calculado 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 *dendrograma* 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 su clada superior. Las cladas que tenga 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 las Ilustraciones 16 y 17, se muestra cómo se pueden identificar cuatro clústeres en una red con una sola fuente de abastecimiento, usando la herramienta *pv-clust*.

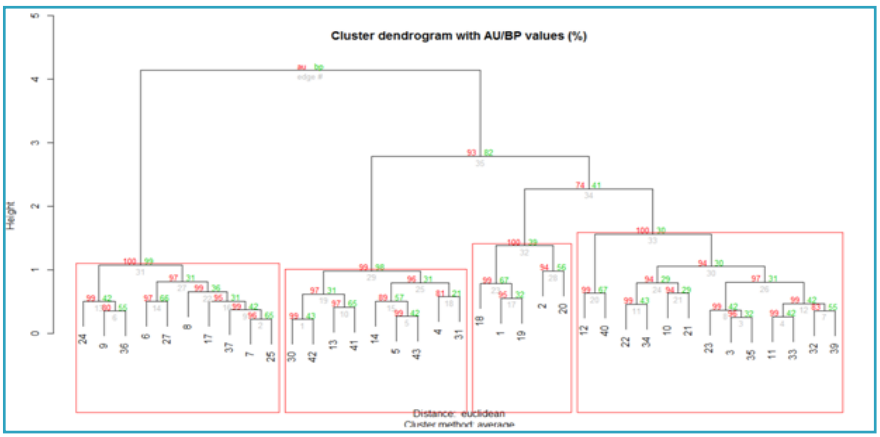


Ilustración 16: Definición de sectores válidos mediante *p-values*

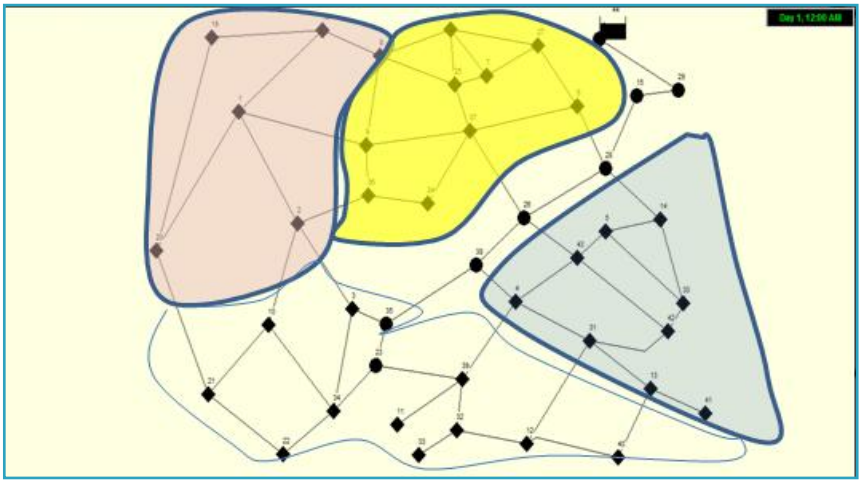


Ilustración 17: Sectores válidos definidos mediante *p*-values

Pese a la precisión que ofrece este método, tiene como desventaja el tiempo que emplea para hacer el cálculo. Un remuestreo de 10,000 iteraciones puede tomar muchas horas, en función de la cantidad de nodos de la que esté compuesta la RDAP.

5.5.3.2.6.7 Decisión final en torno al número de sectores

En la práctica, los seis índices descritos previamente pueden arrojar un resultado distinto con respecto al número de sectores a seleccionar; sin embargo, todos los resultados forman un rango a partir del que se puede tomar una decisión. En particular, la obtención de *p-values* puede tomar mucho tiempo, que puede estar muy bien justificado si el resto de los índices dieran resultados muy dispares. El mejor camino a seguir es emplear una media de los resultados de todos los índice, incluyendo los *p-values*, pero calculándolos con un número bajo de iteraciones (ver Tabla 8).

Medida	Ventaja	Desventaja
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	
Criterio de codo	Rápido y sencillo	Algunas veces no se aprecia un codo claramente en la gráfica

Medida	Ventaja	Desventaja
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

Tabla 8: Ventaja y desventaja de los indicadores para evaluación de particiones en un dendrograma.

5.5.4 R studio

El programa R (o su entorno de trabajo R studio), es un ambiente-lenguaje de programación libre[10] que fue concebido con una orientación hacia el campo de la estadística. Es un proyecto GNU[11] de libre distribución con licencia pública general de la fundación de software libre. Este provee una amplia variedad de herramientas estadísticas, tales como modelización lineal y no lineal, test estadísticos clásicos, análisis de series de tiempo, clasificación, clústering, etc.

Los usuarios de R tienen acceso a una extensa librería[12] de paquetes que se pueden descargar de manera gratuita. Una vez descargados estos paquetes, se pueden ejecutar en el programa las rutinas deseadas, ya que cada paquete cuenta con funciones para efectuar dichas rutinas además de documentación explicativa. Gran parte de estas herramientas son desarrolladas por lo mismo usuarios. Para los análisis descritos previamente, R, ofrece los siguientes paquetes.

[10] http://www.r-project.org/
[11] The GNU free operating system.
[12] CRAN-Comprehensive R Archive Network (red de archivos de R).

Paquete	**Función**
Kernlab (Karatzoglou *et al*., 2004)	*Clústering Espectral, análisis kernel*
Igraph (Csardi y Nepusz, 2006)	*Análisis de grafos*
Cluster (Maechler *et al*., 2013)	*Matrices de disimilaridad*
Clvalid (Brock *et al*., 2008)	*Medidas de validación de clústering jerárquico*
Stats (RCT, 2013)	*Clústering Jerárquico*
pvClust (Suzuki y Shimodaira, 2006)	*Remuestro multiescala para obtención de pv-value AU/BP*

Tabla 9: Paquetes de R empleados

5.5.5 Aplicación de AHP [13] para ponderar aspectos hidráulicos

AHP es una metodología para evaluar alternativas cuando se tienen en consideración varios criterios. En este caso, las alternativas están referidas a las características almacenadas en los nodos de las RDAP: c*oordenadas geográficas (x,y), demanda y elevación, coeficiente de emisor*. El método está basado en el principio de que la experiencia y el conocimiento de los actores son tan importantes como los datos utilizados en el proceso (Osorio y Orejuela, 2008). Dentro del mundo de la hidráulica, este método ha sido empleado por Delgado-Galvan *et al*. (2010) para realizar un análisis de las opciones de gestión de fugas en una RDAP, incluyendo dentro del análisis, externalidades sociales y

[13] Analytic Hierarchy Process o Proceso de Análisis Jerárquico.

ambientales asociadas a las fugas. El método presenta la ventaja de permitir incluir además del punto de vista del operador de la red, el punto de vista de otros actores, tales como los usuarios, organizaciones sociales, etc.

El proceso de AHP se compone de los siguientes pasos (Kasperczyk y Knickel, 2006):

- *Se estructura el problema de decisión y se seleccionan los criterios*
- *Se establecen prioridades de los criterios por comparación de pares (comparando pesos)*
- *Se hace comparación por parejas de las opciones de cada criterio (puntuación)*
- *Se obtiene una puntuación general relativa por cada opción.*

Para hacer una comparación entre las alternativas y opciones, se emplea una escala numérica (ver Tabla 10) que sirve para indicar cuánto más importante es un elemento sobre otro elemento con respecto al criterio o propiedad con el cual está siendo comparado.

Intensidad de Importancia	**Definición**	**Explicación**
1	Igual importancia	Las dos actividades contribuyen de manera igualitaria al objetivo
2	Débil o leve	
3	Importancia moderada	Experiencia y juicio levemente a favor de una actividad sobre otra
4	Más moderado	
5	Importancia fuerte	Experiencia y juicio fuertemente a favor de una actividad sobre otra
6	Más fuerte	
7	Muy fuerte o importancia demostrada	Una actividad se favorece fuertemente sobre otra; su predominancia está comprobada en la práctica
8	Muy, muy fuerte	

Intensidad de Importancia	Definición	Explicación
9	Importancia extrema	La evidencia que favorece una actividad sobre la otra es del orden de afirmación y asunción razonable máxima posible
Reciprocidad de los de arriba	Si la actividad *i* cuando se compara con la actividad *j* tiene asignada alguno de los números no nulos de arriba, entonces *j* tiene un valor reciproco cuando se compara con *i*	
1.1-1.9	Si las actividades son muy cercanas	

Tabla 10: La escala absoluta de números absolutos (Saaty, 2008)

El AHP utiliza comparaciones entre pares de elementos, construyendo matrices a partir de estas comparaciones, y usando elementos de álgebra matricial para establecer prioridades entre los elementos de un nivel, con respecto a elementos de un nivel inmediatamente superior (Osorio y Orejuela, 2008). Para el caso de una RDAP, a partir del método, se obtiene un vector de prioridad en el que se representan los distintos pesos (importancia) para cada una de las características de la red que se emplean para llevar a cabo su partición en sectores. Una de las propiedades que se toma en cuenta para validar esta técnica es la *transitividad*, la cual se refiere a la consistencia (o el orden lógico) que se conserva a la hora de juzgar los elementos de una muestra. Esta se mide mediante un índice de inconsistencia (*IIC*). Dado que las evaluaciones son hechas por humanos, existe un nivel de *inconsistencia* que puede ser aceptado. Para definir este nivel, se compara el *IIC* con un índice de inconsistencia aleatoria (*IA*) medido en matrices generadas por *remuestreo* a partir de la matriz inicial (boostrap resampling). El resultado de esta comparación es un ratio de inconsistencia (*RI*).

$$RI = \frac{IIC}{IA}$$

El valor de *RI* no puede ser superior al 10% para poder definir que el nivel de *inconsistencia* es aceptable. En caso contrario, se hace necesario hacer una

revisión de las comparaciones.

5.5.6 Fugas en las RDAP

5.5.6.1 Aspectos económicos de fugas en RDAP

Se han resumido los aspectos principales para el mal funcionamiento de las RDAP en algo que se denomina como el círculo vicioso (Farley *et al*., 2008), en el cual, las pérdidas físicas desvían el agua potable del camino hacia los consumidores, aumentando los costos de operación. Esto, a su vez, hace que los operadores se vean en la necesidad de aumentar la capacidad del sistema, para lo cual se tienen que hacer inversiones más elevadas que las necesarias para operar el sistema bajo un buen esquema de gestión. En adición, las pérdidas comerciales, causadas por imprecisiones de los equipos de medición, mal manejo de datos y conexiones ilegales, reducen los ingresos y por ende la generación financiera de recursos. El reto para los operadores de agua es lograr transformar el círculo vicioso en un círculo virtuoso, en el cual, la reducción de las pérdidas físicas reduzca la necesidad de explotar nuevas fuentes y mejore las finanzas, ya que la reducción de pérdidas por sí sola, implica mayor disponibilidad del recurso.

El primer paso para la definición de una estrategia para reducir pérdidas físicas en los acueductos es definir una meta de reducción, en la que se tengan en cuenta otras metas de la compañía operadora y/o las políticas que se complementarán o entrarán en conflicto. Adicionalmente, es importante la presencia de un ente regulador que fije metas de desempeño. Pese a la importancia que revisten las fugas en las RDAP, la definición de metas en torno a su reducción sigue sin seguir un rigor económico técnico apropiado. En algunos casos, ni si quiera se considera si las metas son alcanzables.

Más allá de los costes de marginales (es decir, los costes que varían con la producción del agua), que suelen tenerse en cuenta en el estudio del nivel económico de fugas, existen otros costes, que no están directamente asociados a la operación de las RDAP; no obstante, representan un coste que es cubierto al final

de manera indirecta por los usuarios. Este tipo de coste, puede ser el que representan los daños causados por la presencia de una fuga sobre las infraestructuras urbanas, las emisiones de CO_2 generado por los equipos de bombeos en la red, así como el que se asocia al uso de energía eléctrica en las plantas potabilizadoras, o el coste de oportunidad asociada a la explotación de un nuevo cuerpo de agua. Tales costes se denominan externalidades, y pese a que su valoración es más compleja que la de los otros costes de operación mencionados previamente, su inclusión en la toma de decisiones respecto al nivel económico de fugas es importante.

En la Ilustración 18, se muestra lo que ha sido definido como los cuatro pilares de la gestión de fugas en una RDAP. En ella, se puede ver que existe un volumen de pérdidas que no es económicamente viable recuperar. El coste de este volumen es igual al coste de las inversiones para su reducción, tal y como se plantea en la Ilustración 1, en donde se aprecia como las inversiones en gestión de fugas sigue una ley de rendimientos decrecientes. La limitación de este modelo, radica en que los costes asociados a las pérdidas de agua, sólo toman en cuenta el coste marginal de esta, sin tomar en cuenta las externalidades asociadas a la presencia de fugas. De tener en cuenta este tipo de coste, evidentemente, el coste del agua se incrementa, con lo cual el punto de equilibrio se sitúa en un punto más a la izquierda de la Ilustración 1, que a su vez representa la necesidad de mayor inversión en gestión.

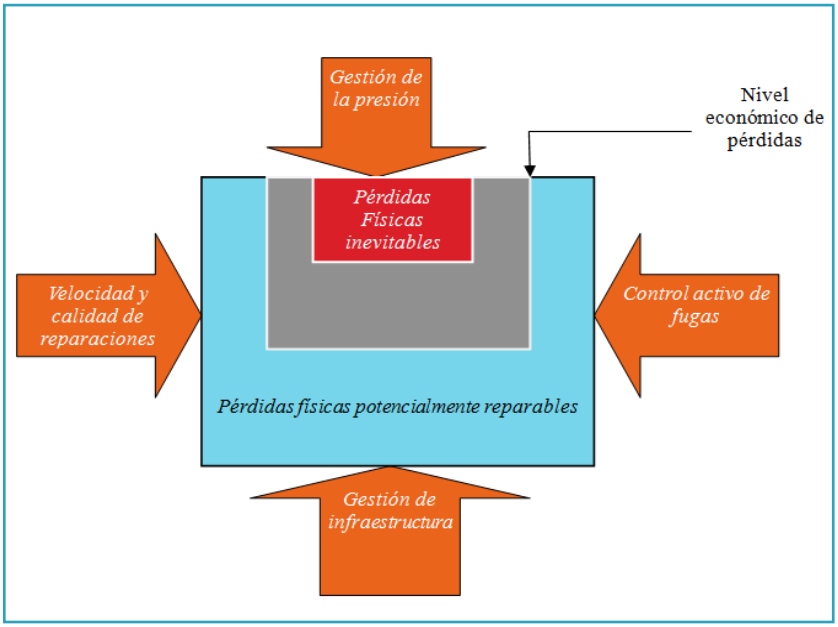


Ilustración 18: Los cuatro pilares de la gestión de fugas (Pilcher *et al.*, 2007)

5.5.6.2 Modelización matemática de fugas en una RDAP

El caudal de fuga en una tubería, es igual al caudal que circula por la misma multiplicado por un factor de fugas.

$$Q_f = f * Q_d$$

Donde:

$f = Fracción\ fugada$

$Q_d = Caudal\ demandado$

El caudal de fugas, puede ser calculado multiplicando un coeficiente de fugas por la presión elevada un exponente N_1. Antes de 1994, se suponía que el caudal fugado a través de una ruptura en una tubería variaba de manera directamente proporcional a la presión elevada al cuadrado. La teoría FAVAD (Mays, 2000), demostró que, pese a que algunas tuberías siguen este modelo de área fijo (tal es el caso de las tuberías metálicas), otros tipo de tuberías (*pvc* o polietileno), el exponente puede varia hasta 2.5. Para áreas tales como orificios en las paredes de las tuberías metálicas, el exponente N_1 es 0.5; en áreas que varían a lo largo de un

eje, como las fugas en uniones o accesorios, el exponente N_1 es 1.5 y para áreas que varían a lo largo de dos ejes, como grietas en tuberías de plástico, el exponente N_1es 2.5 (Pearson *et al*., 2005).

$$Q_f = C_f * P^{N_1}$$

La variable C_f , representa el coeficiente de fugas

A partir de las dos ecuaciones anteriores, se puede estimar el factor de fuga de una tubería.

$$f = \frac{C_f (P)^{N_1}}{Q_d}$$

Determinar el volumen total de fugas resulta complicado, pero mientras se consideren de forma más precisa los factores que influyen en el cálculo, se podrá tener una estimación más cercana a la realidad.

