UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

EXIGENCIAS DE CALIDAD DE SUMINISTRO EN BASE A DENSIDAD DE CONSUMO MEDIANTE TÉCNICAS DE MINERÍA DE DATOS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL ELECTRICISTA

CLAUDIO NICOLÁS FLORES CARTES

PROFESOR GUÍA:
FERNANDO FLATOW GARRIDO

MIEMBROS DE LA COMISIÓN:
JUAN ALBERTO BRAVO
NELSON MORALES OSORIO

SANTIAGO DE CHILE
2014

RESUMEN DE LA MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL ELECTRICISTA
POR: CLAUDIO FLORES CARTES
FECHA: 31/12/2013
PROF. GUÍA: FERNANDO FLATOW G.

EXIGENCIAS DE CALIDAD DE SUMINISTRO EN BASE A DENSIDAD DE CONSUMO MEDIANTE TÉCNICAS DE MINERÍA DE DATOS

Se entenderá por calidad de servicio el conjunto de propiedades y estándares normales que, conforme la Ley y el Reglamento Eléctrico son inherentes a la actividad de distribución concesionada, y constituyen las condiciones bajo las cuales dicha actividad debe desarrollarse. Esta incluye la calidad de servicio comercial, la calidad de suministro y dentro de ésta la calidad de producto y la disponibilidad del servicio eléctrico. Se entenderá por calidad de suministro la disponibilidad del servicio, la cual se medirá a través de las interrupciones de suministro: cantidad y duración de ellas. Estas definiciones encuentran su sustento legal en la Ley General de Servicios Eléctricos (DFL 4/2006), el Reglamento Eléctrico (DS 327/1997) y las normas técnicas.

Actualmente las exigencias de calidad de suministro establecen índices ligados a las empresas y zonas de concesión de distribución, estableciéndose exigencias en base a índices poblacionales, kilómetros de redes de distribución y criterios que no dan cuenta de la densidad de consumo, criterio que determina necesariamente la calidad que se necesita. El presente trabajo propone una metodología de asignación de grupos de consumidores dentro del territorio nacional a índices de calidad representativos mediante técnicas de minería de datos (clustering), con independencia de la empresa suministradora, la topología de las redes existentes o las distinciones demográficas. La agrupación se basa en parámetros geográficos y de consumo anual de energía.

La metodología se deriva de una serie de tentativas de agrupación de datos reales correspondientes a los consumos de la octava y novena región, realizada con tres programas computacionales que implementan una serie de algoritmos de clustering.

La metodología se basa en tres etapas: una etapa de pre-procesamiento donde se llevan los datos a un formato manejable y se filtran para eliminar datos no pertinentes para el análisis, una etapa de clustering en donde los datos son agrupados a través del algoritmo K-means. Luego se realiza el cálculo de los vecinos más cercanos para cada cluster, el cual orienta en la elección del parámetro épsilon para el algoritmo DBSCAN utilizado para realizar una nueva agrupación basada en densidad para cada uno de los cluster, produciéndose subclusters cuya característica es tener distintas densidades. Para todas las etapas de clustering se utiliza una métrica basada en datos geográficos (coordenadas x e y de los datos) y el consumo anual de energía. La tercera etapa de post-procesamiento permite asociar a cada uno de los subclusters un índice de densidad de consumo por área. Finalmente, se escogió un esquema de regresiones lineales con los índices de densidad de consumo para determinar las zonas de exigencias.

Se aplica la metodología y se contrasta con el esquema actual, a través del análisis de algunos casos de interés. Se comprueba que esta metodología corrige limitantes que el esquema actual no considera.

Tabla de contenido

CAPÍTULO 1 - INTRODUCCIÓN 1
1.1 Motivación 1
1.2 Antecedentes 1
1.2.1 Antecedentes generales 1
1.2.2 Antecedentes específicos 1
1.3 Objetivos 2
CAPÍTULO 2 - SECTOR ELÉCTRICO EN CHILE 3
2.1 Introducción 3
2.2 Segmentos del mercado eléctrico chileno 3
2.2.1 Generación 3
2.2.2 Transmisión 5
2.2.3 Distribución 6
2.3 Marco Legal, Reglamentario y Normativo 7
2.4 Institucionalidad 8
2.5 Calidad de suministro 11
2.5.1 Parámetros de continuidad de suministro 11
2.5.2 Índices de confiabilidad de la red usados en el mundo 12
2.5.3 Índices de confiabilidad de la red usados en Chile 13
CAPÍTULO 3 - MINERÍA DE DATOS 18
3.1. Introducción 18
3.2. Orígenes y Tareas de la Minería de datos 18
3.3. Clustering 19
3.3.1 Métodos de Clustering. 20
3.3.2 Técnicas de Clustering 21
3.4. Validación de Clusters 26
3.5. Software de Minería de Datos 27
3.5.1. CLUTO 27
3.5.2. RapidMiner 27
3.5.3. Weka 27
CAPÍTULO 4 – METODOLOGÍA 28
4.1. Pre-procesamiento de los datos 28
4.2. Evaluación del desempeño del software de minería de datos y algoritmos de clustering. 29
4.2.1. Selección de datos de prueba. 30
4.3 Post-procesamiento de los datos 31
CAPÍTULO 5 – RESULTADOS Y ANÁLISIS 32
5.1. Pre-procesamiento de los datos. 32
5.1.1. Datos no considerados 32
5.1.2. Información de los datos escogidos 32
5.2. Evaluación del desempeño de software de minería de datos y algoritmos de clustering 35
5.3. Resultados de clustering con el total de los datos. 38
5.3.1. Resultados 1 38
5.3.2. Resultados 2 45
5.3.3. Resultados 3 51
5.3.4. Resultados 4 51
5.3.5. Resultados 5 60
5.3.6. Resultados 6 67
5.4. Metodología Propuesta 86
CAPÍTULO 6 - CONCLUSIONES 90
CAPÍTULO 7 - BIBLIOGRAFÍA 92

CAPÍTULO 8 – ANEXOS **94**
ANEXO A: LISTADO DE ARCHIVOS ENTREGADOS 94
Resultados 1 ..... *94*
Resultados 2 ..... *95*
Resultados 3 ..... *96*
Resultados 4 ..... *96*
Resultados 5 ..... *97*
Resultados 6 ..... *97*
Otros *98*

Índice de tablas

Tabla 2.1: Límites máximos para índices de interrupción de suministro. 17
Tabla 3.1: Algoritmo básico de K-means 21
Tabla 3.2: Algoritmo DBSCAN 25
Tabla 4.1: Métodos de clustering presentes en los programas computacionales a utilizar....29
Tabla 5.1: Distribución de la cantidad de datos según región 33
Tabla 5.2: Estadísticos de los consumos seleccionados....34
Tabla 5.3: Resultados de las pruebas de software y algoritmos en tiempo y calidad, para la selección de datos. 36
Tabla 5.4: Resultados de la selección de algoritmos....37
Tabla 5.5: Resultados clustering en tiempo y calidad, para el total de datos. 39
Tabla 5.6: Consumo neto de cada cluster, en orden decreciente 48
Tabla 5.7: Área equivalente de cada cluster, en orden decreciente....49
Tabla 5.8: Área equivalente, consumo neto y densidad de consumo por área para cada cluster. 49
Tabla 5.9: Resumen de la obtención de parámetros épsilon para la solución con atributos “x” e “y” 55
Tabla 5.10: Resultados subclustering cluster 1, para ε = 145 fijo. 61
Tabla 5.11: Soluciones para el cluster c1 variando parámetros MinPts y eps. 62
Tabla 5.12: Valor de épsilon considerado para cada cluster 70
Tabla 5.13: Regresión lineal para rectas de zonas de exigencia 78
Tabla A.1 - Ruta: Clustering\Resultados 1\Formato CLUTO\Resultados\....92
Tabla A.2 - Ruta: Clustering\Resultados 1\Formato CLUTO\....92
Tabla A.3 - Ruta: Clustering\Resultados 1\Formato RapidMiner\Resultados\....92
Tabla A.4 - Ruta: Clustering\Resultados 1\Formato RapidMiner\....92
Tabla A.5 - Ruta: Clustering\Resultados 1\Formato WEKA\Resultados\....93
Tabla A.6 - Ruta: Clustering\Resultados 1\Formato WEKA\.... 93
Tabla A.7 - Ruta: Clustering\Resultados 1\.... 93
Tabla A.8 - Ruta: Clustering\Resultados 2\Área.... .93
Tabla A.9 - Ruta: Clustering\Resultados 2\cambio de distancia....93
Tabla A.10 - Ruta: Clustering\Resultados 2\Distancia Momento....93
Tabla A.11 - Ruta: Clustering\Resultados 2\....94
Tabla A.12 - Ruta: Clustering\Resultados 3\....94
Tabla A.13 - Ruta: Clustering\Resultados 4\Clusters....94
Tabla A.14 - Ruta: Clustering\Resultados 4\resultados\Gráficos k-nn....94
Tabla A.15 - Ruta: Clustering\Resultados 4\resultados\....94
Tabla A.16 - Ruta: Clustering\Resultados 4\....94
Tabla A.17 - Ruta: Clustering\Resultados 5\Clusters....95

Tabla A.18 - Ruta: Clustering\Resultados 5\Soluciones……95
Tabla A.19 - Ruta: Clustering\Resultados 5\……95
Tabla A.20 - Ruta: Clustering\Resultados 6\Análisis de Casos……95
Tabla A.21 - Ruta: Clustering\Resultados 6\Clusters……95
Tabla A.22 - Ruta: Clustering\Resultados 6\……95
Tabla A.23 - Ruta: Preprocesados\……96
Tabla A.24 - Ruta: \……96

Índice de figuras

Figura 3.1: El proceso de descubrimiento de conocimientos en bases de datos (KDD). 18
Figura 3.2: La minería de datos como una confluencia de varias disciplinas. 19
Figura 3.3: Análisis de *Clusters* 19
Figura 3.4: *Clustering* Particional 20
Figura 3.5: *Clustering* Jerárquico 20
Figura 3.6: Algoritmo K-means para encontrar tres clusters en datos de prueba 22
Figura 3.7: Puntos de núcleo, de borde y de ruido 24
Figura 3.8: Distancia a los k-nn ordenados ascendentemente. 25
Figura 3.9: Ejemplo de utilización del SSE como indicar para determinar el número de clusters (K) 26
Figura 4.1: Datos seleccionados, (a) vista 2D, (b) vista 3D. 30
Figura 5.1: Distribución espacial de los datos considerados. 33
Figura 5.2: Distribución de la cantidad de consumos según región. 33
Figura 5.3: Distribución espacial del 5% superior de los consumos. 34
Figura 5.4: Histograma de consumos año 2012 de la octava y novena región 35
Figura 5.5: Histogramas de consumos para la solución CLUTO - direct 40
Figura 5.6: Histogramas de consumos para la solución RapidMiner – K-means. 40
Figura 5.7: Distribución espacial de los consumos según la solución RapidMiner – K-means. 41
Figura 5.8: Histogramas de consumos para la solución Weka – FarthestFirst 42
Figura 5.9: Distribución espacial de los consumos según la solución Weka – FarthestFirst. 42
Figura 5.10: Distribución de la cantidad de consumos por clusters para la solución Weka – FarthestFirst con K = 10 43
Figura 5.11: Histograma de consumos para la solución Weka – FarthestFirst con K = 10 43
Figura 5.12: Distribución espacial de los consumos según la solución Weka – FarthestFirst con K = 10. . 44
Figura 5.13: (a) Distribución espacial de los clusters de alto consumo (b) Distribución espacial de los clusters de bajo consumo. 45
Figura 5.14: Agrupación de clusters por zonas 45
Figura 5.15: Distribución de la cantidad de datos por clusters para la solución geográfica y cálculo de consumo por área 47
Figura 5.16: Histogramas de consumo por clusters para la solución geográfica y cálculo de consumo por área. 48
Figura 5.17: Gráfico de densidad de consumo por área, ordenado decrecientemente 50
Figura 5.18: Disposición geográfica de los consumos con la representación de su área equivalente. 50
Figura 5.19: Distribución de la cantidad de datos por clusters para la solución con atributos “x” e “y” 52
Figura 5.20: Histogramas de consumo por clusters para la solución con atributos “x” e “y” 52
Figura 5.21: Disposición geográfica de los consumos con la representación de su área equivalente para la solución (X, Y). 53
Figura 5.22: Vecinos más cercanos para el cluster 1, solución con atributos “x” e “y” 54

Figura 5.23: Cluster 1 en la solución (X.Y) 55
Figura 5.24: Solución de subclustering C01 -1 (MinPts = 500, ε = 1.200), para solución con atributos “x” e “y”. 56
Figura 5.25: Solución de subclustering C01 – 2 (MinPts = 500, ε = 2.800), para solución con atributos “x” e “y”. 56
Figura 5.26: Solución de subclustering C01 - 3 (MinPts = 500, ε = 5.200), para solución con atributos “x” e “y” 57
Figura 5.27: Subclustering C06 – 1.....58
Figura 5.28: Subclustering C09 - 1 58
Figura 5.29: Subclustering C10 - 1 59
Figura 5.30: Distancia ordenada de los vecinos más cercanos, para el cluster 1. (a) Utilizando parámetros “x”,”y” y consumo con distancia 3d. (b) Utilizando sólo atributos “x”,”y” y distancia 2d.....61
Figura 5.31: Resultados de las soluciones para el cluster c1 variando parámetros MinPts y eps. 62
Figura 5.32: Densidad de consumo por área para solución calculada en Resultados 5.....63
Figura 5.33: Densidad de consumo por área, divida en terciles.....64
Figura 5.34: Densidad de consumo por área, divida en mitades, sin considerar a los subcluster de ruido64
Figura 5.35: Código de colores para la comparación entre esquema actual y esquema propuesto.....64
Figura 5.36: Distribución de la cantidad de datos en cada zona de exigencia en el esquema actual y en el esquema propuesto por terciles.....65
Figura 5.37: Disposición geográfica de los consumos en el esquema actual y en el esquema propuesto por terciles. 65
Figura 5.38: Distribución de la cantidad de datos en cada zona de exigencia en el esquema actual y en el esquema propuesto por mitades y ruido como zona de menor exigencia 66
Figura 5.39: Disposición geográfica de los consumos en el esquema actual y en el esquema propuesto por mitades y ruido como zona de menor exigencia.....66
Figura 5.40: Distribución de la cantidad de datos por clusters para la solución k-means, k=6, considerando parámetros geográficos y de consumo. 68
Figura 5.41: Disposición geográfica de los consumos con la representación de su área equivalente para la solución k-means, k=6, considerando parámetros geográficos y de consumo. 68
Figura 5.42: Disposición geográfica de los consumos en las capitales provinciales.....69
Figura 5.43: Resultados de graficar ascendentemente los vecinos más cercanos para los 6 clusters 69
Figura 5.44: Resultados DBSCAN (subclustering por densidad) cluster c0. 71
Figura 5.45: Resultados DBSCAN (subclustering por densidad) cluster c1. 72
Figura 5.46: Resultados DBSCAN (subclustering por densidad) cluster c2. 73
Figura 5.47: Resultados DBSCAN (subclustering por densidad) cluster c3. 74
Figura 5.48: Resultados DBSCAN (subclustering por densidad) cluster c4. 75
Figura 5.49: Resultados DBSCAN (subclustering por densidad) cluster c5. 76
Figura 5.50: Densidad de consumo por área. 77
Figura 5.51: Densidad de consumo por área con rectas de zonas de exigencia. 78
Figura 5.52: Distribución de la cantidad de datos en cada zona de exigencia en el esquema actual y en el esquema propuesto por pendientes. 79

Figura 5.53: Disposición geográfica de los consumos en el esquema actual y en el esquema propuesto por pendientes. 79
Figura 5.54: Comparación de la zona Temuco - Padre las Casas con el esquema vigente y propuesto, vista general 80
Figura 5.55: Comparación zona Temuco - Padre las Casas: esquema vigente 81
Figura 5.56: Comparación zona Temuco - Padre las Casas: esquema propuesto 81
Figura 5.57: Comparación de la zona Villarrica - Pucón con el esquema vigente y propuesto, vista general 82
Figura 5.58: Comparación zona Villarrica - Pucón: esquema vigente 83
Figura 5.59: Comparación zona Villarrica - Pucón: esquema propuesto 83
Figura 5.60: Comparación del Gran Concepción con el esquema vigente y propuesto, vista general 84
Figura 5.61: Comparación zona Gran Concepción: esquema vigente 85
Figura 5.62: Comparación zona Gran Concepción: esquema propuesto 85
Figura 5.63: Cálculo del área equivalente de un subcluster 88
Figura 5.64: Diagrama de la metodología propuesta. 89

Glosario

AFR:	Aporte Financiero Reembolsable
Asai:	Average service availability index
ATD:	Área Típica de Distribución
CDEC:	Centro de Despacho Económico de Carga
CNE:	Comisión Nacional de Energía
FMIK:	Frecuencia media de interrupción por kVA
FMIT:	Frecuencia media de interrupción por transformador
KDD:	Knowledge Discovery in database
LGSE:	Ley General de Servicios Eléctricos
NTCO	Norma Técnica de Conexión y Operación
NTSyCS	Norma Técnica de Calidad y Seguridad de Servicio
PMGD	Pequeños Medios de Generación Distribuidos
Saidi:	System average interruption duration index
Saifi:	System average interruption frequency index
SEC:	Superintendecia de Electricidad y Combustibles
SIC:	Sistema Interconectado Central
SING:	Sistema Interconectado del Norte Grande
SSE:	Sum of squares errors
TTIK:	Tiempo total de Interrupción por kVA
TTIT:	Tiempo total de Interrupción por transformador
VAD:	Valor Agregado de Distribución

Capítulo 1 - Introducción

1.1 Motivación

Actualmente las exigencias de calidad de suministro para usuarios de distribución está determinada someramente en el Reglamento Eléctrico (Decreto Supremo N°327/97) y en la Resolución Ministerial Exenta Nº 53, del Ministerio de Economía, Fomento y Reconstrucción, de 23 octubre de 2006. Estas disposiciones establecen criterios ligados a las empresas y zonas de concesión de distribución, estableciéndose exigencias de calidad de suministro en base a índices poblacionales y kilómetros de redes de distribución. Estos criterios hacen referencia a variables asociadas a la actividad humana, los que no dan cuenta de la intensidad del recurso electricidad en un espacio geográfico determinado. Por otra parte, la utilización de kilómetros de redes reconoce la infraestructura de las empresas existentes entregándoles índices que deben cumplir, lo que no se condice con el criterio de eficiencia asociado a la tarificación por empresa modelo donde se define una red eficiente y sobre ésta se asocian índices de calidad con independencia de las inversiones que el concesionario ha realizado.

1.2 Antecedentes

1.2.1 Antecedentes generales

La calidad de suministro está determinada en el Reglamento Eléctrico para cada área típica de distribución (ATD), las cuales agrupan a las empresas concesionarias cuyos costos medios son similares entre sí, sin considerar las características de las comunas a las cuales dan servicio, pudiendo suceder que dos comunas semejantes estén bajo concesiones en distintas áreas típicas con lo cual se le imponen exigencias de calidad distintas o, aún peor, una misma comuna bajo distintas áreas de concesión y distintos índices.

Se plantea crear un nuevo esquema para agrupar a los clientes con el fin de evaluar los niveles de calidad de suministro en base a parámetros geográficos y de densidad de consumo. En ese sentido, este trabajo puede ser aprovechado para futuros cambios en la reglamentación del sector distribución y que al dar cuenta de lo discutido anteriormente, lo corrijan considerando una calidad de suministro acorde con la realidad de las comunas, independiente de la concesionaria que le preste servicios.

1.2.2 Antecedentes específicos

Se dispone de datos geo-referenciados de consumos de opción tarifaria BT1 (Para mayor información sobre las opciones tarifarias para clientes regulados remitirse a [1]) para clientes de la octava y novena región, entregados por las empresas concesionaras a la Superintendencia de Electricidad y Combustibles (SEC), por su obligación legal de trasparencia [2]

1.3 Objetivos

Objetivo General: Proponer una metodología de asignación de grupos de consumidores dentro del territorio nacional a índices de calidad de suministro representativos mediante técnicas de minería de datos, reconociendo la densidad de carga con independencia de la empresa suministradora, la topología de las redes existentes o las distinciones demográficas.

Objetivo Específicos:

- Determinar la o las técnicas de clustering más apropiadas y sus respectivos parámetros con el fin de agrupar a los clientes en conjuntos de distintos niveles de exigencias de calidad de suministro, independientemente de las áreas de concesión de la cual formen parte.

- Agrupar los consumos en base a parámetros geográficos y densidad de consumo.

- Definir, la cantidad adecuada de grupos de exigencia distinta y, para cada grupo, definir los valores máximos para índices de calidad de suministro acorde con las características de densidad de consumo del grupo.

Capítulo 2 - Sector Eléctrico en Chile

A continuación se expone una revisión bibliográfica que describe el funcionamiento del sector eléctrico en Chile revisando los segmentos que lo componen, el marco legal, regulatorio y normativo, la institucionalidad vigente y profundiza en el tema de calidad de servicio en distribución.

2.1 Introducción

El sector eléctrico en Chile está compuesto por las actividades de generación, transmisión y distribución de suministro eléctrico. Estas actividades son desarrolladas por empresas que son controladas en su totalidad por capitales privados, mientras que, de acuerdo al orden constitucional y a la legislación vigente, el Estado ejerce funciones de regulación y fiscalización.

Este trabajo se centrará en el ámbito de la distribución. Sin embargo es necesario dar un contexto general por lo que se desarrollarán reseñas descriptivas de los 3 segmentos del mercado eléctrico mencionados anteriormente, de las normativas asociadas para el desarrollo de estas actividades y de la institucionalidad vigente.

Finalmente, se ahondará en los índices asociados a la calidad de suministro en distribución.

2.2 Segmentos del mercado eléctrico chileno

2.2.1 Generación

La actividad de generación corresponde al sector del mercado eléctrico encargado de la producción de la electricidad, a través de la transformación de los distintos energéticos primarios en energía eléctrica. Esta función se lleva a cabo a través de diferentes centrales, de propiedad privada, construidas acorde con las características geográficas y recursos naturales disponibles, así como también a partir de los factores económicos y tecnológicos presentes en el país [3].

En Chile esta actividad es realizada principalmente mediante centrales hidráulicas y térmicas. Sin embargo, debido a los cambios normativos recientes y a las crecientes exigencias en materia de responsabilidad ambiental, las empresas se han visto en la necesidad de introducir generación a partir de energías renovables no convencionales, como lo son la eólica, solar fotovoltaica, biomasa y minihidro.

En Chile, con la excepción de los pequeños sistemas aislados de Aysén y Punta Arenas, las actividades de generación se desarrollan en torno a dos sistemas eléctricos: el Sistema Interconectado Central (SIC), que cubre desde el sur de la II Región (quebrada de Paposo) a la X Región (localidad de Quellón), abasteciendo el consumo de aproximadamente 92% de la población nacional; y el Sistema Interconectado del Norte Grande (SING), que abarca la I y II regiones, y cuyos principales usuarios son empresas mineras e industriales. En cada uno de estos grandes sistemas, la generación eléctrica es coordinada por su respectivo e independiente Centro de Despacho Económico de Carga (CDEC).

El modelo de negocio de la actividad de generación está basado principalmente en contratos de largo plazo entre generadores y clientes, que especifican el volumen, el precio y las condiciones para la venta de energía y potencia, aunque también participan del llamado mercado spot en el cual la empresas generadoras pueden comercializar sus excedentes de energía, al precio spot correspondiente al costo marginal horario y los de potencia al precio de nudo de la potencia.

La Ley General de Servicios Eléctricos (LGSE) [4], en su artículo 147, establece dos tipos de clientes: clientes libres y clientes regulados.

Son clientes libres principal y obligatoriamente los consumidores cuya potencia conectada es superior a 2 MW, por lo general de tipo industrial o minero, y adicionalmente aquellos con potencia conectada de entre 500 kW y 2 MW que hayan optado -por un período de cuatro años- por la modalidad de precio libre. Estos clientes no están sujetos a regulación de precios, y por lo tanto las empresas generadoras y distribuidoras pueden negociar libremente con ellos los valores y condiciones del suministro eléctrico.

Son clientes regulados, por su parte, los consumidores cuya potencia conectada es igual o inferior a 2 MW, y adicionalmente aquellos clientes con potencia conectada de entre 500 kW y 2 MW que no hayan optado -también por cuatro años- por un régimen de tarifa libre. Estos clientes reciben suministro desde las empresas distribuidoras, las cuales deben desarrollar licitaciones públicas para asignar los contratos de suministro de energía eléctrica que les permitan satisfacer su consumo. (Art. 131 de la Ley)

De acuerdo a los cambios introducidos a la ley eléctrica en mayo del año 2005, a través de la ley corta 2 (LC2 [5]), los nuevos contratos que asignen las empresas distribuidoras para el consumo de sus clientes a partir de 2010, deben ser adjudicados a las empresas generadoras que ofrezcan en licitaciones públicas reguladas el menor precio de suministro. Estos precios toman el nombre de precios de nudo de largo plazo (PNLP), contemplan fórmulas de indexación y son válidos para todo el período de vigencia del respectivo contrato, hasta un máximo de 15 años.

En términos más precisos, el precio de nudo de la energía de largo plazo para un determinado contrato corresponde al más bajo precio de energía ofrecido por las generadoras participantes del respectivo proceso de licitación, en tanto el precio de nudo de la potencia de largo plazo corresponde al precio de nudo de la potencia fijado en el decreto de precio de nudo vigente al momento de la licitación.

Sin embargo, puesto que los nuevos contratos de suministro asignados según esta modalidad comenzaron a regir gradualmente desde el 2010, los contratos que se encontraban vigentes al momento de aprobarse la LC2 deberán seguir considerando como tarifa, hasta el momento de su expiración, los precios de nudo fijados semestralmente por la autoridad, que ahora son denominados precios nudos de corto plazo (PNCP)

Los PNCP son determinados cada seis meses por la Comisión Nacional de Energía (CNE) sobre la base de una comparación entre los precios proyectados y el precio medio ofrecido por las generadoras a clientes libres y a distribuidoras a PNLP. En primera instancia, el precio de nudo de energía es fijado sobre la base de las proyecciones de los costos marginales esperados del sistema para los siguientes 48 meses, en el caso del SIC, y 24 meses, en el caso del SING; y el precio de nudo de potencia es determinado a partir del cálculo del precio básico de la potencia de punta. Sin embargo, en segunda instancia, para asegurar que los precios de nudo se mantengan en torno a valores de mercado, se aplica un mecanismo de banda de precios en el caso que los valores teóricos resultantes de esos cálculos de la autoridad, en términos monómicos (por concepto tanto de energía como de potencia), disten 5% o más de los precios medios de mercado. Dicha banda puede fluctuar entre 5% y 30%, dependiendo de la diferencia entre el precio de nudo teórico y el precio medio de suministro que enfrentan los clientes no sometidos a regulación de precios.

Cabe destacar que el precio que ve el usuario final corresponde al Precio Nudo Promedio (PNP) de energía y potencia, el cual representan el nuevo precio de nudo resultante del procedimiento que combina los 2 anteriores, y que debe traspasar cada empresa distribuidora a sus clientes regulados.

2.2.2 Transmisión

Corresponde al sector del mercado eléctrico encargado del transporte de la electricidad desde el punto de generación hacia los puntos de consumo o distribución. Su composición queda determinada por las líneas y subestaciones utilizadas para realizar aquella función que cumplan con voltaje o tensión superior a 23 kV.

Motivo de sus grandes economías de escala, la transmisión se considera un monopolio natural encargado de realizar un servicio público concesionado y de libre acceso. Dada su característica de servicio público monopólico, las instalaciones de los sistemas de transmisión están sometidas a un régimen de acceso abierto, debiendo los usuarios efectuar un pago que es regulado por el Estado.

Dentro de su mercado se distinguen 3 sectores, los sistemas de transmisión troncal, los sistemas de subtransmisión y sistemas adicionales.

- Sistema de transmisión troncal: Correspondiente al conjunto de líneas y subestaciones eléctricas principales, de tensiones nominales igual o superiores a 220 kV, en los cuales sus flujos según sus tramos pueden ser bidireccionales y no son atribuidos a clientes específicos, o a grupos de centrales, sino que a la totalidad de la demanda. (Art. 74 ley eléctrica [4])

- Sistemas de subtransmisión: Constituidos por las líneas y subestaciones eléctricas que, dispuestas para el abastecimiento exclusivo de grupos de consumidores finales libres o regulados, territorialmente identificables que se encuentren en zonas de concesión de empresas distribuidoras.

- Sistemas adicionales: Constituidos por las instalaciones de transmisión destinadas esencialmente al suministro de usuarios no sometidos a regulación de precios, y además por instalaciones cuyo objetivo principal es permitir a los generadores inyectar su producción al sistema eléctrico, sin que formen parte del sistema de transmisión troncal ni de los sistemas de subtransmisión.

En las tarifas a clientes finales se incorporan recargos por el uso de las redes troncales y de subtransmisión. El primero se traduce en un cargo único por el uso del sistema troncal y para subtransmisión como un peaje.

2.2.3 Distribución

El segmento de distribución se caracteriza por la compra de grande bloques de energía y potencia en el mercado mayorista, a través del esquema de licitaciones que fue explicado anteriormente y la posterior venta en el mercado minorista para llevar la electricidad a los consumidores finales. En Chile, se denomina distribución a todo sistema de transmisión (líneas de transmisión, subestaciones y equipos destinados a la distribución de electricidad) cuya tensión sea de 23 kV o menos, localizados en cierta zona geográfica explícitamente delimitada, a la que se denomina área de concesión.

Actualmente existen 26 empresas de distribución y 9 cooperativas, por lo cual hay 35 concesiones de distribución dispersas a lo largo y ancho del país [1].

Los clientes en distribución pueden ser clasificados en dos tipos: clientes libres y clientes regulados, según fue explicado en la descripción del segmento de generación. Los clientes libres pueden negociar sus contratos con las empresas generadoras o distribuidoras. Los clientes regulados poseen una serie de opciones tarifarias a las cuales pueden optar libremente aceptando las condiciones y limitaciones que éstas establezcan dentro del nivel de tensión que les correspondan. Las empresas distribuidoras están obligadas a aceptar la opción tarifaria escogida por el cliente.

El documento vigente que fija las fórmulas tarifarias aplicables a los suministros sujetos a precios regulados, para el cuadrienio noviembre de 2012 a noviembre de 2016, es el Decreto N° 1T de 2012 emitido por el Ministerio de Energía [1].

La tarifa que se le cobra al cliente final se compone de: Los precios a nivel de generación (precios de nudo), los precios a nivel de transmisión (cargo único por uso del sistema troncal y peajes de subtransmisión) y del valor agregado de distribución (VAD)

Parte importante del proceso de fijación de tarifa consiste en determinar el VAD, el cual es calculado cada 4 años por estudios presentados tanto por la CNE como por parte de las empresas concesionarias, las cuales son agrupadas en áreas típicas de distribución (ATD). Para cada área típica se calcula el VAD por 2 consultoras, una contratada por la empresa y otra por la CNE., El resultado es la ponderación de 2/3 el resultado de la CNE y 1/3 el resultado de las empresas concesionarias.

De acuerdo con lo señalado en el DFL 4/2006, el procedimiento de definición de las ATD debe ser tal que queden agrupadas las empresas o sectores de ellas, cuyos valores agregados por la actividad de distribución (Costos Medios) sean parecidos entre sí (Art. 225° letra m [4]). Esta clasificación obedece a un problema de índole práctico, que permite entregar señales de precio homogéneas a los clientes finales pero que a su vez no compromete los ingresos de las empresas concesionarias de distribución. Actualmente existen 6 ATD.

Dentro de las obligaciones de las empresas distribuidoras se encuentra la de proporcionar suministro a sus clientes a un precio regulado cumpliendo con las exigencias de calidad de servicio y suministro. Por otro lado, las empresas concesionarias tienen el derecho de hacer uso de la vía pública para construir, operar y mantener las instalaciones necesarias y a cobrar una tarifa regulada fijada por la autoridad a sus clientes regulados.

2.3 Marco Legal, Reglamentario y Normativo

El objetivo de esta sección es entregar una síntesis de los aspectos centrales del marco regulatorio vigente en Chile para el segmento de distribución de electricidad, contexto en que destacan por su importancia la LGSE y el Reglamento de la Ley.

El cuerpo legal que regula la actividad del sector eléctrico actualmente es el DFL Nº4 promulgado en el año 2006 por el Ministerio de Economía, Fomento y Reconstrucción, el que fija texto refundido, coordinado y sistematizado del DFL Nº1 de 1982, LGSE, en materia de energía eléctrica. El DFL Nº4 regula la producción, transporte, distribución, concesiones y tarifas de energía eléctrica y las funciones del Estado relacionadas con estas materias, esta ley y su reglamentación complementaria determinan las normas técnicas y de seguridad por las cuales debe regirse cualquier instalación eléctrica en el país. [4]

Existen diversos reglamentos que regulan los aspectos normados en la LGSE, siendo el más importante para efectos de esta memoria el Decreto Supremo 327 (DS 327), vigente desde 1997, derogando disposiciones contenidas en normativas dispersas y parciales. Esta reglamentación comprende los aspectos de concesiones, permisos y servidumbres, así como referentes a la interconexión de instalaciones. En materia de distribución, entrega a los concesionarios de este servicio público la responsabilidad del cumplimiento de los estándares y normas de calidad de servicio establecidos en la ley y los reglamentos, delegando a todo quien entregue suministro eléctrico el compromiso de cumplir con los estándares de calidad establecidos. [6]

Dentro de la normativa técnica que rige el segmento de distribución eléctrica, se encuentra la Norma Técnica sobre Conexión y Operación de Pequeños Medios de Generación Distribuidos en Instalaciones de Media Tensión (NTCO) [7] la cual establece los procedimientos, metodologías y demás exigencias para la conexión y operación de los Pequeños Medios de Generación Distribuida (PGMD) en redes de MT de empresas distribuidoras o empresas de distribución que utilicen bienes nacionales de uso público.

Hasta el momento no existe en Chile un marco ordenador o Código de Red que defina estándares de calidad de servicio para los sistemas de distribución, a diferencia de los sectores de generación y transmisión que si lo tienen y está establecido en la Norma Técnica de Seguridad y Calidad de Servicio (NTSyCS), la cual determina el conjunto de exigencias mínimas de seguridad y calidad de servicio asociadas al diseño de las instalaciones y a la coordinación de la operación de los sistemas eléctricos que operan interconectados.

2.4 Institucionalidad

Dentro del sector eléctrico existen una serie de organismos encargados de regular, fiscalizar, coordinar y entregar las normativas y leyes con las cuales deben desempeñarse los integrantes del mercado [8].

- Ministerio de Energía: El Ministerio de Energía es el órgano superior de colaboración del Presidente de la República en las funciones de gobierno y administración del sector de energía. Su objetivo general es elaborar y coordinar los planes, políticas y normas para el buen funcionamiento y desarrollo del sector, velar por su cumplimiento y asesorar al Gobierno en todas aquellas materias relacionadas con la energía.

Dentro de sus funciones se destacan:

a) Preparar, dentro del marco del plan nacional de desarrollo, los planes y políticas para el sector energía.

b) Estudiar y preparar las proyecciones de la demanda y oferta nacional de energía que deriven de la revisión periódica de los planes y políticas del sector.

c) Elaborar, coordinar, proponer y dictar según corresponda, las normas aplicables al sector energía que sea necesario dictar para el cumplimiento de los planes y políticas energéticas de carácter general.
d) Velar por el efectivo cumplimiento de las normas sectoriales, sin perjuicio de las atribuciones que correspondan a los organismos en ella mencionados, a los que deberá impartir instrucciones, pudiendo delegar las atribuciones y celebrar con ellos los convenios que sean necesarios.

f) Cumplir las demás funciones y tareas que las leyes o el Gobierno le encomienden concernientes a la buena marcha y desarrollo del sector energía.

- Comisión Nacional de Energía (CNE): La CNE es un organismo público y descentralizado, con patrimonio propio y plena capacidad para adquirir y ejercer derechos y obligaciones, que se relaciona con el Presidente de la República por intermedio del Ministerio de Energía.

El objetivo principal es analizar tarifas y normas técnicas a las que deben ceñirse las empresas de producción, generación, transporte y distribución de energía, con el objeto de disponer de un servicio suficiente, seguro y de calidad, compatible con la operación más económica.

La CNE es quien dictamina las Bases Técnicas de los estudios de los Valores Agregados de Distribución y realiza la tipificación de las áreas típicas, y responde a las observaciones efectuadas por las empresas. También es quien establece el listado de consultoras que pueden realizar los estudios y contratar a uno de ellos para encargárselo. Además, en caso que las empresas contraten un estudio, es quien recibe sus resultados y pondera sus valores con los de su estudio para la obtención final de las componentes de los Valores Agregados de Distribución. Finalmente es quien presenta una propuesta de fijación de tarifas al Ministerio de Energía.

- Superintendencia de Electricidad y Combustibles (SEC): La Superintendencia de Electricidad y Combustibles tiene por misión vigilar la adecuada operación de los servicios de electricidad, gas y combustibles, en términos de su seguridad, calidad y precio.

Para ello fiscaliza y supervigila el cumplimiento de las disposiciones legales, reglamentarias, y de normas técnicas sobre generación, producción, almacenamiento, transporte y distribución de combustibles líquidos, gas y electricidad, para verificar que la calidad de los servicios que se presten a los usuarios sea la señalada en dichas disposiciones, y que las operaciones y el uso de los recursos energéticos no constituyan peligro para las personas o sus cosas.

Con respecto a los estudios de componentes del Valor Agregado de Distribución, la SEC es quien fija los valores nuevos de reemplazo y los costos de explotación, a partir de los informes auditados que sobre estos ítems entregan las empresas.

- Centro de Despacho Económico de Carga: El Centro de Despacho Económico de Carga es un ente creado para la coordinación de la operación de las instalaciones eléctricas de los concesionarios que operen interconectados entre sí, con el fin de:

  - Preservar la seguridad del servicio en el sistema eléctrico.

  - Garantizar la operación más económica para el conjunto de las instalaciones del sistema eléctrico.

  - Garantizar el derecho de servidumbre sobre los sistemas de transmisión establecidos mediante concesión.

