

UNIVERSIDAD POLITÉCNICA SALESIANA
SEDE CUENCA

CARRERA DE INGENIERÍA ELÉCTRICA

**TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL
TÍTULO DE INGENIERO ELÉCTRICO**

PROYECTO TÉCNICO DE ENFOQUE INVESTIGATIVO:

**“APLICACIÓN DE TÉCNICAS DE CLUSTERING Y
DETECCIÓN DE ANOMALÍAS PARA LA CLASIFICACIÓN
DE CONSUMOS DE ENERGÍA ELÉCTRICA”**

AUTORES:

PEDRO SEBASTIÁN CONTRERAS ORTIZ
MILTON XAVIER MORA ALVAREZ

TUTOR:

ING. JOHNNY XAVIER SERRANO GUERRERO, MSc.

CUENCA – ECUADOR

2018

CESIÓN DE DERECHOS DE AUTOR

Nosotros; **Pedro Sebastián Contreras Ortiz**, con documento de identificación N° 010423146-9 y **Milton Xavier Mora Alvarez**, con documento de identificación N° 010573691-2, manifestamos nuestra voluntad y cedemos a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que somos autores del trabajo de titulación: “**APLICACIÓN DE TÉCNICAS DE CLUSTERING Y DETECCIÓN DE ANOMALÍAS PARA LA CLASIFICACIÓN DE CONSUMOS DE ENERGÍA ELÉCTRICA**”, mismo que ha sido desarrollado para optar por el título de: Ingeniero Eléctrico, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En aplicación a lo determinado en la Ley de Propiedad Intelectual, en nuestra condición de autores nos reservamos los derechos morales de la obra antes citada. En concordancia, suscribimos este documento en el momento que hacemos entrega del trabajo final en formato impreso y digital a la Biblioteca de la Universidad Politécnica Salesiana.

Cuenca, Enero del 2018



Pedro Sebastián Contreras Ortiz

010423146-9



Milton Xavier Mora Alvarez

010573691-2

CERTIFICACIÓN

Yo; Ing. Johnny Xavier Serrano Guerrero, MSc., declaro que bajo mi tutoría fue desarrollado el trabajo de titulación “**APLICACIÓN DE TÉCNICAS DE CLUSTERING Y DETECCIÓN DE ANOMALÍAS PARA LA CLASIFICACIÓN DE CONSUMOS DE ENERGÍA ELÉCTRICA**”, realizado por los autores **Pedro Sebastián Contreras Ortiz** y **Milton Xavier Mora Alvarez**, obteniendo el Proyecto Técnico de Enfoque Investigativo que cumple con todos los requisitos estipulados por la Universidad Politécnica Salesiana.

Cuenca, Enero del 2018



Ing. Johnny Xavier Serrano Guerrero, MSc.
010498338-2

DECLARATORIA DE RESPONSABILIDAD

Nosotros; **Pedro Sebastián Contreras Ortiz**, con número de cédula 010423146-9 y **Milton Xavier Mora Alvarez**, con número de cédula 010573691-2, autores del trabajo de titulación: **“APLICACIÓN DE TÉCNICAS DE CLUSTERING Y DETECCIÓN DE ANOMALÍAS PARA LA CLASIFICACIÓN DE CONSUMOS DE ENERGÍA ELÉCTRICA”** certificamos que el total contenido de este Proyecto Técnico de Enfoque Investigativo es de nuestra exclusiva responsabilidad y autoría.

Cuenca, Enero del 2018



Pedro Sebastián Contreras Ortiz

010423146-9



Milton Xavier Mora Alvarez

010573691-2

DEDICATORIA

Consagro este proyecto técnico primeramente a Dios por haberme brindado salud, fortaleza y sabiduría.

A mi padre el Sr. Milton Salvador Mora Beltrán, quien me han manifestado su apoyo incondicional, gracias a sus consejos y recomendaciones.

A mi madre la Sra. Ruth Yolanda Alvarez Peralta (+), quien desde pequeño me enseñó la responsabilidad, dedicación y el afán que se debe poner a las cosas cuando se desea lograr una meta.

A mi hermano Juan Carlos Mora Alvarez y a toda mi familia por ser parte esencial de mi vida, con quienes he compartido grandes momentos.

A mi enamorada Katherine Elizabeth Ochoa Carchi quien me apoyó y animó mediante su incansable frenesí para continuar en el desarrollo y cumplimiento de mis metas en todo momento.

Milton Xavier Mora Alvarez.

*Primeramente, quisiera dedicar este trabajo a Dios, por haberme dado la fortaleza y sabiduría
para culminar con éxito esta etapa.*

*De igual manera, dedicar este trabajo a mis padres, Bolívar y Magdalena por ser los pilares más
importantes en mi formación profesional.*

A mis hermanos, José y Felipe quiénes han sido y serán mi más claro ejemplo a seguir.

*A toda mi familia quiénes en algún momento me supieron guiar, aconsejar y ayudar para cumplir
con todas las metas que me he propuesto.*

Pedro Sebastián Contreras Ortiz.

AGRADECIMIENTOS

Lo primero de todo, me gustaría agradecer a Dios por brindarme salud, fortaleza y sabiduría; y permitirme cumplir uno de mis objetivos y finalizar esta meta crucial de mi vida.

Quisiera dedicar este trabajo a mis padres por el sustento brindado para culminar y sobresalir en mis estudios de Pregrado.

En particular, me gustaría dar las gracias al Ing. Johnny Xavier Serrano Guerrero, MSc.; por despertar mi interés en la temática propuesta, por el conocimiento transmitido, por su supervisión y dirección de inicio a fin en el desarrollo del presente proyecto de titulación.

Estoy especialmente agradecido con mi compañero trabajo Pedro Sebastián Contreras Ortiz por su dedicación y responsabilidad en la elaboración del presente proyecto de titulación.

De igual manera, agradezco a las empresas CENTROSUR C.A. Y CONTINENTAL TIRE ANDINA S.A. por su ayuda mediante la entrega de información y recursos necesarios para el desarrollo de este proyecto.

Finalmente, doy las gracias personalmente a la Universidad Politécnica Salesiana, la cual, me instruyó como buen cristiano, honrado ciudadano y excelente profesional en los diversos ámbitos de la Carrera de Ingeniería Eléctrica.

Milton Xavier Mora Alvarez.

Quiero dar infinitamente gracias a Dios, por haberme dado las fuerzas suficientes para cumplir con lo que me propuse hace 6 años.

Agradecer de una manera muy especial a mis padres quiénes me han brindado un apoyo incondicional durante todo este tiempo, sin ellos no estaría en donde estoy, ellos han sido el pilar fundamental para cumplir con esta meta y de ahora en adelante poder retribuirles un poco de lo mucho que me han brindado para que así se sientan orgullosos de la persona en que me he convertido.

Quiero dar gracias a mis hermanos, que desde pequeño me dieron el ejemplo de seguir adelante y cumplir con lo que me proponga.

De manera particular, agradecer al Ing. Johnny Xavier Serrano Guerrero, por habernos guiado con sus conocimientos durante este proceso de aprendizaje.

Agradecer de igual manera a mi amigo y compañero de trabajo, Milton Xavier Mora Alvarez quién a pesar de todas las dificultades que se presentaron, siempre estuvo ahí para ayudarme desde primer ciclo.

A los docentes de la Universidad Politécnica Salesiana, que supieron transmitirnos sus conocimientos y hacernos ver que la Ingeniería es una carrera que debe estar al servicio de la sociedad bajo el lema de “Buenos cristianos y honrados ciudadanos”.

Pedro Sebastián Contreras Ortiz.

RESUMEN

El presente proyecto técnico de enfoque investigativo tiene como objetivo aplicar técnicas adecuadas de segmentación de datos (clustering) y detección de anomalías para la clasificación de consumos de energía eléctrica. Para este propósito; se desarrolla un programa en el Software MATLAB; que, en primer lugar, aplica técnicas de minería de datos para el preprocesamiento de la selección y preparación de información útil correspondiente a las mediciones de energía eléctrica del consumidor o los consumidores que sean estudiados. Una vez con la información útil, se realiza la identificación y transformación de variables (factores de forma) que permitan caracterizar a los consumos de energía eléctrica. Con las variables identificadas, se elige adecuadamente las técnicas de segmentación de datos para posteriormente aplicar una nueva metodología con respecto a la detección de outliers en los perfiles de consumo de energía eléctrica. Finalmente, esto conlleva a identificar patrones de comportamiento del consumo de energía eléctrica; sintetizando y evaluando el uso de las técnicas aplicadas.

ABSTRACT

This technical project in an investigative approach aims to apply appropriate data clustering techniques and detection of anomalies for the classification of electric power consumption. For this purpose; an algorithm is developed in MATLAB; this algorithm applies data mining techniques for preprocessing and preparation of useful data that corresponds to the measurements of electrical energy of the consumer or the consumers that are studied in this project. Once the useful information is obtained, the identification and transformation of form factors that allow the characterization of the diary electric power consumption. With the form factors identified, the clustering techniques are chosen appropriately to subsequently apply a new methodology to the detection of outliers in the electric power consumption profiles. Finally, this leads to the behavior patterns identification of electric power consumption; synthesizing and evaluating the use of applied techniques.

PREFACIO

El consumo de energía eléctrica es una característica significativa en la sociedad actual, pues es un factor que está ligado al desarrollo de los sectores residenciales, comerciales e industriales.