En EPANET 2.0, para representar las fugas se emplean dos variables en los nodos. La primera es el coeficiente de fugas, que en EPANET lleva el nombre de coeficiente de emisor y el exponente de fugas N_1, que en EPANET 2.0 lleva el nombre de exponente de emisor. En concreto, el coeficiente de emisor hace la función de un orificio en las tuberías, el cual descarga a la atmósfera (Rossman, 2000). El caudal que circula por un emisor, es proporcional a la presión disponible en el nodo, afectado por el exponente que se relaciona con el tipo de ruptura.

La desventaja de modelizar las fugas a través de emisores, radica en que EPANET 2.0 sólo permite incluir un único exponente para toda la red, no logrando hacer una distinción entre los distintos materiales de tuberías que pueden estar disponible en una RDAP. También, al igual que en el caso de la demanda, la correcta asignación de un valor de emisor, depende de la disponibilidad de datos de medición de caudal que sean fiables. El proceso de asignación parte de lograr separar el consumo que no depende de la presión (el consumo de los usuarios) y el

consumo dependiente de la presión (el consumo de fugas representado a través de los emisores). A continuación se describe una metodología para asignar emisores en una red en dónde los otros parámetros de calibración han sido apropiadamente ajustados.

5.5.6.3 Método para asignar coeficientes de fugas en EPANET 2.0

Este método consiste en asignar un coeficiente de emisor a todos los nudos de la red a fin de lograr simular el caudal de fugas. Para ello, se calcula un coeficiente de emisor medio para toda la red con base en el caudal de fugas y la presión promedio en el punto representativo de cota media de la red.

$$Ce = \frac{Q}{P_{(media)}^{N_1}}$$

Luego, el coeficiente de emisor se reparte a través de todos los nodos, según el criterio que se estime conveniente. Una manera relativamente común de proceder, cuando no se tienen buenos datos de medición de caudal, es hacerlo de manera uniforme entre todos los nodos. Otra manera, es distribuirlos en proporción a la longitud de cada tubería; para ello, el coeficiente de emisor obtenido para toda la red se multiplica por la media ponderada de la longitud correspondiente a cada nodo. Para obtener la media ponderada de longitud para cada nodo, se suma la media de las longitudes de todas las tuberías que llegan a cada nodo, y se divide entre la longitud total de la red.

$$MPL_{nodo\ i} = \frac{\sum L_i/2}{L_n}$$

En donde:

$L_i = Longitud\ de\ todas\ las\ tuberías\ que\ llegan\ al\ nodo\ i$

$L_n = Longitud\ de\ todas\ las\ tuberías\ de\ la\ red$

Al efectuar la asignación de su correspondiente coeficiente de emisor a cada nodo, se puede hacer la comprobación del caudal total del día; no obstante, la desviación

estándar de la curva de caudal medido será distinta que la desviación estándar de la curva de caudales de entrada en EPANET. Dicho en otras palabras, el volumen total medido en un periodo de tiempo (24 horas por ejemplo) es igual la suma de los caudales horario de entrada a la red de acuerdo al modelo matemático; no obstante, los dos tipos de caudales hora a hora son diferentes. Para poder igualar los caudales de la medición hora a hora, con los indicados hora a hora por el modelo matemático (igualar la desviación estándar de las dos curvas de caudales), se tiene que adaptar la curva del patrón de demanda, lo que se hace mediante un proceso iterativo, en el que se evalúa la presión inicial de la red anterior (con los coeficiente de emisor ya asignados), tomando como referencia el punto de elevación media de la red. A partir de los datos de presión se calculan los coeficientes horarios del patrón de demanda. Con este nuevo patrón de demanda se vuelve a realizar la simulación de la red y se vuelven obtener datos de presión a partir de los cuales se recalcula el patrón de demanda, y así sucesivamente, hasta que la diferencia entre las desviaciones estándar de las dos curvas tome un valor mínimo (aceptable).

5.5.7 Energía en las RDAP

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, así, 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).

$$P_{inp} = P_{out} + P_{int}$$

También, la potencia de entrada es equivalente a la suma de la potencia suministrada por bombeos y por embalses

$$P_{inp} = \sum_{e=1}^{n_e} Q_e * H_e + \sum_{j=1}^{n_p} P_J$$

Por otro lado, la potencia de entrega a los consumidores 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 y en otro se emplea la altura piezométrica máxima H_j^* requerida para satisfacer un requerimiento mínimo de presión.

$$P_{out}^{real} = \sum_{j=1}^{n_n} Q_j * H_j$$

$$P_{out}^{max} = \sum_{j=1}^{n_n} Q_j * H_j^*$$

Luego, 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 poderse expresar la potencia de entrega en dos términos (máximo y real) la potencia operacional también se puede expresar en los mismos dos términos.

$$P_{int} = P_{inp} - P_{out}$$

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

$$P_{int}^{max} = P_{inp} - P_{out}^{max}$$

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).

También, Todini (2000) presenta un índice de resiliencia (Ir) para evaluar la eficiencia energética de las RDAP. 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.

$$Ir = 1 - \frac{P_{int}^{real}}{P_{int}^{max}}$$

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

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

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

Con este índice se puede estimar qué tan fiable es la red ante la falla de uno de sus elementos. Éste 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, representaría una red sobredimensionada, que además de costosa, probablemente presentaría valores bajos de velocidad, que luego pueden desencadenar problemas de calidad de agua.

Sobre esto, Di Nardo *et al.* (2013a) presenta un índice de desviación de resiliencia que permite evaluar que tanto disminuye la resiliencia un nuevo esquema de una RDAP con respecto al que tenía originalmente. Es decir, evalúa, cuanta energía extra se disipa o no como consecuencia de la implementación de un nuevo

esquema de red. El índice en cuestión se calcula mediante la siguiente expresión.

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

En donde:

I_{rd} = Desviación del índice de resiliencia

I^*_r = Índice de resiliencia de la RDPA con el nuevo esquema

I_r = Índice de resiliencia de la RDPA con el esquema original

De acuerdo con Araque y Saldarriaga (2011), al minimizar 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 presiones. La definición de energía óptima disipada hace referencia a cuanta energía se espera que el sistema de distribución de agua potable disipe en cada uno de los tubos que la conforman. Para evaluar el grado de uniformidad de presiones de una RDAP, estos autores proponen el coeficiente que se presenta en la siguiente ecuación.

$$CU = \frac{\sum_{j=1}^{n} P_j}{n * max[P_j]}$$

En donde:

n = número de nodos de la red

P_j = presión en cada uno de los nodos de la red

En algunas RDAP (redes 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 una de presión mínima muy baja.

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 quedará definida con el Coeficiente de Pérdida de Potencia (*CPP*), que resulta mayor para los mejores esquemas de red sectorizada.

$$CPP = \frac{\sum_{i=1}^{n} Q_{i_{df}}^{*} * P_i^{*}}{\sum_{i=1}^{n} Q_{i_{df}} * P_i}$$

donde:

n = número de nodos de la red

$Q_{i_{df}}^{*}$ *= Caudal de demanda y fuga en cada nodo en el nuevo esquema de RDAP*

P_i^{*} *= presión en cada uno de los nodos de la red en el nuevo esquema de RDAP*

$Q_{i_{df}}$ *= Caudal de demanda y fuga en cada nodo en el esquema original*

P_i *= 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*.

5.6 Investigaciones Previas

A continuación se presentan más detalles de las investigaciones mencionadas en la *subsección 1.1.*

5.6.1 Verificación de planos de sectorización.

En Tzatchkov *et al*. (2008) se hace una descripción de los algoritmos empleados en SCARED para verificar esquemas de sectorización en RDAP. Tal y como ya ha sido mencionado previamente, para ejecutar el proceso de verificación de sectores se emplea la teoría de grafos. El objetivo es identificar las zonas de influencia de cada una de las fuentes de la red, por lo que al final sólo se pueden identificar sectores-aislados (sectores con una fuente de agua exclusiva). Para el caso de los grafos no dirigidos se emplea el método de pila, y a partir de allí se procede con un algoritmo BA o un algoritmo BP. De manera distinta se procede para el caso de los grafos dirigidos (digrafos), en el cual se emplea el método de EPANET 2.0 para calcular la concentración de una sustancia química conservativa a través de la red.

5.6.2 Herramientas de ayuda a la sectorización.

Vega (2012) describe una herramienta para sectorización implementada en el SIG GISRED desarrollado por el grupo REDISHP del Instituto del Agua y Medioambiente (IIAMA) en la Universidad Politécnica de Valencia (UPV). GISRED se ha concebido como una extensión de ArcView 3.2 que integra en dicho entorno el programa EPANET 2.0. El objetivo de esta aplicación es construir modelos de redes hidráulicas a presión directamente desde un SIG y analizar los resultados sobre el propio SIG, evitando así las incómodas conexiones que ofrecen otras aplicaciones (REDHISP, 2011).

En cuanto a la simulación del comportamiento de la red, GISRED genera un fichero con los datos de entrada requeridos por el simulador y se llama al módulo de herramienta (toolkit) de EPANET 2.0. El simulador proporciona un conjunto de resultados, que la aplicación se encarga de almacenar en su propia base de datos para consulta posterior.

La herramienta de sectorización presentada en este trabajo hace uso de la teoría de grafos y, a través de una serie de pasos consecutivos, es capaz de dividir la red en

sectores que también pueden ser sectores-aislados o cascadas de sectores. En una fase inicial, el procedimiento se basa en la metodología de sectorización descrita en la sección anterior, aplicando el algoritmo de determinación del área de influencia de las fuentes para definir los macrosectores de la red. Luego, dentro de estos, se establecen los sectores.

En esta herramienta se pueden emplear múltiples criterios para efectuar la división de la red, entre ellos, el número máximo de acometidas, el número máximo de abonados, demanda máxima, longitud de la red y estratos de presión. La selección de una zona como sector se basa en una medida acumulativa de estos criterios en los nodos y/o líneas de la red. Los pasos del método propuesto son: identificación de los macrosectores, identificación de la red arterial, creación de árboles dirigidos de mínimo coste, ordenación topológica y conjunto de corte de cada rama, creación de sectores.

5.6.3 Sectorización de redes existentes.

En Saldarriaga *et al*. (2008) se prueban cinco métodos de sectorización en cuatro redes distintas, empleando como criterio de evaluación el índice I_r de la RDAP (ver sección 5.5.7)

Las metodologías de sectorización que se emplean son:

- ***Sectorización mediante límites naturales***: Se aprovechan límites naturales para definir sectores
- ***Sectorización mediante pérdidas de energía***: Se divide la red a través de aquellas tuberías que presentan la menor pérdida de altura de presión o las que presentan menor diferencia de presión entre sus extremos.
- ***Sectorización mediante caudal***: Divide la red a través de aquellas tuberías que presentan el menor caudal.
- ***Sectorización mediante presión***: Genera los límites sobre la superficie de presión del modelo, buscando las líneas de máxima pendiente y haciendo la división del modelo a través de ellas.

- ***Sectorización mediante potencia***: Se divide la red a través de aquellas tuberías que presentan el menor valor de potencia unitaria. Siendo la potencia unitaria el resultado de multiplicar el valor de caudal de la tubería por el valor de la pérdida de altura piezomética que se da a través de ella.

El trabajo demuestra que todas las metodologías afectan en alguna medida al índice Ir de la red.

5.6.4 Partición de grafo para sectorización automática.

En Di Nardo *et al*. (2011b) se propone una metodología basada en técnicas particionales de grafos combinadas con índices de energía para encontrar los sectores más eficientes en términos energéticos. Para ello se emplea el algoritmo MLRB[14] implementado en el software METIS[15]. Con esta aproximación se generan particiones a diferentes niveles. Con base en criterios energéticos se selecciona la mejor partición y luego, siempre con criterios energéticos, se define cual de las líneas límites del sector corresponderá a la línea de alimentación del sector. Como resultado se logra una partición de sectores; no obstante, al no haber una definición clara de red de alta a priori, se formar cascadas de sectores.

5.6.5 Sectorización de sistemas de abastecimiento de agua.

En Di Nardo *et al*. (2013a) se emplea la metodología propuesta por Tzatchkov *et al*. (2008) para lograr conformar sectores-aislados, y a partir de los índices energéticos explicados en la sección 5.5.7 se redefinen las fronteras de los sectores, de tal manera que se logre maximizar el índice Ir de la red.

5.6.6 Metodología heurística de diseño de distritos métricos.

Di Nardo y Di Natale (2011a) plantean identificar configuraciones de sectores

[14] Multilevel Recursive Bisection o Bisección Recurisva Multinivel.
[15] METIS es una serie de programas para partición de grafos y partición finita de elementos mallados. Disponible en: http://glaros.dtc.umn.edu/gkhome/metis/metis/overview

efectivas en términos de costes. En particular, ayuda a los diseñadores a definir sectores mediante comparación de muchas opciones. Los pasos de la propuesta son:

1. Hacer una simulación de la RDAP para obtener el caudal en las tuberías y la altura piezométrica en los nodos.
2. Definir una matriz de adyacencia, incidencia y pesos.
3. Calcular los caminos más cortos de la fuente a cada nodo y la frecuencia con que cada tubería aparece en todos los caminos. Borrar las tuberías con frecuencia nulas. El resultado será un esquema principal de la red.
4. Transformar el esquema anterior en un grafo.
5. Escoger un número de sectores y sus dimensiones.
6. Insertar las UOC.
7. Colocar las válvulas para aislar los sectores.
8. Volver a hacer la simulación de la red con el nuevo esquema.
9. Evaluar la alternativa con índices de eficiencia energética. En caso que no sea satisfactorio, se replantea el paso 5. En caso que sea satisfactorio, se elimina una UOC para ver si los índices de eficiencia energética se mantienen en los rangos de aceptabilidad. La idea es llegar un punto con el mínimo número de UOC posibles.

5.6.7 Sectorización de redes de agua basadas en algoritmos genéticos.

La metodología propuesta por Di Nardo *et al*. (2013b) es una variación de la descrita en la sección anterior. En esta se persigue crear sectores-aislados, para lo cual, inicialmente, tal y como fue descrito anteriormente, se busca el camino más corto de disipación de energía de cada nodo a cada fuente. La variación de esta propuesta parte de el hecho que existen ciertos nodos que pueden estar en el camino de mínima disipación de dos fuentes, con lo que se sigue un procedimiento heurístico, basado en algoritmos genéticos, para evaluar en qué sector-aislado se debe encontrar cada uno de estos nodos, de manera tal que se optimicen los índices energéticos ya presentados.

5.6.8 Definición de clústeres en RDAP.

Herrera (2011a) plantea el uso de aprendizaje semisupervisado para definir clústeres en las RDAP. Para efectos de este trabajo se considera la red como un grafo no dirigido. La idea central del proceso es combinar una matriz del grafo (matriz ***L***) con las matrices de *disimilaridad* de las características de la red que se deseen incluir en la partición. Esto se hace mediante la expresión presentada en la sección 5.5.3.2.2, una vez que se transforman las matrices de características de la RDAP (incluyendo la matriz ***L*** del grafo) en matrices *kernel*. Es a partir de esta combinación que se hace la partición en sectores mediante la aplicación de clústering espectral.