Los CDEC están integrados por todas aquellas empresas de generación, transmisión y consumidores de precio no regulado (clientes libres) que cumplen con los requisitos establecidos por la Ley y se diferencian entre ellos según el sistema al cual corresponden.

A lo largo de Chile existen 4 sistemas eléctricos aislados entre sí, correspondientes a:

- El Sistema Interconectado del Norte Grande (SING) se extiende entre Arica-Parinacota, Tarapacá y Antofagasta, Decimoquinta, Primera y Segunda regiones de Chile, respectivamente, cubriendo una superficie de 185.142 km^2, equivalente a 24,5% del territorio continental.

- Sistema Interconectado Central el cual comprende el área ubicada desde la rada de Paposo por el norte (en la II Región) y la localidad de Quellón por el sur, en la isla de Chiloé (X Región), cubriendo cerca del 93% de la población de la República de Chile.

- Sistema Aysén

- Sistema Magallanes

Solo los 2 primeros poseen CDEC. Los dos últimos son sistemas medianos y están verticalmente integrados.

- Panel de Expertos: Órgano colegiado autónomo creado en el año 2004 por la Ley Corta I [9], de competencia estricta y regulada. Su función es pronunciarse, mediante dictámenes de efecto vinculante, sobre aquellas discrepancias y conflictos que, conforme a la Ley, se susciten con motivo de la aplicación de la legislación eléctrica y que las empresas eléctricas sometan a su conocimiento.

Se considera adecuado destacar que la fijación de tarifas de distribución no es de las discrepancias que son sometidas a este panel.

2.5 Calidad de suministro

La LGSE en su artículo 225°, literal u) define la calidad de servicio como el atributo de un sistema eléctrico determinado conjuntamente por la calidad del producto, la calidad de suministro y la calidad de servicio comercial, entregado a sus distintos usuarios y clientes. Dentro de este mismo artículo define, en su literal v), la calidad del suministro como la componente de la calidad de servicio que permite calificar el suministro entregado por los distintos agentes del sistema eléctrico y que se caracteriza, entre otros, por la frecuencia, la profundidad y la duración de las interrupciones de suministro.

El Reglamento Eléctrico (DS 327) viene a profundizar las definiciones dadas en la Ley, definiendo calidad de servicio como el conjunto de propiedades y estándares normales que son inherentes a la actividad de distribución de electricidad concesionada. [6]

Dentro de los parámetros que engloba este concepto se encuentran los estándares de calidad de suministro. Respecto a la calidad de suministro, el artículo 223° del Reglamento la define como el conjunto de parámetros físicos y técnicos que debe cumplir el producto electricidad. Dichos parámetros son, entre otros, tensión, frecuencia y disponibilidad

En caso de existir una calidad de servicio reiteradamente deficiente en una empresa, la SEC está facultada para amonestar, multar o adoptar otras medidas pertinentes.

2.5.1 Parámetros de continuidad de suministro

La continuidad de suministro hace referencia a la existencia o no de tensión en el punto de conexión. Cuando la continuidad de suministro falla, se habla de una interrupción de suministro. La norma europea UNE-EN 50160 [10] define una interrupción de suministro cuando la tensión está por debajo del 1% de la tensión nominal en cualquiera de sus fases de alimentación. Se ha tomado como referencia esta norma, ya que como se señaló anteriormente, en Chile no existe una norma técnica que establezca los estándares de calidad de suministro a nivel de distribución.

En una red de media tensión, el DS 327 [6] indica que el valor estadístico de la tensión deberá estar dentro del rango +6,0% a –6,0% durante el 95% del tiempo de cualquiera semana del año o de siete días consecutivos de medición y registro. Este período excluye los momentos en que ocurre una interrupción de suministro.

Las interrupciones de servicio se clasifican según su duración. Para el problema de continuidad de suministro, se considerarán aquellas mayores a tres minutos [6] [10], las cuales son llamadas interrupciones largas.

Las interrupciones menores a tres minutos, llamadas interrupciones breves, se consideran un problema de calidad de onda, ya que se deben a un problema en la operación de los sistemas de protección de las redes: reconexiones rápidas debido a fallas y operación de zonas aisladas, entre otras. Las interrupciones largas, en cambio, se entiende que requieren de la reparación de algún material defectuoso de la red, o de la inspección de los tramos con problemas, así como la reposición manual de la tensión.

En la literatura se clasifican las interrupciones según su duración y origen [11]. Esta clasificación tiene implicancias regulatorias e informativas. En el caso de su origen, se hacen dos distinciones para una interrupción larga: interrupciones programadas e interrupciones imprevistas o fallas.

2.5.2 Índices de confiabilidad de la red usados en el mundo

En el contexto internacional existen los índices individuales de cliente (costumer) y los índices de sistema (system).

Algunos de los índices de confiabilidad más usados para fallas sostenidas en el tiempo (mayor a 5 minutos) recomendados por el subcomité de distribución y transmisión del IEEE Power Engineering Society se describen a continuación [12]:

- Saifi: (System average interruption frequency index) Corresponde a la frecuencia promedio de interrupciones por cliente en un área predefinida.

$$\text{SAIFI} = \frac{\text{Número total de clientes interrupidos}}{\text{Número total de clientes}}$$

  Para el cálculo, se utiliza la siguiente fórmula:

$$\text{SAIFI} = \frac{\sum \text{N}_\text{i}}{\text{N}_\text{T}}$$

  Donde:
  - N_i : Número de clientes interrumpidos para cada interrupción i durante el período analizado dentro del área analizada.
  - N_T : Número total de clientes en el área que está siendo analizada.

- Saidi: (System average interruption duration index) Corresponde a la cantidad de tiempo promedio de interrupciones por cliente en un área predefinida.

$$\text{SAIDI} = \frac{\sum \text{Duración de la interrupción a clientes}}{\text{Número total de clientes}}$$

Para el cálculo, se utiliza la siguiente fórmula:

$$\text{SAIFI} = \frac{\sum r_i N_i}{N_T}$$

Donde:

- r_i : Tiempo necesario para reponer el servicio dada una interrupción i.

- Asai: (Average service availability index) Este índice representa la fracción de tiempo (o porcentaje) que un cliente ha sido provisto de energía eléctrica durante un año (o periodo definido)

$$\text{ASAI} = \frac{\text{Horas de servicio disponibles para el cliente}}{\text{Horas de demanda de servicio del cliente}}$$

Para el cálculo, se utiliza la siguiente fórmula:

$$\text{ASAI} = \frac{N_T \cdot 8760 - \sum r_i N_i}{N_T \cdot 8760} = 1 - \frac{\sum r_i N_i}{N_T \cdot 8760}$$

Estos índices estan basados en los clientes.También existen otros índices basados en la carga como el ASIDI (Average system interruption duration) que indica el tiempo en horas de interrupción de la potencia total conectada en el sistema y el ASIFI (Average system interruption frecuency) que representa la cantidad de potencia interrumpida, con respecto a la potencia total servida.

A modo de ejemplo, podemos decir que en America Latina, Perú adoptó los índices Saidi y Saifi para evaluar su calidad de suministro. [13]

2.5.3 Índices de confiabilidad de la red usados en Chile

Los índices de confiabilidad intentan medir la continuidad del suministro, es decir, el número de veces que se ve interrumpido un cliente o un grupo de clientes y por cuánto tiempo.

El Reglamento Eléctrico exige, en su artículo 227, que la calidad de suministro deberá ser evaluada separadamente en los sistemas de generación, transporte, distribución y en los propios del consumidor final. Las mediciones deben ser efectuadas bajo las dos siguientes modalidades:

• En un punto específico de la red, para determinar el nivel de calidad del suministro entregado. (Calidad Individual)

• En un conjunto de puntos de la red o de usuarios, seleccionados de acuerdo a procedimientos estadísticos y metodología determinada por la Superintendencia. (Calidad Global) [6]

El artículo 246, define cuales son los índices que se desea medir para determinar la calidad global de suministro. El Reglamento no da las fórmulas para el cálculo de los índices, lo deja propuesto para la norma técnica de calidad de servicio, la cual se está elaborando pero aún no ha entrado en vigencia, por lo que típicamente la CNE ha definido estos valores en las bases del estudio del VAD [14]. Además señala que estos valores dependerán de cada ATD pero fija valores máximos.

A continuación se presentan los índices considerados respecto al parámetro de interrupciones de suministro en instalaciones de servicio público de distribución y se presentan sus fórmulas de cálculo [14] [15]:

a) Frecuencia media de interrupción por transformador, FMIT: Corresponde a la frecuencia media de transformadores afectados cuando ocurre una falla, superior a tres minutos, en un alimentador. Se rige según la siguiente expresión:

$$FMIT = \frac{\sum_{i}^{n} Q_{fsi}}{Q_{inst}}$$

Donde:

- FMIT, Frecuencia Media de Interrupción por Transformador
- $Q_{fsi,}$ cantidad de transformadores fuera de servicio a causa de la interrupción i.
- $Q_{inst,}$ cantidad de transformadores instalados.
- n: número de interrupciones en el período de control

b) Frecuencia media de interrupción por kVA, FMIK: Corresponde a la frecuencia media de potencia aparente no suministrada en la red, cuando ocurre una falla mayor a tres minutos en un alimentador. Se calcula utilizando la siguiente fórmula:

$$FMIK = \frac{\sum_{i}^{n} kVA_{fsi}}{kVA_{inst}}$$

Donde:

- FMIK, Frecuencia Media de interrupción por kVA
- $kVA_{fsi,}$ Potencia Nominal fuera de servicio a causa de la interrupción i
- $kVA_{inst,}$ Potencia Nominal Total Instalada.
- n: número de interrupciones en el período de control

c) Tiempo total de interrupción por transformador, TTIT: Es el tiempo total de interrupción debido a fallas superiores a tres minutos, indexando la ponderación de transformadores no suministrados y el número total de transformadores de la red. Se calcula utilizando la siguiente fórmula:

$$TTIT = \frac{\sum_{i}^{n} Q_{fsi} \cdot T_{fsi}}{Q_{inst}} [h]$$

Donde:

- TTIT, Tiempo total de interrupción por transformador
- $Q_{fsi,}$ Cantidad de transformadores fuera de servicio a causa de la interrupción i.
- $T_{fsi,}$ Tiempo fuera de servicio del alimentador primario a causa de la interrupción i.
- n: número de interrupciones en el período de control

d) Tiempo total de interrupción por kVA, TTIK: Tiempo total de interrupción, debido a fallas mayores a tres minutos, indexando además la ponderación de potencia no suministrada y la potencia total de la red. Se calcula utilizando la siguiente fórmula:

$$TTIK = \frac{\sum_{i}^{n} kVA_{fsi} \cdot T_{fsi}}{kVA_{inst}} [h]$$

Donde:

- TTIK, Tiempo total de interrupción por kVA instalado
- T_{fsi}: Tiempo fuera de servicio del alimentador primario a causa de la interrupción i.
- $kVA_{inst,}$ Potencia Nominal Total Instalada.
- n: número de interrupciones en el período de control

Según el Reglamento los valores exigidos de estos índices dependerán del área típica de distribución de que se trate y serán definidos por la CNE con ocasión del cálculo de VAD y los fijará en las bases del estudio de cada ATD. De todos modos, define valores máximos que no pueden ser sobrepasados, considerando sólo las interrupciones internas de la red, los cuales son:

- FMIT entre 5 y 7 veces al año;
- FMIK entre 3,5 y 5 veces al año;
- TTIT entre 22 y 28 horas al año;
- TTIK entre 13 y 18 horas al año.

El artículo 247 del Reglamento establece que se pueden aplicar holguras para zonas definidas como rurales. Estas zonas están definidas en la resolución ministerial exenta N° 53 del Ministerio de Economía, Fomento y Reconstrucción [16] y sus posteriores modificaciones a través de la Resolución Ministerial Exenta N°30 de 2008 y Resolución Ministerial Exenta N°36 de 2009 del mismo ministerio. Además se fijan sus límites máximos.

Según la resolución, las zonas rurales tipo 1 quedan definidas como aquellas que cumplen simultáneamente las dos siguientes condiciones:

Condición 1:

- Población total inferior a 70.000 habitantes;
- Población total mayor a 70.000 habitantes y relación entre viviendas urbanas y superficie total de la comuna inferior a 350 viviendas/km2 (N°Viv.Urb/km2 < 350)

Condición 2:

- Número de clientes de la empresa dentro de la comuna inferior a 10.000;
- Número de clientes de la empresa dentro de la comuna mayor a 10.000 y una relación entre la potencia total vendida y los kilómetros de línea de media tensión inferior a 15 kW/km (kW/kmMT < 15)

Las zonas rurales tipo 2, quedan definidas como aquellas zonas que cumplen con las condiciones establecidas para ser clasificadas como zonas rural tipo 1 y, adicionalmente, en forma simultánea, cumplen las siguientes condiciones:

Condición 1: Ser suministradas por un alimentador cuya longitud total conectado a través de líneas de media tensión sea superior a 75 km, límite mínimo que no será aplicable a los territorios insulares;

Condición 2: Ser suministradas por un alimentador cuya relación entre la suma de las potencias de las subestaciones de distribución (transformación MT/BT), conectadas a dicho alimentador mediante líneas de media tensión y medida en kVA, respecto de la suma de las longitudes de esas mismas líneas de media tensión expresada en kilómetros, sea inferior a 50 kVA/km.

Se resumen los valores máximos que no pueden ser sobrepasados en las distintas zonas en la siguiente tabla:

	FMIT [veces al año]	FMIK [veces al año]	TTIT [horas al año]	TTIK [horas al año]
Zona urbana	5	3,5	22	13
Zona rural tipo 1	7	5	28	18
Zona rural tipo 2	11	8	42	27

Tabla 2.1: Límites máximos para índices de interrupción de suministro.

Lo que se busca con este trabajo es encontrar disponer de una metodología de asignación de grupos de consumidores que permita aplicar estos índices de calidad de suministro, reconociendo la densidad de carga con independencia de la empresa suministradora, la topología de las redes existentes o las distinciones demográficas expuestas anteriormente.

Capítulo 3 - Minería de Datos

En este capítulo se describe el proceso de la minería de datos, ahondando en el estado del arte de las técnicas de clustering y la descripción del software utilizado.

3.1. Introducción

La minería de datos forma parte integral del descubrimiento de conocimientos en bases de datos (KDD por sus siglas en inglés, knowledge discovery in database). Tal como lo dice su nombre, el proceso de KDD se refiere al proceso global de conversión de los datos en bruto, dentro de las bases de datos, en información útil [17].

Este proceso consiste en una serie de transformaciones desde el pre-procesamiento de los datos hasta el post-procesamiento de los datos obtenidos con técnicas de minería de datos. La metodología estándar de este tipo de procesos es la que se muestra en la Figura 3.1

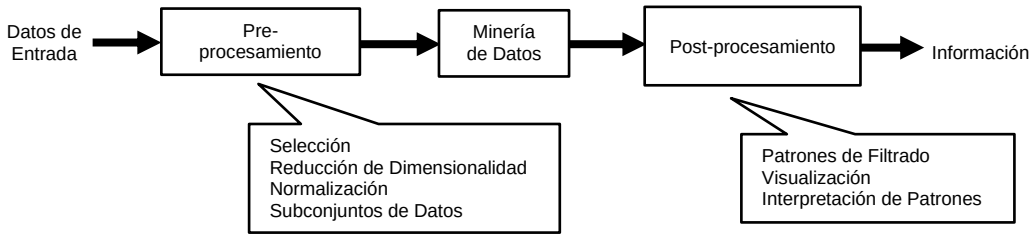


Figura 3.1: El proceso de descubrimiento de conocimientos en bases de datos (KDD).

3.2. Orígenes y Tareas de la Minería de datos

La minería de datos nace tratando de resolver desafíos que no habían logrado ser resueltos con las herramientas tradicionales conocidas a la fecha. Algunos de estos desafíos son la escalabilidad de los algoritmos cuando se procesaban grandes volúmenes de datos, la alta dimensionalidad de los mismos (gran cantidad de atributos), los datos de carácter heterogéneos y/o complejos entre otros.

Con el fin de poder resolver estos desafíos, los investigadores de distintas disciplinas desarrollaron herramientas cada vez más eficientes y escalables que podían manejar distintos tipos de datos. Esto culminó en el campo de la minería de datos, construida sobre las metodologías y los algoritmos que habían utilizado anteriormente. En particular, la minería de datos se basa en ideas tales como muestreo, estimación, test de hipótesis de la estadística y algoritmos de búsqueda, técnicas de modelamiento de la inteligencia artificial, reconocimiento de patrones y aprendizaje de máquinas. La minería de datos rápidamente adquirió ideas desde otras áreas, incluyendo optimización, computación evolutiva, teoría de la información, procesamiento de señales, visualización y recuperación de información. Además toma soporte en los sistemas de almacenamientos de bases de datos, la computación paralela y distribuida (ver Figura 3.2)

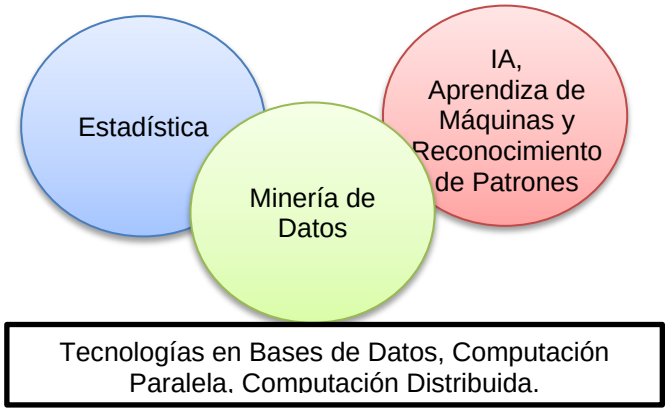


Figura 3.2: La minería de datos como una confluencia de varias disciplinas.

Las tareas en minería de datos se suelen dividir en dos grandes categorías:

- Tareas predictivas: El objetivo de este tipo de tareas es predecir el valor de un atributo en particular basado en los valores de los otros atributos.

- Tareas descriptivas: Aquí el objetivo es obtener patrones (correlaciones, tendencias, agrupamientos o clusters, trayectorias y anomalías) que resuman las relaciones subyacentes entre los datos.

En este trabajo en particular se utilizará la minería de datos de manera descriptiva, en particular se utilizará el uso de agrupamientos (clustering) que se analiza a continuación.

3.3. Clustering

El análisis de clusters agrupa objetos basados solamente en la información encontrada en los datos que describen a los objetos y sus relaciones [18]. La meta es que los objetos dentro de un grupo sean similares (o relacionados) entre sí y diferentes (o no relacionados con) los objetos en otros grupos. A mayor similitud (u homogeneidad) dentro de un grupo y a mayor diferencia entre grupos, mejor o más distinto es el clustering. (Figura 3.3)

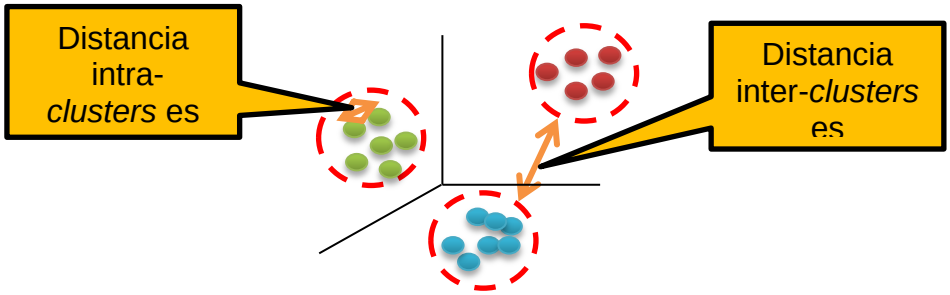


Figura 3.3: Análisis de *Clusters*

3.3.1 Métodos de Clustering.

Entenderemos por clustering una colección completa de clusters, algunos de los distintos tipos de clustering que se pueden distinguir son: jerárquico (anidado) versus particional (no anidado), exclusivo versus traslapado versus difuso y completo versus parcial.

El clustering particional divide los objetos en subconjuntos (cluster) sin traslape, es decir cada objeto está exactamente en un solo grupo. El cluster jerárquico es un conjunto de clusters anidados organizados como un árbol jerárquico.

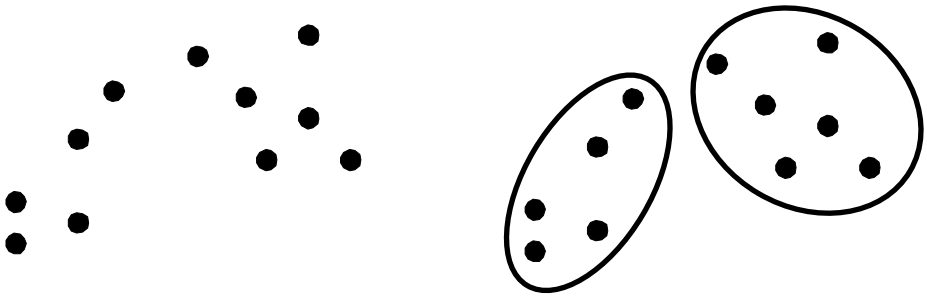
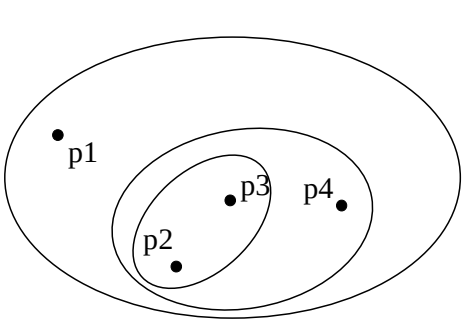
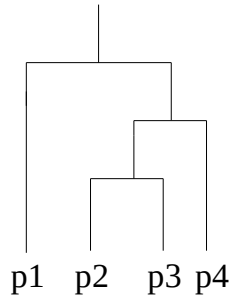


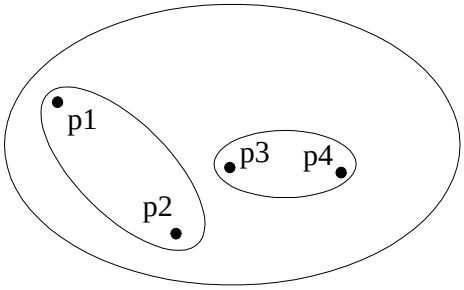
Figura 3.4: *Clustering* Particional



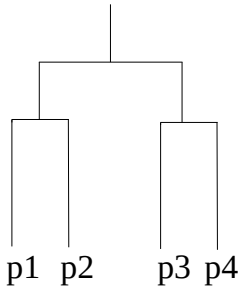
***Clustering* Jerárquico Tradicional**



Dendograma Tradicional



***Clustering* Jerárquico No Tradicional**



Dendograma No Tradicional

Figura 3.5: *Clustering* Jerárquico

El clustering exclusivo es aquel en donde cada objeto puede pertenecer sólo a un cluster (al igual que el clustering particional). En el clustering traslapado (o no exclusivo) los elementos pueden pertenecer a más de un cluster. En el clustering difuso cada objeto pertenece a todos los clusters con algún grado de pertenencia el que va entre 0 (absolutamente no perteneciente) a 1 (absolutamente perteneciente). Similar a este existe un clustering probabilístico donde cada elemento posee una probabilidad de pertenecer a cada uno de los cluster y esas probabilidades deben sumar 1, este tipo de clustering se utiliza para evitar la arbitrariedad de asignar un objeto únicamente a un cluster cuando puede estar cercano a varios. En la práctica este tipo de clustering se puede convertir en un clustering exclusivo asignando a cada objeto al cluster al cual tiene un mayor grado de pertenencia (o probabilidad de pertenencia).

El clustering completo es aquel en que cada objeto debe ser asignado a un cluster, mientras que en el parcial esto no tiene por qué ser así. La motivación para el clustering parcial es que algunos objetos en los set de datos pueden no pertenecer a ningún grupo bien definido. Algunos objetos dentro de los set de datos representan ruido u outliers.

3.3.2 Técnicas de Clustering

En esta sección se comentaran los tres enfoques principales para realizar agrupaciones. La mayoría de las técnicas empleadas se derivan o se basan en estas. Se verán con mayor detalle las técnicas K-means y DBSCAN por ser las empleadas en este trabajo.

3.3.1.1. K-means

Esta técnica está basada en el clustering particional que intenta encontrar un número de clusters (K) especificados por el usuario, los cuales son representados por sus centroides.

El algoritmo básico se describe a continuación: Primero se eligen K centroides iniciales, donde K es un parámetro especificado por el usuario y corresponde al número de clusters deseados. Cada punto es asignado a su centroide más cercano y cada colección de puntos asignado a un centroide representa un cluster. El centroide de cada cluster se actualiza basado en la asignación de puntos al cluster. Se repiten los pasos de asignación y actualización hasta que los puntos dentro del cluster no cambien, o equivalentemente, hasta que los centroides dejen de cambiar.

Formalmente el algoritmo se describe como sigue:

Algoritmo Básico de K-means
1: Seleccionar K puntos iniciales como centroides 2: **Repetir** 3: Formar K cluster asignando cada punto a su centroide más cercano 4: Recalcular el centroide de cada cluster 5: **Hasta** que el centroide no cambie

Tabla 3.1: Algoritmo básico de K-means

La operación del algoritmo es mostrada en la Figura 3.6, se observa cómo partiendo desde tres centroides (marcados con una +), los clusters finales se encuentran en 6 pasos de asignación-actualización. Cada gráfico muestra en 3 colores la asignación de puntos a los clusters para cada iteración.

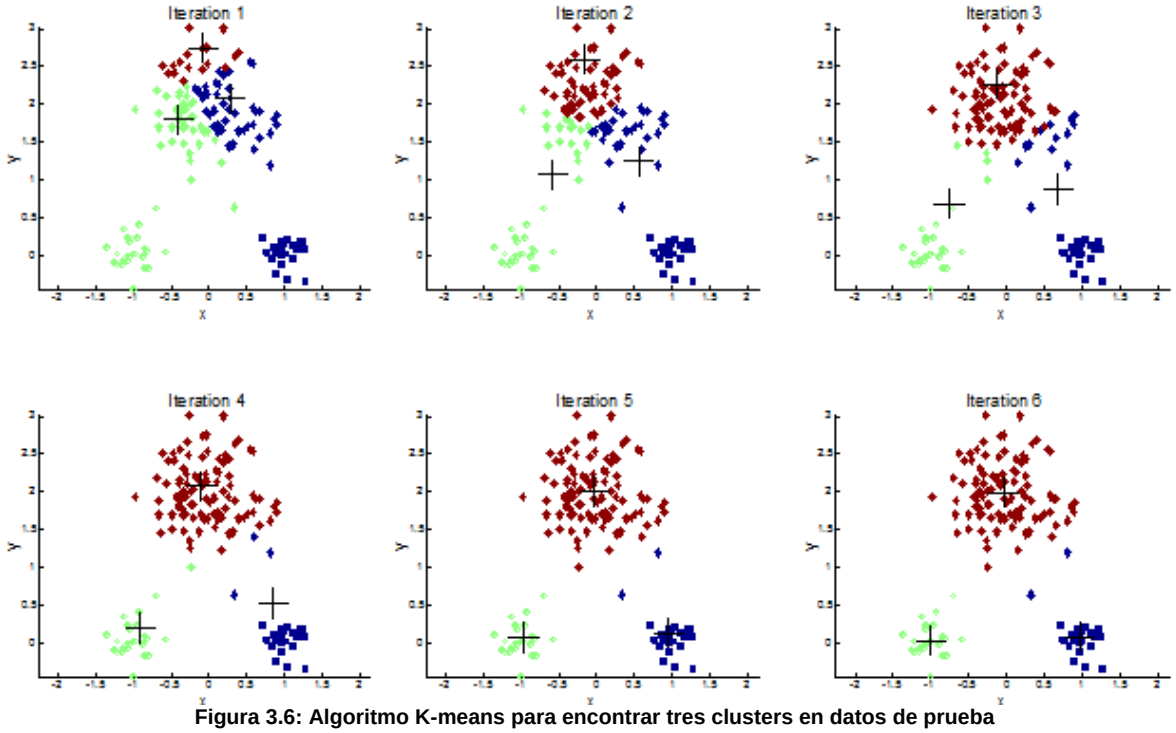


Figura 3.6: Algoritmo K-means para encontrar tres clusters en datos de prueba

En el primer paso, los puntos son asignados a sus centroides iniciales los cuales se encuentran todos sobre el grupo más grande de datos. Luego de que los puntos son asignados a su centroide, el centroide es actualizado. En el segundo paso los puntos son asignados a los centroides actualizados y los centroides son actualizados nuevamente. Se aprecia como a medida que se va iterando los centroides se van desplazando a los grupos más pequeños hacia la parte de debajo de las figuras. El algoritmo K-means termina en la iteración 6, porque ya no están ocurriendo más cambios, los centroides han identificado la agrupación natural de los puntos.

Para asignar cada punto a su centroide más cercano, es necesaria una medida de distancia que cuantifique la noción de “cercano” para los datos específicos que se estén considerando. La distancia Euclidiana (L_2) es usada comúnmente cuando los datos se encuentran en un espacio Euclidiano, mientras que la distancia coseno es más apropiada para documentos. Sin embargo, existen distintos tipos de medidas de similitud o distancia que son apropiadas para un tipo de datos dado. Por ejemplo la distancia Manhattan (L_1) puede ser usada en datos Euclidianos, así como también se emplea a veces la medida de Jaccard en documentos (mayor información sobre estas medidas se puede encontrar en [18])

Fortalezas y Debilidades: K-means es simple y puede ser usado para una amplia variedad de tipos de datos. También es bastante eficiente, pese a que a menudo se debe ejecutar varias veces. Algunas variantes, como el bisecting K-means son aún más eficientes y menos susceptibles a problemas de inicialización. Sin embargo, K-means no es apropiado para todo tipo de datos, no puede manejar clusters que no sean esféricos o que sean de distintos tamaños y densidades, aunque típicamente puede encontrar subclusters puros si se especifica un número suficientemente grande de clusters. K-means también tiene problemas agrupando datos que contengan outliers. La detección y remoción de estos datos puede ayudar significativamente en estas situaciones. Finalmente, K-means se limita a los datos para los cuales hay una noción de centro (centroide). Una técnica relacionada K-mediod clustering no tiene esta restricción, pero es computacionalmente más pesada.

3.3.1.2. Clustering Jerárquico Aglomerativo:

Este enfoque de clustering se refiere a una colección de técnicas de agrupamiento estrechamente relacionadas que producen un agrupamiento jerárquico comenzando con cada punto como un cluster singleton (con un solo elemento) e iterativamente lo agrupa con los dos clusters más cercanos hasta que un único cluster que abarca a los todos los demás permanece.

Fortalezas y Debilidades: Este tipo de algoritmos se utilizan normalmente porque la aplicación subyacente, por ejemplo, la creación de una taxonomía, requiere una jerarquía. Además, hay algunos estudios que sugieren que estos algoritmos pueden producir mejor calidad de clusters. Sin embargo, los algoritmos de clustering jerárquicos son costosos en términos de sus requerimientos computacionales y de almacenamiento. El hecho de que todas las clusters terminen finalmente unidos también puede causar problemas para datos ruidosos o de alta dimensionalidad. Estos dos problemas se pueden resolver en cierta medida agrupando los primeros datos parcialmente utilizando otra técnica, como K-means.

3.3.1.3. DBSCAN:

Este es un algoritmo de clustering basado en densidad que produce un clustering particional, en el cual el número de cluster es determinado automáticamente por el algoritmo. Puntos con baja densidad son clasificados como ruido y son omitidos, por lo que DBSCAN no produce un clustering completo.

DBSCAN opera en un enfoque de densidad en base a centros, el cual se describirá a continuación: En el enfoque basado en centros, la densidad es estimada para un punto en particular contando el número de puntos que se encuentra al interior de un radio especificado ε, épsilon o Eps centro en este punto.

El enfoque basado en centros permite clasificar los puntos como aquellos que se encuentran al interior de una región densa (puntos de núcleo), aquellos en el borde de una región densa (puntos de borde) y aquellos que se encuentran dispersos (puntos de ruido o de fondo).

Una definición más precisa se dará a continuación:

- Puntos de Núcleo: Son aquellos puntos que tienen más de MinPts vecinos dentro de su vecindario de Radio Eps.

- Puntos de Borde: Son los puntos que tienen menos de MinPts vecinos dentro de su vecindario de radio Eps, pero están en la vecindad de un punto de núcleo.

- Puntos de Ruido: Son aquellos puntos que no caen en ninguna de las dos categorías anteriores.

La Figura 3.7 representa un ejemplo de puntos de núcleo, de borde y de ruido considerando un Eps de valor 1 y un MinPts de valor 4.

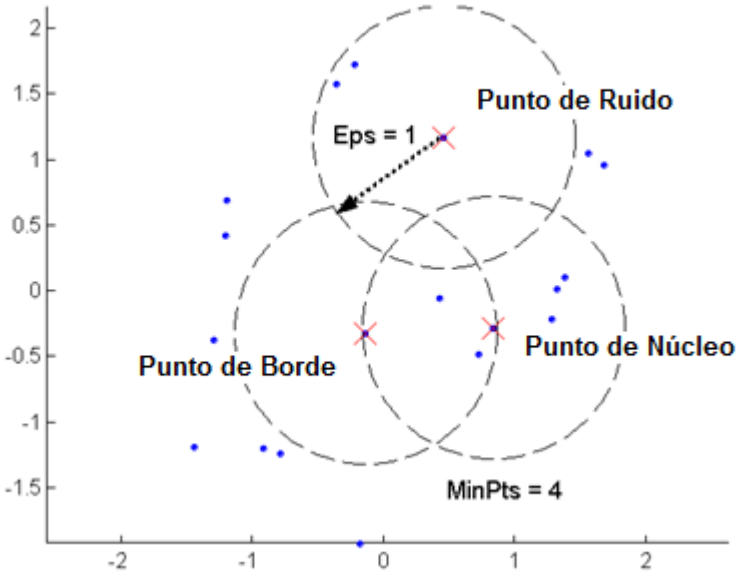


Figura 3.7: Puntos de núcleo, de borde y de ruido

Dada las definiciones anteriores, el algoritmo DBSCAN se describe como sigue: Cualquier par de puntos de núcleo que estén lo suficientemente cerca (menor a una distancia Eps) son asignados a un mismo cluster. Así mismo, cualquier punto de borde que este lo suficientemente cerca de un punto de núcleo es puesto en el mismo cluster que el punto de núcleo. (Podría ser necesario resolver “empates” cuando un punto de borde se encuentra a la misma distancia de dos puntos de núcleo de distintos clusters). Los puntos de ruido se descartan.

Una descripción formal del algoritmo se encuentra en la Tabla 3.2

Algoritmo DBSCAN
1: Etiquetar todos los puntos como de núcleo, de borde o de ruido
2: Eliminar puntos de ruido
3: Poner un borde entre todos los puntos de ruido que estén dentro de una distancia Eps de cada cluster
4: Hacer de cada grupo de puntos de borde que estén conectados un cluster separado
5: Asignar cada punto de borde a uno de los clusters asociados con sus puntos de núcleo.

Tabla 3.2: Algoritmo DBSCAN

Este método es simple de implementar, pero la densidad dependerá del radio épsilon especificado. Por ejemplo, si el radio es lo suficientemente grande, entonces todos los puntos tendrán una densidad igual al número total de datos. Así mismo, si el radio es demasiado pequeño, todos los puntos tendrán una densidad 1.

Por lo tanto, una manera para determinar los parámetros Eps y MinPts es realizar el cálculo de los k vecinos más cercanos (k-nn por k nearest neigthboor en inglés).

Para los puntos al interior de un cluster, su k-nn se encuentra aproximadamente a la misma distancia y los puntos de ruido tendrán su k-nn a mayor distancia.

Se grafica entonces, para un k = MinPts fijo, las distancia de los puntos a sus k-nn ordenadas de manera ascendente y se obtendrá un gráfico como el de la Figura 3.8 donde se observa un codo a partir del cual las distancias a los k-nn se empiezan a hacer mucho mayores, por lo tanto estos puntos corresponden a ruido y se debe considerar un Eps cerca del codo.

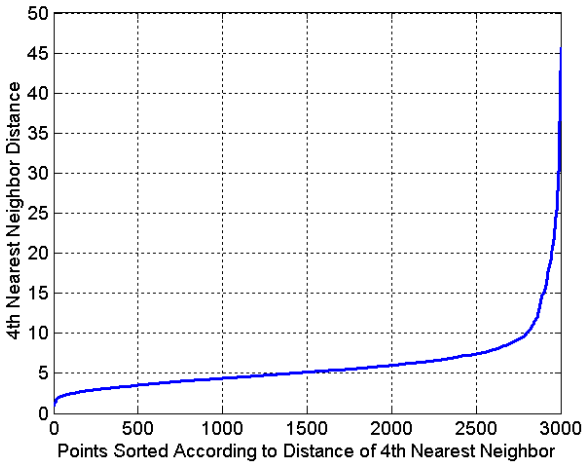


Figura 3.8: Distancia a los k-nn ordenados ascendentemente.

Fortalezas y Debilidades: Dado que DBSCAN usa una definición de cluster basada en densidad, es relativamente resistente al ruido y puede manejar cluster de forma y tamaños arbitrarios. Así, DBSCAN puede encontrar muchos clusters que no pueden ser encontrados utilizando K-means. Sin embargo, DBSCAN tiene problemas cuando los clusters tienen una amplia variedad de densidades. También tiene problemas con la alta dimensionalidad porque la densidad es más difícil de definir para este tipo de datos. Finalmente, DBSCAN puede ser computacionalmente costoso por ejemplo en el cálculo de los “vecinos más cercanos” para datos de alta dimensión (muchos atributos).

3.4. Validación de Clusters

Por su propia naturaleza, la evaluación de clusters no es una parte muy desarrollada o utilizada comúnmente de análisis de clustering. Sin embargo, la evaluación o validación de clusters como se le llama tradicionalmente, es importante, pues permite: Encontrar patrones de ruido, comparar algoritmos de clustering diferentes, comparar conjuntos de clusters diferentes, comparar dos clusters. La validación busca establecer una tendencia de agrupamiento (es decir si existen estructuras de agrupamiento no aleatorias), establecer el número correcto de clusters, evaluar si los resultados se ajustan a los datos, comparar los resultados del clustering contra resultados externos, comparar dos conjuntos de clustering para ver cuál es mejor. En esta sección se describen brevemente uno de los enfoques más comunes y de fácil aplicación.