La regularización de los sistemas de energía eléctrica ha fomentado la necesidad de monitorizar y controlar la demanda de los usuarios. Debido a esto, la cantidad de datos almacenados por parte de las empresas encargadas de comercializar y distribuir la energía eléctrica han incrementado sustancial y mayoritariamente, lo que supone un problema de utilización para los casos de análisis; siendo estos datos dimensionales y con alto grado de dificultad para su modelamiento.

En base a esto se han desarrollado algunas técnicas para la segmentación e identificación de la demanda eléctrica, a partir de los perfiles de consumo eléctrico obtenidos de los diferentes tipos de usuarios.

Dado lo expuesto, el presente proyecto técnico de enfoque investigativo pretende evidentemente realizar la segmentación de datos y detección de anomalías, y a partir de los resultados obtenidos realizar la toma de decisiones que permitan una correcta clasificación de consumos de energía eléctrica y la definición de patrones de consumo y producción que contribuyan a la solución de problemas encontrados en los diversos consumidores de energía eléctrica.

Este proyecto técnico es presentado a través de 5 capítulos, revelando de una forma secuencial y sistemática la manera en que se llegó a obtener los resultados de algunos consumidores de energía eléctrica, de tal modo que cualquier persona consiga comprender claramente el desarrollo de este trabajo. Es así, que en los capítulos 1 y 2 se definen los fundamentos teóricos que permitieron el desarrollo del presente proyecto técnico; así como una breve descripción de las diversas técnicas de clustering existentes.

Posteriormente, en el capítulo 3 se detalla el desarrollo del programa desarrollado en el Software MATLAB que permitirá la aplicación de técnicas de clustering k -Means y k -Medoids; así como, la detección de anomalías en los perfiles de consumo de energía eléctrica para su respectiva clasificación y posterior identificación de patrones de consumo energético.

En el capítulo 4, se presenta los resultados obtenidos mediante el programa desarrollado en el Software MATLAB de las instalaciones pertenecientes a la Universidad Politécnica Salesiana – Sede Cuenca, Universidad Politécnica de Valencia – Madrid y Continental Tire Andina S.A.; así como, la verificación de la segmentación de datos en el software Weka y SPSS, respectivamente.

Finalmente, en el capítulo 5 se efectúa un análisis de los resultados obtenidos que permiten llegar a ciertas conclusiones y recomendaciones expuestas al final del presente trabajo.

GLOSARIO

UPS	Universidad Politécnica Salesiana
GIE	Grupo de Investigación de Energías
UPV	Universidad Politécnica de Valencia
PTEI	Proyecto Técnico de Enfoque Investigativo
CEE	Consumo de Energía Eléctrica
PCEE	Perfiles de Consumo de Energía Eléctrica
TC	Técnicas de Clustering
VA	Valores Atípicos
VS	Valores de Silueta
DE	Distancias Euclidianas
FDP	Función de Densidad de Probabilidad
FDAI	Función de Distribución Acumulativa Inversa
MSV	Máquinas de Soporte Vectorial
DA	Detección de Anomalías
PCPEE	Patrón de Consumo Promedio de Energía Eléctrica
RPC	Reconocimiento de Patrones de Consumo

JUSTIFICACIÓN

Se aplican técnicas estadísticas para el análisis y modelamiento de los datos recolectados por parte de las empresas sobre el consumo de energía eléctrica de los usuarios, sugiriendo modelos desarrollados para asistir en la resolución de los problemas, soporte de ayuda en la toma de decisiones y de la información cualitativa (Pérez, 2013).

Por esta razón las empresas encargadas de proveer el servicio eléctrico en sus diferentes instancias están interesadas en desarrollar nuevas estrategias y productos orientados hacia la demanda, con el objetivo de descubrir curvas de demanda para ofrecer un mejor servicio a cada usuario o de encontrar nuevas oportunidades en el mercado (Valero & Senabre, 2015).

Sin embargo, no sólo es útil descubrir curvas de demanda, es importante además revelar comportamientos, patrones y características similares de los consumidores finales a partir de perfiles de carga; y métodos de segmentación de datos para poder clasificar de la manera más confiable posible a los usuarios del sistema (Valero & Senabre, 2015).

De esta manera, para dicha clasificación se desarrollan estrategias sobre detección eficiente de anomalías en los diferentes grupos de usuarios que oriente a la planificación adecuada de inspecciones de suministro de energía eléctrica por parte de las empresas encargadas de proveer el servicio (Gómez & Arco, 2015).

GRUPO OBJETIVO (BENEFICIARIOS)

- Universidad Politécnica Salesiana – Grupo de Investigación de Energías.
- Empresas comercializadoras y distribuidoras de energía eléctrica.
- Consumidores de energía eléctrica en general.
- Ciudadanía del Cantón Cuenca.

OBJETIVOS

Objetivo general.

- Aplicar técnicas adecuadas de segmentación de datos (clustering) y detección de anomalías para la clasificación de consumos de energía eléctrica.

Objetivos específicos.

- Recopilar información sobre técnicas de segmentación, agrupación de datos y detección de anomalías de perfiles de consumo de energía eléctrica.
- Realizar pruebas aplicativas de las técnicas de clustering y detección de anomalías.
- Aplicar las técnicas de clustering y detección de anomalías con la finalidad de identificar patrones de comportamiento del consumo de energía eléctrica.
- Sintetizar y evaluar el uso de las técnicas aplicadas en el consumo de energía eléctrica.

ÍNDICE GENERAL

CESIÓN DE DERECHOS DE AUTOR	II
CERTIFICACIÓN	III
DECLARATORIA DE RESPONSABILIDAD	IV
DEDICATORIA	IV
AGRADECIMIENTOS	VII
RESUMEN	IX
ABSTRACT.....	X
PREFACIO	XI
GLOSARIO	XIII
JUSTIFICACIÓN	XIV
GRUPO OBJETIVO (BENEFICIARIOS)	XV
OBJETIVOS	XV
ÍNDICE GENERAL.....	XVI
ÍNDICE DE FIGURAS.....	XXI
ÍNDICE DE TABLAS	XXVI
ÍNDICE DE ECUACIONES.....	XXVIII

CAPÍTULO I..... 1

1.1. Minería de datos..... 1

1.1.1. Introducción a la minería de datos..... 1
1.1.2. Definición 1
1.1.3. Proceso de la minería de datos 2
 1.1.3.1. Requerimientos del análisis..... 2
1.1.4. Funcionalidades de la minería de datos 3
 1.1.4.1. Concepto/Clase: Caracterización y Discriminación 3
 1.1.4.2. Patrones frecuentes, asociación y correlación 4
 1.1.4.3. Clasificación y regresión para análisis predictivo 4
 1.1.4.4. Análisis de clusters 4
1.1.5. Aplicaciones de la Minería de Datos 5
 1.1.5.1. Perfil de consumo de energía eléctrica de clientes 5
 1.1.5.2. Segmentación de clientes 6
 1.1.5.3. Identificación de outliers y detección de fraudes 6

CAPÍTULO II 8

2.1. Análisis cluster y detección de outliers..... 8

2.1.1. Aprendizaje automático 8
2.1.2. Definición y Objetivo del Análisis Cluster..... 8
2.1.3. Requerimientos para el análisis cluster 9
2.1.4. Introducción a las técnicas de clustering 10
 2.1.4.1. Métodos de partición 11
 2.1.4.2. Métodos jerárquicos 11
 2.1.4.3. Métodos basados en densidad 11
 2.1.4.4. Métodos basados en malla (Grid-Based Methods)..... 12
 2.1.5. Métodos de partición 12
 2.1.5.1. k-Means: Una técnica basada en los centroides 12
 2.1.5.3. k-Medoids: Una técnica representativa basada en los objetos 14
 2.1.5.4. Introducción a k-Medoids en MATLAB 15
2.1.6. Métodos jerárquicos 16
 2.1.6.1. Comparación entre el Clustering Jerárquico Aglomerativo y
 Divisivo..... 16
2.1.7. Máquinas de soporte vectorial para la detección de anomalías 17
 2.1.7.1. Introducción a las máquinas de soporte vectorial 17
 2.1.7.2. Principios de las máquinas de soporte vectorial..... 17
2.1.8. Distribuciones de Kernel 18
 2.1.8.1. Introducción 18
 2.1.8.2. El histograma..... 18
 2.1.8.3. Polígonos de frecuencia..... 19
 2.1.8.4. Estimadores de densidad por Kernel 20
2.1.9. Gráfica de silueta para la validación del cluster 23