Los pasos del método propuesto son:

- *Abstracción de la RDAP como un grafo.*
- *Construcción de una matriz laplaciana y de disimiliaridad.*
- *Introducción de las restricciones hidráulicas.*
- *Transformación de datos en una única matriz kernel.*
- *Cálculo del espectro de la matriz.*
- *k-means de los valores propios más altos.*
- *Reasignación de clúster dentro de los datos originales.*
- *Validación hidráulica (EPANET).*

6 METODOLOGIA PROPUESTA

6.1 Descripción de la Metodología

La metodología propuesta sigue la línea de Aprendizaje Automático Computacional (*ML*, Machine Learning) ya propuesta por Herrera. (2011a); no obstante, en esta última propuesta, no se toman en cuenta las tuberías de la red que forman la red de alta. En este contexto, la resolución del problema sólo tiene en cuenta los nodos de la red de distribución y los cortes de tuberías propuestos se hacen en las líneas de esta misma. En un primer momento se hace uso del método de aprendizaje *no supervisado*, específicamente, clústering jerárquico. Con este método se hace una primera exploración de las características de la red y mediante un algoritmo de formación de clústeres en jerarquía, en combinación con medidas de validación de clústeres, se determina el número de clústeres en que se puede realizar la partición de la red. A continuación se emplea otro método de aprendizaje *no supervisado*, en específico clústering espectral, con el cual se logra una partición en la que se optimice (minimice) el número de válvulas necesarias. Esta optimización también tiene en cuenta los pesos que se asocian a las características que se emplean en el partición, de manera tal que si una de las características es muy importante para la partición, a esta se le asigna un peso muy alto con respecto a las otras y la forma de los sectores se verá altamente influida por ella. Los pesos que se asocian a la suma de matrices *kernel* en el paso anterior, son estimados con la herramienta AHPcalc (Goepel, 2013), que está disponible en la página web de la empresa "*Bussiness Performance Management*". Otra novedad que se incluye en este método con respecto a la propuesta de Herrera (2011a), es la inclusión de emisores de fugas dentro de las características a tomar en cuenta; así, las zonas pueden ser aisladas en función del nivel de fugas que reporten. El valor de λ_A de la suma de matrices *kernel* se estima por un análisis de costes para diferentes particiones, tal y como ha sido descrito en la sección 5.5.3.2.2. Una vez hecha la partición, se elige la entrada de cada sector. Para elegir

la entrada de un sector dado, lo primero es determinar las líneas que, de manera independiente (mientras el resto de las líneas que conectan al sector con la red de alta están cerradas), pueden abastecer al mismo por 24 horas. Estas se definen como líneas candidatas. Luego, se cierran todas las líneas del sector en que se hará la primera evaluación, mientras una de las líneas candidatas se deja abierta. Las restantes líneas de la red se dejan abiertas. Bajo este escenario, se calcula para el sector dado el Coeficiente de Pérdida de Potencia (*CPP*) y el Índice de Resiliencia (I_r). Luego se procede de la misma manera para el resto de las líneas candidatas. Como resultado, se obtienen dos *rankings* de valores de eficiencia energética del sector, a partir de los cuales se puede seleccionar la mejor opción de alimentación del mismo.

En algunos contextos, las particularidades urbanísticas hacen muy difícil la instalación de Unidades Operativas de Control (UOC) en algunos sitios. Mediante los *ranking*s anteriores se pueden estimar entradas alternativas en las que el operador no tenga que lidiar con aspectos constructivos muy complejos. En algunos casos, se podría considerar un aumento de diámetro de alguna línea que no sea la primera en ambos *rankings*, y así lograr que las presiones en el sector sean similares a las que se obtendrían si se emplease la primera alternativa.

A continuación se valida la propuesta mediante simulación hidráulica en EPANET 2.0. El parámetro que se consideran para tal valoración es la presión en los nodos, tanto a la hora de mayor consumo, así como a la hora en la cual el consumo es menor.

A continuación se enumeran los pasos del método:

1. *Preparación del modelo matemático de la red en EPANET 2.0 con distribución de coeficiente de emisores y balance de caudales.*
2. *Clústering jerárquico de la red, para obtención del dendrograma.*
3. *Validación del procedimiento de clústering jerárquico mediante el CPCC.*
4. *Ponderación de las características de la red mediante AHP.*
5. *Transformación de la red en un grafo no dirigido.*

6. *Obtención de la matriz kernel de la matriz* ***L*** *del grafo.*
7. *Obtención de las matrices kernel de las matrices de disimilaridad de las características de la red.*
8. *Suma de las matrices kernel de los pasos 6 y 7, con las ponderaciones obtenidas en el paso 4.*
9. *k-means de los valores propios de la nueva matriz kernel.*
10. *Selección de las entradas de cada sector (en el escenario de mayor consumo).*
11. *Modelización matemática de la propuesta de sectorización.*

Como alternativa, se plantea estudiar una mejora en los índices energéticos de toda la red (*CPP*, *CU*[16], I_r, I_{rd}). Esto se logra variando las fronteras entre los sectores. En el ejemplo posterior, se muestra una variación al esquema de sectorización inicial que reduce levemente la disipación de energía. Dado que las posibilidades de variación son muchas, sería una tarea exhaustiva probar todas estas sin contar con una herramienta informática para tal fin. Se espera en el futuro poder desarrollar un procedimiento heurístico que permita probar muchos escenarios, permitiendo encontrar el mejor esquema posible.

6.2 Ventajas comparativas de la metodología propuesta

De lo expuesto, a lo largo del trabajo, se puede concluir que las metodologías de sectorización previamente propuestas se han centrado principalmente en establecer estrategias para definir sectores alrededor de una fuente de abastecimiento (sectores con fuentes de abastecimiento exclusiva). Esto puede ser válido para RDAP con muchas fuentes distribuidos dentro de ellas. Los trabajos

[16] Coeficiente de Uniformidad

en que se ha abordado la sectorización de redes con un número limitado de fuentes, no consideran la segregación de una red troncal y por ende se generan cascadas de sectores. La metodología que se propone en este trabajo hace una innovación muy importante en este sentido, al considerar una segregación de la red troncal (red de alta) de la red secundaria. Esta innovación implica una serie de ventajas, con respecto a metodologías previamente planteadas, para la gestión de RDAP de ciudades de mediana y gran extensión. A continuación se enumeran estas ventajas:

Aplicabilidad en redes de abastecimiento de agua potables de mediana y gran extensión:

En muchas RDAP de mediana y gran extensión (más de 1000 km de extensión de tuberías), las fuentes de abastecimiento no se encuentran dentro de la ciudad y/o cuentan con un número de fuentes muy limitados. Esto hace inviable el establecimiento de microsectores con una fuente exclusiva. Se pueden establecer únicamente macrosectores con una fuente exclusiva, pero luego estos serían poco útiles para los objetivos de un plan de control activo de fugas, en vista que se dispondría de menor grado de sensibilidad para registrar cambios anómalos de caudal. De allí que sea necesario establecer sectores de menor tamaño que se abastezcan partir de puntos conectados a una red troncal.

También, al segregar la red troncal de la sectorización, se forman sólo sectores conectados individualmente a la red troncal y no cascadas de sectores. Al contar cada sector únicamente con un caudalímetro, se facilita el monitoreo de los caudales y por ende las actividades relativas al control activo de fugas.

Mayor eficiencia en el control activo de fugas:

La otra ventaja se relaciona con el control activo de fugas. Es ampliamente reconocido, que el mayor volumen de fugas en las RDAP se reporta en las líneas y accesorios de conexión de menor diámetro (al ser estas líneas más sensibles a cambios de presión, ser manipuladas en muchos casos por personal

ajeno a las compañías de agua o simplemente recibir poco mantenimiento). También, en estas líneas es más complicada la detección de una fuga. Al separar la red troncal de la sectorización, se enfocan los esfuerzos de gestión de fugas sobre el punto central en donde se produce el problema.

Ahorro de costes por implementación:

La metodología también presenta ventajas en términos económicos, ya que al no incluir la red troncal, los caudalímetros tendrán menor diámetro. Es importante tener en cuenta que es recomendable ubicar los caudalímetros dentro de una caja de registro (UOC, unidad operativa de control). Al segregar la red troncal, las UOC son más pequeñas (por ende más económicas) y se ubican en vías menos traficadas, con lo cual, su construcción implica un menor grado de complejidad.

Flexibilidad operacional:

Las RDAP cambian constantemente en respuesta a la demanda. Esta respuesta en muchas ciudades se produce de una manera anárquica. Al mantener la red de troncal segregada, se hace más simple la modificación del esquema de sectorización para enfrentar las necesidades eventuales.

Rankings de alternativas para colocación de Unidades Operativas de Control (UOC):

La metodología establece un ranking de soluciones para la colocación de caudalímetros. Así, los operadores pueden considerar aspectos urbanísticos a la hora de decidir en qué punto se colocarán las entradas de cada sector.

Consideración del nivel de fugas para establecer el esquema de sectorización:

Esta metodología es la primera que permite considerar el nivel de fugas en la red a la hora de establecer el diseño de los sectores. Esto tiene como ventaja la capacidad de aislar zonas críticas (en términos de nivel de fugas) en las que los operadores quieran centrar esfuerzos especiales (por ejemplo el caso antiguo de una ciudad).

6.3 Implementación de la Metodología

Para ejemplificar el método, se toma una red de 100 km de longitud de tubería, con una red de alta de 10 km (ver Ilustración 19). La demanda media de la red corresponde a 523.8 m^3/h y el caudal medio de fugas igual 210 m^3/h. En la Ilustración 20 se muestra las curvas de caudales de la red en cuestión.

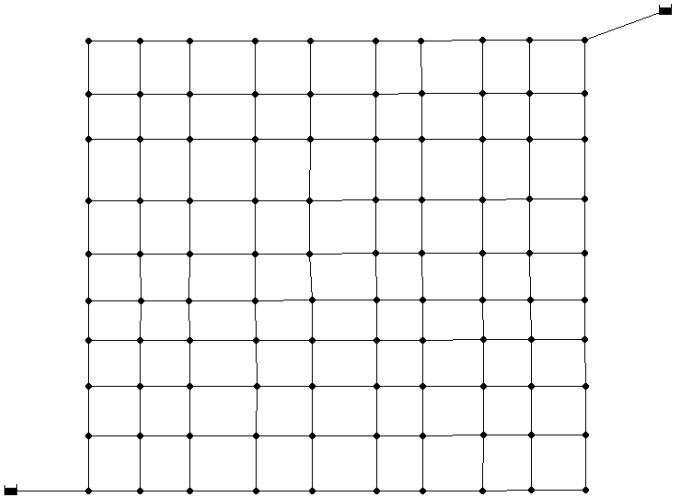


Ilustración 19: Red de Estudio

Los caudales de fugas son representados mediante coeficiente de emisores en los nodos de la red. Se ha adaptado el patrón de demanda, en función de las presiones resultantes, de tal manera que la curva de caudal de entrada coincidiera con la curva de caudal de medición.

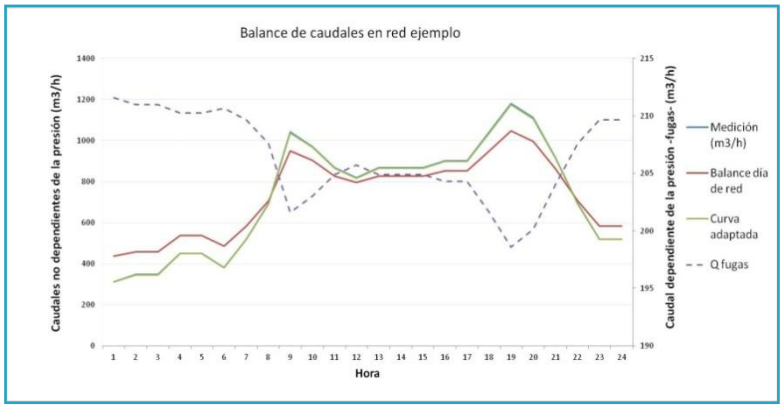


Ilustración 20: Balance de caudales en red ejemplo

6.3.1 Selección del número de sectores

Como primer paso, se hace el análisis de los datos mediante la técnica de clústering jerárquico. Para la creación de la matriz de *disimilaridad* se evalúa el CPCC para 8 combinaciones posibles. A cómo se puede ver en la Tabla 11 y la

Ilustración 21 el valor de CPCC más alto se obtiene para el uso de métrica *euclidiana* y método de fusión *promedio*. Como resultado, se obtiene el *dendrograma* de la Ilustración 22.

Métrica	***Método***	***Promedio***	***Centroide***	***Completo***	***Individual***
Euclidiana		**0.9477671**	0.9023773	0.9412087	0.9380865
Manhattan		0.940827	0.9129212	0.9400916	0.9361117

Tabla 11: Evaluación del CPCC

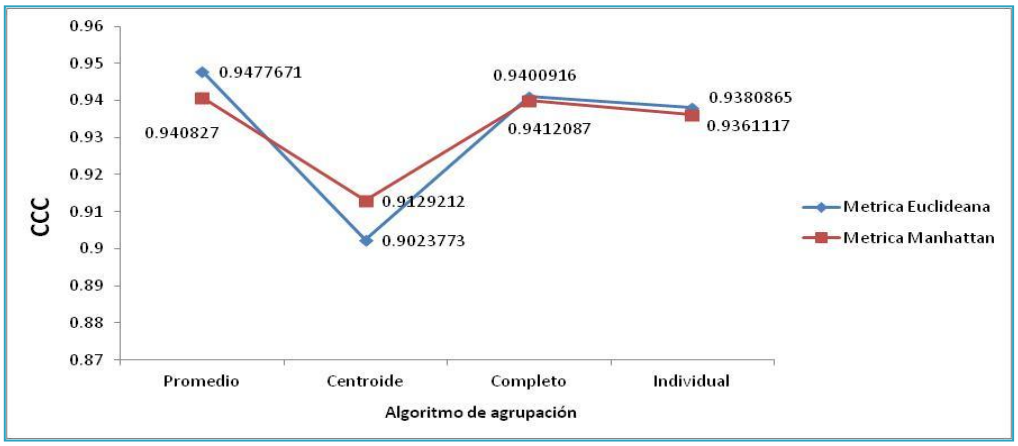


Ilustración 21: CPCC obtenidos

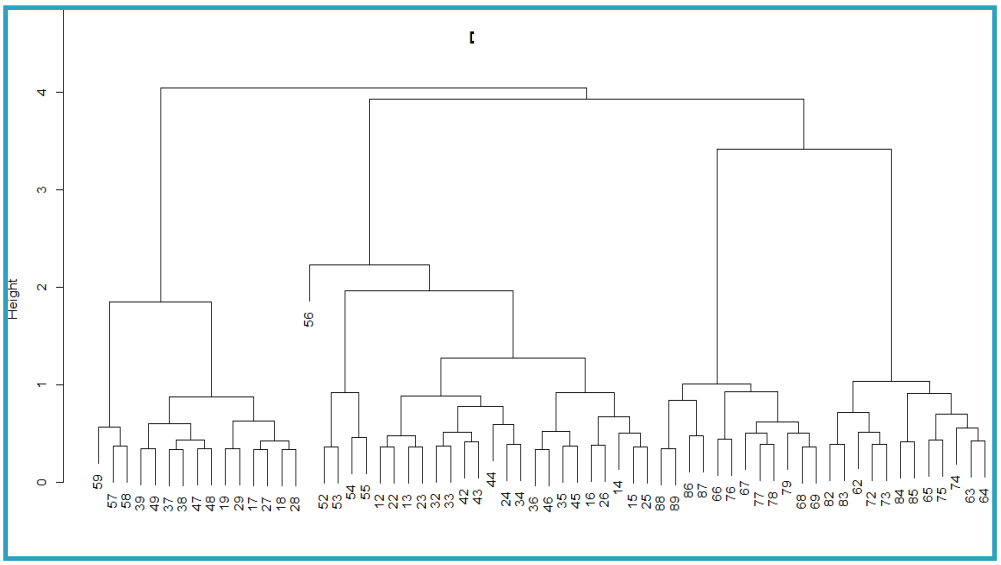


Ilustración 22: *Dendrograma* de la red obtenido

A continuación se procede a estimar el número de clústeres en que se hará la partición. Los criterios a considerar serán: el criterio del codo en la gráfica de altura vs número de clústeres; promedio de *ancho de siluetas*; *índice de conectividad*; *índice de Dunn* y los *pv-values*, AU[17] y BP[18].