- Medidas Internas, SSE: Los clusters en figuras más complejas no están tan bien separados, se utiliza un índice interno llamado SSE (sum of squares errors) para medir que tan buena es una estructura de clustering sin utilizar información externa. Se utiliza para comparar dos soluciones de clustering distintas y también puede ser utilizado para estimar el número de clusters. (Figura 3.9)

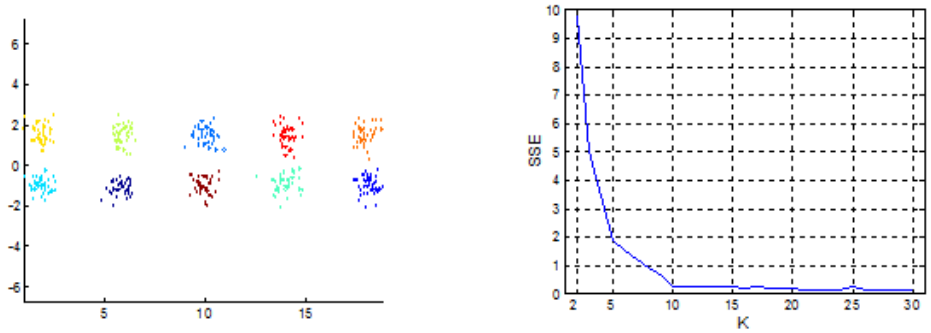


Figura 3.9: Ejemplo de utilización del SSE como indicar para determinar el número de clusters (K)

3.5. Software de Minería de Datos

En esta sección se dará una breve reseña de los tres software gratuitos de minería de datos que fueron utilizados en este trabajo.

3.5.1. CLUTO

CLUTO es un paquete de software para realizar clustering en datos de baja y alta dimensionalidad y analizar las características de los cluster. Fue desarrollado por George Karypis, profesor del Departamento de Ciencias de la Computación de la Universidad de Minnesota

CLUTO es propiedad intelectual de los Regentes de la Universidad de Minnesota. Puede ser utilizado libremente para propósitos educativos y de investigación de las instituciones sin fines de lucro y agencias del gobierno de Estados Unidos solamente.

CLUTO provee tres clases diferentes de algoritmos de clustering están basados en paradigmas particional, aglomerativo y basado en grafos [19].

La versión utilizada de CLUTO corresponde a la 2.1.1

3.5.2. RapidMiner

RapidMiner es un programa informático para el análisis y minería de datos. Permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico.

La versión inicial fue desarrollada por el departamento de inteligencia artificial de la Universidad de Dortmund en 2001. Se distribuye bajo licencia AGPL y está hospedado en SourceForge desde el 2004.

RapidMiner proporciona más de 500 operadores orientados al análisis de datos, incluyendo los necesarios para realizar operaciones de entrada y salida, pre-procesamiento de datos y visualización.

La versión utilizada de RapidMiner corresponde a la 5.3.013 distribuida por Rapid-I.

3.5.3. Weka

Weka (Waikato Environment for Knowledge Analysis - Entorno para Análisis del Conocimiento de la Universidad de Waikato) es una plataforma de software para aprendizaje automático y minería de datos escrito en Java y desarrollado en la Universidad de Waikato. Weka es un software libre distribuido bajo licencia GNU-GPL.

La versión empelada de Weka corresponde a la 3.6.8.

Capítulo 4 – Metodología

En este capítulo se dan a conocer los pasos que se llevaron a cabo para lograr el objetivo de la obtención de una metodología para clasificar consumos en distintas zonas de exigencia de calidad de suministro basado en criterios geográficos y de consumo de energía eléctrica.

4.1. Pre-procesamiento de los datos

Para lograr hacer el agrupamiento de los clientes se necesitó contar con datos reales de sus consumos.

Se tuvo acceso reservado a datos de la energía demandada por los consumos de la octava y novena región (correspondientes a un total de 86 comunas) durante el año 2012. Estos datos fueron recibidos como dos archivos de texto simple (formato TXT).

En el archivo CLIENTES.txt se identifican los siguientes datos para el cliente:

- Región
- Código de identificación (ID) de empresa,
- ID de punto de consumo
- ID de comuna
- coordenada X
- coordenada Y
- Datos de facturación.

En archivo TMP_PCONSUMO_ENERGIA.txt, se identifica, para cada mes (periodo):

- ID de empresa
- ID de punto de consumo
- Energía consumida
- Energía adicional de invierno consumida

Esta información se montó en una base de datos Access que permite manipular los datos, en particular obtener el consumo anual para cada cliente, de un total de 952.335 clientes.

A través de esta base de datos fue posible seleccionar subconjuntos de los datos o de sus atributos.

Se tuvo especial cuidado con los puntos que son demasiado disímiles del resto (outliers) porque pueden afectar el desempeño de algunos de los algoritmos. Se descartaron del análisis aquellos puntos que no ayudan para encontrar las zonas de exigencia de calidad de suministro que se espera encontrar.

4.2. Evaluación del desempeño del software de minería de datos y algoritmos de clustering.

Todas las pruebas y experimentos se realizaron en un computador personal Samsung RV409 con un procesador Intel® Pentium Dual Core P6200 (2.13 GHz, 3 MB) y memoria RAM DDR3 de 4 GB a 1,066 MHz con sistema operativo Microsoft Windows 7 de 64 bits.

Existen variados programas computacionales gratuitos que implementan los principales métodos de clustering. Se utilizaron tres: CLUTO, RapidMiner y Weka.

En la siguiente tabla se resumen los algoritmos presentes en cada uno de los programas, los cuales corresponden a los explicados en el capítulo 3 o a variaciones de ellos.

	CLUTO	RapidMiner	Weka
Bisecting K-Means	√		
Rbr	√		
Bagglo	√		
K-Means	√	√	√
K-Mediods		√	
K-Means kernel		√	
X-Means		√	√
DBSCAN		√	√
Expectation Maximization		√	√
Aglomerativo	√	√	√
Top-Down Clustering		√	
COBWEB			√
Farthest First			√
MakeDensityBased			√
sIB			√

Tabla 4.1: Métodos de clustering presentes en los programas computacionales a utilizar

Se evaluó el desempeño de los algoritmos presentes en cada uno de los programas computacionales utilizando un set de prueba. Para la evaluación se consideró tanto el tiempo empleado para generar la solución de clustering como la calidad de la solución, en base a criterios de resultados que se esperaría obtener.

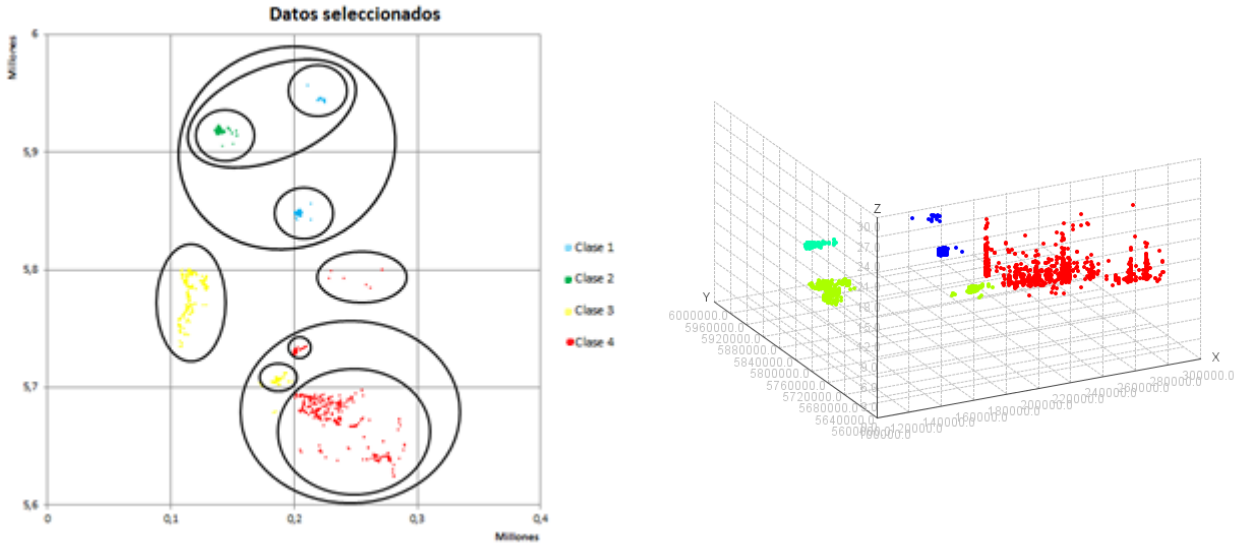
Cada solución de clustering se calculó utilizando 3 atributos de los datos: su coordenada X, su coordenada Y y su consumo anual de energía en kWh.

De una exploración de los datos, se buscó un grupo reducido pero lo suficientemente representativo para evaluar el desempeño de los programas computacionales y sus algoritmos. El set de pruebas corresponde a un subconjunto de 4.000 datos de consumos con características conocidas.

Posteriormente, para aquellos algoritmos que mostraron un buen desempeño, en términos de calidad de los resultados y tiempo de ejecución, se realizaron pruebas con la totalidad de los datos.

4.2.1. Selección de datos de prueba.

Se seleccionaron 4.000 datos, separando los consumos en 4 grupos: consumos bajos (celeste), medio-bajos (verde), medio-altos (amarillos) y altos (rojos). De cada grupo se tomaron 1.000 datos, además, se procuró que los grupos estuvieran suficientemente separados paro que los algoritmos pudieran clasificarlos en grupos distintos (Figura 4.1)



(a) b)

Figura 4.1: Datos seleccionados, (a) vista 2D, (b) vista 3D.

Los ejes X e Y corresponden a la ubicación geográfica de los consumos, cuyas unidades de medida se desconocen.

El eje Z de la Figura 4.1 (b) representa el consumo anual en kWh. Se observa que todos los datos están en un nivel de consumo parecido excepto los de rango alto, donde los consumos tienen variaciones significativas.

Los trazos marcados con negro en la Figura 4.1 (a) representan algunas agrupaciones que se esperaría obtener como resultados.

4.3 Post-procesamiento de los datos

Se analizó el resultado de los clusters encontrados y se realizaron visualizaciones para chequear que los clusters tengan sentido, por ejemplo, al encontrar puntos pertenecientes a un cluster en medio de muchos puntos de otro cluster, estos deben ser absorbidos por la mayoría, de modo que las exigencias en calidad que se asignaran para estos puntos en particular tengan oportunidad de ser aplicadas (resultaría impráctico tener disposiciones especiales para muy pocos puntos dentro de un grupo).

Una vez obtenido el clustering final, se analizó, en base a los resultados, la manera más apropiada de definir las zonas de exigencias de calidad de suministro para los consumos.

Cabe destacar que este fue un proceso iterativo, en dónde continuamente se estuvo obteniendo soluciones y realizando su post-procesamiento para comparar la calidad de una solución con respecto a otra. Además se incorporaba la información y los descubrimientos realizados en las etapas anteriores.

Capítulo 5 – Resultados y Análisis

5.1. Pre-procesamiento de los datos.

5.1.1. Datos no considerados

Al analizar los archivos de texto obtenidos y montarlos en la base de datos se observó que no todos los consumos poseían un registro completo para los 12 meses, lo que se puede explicar debido a que existen registros que se incorporaron durante el año 2012 y por lo tanto no se cuenta con un registro del año completo. Este tipo de datos no fue considerado, con lo que se obtuvo un primer total de 952.335 registros.

No se consideraron registros cuya suma anual de consumo de energía (en kWh) era negativa o cero. El hecho de tener consumos negativos se puede explicar por error en la toma de los datos o error al enviar la información por parte de la empresa concesionaria a la SEC, también por consignar los Aportes Financieros Reembolsables (AFR), en el cual la empresa hace devoluciones que se ven reflejadas como consumos negativos. Para mayor detalle, analizar el artículo 128 de la LGSE [4].

Tampoco se consideraron aquellos datos que sus coordenadas geográficas se encontraban notoriamente fuera de rango, estos son consumos que pertenecen a otras regiones pero que están mal etiquetados.

Finalmente, se eliminaron todos los datos cuyo consumo anual de energía era menor a 60 kWh. Este número de corte se basa en considerar que un consumo promedio de 5 kWh mensual puede corresponder a un consumo bajo, como ejemplo el de un hogar donde viva una sola persona. Así, valores menores dificultarían el análisis de los datos y la representatividad de estos en el contexto de un trabajo de título. Corresponde, sin embargo, para una aplicación completa y robusta, incluir este tipo de datos en desarrollos futuros.

Con ello, el total de datos considerados es de 901.064 registros, es decir, se descartó un 5,38% del total de los datos que si registraban consumo para los 12 meses del año 2012.

5.1.2. Información de los datos escogidos

A continuación se exponen visualizaciones y estadísticas de los datos que serán utilizados en los análisis posteriores.

La Figura 5.1 corresponde a una visualización de los datos donde el color azul representa aquellos consumos que están etiquetados como pertenecientes a la octava región y los de color rojo representan aquellos que están etiquetados como pertenecientes a la novena.

La Tabla 5.1 y la Figura 5.2 muestra cómo se distribuye porcentualmente la cantidad de consumos etiquetados como octava región y como novena región.

Se presentan además, en la Tabla 5.2, algunos estadísticos para la energía anual consumida.

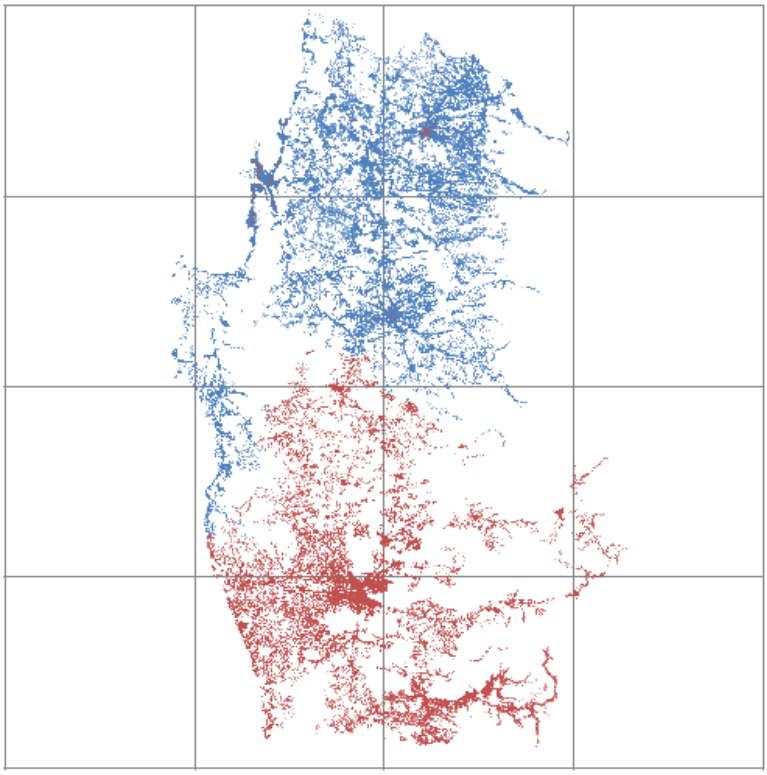


Figura 5.1: Distribución espacial de los datos considerados.

Región	Cantidad de datos	Porcentaje
octava	610.774	67,78%
novena	290.290	32,22%
Total:	901.064	100,00%

Tabla 5.1: Distribución de la cantidad de datos según región

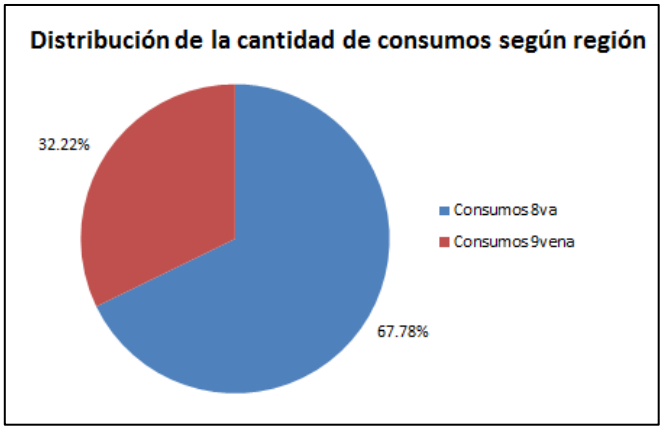


Figura 5.2: Distribución de la cantidad de consumos según región.

Estadísticos	Valor [kWh]
Min	60
Max	38.763.986
Promedio	5.830
Percentil 95%	5.598
Percentil 99%	39.538

Tabla 5.2: Estadísticos de los consumos seleccionados

Originalmente se trabajó descartando el 5% superior de los datos (mayores consumos). En medio del proceso de cálculo de soluciones se decidió no recortar los datos hacia arriba, es decir, trabajar con todos los datos mayores o iguales a 60 kWh al año, pues los consumos altos son determinantes a la hora de hablar de densidad de consumo. En la Figura 5.3 se muestra la distribución espacial de los consumos superiores a 5.598 kWh al año, se observa que tienden a estar agrupados en torno a los principales centros urbanos de la zona (el Gran Concepción, Temuco, Los Ángeles y Chillán entre otras), por lo tanto, al eliminar estos consumos se estaría disminuyendo artificialmente la densidad de consumo de estas zonas. Por esta razón se tomó la decisión metodológica de no eliminar los consumos altos (superiores a 5.598 kWh).

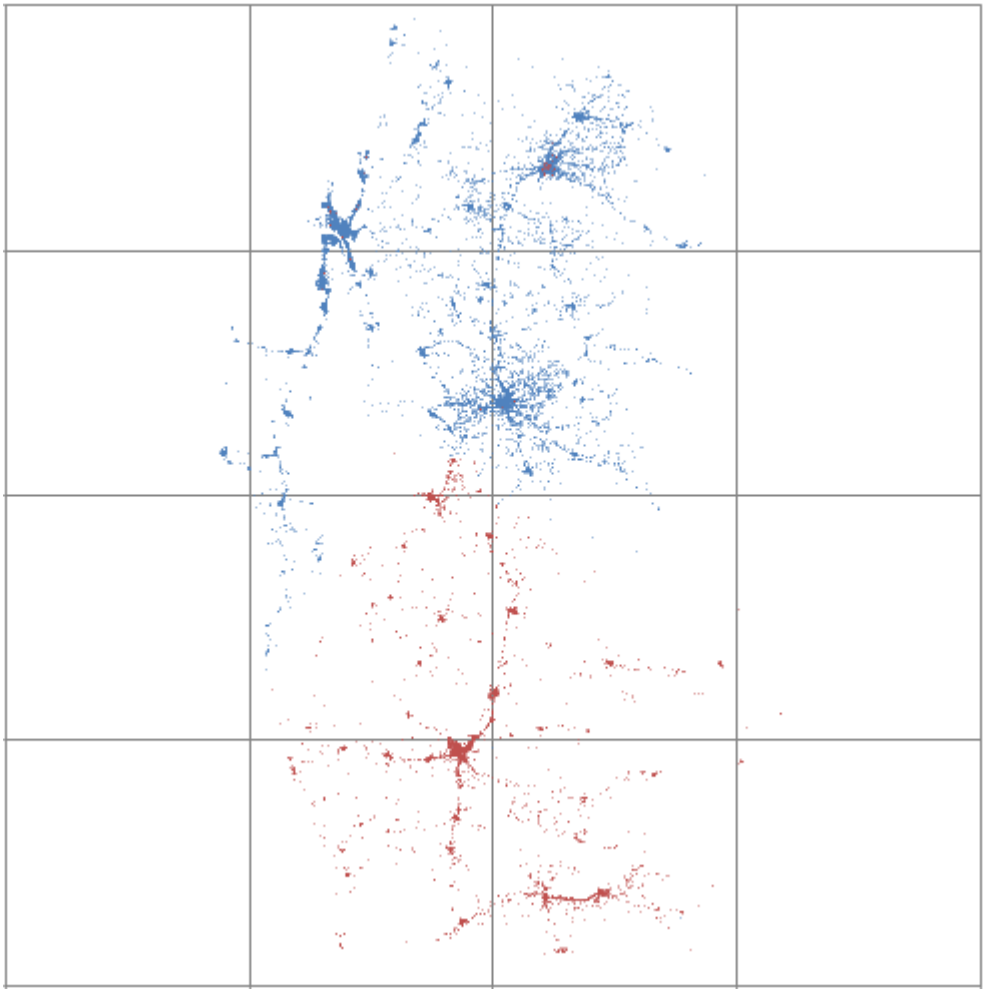


Figura 5.3: Distribución espacial del 5% superior de los consumos.

Finalmente, se elaboró un histograma (Figura 5.4) en el cual se graficaron sólo los consumos menores o iguales a 40.000 kWh. Los consumos mayores a este valor resultaron ser 8.931 (0,99% del total) y se repartían en el rango desde los 40.001 (kWh) hasta los 38.763.986 (kWh). Para no contraer demasiado el eje X del histograma estos datos fueron dejados de lado para la visualización (mas no del análisis)

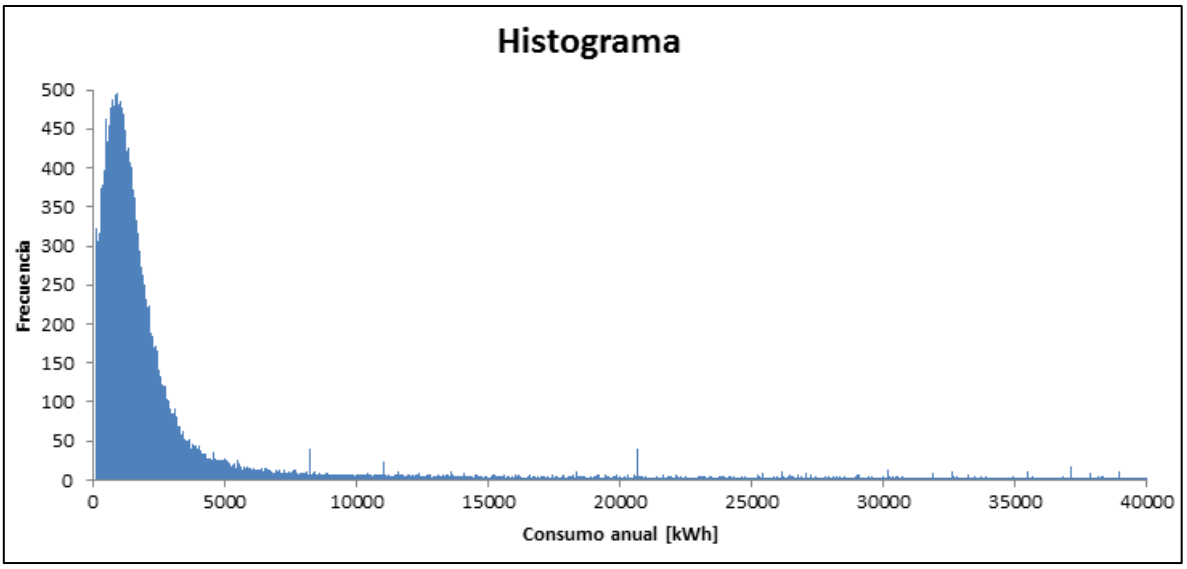


Figura 5.4: Histograma de consumos año 2012 de la octava y novena región

En el histograma se aprecia que la mayor cantidad de datos se concentra en torno a consumos bajos, pero se registra un amplio espectro de consumos.

5.2. Evaluación del desempeño de software de minería de datos y algoritmos de clustering.

En esta sección se mostraran los resultados de las pruebas realizadas para seleccionar los algoritmos de clustering y los programas computacionales escogidos para llevar a cabo la clasificación de los consumos.

Se realizaron pruebas con cada uno de los algoritmos presentes en los tres programas. Para cada uno se utilizaron los parámetros por defecto. En el caso de ser necesario se utilizó K = 4 (número de clusters) y los atributos “x”, “y” y consumo anual de energía de los datos.

Se expone una síntesis de los resultados en la Tabla 5.3.

El número dentro de cada celda es el tiempo (en segundos) empleado por el algoritmo (filas) y el software (columnas) correspondiente en calcular una solución de clustering.

Los colores representan, de manera cualitativa, la calidad de la solución, para la cual se consideraron los siguientes criterios:

- Que los grupos formados tengan sentido (cercanos considerando distancia geográfica y de consumo similar).
- Detección outliers.
- Que no se dividan cluster de manera arbitraria.

	CLUTO	RapidMiner	Weka
Bisecting K-Means	0,253		
Rbr	0,266		
Bagglo	8,476		
K-Means	0,283	0	0,17
K-Mediods		**589**	
K-Means kernel		ver tabla	
X-Means		0	0,07
DBSCAN		10	15,08
Expectation Maximization		**60**	**247,32**
Aglomerativo	7,373	**189**	
Top-Down Clustering		3	
COBWEB			7,52
Farthest First			0,03
MakeDensityBased			0,24
sIB			**654,5**

K-Means kernel	Radial	30
	Dot	78
	Polynomial	57
	Sigmoid	36
	Anova	97
	Epanechnikov	28
	gausian_combination	**2121**
	Multiqudric	**1315**

Tabla 5.3: Resultados de las pruebas de software y algoritmos en tiempo y calidad, para la selección de datos.

El color verde representa una solución de clustering buena, es decir, cumple con los tres criterios anteriores, el color amarillo quiere decir que el algoritmo falla en al menos uno de los aspectos antes descritos, pero que podría mejorar su desempeño al variar sus parámetros, por ejemplo,. el parámetro K (número de clusters), o el tipo de distancia utilizado. El color rojo representa una solución mala, que falla en más de uno de los aspectos y no se puede arreglar a través de la variación de sus parámetros.

Se destacó con negrita aquellos tiempos que se consideraron prohibitivos, dado que si para 4.000 datos se demora un tiempo considerable, para el total de datos será imposible calcular una solución.

La Tabla 5.4 muestra cuáles son los algoritmos seleccionados para probar soluciones de clustering con el total de los datos.

Las celdas marcadas con un ticket (√) corresponden a aquellas que fueron probadas con el total de los datos

Aquellas celdas que se encuentran en gris y con la marca √ corresponden a soluciones que por tiempo y calidad no son de las mejores, pero de todos modos se intentarán utilizar con el total de los datos.

De la Tabla 5.4 se puede ver que son un total de 15 los algoritmos que se probaran con el total de los datos.

	CLUTO	RapidMiner	Weka
Bisecting K-Means	√		
Rbr	√		
Bagglo	√		
K-Means	√	√	√
K-Mediods			
K-Means kernel		ver tabla	
X-Means		√	√
DBSCAN			
Expectation Maximization		√	√
Aglomerativo			
Top-Down Clustering		√	
COBWEB			
Farthest First			√
MakeDensityBased			√
sIB			

K-Means kernel	Radial	
	Dot	√
	polynomial	√
	Sigmoid	
	Anova	
	epanechnikov	
	gausian_combination	
	multiqudric	

Tabla 5.4: Resultados de la selección de algoritmos

5.3. Resultados de clustering con el total de los datos.

En esta sección se realizará una síntesis de la información obtenida cada vez que se ejecutaba un conjunto de soluciones (agrupados en este texto como “Resultados”). En el cuerpo principal de este documento no se realizará un análisis detallado del resultado de correr cada una de las múltiples soluciones obtenidas al variar los algoritmos y los parámetros, dado que no todas las soluciones fueron útiles al aportar información, pero se mostrarán algunas, a modo de ejemplo, cuando sea necesario.

5.3.1. Resultados 1

En los primeros resultados se había considerado descartar los consumos altos, trabajando sólo hasta el percentil 95%, es decir, consumos de energía menores o iguales a 5.598 kWh al año. Cuando se diga el “total de los datos” se refiere a este 95%.

El primer enfoque adoptado para resolver el problema era encontrar un algoritmo que encontrara agrupaciones de consumos en zonas geográficas bien definidas, como por ejemplo norte, sur, centro, cordillera o costa y que contuvieran un sólo tipo de consumos: bajos, medios o altos. Esto se traducía, para la visualización, en que el cluster abarcara sólo una parte del histograma. De este modo se esperaba encontrar clusters compuestos por el par zona - consumo al estilo de: “norte - alto consumo”, “norte - medio consumo”, “norte - bajo consumo”, “centro – medio consumo” y así sucesivamente.

Las soluciones se calcularon utilizando los atributos “x”, “y” y el consumo anual de energía de los datos. Nuevamente los parámetros fueron dejados por defecto y se utilizó K = 4.

La Tabla 5.5 muestra una síntesis de los resultados obtenidos. Se observa que sólo fue posible correr 9 de los algoritmos, para el resto de los algoritmos seleccionados los programas no funcionaron adecuadamente ya que terminaban abruptamente o demoraban demasiado tiempo en su ejecución (marca morada).

Las soluciones corridas con el software CLUTO fueron muy similares entre sí, siendo la solución direct (k-means) la que se observó mejor (menor tiempo y cualitativamente mejor agrupación). Las soluciones se clasificaron de mediana calidad puesto que no se reconocían agrupaciones geográficas. Los puntos pertenecientes a los clusters estaban esparcidos por todo el espacio y no se concentraban en una sola zona, pero algunas zonas de los histogramas se consideraban exclusivas, es decir, sin traslapes entre los distintos clusters. A modo de ejemplo, en la Figura 5.5 se muestran los histogramas para una de las soluciones encontradas con CLUTO. Se cree que el aumentar el número K podría producir que estas zonas exclusivas del histograma queden contenida en clusters distintos.

La solución aglomerativa del software CLUTO no fue capaz de correrse con el total de los datos.

.	CLUTO	RapidMiner	Weka
Bisecting K-Means	44,11		
Rbr	32,50		
Bagglo			
K-Means	10,87	68,00	427,7*
K-Mediods			
K-Means kernel		ver tabla	
X-Means		114,00	36,25
DBSCAN			
Expectation Maximization			
Aglomerativo			
Top-Down Clustering		>3600	
COBWEB			
Farthest First			5,01
MakeDensityBased			128,39
sIB			

K-Means kernel	Radial	
	Dot	>3600
	Polynomial	
	Sigmoid	
	Anova	
	Epanechnikov	
	gausian_combination	
	Multiqudric	

Tabla 5.5: Resultados clustering en tiempo y calidad, para el total de datos.

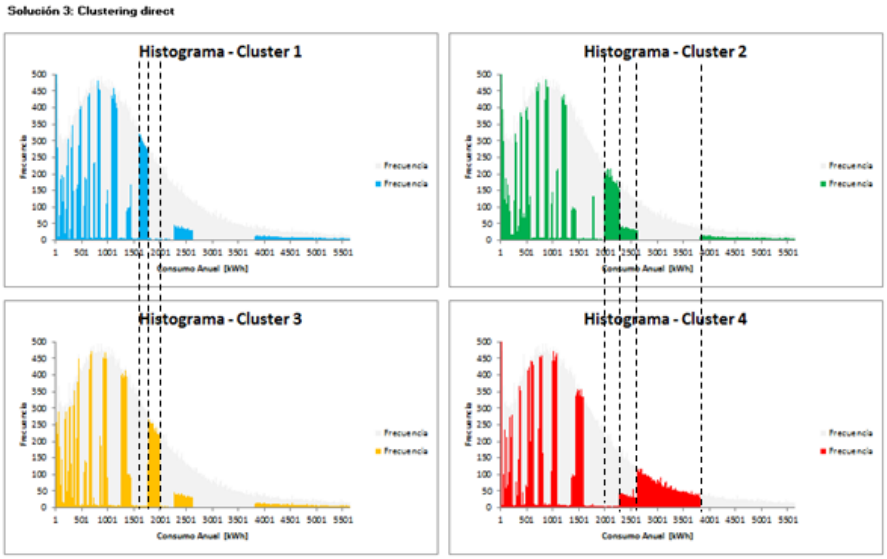


Figura 5.5: Histogramas de consumos para la solución CLUTO - direct

Con el software RapidMiner sólo fue posible correr dos algoritmos: K-means y X-means. Las soluciones aglomerativa Top-Down y la basada en un enfoque probabilístico Expectation Maximization no pudieron ser corridas con el total de los datos.

Las soluciones obtenidas con los algoritmos K-means y X-means resultaron ser muy similares entre sí y se consideraron como malas soluciones puesto que sólo agrupaba geográficamente, sin considerar una separación por tipos de consumos, existía presencia de consumos bajos medios y altos en todos los clusters (ver Figura 5.6 y Figura 5.7). Además los tiempos de procesamiento eran más altos que las soluciones calculadas en CLUTO.

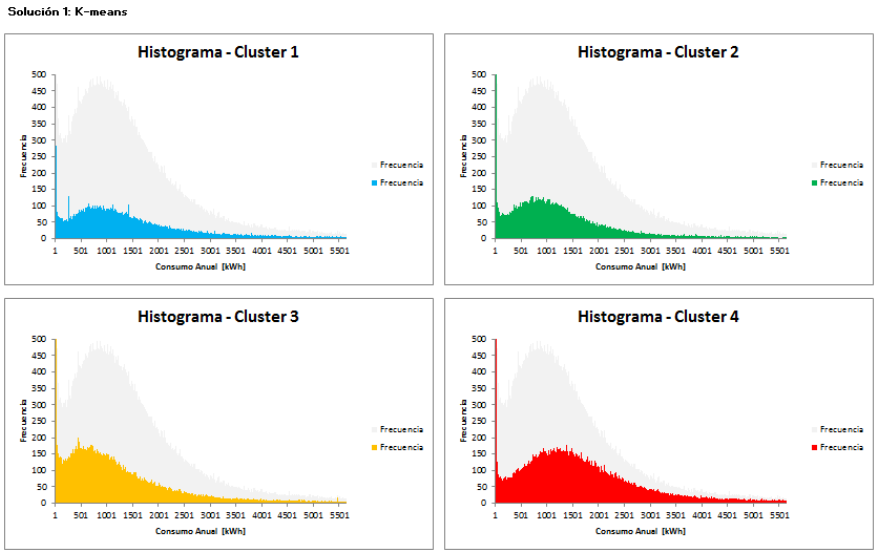


Figura 5.6: Histogramas de consumos para la solución RapidMiner – K-means.

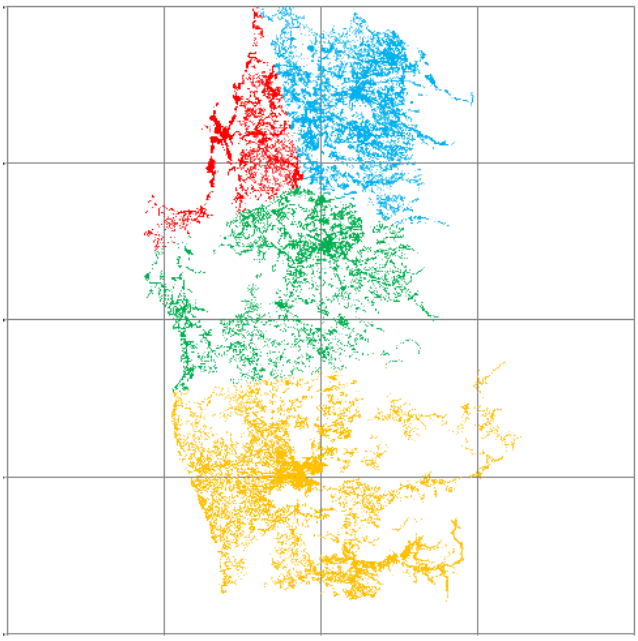


Figura 5.7: Distribución espacial de los consumos según la solución RapidMiner – K-means.

3 de las 4 soluciones obtenidas con el software Weka se consideraron como buenas dado que detectaban “capas” de alto, bajo y medio consumo para un mismo lugar. La cuarta solución corresponde a la ejecutada con algoritmo X-means, que sólo encontró 2 clusters, separando los consumos sólo en “norte” y “sur” sin tomar en consideración los distintos tipos de consumos (altos, medios y bajos).

Cabe destacar que para poder ejecutar la solución con el algoritmo simple K-means (mostrado como K-means en la Tabla 5.5: Resultados clustering en tiempo y calidad, para el total de datos. y marcado con un asterísco) se tuvo que realizar un muestreo aleatorio de los datos considerando sólo un 75% del total. Al intentar correr con porcentajes superiores, el programa se cerraba subitamente sin llegar a entregar resultados. Esto pudo haber influido en que la solución encontrada con este algoritmo y un porcentaje menor de los datos fuera mejor que otras soluciones.

Dentro de las soluciones consideradas como buenas se decidió utilizar la solución FarthestFirts para seguir avanzando en las pruebas (cambiando los parámetros) dado que la agrupación geográfica tenía sentido (Figura 5.9), separaba por grupos con distintos tipos de consumo (Figura 5.8), se pudo correr con el 100% de los datos y en un tiempo muy bajo.

Es importante señalar de esta solución que encontró una zona de alto consumo, que representaba a algunas ciudades, y tres capas que cubrían todo el mapa: una de baja consumo en la cordillera, y otras dos al norte y al sur, siendo la zona norte (celeste) la concentra la mayor parte de los consumos.