CAPÍTULO III..... 24

3.1. Metodología para el desarrollo del programa en MATLAB 24

3.1.1. Descripción de metodologías previas y tipos de análisis.....	24
3.1.2. Bases de datos para el análisis clustering	27
3.1.3. Enfoque investigativo	28
3.2. Flujograma del programa	28
3.3. Desarrollo del programa en el software MATLAB	31
3.3.1. Lectura de los datos de medición de energía eléctrica	31
3.3.2. Preprocesamiento de datos de medición de energía eléctrica.....	32
3.3.2.1. Reconocimiento mediante días de la semana y rangos horarios	32
3.3.2.2. Obtención de los Factores de Forma que caracterizan los PCEE diarios.....	37
3.3.2.3. Creación del archivo de Microsoft Excel que posee los factores de forma.....	41
3.3.3. Visualización de las curvas de perfiles de carga diarios.....	43
3.3.4. Visualización de factores de forma de los perfiles de carga diarios.....	46
3.3.5. Aplicación de las Técnicas de Clustering k-Means y k-Medoids.....	50
3.3.5.1. Regla de Sturges.....	50
3.3.5.2. Datos de entrada para las técnicas de clustering k-Means y k- Medoids.....	51
3.3.5.3. Opciones de control para las técnicas de clustering k-Means y k-Medoids	51
3.3.5.4. Distancia euclidiana	52
3.3.5.5. Algoritmo k-means++	52
3.3.5.6. Datos de salida de las técnicas de clustering k-Means y k- Medoids.....	54
3.3.5.7. Valor de silueta.....	54
3.3.6. Evaluación del número óptimo de clusters.....	55
3.3.6.1. Análisis basado con respecto a los valores silueta	55
3.3.6.2. Análisis basado con respecto a los tiempos de ejecución del programa	56
3.3.7. Agrupamiento de Datos en base al número óptimo de clusters.....	57
3.3.8. Visualización de los factores de forma agrupados en su respectivo cluster	57
3.3.9. Detección de outliers y clasificación de consumos de energía eléctrica diarios	59
3.3.9.1. Análisis estadístico descriptivo de las distancias euclidianas y los valores de siluetas de los factores de forma	60
3.3.9.2. Probabilidad por Kernel de las distancias euclidianas y valores de silueta.....	63
3.3.9.3. Indicador de outliers en base a las distancias euclidianas y valores de silueta.....	64
3.3.9.4. Aplicación de máquinas de soporte vectorial no supervisadas	65
3.3.9.5. Condiciones base para la detección de outliers y la respectiva clasificación de perfiles de carga de energía eléctrica diarios con y sin valores atípicos de cada cluster	66
3.3.9.6. Visualización de perfiles de carga diarios sin outliers y con outliers de cada cluster.....	67
3.3.9.7. Obtención del patrón de consumo promedio de energía eléctrica diario de cada cluster.....	70
3.3.10. Almacenamiento de los perfiles de carga sin outliers y con outliers.....	71

3.3.11. Almacenamiento del patrón diario de consumo de energía eléctrica	74
--	----

CAPÍTULO IV 76

4.1. Resultados obtenidos mediante el programa desarrollado en MATLAB 76

4.1.1. Consumo de energía eléctrica de la Universidad Politécnica Salesiana (Acometida Principal)	76
4.1.1.1. Valores promedios de las siluetas según el número de clusters con los que se segmentaron los factores de forma	84
4.1.1.2. Evaluaciones del número de clusters más óptimo para los datos en análisis	84
4.1.1.3. Valores de los tiempos de ejecución para cada número de cluster con los que se segmentaron los factores de forma	85
4.1.1.4. Indicador del número de cluster a cuál pertenece cada PCEE diario	87
4.1.1.5. Ubicación de los centroides para cada cluster	88
4.1.1.6. Distancia de los factores de forma que caracterizan cada PCEE diario hasta el centroide de cada cluster	88
4.1.1.7. Análisis estadístico de media aritmética y desviación estándar de las distancias a los centroides y de los valores de silueta de cada punto para la detección de perfiles de carga diarios con outliers y sin outliers de cada cluster	96
4.1.2. Consumo de energía eléctrica de la Universidad Politécnica Salesiana (Edificio Cornelio Merchán)	107
4.1.2.1. Valores promedios de las siluetas según el número de clusters con los que se segmentaron los factores de forma	115
4.1.2.2. Evaluaciones del número de clusters más óptimo para los datos en análisis	115
4.1.2.3. Valores de los tiempos de ejecución para cada número de cluster con los que se segmentaron los factores de forma	116
4.1.2.4. Indicador del número de cluster a cuál pertenece cada PCEE diario	118
4.1.2.5. Ubicación de los centroides para cada cluster	119
4.1.2.6. Distancia de los factores de forma que caracterizan cada PCEE diario hasta el centroide de cada cluster	119
4.1.2.7. Análisis estadístico de media aritmética y desviación estándar de las distancias a los centroides y de los valores de silueta de cada punto para la detección de perfiles de carga diarios con outliers y sin outliers de cada cluster	127
4.1.3. Consumo de Energía Eléctrica de CONTINENTAL TIRE ANDINA S.A.	138
4.1.3.1. Valores promedios de las siluetas según el número de clusters con los que se segmentaron los factores de forma	146
4.1.3.2. Evaluaciones del número de clusters más óptimo para los datos en análisis	146
4.1.3.3. Valores de los tiempos de ejecución para cada número de cluster con los que se segmentaron los factores de forma	147
4.1.3.4. Indicador del número de cluster a cuál pertenece cada PCEE diario	149
4.1.3.5. Ubicación de los centroides para cada cluster	150

4.1.3.6. Distancia de los factores de forma que caracterizan cada PCEE diario hasta el centroide de cada cluster.....	150
4.1.3.7. Análisis estadístico de media aritmética y desviación estándar de las distancias de cada perfil de carga de energía eléctrica a su centroide y de los valores de silueta para la clasificación de dichos perfiles de carga con outliers y sin outliers.....	163
4.1.4. Consumo de Energía Eléctrica de la Universidad Politécnica de Valencia.....	174
4.1.4.1. Valores promedios de las siluetas según el número de clusters con los que se segmentaron los factores de forma.....	182
4.1.4.2. Evaluaciones del número de clusters más óptimo para los datos en análisis.....	182
4.1.4.3. Valores de los tiempos de ejecución para cada número de cluster con los que se segmentaron los factores de forma.....	183
4.1.4.4. Indicador del número de cluster a cuál pertenece cada PCEE diario.....	185
4.1.4.5. Ubicación de los centroides para cada cluster.....	186
4.1.4.6. Distancia de los factores de forma que caracterizan cada PCEE diario hasta el centroide de cada cluster.....	186
4.1.4.7. Análisis estadístico de media aritmética y desviación estándar de las distancias de cada perfil de carga de energía eléctrica a su centroide y de los valores de silueta para la clasificación de dichos perfiles de carga con outliers y sin outliers.....	198
4.2. Análisis de resultados de segmentación mediante el software WEKA.....	214
4.2.1. Análisis de resultados de la acometida principal de la UPS.....	214
4.2.2. Análisis de resultados de la del Edificio Cornelio Merchán - UPS.....	215
4.2.3. Análisis de resultados de CONTINENTAL GENERAL TIRE ANDINA.....	217
4.2.4. Análisis de resultados de la UPV.....	219
4.3. Análisis de resultados de segmentación mediante el software SPSS.....	220
4.3.1. Análisis de resultados de la acometida principal de la UPS.....	220
4.3.2. Análisis de resultados de la del Edificio Cornelio Merchán-UPS.....	221
4.3.3. Análisis de resultados de CONTINENTAL GENERAL TIRE ANDINA.....	222
4.3.4. Análisis de resultados de la UPV.....	223
<i>CAPÍTULO V.....</i>	<i>225</i>
5.1. CONCLUSIONES.....	225
5.2. RECOMENDACIONES.....	226
<i>BIBLIOGRAFÍA.....</i>	<i>228</i>
<i>ANEXOS.....</i>	<i>230</i>

ÍNDICE DE FIGURAS

Figura 1. Proceso KDD (Knowledge Discovery in Databases) [3].	3
Figura 2. Gráfico 2D de una serie de datos generados aleatoriamente.	5
Figura 3. Base de Datos aplicando el análisis cluster.	5
Figura 4. Perfil de consumo de energía eléctrica.	6
Figura 5. Perfiles de carga segmentados por días de la semana.	6
Figura 6. Clasificación de los tipos de aprendizaje automático.	8
Figura 7. Clasificación de las técnicas de clustering.	10
Figura 8. Ejemplo de hiperplano de separación [10].	17
Figura 9. Ejemplo de MSV sin supervisión para la detección de outliers.	18
Figura 10. Polígonos de frecuencia.	20
Figura 11. Comparación de las funciones de densidad de probabilidad ajustadas a un histograma.	22
Figura 12. Clusters formados a partir de datos aleatorios.	23
Figura 13. Valores de silueta para cada uno de los clusters.	23
Figura 14. Objetos del conjunto de datos X.	24
Figura 15. Objetos vinculados en grupos según la jerarquía de clusters.	25
Figura 16. Dendrograma de los objetos según la jerarquía de clusters.	26
Figura 17. Flujograma del Programa desarrollado en al Software MATLAB.	29
Figura 18. Flujograma del Programa desarrollado en al Software MATLAB.	30
Figura 19. Curvas de carga - E.E. Ambato [21].	33
Figura 20. Curvas de carga - E.E. Centrosur [21].	33
Figura 21. Curvas de carga - CNEL-Sucumbíos [21].	34
Figura 22. Curvas de Carga - E.E. Quito [21].	34
Figura 23. Curvas de Carga – Eléctrica de Guayaquil [21].	34
Figura 24. Perfiles de Carga de Energía Eléctrica (Lunes).	77
Figura 25. Perfiles de Carga de Energía Eléctrica (Martes).	78
Figura 26. Perfiles de Carga de Energía Eléctrica (Miércoles).	78
Figura 27. Perfiles de Carga de Energía Eléctrica (Jueves).	78
Figura 28. Perfiles de Carga de Energía Eléctrica (Viernes).	79
Figura 29. Perfiles de Carga de Energía Eléctrica (Sábados).	79
Figura 30. Perfiles de Carga de Energía Eléctrica (Domingos).	79
Figura 31. Factores de Forma F1 vs F2.	80
Figura 32. Factores de Forma F3 vs F4 vs F5.	81
Figura 33. Factores de Forma F6 vs F7 vs F8.	82
Figura 34. Factores de Forma F9 vs F10 vs F11.	82
Figura 35. Factores de Forma F12 vs F13 vs F14.	82
Figura 36. Factores de Forma F15 vs F16 vs F17.	83
Figura 37. Factores de Forma F18 vs F19 vs F20.	83
Figura 38. Factores de Forma F21 vs F22 vs F23.	83
Figura 39. Evaluación según el valor promedio de Silueta del cluster k-Means.	84
Figura 40. Tiempos de ejecución según el número de cluster k-Means.	85
Figura 41. Tiempos de ejecución según el número de cluster k-Medoids.	85
Figura 42. Gráfica de Silueta con el número de clusters óptimo.	86
Figura 43. Segmentación de los datos en dos clusters.	92
Figura 44. Segmentación de los datos en dos clusters.	93
Figura 45. Segmentación de los datos en dos clusters.	94
Figura 46. Segmentación de los datos en dos clusters.	94
Figura 47. Segmentación de los datos en dos clusters.	94
Figura 48. Segmentación de los datos en dos clusters.	95