En la Ilustración 23 se observa que a partir de 3 clústeres, los valores de altura de fusión no varían mucho más. Por otro lado, de acuerdo a la Ilustración 24, las medidas de evaluación interna se optimizan cuando la partición se hace en tres clústeres. La Ilustración 25 muestra el ancho de silueta para una partición en tres clústeres. Finalmente, los resultados de *p*-values AU/BP muestran también que la partición mejor sustentada por los datos implica tres clústeres, tal y como se puede ver en la Ilustración 26.

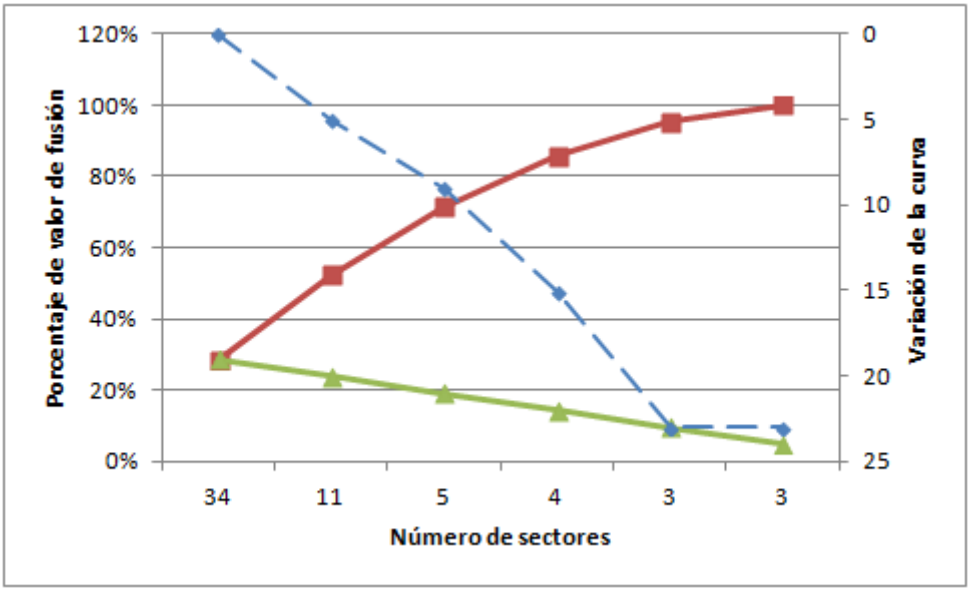


Ilustración 23: Curva de altura vs número de clústeres

[17] Approximately Unbiased
[18] Bootstrap Probability

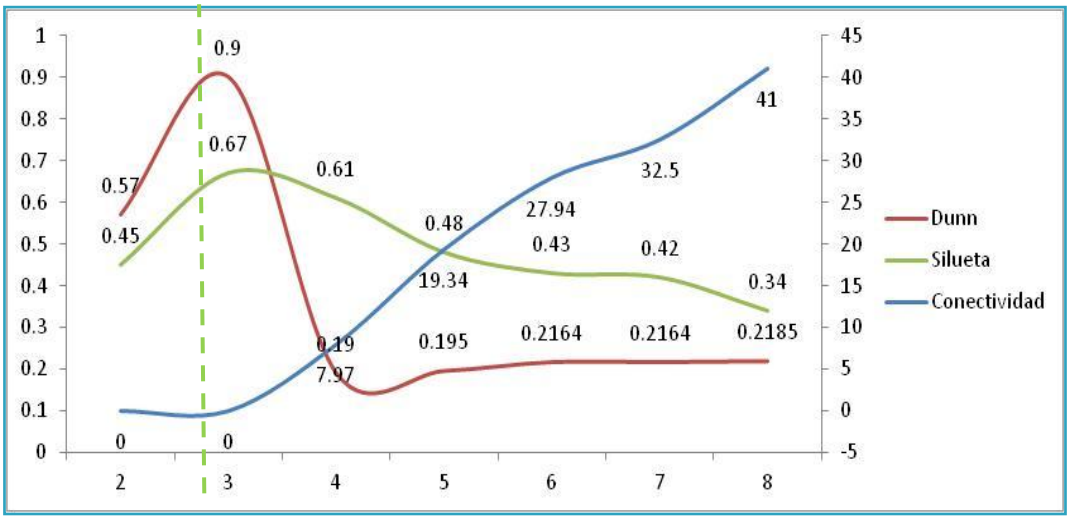


Ilustración 24: Evaluación de medidas internas

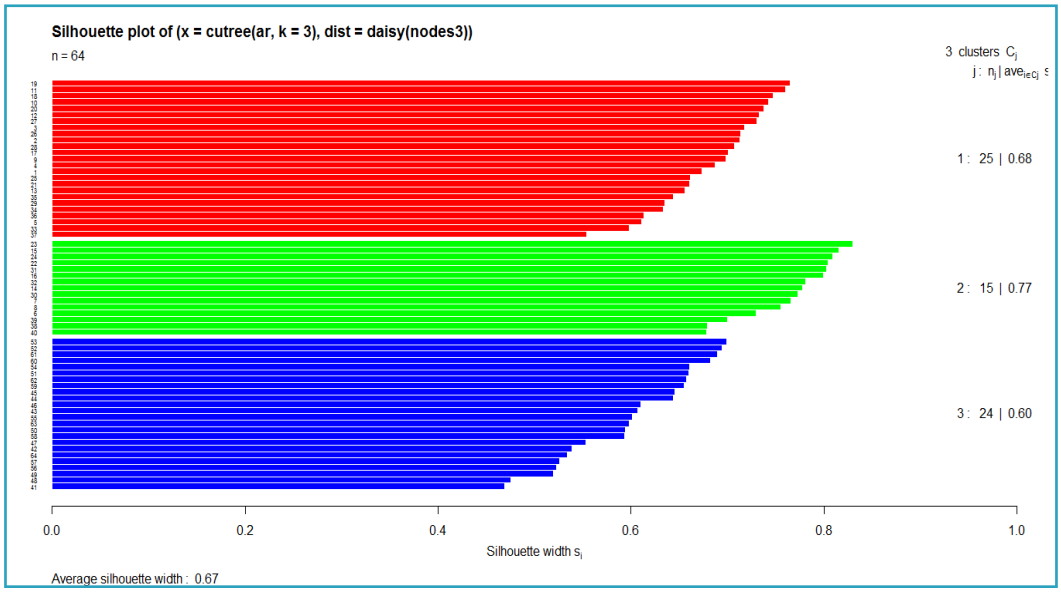


Ilustración 25: Ancho de silueta para una partición en tres sectores

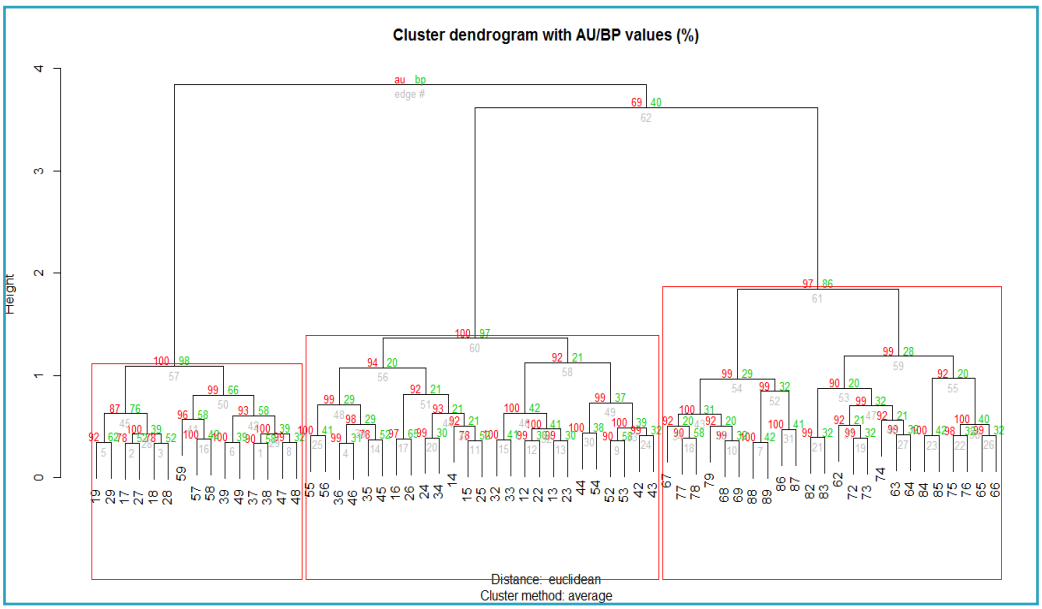


Ilustración 26: *p*-values AU/BP obtenidos por *remuestreo multiescala* a partir de los mismos datos

6.3.2 Subdivisión de la red en sectores

Habiendo tomado la decisión de hacer la partición en tres clústeres, se proceden a ponderar, mediante AHP, las características de la RDAP que serán empleadas como criterios de partición (*coordenada x, coordenada y, coeficiente de emisor, demanda, elevación*). Como resultado se obtiene el siguiente vector de prioridad: $(0.044,\ 0.054,\ 0.265,\ 0.382,\ 0.254)^{0.04}$, en el que 0.04, corresponde al Ratio de Inconsistencia (*RI*), el cual, al ser menor al 10% valida el resultado obtenido.

	Coor-x	***Coor-y***	***Emisor***	***Demanda***	***Elevación***
Coor-x	1	1	1/6	1/7	1/8
Coor-y	1	1	1/4	1/5	1/6
Emisor	6	4	1	1	1
Demanda	7	5	1	1	3
Elevación	8	6	1	1/3	1

Tabla 12: Comparación de criterios a tomar en cuenta en la partición

A continuación se muestra la suma de las matrices *kernel* del grafo de la red y de

las matrices de *disimilaridad* de las características empleadas.

$$K = 0.4 * k_A + \{(1 - 0.4) * [(0.044 * coorx) + (0.054 * coory) + (0.265 * emisor) + (0.382 * demanda) + (0.254 * elevación)]\}$$

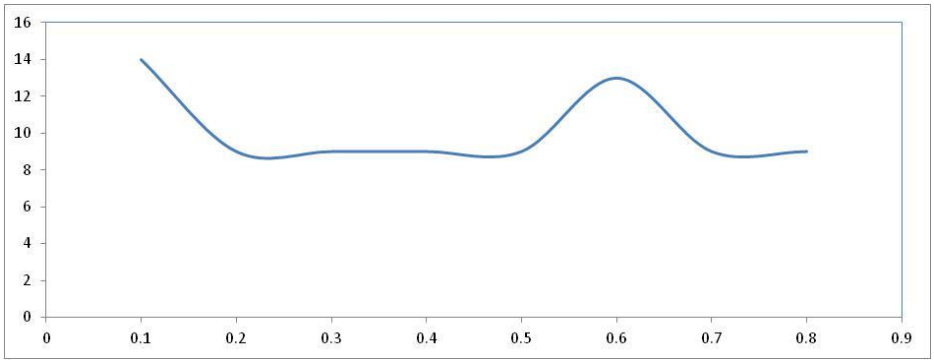


Ilustración 27: Gráfica para determinar el valor alfa de la suma de matrices *kernel*

Como se puede ver, el rango de valores "*económicos*" de λ_A va de 0.2-0.5 (ver Ilustración 27). En este caso se tomó el valor 0.4 a fin de darle una importante ligeramente menor al grafo con respecto a las características de las RDAP.

En la Ilustración 28 se puede apreciar la subdivisión de los nodos de la red en tres clústeres. Es notorio que la división está más influenciada por las características hidráulicas de la red (demanda, elevación, emisor), y mínimamente por las características geográfica, tal y como es de esperar de acuerdo a los pesos empleado en la suma de las matrices *kernel.*

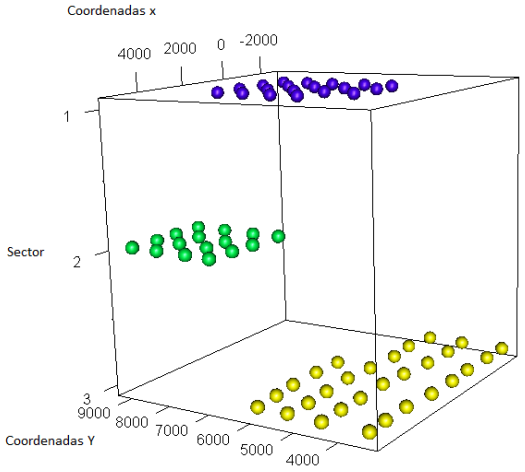


Ilustración 28: Clústeres en la red ejemplo

En la Ilustración 29 se observa la división de la red en tres subsectores. Las líneas de color negro indican tramos de cierre, o bien con corte de tuberías o bien la colocación de válvulas cerradas.

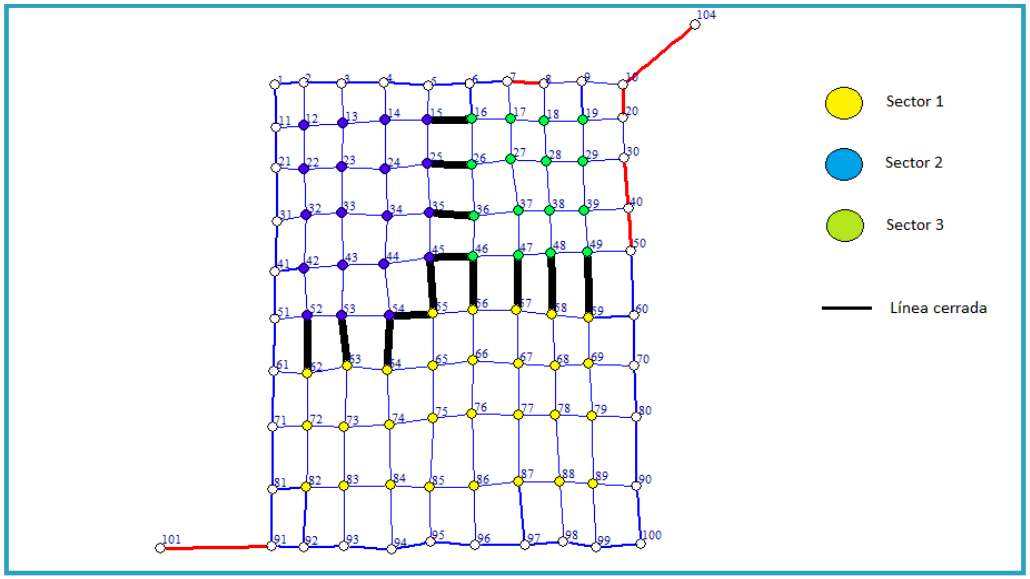


Ilustración 29: Partición de la red ejemplo

Si se colocara una variación muy fuerte en algunas de las características de una determinada zona de la red, esto modificaría el resultado de la partición. Para ejemplificar esto, se toma la red anterior, y se agrega un coeficiente de emisor muy elevado a una zona en donde las características en los nodos son muy

similares (ver zona sombreada de la Ilustración 30).

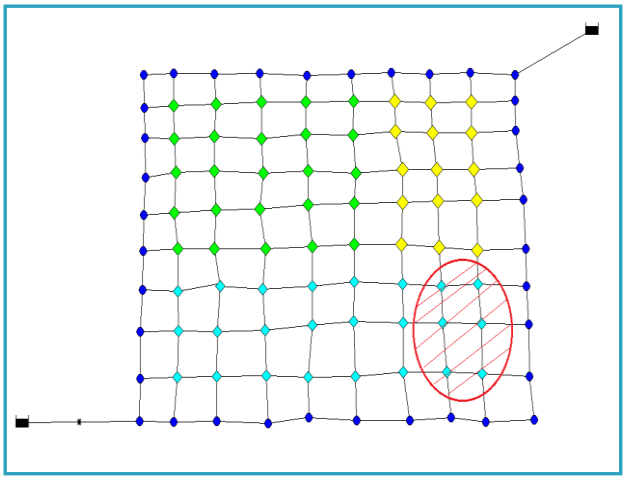


Ilustración 30: Red con un valor de coeficiente de emisor elevado en algunos nodos

En este nuevo caso, los *p*-values (AU/BP) indican que la partición mejor soportada por los datos se da en cuatro sectores (ver Ilustración 31), que a su vez coincide con los máximos valores del *índice de Dunn* y el promedio de *ancho de silueta* (ver Ilustración 32). De acá que en este caso sea lógico seleccionar una partición en cuatro clústeres.