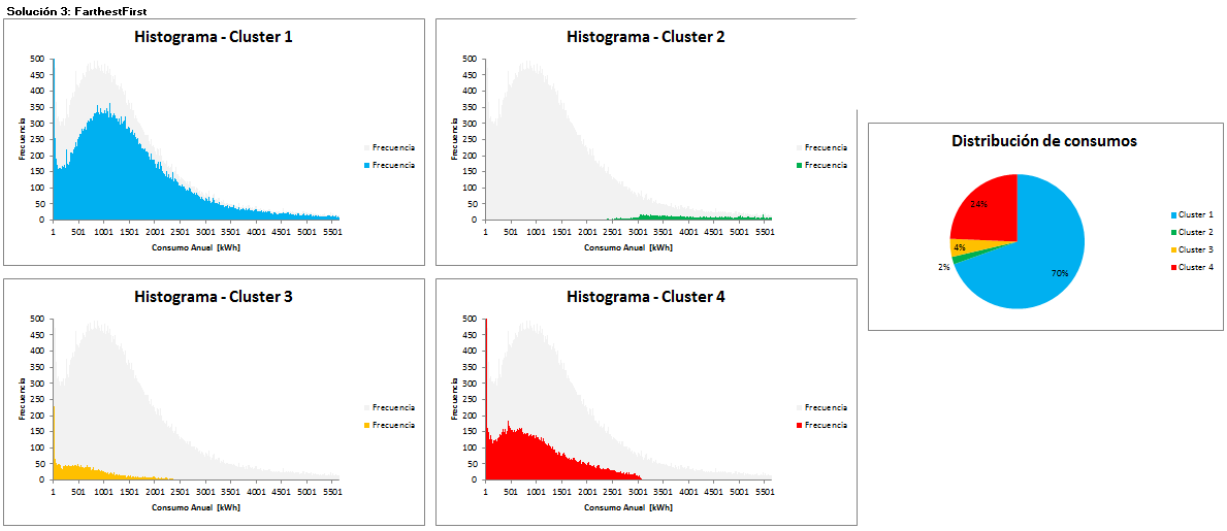


Figura 5.8: Histogramas de consumos para la solución Weka – FarthestFirst

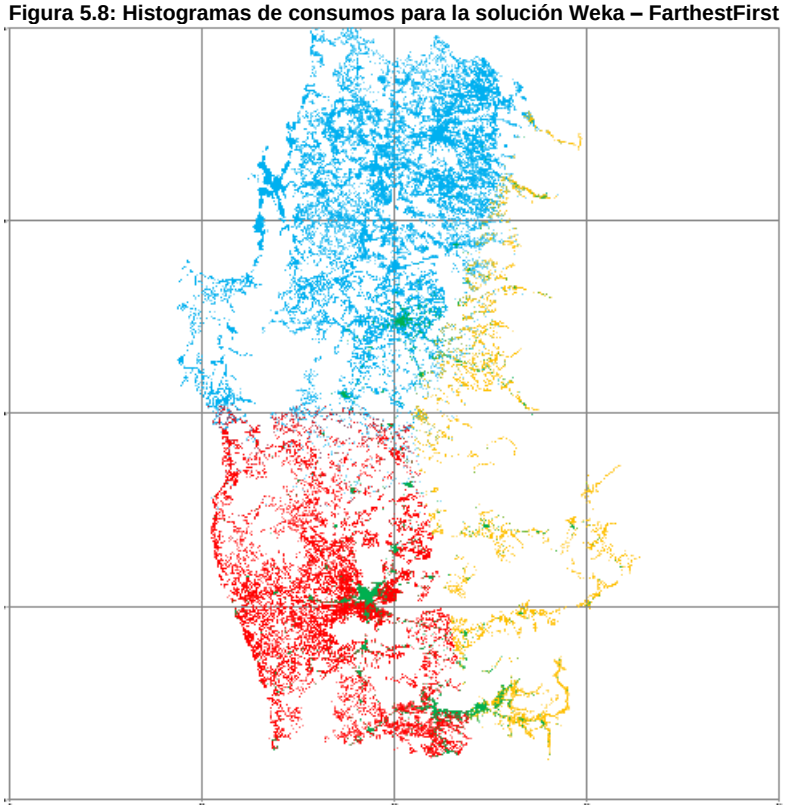


Figura 5.9: Distribución espacial de los consumos según la solución Weka – FarthestFirst.

El siguiente resultado se obtuvo utilizando el algoritmo (FarthestFirts) y el software (Weka) escogido y llevando el parámetro K (número de cluster) a 10, obteniéndose el resultado que se ilustra en la Figura 5.10, Figura 5.11 y Figura 5.12.

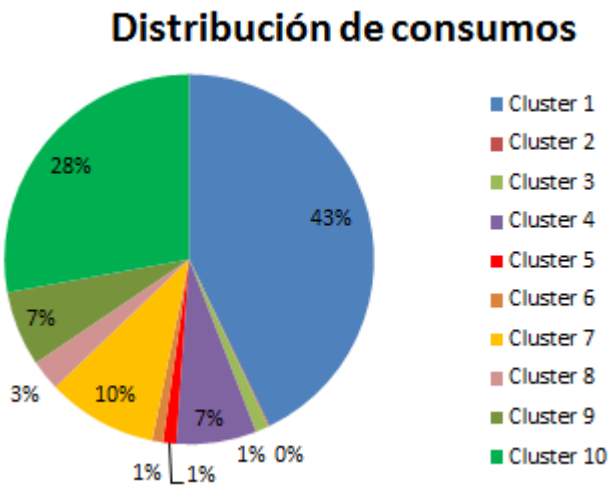


Figura 5.10: Distribución de la cantidad de consumos por clusters para la solución Weka – FarthestFirst con K = 10

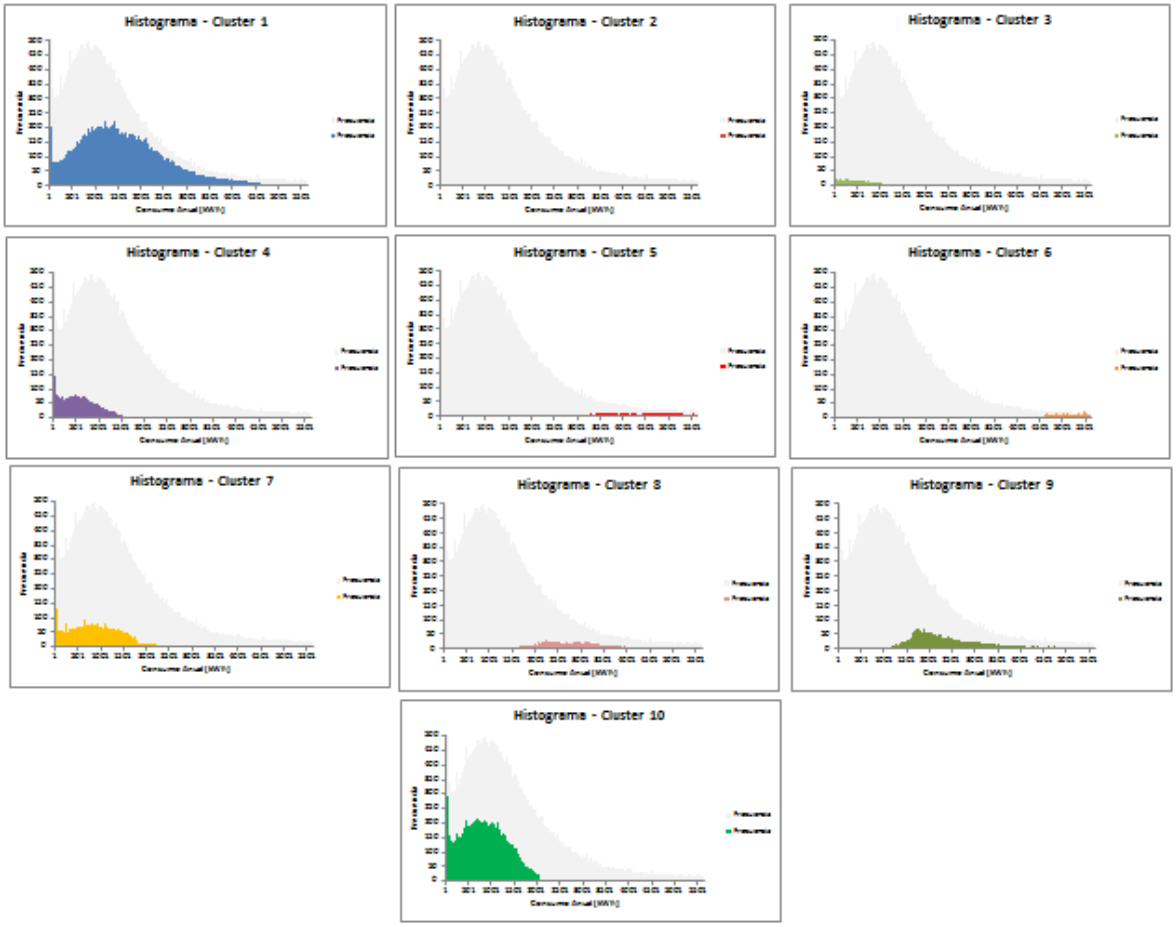


Figura 5.11: Histograma de consumos para la solución Weka – FarthestFirst con K = 10

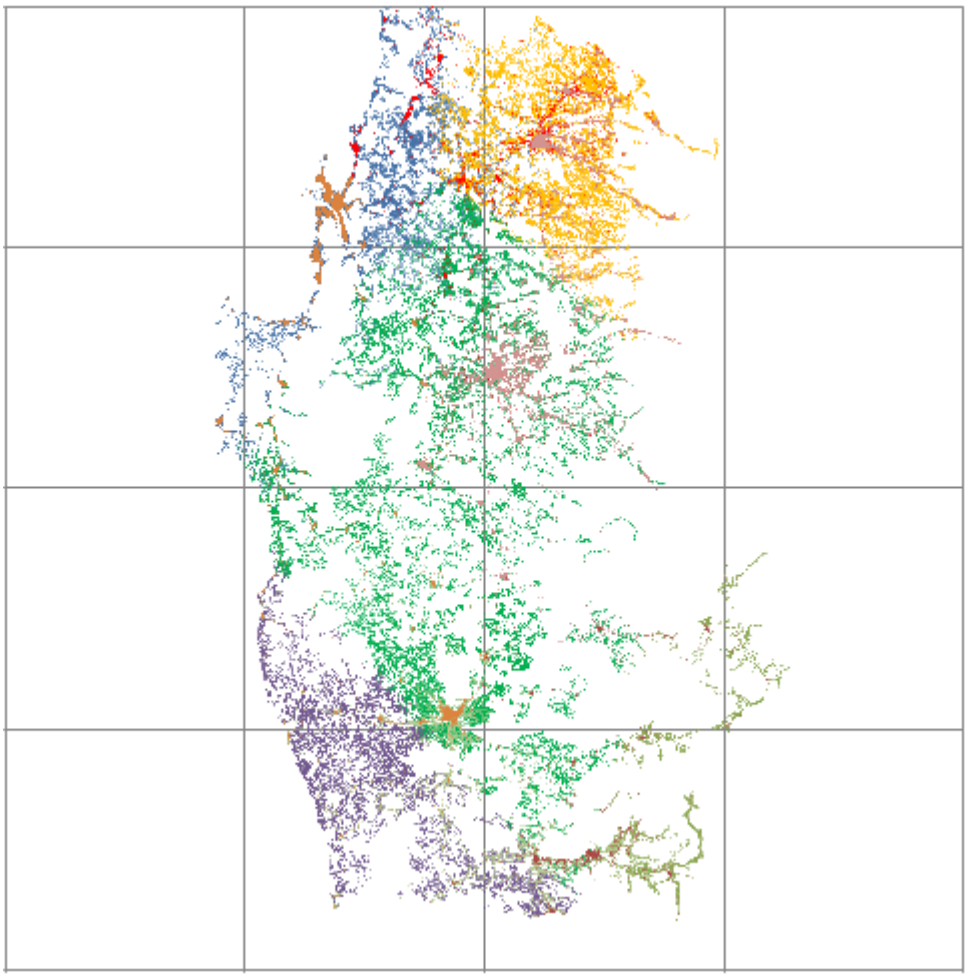
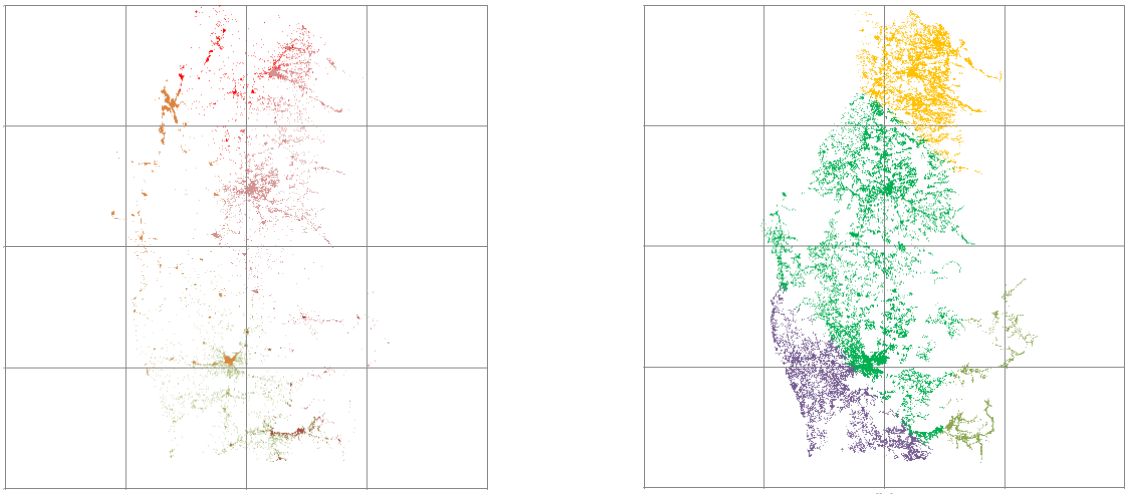


Figura 5.12: Distribución espacial de los consumos según la solución Weka – FarthestFirst con K = 10.

En la Figura 5.11 se puede observar en los histogramas que existen algunos cluster de alto consumo: 2, 5, 6, 8 y 9 (medio-alto) y otros que son de consumos bajos: 3, 4, 7 y 10, Además del cluster 1 que concentra gran cantidad de los datos (43%) que distribuyen en todo el espectro de consumos (desde bajos a altos)

En la Figura 5.13 se grafican por separado los clusters de alto consumo (a) y los de bajo consumo (b). La superposición de estas dos imágenes más la inclusión del cluster 1 es lo que se encuentra graficado en la Figura 5.12.

Se aprecia la formación de “capas” de alto consumo y de bajo consumo, las cuales pueden ser separadas en zonas como se muestra en la Figura 5.14. Esto era lo que se había estado buscando. En esta figura, los clusters cuyos nombres se encuentran escritos con letra roja representan cluster de alto consumo y los escritos con letra azul representan a aquellos de bajo consumo.



(a) (b)

Figura 5.13: (a) Distribución espacial de los clusters de alto consumo (b) Distribución espacial de los clusters de bajo consumo.

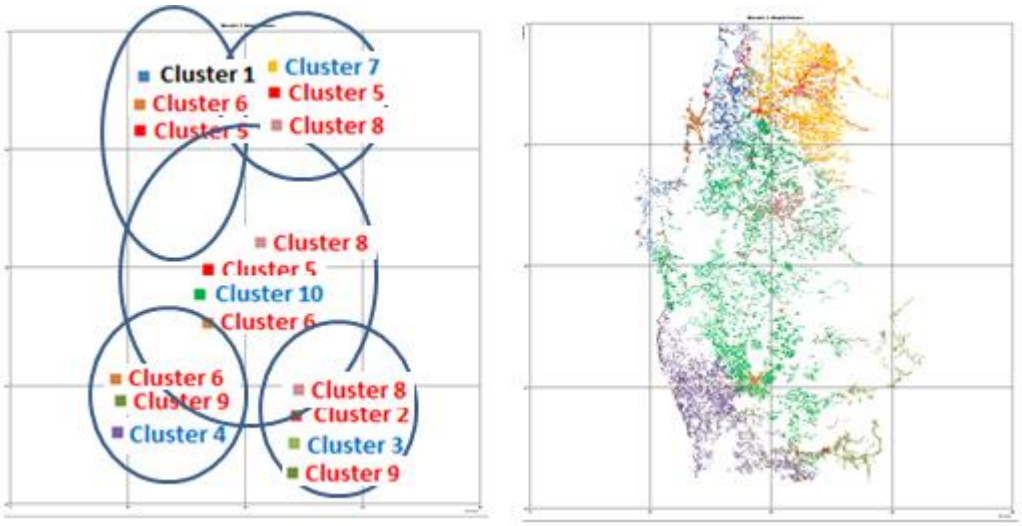


Figura 5.14: Agrupación de clusters por zonas

5.3.2. Resultados 2

Dado que los resultados finales expuestos en la sección Resultados 1 mostraban agrupaciones coherentes con los criterios de separación por zonas de consumo, se siguió experimentando con estas soluciones. Esta vez se consideró cambiar la distancia (similitud o métrica) utilizada por el algoritmo FarthestFirst de Weka.

Weka sólo tiene implementada dos métricas, la distancia Manhattan (distancia del taxi o distancia módulo) y la distancia Euclidiana, con la cual se había realizado la solución de clustering anterior.

La solución con distancia Manhattan resultó ser similar a la con distancia Euclidiana por lo que no se darán mayores detalles.

Se implementó el algoritmo K-means en MATLAB considerando una nueva distancia llamada “momento de carga”, definida como la distancia euclidiana multiplicada por el consumo de energía

$$Momento\ de\ carga = \sqrt{(x_i - Cx_i)^2 + (y_i - Cy_i)^2} \cdot kWh_i$$

Dónde:

x_i: Coordenada x del consumo i.
Cx_i: Coordenada x del centroide del cluster al cual pertenece el consumo i.
y_i: Coordenada y del consumo i.
Cy_i: Coordenada y del centroide del cluster al cual pertenece el consumo i.
kWh_i: Cantidad anual de energía consumida por el consumo i.

Se calcularon tres soluciones con este método, considerando el total de datos, K=10, los centroides iniciales tomados como un punto al azar dentro de la muestra y criterio de parada del algoritmo cuando el centroide se reajustará menos de un 5% (en módulo) que el paso anterior.

Las tres soluciones fueron bastante distintas entre sí, y además no se estaba tomando en cuenta el consumo a la hora de agrupar.

El hecho de que las soluciones fueran distintas entre sí, se atribuye a que los centroides escogidos al azar convergen a soluciones distintas. Para ello se propuso la mejora de correr un cierto número de veces (parámetro) el algoritmo y quedarse con la mejor, entendida como la solución que posea un menor SSE (Ver sección Validación de Clusters)

El problema de no tomar en cuenta el consumo se presume que radica en los órdenes de magnitud de los datos: La distancia euclidiana se encuentra en el orden de los 10^5 mientras que el consumo en kWh está en ordenes entre 10^0 y 10^3, por lo tanto, al hacer el producto, predomina el atributo de distancia por sobre el de consumo. Se plantea como mejora el normalizar los atributos, dividiendo todos los valores de cada uno de los registros por la norma.

Se aplicaron las mejoras y se corrió 100 veces el algoritmo, se elige este número porque el algoritmo era rápido en su ejecución y se ve una convergencia a cierto SSE y a cierta solución cada vez que se corre 100 veces.

Aun así, este método solo encontró agrupaciones geográficas, sin encontrar agrupaciones por consumo.

Se calculó una solución basada en agrupar los consumos sólo por coordenadas geográficas (utilizando los atributos “x” e “y”) para luego calcular un índice de consumo por área (kWh/Área). Cabe destacar que no interesa en que unidad de medida están representados los atributos x e y, por lo que se hablará simplemente de área (sin unidades). Para esto se escogió utilizar el software Weka y el algoritmo simpleKmeans con distancia euclidiana y K = 10.

Para calcular el área equivalente de cada cluster, se consideró un círculo cuyo radio es la distancia promedio de los puntos al centroide.

$$Área_{cluster} = \pi \cdot \overline{distancia}^2(punto_j, Centroide) \quad \forall punto_j \in Cluster$$

Este resultado se expone en las siguientes figuras y tablas:

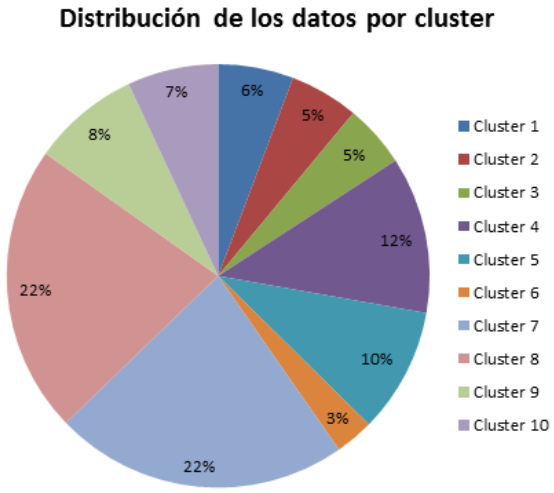


Figura 5.15: Distribución de la cantidad de datos por clusters para la solución geográfica y cálculo de consumo por área

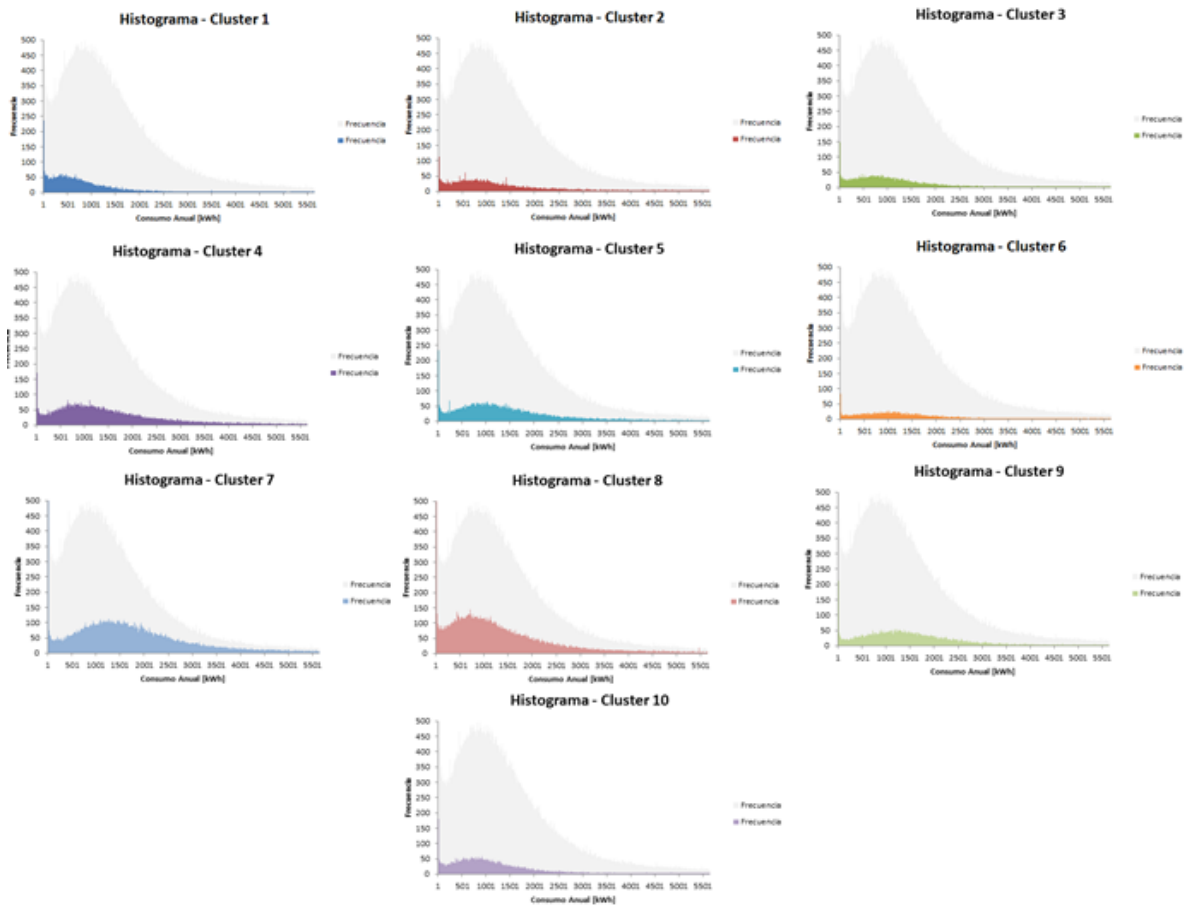


Figura 5.16: Histogramas de consumo por clusters para la solución geográfica y cálculo de consumo por área.

Cluster	Consumo [kWh]
cluster7	337.277.374
cluster8	246.958.062
cluster4	154.559.572
cluster5	116.471.811
cluster9	107.220.462
cluster10	66.915.620
cluster2	57.026.495
cluster3	43.762.537
cluster1	38.665.489
cluster6	32.128.928

Tabla 5.6: Consumo neto de cada cluster, en orden decreciente

					
Cluster	cluster8	cluster10	cluster2	cluster1	cluster3
Área	4.202.404.780	2.620.729.245	2.415.144.529	2.224.649.227	1.942.132.040
					
Cluster	cluster5	cluster4	cluster6	cluster9	cluster7
Área	1.349.615.231	718.996.372	466.408.237	311.346.226	205.152.733

Tabla 5.7: Área equivalente de cada cluster, en orden decreciente

Cluster	Área	Consumo [kWh]	**Consumo/Área**
cluster7	205,152,733	337,277,374	**1.6440**
cluster9	311,346,226	107,220,462	**0.3444**
cluster4	718,996,372	154,559,572	**0.2150**
cluster5	1,349,615,231	116,471,811	**0.0863**
cluster6	466,408,237	32,128,928	**0.0689**
cluster8	4,202,404,780	246,958,062	**0.0588**
cluster10	2,620,729,245	66,915,620	**0.0255**
cluster2	2,415,144,529	57,026,495	**0.0236**
cluster3	1,942,132,040	43,762,537	**0.0225**
cluster1	2,224,649,227	38,665,489	**0.0174**

Tabla 5.8: Área equivalente, consumo neto y densidad de consumo por área para cada cluster.

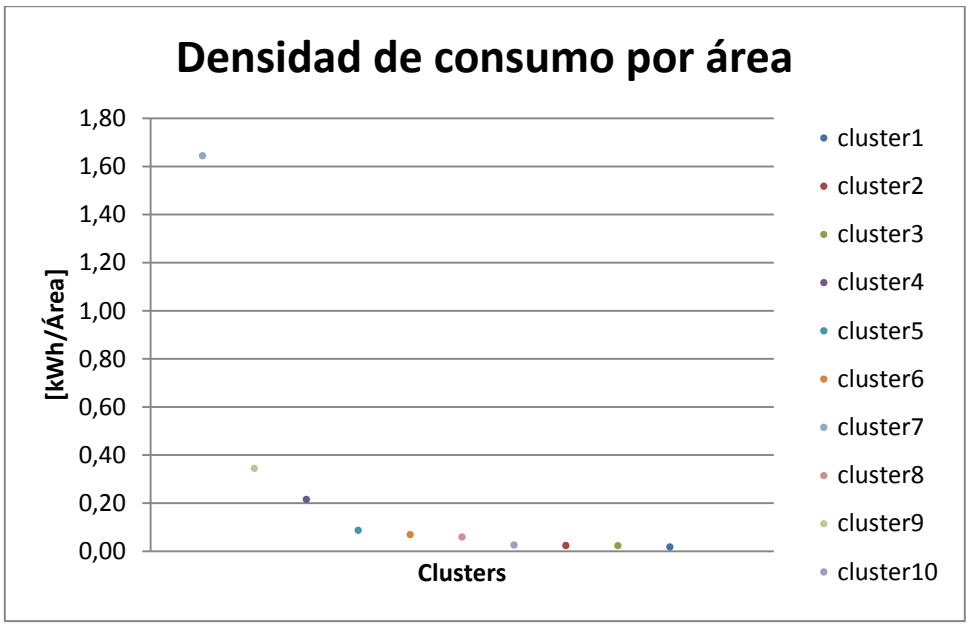


Figura 5.17: Gráfico de densidad de consumo por área, ordenado decrecientemente

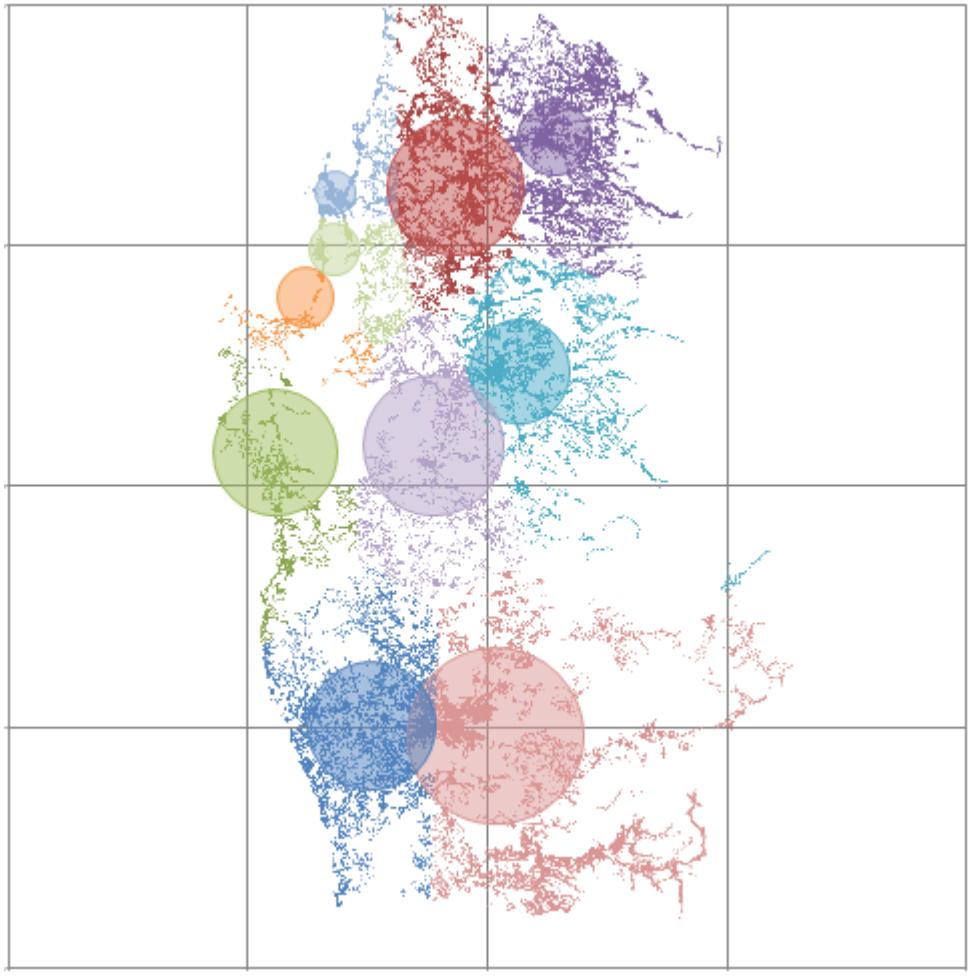


Figura 5.18: Disposición geográfica de los consumos con la representación de su área equivalente.

En esta solución se observa que el cálculo de este índice puede servir para separar en zonas de mayores o menores exigencias desde el punto de vista de calidad de suministro.

Es en este punto donde se toma la decisión de utilizar el 100% de los datos después del filtro, incluyendo el 5% superior que se había eliminado a priori, pues pueden ser determinante a la hora de calcular índices de densidad de consumo (considerando por ejemplo un número bajo de consumos pero que sumen mucha energía consumida en ciertas áreas).

5.3.3. Resultados 3

El conjunto de resultados denominados como resultados 3 consistió en repetir algunos de los resultados 2 pero esta vez utilizando el 100% de datos. Se corrieron los siguientes algoritmos:

- SimpleKmeans en Weka con distancia Euclidiana (muestro al 75% de los datos)
- K-means con momento de carga
- Clustering geográfico considerando sólo atributos “x” e “y” y cálculo de índices.

Las soluciones fueron similares a las obtenidas en los Resultados 2, en la solución de clustering geográfico se obtuvo un reordenamiento de las zonas de mayor consumo, al considerar los grandes consumos que anteriormente habían sido omitidos.

Una vez finalizados estos resultados se consideró que el esquema de cálculo de índices era un buen criterio para comparar la densidad de las distintas zonas y con ello poder asignar metas para los índices de calidad de suministro, así que se continuó avanzando en este rumbo.

5.3.4. Resultados 4

Se comenzó a trabajar en un esquema pensando en dos etapas, la primera basada en separar los datos geográficamente, i.e. solamente utilizando los atributos “x” e “y”, a través de el algoritmo k-means y luego, para cada cluster, correr un algoritmo basado en densidad (DBSCAN) para encontrar subclusters de distintas densidades a los cuales se les calcularía su índice de densidad de consumo por área.

Se utilizó esta estrategia sobre la solución con atributos “x” e “y” calculada en Resultados 3 (K=10) la cual se expone en la Figura 5.19, Figura 5.20 y Figura 5.21

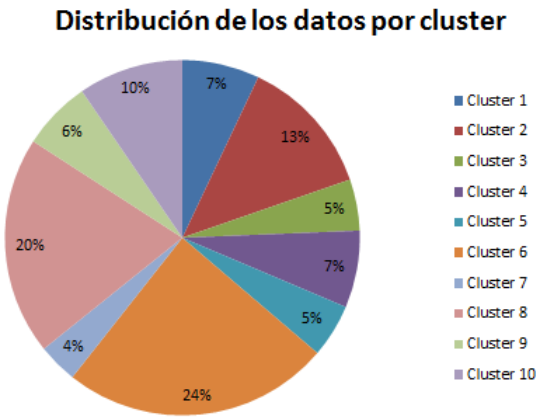


Figura 5.19: Distribución de la cantidad de datos por clusters para la solución con atributos "x" e "y"

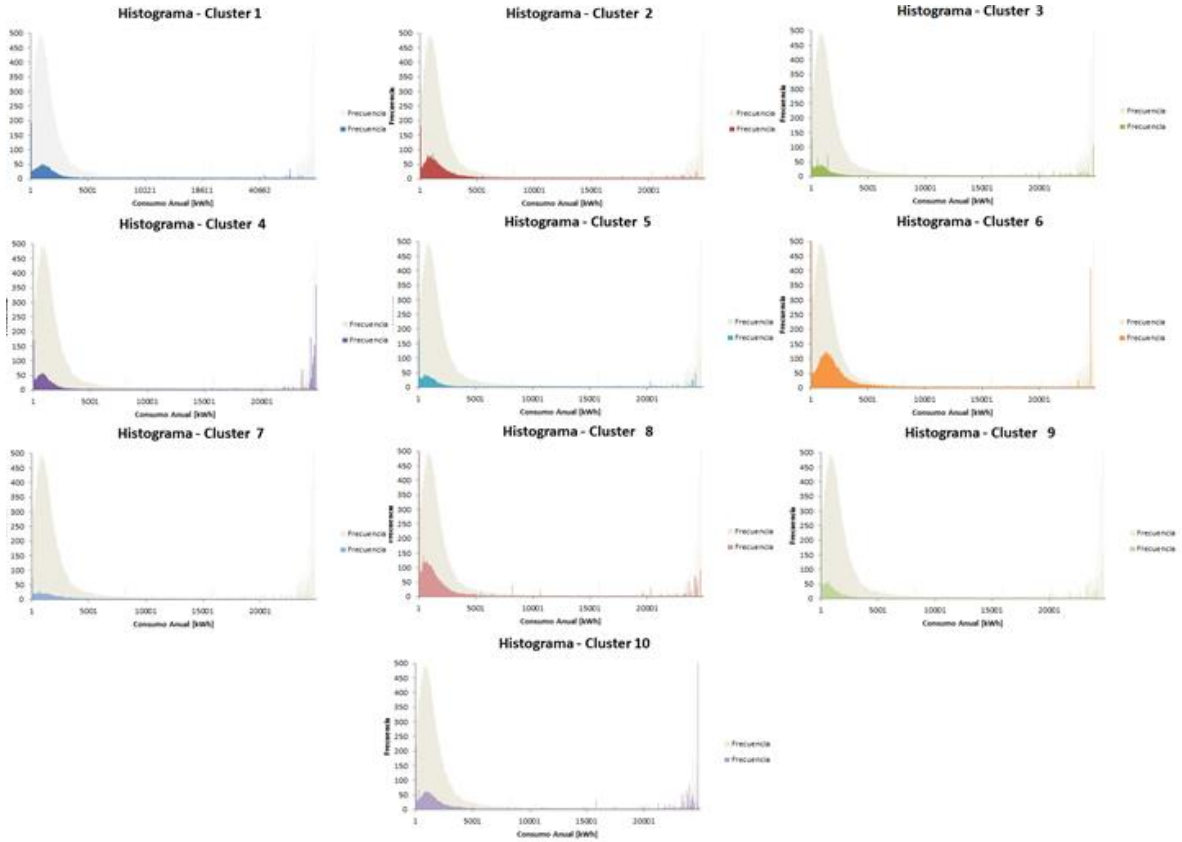


Figura 5.20: Histogramas de consumo por clusters para la solución con atributos "x" e "y"

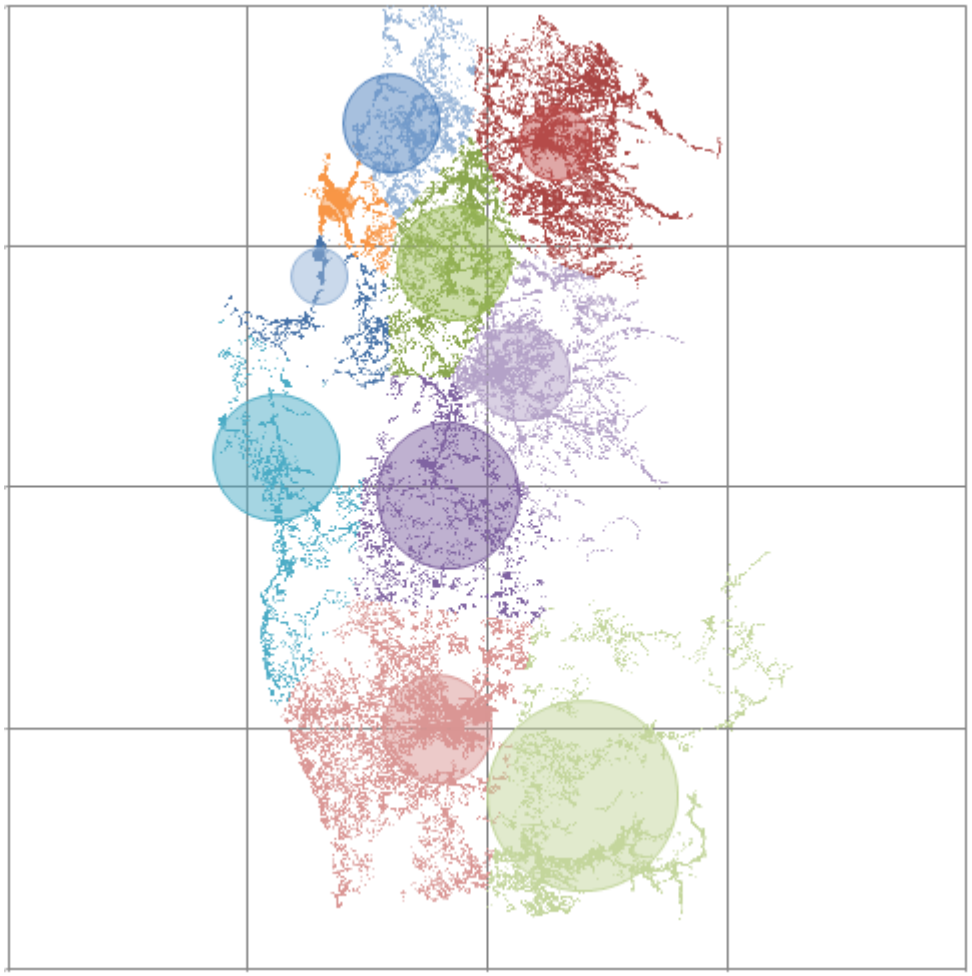


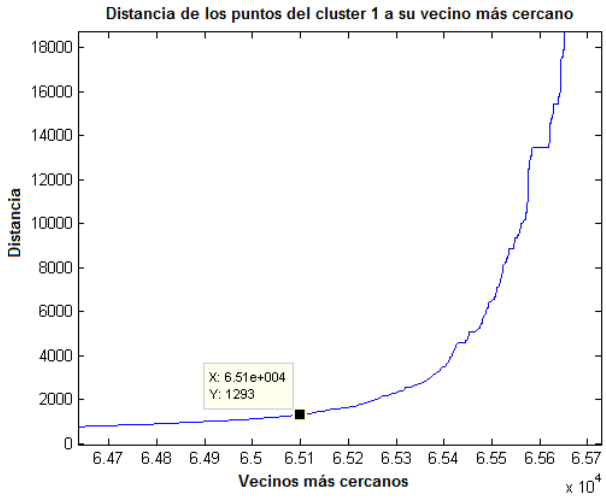
Figura 5.21: Disposición geográfica de los consumos con la representación de su área equivalente para la solución (X, Y).