Figura 49. Segmentación de los datos en dos clusters.	95
Figura 50. Segmentación de los datos en dos clusters.	95
Figura 51. Histograma con las distancias al centroide del cluster 1.	98
Figura 52. Histograma con las distancias al centroide del cluster 2.	98
Figura 53. FDP de las distancias al centroide del cluster 1.....	99
Figura 54. FDP de las distancias al centroide del cluster 2.....	99
Figura 55. FDAI de las distancias al centroide del cluster 1.....	99
Figura 56. FDAI de las distancias al centroide del cluster 2.....	100
Figura 57. Histograma de los valores de Silueta del cluster 1.	100
Figura 58. Histograma de los valores de Silueta del cluster 2.	100
Figura 59. FDP de los valores de silueta del cluster 1.	101
Figura 60. FDP de los valores de silueta del cluster 2.	101
Figura 61. FDAI de los valores de silueta del cluster 1.	101
Figura 62. FDAI de los valores de silueta del cluster 2.	102
Figura 63. PCEE del cluster 1 (Perfiles sin outliers, 77 - Perfiles con outliers, 10).	103
Figura 64. PCEE del cluster 2 (Perfiles sin outliers, 28 - Perfiles con outliers, 11).	104
Figura 65. Patrón de consumo de PCEE diarios del cluster 1.....	105
Figura 66. Patrón de consumo de PCEE diarios del cluster 2.....	106
Figura 67. Perfiles de Carga de Energía Eléctrica (Lunes).	108
Figura 68. Perfiles de Carga de Energía Eléctrica (Martes).	109
Figura 69. Perfiles de Carga de Energía Eléctrica (Miércoles).	109
Figura 70. Perfiles de Carga de Energía Eléctrica (Jueves).	109
Figura 71. Perfiles de Carga de Energía Eléctrica (Viernes).	110
Figura 72. Perfiles de Carga de Energía Eléctrica (Sábados).	110
Figura 73. Perfiles de Carga de Energía Eléctrica (Domingos).	110
Figura 74. Factores de Forma F1 vs F2.....	111
Figura 75. Factores de Forma F3 vs F4 vs F5.....	112
Figura 76. Factores de Forma F6 vs F7 vs F8.....	113
Figura 77. Factores de Forma F9 vs F10 vs F11.....	113
Figura 78. Factores de Forma F12 vs F13 vs F14.....	113
Figura 79. Factores de Forma F15 vs F16 vs F17.....	114
Figura 80. Factores de Forma F18 vs F19 vs F20.....	114
Figura 81. Factores de Forma F21 vs F22 vs F23.....	114
Figura 82. Evaluación según el valor promedio de Silueta del cluster k-Means.	115
Figura 83. Tiempos de ejecución según el número de cluster k-Means.	116
Figura 84. Tiempos de ejecución según el número de cluster k-Medoids.	116
Figura 85. Gráfica de Silueta con el número de clusters óptimo.	117
Figura 86. Segmentación de los datos en dos clusters.	123
Figura 87. Segmentación de los datos en dos clusters.	124
Figura 88. Segmentación de los datos en dos clusters.	125
Figura 89. Segmentación de los datos en dos clusters.	125
Figura 90. Segmentación de los datos en dos clusters.	125
Figura 91. Segmentación de los datos en dos clusters.	126
Figura 92. Segmentación de los datos en dos clusters.	126
Figura 93. Segmentación de los datos en dos clusters.	126
Figura 94. Histograma con las distancias al centroide del cluster 1.	129
Figura 95. Histograma con las distancias al centroide del cluster 2.	129
Figura 96. FDP de las distancias al centroide del cluster 1.....	130
Figura 97. FDP de las distancias al centroide del cluster 2.....	130
Figura 98. FDAI de las distancias al centroide del cluster 1.....	130
Figura 99. FDAI de las distancias al centroide del cluster 2.....	131

Figura 100. Histograma de los valores de Silueta del cluster 1.	131
Figura 101. Histograma de los valores de Silueta del cluster 2.	131
Figura 102. FDP de los valores de silueta del cluster 1.	132
Figura 103. FDP de los valores de silueta del cluster 2.	132
Figura 104. FDAI de los valores de silueta del cluster 1.	132
Figura 105. FDAI de los valores de silueta del cluster 2.	133
Figura 106. PCEE del cluster 1 (Perfiles sin outliers, 36 - Perfiles con outliers, 8).	134
Figura 107. PCEE del cluster 2 (Perfiles sin outliers, 81 - Perfiles con outliers, 10).	135
Figura 108. Patrón de consumo de PCEE diarios del cluster 1.	136
Figura 109. Patrón de consumo de PCEE diarios del cluster 2.	137
Figura 110. Perfiles de Carga de Energía Eléctrica (Lunes).	139
Figura 111. Perfiles de Carga de Energía Eléctrica (Martes).	140
Figura 112. Perfiles de Carga de Energía Eléctrica (Miércoles).	140
Figura 113. Perfiles de Carga de Energía Eléctrica (Jueves).	140
Figura 114. Perfiles de Carga de Energía Eléctrica (Viernes).	141
Figura 115. Perfiles de Carga de Energía Eléctrica (Sábados).	141
Figura 116. Perfiles de Carga de Energía Eléctrica (Domingos).	141
Figura 117. Factores de Forma F1 vs F2.	142
Figura 118. Factores de Forma F3 vs F4 vs F5.	143
Figura 119. Factores de Forma F6 vs F7 vs F8.	144
Figura 120. Factores de Forma F9 vs F10 vs F11.	144
Figura 121. Factores de Forma F12 vs F13 vs F14.	144
Figura 122. Factores de Forma F15 vs F16 vs F17.	145
Figura 123. Factores de Forma F18 vs F19 vs F20.	145
Figura 124. Factores de Forma F21 vs F22 vs F23.	145
Figura 125. Evaluación según el valor promedio de Silueta del cluster k-Means.	146
Figura 126. Tiempos de ejecución según el número de cluster k-Means.	147
Figura 127. Tiempos de ejecución según el número de cluster k-Medoids.	147
Figura 128. Gráfica de Silueta con el número de clusters óptimo.	148
Figura 129. Segmentación de los datos en dos clusters.	159
Figura 130. Segmentación de los datos en dos clusters.	160
Figura 131. Segmentación de los datos en dos clusters.	161
Figura 132. Segmentación de los datos en dos clusters.	161
Figura 133. Segmentación de los datos en dos clusters.	161
Figura 134. Segmentación de los datos en dos clusters.	162
Figura 135. Segmentación de los datos en dos clusters.	162
Figura 136. Segmentación de los datos en dos clusters.	162
Figura 137. Histograma con las distancias al centroide del cluster 1.	165
Figura 138. Histograma con las distancias al centroide del cluster 2.	165
Figura 139. FDP de las distancias al centroide del cluster 1.	166
Figura 140. FDP de las distancias al centroide del cluster 2.	166
Figura 141. FDAI de las distancias al centroide del cluster 1.	166
Figura 142. FDAI de las distancias al centroide del cluster 2.	167
Figura 143. Histograma de los valores de Silueta del cluster 1.	167
Figura 144. Histograma de los valores de Silueta del cluster 2.	167
Figura 145. FDP de los valores de silueta del cluster 1.	168
Figura 146. FDP de los valores de silueta del cluster 1.	168
Figura 147. FDAI de los valores de silueta del cluster 1.	168
Figura 148. FDAI de los valores de silueta del cluster 2.	169
Figura 149. PCEE del cluster 1 (Perfiles sin outliers, 300 - Perfiles con outliers, 37).	170
Figura 150. PCEE del cluster 2 (Perfiles con outliers, 26).	171