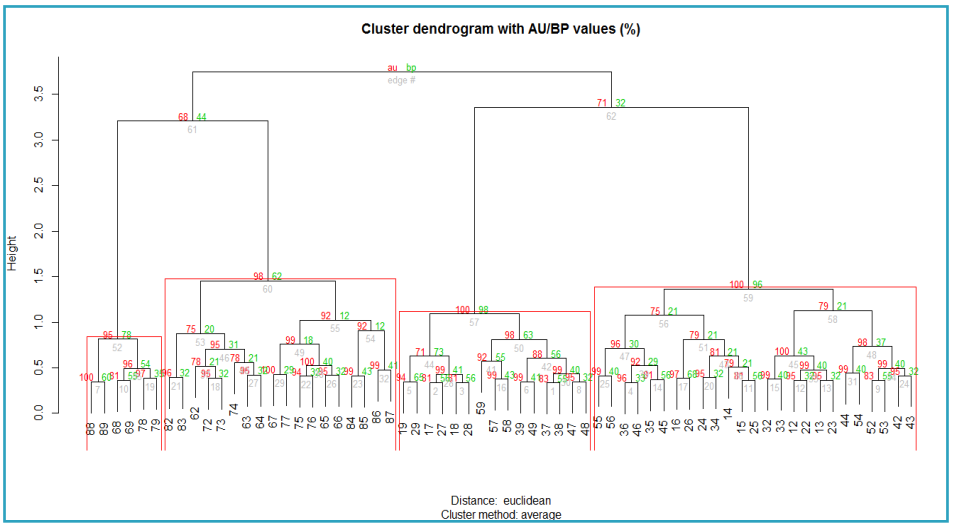


Ilustración 31: *p*-values AU/BP obtenidos por *remuestreo multiescala* a partir de los mismos datos

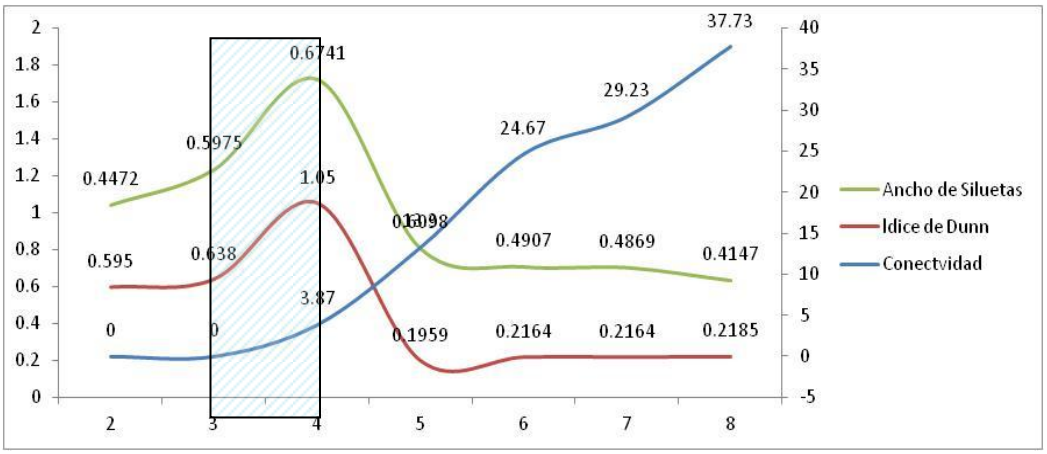


Ilustración 32: Evaluación de medidas internas

Manteniendo los mismos pesos en todas las características de la red al momento de hacer la suma de las matrices *kernel*, se obtiene la partición que aparece en la Ilustración 33. En ella se puede notar la subdivisión del sector 3 del ejemplo anterior en dos nuevos subsectores; no obstante, no todos los nodos que constituyen el nuevo subsector tienen un elevado valor de coeficiente de emisor.

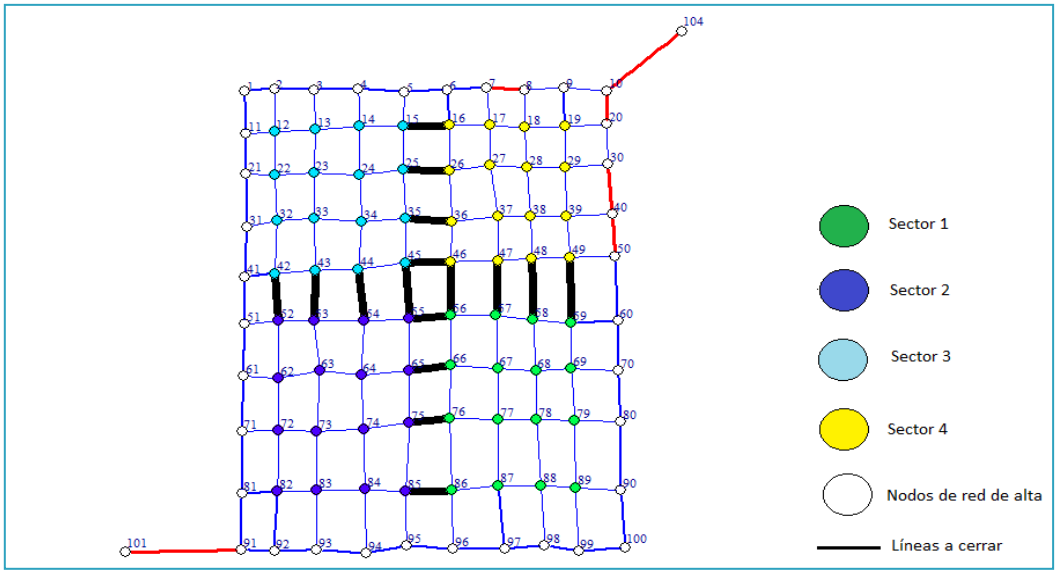


Ilustración 33: Nueva partición de la red usando pesos del ejemplo anterior

Si en el momento de hacer la suma de las matrices *kernel* se le da importancia

máxima a la característica emisores, la partición variará, limitando todas las zonas en la que el coeficiente de emisor tenga el mismo valor.

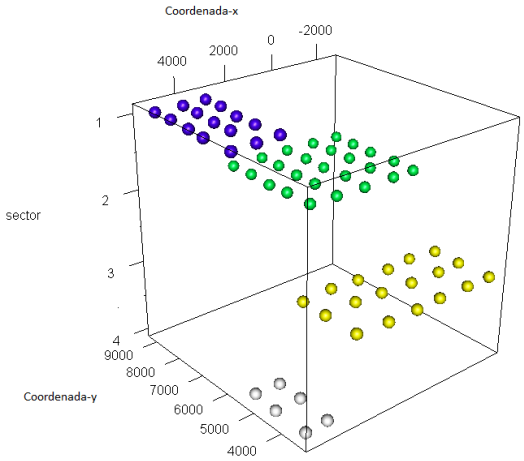


Ilustración 34: Clústeres en la nueva partición

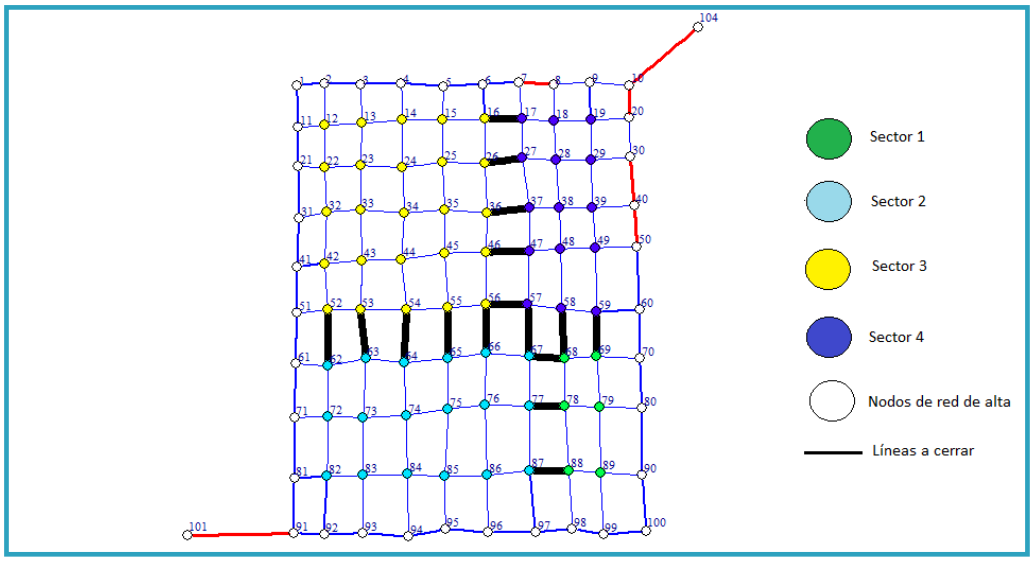


Ilustración 35: Nueva partición de la red dando importancia máxima al coeficiente de emisor

6.3.3 Ubicación de Unidades Operativas de Control (UOC).

Dada la primera partición, se hace, para cada uno de los escenarios de sectorización, un *ranking* de los *CPP* por sectorización, así como de los I_r. En el

caso del I_r, se calcula un valor con la presión (*P*) en los nodos y otro valor con su altura piezométrica (*H*). En la Ilustración 36, se muestra la partición en la que se realizará la selección de entradas.

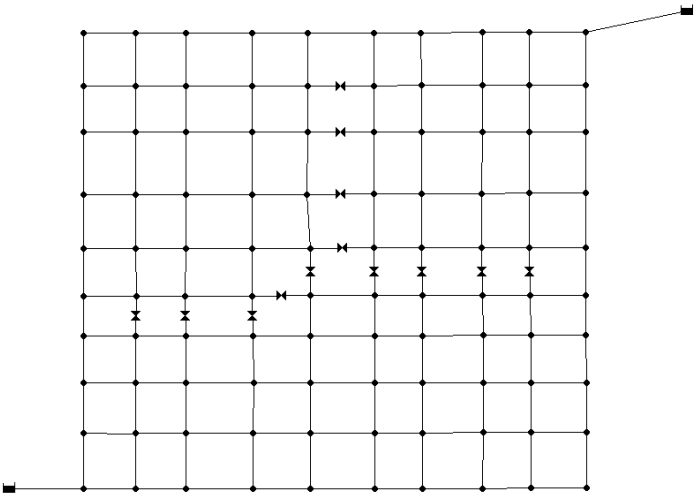


Ilustración 36: Esquema de sectorización antes de la selección de UOC

6.3.3.1 Sector 1

En el sector 1 se definieron dos líneas candidatas 191 y 153 con capacidad de abastecer al sector de manera continua 24 horas. Se estableció como presión mínima deseada 10 mca para el cálculo del I_r.

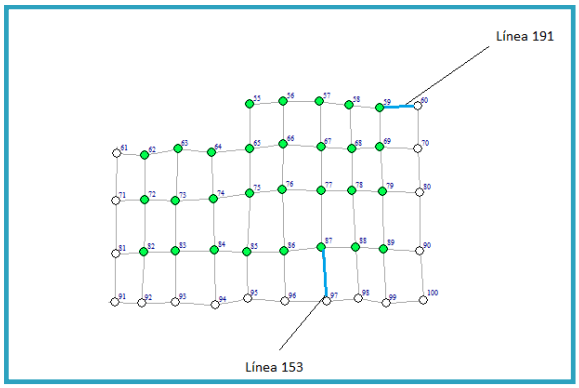


Ilustración 37: Líneas candidatas para UOC del sector 1

En la Tabla 13 se puede apreciar que la primera alternativa correspondería a la línea 191; no obstante, con esta no se garantiza la presión mínima de 10 mca. Subiendo el diámetro de esta misma línea a 400 mm se logra que la presión se ubique sobre 10 mca. Se podría considerar un aumento de la línea 153, pero al subirla a 400 mm, siguen registrando presiones inferiores a 10 mca.

	L-191 (300 mm)	L-191(400 mm)	L-153 (300 mm)	L 153 (400 mm)
Cociente de Pérdida de Potencia (*CPP*)	20.9%	36.4%	6.3%	15.0%
Índice de Resiliencia Ir calculado con P	-0.0068	0.2	-0.47	-0.17
Índice de Resiliencia Ir calculado con H	-0.0078	0.27	-0.42	-0.19

Tabla 13: Cálculo de *CPP* e I_r-Sector 1

6.3.3.2 Sector 2

En el sector 2 se definieron dos líneas candidatas (127 y 73) con capacidad de abastecer al sector de manera continua 24 horas. Se estableció como presión mínima deseada 10 mca para el cálculo del I_r.

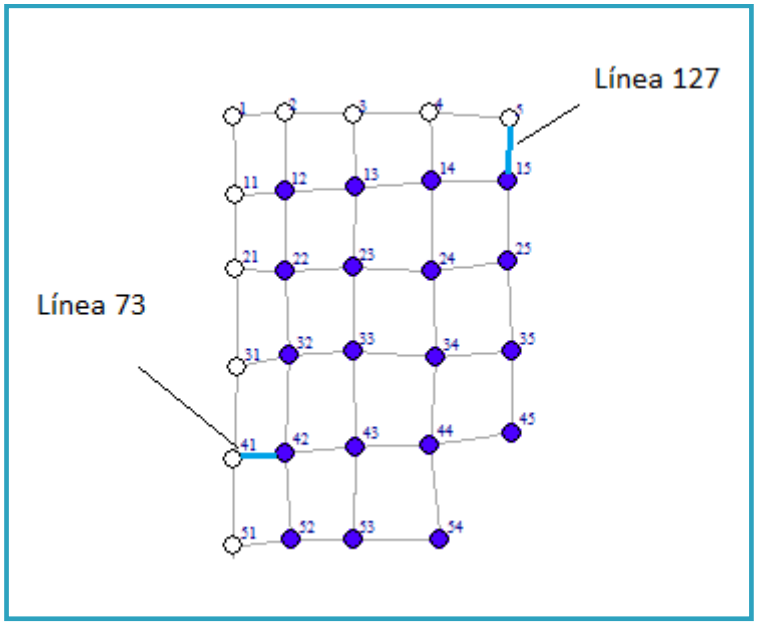


Ilustración 38: Líneas candidatas para UOC del sector 2

En la Tabla 14 se puede apreciar que la primera alternativa correspondería a la línea 73. Con esta línea se puede garantizar una presión mínima de 10 mca. Otra alternativa que se podría considerar es la línea 127, que también garantiza presiones sobre 10 mca, pero más bajas que el caso anterior. En caso de que sea muy complicado construir la UOC en la línea 73, se podría considera como alternativa válida la línea 127.

	L-127 (250 mm)	L-73(300 mm)
Cociente de Pérdida de Potencia (CPP)	39.0%	65.0%
Índice de Resiliencia Ir calculado con P	0.17	0.55
Índice de Resiliencia Ir calculado con H	0.18	0.55

Tabla 14: Cálculo de *CPP* e I_r-Sector 2

6.3.3.3 Sector 3

En el sector 3 se definieron cuatro líneas candidatas (145, 163, 17 y 183) con capacidad de abastecer al sector de manera continua 24 horas. Se estableció como presión mínima deseada 10 mca para el cálculo del I_r.