El algoritmo DBSCAN, como fue explicado en la sección 3.3.1.3, posee dos parámetros, el radio de la vecindad épsilon (Eps) y el número mínimo de puntos MinPts. En esa sección también se explica un método para estimar el parámetro épsilon en base a graficar de manera decreciente los vecinos más cercanos, este método se programó en MATLAB.

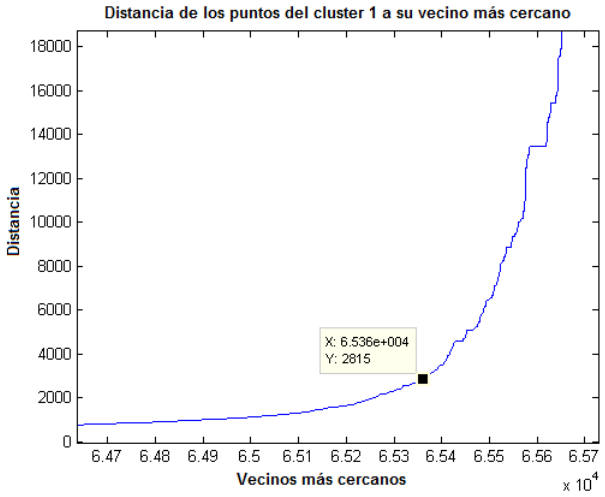
Para cada cluster, se eligieron 3 valores de épsilon para experimentar: Una antes del codo, uno en el codo y uno posterior al codo.

En la Figura 5.22 se ejemplifica como se tomaron estos valores (para el cluster 1) y en la Tabla 5.9 se muestran los 3 resultados obtenidos para cada cluster, donde 1 significa antes del codo, 2 significa en el codo y 3 después del codo.

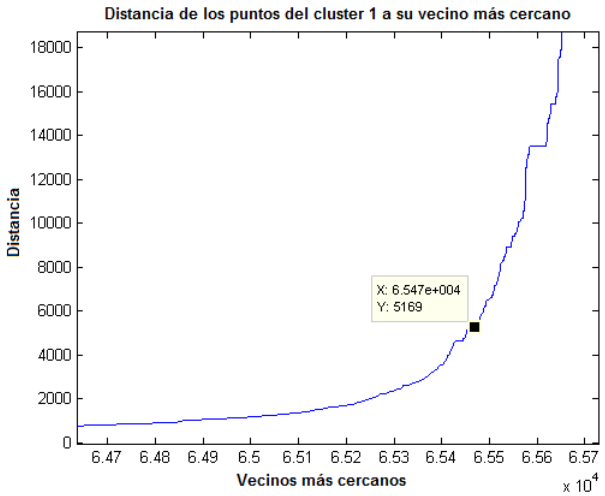
Se reitera que no interesa saber las unidades de distancia, es por esta razón que no se señalan en los gráficos de la Figura 5.22 ni tampoco en la Tabla 5.9.



(1)



(2)



(3)

Figura 5.22: Vecinos más cercanos para el cluster 1, solución con atributos “x” e “y”.

	1	2	3
C01	1.293	2.815	5.169
C02	2.084	3.907	9.942
C03	1.341	4.057	11.050
C04	1.820	5.745	12.860
C05	1.191	2.937	7.414
C06	920	2.975	8.686
C07	2.240	5.735	10.580
C08	2.082	5.139	14.920
C09	2.413	6.123	13.930
C10	4.717	16.170	28.590

Tabla 5.9: Resumen de la obtención de parámetros épsilon para la solución con atributos "x" e "y"

Se experimentó con el cluster 1 calculando tres soluciones de subclustering utilizando DBSCAN con los parámetros MinPts = 500 (un poco menor al 1% de los datos) y los 3 valores de épsilon mostrados en la tabla: ϵ_1=1.200 (solución llamada C01 - 1), ϵ_2=2.800 (C01 -2) y ϵ_3=5.200 (C01 - 3). Cada gráfica de solución se muestra con dos resoluciones, la primera, denominada (a), con la misma resolución que el gráfico general que muestra ambas regiones y la segunda, denominada (b) como un zoom en específico a la zona graficada.

En cada solución calcula con el algoritmo DBSCAN el subcluster 0 (color azul) representa puntos de ruido.

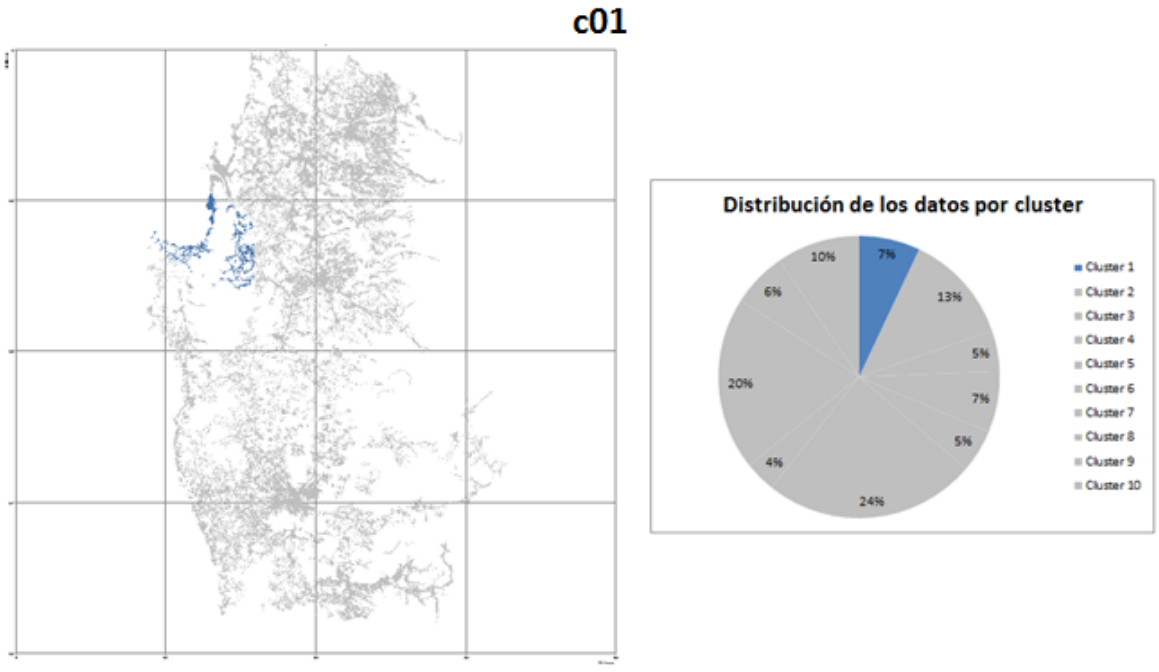
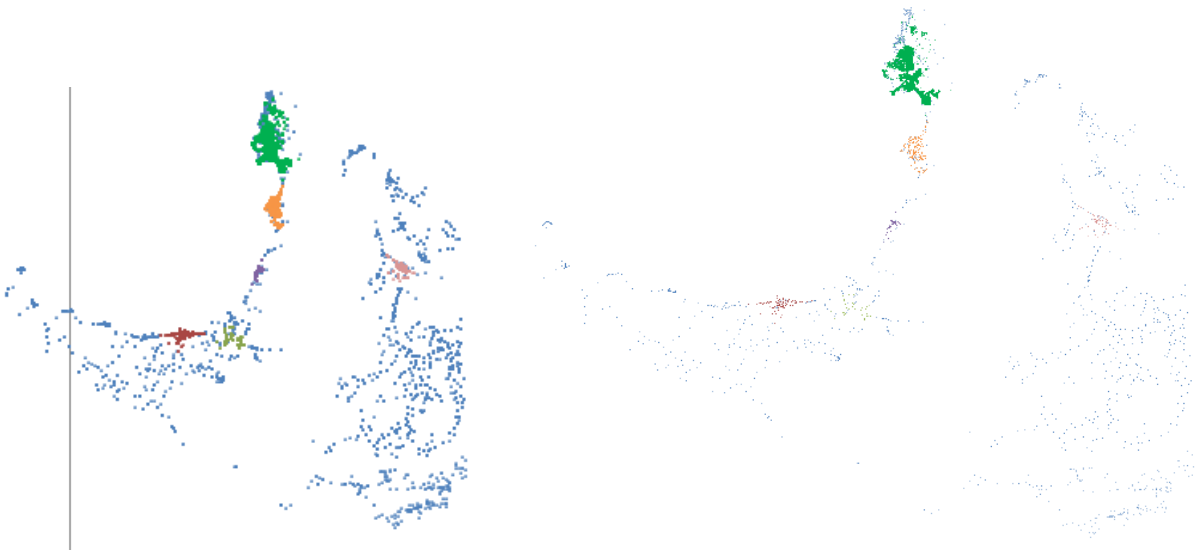
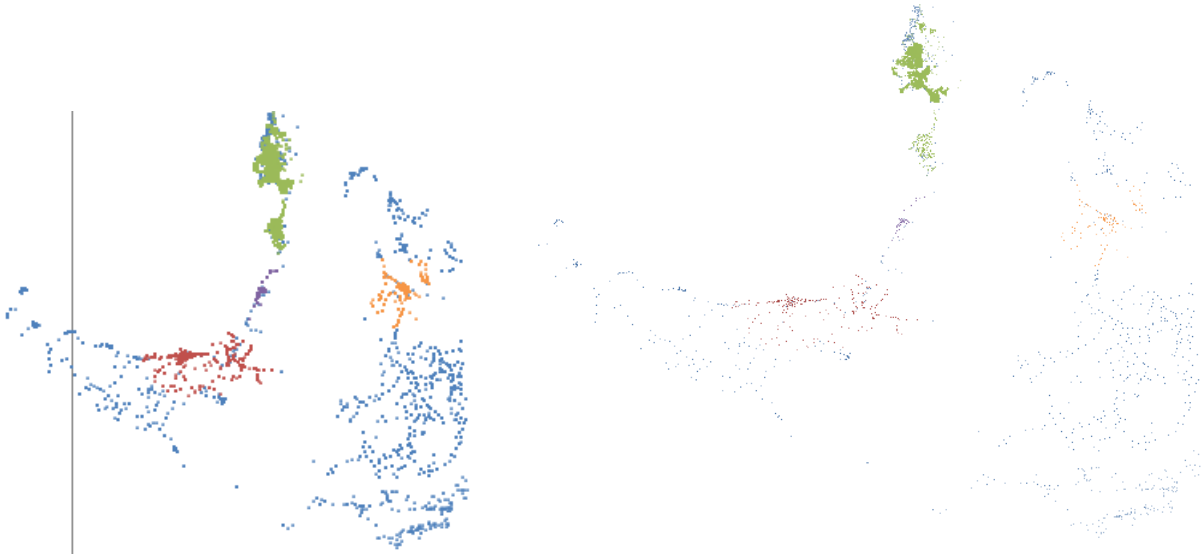


Figura 5.23: Cluster 1 en la solución (X.Y)



(a) (b)

Figura 5.24: Solución de subclustering C01 -1 (MinPts = 500, ε = 1.200), para solución con atributos “x” e “y”.



(a) (b)

Figura 5.25: Solución de subclustering C01 – 2 (MinPts = 500, ε = 2.800), para solución con atributos “x” e “y”.

(a) (b)

Figura 5.26: Solución de subclustering C01 - 3 (MinPts = 500, ε = 5.200), para solución con atributos “x” e “y”

En la solución C01 - 3 (Figura 5.26) se han marcado algunos trazos sobre el gráfico b. Lo que está marcado en verde representa 3 grupos que pertenecen al mismo subcluster pero tienen densidades distintas, por otro lado la marca naranja muestra un cluster con puntos de ruido alrededor, pero que parecieran tener la misma densidad. Por este motivo se descarta utilizar las soluciones con ϵ_3 (por sobre el codo en el gráfico de los vecinos más cercanos).

Se siguió experimentando realizar subclustering con DBSCAN con ϵ_1 y ϵ_2 y MinPts cercano al 1% de la cantidad de datos de cada cluster hasta el cluster 4 y se observó que los resultados empleando ϵ_2 tendían a realizar pocos grupos de densidades mezcladas. Por lo tanto se prefirieron las soluciones con ϵ_1 (bajo el codo) donde se encontraban un mayor número de cluster pero cada uno tenía una densidad más homogénea.

Finalmente se obtuvo el resto de las soluciones de sublcustering con DBSCAN empleando ϵ_1 y MinPts cercano al 1% de la cantidad de datos de cada cluster.

Algunas observaciones realizadas una vez obtenidos todos los resultados fueron las siguientes:

Existen algunos cluster (resultado de correr una solución k-means considerando sólo los atributos X e Y) que contienen grupos densos y para estos cluster hay puntos que son considerados como ruido y que en otros cluster podrían ser considerados como densos. Ejemplo de ellos se pueden ver en la Figura 5.27 dónde el ruido se ve más denso que el ruido de otros cluster y la Figura 5.28 dónde en la parte sur se ve que hay puntos densos que son considerados como ruido, se presume que si estuvieran en otro cluster serían considerados como puntos densos.

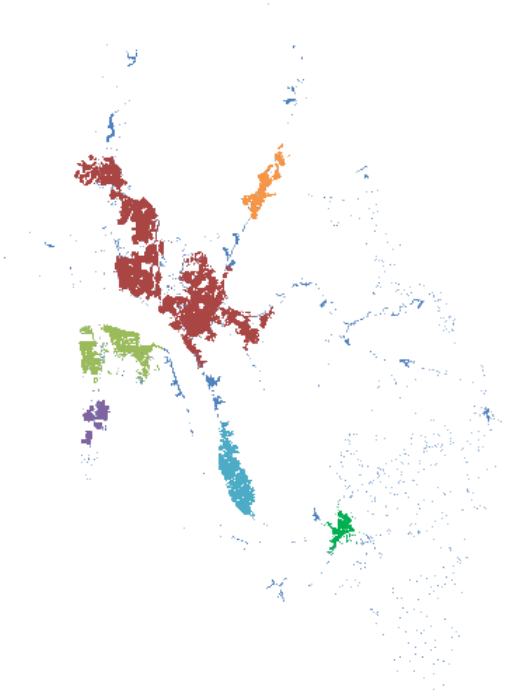


Figura 5.27: Subclustering C06 – 1

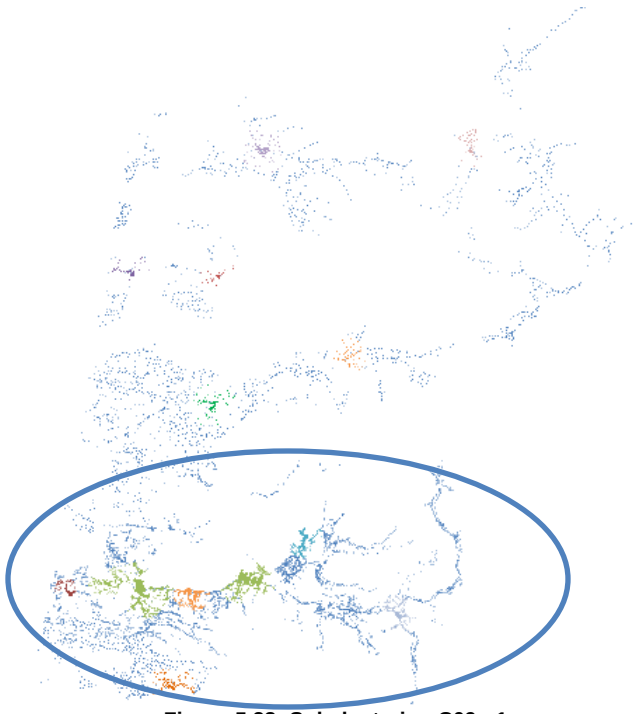


Figura 5.28: Subclustering C09 - 1

Hay soluciones que deben ser analizadas en mayor detalle. La visualización no siempre ayuda dado que el algoritmo DBSCAN basa su noción de distancia en espacio tridimensional “x”, “y” y consumo anual de energía, luego, si bien algunos cluster no se ven densos en las visualizaciones puede que realmente si lo sean.

La última observación que se realiza es que si bien ayuda el método de graficar las distancias a los vecinos más cercanos, ordenadas de manera creciente, no es infalible, ya que finalmente definir dónde está el codo y que punto tomar antes del codo es arbitrario, como ejemplo, en la Figura 5.29, se puede ver que es necesario tomar un épsilon aún menor.

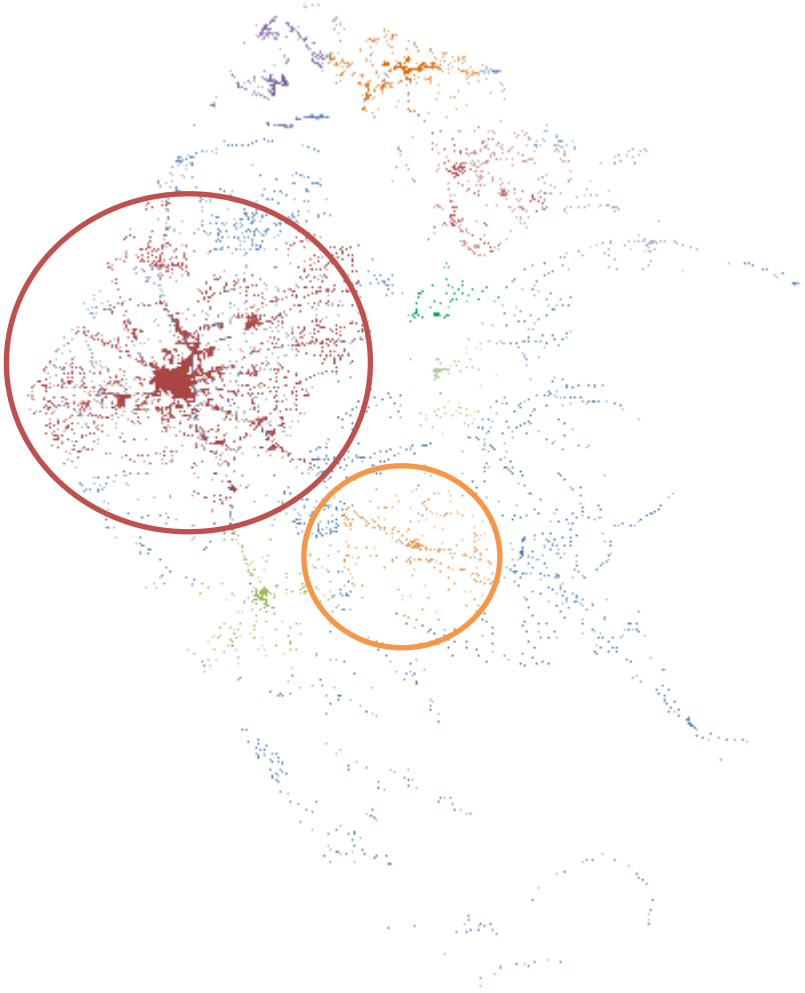


Figura 5.29: Subclustering C10 - 1

Para el siguiente conjunto de soluciones se decide evaluar un K = 6 = número de provincias de la octava región, más el número de provincias de la novena región con el fin de que los centroides de los cluster estén entorno a las capitales provinciales, las cuales se suponen (a priori) como las ciudades con altas densidad de consumos. De esta manera se espera que los puntos densos y de ruido sean más uniformes dentro de los distintos subcluster y no existan puntos de ruido que en otro cluster si serían considerados como puntos densos.

Se decide seguir utilizando el método de los vecinos más cercanos para intentar acercarse al parámetro épsilon y ajustarlo en caso que las soluciones obtenidas no sean satisfactorias.

Se experimentará variar el parámetro MinPts dado un épsilon fijo.

Una vez obtenida las soluciones (cuyo parámetros épsilon y MinPts ya han sido ajustados de modo de tener una buena solución) se calculará el índice Consumo/Área y se graficará de manera ordenada. Similar a lo obtenido en los Resultados 2, pero esta vez considerando los subcluster en vez de cluster. (Ver Figura 5.17)

Esto último plantea un nuevo problema: ¿cómo medir el área de los puntos de ruido?

Finalmente, una vez obtenido el gráfico de densidad de consumo por área, se plantea dividir de alguna forma los subcluster, de manera de formar grupos a los cuales asignarles metas de calidad de suministro y comparar con el esquema actual.

5.3.5. Resultados 5

Se obtuvo una solución de clustering planteada en 3 etapas, la primera a través del algoritmo K-means en Weka, considerando K = 6, luego, el cálculo del parámetro épsilon para cada cluster, que servirá como entrada para la tercera etapa donde se calcula una nueva solución de clustering, dentro de cada uno de los clusters, basada en densidad a través del algoritmo DBSCAN.

En esta oportunidad el algoritmo k-means consideró los atributos “x” e “y” (al igual que en los resultados anteriores) y el algoritmo DBSCAN también consideró sólo los atributos “x” e “y” (a diferencia de los resultados anteriores, dónde se estaban considerando los atributos “x”, “y” y el consumo anual en kWh)

Esta decisión se tomó para intentar arreglar el problema de los cluster que a simple vista se veían con densidades similares pero algunos de ellos considerados como puntos de ruido y otros como puntos densos, debido a las diferencias en consumo.

El hecho de sólo emplear los atributos X e Y en la etapa con DBSCAN trajo más dificultades que beneficios.

Se calcularon las distancias a las vecinos más cercanos a través del algoritmo implementado en MATLAB y se graficaron ordenadas de manera creciente. A simple viste se notaba el contraste con la solución anterior (que consideraba consumo) puesto que el codo obtenido en las gráficas de las nuevas soluciones se encontraban para un épsilon mucho más bajo. A modo de ejemplo, en la Figura 5.30 (a) se graficó la distancia considerando los 3 parámetros para el cluster 1 donde se ve que el codo está entre los 2.000 y los 4.000, en cambio en la Figura 5.30 (b) se gráfica la distancia para el cluster 1 considerando sólo los parámetros X e Y el codo está por debajo de los 500.

Esta diferencia se observaba al graficar cualquier par de clusters, que el codo se encontraba al menos un orden de magnitud más abajo en las nuevas soluciones.

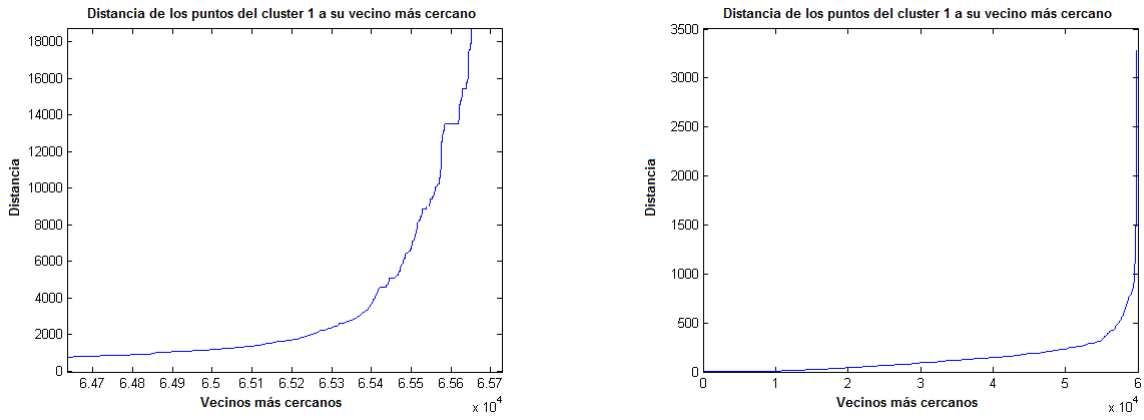


Figura 5.30: Distancia ordenada de los vecinos más cercanos, para el cluster 1. (a) Utilizando parámetros "x","y" y consumo con distancia 3d. (b) Utilizando sólo atributos "x","y" y distancia 2d

Se continuó calculando la solución para el cluster 1, considerando un punto antes del codo de saturación en la gráfica (b) de la Figura 5.30 como épsilon, tal como se había hecho en la solución anterior. Se asumió que este valor era adecuado y se experimentó variando el parámetro MinPts.

Se eligió este cluster debido a que contenía poco elementos y el tiempo de cálculo de solución no era demasiado grande, en comparación al resto de los clusters (alrededor de 20 minutos).

A continuación se exponen los resultados de las primeras 3 pruebas considerando ϵ = 145 fijo y variando MinPts.

Solución	Eps	MinPts	Resultado
sol1	145	6 (0,01% de los datos)	Solución con 860 clusters
sol2	145	60 (0,1% de los datos)	Solución con 147 clusters
sol3	145	598 (1 % de los datos)	Solución con 3 clusters. El 97% de los datos fue considerado ruido

Tabla 5.10: Resultados subclustering cluster 1, para ε = 145 fijo.

De estas 3 pruebas se concluye que tomar un épsilon bajo el codo, como se había hecho en las pruebas anteriores (cuando se consideraba una distancia euclidiana con 3 parámetros) no resultó.

Luego de esto se realizaron varias pruebas. Se probó variar el parámetro épsilon en las soluciones que siguen, dejando fijo el número MinPts = 598 (1%), se observó que a medida que se aumentaba el épsilon sobre 1.000 las soluciones iban siendo mejores, se volvió a cambiar el parámetro a MinPts = 2.290 (5%), incluso se utilizaron valores extremadamente altos de épsilon. El resumen de las soluciones se entrega en la Tabla 5.11 y un gráfico del espacio de soluciones se entrega en la Figura 5.31.

Notar que en la Tabla 5.11 no se incluye el resultado de la solución 8, por tratarse de una solución repetida con las anteriores.

	MinPts	eps	calidad
sol1	6	145	
sol2	60	145	
sol3	598	145	
sol4	598	300	
sol5	598	1000	
sol6	598	1500	
sol7	2990	1000	
sol9	598	1640	
sol10	2990	1640	
sol11	598	3000	
sol12	2990	3000	
sol13	598	4720	

Tabla 5.11: Soluciones para el cluster c1 variando parámetros MinPts y eps.

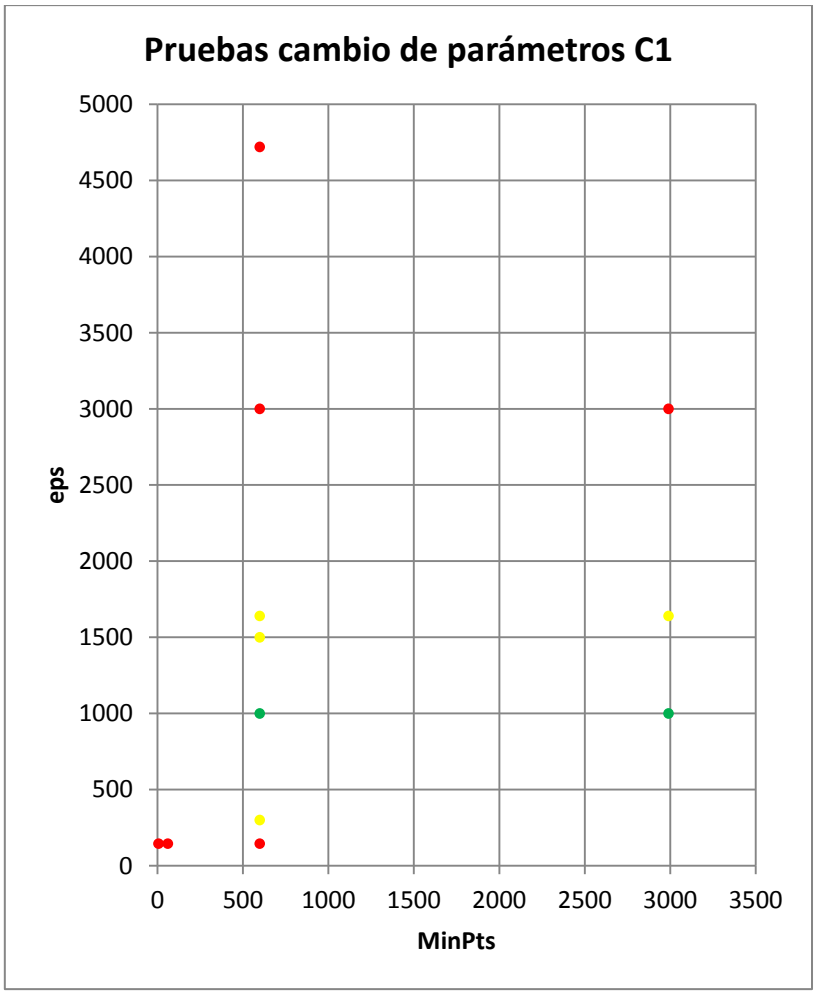


Figura 5.31: Resultados de las soluciones para el cluster c1 variando parámetros MinPts y eps.

Las soluciones se clasificaron nuevamente según criterios. Se consideraron como buenas soluciones aquellas que lograron:

- Tener poco porcentaje del total de los datos considerados como ruido (60% o menos).
- Un número de subclusters entre 5 y 15.
- Que los subclusters consideren puntos con la misma densidad.

Aquellas soluciones que lograron satisfacer 2 de los criterios fueron consideradas como soluciones regulares. Finalmente aquellas soluciones que no cumplían al menos dos de los criterios fueron consideradas como deficientes.

El resultado de estas pruebas fue desalentador: en esos momentos ya no se poseía ni si quiera un método para encontrar uno de los parámetros del algoritmo DBSCAN.

Se prosiguió con el cálculo de soluciones, ajustando manualmente los parámetros de modo de obtener buenas soluciones, según los criterios mencionados anteriormente.

Finalmente para cada cluster, se calculó el índice de área/consumo de sus respectivos subcluster como fue explicado en la sección 5.3.2 Resultados 2

En Figura 5.32 se ve cómo quedan clasificados los consumos. Se propone inicialmente un esquema de separar por percentiles, en específico esta vez se ha escogido terciles de manera de hacer más directa la comparación con el esquema actual que posee 3 zonas, pero se deja abierta la posibilidad de cambiar la forma de clasificar los grupos (Figura 5.33). Se propone también, considerar todos los subclusters 0 (ruido) como zona de exigencia 3 (las de más baja calidad) y separar los subclusters restantes en dos mitades, siendo la mitad superior zona de exigencia 1 (la de más alta calidad) y la de abajo zona de exigencia 2. (Figura 5.34)

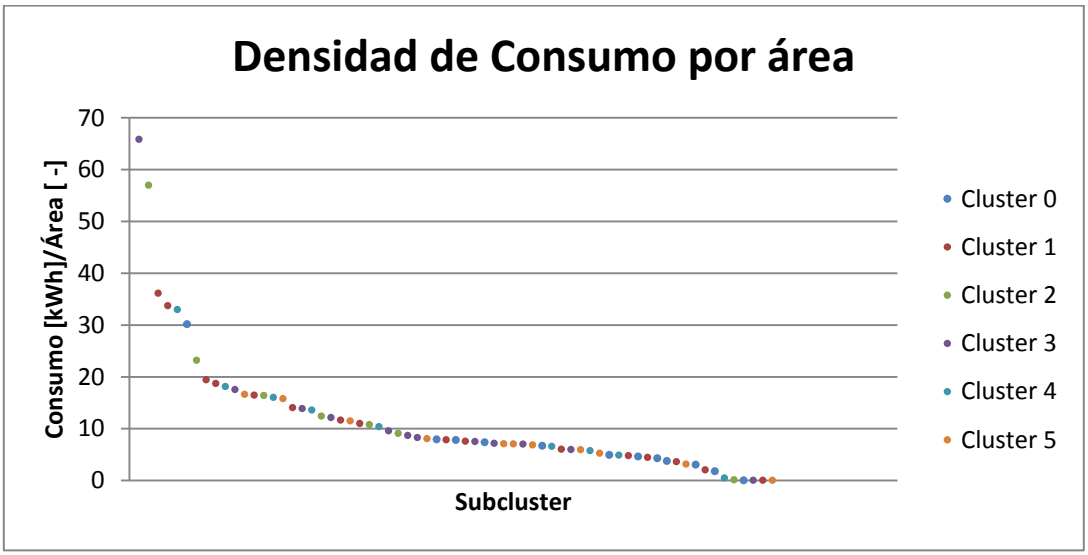


Figura 5.32: Densidad de consumo por área para solución calculada en Resultados 5

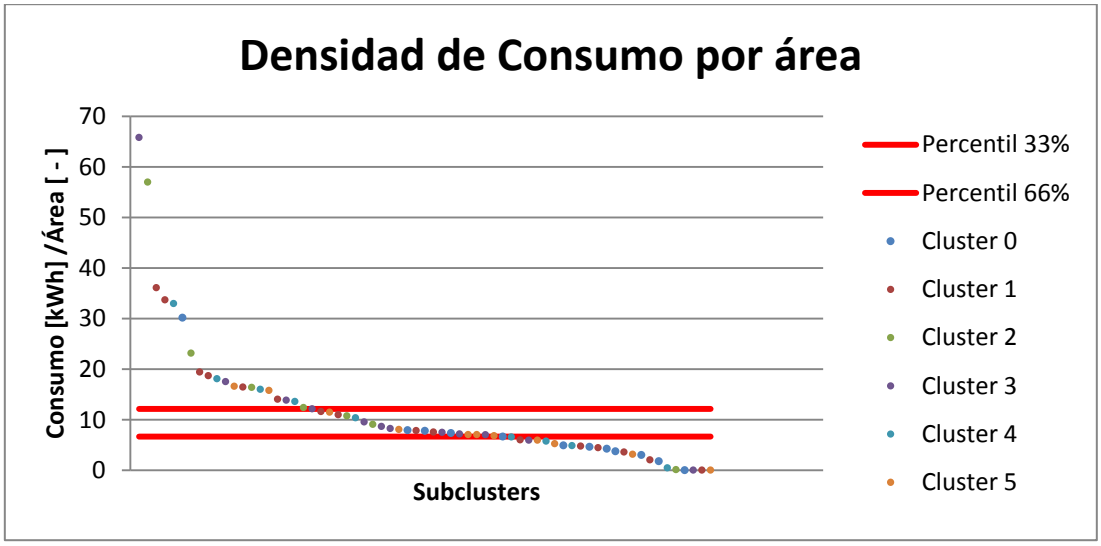


Figura 5.33: Densidad de consumo por área, divida en terciles

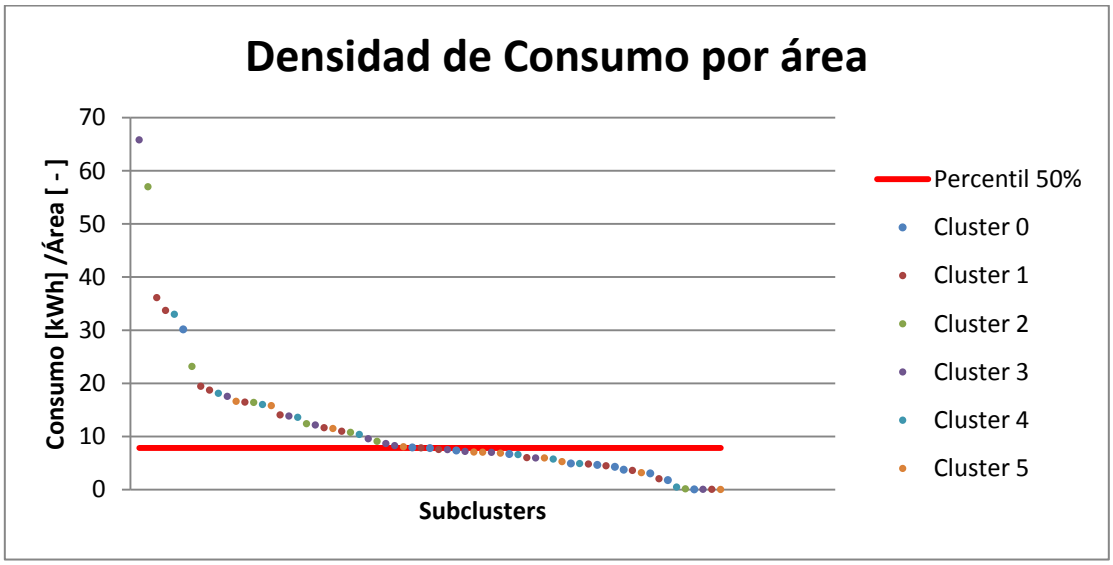


Figura 5.34: Densidad de consumo por área, divida en mitades, sin considerar a los subcluster de ruido

Finalmente se realiza una comparación de las 2 propuestas de clasificación con el esquema actual. La información para clasificar a las comunas actuales fue extraída de la Resolución Exenta 303 de 2012 de la CNE [20]

Se utilizará la siguiente convención de colores:

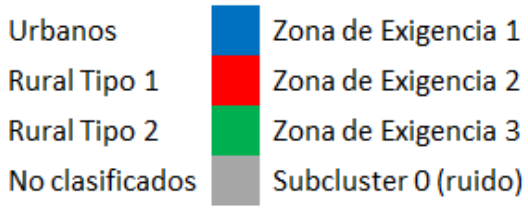


Figura 5.35: Código de colores para la comparación entre esquema actual y esquema propuesto

En primer lugar se muestra la comparación realizada con el esquema considerando terciles (Figura 5.36 y Figura 5.37)

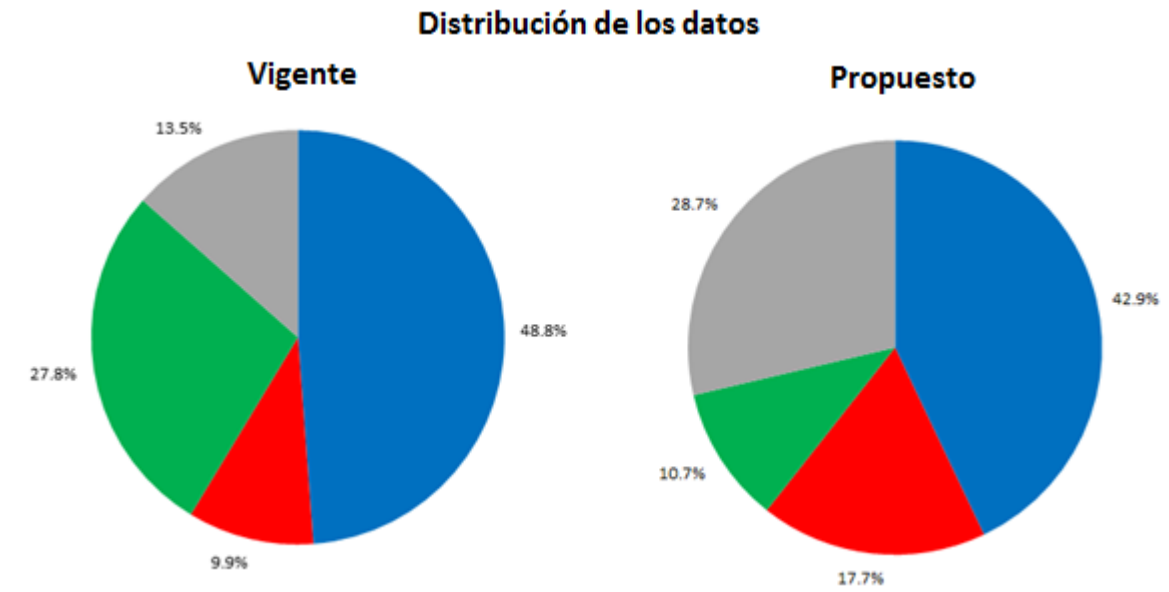


Figura 5.36: Distribución de la cantidad de datos en cada zona de exigencia en el esquema actual y en el esquema propuesto por terciles

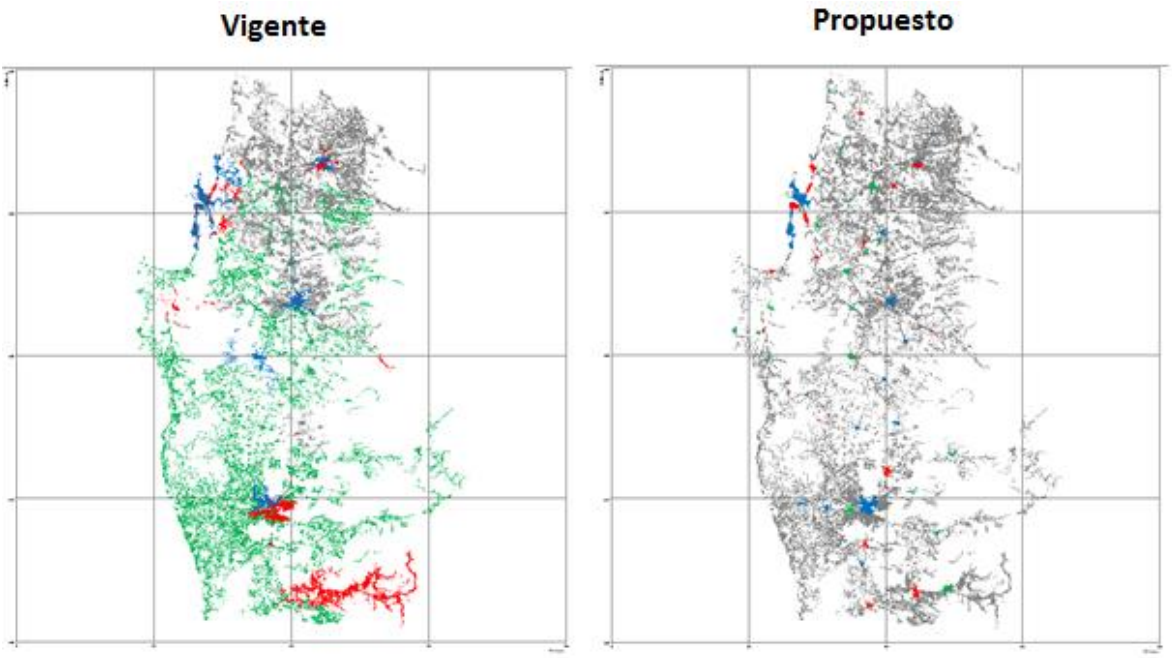


Figura 5.37: Disposición geográfica de los consumos en el esquema actual y en el esquema propuesto por terciles.