Figura 151. Patrón de consumo de PCEE diarios del cluster 1.....	172
Figura 152. Patrón de consumo de PCEE diarios del cluster 2.....	173
Figura 153. Perfiles de Carga de Energía Eléctrica (Lunes).....	175
Figura 154. Perfiles de Carga de Energía Eléctrica (Martes).....	176
Figura 155. Perfiles de Carga de Energía Eléctrica (Miércoles).....	176
Figura 156. Perfiles de Carga de Energía Eléctrica (Jueves).....	176
Figura 157. Perfiles de Carga de Energía Eléctrica (Viernes).....	177
Figura 158. Perfiles de Carga de Energía Eléctrica (Sábados).....	177
Figura 159. Perfiles de Carga de Energía Eléctrica (Domingos).....	177
Figura 160. Factores de Forma F1 vs F2.....	178
Figura 161. Factores de Forma F3 vs F4 vs F5.....	179
Figura 162. Factores de Forma F6 vs F7 vs F8.....	180
Figura 163. Factores de Forma F9 vs F10 vs F11.....	180
Figura 164. Factores de Forma F12 vs F13 vs F14.....	180
Figura 165. Factores de Forma F15 vs F16 vs F17.....	181
Figura 166. Factores de Forma F18 vs F19 vs F20.....	181
Figura 167. Factores de Forma F21 vs F22 vs F23.....	181
Figura 168. Evaluación según el valor promedio de Silueta del cluster k-Means.....	182
Figura 169. Tiempos de ejecución según el número de cluster k-Means.....	183
Figura 170. Tiempos de ejecución según el número de cluster k-Medoids.....	183
Figura 171. Gráfica de Silueta con el número de clusters óptimo.....	184
Figura 172. Segmentación de los datos en tres clusters.....	194
Figura 173. Segmentación de los datos en tres clusters.....	195
Figura 174. Segmentación de los datos en tres clusters.....	196
Figura 175. Segmentación de los datos en tres clusters.....	196
Figura 176. Segmentación de los datos en tres clusters.....	196
Figura 177. Segmentación de los datos en tres clusters.....	197
Figura 178. Segmentación de los datos en tres clusters.....	197
Figura 179. Segmentación de los datos en tres clusters.....	197
Figura 180. Histograma con las distancias al centroide del cluster 1.....	201
Figura 181. Histograma con las distancias al centroide del cluster 2.....	202
Figura 182. Histograma con las distancias al centroide del cluster 3.....	202
Figura 183. FDP de las distancias al centroide del cluster 1.....	202
Figura 184. FDP de las distancias al centroide del cluster 2.....	203
Figura 185. FDP de las distancias al centroide del cluster 3.....	203
Figura 186. FDAI de las distancias al centroide del cluster 1.....	203
Figura 187. FDAI de las distancias al centroide del cluster 2.....	204
Figura 188. FDAI de las distancias al centroide del cluster 3.....	204
Figura 189. Histograma de los valores de Silueta del cluster 1.....	204
Figura 190. Histograma de los valores de Silueta del cluster 2.....	205
Figura 191. Histograma de los valores de Silueta del cluster 3.....	205
Figura 192. FDP de los valores de silueta del cluster 1.....	205
Figura 193. FDP de los valores de silueta del cluster 2.....	206
Figura 194. FDP de los valores de silueta del cluster 3.....	206
Figura 195. FDAI de los valores de silueta del cluster 1.....	206
Figura 196. FDAI de los valores de silueta del cluster 2.....	207
Figura 197. FDAI de los valores de silueta del cluster 3.....	207
Figura 198. PCEE del cluster 1 (Perfiles sin outliers, 132 - Perfiles con outliers, 7).....	208
Figura 199. PCEE del cluster 2 (Perfiles sin outliers, 113 - Perfiles con outliers, 20).....	209
Figura 200. PCEE del cluster 3 (Perfiles sin outliers, 53 - Perfiles con outliers, 7).....	210
Figura 201. Patrón de consumo de PCEE diarios del cluster 1.....	211

Figura 202. Patrón de consumo de PCEE diarios del cluster 2.....	212
Figura 203. Patrón de consumo de PCEE diarios del cluster 3.....	213

ÍNDICE DE TABLAS

Tabla 1. Medidas de distancia en MATLAB [8].....	13
Tabla 2. Funciones por Kernel [12].	21
Tabla 3. Rangos Horarios para la Clasificación de las Mediciones de Energía Eléctrica.....	34
Tabla 4. Factores de Forma Normalizados para la Caracterización de las Curvas de Carga diarias.	38
Tabla 5. Asociación de los Factores de Forma normalizados para su representación mediante gráficas.	49
Tabla 6. Valores promedios de las Siluetas para el número de clusters analizados.....	84
Tabla 7. Tiempos de Ejecución para el número de clusters analizados	85
Tabla 8. Indicador del número de cluster al que pertenece cada PCEE.....	87
Tabla 9. Centroides para cada cluster	88
Tabla 10. Cuadro de distancias de cada Factor de Forma al centroide de cada cluster	88
Tabla 11. Valores promedios de las Siluetas para el número de clusters analizados.....	115
Tabla 12. Tiempos de ejecución para el número de clusters analizados.....	116
Tabla 13. Indicador del número de cluster de cada PCEE.....	118
Tabla 14. Centroides para cada cluster	119
Tabla 15. Cuadro de distancias de cada Factor de Forma al centroide de cada cluster	119
Tabla 16. Valores promedios de las Siluetas para el número de clusters analizados.....	146
Tabla 17. Tiempos de Ejecución para el número de clusters analizados	147
Tabla 18. Indicador del cluster al que pertenece cada PCEE.....	149
Tabla 19. Centroides para cada cluster	150
Tabla 20. Cuadro de distancias de cada Factor de Forma al centroide de cada cluster	150
Tabla 21. Valores promedios de las Siluetas para el número de clusters analizados.....	182
Tabla 22. Tiempos de Ejecución para el número de clusters analizados	183
Tabla 23. Indicador del número de cluster al que pertenece cada PCEE.....	185
Tabla 24. Centroides para cada cluster	186
Tabla 25. Cuadro de distancias de cada Factor de Forma al centroide de cada cluster	186
Tabla 26. Centroides de los Clusters generados de la base de datos de la acometida principal de la UPS.	214
Tabla 27. Instancias agrupadas para la base de datos de la acometida principal de la UPS en cada cluster.	215
Tabla 28. Centroides de los Clusters generados de la base de datos del edificio CORNELIO MERCHÁN de la UPS.	216
Tabla 29. Instancias agrupadas para la base de datos del edificio Cornelio Merchán de la UPS en cada cluster.	217
Tabla 30. Centroides de los Clusters generados de la base de datos de CONTINENTAL GENERAL TIRE ANDINA C.A.	218
Tabla 31. Instancias agrupadas para la base de datos del edificio Cornelio Merchán de la UPS en cada cluster.	218
Tabla 32. Centroides de los Clusters generados de la base de datos de la UPV.....	219
Tabla 33. Instancias agrupadas para la base de datos del edificio Cornelio Merchán de la UPS en cada cluster.	220
Tabla 34. Centros de clusters finales de la acometida principal de la UPS.	220
Tabla 35. Número de casos en cada cluster.	221
Tabla 36. Centros de clusters finales de la base de datos del Edificio Cornelio Merchán.....	221
Tabla 37. Número de casos en cada cluster.	222
Tabla 38. Centros de clusters finales de la base de datos de CONTINENTAL GENERAL TIRE ANDINA.	222
Tabla 39. Número de casos en cada cluster.	223

Tabla 40. Centros de clusters finales de la base de datos de la UPV.223
Tabla 41. Número de casos en cada cluster.224

ÍNDICE DE ECUACIONES

Ecuación 1 – Distancia Euclidiana.....	13
Ecuación 2 – Distancia Manzana de Ciudad.....	13
Ecuación 3 – Distancia de Coseno.....	14
Ecuación 4 – Distancia de Correlación.....	14
Ecuación 5 – Distancia de Hamming.....	14
Ecuación 6 – Error absoluto.....	14
Ecuación 7 – Límite de un histograma de barras centradas en X.....	20
Ecuación 8 – Estimador de densidad simple.....	20
Ecuación 9 – Estimador de densidad simple con función ponderal $S^{(*)}$	21
Ecuación 10 – Estimador de densidad simple con función ponderal $K^{(*)}$	21
Ecuación 11 – Función por Kernel Epanechnikov.....	21
Ecuación 12 – Función por Kernel Bponderado.....	21
Ecuación 13 – Función por Kernel Triangular.....	21
Ecuación 14 – Función por Kernel Gaussiano.....	21
Ecuación 15 – Función por Kernel Rectangular.....	21
Ecuación 16 – Selección de amplitud de intervalo en histogramas.....	22
Ecuación 17 – Amplitud de ventana para estimadores de densidad por kernel.....	22
Ecuación 18 – Transformación de energía eléctrica a potencia eléctrica.....	36
Ecuación 19 – Transformación de energía eléctrica a potencia eléctrica.....	36
Ecuación 20 – Factor de Carga (F1).....	38
Ecuación 21 – Coeficiente de no uniformidad (F2).....	38
Ecuación 22 – Impacto promedio en la mañana (F3).....	38
Ecuación 23 – Impacto máximo en la mañana (F4).....	38
Ecuación 24 – Impacto mínimo en la mañana (F5).....	38
Ecuación 25 – Impacto promedio del período de almuerzo (F6).....	39
Ecuación 26 – Impacto máximo del período de almuerzo (F7).....	39
Ecuación 27 – Impacto mínimo del período de almuerzo (F8).....	39
Ecuación 28 – Impacto promedio de la tarde (F9).....	39
Ecuación 29 – Impacto máximo de la tarde (F10).....	39
Ecuación 30 – Impacto mínimo de la tarde (F11).....	39
Ecuación 31 – Impacto promedio del horario matutino y vespertino (F12).....	39
Ecuación 32 – Impacto máximo del horario matutino y vespertino (F13).....	39
Ecuación 33 – Impacto mínimo del horario matutino y vespertino (F14).....	39
Ecuación 34 – Impacto promedio del atardecer (F15).....	39
Ecuación 35 – Impacto máximo del atardecer (F16).....	39
Ecuación 36 – Impacto mínimo del atardecer (F17).....	39
Ecuación 37 – Impacto promedio de la noche (F18).....	39
Ecuación 38 – Impacto máximo de la noche (F19).....	39
Ecuación 39 – Impacto mínimo de la noche (F20).....	39
Ecuación 40 – Impacto máximo del día completo (F21).....	40
Ecuación 41 – Impacto promedio del día completo (F22).....	40
Ecuación 42 – Impacto mínimo del día completo (F23).....	40
Ecuación 43 – Regla de Sturges.....	50
Ecuación 44 – Distancia Euclidiana.....	52
Ecuación 45 – Probabilidad de Selección del Centroides.....	53
Ecuación 46 – Probabilidad de Selección aleatoria del Centroides.....	53
Ecuación 47 – Valor de Silueta.....	54
Ecuación 48 – Valor Promedio de las Distancias Euclidianas.....	61
Ecuación 49 – Desviación Estándar de las Distancias Euclidianas.....	61

Ecuación 50 – Valor Promedio de los Valores de Silueta	62
Ecuación 51 – Desviación Estándar de los Valores de Silueta	63

CAPÍTULO I

1.1. Minería de datos

1.1.1. Introducción a la minería de datos

Cada vez es más difícil ignorar la importancia que tienen los clientes dentro del enfoque de desarrollo de estrategias de mercado para atraer su interés en nuevos productos o servicios.