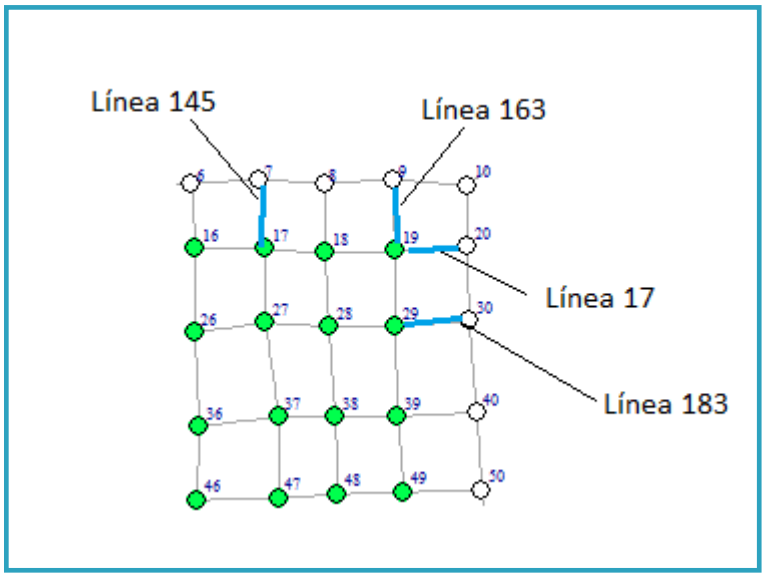


Ilustración 39: Líneas candidatas para UOC del sector 3

En la Tabla 15 se puede apreciar que la primera alternativa correspondería a la

línea 163. Con esta línea se puede garantizar una presión mínima de 10 mca.

	L-145 (250 mm)	L-163 (300 mm)	L-17 (250 mm)	L 183 (250 mm)
Cociente de Pérdida de Potencia (*CPP*)	9.0%	56.0%	17.0%	20.0%
Índice de Resiliencia I_r calculado con *P*	-0.19	0.44	-0.04	-0.01
Índice de Resiliencia I_r calculado con *H*	-0.18	0.34	-0.04	-0.01

Tabla 15: Cálculo de *CPP* e I_r-Sector 3

6.3.3.4 Esquema de sectorización final

A continuación se muestra el esquema de sectorización final (ver Ilustración 40). Las entradas para el sector 2 y 3, se mantienen con el diámetro original. En tanto, en el caso de la entrada del sector 1, se aumenta el diámetro de 300 mm a 400 mm.

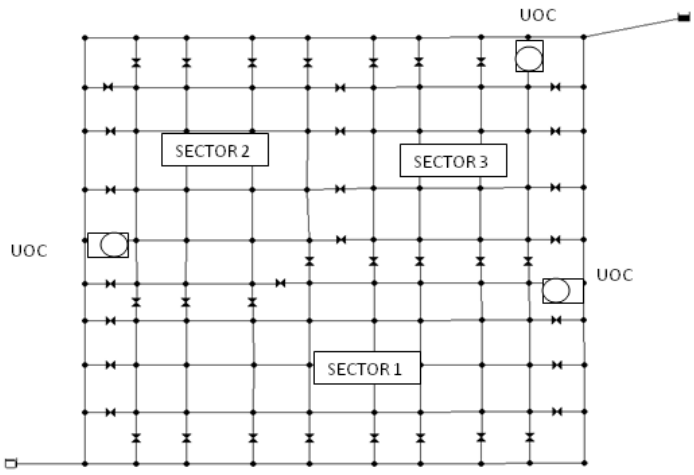


Ilustración 40: Propuesta de sectorización final

6.3.4 Evaluación de propuesta de sectorización

Habiendo definido el esquema de sectorización, se procede a hacer una evaluación de los parámetros hidráulicos en él. Para ello, se estudia la presión en el punto más alto y más bajo de la curva de consumo (18:00 horas y 0:00 horas).

Mediante la Ilustración 41 y la 42 se muestra que las presiones se ubican sobre 10 mca y por debajo 55 mca en los dos periodos (mayor y menor consumo). Más detalles de esta evaluación se pueden apreciar en la Tabla 16. En esta, se puede ver que la mayor desviación estándar entre curvas de presión antes y después sectorización se da para el sector 1, dado que es el sector en el que más baja la presión.

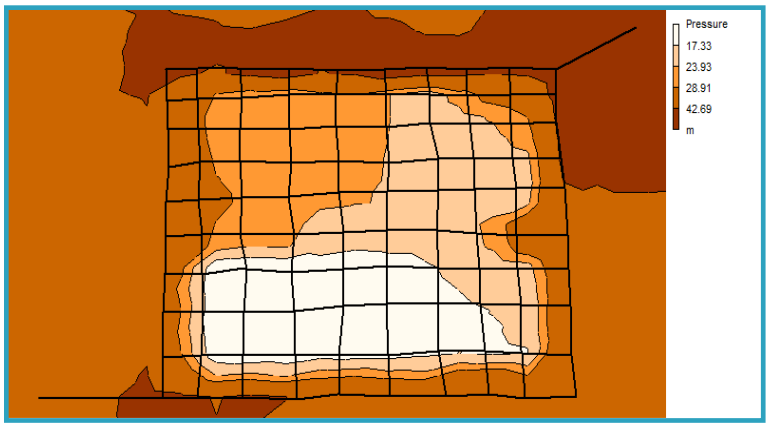


Ilustración 41: MDT de presiones para las 18:00 horas. Periodo de mayor consumo

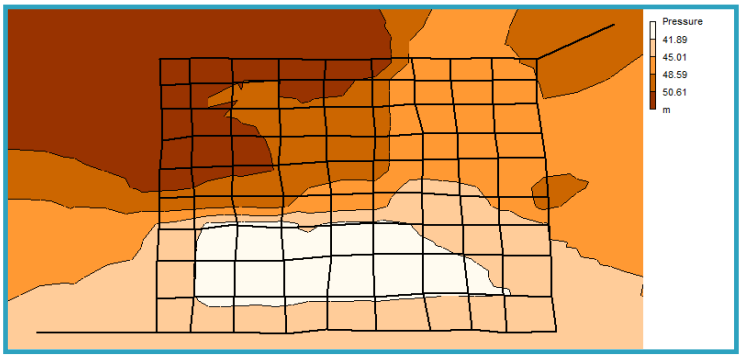
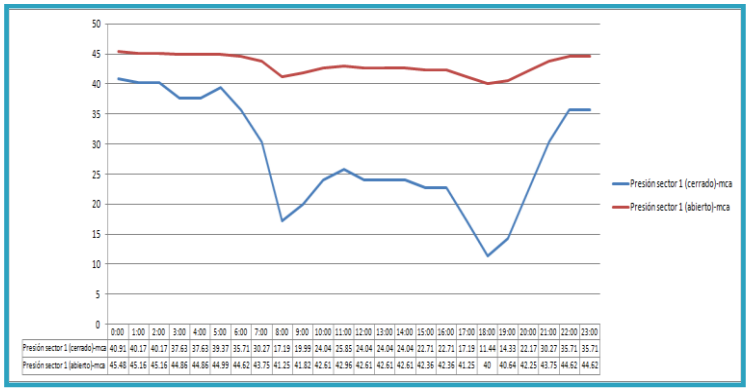
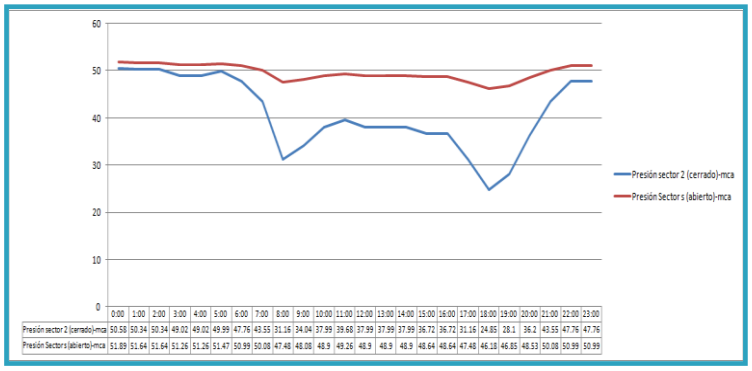


Ilustración 42: MDT de presiones para las 00:00 horas. Periodo de menor consumo



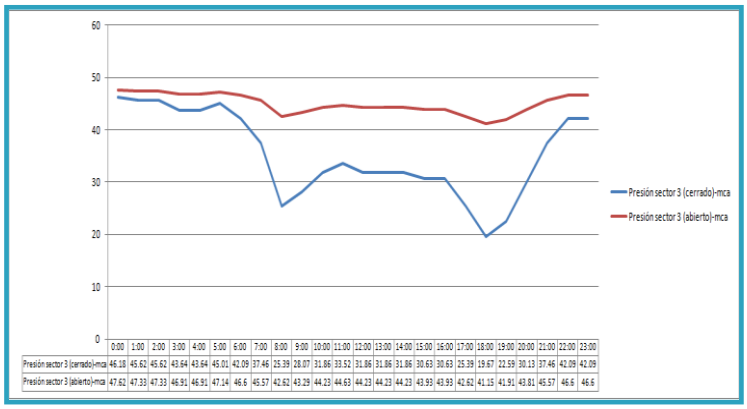
	0:00	1:00	2:00	3:00	4:00	5:00	6:00	7:00	8:00	9:00	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00	19:00	20:00	21:00	22:00	23:00
Presión sector 1 (cerrado)-mca	40.91	40.17	40.17	37.63	37.63	39.37	35.71	30.27	17.19	19.99	24.04	25.85	24.04	24.04	24.04	22.71	22.71	17.19	11.44	14.33	22.17	30.27	35.71	35.71
Presión sector 1 (abierto)-mca	45.48	45.16	45.16	44.86	44.86	44.99	44.62	43.75	41.25	41.82	42.61	42.96	42.61	42.61	42.61	42.61	42.36	42.36	41.25	40	40.64	42.25	43.75	44.62	44.62

Ilustración 43: Curva de presión pre y post sectorización en el nodo crítico del sector 1



	0:00	1:00	2:00	3:00	4:00	5:00	6:00	7:00	8:00	9:00	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00	19:00	20:00	21:00	22:00	23:00
Presión sector 2 (cerrado)-mca	50.58	50.34	50.34	49.02	49.02	49.99	47.76	43.55	31.16	34.04	37.99	39.68	37.99	37.99	37.99	36.72	36.72	31.16	24.85	28.1	36.2	43.55	47.76	47.76
Presión Sectors (abierto)-mca	51.89	51.64	51.64	51.26	51.26	51.47	50.99	50.08	47.48	48.08	48.9	49.26	48.9	48.9	48.9	48.64	48.64	47.48	46.18	46.85	48.53	50.08	50.99	50.99

Ilustración 44: Curva de presión pre y post sectorización en el nodo crítico del sector 2



	0:00	1:00	2:00	3:00	4:00	5:00	6:00	7:00	8:00	9:00	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00	19:00	20:00	21:00	22:00	23:00
Presión sector 3 (cerrado)-mca	46.18	45.62	45.62	43.64	43.64	45.01	42.09	37.46	25.39	28.07	31.86	33.52	31.86	31.86	31.86	30.63	30.63	25.39	19.67	22.59	30.13	37.46	42.09	42.09
Presión sector 3 (abierto)-mca	47.62	47.33	47.33	46.91	46.91	47.14	46.6	45.57	42.62	43.28	44.23	44.63	44.23	44.23	44.23	43.93	43.93	42.62	41.15	41.91	43.81	45.57	46.6	46.6

Ilustración 45: Curva de presión pre y post sectorización en el nodo crítico del sector 3

	Nodo 62		Nodo 15		Nodo 47	
	Presión sector 1 (cerrado)-mca	Presión Sector 1 (abierto)-mca	Presión sector 2 (cerrado)-mca	Presión sector 2 (abierto)-mca	Presión sector 3 (cerrado)-mca	Presión sector 3 (abierto)-mca
mínimo	11.44	40	24.85	46.18	19.67	41.15
máximo	40.91	45.48	50.58	51.89	46.18	47.62
Desviación estándar	10		7.6		7.6	

Tabla 16: Valores de presión máximos y mínimos en los sectores antes y después de la sectorización

Finalmente, esta propuesta de sectorización se traduce en un beneficio a corto plazo, tal y como lo es la reducción del caudal de fugas asociado a la disminución de la presión. Sólo por el hecho de sectorizar, se espera que el caudal de fugas baje de 4951 m^3/día a 4256 m^3/día, es decir, un 14% (ver Ilustración 46). A esto, se le sumará el beneficio de las facilidades sobre el CAF que se gana al tener la red sectorizada, lo que permitirá ir reduciendo las fugas hasta alcanzar el nivel económico de fugas característico del sitio.

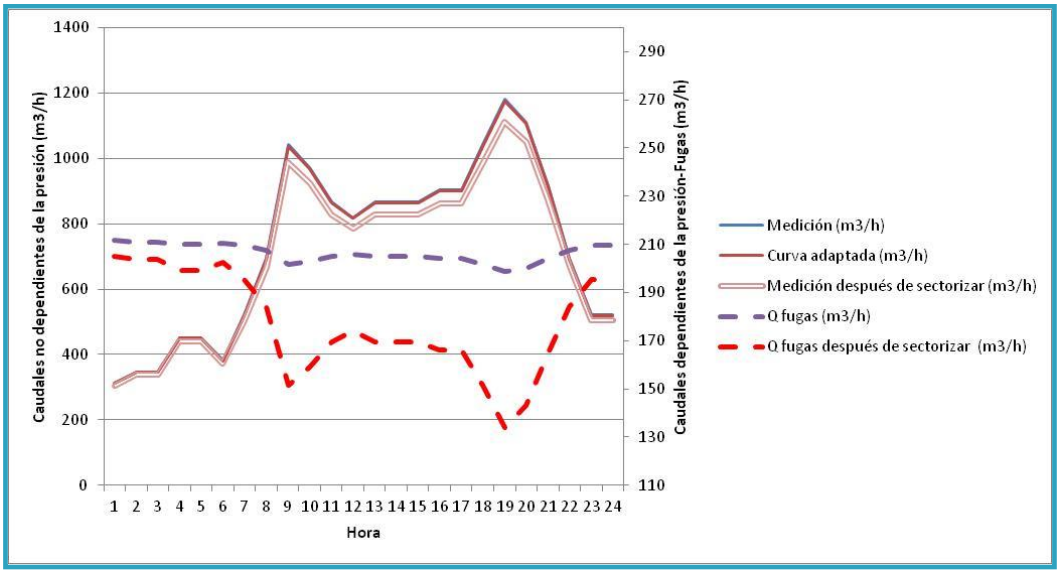


Ilustración 46: Curvas de consumo y caudales de fugas antes y después de la sectorización

6.3.5 Mejora de la eficiencia energética de la RDAP sectorizada

Teniendo el plano de sectorización final, se varió un poco la frontera de los sectores 1 y

3. Para ello, algunos de los nodos del sector 1 se pasaron al sector 3, reduciendo así el tamaño del primer sector y aumentando el tamaño del segundo (ver Ilustración 47). Como resultado de esta variación, se obtuvo un leve aumento del I_r y del *CPP*, a la vez que se aumentó el *CU*, indicativo de una mayor nivel de uniformidad de presiones.

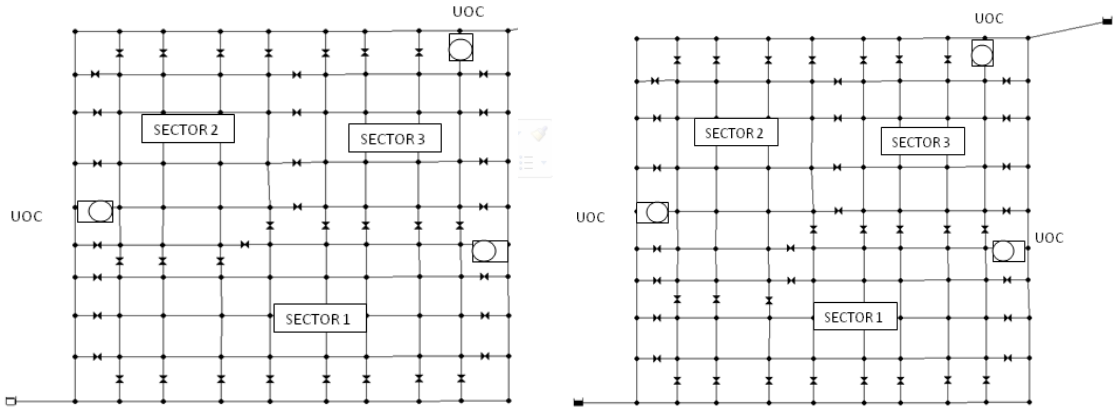


Ilustración 47: Modificación de la primera propuesta de sectorización. Original (izquierda), modificada (derecha).