Se presenta además una comparación del esquema en el cual se consideró al ruido (subclusters 0) como la zona de menor exigencia y luego los subcluster restantes fueron divididos en una mitad superior y una mitad inferior.

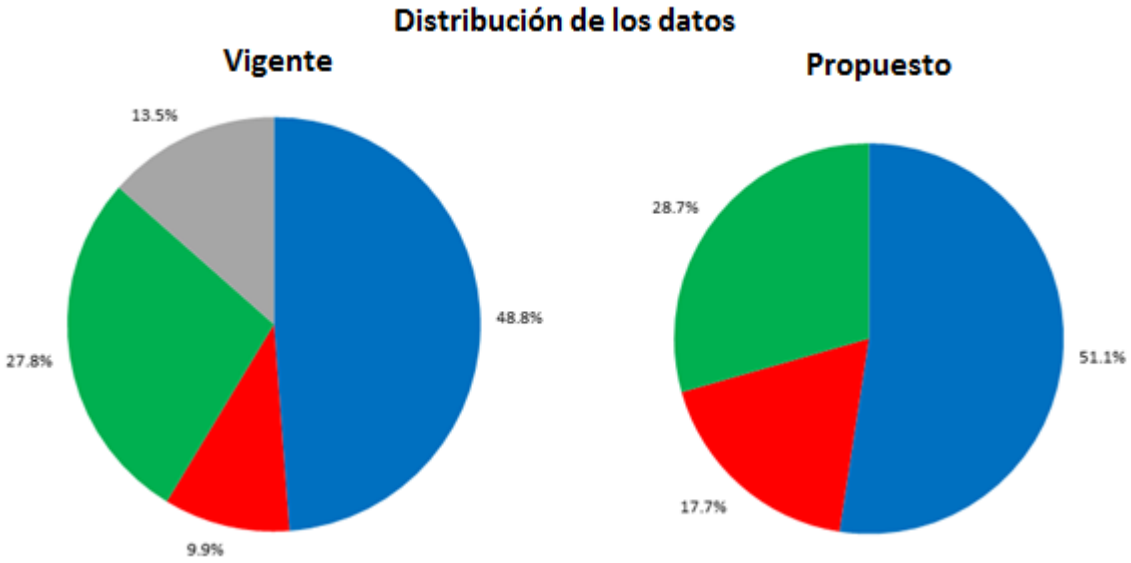


Figura 5.38: Distribución de la cantidad de datos en cada zona de exigencia en el esquema actual y en el esquema propuesto por mitades y ruido como zona de menor exigencia

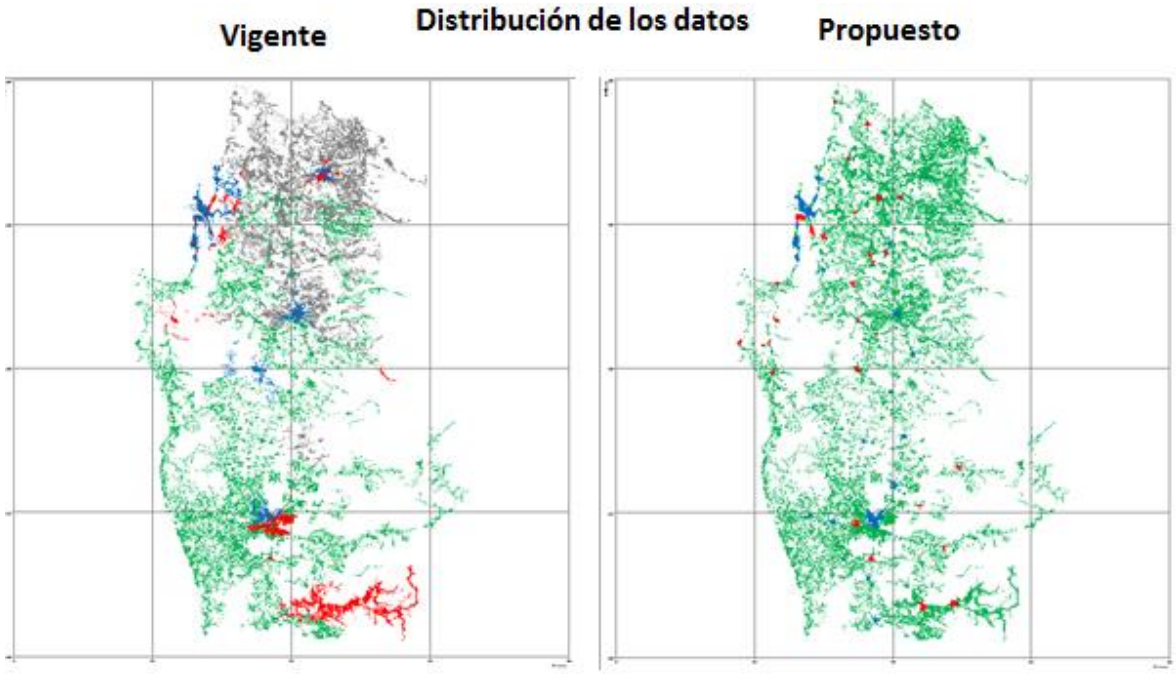


Figura 5.39: Disposición geográfica de los consumos en el esquema actual y en el esquema propuesto por mitades y ruido como zona de menor exigencia

Si se quiere seguir este esquema, se deben seguir ajustando los parámetros para que las soluciones encontradas por DBSCAN abarquen más territorio.

Más que un aporte fue un problema haber sacado la componente de consumo en los cálculos de clustering a través de DBSCAN, por lo que los resultados que se presentarán a continuación corresponde a rehacer los resultados 5, es decir, tres etapas, la primera K-means en Weka con K=6, cálculo de épsilon con 3 componentes y DBSCAN basado en 3 componentes.

5.3.6. Resultados 6

Estos resultados son similares a los anteriores, pero está vez se decidió considerar los atributos tanto geográficos como los atributos de consumo.

Esta vez los resultados serán analizados en detalle ya que en base a estos es como se definió la metodología final.

En una primera etapa, se separaron los datos utilizando el algoritmo K-means del programa Weka en 6 clusters. Se elige el número 6 por la cantidad de provincias que existe en estas dos regiones, con el fin de que los centroides queden cerca de las capitales provinciales, lugares que se asumen como intensivos en consumo. Para agrupar se utilizaron criterios geográficos y de consumo, es decir, se consideró como métrica la distancia euclidiana entre los puntos considerando los atributos “x”,”y” y el consumo anual de energía medido kWh.

En la Figura 5.40 se puede ver la distribución de la cantidad de datos en cada cluster, y en la Figura 5.41 se visualiza la disposición física de los consumos de cada uno de los clusters.

En la Figura 5.42 se ve cómo quedan distribuidos especialmente los consumos de las capitales provinciales, donde se observa que la mayoría de ellas pertenecen casi en su totalidad a un solo cluster a excepción de Angol cuyos consumos pertenecen al cluster 3 (morado) y cluster 5 (naranjo)

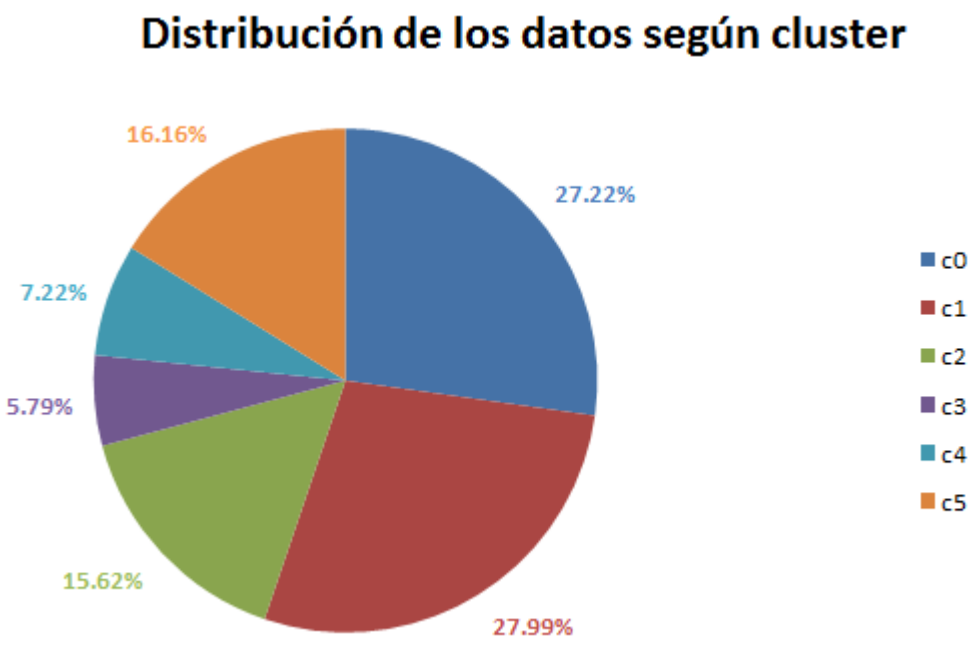


Figura 5.40: Distribución de la cantidad de datos por clusters para la solución k-means, k=6, considerando parámetros geográficos y de consumo.

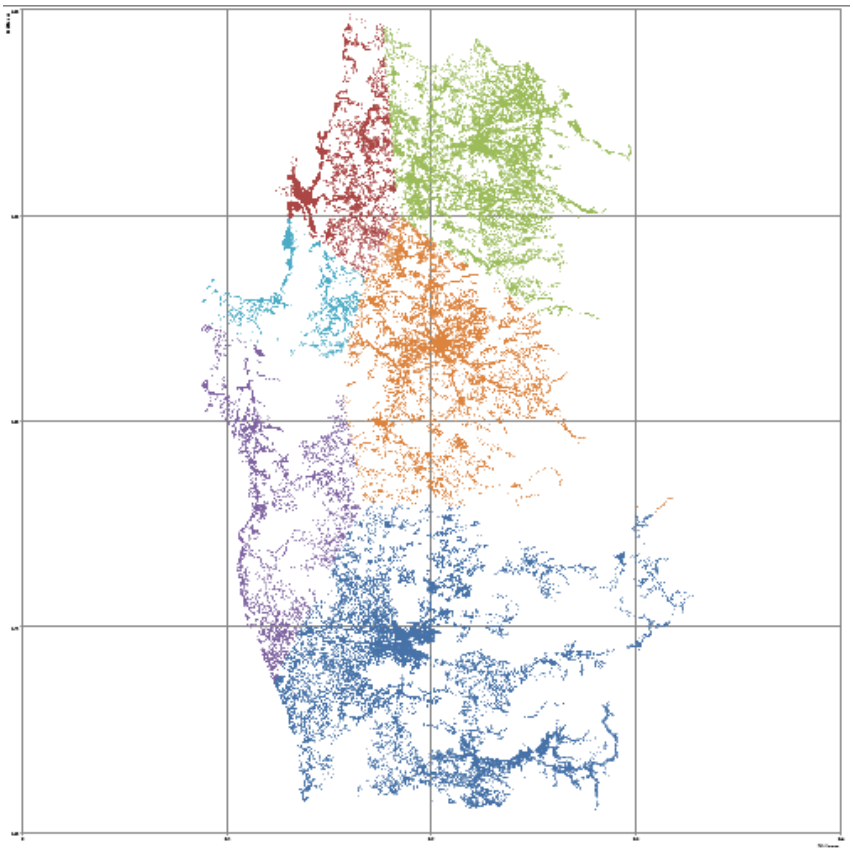


Figura 5.41: Disposición geográfica de los consumos con la representación de su área equivalente para la solución k-means, k=6, considerando parámetros geográficos y de consumo.

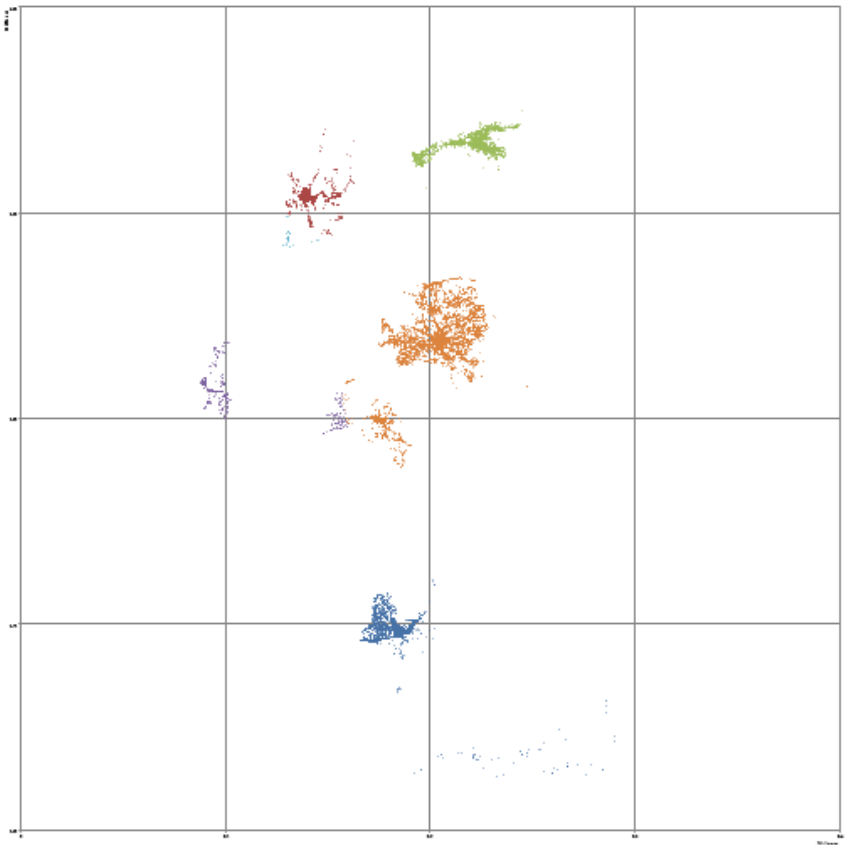


Figura 5.42: Disposición geográfica de los consumos en las capitales provinciales.

Pensando en utilizar DBSCAN, en una segunda etapa se calcularon las distancias a los vecinos más cercanos para los 6 clusters utilizando un método programado en MATLAB, para el cual también se utiliza la distancia euclidiana basada en los atributos geográficos y de consumo, los resultados se aprecian en la Figura 5.43 y la Tabla 5.12.

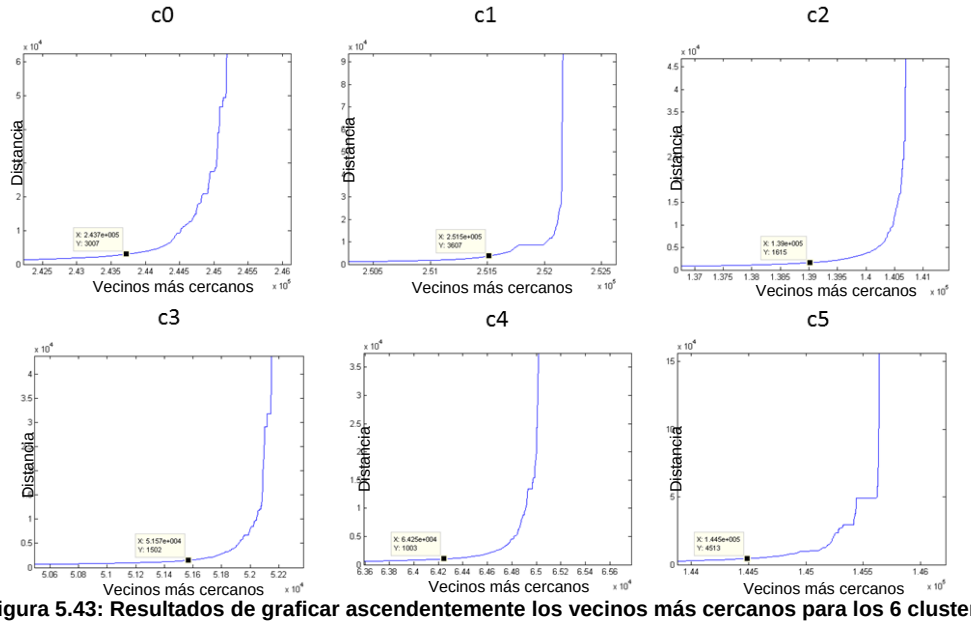


Figura 5.43: Resultados de graficar ascendentemente los vecinos más cercanos para los 6 clusters

Cluster	Épsilon
c0	3.000
c1	3.600
c2	1.600
c3	1.500
c4	1.000
c5	4.500

Tabla 5.12: Valor de épsilon considerado para cada cluster

Posteriormente, en una tercera etapa, se utilizó DBSCAN con el software RapidMiner, tomando como parámetro épsilon los encontrados de la etapa anterior (Tabla 5.12) y como parámetro MinPts el 1% de la cantidad de elementos que posee el cluster. Esto se realizó sobre los 6 clusters obtenidos en la primera etapa, nuevamente considerando los atributos X, Y y kWh con lo que se obtiene un nuevo agrupamiento, para cada cluster, cuyos elementos son llamados subcluster.

Los resultados de esta etapa se aprecian en la Figura 5.44, Figura 5.45, Figura 5.46, Figura 5.47, Figura 5.48 y Figura 5.49.

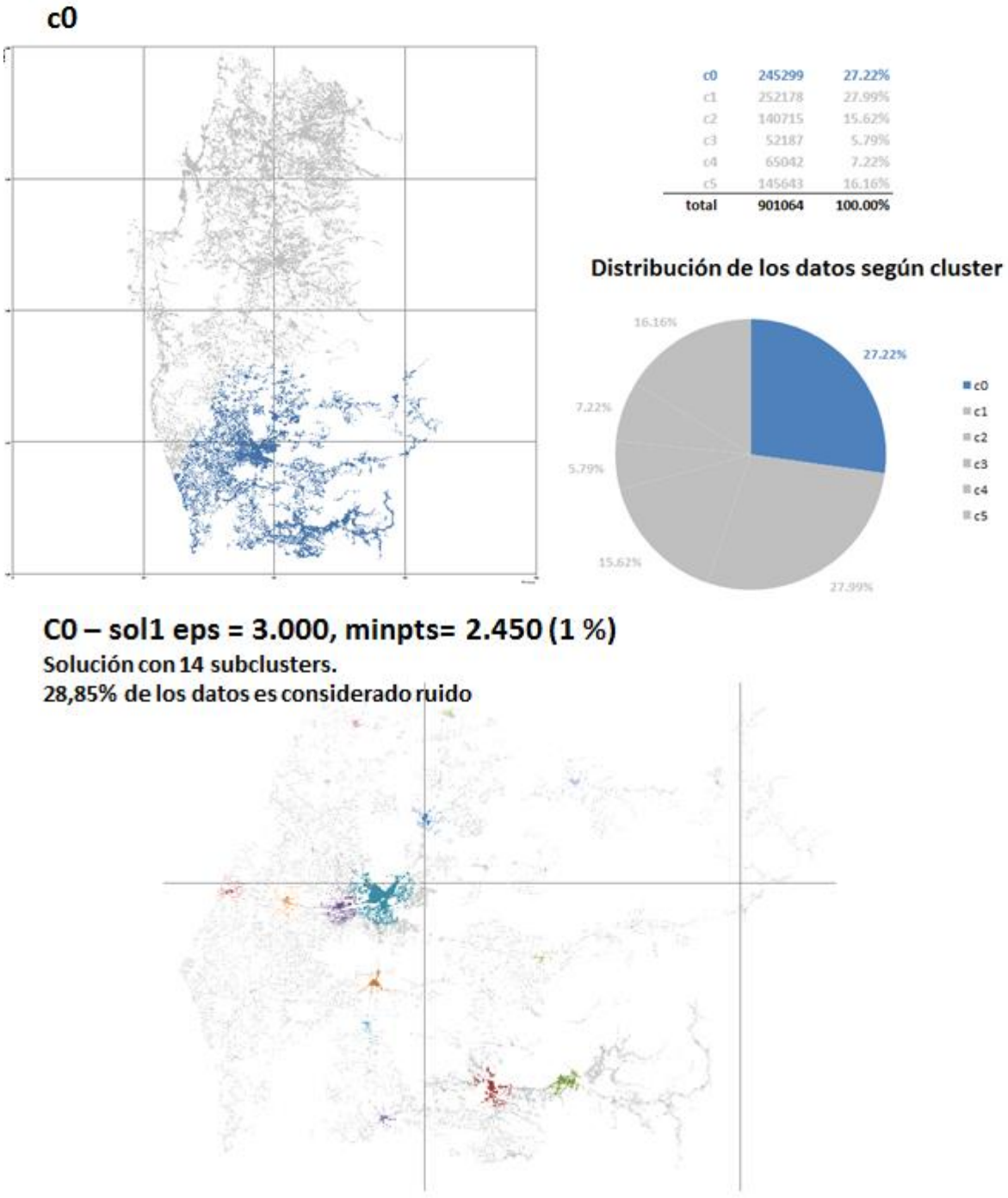


Figura 5.44: Resultados DBSCAN (subclustering por densidad) cluster c0.

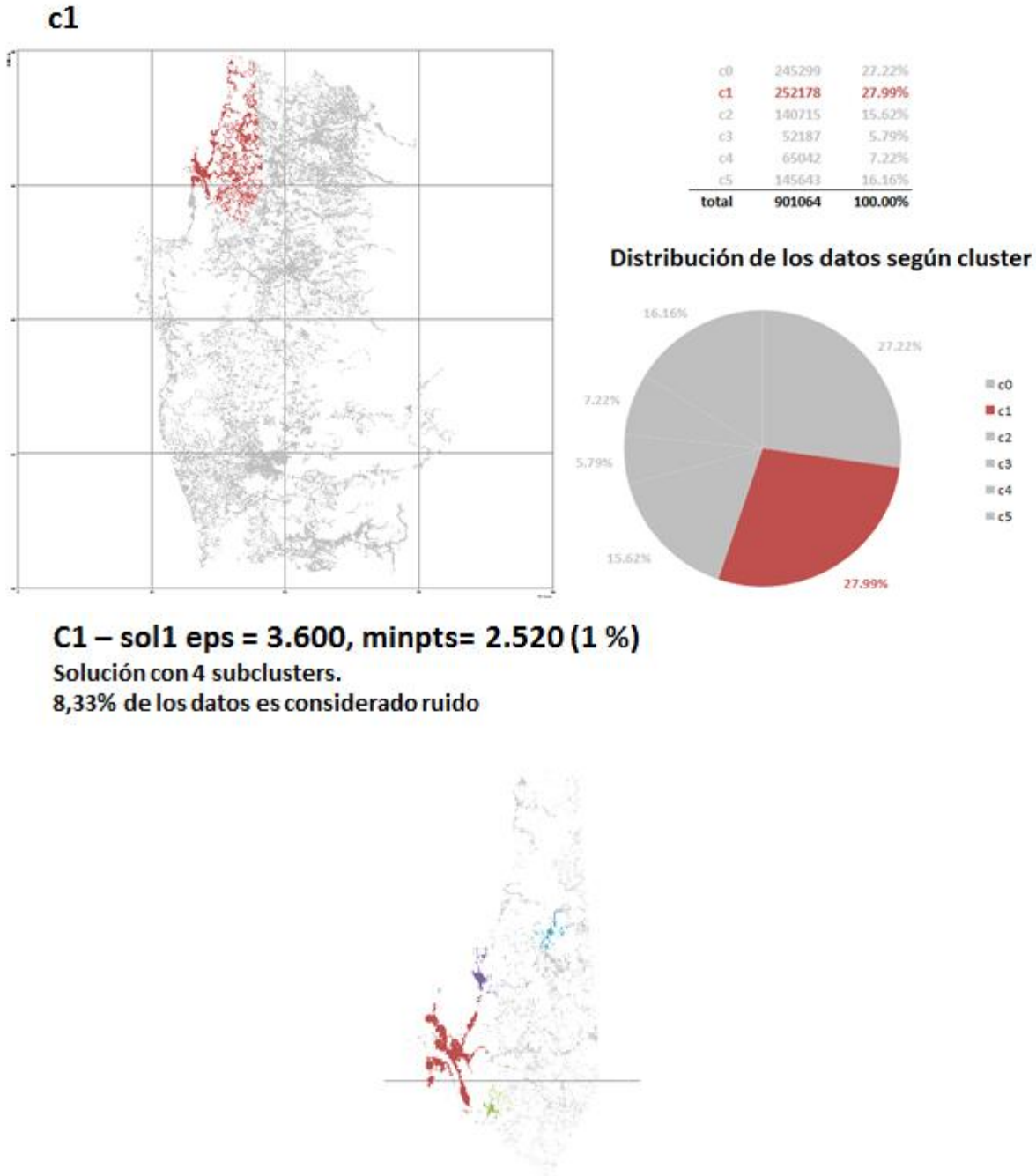
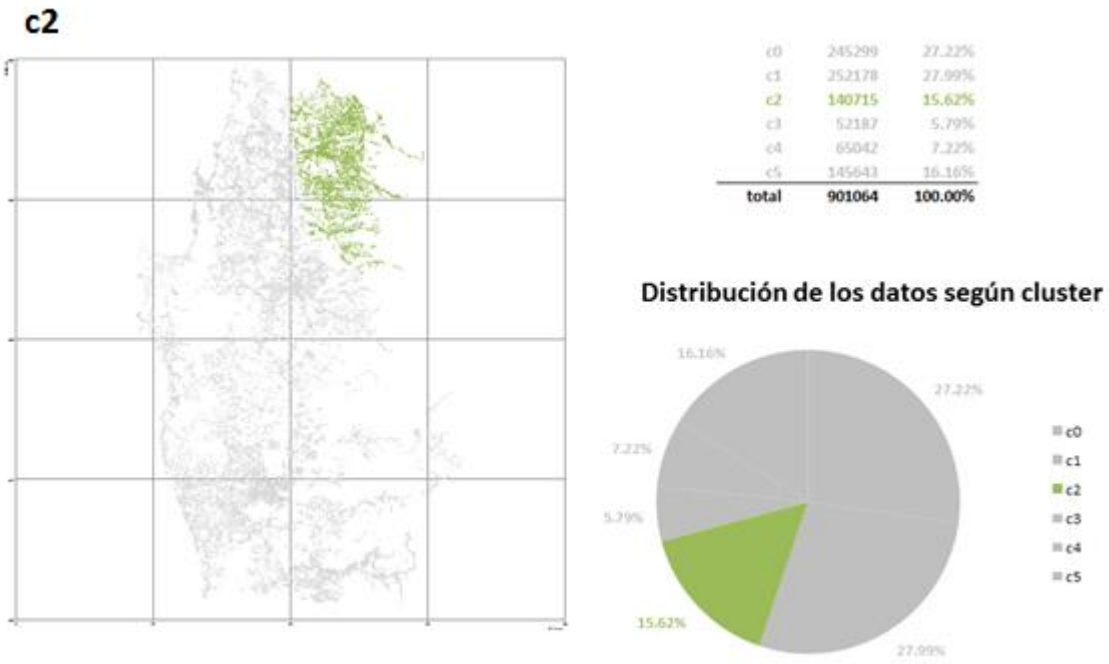


Figura 5.45: Resultados DBSCAN (subclustering por densidad) cluster c1.



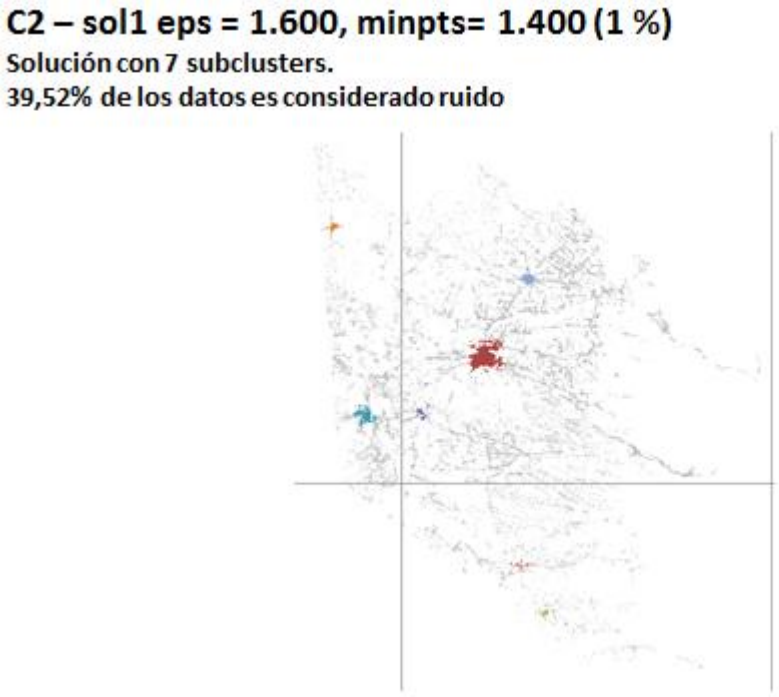


Figura 5.46: Resultados DBSCAN (subclustering por densidad) cluster c2.

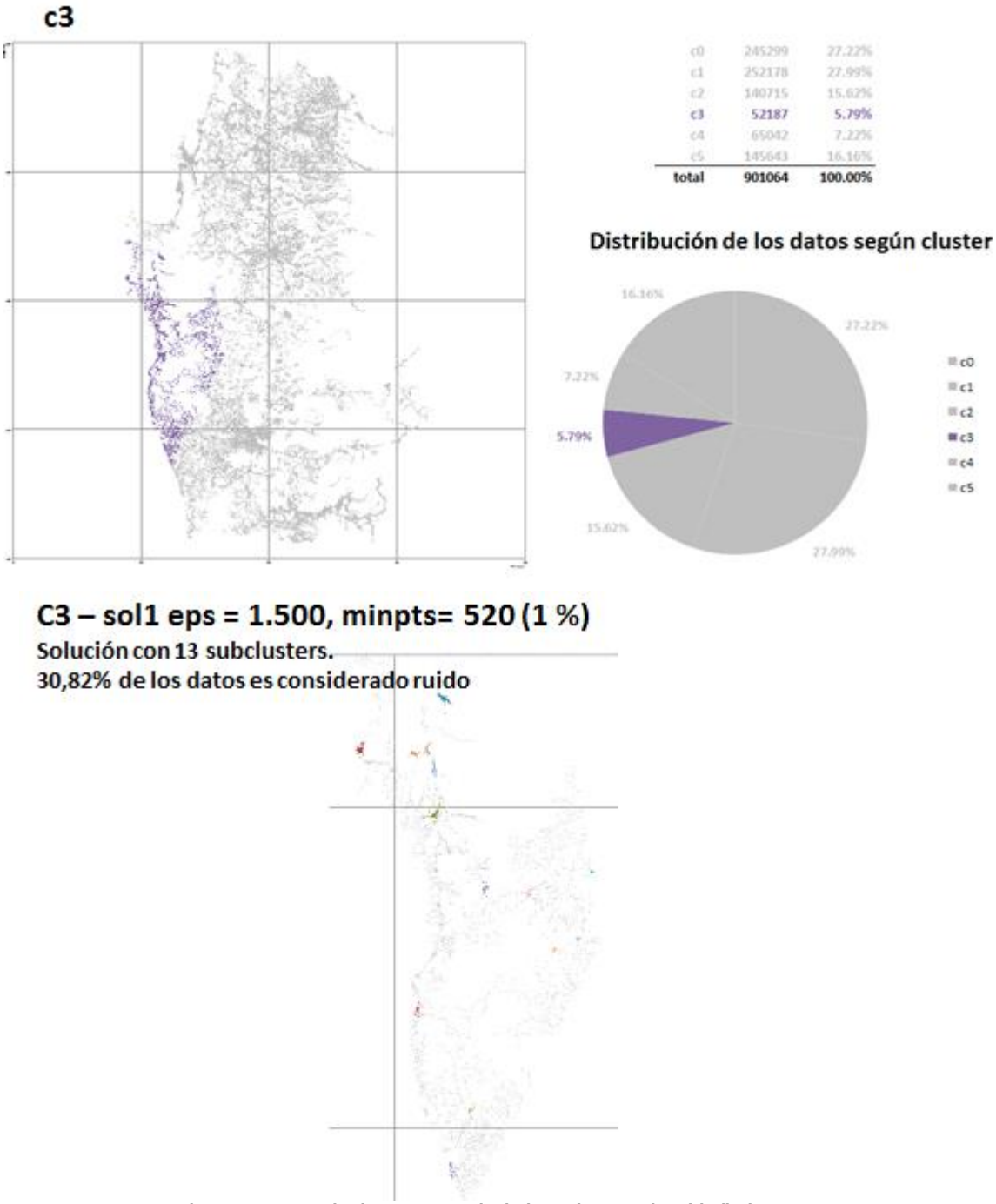


Figura 5.47: Resultados DBSCAN (subclustering por densidad) cluster c3.