El cliente de energía eléctrica forma parte de este enfoque, ya que, en los últimos años, ha existido un interés creciente en la participación de este tipo de cliente como objeto central en el negocio de la electricidad.

El progreso en el planteamiento de mecanismos más minuciosos para involucrar al lado de la demanda también pasa por la evolución del análisis de datos, con lo que el operador puede obtener un mayor conocimiento sobre la minería de datos [1].

Los patrones de variación de las cantidades que reflejan el comportamiento de los clientes tienen que ser interpretados y analizados, sintetizar y utilizar las características subyacentes para predecir las tendencias esperadas sobre la base de condiciones externas dadas [1].

Hoy en día la cantidad de datos almacenados puede llegar a exceder los límites de las posibilidades de analizarlos sin el uso de técnicas automatizadas. Es decir, llega un momento donde se dispone de tantos datos y es ahí en donde se vuelve difícil obtener algún tipo de información útil, es por ello que se recurre a la **Minería de Datos (Data Mining)** la cual se refiere a un proceso que proporciona información útil usando técnicas de aprendizaje inteligente que analizan los datos mediante procesos automatizados, y la información obtenida de este análisis debe ser de utilidad para procesos posteriores de estudio.

1.1.2. Definición

A la Minería de Datos se le puede proporcionar varias definiciones que a lo largo del tiempo se han desarrollado, entre las más aceptadas están las siguientes [2]:

- “Minería de Datos es el proceso de extracción y refinamiento de conocimiento útil desde grandes bases de datos”.
- “Minería de Datos es el proceso de extracción de información previamente desconocida, válida y procesable desde grandes bases de datos para luego ser utilizada en la toma de decisiones”.
- “Minería de Datos es la exploración y análisis, a través de medios automáticos y semiautomáticos, de grandes cantidades de datos con el fin de descubrir patrones y reglas significativos”.
- “Minería de Datos es el proceso de planteamiento de distintas consultas y extracción de información útil, patrones y tendencias previamente desconocidas desde grandes cantidades de datos posiblemente almacenados en bases de datos”.
- “Minería de Datos es el proceso de descubrir modelos en los datos”.

La Minería de Datos es un proceso de desarrollo y análisis de técnicas para extraer información útil del conjunto de datos que se dispone [3].

La Minería de Datos plantea de una forma general el análisis e interpretación de los resultados de la información como método para obtener una interpretación de la misma, este método ha sido usado por muchos años en varios campos de la ingeniería, medicina, marketing, etc., para encontrar y analizar tendencias, sus cambios y afectaciones dentro de cada área y posteriormente tomar decisiones en base a esta información construyendo una gestión más adecuada en lo que se refiere a prestación de servicios.

A menudo las bases de datos pueden incrementar su tamaño exponencialmente tanto en información como en los atributos de las mismas por lo que realizar una Minería de Datos manualmente sería algo tedioso y poco eficiente para los propósitos requeridos de alguna área. Es por esto que a través del tiempo se ha visto la necesidad de escalar un peldaño más en el proceso de extracción de información útil de una base de datos, es decir, automatizar total o parcialmente el análisis de los objetos con el objetivo de proveer servicios más eficientes en base a un estudio detallado de esta información encontrada con la aplicación de la Minería de Datos.

1.1.3. Proceso de la minería de datos

Para que el proceso sea efectivo, se debe tener mucho cuidado con los objetivos que se desean alcanzar, además de tener una base de datos bien definida para el análisis [4].

1.1.3.1. Requerimientos del análisis

Como requerimientos para la Minería de Datos se debe tener bien definido que se busca con el proceso, es decir si se tienen claros cuales son los objetivos del estudio, más adelante será más fácil analizar los resultados que se obtengan.

• Selección de información

Lo siguiente es tener presente cuál es la mejor fuente para obtener la información, es decir, es necesario encontrar una base de datos donde se encuentre la información que se necesita para la Minería de Datos, algunas veces las empresas poseen información que no es útil para el estudio que se desea realizar, así que para ello los requerimientos deben estar bien definidos y saber qué tipo de información se necesita y bajo qué atributos se la debe adquirir.

• Preparación de los datos

Por lo general, las empresas realizan un almacenamiento de datos que ocupan para realizar sus estudios internos, así que, para realizar una Minería de Datos, se debe analizar si la información proporcionada por la empresa es de utilidad para el estudio, evitando que más adelante se deba lidiar con problemas como identificación de la información, información incompleta o ambigüedad en los datos.

• Validación e implementación de la minería de datos

En este punto, una vez que se tiene toda la información arreglada y sin inconvenientes, se debe elegir las técnicas más prometedoras, escogiendo de entre las que más se ajustan a nuestra base de datos. Para cada técnica se puede observar e interpretar los resultados obtenidos. Es decir, una vez que se

haya elegido un método de Minería de Datos, este puede contener el desarrollo de algoritmos necesarios para realizar la evaluación automática de la información.

- **Visualización y análisis de resultados**

En esta etapa se procede a revisar los resultados obtenidos debiendo presentarlos de una manera que sean interpretables en caso de que se desee hacer uso de los mismos.

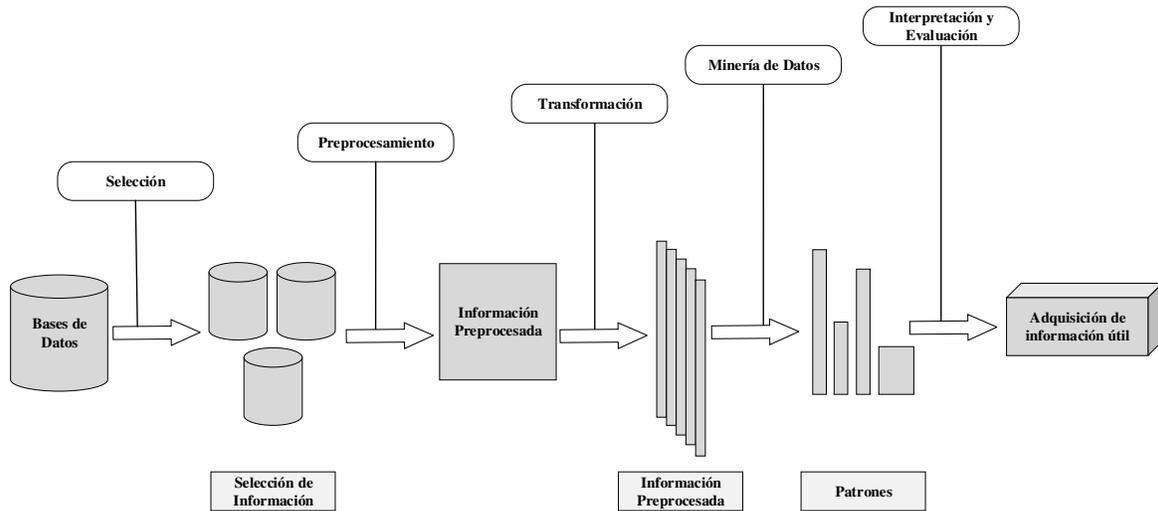


Figura 1. Proceso KDD (Knowledge Discovery in Databases) [3].

1.1.4. Funcionalidades de la minería de datos

Las funcionalidades de una Minería de Datos son usadas para especificar el tipo de patrón a encontrar en una tarea específica. Por lo general, las tareas de la Minería de Datos se clasifican en: **descriptivas**, que caracterizan las propiedades generales de los datos en la base de datos; y las **predictivas**, que realizan inferencias en los datos actuales con el fin de realizar predicciones [5].

Algunas de las funcionalidades de la minería de datos y tipos de patrones que se pueden encontrar se describen a continuación.

1.1.4.1. Concepto/Clase: Caracterización y Discriminación

Los datos pueden ser asociados a clases o conceptos. Como un ejemplo práctico se puede suponer que un almacén electrónico, hay productos de distintas clases como computadoras e impresoras; también se puede decir que existen conceptos de clientes como mayoristas y minoristas. Un tema de interés y que sería de gran utilidad es describir el comportamiento individual de clases y conceptos de forma concisa, resumida y en términos precisos [6].

Estas descripciones pueden derivar en [6]:

- **Caracterización:** Resumen de las características generales de una clase objetivo.
- **Discriminación:** Es una comparación entre las características generales de los objetos de la clase objetivo y las características generales de los objetos de una o varias clases comparables.