	Propuesta 1	**Propuesta 2**
I_r	0.29	0.30
CPP	54.61%	55.06%
I_{rd}	65.8%	64.6%
CU	0.58	0.59
$P_{máxima}$	48.82	48.82
$P_{mínima}$	11.02	15.02

Tabla 17: Índices de energía para dos propuestas de sectorización

7 CONCLUSIONES Y LINEAS FUTURAS

7.1 Conclusiones

La sectorización de RDAP representa una opción estratégica que puede llevarse a cabo con múltiples objetivos, que van desde la mejora del CAF hasta el control de la calidad del agua. Disponer de una red sectorizada favorece el tratamiento de algunos problemas en la misma (fugas, calidad del agua, reparación, etc.), debido a la reducción dimensional implícita en ella (sectorización); sin embargo, su implantación cambia su comportamiento hidráulico, dado que al implicar el cierre de válvulas (tuberías), se rompe el principio de redundancia hidráulica que tienen las redes malladas, haciéndolas más vulnerables a entrar en escenarios de desabastecimiento ante la falla de uno (o más) de sus elementos. Esta alternativa de gestión, que ha cobrado un buen nivel de popularidad en acueductos de pequeña, mediana y gran extensión en América Latina y Europa, por lo general se lleva a cabo siguiendo aproximaciones de prueba y error, que luego pueden acarrear consecuencias negativas tales como el desabastecimiento y problemas de calidad de agua. Las recomendaciones existentes al respecto son generales, estableciendo el tamaño de los sectores en función del número de acometidas y/o extensión de longitud de tubería. En los últimos años se han desarrollado investigaciones en torno a este tema, pero en su mayoría están enfocadas hacia la generación de sectores sin tener presente una división de las RDAP en red de distribución y red de conducción principal, pese a que en muchas ciudades existe una división bastante marcada entre estas dos. El método propuesto en esta tesis aborda este problema, iniciando el proceso de sectorización con una segregación de la red de conducción principal del resto de la red. Con esta segregación, se conserva la flexibilidad de la red en caso que en el futuro se requiera variar el esquema de sectorización seleccionado y también se ahorran costes por la instalación de válvulas y caudalímetros de gran diámetro. En este trabajo, se demuestra la aplicabilidad del Aprendizaje Automático Computacional (*ML*,

Machine Learning) para abordar la tarea propuesta. Mediante clústering jerárquico, se logró estimar un número de sectores en los que se conserve el mejor grado de homogeneidad posible de las características de los sectores, lo que luego facilita que la red presente un buen rendimiento energético. Mediante el proceso de clústering espectral se logran mejorar los resultados de clústering jerárquico, encontrando una partición que además de mantener la homogeneidad de las características de cada uno de los sectores, minimice el número de válvulas que se deben emplear para hacer la partición. El empleo de indicadores de disipación de energía a través de la red, ha permitido encontrar las entradas a cada sector, de manera tal que se minimice la energía disipada por la red y se garantice la mayor presión posible en los nodos de consumo. Con esta propuesta, se logró obtener un plano de red sectorizada (100 km de red, división en tres sectores) que mantiene la presión dentro de los rangos establecidos como apropiados (10-55 mca) y a su vez implican una disminución del nivel de fugas tan sólo por implementarla. Es de esperar que esta propuesta de sectorización sea mejor en términos operacionales que propuestas basadas en principios empíricos. A partir de este esquema, se pueden incluir otros objetivos de sectorización, de manera tal que se puedan emprender acciones alrededor de estos objetivos, mientras se conserva un grado de funcionamiento apropiado de la RDAP.

Es muy importante destacar que esta propuesta de sectorización, al tomar en cuenta una segregación de la red de alta con respecto a la red de distribución, implica una serie de ventajas comparativas, tales como: aplicabilidad en ciudades de mediana y gran extensión que dependan de una red troncal, reducción de costes en la implementación, facilidades para las actividades relativas al control activo de fugas, conservación de la flexibilidad de la estructura de la red que permite modificaciones al esquema de sectorización implementado. También es la única propuesta que puede tener en cuenta el nivel de fugas de la red dentro del diseño de sectores.

7.2 Líneas Futuras

El método planteado permite evaluar la eficiencia energética de una o más

propuestas de sectorización, teniendo como referencia la red original. Al emplear clústering espectral se generan sectores en los que se minimiza el número de válvulas requeridas; sin embargo, las fronteras de estos sectores no garantizan los mayores valores de índice de eficiencia energética. La cantidad de escenarios dentro de los que se encuentra el escenario que puede reportar mayor eficiencia energética es muy grande, de allí que se recomiende aplicar algún método heurístico que permita hacer una negociación de fronteras de sectores a fin de encontrar el mejor de todos los posibles.

El primer paso de esta propuesta es una segregación de la red de alta con respecto a la red de conducción. En general, las redes de alta son definidas por los operadores de red; sin embargo, sería interesante incluir un análisis para determinar hasta qué punto se debe definir la frontera entre red de alta y red de distribución.

Dado que las redes no sectorizadas pueden tener problemas de diseño, al tomar como referencia el I_r de la red inicial, luego puede ser muy difícil encontrar el mejor escenario de red sectorizada. Esto podría ser mejorado si en los esquemas de sectorización propuestos, se contempla la posibilidad de corregir el diámetro de algunas tuberías. De esta manera, sería recomendable llevar a cabo un análisis de coste-beneficio de todas las posibles actuaciones que se puedan llevar a cabo para establecer una propuesta de sectorización apropiada, eligiendo la que con menor coste brinde el mejor resultado en términos energéticos.

Establecer un único número de sectores para la partición de la red con clústering espectral, podría limitar excesivamente el espacio de búsqueda de una buena propuesta de sectorización. Es importante tener en cuenta que las medidas empleadas para determinar el número de sectores en que se hace la partición a partir de un dendrograma siguen siendo en la actualidad objeto de investigación. Por esto se recomienda probar también soluciones con números de sectores cercanos al que se determina como "apropiado".

La sectorización puede perseguir muchos objetivos más allá del CAF. Se

recomienda combinar esta propuesta con otros objetivos de gestión, tal como lo es el control de la calidad del agua, calendario de reparación y/o reemplazo de tuberías.

8 REFERENCIAS

8.1 Referencias Propias

Campbell, G., Pérez, R., Izquierdo, J., Ayala, D. (2013). Aproximación a la sectorización automatizada en redes de abastecimiento de agua. En *III Jornada de Investigación del Agua (JIA*) .Valencia: Marcombo.

8.2 Referencias

Abony, J. and Balazs, F. (2007). *Cluster Analysis for Data Mining and System Identification*. Basel: Birkhäuser Verlag AG.

Aggarwal, C. C. (2011). *Social Network Data Analytics.* New York: Springer-Verlag.

Akhiezer, N. I., and Glazman, I. M. (1993). *Theory of linear operators in Hilbert space*. New York: Dover Publishing.

Albalate, A. and Minker, W. (2011). *Semi-supervised and unsupervised machine learning: novel strategies*. London: John Wiley & Sons, Inc.

Alpaydin, E. (2004). *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*. Massachusetts: The MIT Press.

Amar, N., Tazi, L. and Bensaid, A. (1997). Semi-supervised hierarchical clustering algorithms. In G. Grahne (Ed.), *Proceedings of the sixth Scandinavian conference on Artificial* intelligence *(SCAI '97)*: *Vol 1*, (*pp. 232-239*). Amsterdam: IOS Press.

Ando, R. and Zhang, T. (2005). A high-performance semi-supervised learning method for text chunking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL '05)*: (pp. 1-9). Michigan: Omnipress.

Arabie, P., and Hubert, L. J. (1996). An Overview of Combinatorial Data Analysis. In P. Arabie, L. Hubert, G. De Soete, P. Arabie, L. Hubert, and G. De Soete (Edits.), *Clustering and Classification:* (*pp. 8-17*). New Jersey: World Scientific Publishing.

Araque, A. and Saldarriaga, J. (2011). *Optimización Operacional de Redes de Distribución de Agua Potable con el Fin de Maximizar la Uniformidad de Presiones en una Red de Agua Potable*.Catedra PAVCO UNIANDE. Disponible en: http://pavco.com.co/index.php?view=page&id=155 (acceso Febrero 2 2013).

AVSA (Aguas de Valencia S.A). (2013). *Sectorizacion.*
Disponible en:

https://www.aguasdevalencia.es/portal/web/Tecnologia/Tecnologia/GestionRedes/Sectorizacion.html (acceso Junio 1 2013).

Biggs, N., Lloyd, E. and Wilson, R. (1986). *Graph Theory, 1736-1936*. New York: Oxford University Press.

Biswal, P.C. (2005). *Discrete Mathematics and Graph Theory*. New-Dehli: PHI learning Private Limited.

Bollobás, B. (1998). *Modern Graph Theory*. New York: Springer-Verlag.

Boswell. (2002). *Introduction to Support Vector Machine*. Disponible en: http://dustwell.com/PastWork/IntroToSVM.pdf (acceso 2 Mayo 2013).

Brock, G., Pihur, V., Datta, S. and Datta, S. (2008). ClValid. An R Package for Cluster Validation. *Journal of Statistical Software* , 25 (4): 1-22.

Chen, D and Xu, B. (2003). Geometric Algorithms for Agglomerative Hierarchical Clustering. In Warnow, T.and Zhu, B. (Eds.), *Proceedings of the 9th annual international conference on Computing and combinatorics (COCOON'03)*: *Vol 9*, (pp. 30-39). Montana: Springer-Verlag.

Cimiano, P., Hotho, A., and Staab, S. (2004). Comparing Conceptual, Divise and Agglomerative Clustering for Learning Taxonomies from Text. In R, López de Mántaras. and L, Saitta. (Eds.), *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004): (pp 435-439)*. Valencia: IOS press

Cord, M and Cunningham, P. (2008). *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval (Cognitive Technologies)*. Santa Clara, CL: TELOS.

Cover, T.M. (1965). Geometrical and Statistical Propierties of Systme of Linear Inequalities with Applications in Pattern Recognition. *IEEE Transactions on Electronic Computers*, EC-14 (3): 326-334.

Cristiani, N. and Shawe-Taylor J. (2000). *An Introduction to Support Vector Machines: And Other Kernel-Based Learning Methods*. Cambridge: Cambridge University Press.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *Inter Journal* , Vol. Complex Systems.

CVIA (Centro Virtual de Información del Agua). (2010). *Agua. Guia para Organismos Operadores. Agua Potable, Alcantarillado y Sanamiento*. IMTA.
Disponible en:
http://www.imta.gob.mx/compaps/images/stories/pdf/guia_para_organismos_operadores.pdf (acceso Febrero 2013).

Daya, P., Sahani, M. and Deback, G. (1999). *Unsupervised Learning*. Massachusetts: The MIT Press.

Delgado-Galván, X., Pérez-García, R., Izquierdo, J. and Mora-Rodríguez, M. (2010). An analytic hierarchy process for assessing externalities in water leakage management. *Mathematical and Computer Modeling*, 52: 1194-1202.

Di Nardo, A. and Di Natale, A. (2011a). A Hueristic Support Methodology Based on Graph Theory for District Metering of Water Supply Networks. *Engineering Optimization*, 43 (2): 193-212.

Di Nardo, A., Di Natale, M., Santonastaso, G. and Ventincinque, S. (2011b). Graph Partioning for automatic sectorization of a water distribution system. In *Proceedings of the 11th International Conference on Computing and Control for Water Industry (CCWI). Urban Water Management: Challagnes and Opportunities: Vol 3,* (*pp. 841-846*). Exeter: Centre for Water Systems, University of Exeter.

Di Nardo, A., Di Natale, A., Santonastaso, G., Tzatchkov, G., and Alcocer-Yamanaka, V., (2013a). Water Network Sectorization Based on Graph Theory and Energy Perfomance Indices. *Journal onf Water Resorces Planning and Management* , 139(1).

DiNardo, A., DiNatale, M., Santonastaso, G., Tzatchkov, V., and Alcocer-Yamanaka, V. (2013b). Water Networks Sectorization based on Genetic Algorithms and Minimum Dissipated Powers Paths. *Water Sience and Technology: Water Supply*, (2):193-211.

Ding, W., Jiamthapthaksin, R., Parmar, R., Jiang, D., Tomasz, F., Stepinski. and Christoph, F.E.(2008). Towards region discovery in spatial datasets. In Washio, T., Inokuchi., Suzuki, E. and Ting, K.M. (Eds.), *Proceedings of the 12th Pacific-Asia conference on Advances in knowledge discovery and data mining (PAKDD'08)*: (pp 88-99). Osaka: Springer-Verlag.

Divakaran, A. (2009). *Multimedia Content Analysis: Theory and Applications*. New Jersey: Springer-Verlag.

Dunn, J.C. (1974). Well-Separated Clusters and Optimal Fuzzy Partitions. *Journal of Cybernetics*, 4 (1): 95-104.

DVGW (German Technical Association of Gas and Water). (2003). *Arbeitsblatt W 392 Rohrnetzinspektion und Wasserverluste - Maßnahmen, Verfahren*. Bonn: Deutsche Vereinigung des Gas- und Wasserfaches (DVGW).

Efron, B., Halloran, E. and Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Science*: *Vol. 93*, (*pp. 13429-13434*). Washington: USA National Academy of Science.

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

Farley, M., Wyeth, G., Istandara, A. and Sher, S. (2008). *The Manager's Non-Revenue Water Handbook. A Guide to Understanding Water Losse*. Bangkok: Ranhill Utilities and United States Agency for International Development (USAID). Disponible en:

http://www.waterlinks.org/library/non-revenue-water/nrw-handbook (acceso Marzo 2 2013).

Farris, J.S. (1969). On the Cophenetic Correlation Coefficient. *Systematic Zoology*, Vol. 18(3): 279-285.

Ghahramani, Z. (2004).Unsupervised Learning. In Bousquet, O., Raetsch, G. and Von-Luxburg, U. (Eds.), *Advanced Lectures on Machine Learning LNAI 3176*. Tübingen: Springer-Verlag.

GIZ (Deutsche Gesellschaft für Internationale Zusammenarbeit GmbH)., VAG (VAG Armaturen GmbH)., FHNW (Fachhochschule Nordwestschweiz)., and KIT (Karlsruhe Institute of Technology). (2011). *Guidelines for Water Loss Reduction: A Focus on Pressure Management*. Bonn: Eschborn.
Disponible en: http://www.waterloss-reduction.com/?id=8 (acceso Agosto 12 2012).

Goepel, K. (2013). *AHPcal excel template*. BPMSG (Business Performance Management). Disponible en: http://bpmsg.com/new-ahp-excel-template-with-multiple-inputs/(acceso Noviembre 15 2012).

Gonçalves, L., Rodriguez, R., Amaral, A.J., Karasawa, M. and Sudré, C. (2008). Comparison of Multivariate Statistical Algorithms to Cluster Tomato Heriloon Accessions. *Genetic and Molecular Research* , 7 (4):1289-1297.

Goset, E. (2009). *Discrete Mathematics with Proff*. New Jersey: John Wiley & Sons, Inc.

Hamfelt, A., Karlsson, M., Thierfelder, T. and Valkovsky, V. (2011). Beyond k-means: Clusters identification for GIS. In Popovich, V.V., Claramunt,C., Devogele, T., Schrenk,M. and Korolenko, K. (Eds.), *Information Fusion and Geographic Information Systems, Lecture Notes in Geoinformation and Cartography*: (pp. 93-105). Newport: Springer-Verlag.

Han,J., Kamber, M. and Pei, J. (2006). *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann Publishers Inc.

Han, J. Kamber, M. y Pei, J. (2012). *Data mining concepts and techniques*. San Francisco: Morgan Kaufmann Publishers Inc.

Handl, J., Knowles, J. and Kell, D. (2005). Computational Cluster Validation in Post-Genomic Data Analysis. *Bioinformatics* (21): 3201-3212.