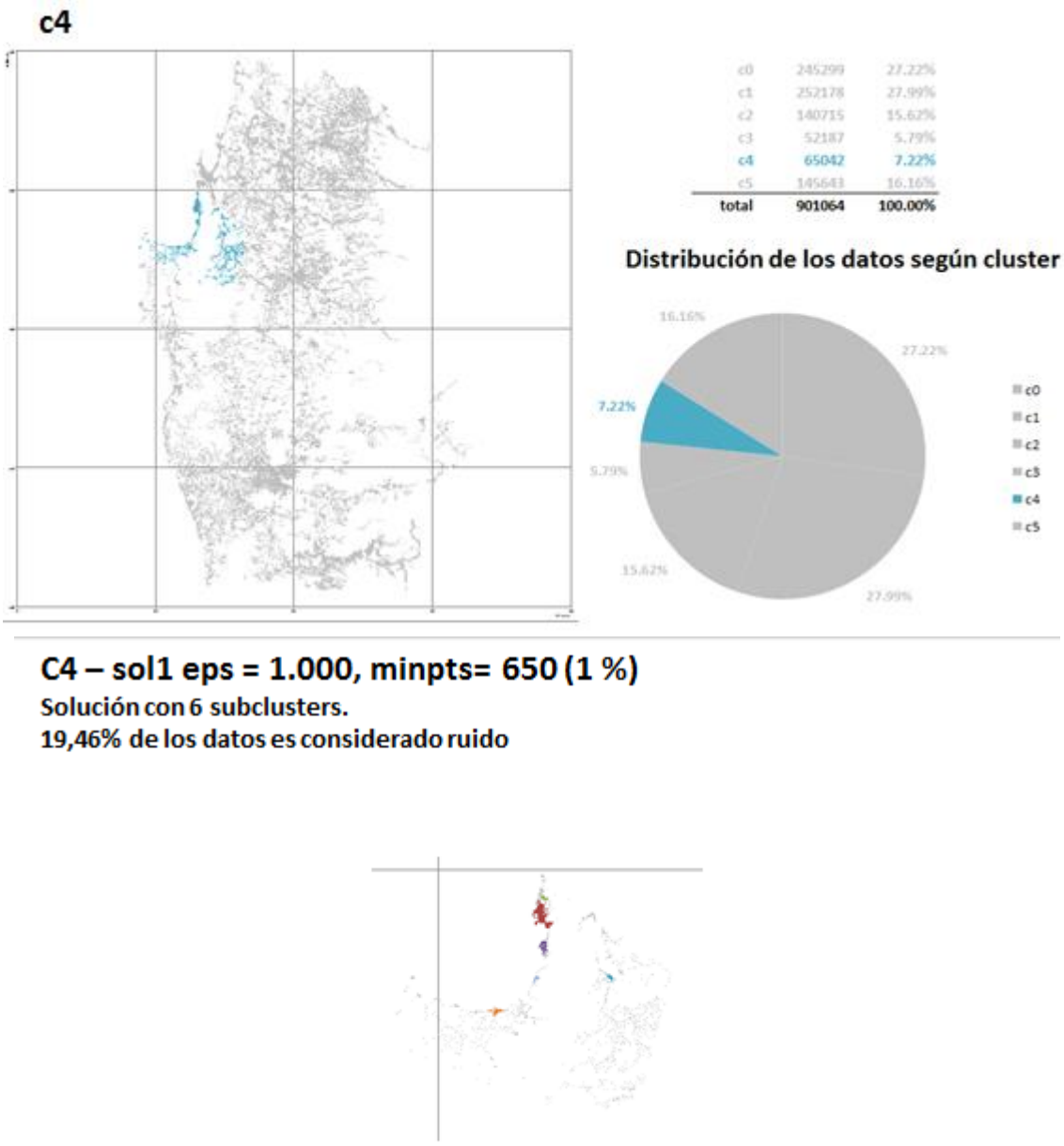


Figura 5.48: Resultados DBSCAN (subclustering por densidad) cluster c4.

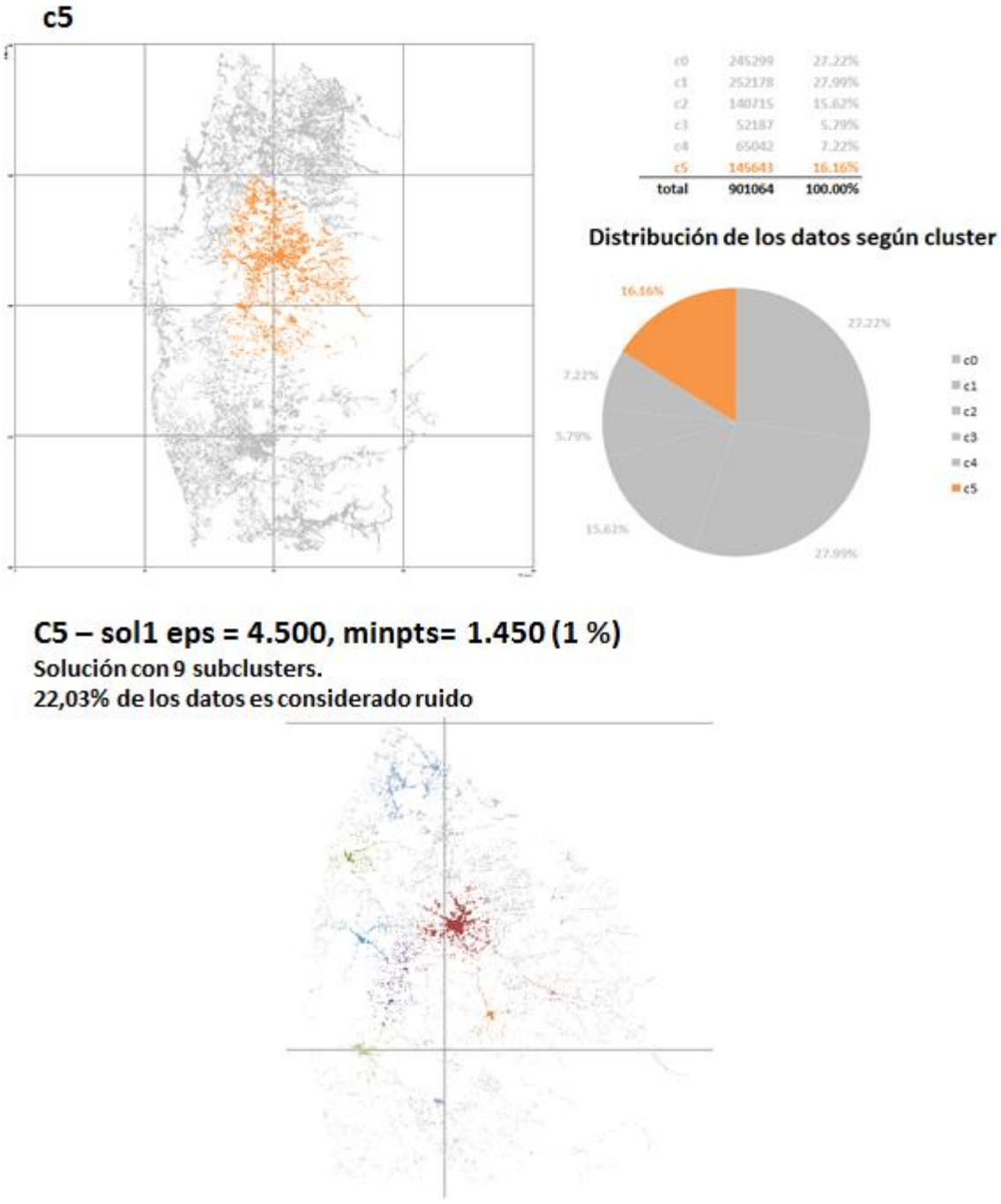


Figura 5.49: Resultados DBSCAN (subclustering por densidad) cluster c5.

Una vez finalizada la etapa de clustering se procesaron los datos obtenidos para calcular los índices de densidad de consumo y la separación en distintas zonas de exigencia.

El resultado de calcular el índice y graficarlos decrecientemente se observa en la Figura 5.50.

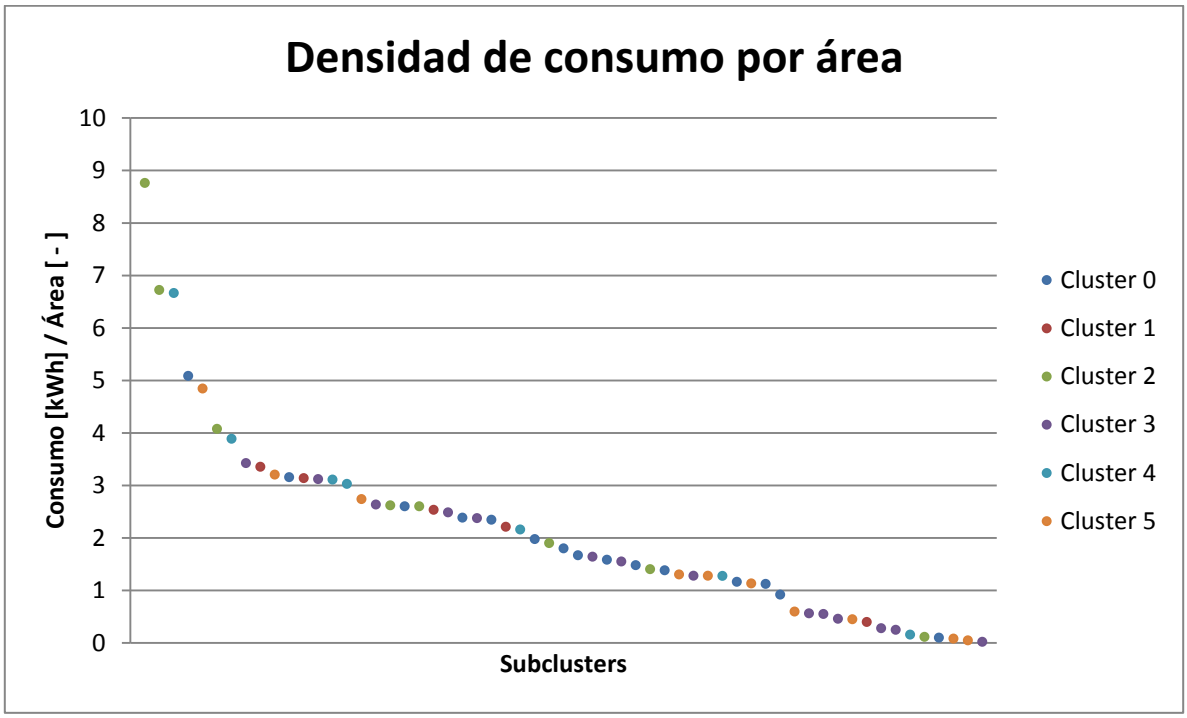


Figura 5.50: Densidad de consumo por área.

Se emplearon 2 esquemas para encontrar a qué zona de exigencia debe pertenecer cada subcluster, uno basado en terciles como el descrito en la sección del Resultados 5 y otro basado en las pendientes observadas en el gráfico de densidad de consumo por área. Se optó por este último dado que en este enfoque se agrupan los subclusters que tienen una misma tendencia, en vez de ser agrupados por la fuerza para que cada zona de exigencia tenga la misma cantidad de subclusters. Se calcularon regresiones lineales con los subcluster que corresponden a las zonas de exigencias. La manera de encontrar si un subcluster pertenece o no una zona de exigencia se determinó cuando al agregar el siguiente subcluster el coeficiente R^2 de la regresión lineal disminuye. En ese caso este subcluster pasa a ser el primero de la siguiente zona de exigencia. El resultado de la agrupación por pendientes se muestra en la Figura 5.51 y el detalle de las regresiones en la Tabla 5.13.

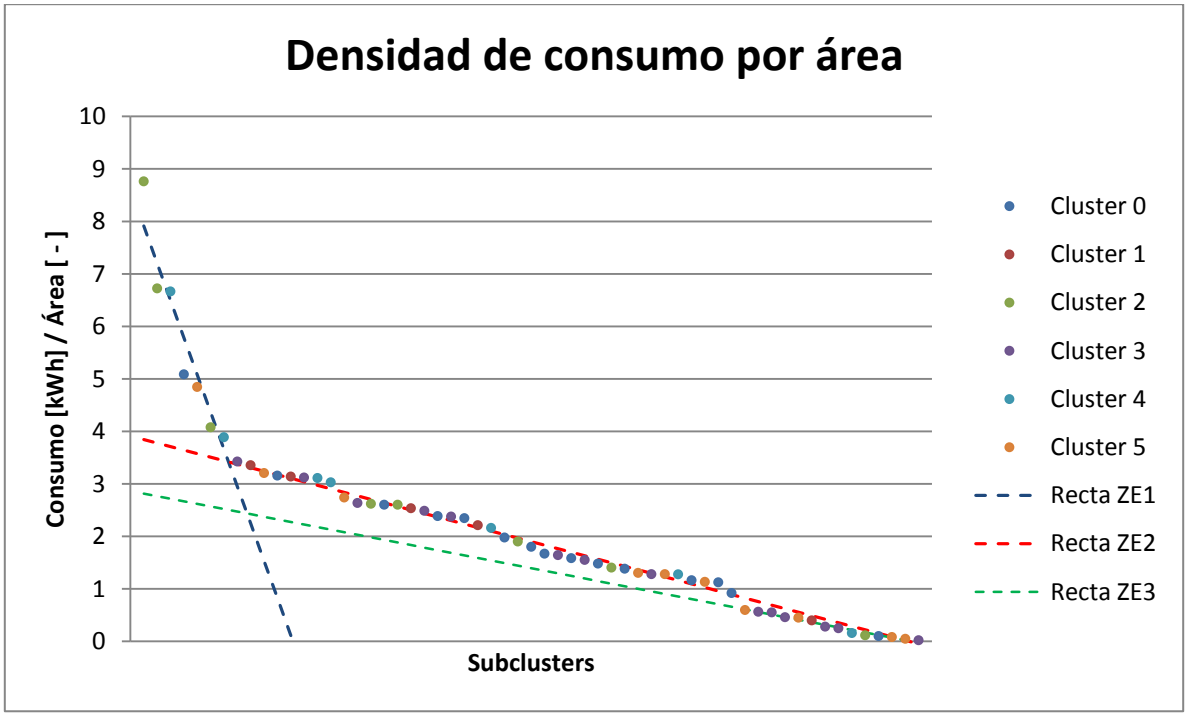


Figura 5.51: Densidad de consumo por área con rectas de zonas de exigencia.

Zona de Exigencia	Coeficiente libre	Coeficiente Variable X	R^2 del ajuste
ZE1	8.623	-0,71	0,904
ZE2	3.911	-0,07	0,986
ZE3	2.864	-0,05	0,970

Tabla 5.13: Regresión lineal para rectas de zonas de exigencia

Posteriormente se compararon los resultados obtenidos con el esquema de zonas de exigencia de calidad de suministro existente.

Siguiendo con el esquema de colores expuesto en la Figura 5.35, se muestra una distribución de la cantidad de datos por zona de exigencia. Se observa que la cantidad de consumos considerados como los de mayor exigencia (urbano en el esquema actual) baja considerablemente de un 49,9% a un 29% y los considerados como exigencia media (rural tipo 1) aumentan 45%.

La Figura 5.53 muestra la distribución geográfica de los consumos distinguiendo las tres zonas de exigencias con sus respectivos colores.

Se aprecia, al comparar, que existen lugares en los cuales el nivel de exigencia se mantiene (marcados con naranja), otros en los cuales aumenta (marcados con morado) y otro en los cuales disminuye (marcados con celeste)

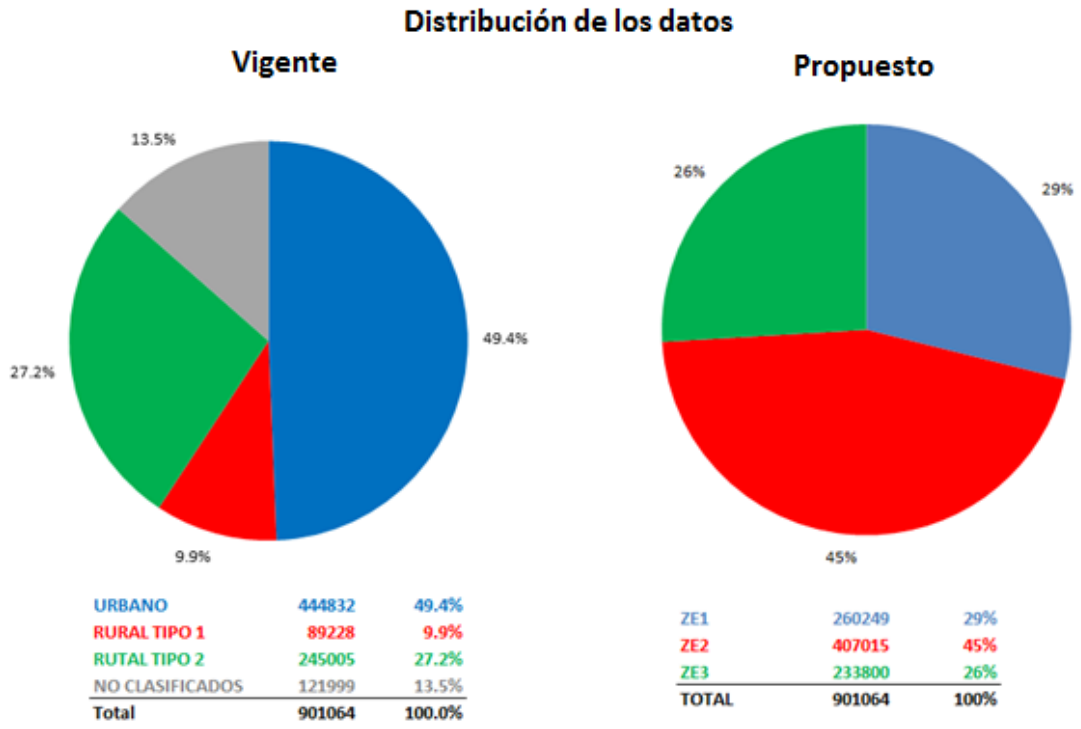


Figura 5.52: Distribución de la cantidad de datos en cada zona de exigencia en el esquema actual y en el esquema propuesto por pendientes.

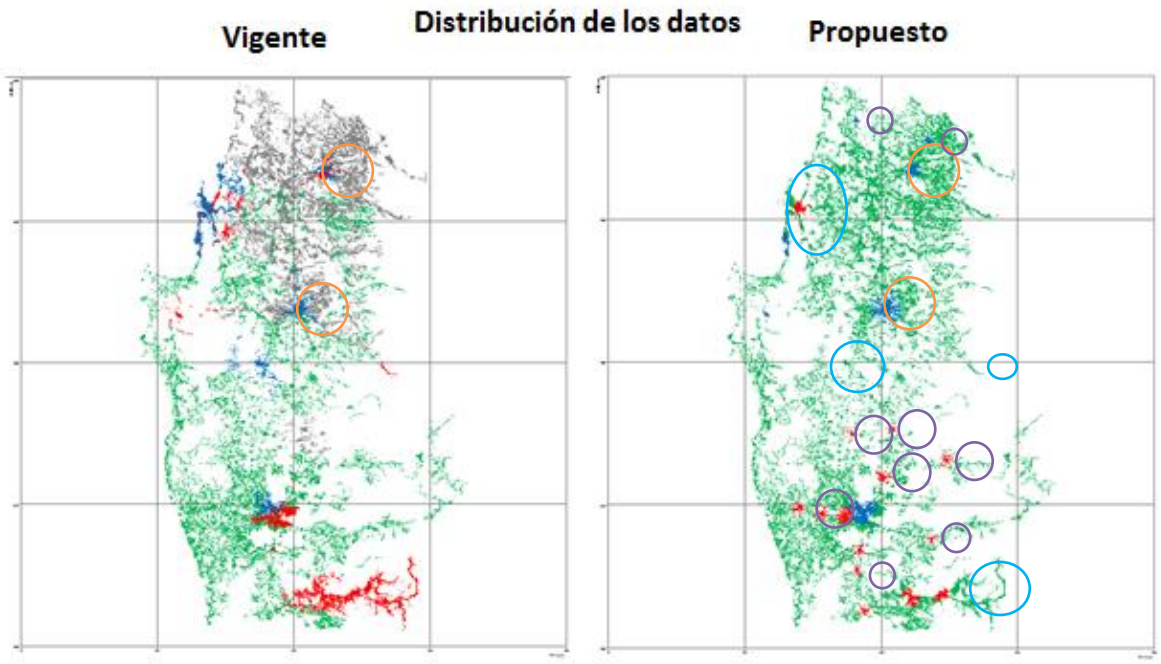


Figura 5.53: Disposición geográfica de los consumos en el esquema actual y en el esquema propuesto por pendientes.

Ciudades como Los Ángeles y Chillán representan el primer caso que conservan su alto nivel de exigencia en calidad de suministro bajo el nuevo esquema basado en densidad de consumo y parámetros geográficos, además, la mayoría de los lugares catalogados como rurales tipo 2 conservan su bajo nivel de exigencia en el esquema propuesto.

En la novena región, se registran varios lugares que en el nuevo esquema deberían tener mayor exigencia que la que tienen consignada actualmente y en la octava también se observan 2 lugares que caen en el mismo caso.

Por otro lado, los resultados muestran que para ciudades como Angol y el Gran Concepción (Concepción y alrededores), el nivel de exigencia requerido no es tan alto como el que poseen actualmente. Recordemos que Angol fue separado en dos clusters en la primera etapa con K-means, lo que puede afectar en el momento de la clasificación de los subclusters.

Finalmente se analizarán 3 casos particulares, correspondientes a la zona de Temuco – Padre las Casas, Villarrica – Pucón y el Gran Concepción.

Temuco – Padre las Casas:

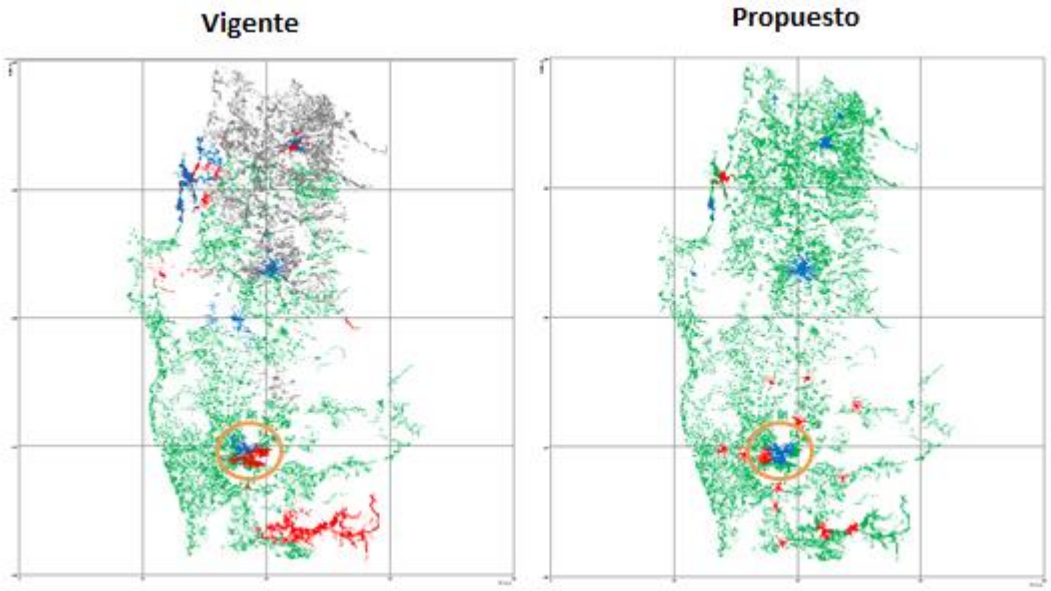


Figura 5.54: Comparación de la zona Temuco - Padre las Casas con el esquema vigente y propuesto, vista general

En la Figura 5.55 se ve gráficamente como es el esquema vigente para las exigencia en calidad de suministro el cual está separado por comunas, para Temuco se considera que los consumos de la concesionaria de distribución CGE son de tipo urbano y los de Frontel son de exigencia rural tipo 2, mientras que para Padre las Casas, los consumos CGE se consideran rural tipo 1 y los consumos Frontel rural tipo 2.

En el nuevo esquema (Figura 5.56), el cual no considera límites comunales ni empresas concesionarias del servicio de distribución, se observa que la clasificación está basada en densidad, siendo los puntos del centro considerados como aquellos de exigencia más altas, otra agrupación importante que se observa a la izquierda de la imagen corresponde a un tipo de exigencia media, y los puntos hacia la periferia se consideran con una menor exigencia, lo cual cumple los objetivos buscados.

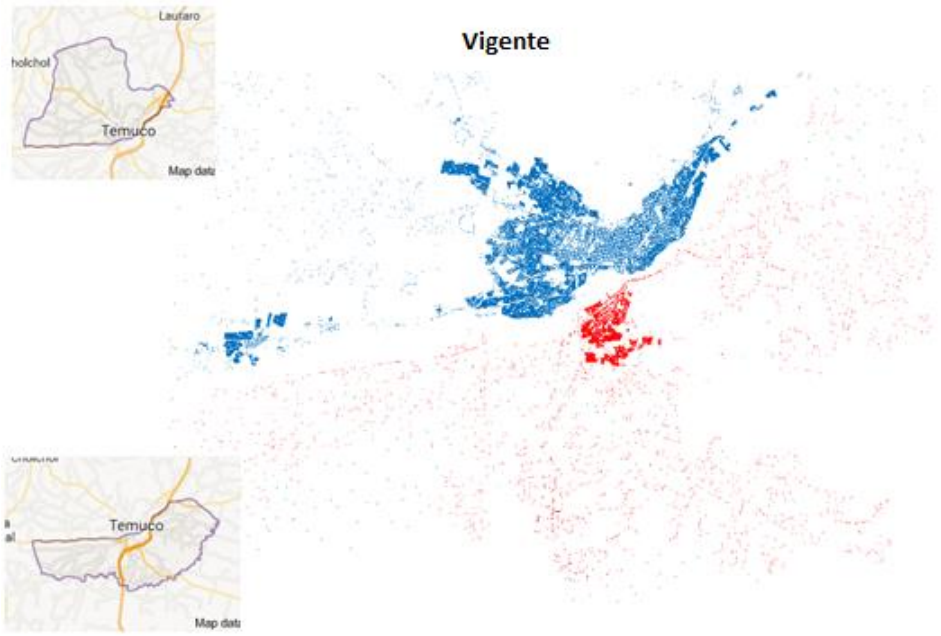


Figura 5.55: Comparación zona Temuco - Padre las Casas: esquema vigente

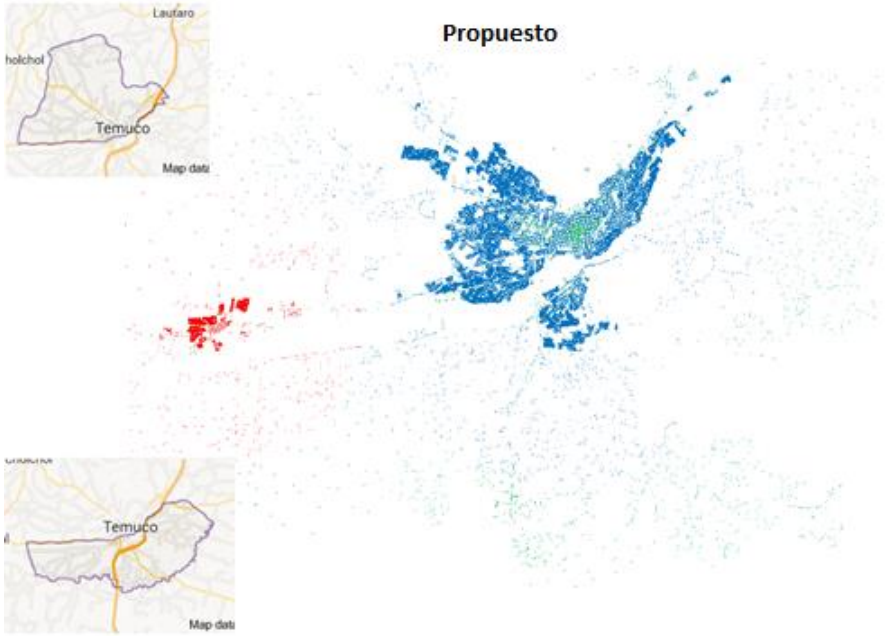


Figura 5.56: Comparación zona Temuco - Padre las Casas: esquema propuesto

Villarrica – Pucón:

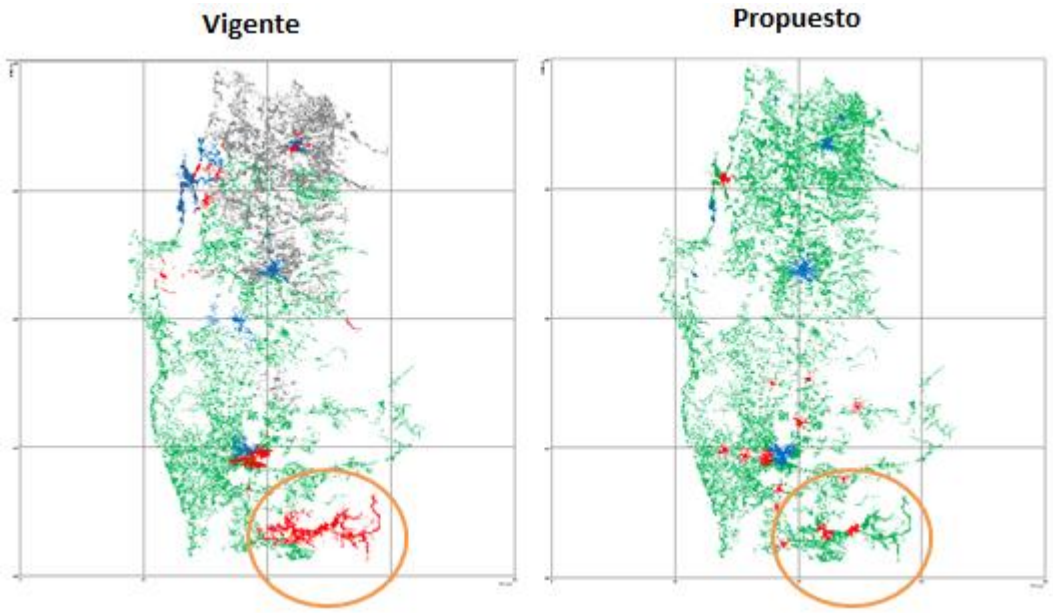


Figura 5.57: Comparación de la zona Villarrica - Pucón con el esquema vigente y propuesto, vista general

La Figura 5.58 muestra el esquema vigente para la zona de Villarrica y Pucón, donde se observa que prácticamente todos los consumos son clasificados como zona rural tipo 1, en el esquema propuesto (Figura 5.59) se ve que solamente las zonas más densas de cada ciudad son clasificadas como zonas de exigencia media, en cuanto los consumos que son menos densos son clasificados en zonas de exigencia baja, es importante notar que los consumos que se encuentran más hacia la cordillera (derecha de la figura), que probablemente corresponden a la cola del alimentador, no necesitan de una calidad tan elevada como si se necesita en las ciudades, lo cual puede traer implicancias en la planificación de redes, en las inversiones asociadas a subir el nivel de calidad y finalmente en la tarifa vista por el usuario final.

Cabe destacar que el método de clasificación funcionó bien a pesar del accidente geográfico del Lago Villarrica, el cuál no repercutió a la hora de que el algoritmo encontrara zonas de consumo más densas y menos densas.

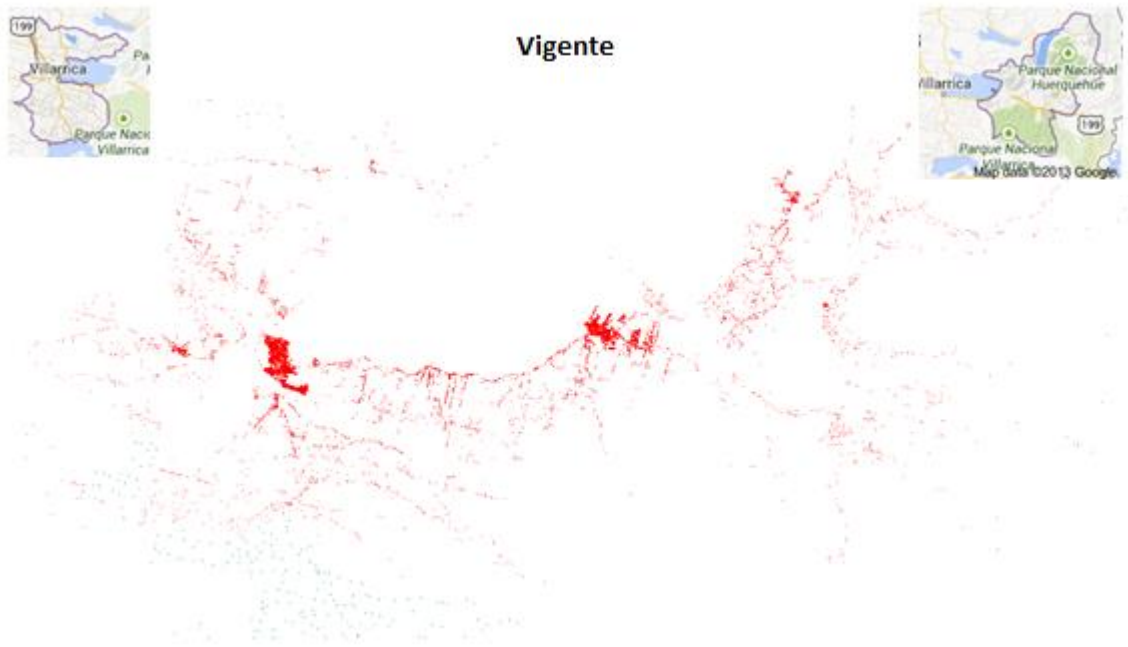


Figura 5.58: Comparación zona Villarrica - Pucón: esquema vigente

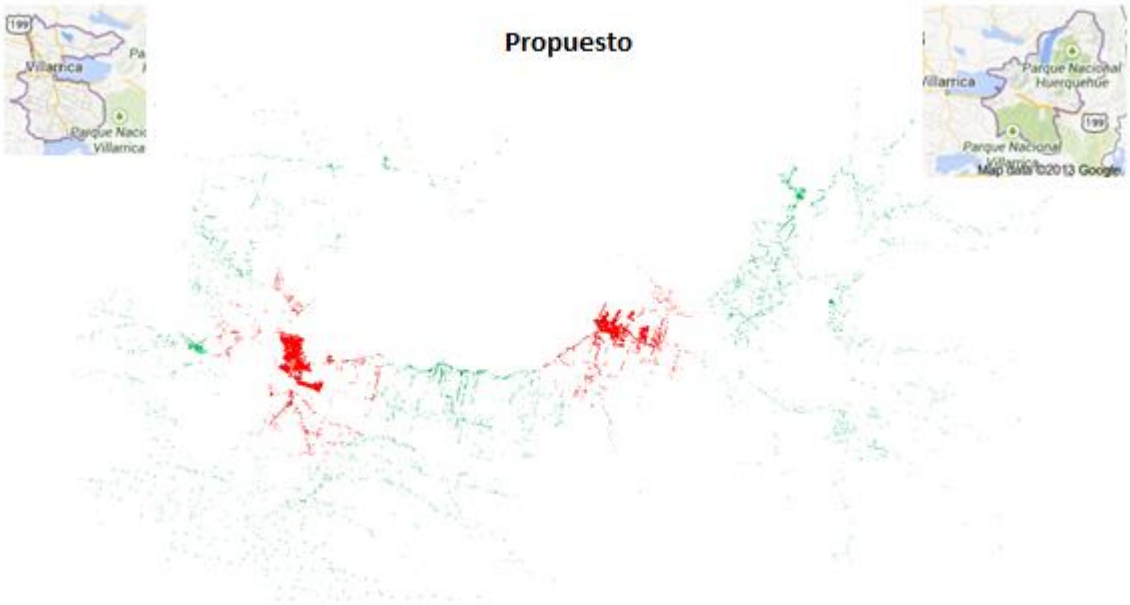


Figura 5.59: Comparación zona Villarrica - Pucón: esquema propuesto.

Gran Concepción:

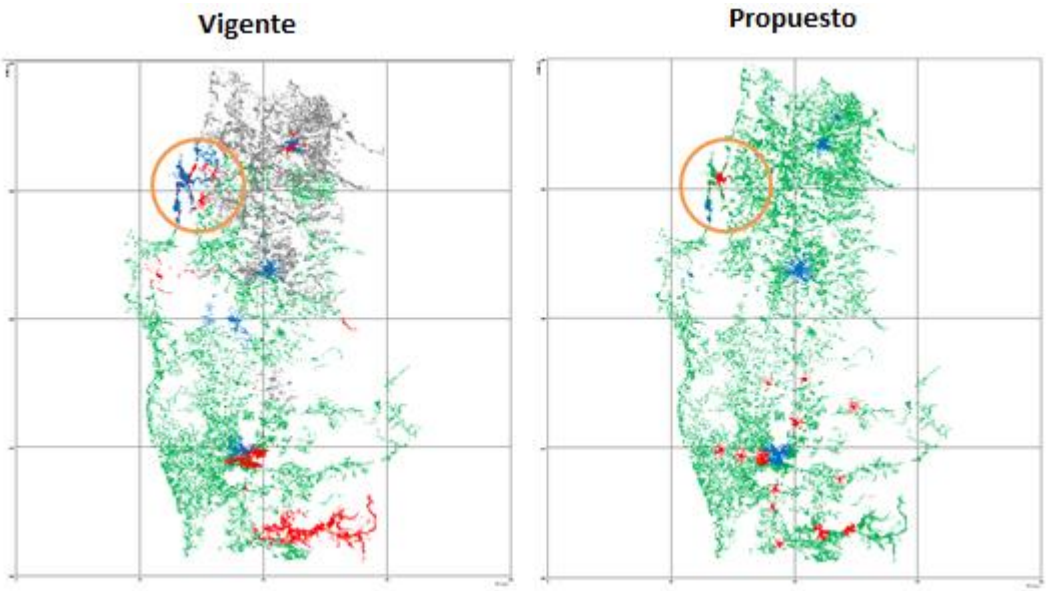


Figura 5.60: Comparación del Gran Concepción con el esquema vigente y propuesto, vista general

La Figura 5.61 muestra el esquema vigente para la zona del Gran Concepción, se observa que la gran mayoría de las comunas (y consumos) están clasificadas como zona de exigencia urbana, la más exigente de todas, a excepción de las comunas de Penco y Hualqui. La Figura 5.61 muestra que en el esquema propuesto, la mayoría de las comunas baja a un nivel de exigencia medio. Este resultado resulta raro a priori, algunas de las cosas que se deben considerar cuando se analizan estos resultados es que no siempre la visualización permite observar bien lo que está pasando, por ejemplo podría pasar que los consumos que se ven igualmente densos al ser graficados con otra resolución no lo sean.