1.1.4.2. Patrones frecuentes, asociación y correlación

Los patrones frecuentes son aquellos que aparecen a menudo en el conjunto de datos. Existen tipos de patrones como los conjuntos de ítems, sub secuencias y sub estructuras. Un conjunto de ítems se refiere a ítems que frecuentemente aparecen juntos en un conjunto de datos. Una sub secuencia frecuente es un patrón como aquel en donde los clientes tienden a comprar primero un determinado producto seguido por el complemento de ese producto y así sucesivamente. Una sub estructura se puede referir a diferentes tipos de estructuras como por ejemplo gráficos, árboles, etc., los cuales pueden ser combinados con conjuntos de ítems o sub secuencias. La búsqueda de patrones frecuentes lleva al descubrimiento de asociaciones y correlaciones interesantes dentro del conjunto de datos [6].

1.1.4.3. Clasificación y regresión para análisis predictivo

Clasificación es el proceso de encontrar un modelo que describe y distingue las clases de datos o conceptos. El modelo es usado para predecir la clase de los objetos la cual es desconocida, ya que este modelo derivado está basado en el análisis de un conjunto de datos de entrenamiento, que pueden ser aquellos de los cuales su clase es conocida.

Mientras que la clasificación predice etiquetas categóricas, la regresión modela funciones continuas de las que se puede extraer un valor concreto. Este quiere decir que la regresión es usada para predecir valores numéricos perdidos o no disponibles en vez de una etiqueta de clase. El análisis de regresión es una metodología estadística que se usa con frecuencia para la predicción numérica. La regresión también abarca la identificación de las tendencias de una distribución desconocida en base a los datos que se encuentran disponibles [6].

1.1.4.4. Análisis de clusters

A diferencia de la clasificación y regresión que analizan objetos de datos que pertenecen a una clase determinada, el análisis de clusters lo hace sin conocer las clases. El análisis cluster puede ser utilizado para generar las etiquetas de clases del conjunto de datos. Los objetos se agrupan según el principio de maximizar la semejanza intracase y minimizar la similitud intercase. Cada cluster puede ser considerado como una clase de objetos y posteriormente derivar otras [6].

Más adelante se realiza un estudio más detallado a este tipo de análisis ya que será el método de trabajo para el presente proyecto de investigación.

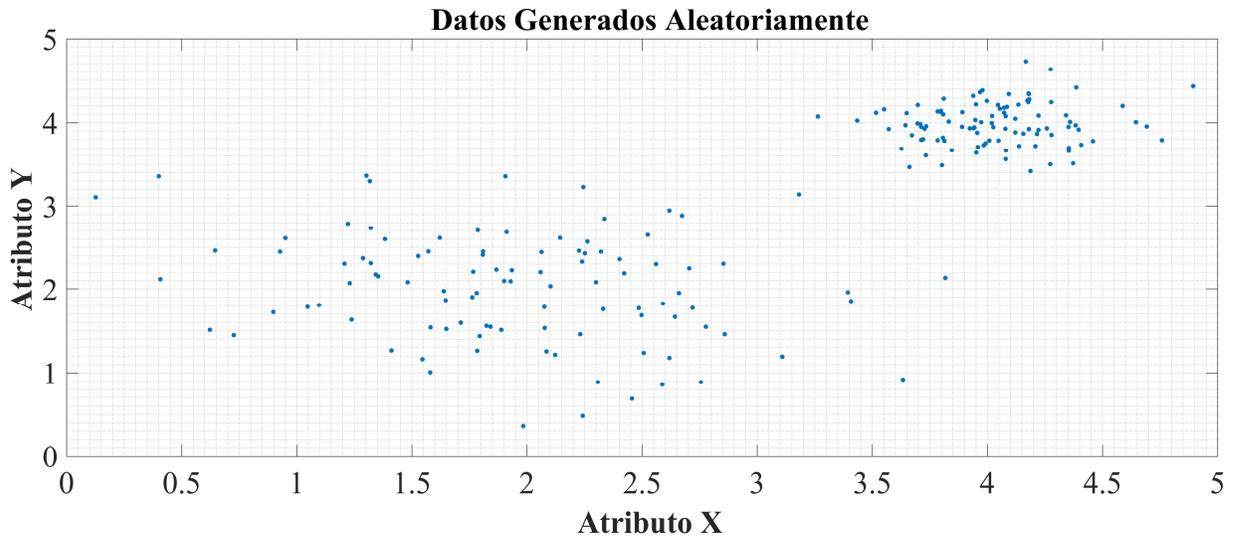


Figura 2. Gráfico 2D de una serie de datos generados aleatoriamente.

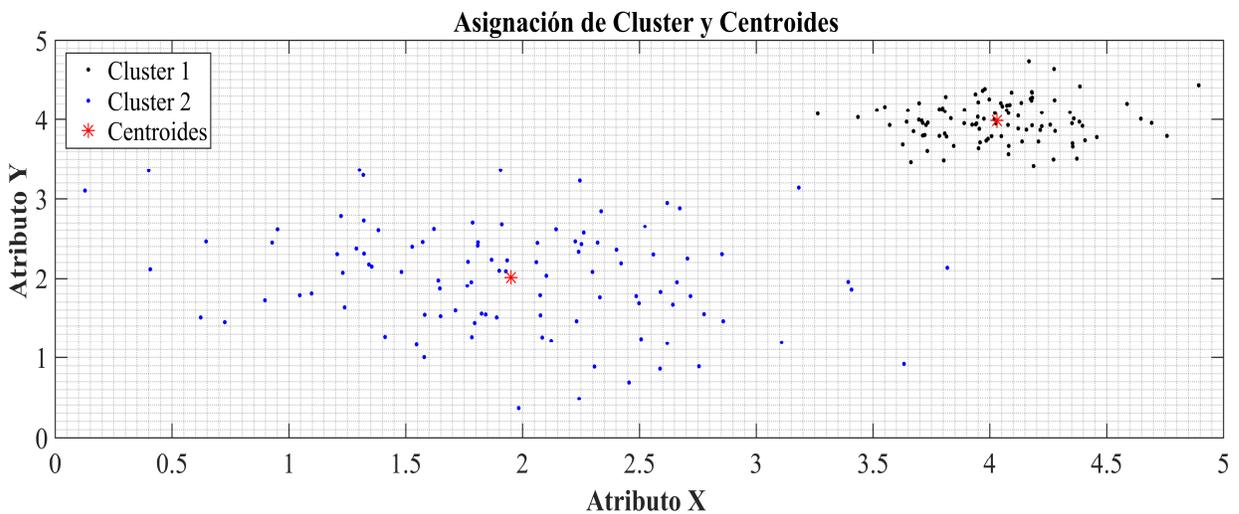


Figura 3. Base de Datos aplicando el análisis cluster.

1.1.5. Aplicaciones de la Minería de Datos

La Minería de Datos se ha utilizado en diferentes aplicaciones entre las cuales las más importantes son:

1.1.5.1. Perfil de consumo de energía eléctrica de clientes

El perfil de **CEE (Consumo de Energía Eléctrica)** del cliente es el proceso de usar información relevante y disponible para describir las características de un grupo de clientes e identificar discriminadores y controladores de otros clientes para sus decisiones de compra [4].

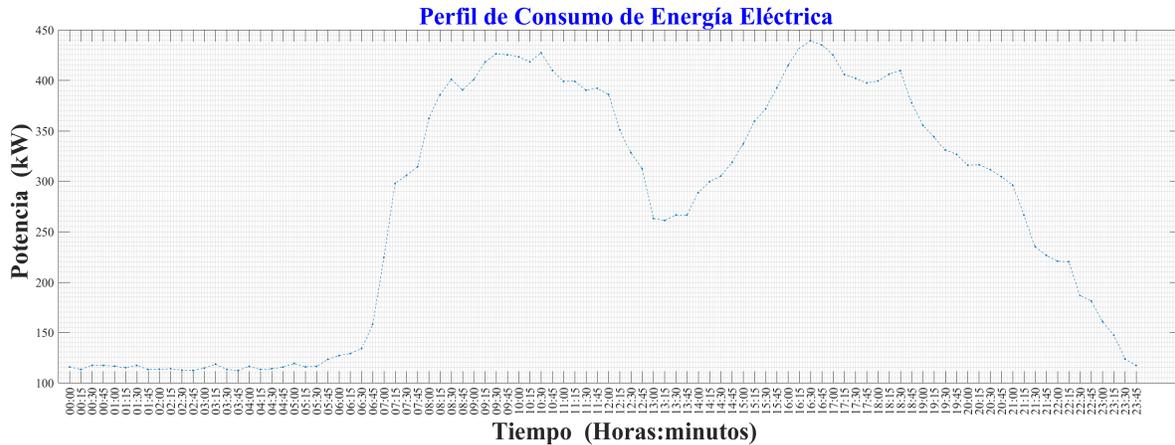


Figura 4. Perfil de consumo de energía eléctrica.

1.1.5.2. Segmentación de clientes

La segmentación es una forma de tener una interacción más específica con los clientes. Es decir, centrarse en un determinado grupo para su análisis. La segmentación es el proceso en el cual significa colocar a la población en segmentos de acuerdo con su afinidad o características similares.

La segmentación de clientes es un paso de preparación para clasificar a cada cliente de acuerdo con los grupos de clientes que se han definido [7].