Hastie, T., Tibshirani, R, and Friedman, J. (2009). *Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer-Verlag.

Herrera, M. (2011a). *Improving Water Networks Management by Efficient Division into Supply Clusters*. PhD Thesis, Universitat Politecnica de Valencia, Valencia, España.

Herrera, M. (2011b). Water Supply Cluster by Multi-Agent Based Aproach. In

Izquierdo,J., and Pérez-García, R. (Eds.), *Aplication of Intelligent Data Analysis in Urban Water System (pp. 254-264)*. Valencia: Universidad Politecnica de Valencia.

Hunaidi, O. and Brothers, K. (2007). Optimun Size of District Metered Areas. In *Water Loss Specialist Conference, International Water Association:* (*pp. 57-66*). Londres: International Water Association (IWA).

IMTA (Instituto Mexicano de Tecnología del Agua). (2008). *Informe Anual 2008. Conocimiento y Tecnologías para la Gestión Sustentable del Agua*. Morelos: IMTA.

IWA(International Water Association). (2000). *Performance Indicators for Water Supply Services*. Londres: IWA.

Izquierdo, J., Montalvo, I., Pérez-García, R. and Herrera, M. (2008). Sensitivy Analisis to Assess the Relative Importance of Pipes in Water Distribution Networks. *Mathematical and Computer Modelling*, 48 (Issues 1–2): 268–278.

Izquierdo, J., Herrera, M., Montalvo, I., Pérez-García, R. (2011). *Agent-Based Division of Water Distribution Systems into District Metered Areas*. In Cordeiro, J., Ranchordas, A., Shishkov., B. (Eds.), In *Software and Data Technologies: 4th International Conference ICSOFT 2009*: (*pp. 167-180*). Sofia: Springer-Verlag.

Jain, A.K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters.* 31(8): 651-666.

Kanungo, T., Mount, D., Netanyahu, N., Piatko, C., Silverman, R. and Wu, A. (2002). An Efficient k-Means Clustering Algorithm: Analysis and Implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* 24(7): 881-892.

Karatzoglou, A., Samola, A., Hornik, K. and Zeileis, A. (2004). kernlab -- An {S4} Package for Kernel Methods in {R}. *Journal of Satatistical Software* , 11 (9): 1-20.

Karlsson, C. (2008). *Handbook of Research on Cluster Analysis*. London: Edward Elgar Publising Limited.

Kasperczyk, N. and Knickel, K. (2006). *Analytic Hierarchy Process.* Amsterdam: Institute for Environmental Studies. University of Amsterdam.
Disponible en:
http://www.ivm.vu.nl/en/Images/MCA3_tcm53-161529.pdf (acceso Noviembre 3 2012).

Kingdom, B., Liemberger, R. and Marin, P. M. (2006). *The Challenge of Reducing Non-Revenue Water (NRW) in Developing Countries. How the Private Sector Can Help: A look at Perfomance-Based Service Contracting*. Washington, DC:
The World Bank Group.

Kleppen, M. (2011). *Optimization of Small Urban Water Services in Developing Countries by Water Loss Management*. Master Thesis, Telemark University College,

Norway.

Koronacki, J., Zbigniew, R. and Slawomir, T. (2010). *Advances in Machine Learning*. Berlin: Springer-Verlag.

Liu, Z., Liu, G., and Zhou, H. (2011). An image-segmentation method based on improved spectral clustering algorithm. In Qi, L. (Ed.), *Information and Automation: Vol 86, Communications in Computer and Information Science* : (*pp. 178-184*). Guangzhou: Springer-Verlag.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. and Hornik, K.(2013). *Cluster: Cluster Analysis Basics and Extensions*. R package version 1.14.4. Disponible: http://cran.r-project.org/web/packages/cluster/index.html (acceso Octubre 3 2012).

Manning, D., Raghavan, P. and Schutze, H. (2009). *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Manning, C., Raghavan, P. and Schütze, H. (2008). *Introduction to Information Retrieval*. New York : Cambridge University Press.

Mays, L.W. (2000). *Water Distribution System Handbook*. New York: McGraw-Hill.

Mirkin, B. (2013). *Clustering. A Data Recovery Approach*. Boca Raton: CRC Press.

Mooi, E. and Sardtedt, M. (2011). Cluster Analysis. In Mooi, E., and Marko, S. (Edits.), *A Concice Guide to Market Research. The Process, Data, and Methods Usuing IBM SPSS Statistics*: (pp. 9). Munich: Springer- Verlag.

Morrison, J., Stephen, T. and Rogers, D. (2007). *District metered areas: Guidance notes*. London: Water Loss Task Force. International Water Association (IWA). Disponible: http://www.waterlinks.org/sites/default/files/District%20Metered%20Areas%20Guidance%20Notes.pdf (acceso Agosto 3 2012).

Nagabhushana, S. (2006). *Data Warehousing. Implementation and OLAP*. New Dehli: New Age International (P) Ltd.

M. E. J. Newman. (2006). Modularity and community structure in networks. In *Proceedings of the National Academy of Sciences*, 103(23): 8577-8582. Washington: National Academy of Sciences.

M. E. J. Newman. and Girvan, M. (2004). Finding and Evaluaing Community Structure in Networks. *Physical Review E* , 69 (2): 4-19.

Osorio, J. and Orejuela, J. (2008). El Proceso de Análisis Jerárquico (AHP) y la toma de decisiones multicriterios-Ejemplo de Aplicación. *Scientia et Technica* , 2 (39): 247-252.

Pearson, D. and Trow, S. (2008). *Identifying Economic Interventions Againts Water Losses*. In Thorton,J., Sturm, R. and George, K. (Eds.), *Water Loss Control* (Second

Edition ed., pp. 106-109). New York: McGrawHill.

Pearson, D., Fantozzi, M. and Soares, D. (2005). Searching for n2: how does pressure reduction reduce burst frequency. In *International Water Association (IWA) conference proceedings, Leakage 2005*. Halifax: IWA.

Pilcher, R., Hamilton, S., Chapman, H., Field, D., Ristovski, B. and Stapely, S. (2007). *Leak Location and Repair Guidance Notes*. Londres: Operation and Management. IWA.

Podani, J. and Dénes, S. (2006). On Dengrogram-Based Measures of Funcional Diversity. *OIKOS*, 115 (1): 179-185.

RCT (R Core Team). (2013). *The R Stats Package*.
Diponibe en: http://stat.ethz.ch/R-manual/R-patched/library/stats/html/stats-package.html (acceso Septiembre 3 2012).

REDHISP. (2011). *Group of Hydraulic Networks and Presurized Systems*. Universidad Politécnica de Valencia (UPV).
http://www.idmh.upv.es/ (acceso Julio 1 2012).

Romesburg, C. (2004). *Cluster analysis for researches.* North Carolina: Lulu press.

Rossman, L.A (2000). *EPANET 2. Users Manual*. Washington: U.S. Environmental Protection Agency (EPA). Disponible en:
http://www.epa.gov/nrmrl/wswrd/dw/epanet.html (acceso Agosto 1 2013).

Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics* , 20 (1): 53-65.

Saaty, T.L. (2008). Decision Making with the Analytic Hierarchy Process. *International Journal of Services Sciences* , 1 (November 2008): 83-89.

Saldarriaga, J., Naranjo, G. and Rothstein, E. (2008). *Metodología para la Sectorización de Redes Existentes de Distribución de Agua Potable*. Catedra PAVCO UNIANDE.
Disponible en: http://pavco.com.co/index.php?view=page&id=155
(acceso Febrero 2 2013).

Sammut, C. and Webb, G. I. (Eds) (2010). *Encyclopedia of Machine Learning*. New York: Springer.

Sander, J. (1999). *Generalized Desity-Based Clustering for Spatial Data Mining*. Munchen: Herbert Utz Verlag Gmbh.

(SAWUADB) The Southeast Asian Water Utilities Network and Asian Development Bank). (2007). *Data Book of Southeast Asian Water Utilities 2005*. Tokio: Asian Development Bank.
Disponible en: http://www.adb.org/publications/data-book-southeast-asian-water-utilities-2005 (acceso Mayo 2 2013).

Scholkopf, B. and Smola, A. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Massachusetts: MIT Press.

Selim, S. and Ismael, M. (1984). K-means type algorithms: a generalization convergence theorem and characterization of local optimality. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 6 (1): 81-77.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge: Cambridge University Press.

Shental, N. (2004). *From Unsupervised to Semi-Supervised Learnings*. PhD Thesis. The Hebrew University of Jerusalem, Jerusalem, Israel.

Shimodaira, H. (2004a). Approximately Unbiased Test of Regions Using Multistep-Multiscale Bootstrap Resampling. *The Annual of Statistics*, 32 (6): 2616-2641.

Shimodaira, H. (2004b). *Technical Details of the Multistep-Multiscale Bootstraap Resampling*. Research Reports on Mathematical and Computing Sciences. Tokyo: Department of Mathematical and Computing Sciences Tokyo Institute of Technology. Disponible en: http://www.is.titech.ac.jp/~shimo/pub/B403.pdf (acceso Enero 1 2013).

Sokal, R. and Rohlf, F. (1962). The Comparison of Dedrograms by Objective Methods. *Taxon* , 11 (2): 33-40.

Soman, K., Loganathan, R. and Ajay, V. (2009). *Machine Learning with SVM and Others Kernel Methods*. New Delhi: PHI Learning Private Limited.

Suzuki, R and Shimodaira, H. (2006). pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22 (12): 1540-1542.

Taiwo, O.A. (2010). Types of Machine Learning Algorithms. In Yagang Zhang (Ed.), *New Advances in Machine Learning*: (pp.19-48). Rijeka: InTech Disponible en: http://www.intechopen.com/books/new-advances-in-machine-learning/types-of-machine-learning-algorithms (acceso: Enero 3 2013).

Thornton, J., Sturm, R. and Kunkel, G. (2008). *Water Loss Control*. New York: McGrawHill.

Thulasiraman, K. and Swamy, M. N. S. (1992). *Graph Algorithms, in Graphs: Theory and Algorithms*. New Jersey: John Wiley & Sons, Inc.

Todini, E. (2000). Looped Water Distribution Networks Desing Using a Resilience Index. *Urban Water* , 2 (1): 115-122.

Tzatchkov, V., Alcocer Yamanaka, V. H. and Bourguett Ortiz, V. (2008). Sectorización de Redes de Distribución de Agua Potable a Través de Algoritmos Basados en la Teoría de Grafos. *Tláloc AMH*, 1 (40): 14-22.

Vegas, O. (2012). *Herramienta de Ayuda a la Sectorización de Redes de Abastecimiento de Agua Basadas en la Teoría de Grafos Aplicando Diferentes Criterios*. Tesis de Master, Universitat Politecnica de Valencia, Valencia, España.

Von-Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing,* 17(4): 395-416.

Walski, T., Gangemi, D., Kaufman, A. and Malos, W. (2001). Establishing a System Submetering Project. *AWWA Annual Conference*. Washington, DC: AWWA

Wang,J., Shen, X. and Pan, W. (2009). On Efficient Large Margin Semisupervised Learning: Method and Theory. *Journal of Machine Learning,* 10: 719-742.

Wishart, D. (2001). K-means clustering with outlier detection mixed variables and missing values. In Optiz, O. and Shwaiger, M. (Eds.), *Exploratory data analysis in empirical research: proceedings of the 25th Annual Conference of the Gesellschaft für Klassifikation eV*: *Vol 1*, (*pp.216-226*). Munich: Springer-Verlag.

Witten, I. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco: Morgan Kaufmann Publishers Inc.

WWC (World Water Council). (2009). Istambul Water Consensus for Local and Regional Authorities. In *Proceeding of the 5th World Water Forum. Istanbul.* Istambul: World Water Council.

Jummin, X. (2003). *Theory and Application of Graphs.* Norwell: Kluwer Academic Publisher.

Zelnik-Manor, L. and Perona, P. (2004). Self-Tuning Spectral Clustering. In Saul, L.k., Weiss, Y. and Bottou, L. (Eds.), *Proceedings of the 18th Annual Conference on Neural Information Processing Systems (NIPS'04)*: *Vol 17,* (*pp. 1601-1608*). Vancouver: The MIT press.

Zhang, X. and Wang, R.. (2008). Optimization Analysis of Modularity Measures for Network Comunity Detecction. In Zhang, X., Chen, L., Wu, L. and Wang, Y. (Eds.), *The Second International Symposium on optimization and Systems Biology (OSB'08): Vol 1, (pp.13-20).* Lijiang: World Publishing Corporation

9 ANEXO I: Fracción del código Implementado

```
# PASO1: ABRIR LIBRERIAS ---------------
library(cluster)
library(kernlab)
library(igraph)

# PASO 2 DIBUJAR RED NORMAL:-------------------

nodesdat2<-nodesdat
for(id in 1:nrow(nodesdat2)){
  nodesdat2$Demand[nodesdat2$Demand==0] <- NA}
grn<-graph.empty()
)<-names
.dist=0.4, edge.arrow.size=0.0, vertex.label.cex=0.5,
main="DEMANDAS",  frame=TRUE)


# PASO 3 PREPARACION DE TABLAS: ------------------------------

nodes1<-nodesdat[!(nodesdat$Demand==0),]
as<-1:nrow(nodes1)
nodes.as<-cbind(nodes1,as)


# PASO 4 DIBUJO DE RED DE DISTIBUCION ----------------------

="red", edge.arrow.size=0.0,vertex.label.cex=0.5, main="RED DE
DISTRIBUCION",  frame=TRUE)

# PASO 5 ESTIMACION DEL NUMERO DE SECTORES ----------------

# standarizar datos

nodoshc<-nodes1
nodos.hc2<-scale(nodos.hc1)
```

```
#Clustering jer??rquico

datos<-daisy(nodos.hc2, metric="euclidean")

x <- rect.hclust(hc1, h = 3.0)

d1 <- datos
d2 <- cophenetic(hc1)
cor(d1, d2)

##creacion de de la matriz de disimilaridad de demandas (con es-
tandarizacion)
ds.demanda<-as.matrix(daisy(demanda.sc, metric = "euclidean",
= FALSE))
##kernel de la matriz de disimilaridad
coorx<-as.kernelMatrix(ds.coordenadasx)
diag(coorx)<-1

##creacion de de la matriz de disimilaridad de elevacion (con
)
ds.emisores<-as.matrix(dist(emisores.sc))
##kernel de la matriz de disimilaridad
emisor<-as.kernelMatrix(ds.emisores)
diag(emisor)<-1


#SECCION DE SUMA DE LOS KERNELS

#Suma de las matrices kernel
K<-(0.4*knet)+((1-
(max(nodes2[ ,7]))
V(g)$color <- colbar[comps2]

plot(g, layout=coor2, ver-
tex.size=4,vertex.label=nodes2[,7],edge.arrow.size=0.0,vertex.la
bel.dist=0.7, vertex.label.cex=0.5,main="CLUSTERS EN RED DE
DISTRIBUCION",  frame=TRUE)


#PLOT RED COMPLETA CON SECTORES

comps2 <-nodesdatf[ ,7]
colbar <- topo.colors(max(nodesdatf[ ,7]))
V(grf2)$color <- colbar[comps2]

plot(grf2, layout=coor1, ver-
tex.size=4,vertex.label=nodesdatf[,7],edge.arrow.size=0.0,vertex
.label.dist=0.4, vertex.label.cex=0.5,main="CLUSTERS EN RED
COMPLETA",  frame=TRUE)


#Dibujar Lineas a Cerrar

$sc[id]}
```

```
comps2 <-nodes3d[ ,3]
colbar <- topo.colors(3)


plot3d(nodes3d, top=FALSE , box= TRUE, axis= TRUE, type="s", #

PASO 8: GRAFICO DE SILUETA DE PARTICION SELECCIONADA (OPCIONAL)
-------------

ar <- hc1
si3 <- silhouette(cutree(ar, k = 3), # k = 4 gave the same as
pam() above
                  daisy(nodes3))

      plot(si3, nmax = 80, cex.names = 0.5, col=rainbow(3))
```