El hecho de estar trabajando con áreas promedio también puede afectar, considerar los consumos más alejados y pequeños puede estar repercutiendo en que se aumente el área del subcluster (denominador del índice) y aumentando muy poco el consumo neto del subcluster (numerador)

Vale la pena mencionar que el subcluster correspondiente a Concepción y alrededores es el que tiene más alto índice de consumo por área dentro de los subcluster clasificados como zona de exigencia 2 (exigencia media). Esto muestra la importancia de revisar la asignación de zonas continuamente, por ejemplo con ocasión del cálculo del valor agregado de distribución, ya que en 4 años las redes se van desarrollando y los consumos van creciendo a distintas tasas, dependiendo de los sectores, por lo que es bastante probable que si se consideran datos de años posteriores este subcluster termine por pasar a ser de alta exigencia.

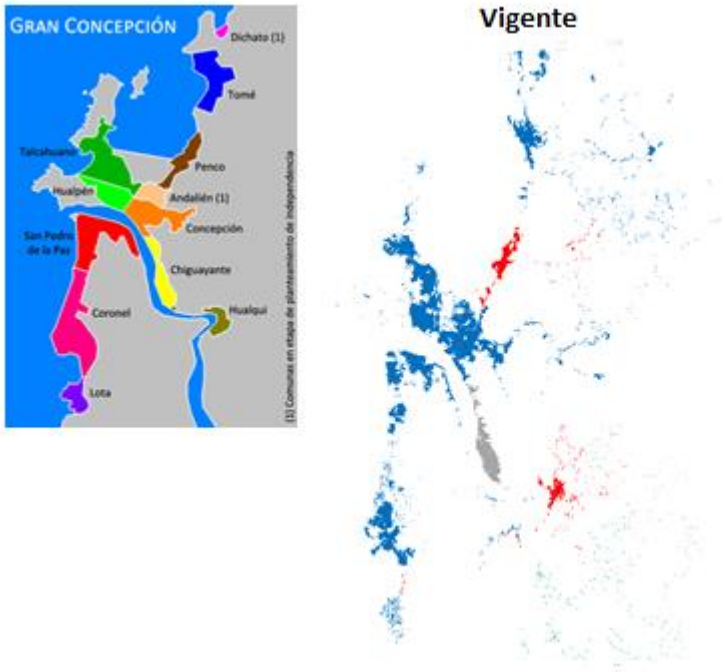


Figura 5.61: Comparación zona Gran Concepción: esquema vigente

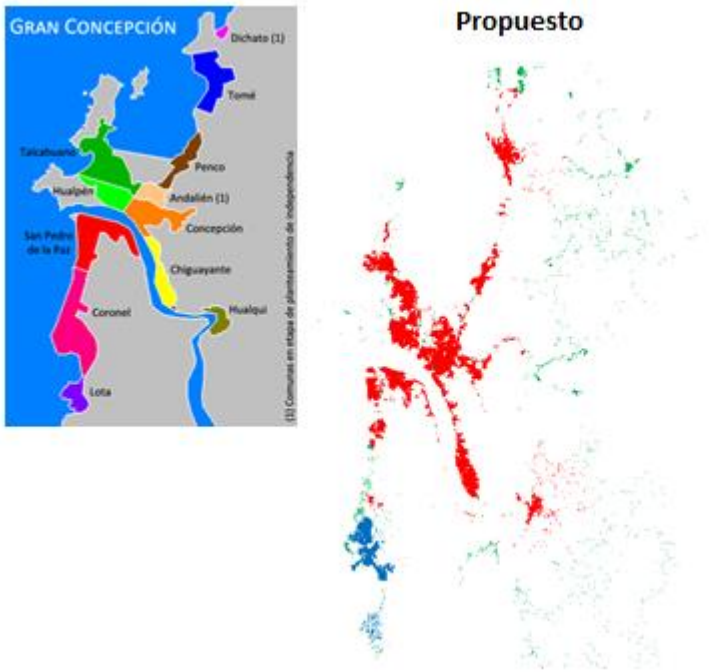


Figura 5.62: Comparación zona Gran Concepción: esquema propuesto

5.4. Metodología Propuesta

En esta sección se sintetiza la metodología propuesta para encontrar zonas de exigencia de calidad de suministro basada meramente en criterios geográficos y de calidad de suministro.

Esta metodología se basa en los resultados expuestos en la sección anterior, la cual fue validada a través de la obtención de los últimos resultados (Resultados 6) que fueron satisfactorios pero perfectibles. Experimentando aún más se puede encontrar algoritmos que se desempeñen mejor y lograr un mejor ajuste de los parámetros de estos algoritmos.

La metodología se basa en tres etapas principales: la etapa de pre-procesamiento, la etapa de clustering y la etapa de post-procesamiento.

La etapa de pre-procesamiento se ocupa de dejar los datos en un formato manejable, luego de aplicar el filtrado de datos en bruto a través de dos filtros y la posterior visualización y estadísticas de los datos a considerar.

Se monta una base de datos con los archivos de texto en los cuales viene la información, luego desde esta base de datos se exportan a formato Excel o los formatos de entrada que necesiten los programas.

Filtro geográfico: Se eliminan todos los datos que estén etiquetados como pertenecientes a las regiones de estudio pero cuyas coordenadas se encuentran notoriamente fuera de rango.Para ello se define un mínimo y un máximo en las coordenadas X e Y y se elimina todo lo que esté fuera de este rango.

Filtro por consumo: Se eliminan todos los consumos debajo de cierto umbral, que en el caso de estudio fue definido en 60 kWh anuales, por considerarse el mínimo consumo de energía que puede tener un cliente. Con este filtro se eliminan también consumos negativos.

Posterior a esto, los datos quedan listos para ser procesados. Se puede obtener distinta información de estos datos que puede ser útil en el posterior procesamiento y comparación del esquema actual con el obtenido. Por ejemplo en el caso de estudio se graficó la distribución porcentual de la cantidad de datos por región, también se obtuvieron algunos estadísticos del consumo anual de energía como el mínimo, el máximo, el promedio, el percentil 95% y el percentil 99% y la gráfica del histograma de consumos. Podrían obtenerse datos para comunas o empresas concesionarias de distribución particulares si así se deseara.

La etapa de clustering a su vez se divide en subetapas.

La primera consiste en una separación de los consumos en K zonas, donde K se ha fijado como el número de provincias existentes en la zona a estudiar. Esta etapa es llevada a cabo a través del software Weka y el algoritmo simpleKmeans, considerando una distancia euclidiana basada en los atributos X,Y y consumo anual de los datos.

La segunda subetapa consiste en el cálculo de los vecinos más cercanos para los K clusters. Una vez obtenidos estos, se grafican de manera creciente y se escoge el parámetro épsilon del algoritmo DBSCAN en torno al codo que presenta esta curva. Para esta etapa se emplea un método programado en MATLAB. Para ello también una distancia euclidiana basada en los atributos X,Y y consumo anual de los datos.

La tercera subetapa consiste en la reclasificación de los consumos por criterios de densidad, la cual se efectúa sobre cada uno de los K clusters calculados anteriormente. Esto se lleva a cabo a través del algoritmo DBSCAN, considerando el parámetro épsilon obtenido de la etapa anterior y el parámetro MinPts como el 1% de los datos del cluster. Este se lleva a cabo a través del software RapidMiner.

La última etapa corresponde al post-procesamiento de los datos, en donde para cada subcluster obtenido se realiza el cálculo de un índice de consumo por área, definido como el cuociente entre el consumo neto del subcluster por el área equivalente.

El consumo neto del subcluster se calcula como la suma del consumo anual de cada uno de los componentes del subcluster.

El área equivalente de cada subcluster, se define como un círculo cuyo radio es la distancia promedio de los puntos al centroide del subcluster.

$$Área\ Equivalente_{\text{subcluster}} = \pi \cdot \overline{\text{distancia}}^2(\text{punto}_j, \text{Centroide}) \quad \forall \text{punto}_j \in \text{Subcluster}$$

Notar que en esta definición, la distancia se refiere netamente a la distancia geográfica, considerando los atributos “x” e “y” y dejando de lado el consumo, a diferencia de como ha sido en el resto de los procesos.

A modo de ejemplo, en la Figura 5.63 se muestra un subcluster de 6 puntos. A la derecha se muestran las distancias de cada uno de los puntos al centroide (punto F), la distancia promedio del conjunto y el área equivalente del subcluster.

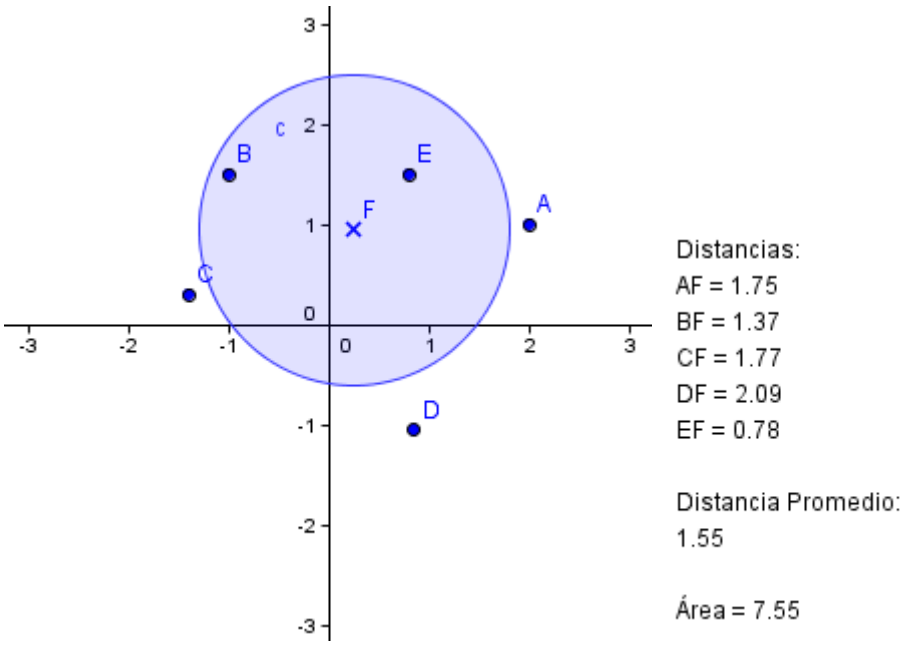


Figura 5.63: Cálculo del área equivalente de un subcluster

Para el cálculo del índice se implementaron métodos en lenguaje VBA sobre Microsoft Excel (macros).

Una vez obtenido los índices, son graficados decrecientemente y se calculan regresiones lineales de modo de dejar agrupado los subcluster similares. En este caso de estudio se encontraron 3 pendientes similares, pero podrían ser menos o más.

Para saber si un subcluster pertenece o no a una recta (y por ende a esa respectiva zona de exigencia), se busca el conjunto que maximice el R^2 de la regresión.

Una vez clasificados los consumos en las distintas zonas de exigencia se pueden obtener visualizaciones y estadísticas; en general y también bajo ciertos criterios como filtros de comunas y empresas concesionarias, por ejemplo.

En la Figura 5.64 se resume la metodología propuesta a través de un esquema.

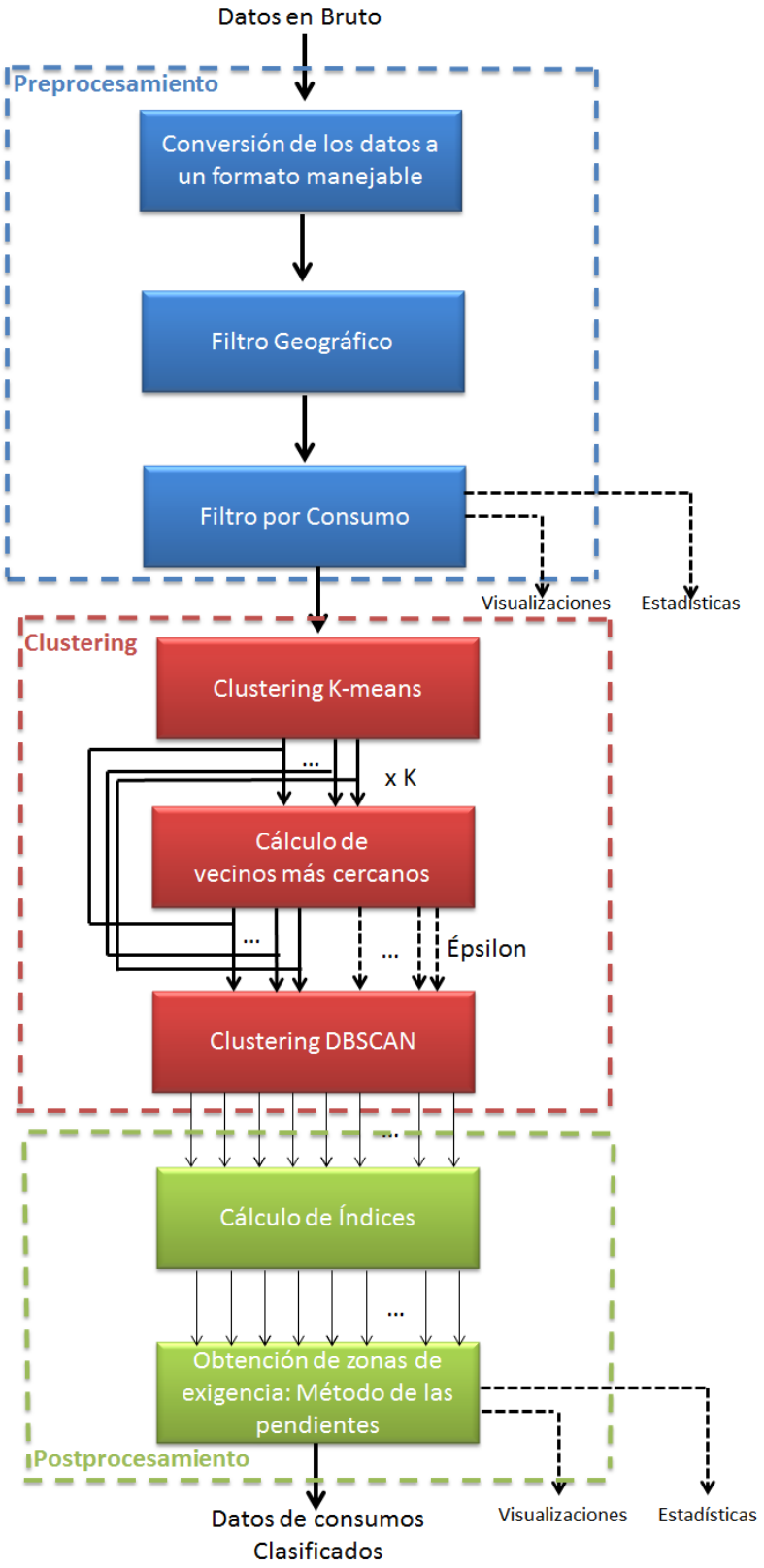


Figura 5.64: Diagrama de la metodología propuesta.

Capítulo 6 - Conclusiones

Sobre la base de múltiples pruebas con los distintos programas computacionales disponibles y los algoritmos que ellos tienen implementados, se elaboró una metodología para encontrar zonas de exigencias de calidad de suministro que no están ligadas a las empresas distribuidoras ni a zonas de concesión de distribución. Esta metodología se elaboró con datos reales de consumo anual de energía de clientes de la octava y novena región, suministrados por las empresas concesionarias de distribución a la Superintendencia de Electricidad y Combustibles. Los datos contienen múltiples atributos, siendo la localización geográfica y el consumo anual de energía la información utilizada para clasificar los clientes en distintas zonas de exigencia.

Los atributos geográficos y de consumo de energía dan mejor cuenta de la calidad de suministro necesaria que el esquema actual basado en índices poblacionales, kilómetros de redes de distribución entre otros, dado que corrige problemas como el mostrado en la zona de Temuco – Padre las Casas donde los clientes de nivel de consumo de energía similar y cercanía geográfica estaban bajo distintas exigencias en calidad de suministro solamente por el hecho de que su energía era suministrada por dos empresas concesionarias de distribución diferentes. Otro problema que esta metodología corrige es la sobre-exigencia de calidad en lugares que no lo necesitan, como se observa en al caso de la zona Villarrica – Pucón: Todo estaba clasificado homogéneamente con una exigencia media, con la nueva metodología se determina que existen dos zonas que deberían tener distinta exigencias en calidad de suministro, una media para el centro de las comunas, donde se encuentra la zona urbana y una más baja para el resto. Esto tiene un efecto directo en las tarifas a clientes finales, puesto que la tarifa se fija dada una calidad de servicio determinada. Una mayor exigencia en calidad de suministro implica para la empresa concesionaria de distribución incurrir en gastos para aumentar la confiabilidad del sistema, como por ejemplo la redundancia de los equipos o el aumento de tamaño de los departamentos de mantención, lo cual es reconocido en los estudios tarifarios como parte del Valor Nuevo de Reemplazo y los Costos de Operación, Mantención y Administración que determinan el Valor Agregado de Distribución.

Respecto a la metodología en sí, es vital realizar una etapa de pre-procesamiento para verificar que la información entregada por las empresas distribuidoras es coherente con el análisis que se desea hacer. Se debe filtrar de modo de quedarse solamente con datos pertinentes.

La etapa de clustering posee 3 parámetros: el número de clusters K, el radio del vecindario épsilon y el número mínimo de puntos por subcluster. Las soluciones varían al variar estos parámetros, pero no existe un método analítico que permita distinguir que solución es mejor que otra. Por lo tanto, es necesario que quien lleve a cabo la metodología tenga conocimiento del problema a resolver que le permita discernir la mejor solución entre un conjunto de distintas soluciones. Así mismo, si se decide adoptar esta metodología para determinar las zonas de exigencia de calidad de suministro, es conveniente que sea llevada a cabo por un organismo independiente que no tenga intereses creados.

La manera de determinar las zonas de exigencias, una vez determinado los índices de densidad de consumo por área, puede ser cambiada, por ejemplo a un enfoque basado en percentiles, dependiendo de los resultados. Para el caso particular analizado con los consumos de la octava y novena región se consideró que un enfoque basado en ajustes de regresiones lineales para agrupar los subcluster resultó más conveniente. El número de zonas de exigencia no tiene por qué ser 3, como lo es en el esquema actual. En el caso de estudio se eligió separar en tres zonas porque se observaron 3 marcados grupos de subclusters.

En caso de implementar esta metodología, se sugiere partir utilizando las mismas metas que existen para las zonas actuales, y ver cómo evolucionan de un período de estudio a otro.

Como trabajo a futuro se plantea:

- Seguir realizando experimentación con los métodos que no fueron descartados y no alcanzaron a ser probados, así como también con los ya probados para seguir ajustando los parámetros. El tiempo de procesamiento de los datos fue una limitante importante para no haber probado más métodos y llegar al ajuste óptimo de parámetros. Un procesamiento normal, corrido sobre un PC con procesador de 2.13 GHz y Memoria RAM de 4 GB tomaba de 20 minutos a 8 horas en converger a una solución, dependiendo de la cantidad de datos que debía clasificar.

- Experimentar con métodos de clustering más específicos que no están implementados en los programas computacionales utilizados.

- Poner a prueba la metodología con consumos de otras regiones o de todo el territorio nacional.

- Desarrollar herramientas que ayuden a manejar la visualización de los datos, en este sentido se podría desarrollar un software para el post-procesamiento que permita el ajuste automático de las escalas de los gráficos, el filtrado por comunas, por empresas, por clusters, entre otros, además de la obtención de estadísticas y, en general, cualquier dato que pueda facilitar el análisis.

Capítulo 7 - Bibliografía

[1] Decreto N° 1T, FIJA FÓRMULAS TARIFARIAS APLICABLES A LOS SUMINISTROS SUJETOS A PRECIOS REGULADOS QUE SE SEÑALAN, EFECTUADOS POR LAS EMPRESAS CONCESIONARIAS DE DISTRIBUCIÓN QUE INDICA, Diario Oficial de la República de Chile, Santiago, Chile, 5 de Noviembre de 2012.

[2] LEY Nº 18.410, CREA LA SUPERINTENDENCIA DE ELECTRICIDAD Y COMBUSTIBLES, Diario Oficial de la República de Chile, Santiago, Chile, 22 de Mayo de 1985.

[3] Generadoras de Chile A.G., «Sector Generación Eléctrica,» [En línea]. Disponible: http://generadoras.cl/energia-electrica/sector-generacion-electrica/. [Último acceso: 14 Noviembre 2012].

[4] D.F.L. Núm. 4/20.018, FIJA TEXTO REFUNDIDO, COORDINADO Y SISTEMATIZADO DEL DECRETO CON FUERZA DE LEY Nº 1, DE MINERIA, DE 1982, LEY GENERAL DE SERVICIOS ELECTRICOS, EN MATERIA DE ENERGIA ELECTRICA, Diario Oficial de la República de Chile, Santiago, Chile, 5 de Febrero de 2007.

[5] LEY NUM. 20.018, MODIFICA EL MARCO NORMATIVO DEL SECTOR ELECTRICO, Diario Oficial de la República de Chile, Santiago, Chile, 19 Mayo de 2005.

[6] DECRETO SUPREMO N°327, FIJA REGLAMENTO DE LA LEY GENERAL DE SERVICIOS, Diario Oficial de la República de Chile, Santiago, Chile, 10 de septiembre de 1998.

[7] Resolución Ministerial Exenta N° 24, DICTA NORMA TECNICA SOBRE CONEXION Y OPERACION DE PEQUEÑOS MEDIOS DE GENERACION DISTRIBUIDOS EN INSTALACIONES DE MEDIA TENSION, Diario Oficial de la República de Chile, Santiago, Chile, 25 de Mayo de 2007.

[8] Comisión Nacional de Energía, LA REGULACIÓN DEL SEGMENTO DISTRIBUCIÓN EN CHILE, Junio de 2006.

[9] Ley N° 19.940, REGULA SISTEMAS DE TRANSPORTE DE ENERGIA ELECTRICA, ESTABLECE UN NUEVO REGIMEN DE TARIFAS PARA SISTEMAS ELECTRICOS MEDIANOS E INTRODUCE LAS ADECUACIONES QUE INDICA A LA LEY GENERAL DE SERVICIOS ELECTRICOS, Diario Oficial de la República de Chile, Santiago, Chile, 13 de Marzo de 2004.

[10] Norma Española UNE-EN 60160, Asociación Española de Normalización y Certificación (AENOR), Características de la tensión suministrada por las redes generales de distribución, Madrid, España, Enero 2001.

[11] J. R. Abbad, Calidad del servicio: Regulación y optimación de inversiones, Madrid, España: Universidad Pontificia Comillas, 2000.

[12] IEEE Std 1366-1998, IEEE Trial-Use Guide for Electric Power Distribution Reliability Indices.

[13] A. M. V. Esteban Inga Llanca, Calidad de Suministro Eléctrico en el Perú.

[14] Comisión Nacional de Energía, Bases para el Cálculo de las Componentes del Valor Agregado de Distribución, Santiago, Chile, 2012.

[15] SYNEX Ingenieros Consultores, «Anexo 1,» de estudio para la CNE denominado PROPUESTA DE NORMA TÉCNICA DE CALIDAD DE SERVICIO PARA SISTEMAS DE DISTRIBUCIÓN, Santiago, Chile, 2010.

[16] Resolución Ministerial Exenta N° 53, DICTA NORMA TECNICA SOBRE DEFINICION DE ZONAS RURALES Y EXIGENCIAS DE CALIDAD DE SERVICIO, Diario Oficial de la República de Chile, Santiago, Chile, 26 de Octubre de 2006.

[17] P.-N. Tan, M. Steinbach y V. Kumar, «Introduction,» de Introduction to Data Mining, Pearson Education, Inc., 2006.

[18] P.-N. Tan, M. Steinbach y V. Kumar, «Cluster Analysis: Basic Concepts and Algorithms,» de Introduction to Data Mining, Pearson Education Inc., 2006.

[19] G. Karypis, CLUTO A Clustering Toolkit, Minneapolis: University of Minnesota, Department of Computer Science, 28 de Noviembre de 2003.

[20] Comisión Nacional de Energía, Resolución Exenta N° 303, Santiago, 07 de Mayo de 2012.

Capítulo 8 – Anexos

Anexo A: Listado de Archivos Entregados

Se entregan adjuntos al texto de esta memoria una serie de archivos correspondientes a los resultados obtenidos en las distintas pruebas. En el siguiente listado se señala la ruta y el nombre del archivo y una breve descripción del contenido.

Resultados 1

Tabla A.1 - Ruta: Clustering\Resultados 1\Formato CLUTO\Resultados\

Nombre del Archivo	Descripción
Resultados Totales.xlsx	Resultados con los tres algoritmos de CLUTO sobre el total de los datos
Seleccion.xlsx	Resultados con 5 algoritmos de CLUTO sobre la selección de datos
Sol1 Gráfico.xlsx	Gráfico de distribución espacial de consumos con algoritmo rb
Sol1.xlsx	Resultados del algoritmo rb con los histogramas de consumo por cluster
Sol2 Gráfico.xlsx	Gráfico de distribución espacial de consumos con algoritmo rbr
Sol2.xlsx	Resultados del algoritmo rbr con los histogramas de consumo por cluster
Sol3 Gráfico.xlsx	Gráfico de distribución espacial de consumos con algoritmo direct
Sol3.xlsx	Resultados del algoritmo direct con los histogramas de consumo por cluster

Tabla A.2 - Ruta: Clustering\Resultados 1\Formato CLUTO\

Nombre del Archivo	Descripción
Instruccion.txt	Archivo de texto que contiene la instrucción de entrada para vcluster
seleccion.mat	Archivo de entrada para el programa vcluster con los datos de selección
seleccion.mat.clabel	Nombres de las columnas de los datos de entrada selección.mat
seleccion.mat.clustering.4	Archivo de salida del clustering entragado por vcluster
total.mat	Archivo de entrada para el programa vcluster con el total de los datos
total.mat.clustering.4	Archivo de salida del clustering entragado por vcluster

Tabla A.3 - Ruta: Clustering\Resultados 1\Formato RapidMiner\Resultados\

Nombre del Archivo	Descripción
Sol1 - Gráfico.xlsx	Gráfico de distribución espacial de consumos con algoritmo K-means
Sol1.xlsx	Resultados del algoritmo K-means con los histogramas de consumo por cluster
Sol2 - Gráfico.xlsx	Gráfico de distribución especial de consumos con algoritmo X-means
Sol2.xlsx	Resultados del algoritmo X-means con los histogramas de consumo por cluster
Total.xlsx	Resultados para el total de los algoritmos ejecutados con RapidMiner

Tabla A.4 - Ruta: Clustering\Resultados 1\Formato RapidMiner\

Nombre del Archivo	Descripción
Seleccion.xlsx	Archivo de entrada para RapidMiner con los datos de selección
Total.xlsx	Archivo de entrada para RapidMiner con el total de los datos

Tabla A.5 - Ruta: Clustering\Resultados 1\Formato WEKA\Resultados\

Nombre del Archivo	Descripción
Sol1 - Gráfico.xlsx	Gráfico de distribución espacial de consumos con algoritmo simpleKmeans
Sol1.xlsx	Resultados del algoritmo simpleKmeans con los histogramas de consumo por cluster
Sol2 - Gráfico.xlsx	Gráfico de distribución espacial de consumos con algoritmo Xmeans
Sol2.xlsx	Resultados del algoritmo Xmeans con los histogramas de consumo por cluster
Sol3 - Gráfico.xlsx	Gráfico de distribución espacial de consumos con algoritmo FarthestFirts
Sol3.xlsx	Resultados del algoritmo FarthestFirts con los histogramas de consumo por cluster
Sol4 - Gráfico.xlsx	Gráfico de distribución espacial de consumos con algoritmo MakeDensityBased
Sol4.xlsx	Resultados del algoritmo MakeDensityBased con los histogramas de consumo por cluster
SolK10 - Gráfico.xlsx	Gráfico de distribución espacial de consumos con algoritmo FarthestFirst y K=10
SolK10.xlsx	Resultados del algoritmo FarthestFirst y K=10 con los histogramas de consumo por cluster

Tabla A.6 - Ruta: Clustering\Resultados 1\Formato WEKA\

Nombre del Archivo	Descripción
Seleccion.csv	Archivo de entrada para WEKA con los datos de selección
Total.csv	Archivo de entrada para WEKA con el total de los datos

Tabla A.7 - Ruta: Clustering\Resultados 1\

Nombre del Archivo	Descripción
Evaluación de Algoritmos - Seleccion.xlsx	Resumen de los resultados para la evaluación de los algoritmos con los datos de selección
Evaluación de Algoritmos -Totales.xlsx	Resumen de los resultados para la evaluación de los algoritmos con los datos totales
Resultados.pdf	Explicación de los resultados obtenidos

Resultados 2

Tabla A.8 - Ruta: Clustering\Resultados 2\Área

Nombre del Archivo	Descripción
Sol - Gráfico.xlsx	Gráfico de distribución espacial de consumos considerando método de índice de consumo por área
Sol.arff	Archivo de salida de WEKA
Sol.xlsx	Gráfico de K-means Weka para evaluar método de índice de consumo por área
Total.arff	Archivo de entrada de WEKA con el total de los datos

Tabla A.9 - Ruta: Clustering\Resultados 2\cambio de distancia

Nombre del Archivo	Descripción
Euc – Gráfico.xlsx	Gráfico de distribución espacial de consumos considerando distancia euclidiana
Euc.xlsx	Resultados de considerar distancia euclidiana con los histogramas de consumo por cluster
Man - Gráfico .xlsx	Gráfico de distribución espacial de consumos considerando distancia Manhattan
Man.xlsx	Resultados de considerar distancia Manhattan con los histogramas de consumo por cluster

Tabla A.10 - Ruta: Clustering\Resultados 2\Distancia Momento

Nombre del Archivo	Descripción
Resultados.xlsx	Resumen de resultados obtenido para las 4 soluciones
Sol1 - Grafico.xlsx	Resultados de correr el algoritmo por primera vez
Sol1.xlsx	Gráfico de distribución espacial de consumos considerando distancia momento
Sol4 - Grafico.xlsx	Resultados de correr el algoritmo al correrlo 100 veces (tras mejoras)
Sol4.xlsx	Gráfico de distribución espacial de consumos considerando distancia momento

Nota: Las soluciones sol2 y sol3, corresponden a corridas individuales del algoritmo, tal como la 1.

Tabla A.11 - Ruta: Clustering\Resultados 2\

Nombre del Archivo	Descripción
kmeansmomento.m	Método en MATLAB para el cálculo de clustering con la distancia momento
Resultados 2.pdf	Explicación de resultados
Resultados 22.pdf	Explicación de resultados tras implementar mejoras
total.mat	Total de datos en formato MATLAB, para correr clustering con distancia momento

Resultados 3

Recordar que los resultados 3 corresponder a repetir el trabajo realizado en 2 pero sin eliminar los datos de consumos altos

Tabla A.12 - Ruta: Clustering\Resultados 3\

Nombre del Archivo	Descripción
data.mat	100% de los datos a considerar en formato MATLAB
Resultados 3.pdf	Explicación de resultados
sol.xlsx	Resultados de correr el algoritmo kmeans con distancia euclidiana
sol - Grafico.xlsx	Gráfico de distribución espacial de consumos considerando distancia euclidiana
Sol XY.xlsx	Resultados de correr el algoritmo kmeans con Weka
Sol XY - Gráfico.xlsx	Gráfico de distribución espacial de consumos con Kmeans y Weka
sol2.xlsx	Resultados de correr el algoritmo kmeans momento normalizado e iterativo
sol2 - Grafico.xlsx	Gráfico de distribución espacial de consumos considerando distancia momento mejorada
Total 100.csv	Dato de entrada para WEKA con el 100% de los datos a considerar.

Resultados 4

Tabla A.13 - Ruta: Clustering\Resultados 4\Clusters

Nombre del Archivo	Descripción
x.csv	Archivo de Entrada para RapidMiner con los datos del cluster x. x= 01, …, 10.

Tabla A.14 - Ruta: Clustering\Resultados 4\resultados\Gráficos k-nn

Nombre del Archivo	Descripción
cx -1.png	Gráfico destacando la distancia épsilon antes del codo para el cluster x. x=01,…,10
cx -2.png	Gráfico destacando la distancia épsilon en el codo para el cluster x. x=01,…,10
cx -3.png	Gráfico destacando la distancia épsilon despues del codo para el cluster x. x=01,…,10
cx.mat	Archivo MATLAB que contiene el resultado del cálculo de los knn para el cluster x.
knn.xlsx	Tabla resumen con los valores de eps1, eps2 y eps3 para cada cluster

Tabla A.15 - Ruta: Clustering\Resultados 4\resultados\

Nombre del Archivo	Descripción
CX -01.xlsx	Resultados de clustering DBSCAN para el cluster x considerando eps1 X=01,…,10
CX -02.xlsx	Resultados de clustering DBSCAN para el cluster x considerando eps2 X=01,…04
CX -03.xlsx	Resultados de clustering DBSCAN para el cluster x considerando eps3 X=01

Tabla A.16 - Ruta: Clustering\Resultados 4\

Nombre del Archivo	Descripción
Ajuste de parametros DBSCAN c01.xlsx	Resultados de experimentación ajustando parámetros del algoritmo DBSCAN para el cluster 01
dist.m	Método MATLAB para el cálculo de los vecinos más cercanos
Resultados 4.pdf	Explicación de resultados

Resultados 5

Tabla A.17 - Ruta: Clustering\Resultados 5\Clusters

Nombre del Archivo	Descripción
cx.csv	Archivo de Entrada para RapidMiner con los datos del cluster x. x= 01, …, 10.
dist2.m	Archivo para el cálculo de vecinos más cercanos considerando sólo distancia geográfica
dist2.m	Archivo para el cálculo de vecinos más cercano considerando distancia geográfica y consumo
knn.m	Resultados de correr el cálculo de los vecinos más cercanos con dist2.m
knn3.m	Resultados de correr el cálculo de los vecinos más cercanos con dist3.m
matlab.m	Datos de entrada para correr cálculo de vecinos más cercanos.

Tabla A.18 - Ruta: Clustering\Resultados 5\Soluciones

Nombre del Archivo	Descripción
cx soly.xlsx	Solución número y para el cluster x
Densidad.xlsx	Cálculo del índice de densidad de consumo por área para los subcluster y los resultados de la clasificación por terciles y mitades.
Resultado Clasificación 1.xlsx	Gráfico de la distribución espacial de los consumos al clasificar con terciles
Resultado Clasificación 1.xlsx	Gráfico de la distribución espacial de los consumos al clasificar por mitades
Resultado XY.xlsx	Gráfico de la distribución espacial de los consumos separados por clusters.

Tabla A.19 - Ruta: Clustering\Resultados 5\

Nombre del Archivo	Descripción
Datos.csv	Archivo de entrada para realizar clustering K-means con WEKA
macro densidad.txt	Código VBA para realizar el cálculo de índices
macro graficar.txt	Código VBA para realizar visualizaciones de la distribución espacial de los consumos
Resultados 5.pdf	Explicación de los resultados
Soluciones c0.xlsx	Resumen del ajuste de parámetros con el cluster c0

Resultados 6

Tabla A.20 - Ruta: Clustering\Resultados 6\Análisis de Casos

Nombre del Archivo	Descripción
Gran Concepción.xlsx	Comparación en la zona del Gran Concepción del esquema actual y esquema propuesto
Temuco - Padre las Casas.xlsx	Comparación en la zona de Temuco-Padre las casas del esquema actual y esquema propuesto
Villarrica - Pucon.xlsx	Comparación en la zona de Villarrica-Pucón del esquema actual y esquema propuesto

Tabla A.21 - Ruta: Clustering\Resultados 6\Clusters

Nombre del Archivo	Descripción
cx.csv	Archivo de entrada para realizar DBSCAN con RapidMiner. x = 0,…, 5.
cx.png	Gráfico de vecinos más cercanos con eps a considerar. x = 0,…, 5.
cx.xlsx	Archivo con la información de comunas y empresas distribuidoras, por cada cluster.
dist.m	Método MATLAB para el cálculo de los vecinos más cercanos
knn	Resultados del cálculo de vecinos más cercanos para los 6 clusters

Tabla A.22 - Ruta: Clustering\Resultados 6\

Nombre del Archivo	Descripción
cx- sol01.xlsx	Solución de subclustering para el cluster x
Clasificación – 2.xlxs	Clasificación de subcluster a zona de exigencia según enfoque de pendientes
Claisifación.xlxs	Clasificación de subcluster a zona de exigencia según enfoque de terciles
Clusters.xlsx	Resultados de clustering
Resultados 6.pdf	Explicación de Resultados

Otros

Tabla A.23 - Ruta: Preprocesados\

Nombre del Archivo	Descripción
Clasificacion consumos.xlsx	Esquema de clasificación vigente de consumos en zonas de exigencias
Consumos.xlsx	Información de los consumos a ser considerados en los cálculos
Histograma.xlsx	Visualización del histograma de consumos (para los datos considerados)
Seleccion.xlsx	Información de los datos de prueba utilizados para evaluar algoritmos

Tabla A.24 - Ruta: \

Nombre del Archivo	Descripción
10 PCR + Consumos.xlsx	Información del total del consumos (sin aplicar filtros) extraída de la base de datos
Códigos.xlsx	Códigos de las empresas concesionarais de distribución.
Comunas.xlsx	Códigos de las comunas de la octava y novena región
Consumos.xlsx	Gráfico de la distribución espacial de los consumos sin aplicar ningún filtro
Histograma.xlsx	Visualización del histograma de consumos para el total de consumos, sin aplicar filtros