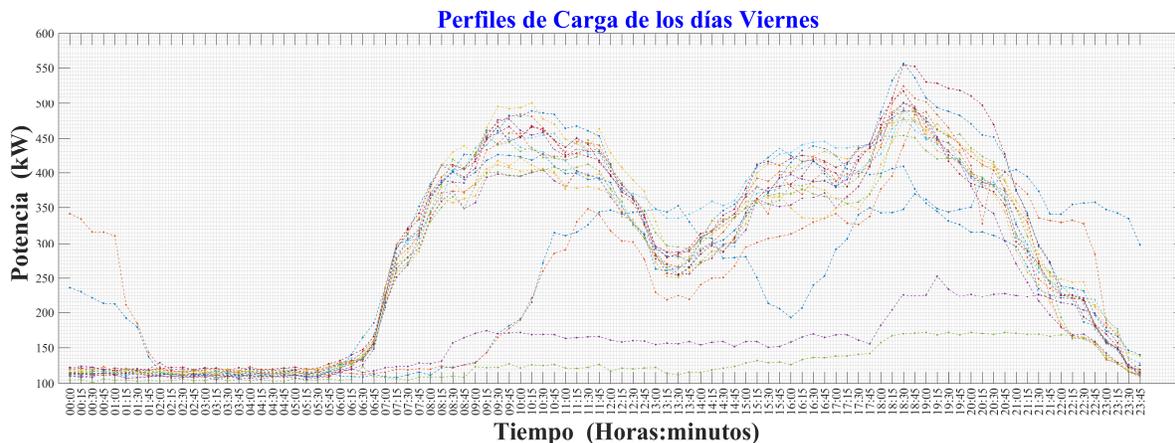


Figura 5. Perfiles de carga segmentados por días de la semana.

1.1.5.3. Identificación de outliers y detección de fraudes

Una de las definiciones más citadas en investigaciones anteriores es la enunciada por David Hawkins en sus monografías sobre estadística y probabilidad aplicadas del año 1980: “*Un outlier es una observación que se desvía tanto de otras observaciones que despierta la sospecha de haber sido generada por un mecanismo diferente*” [5].

Es por ello, que la aplicación de algoritmos sofisticados y automatizados son necesarios para la detección de estos **VA (Valores Atípicos)** que pueden ser un caso de estudio mucho más interesante que las muestras comunes. Con la identificación de outliers se puede beneficiar a aplicaciones en campos científicos, comerciales o estratégicos, como parte de un sistema de detección de actividades inusuales consideradas típicamente como fraude [5].

CAPÍTULO II

2.1. Análisis cluster y detección de outliers

2.1.1. Aprendizaje automático

Existen diferentes tipos de aprendizaje, uno de ellos es el aprendizaje inductivo, que crea modelos de conceptos a partir de generalizar ejemplos simples. Como ejemplo de éste se tiene el Aprendizaje Automático, o Machine Learning, que es una rama de la inteligencia artificial encargada de desarrollar programas capaces de aprender, es decir, de crear programas capaces de generalizar comportamientos a partir de una información no estructurada dada como ejemplos.

Se puede clasificar los algoritmos del aprendizaje automático en:

- aprendizaje inductivo supervisado, y
- aprendizaje inductivo no supervisado

La diferencia entre ambos tipos de aprendizaje es que, en el aprendizaje inductivo supervisado, las muestras se encuentran etiquetadas. Estas etiquetas podrían ser un valor numérico, y el objetivo del algoritmo será encontrar un conjunto de características a partir de las muestras, las cuales se denominan atributos, que permitan predecir con precisión el valor de etiqueta correcto de cada objeto del conjunto de entrenamiento, produciendo así una función que tenga como salida un valor real, como en los problemas de regresión, o una etiqueta de clase, como en los problemas de clasificación.

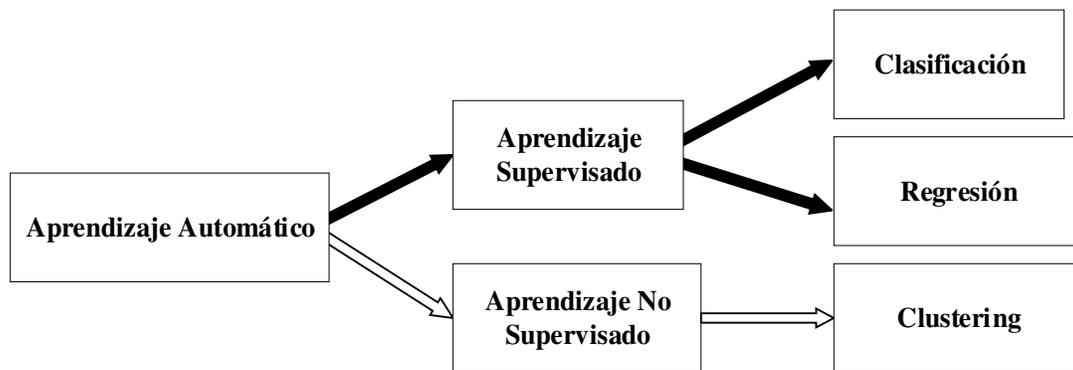


Figura 6. Clasificación de los tipos de aprendizaje automático.

2.1.2. Definición y Objetivo del Análisis Cluster

Como parte de una definición general, [6, p. 444] “*el análisis cluster o también denominado simplemente como clustering, es el proceso de dividir un conjunto de datos en subconjuntos.*” El clustering puede ser tratado como una **segmentación de la información** ya que separa grandes cantidades de conjuntos de información en grupos de acuerdo a su similitud [6].

El objetivo del análisis cluster es determinar si los datos caen naturalmente en grupos significativos con pequeñas variaciones dentro del grupo y con grandes variaciones entre grupos, en base a este concepto cada cluster puede ser tratado como una clase implícita [4, p. 219].

Como una función de la Minería de Datos, el análisis cluster puede ser utilizado como una herramienta independiente para obtener información sobre la distribución de los datos, con el objetivo de observar las características de cada uno de los clusters obtenidos y enfocar el análisis en un conjunto de clusters de interés. En otro contexto puede servir como un precedente en el uso de otro tipo de procesos (algoritmos) como caracterización, selección en base a los atributos de los subconjuntos, y clasificación, el cual operará en los clusters detectados y los atributos o características seleccionadas. Otra de las aplicaciones que se le puede dar al clustering es para la **detección outlier**, donde los outliers (valores alejados de cualquier cluster) pueden ser casos más interesantes de estudiar que cualquier otro caso común de análisis [6].

2.1.3. Requerimientos para el análisis cluster

En cuanto a los requerimientos, existen algunos que son necesarios conocerlos y estudiarlos de una manera en que se pueda realizar un análisis cluster adecuadamente [6]:

- **Escalabilidad:** Se refiere a la propiedad de aumentar características de un determinado sistema (tamaño, capacidad de trabajo) sin que esto incida en la eficiencia de funcionamiento del mismo. Existen algunos algoritmos para realizar un proceso de clustering que trabajan muy bien en conjuntos de datos pequeños, sin embargo, en una base de datos que contiene millones de objetos que pueden ser estudiados, si no se sabe elegir bien el algoritmo a utilizar los resultados pueden llegar a ser imprecisos, así que en un proceso de clustering para una base de datos grande se debe tener en cuenta que se deben utilizar algoritmos que sean altamente escalables.
- **Habilidad para trabajar con diferentes tipos de atributos:** Existen diversos métodos que se utilizan para realizar el clustering de información solamente numérica. Sin embargo, existen aplicaciones que necesitan realizar un clustering a otros tipos de información ya sea binaria, nominal, ordinaria o una mezcla de estos tipos. Hoy en día, muchas más aplicaciones requieren realizar un clustering para información más compleja como son gráficos, imágenes, etc.
- **Habilidad para descubrir clusters con diferentes formas:** Algunos algoritmos de clustering determinan los clusters basados en las mediciones de la distancia Euclidiana o la distancia de Manhattan. Estos algoritmos tienden a encontrar clusters esféricos con similar tamaño y densidad. Pero los clusters pueden adoptar cualquier forma, así que es importante desarrollar **TC (Técnicas de clustering)** que puedan detectar clusters de diferentes formas.
- **Habilidad para tratar con ruido en la información:** La mayoría de la información contiene outliers y/o información desconocida, errónea o sin sentido. Ya que algunos algoritmos utilizados para el proceso de clustering son sensibles a este tipo de información, pueden devolver información imprecisa y de poca ayuda para los casos de estudio, por esta razón los algoritmos deben ser robustos ante estas dificultades.

- **Agrupamientos incrementales e insensibilidad al orden de los datos de entrada:** En algunas aplicaciones las actualizaciones de la información pueden darse en cualquier momento. Algunos de los algoritmos de clustering no pueden manejar nueva información por lo que se deben reprogramar los algoritmos para que trabajen de manera adecuada. Además, cuando estos algoritmos son sensibles al orden de entrada de la información pueden devolver clusters diferentes a los resultados previamente obtenidos, por lo que los algoritmos deben ser insensibles a la entrada de nueva información.
- **Capacidad de realizar un proceso de clustering a información altamente dimensionada:** Se sabe que la información puede contener numerosas dimensiones o atributos. Hoy en día la mayoría de algoritmos de clustering realizan el proceso a información en dos o tres dimensiones. Sin embargo, encontrar clusters de objetos de datos en un espacio de alta dimensión es un desafío, especialmente teniendo en cuenta que tales datos pueden ser muy escasos y altamente sesgados.
- **Clustering basado en restricciones:** Varias aplicaciones necesitan un proceso de clustering que trabaje bajo varios tipos de restricciones. Una tarea difícil consiste en encontrar grupos de datos con buen comportamiento de agrupación que cumplan las restricciones especificadas.
- **Interpretación y utilidad:** Se desea que los resultados de la agrupación (clustering) sean interpretables, comprensibles y utilizables. Es importante estudiar como una de las funciones de la aplicación puede influir en la selección de funciones y métodos de agrupación.

2.1.4. Introducción a las técnicas de clustering

Existen diferentes métodos para realizar un proceso de clustering, en general los métodos más fundamentales se pueden clasificar dentro de las siguientes categorías [6].

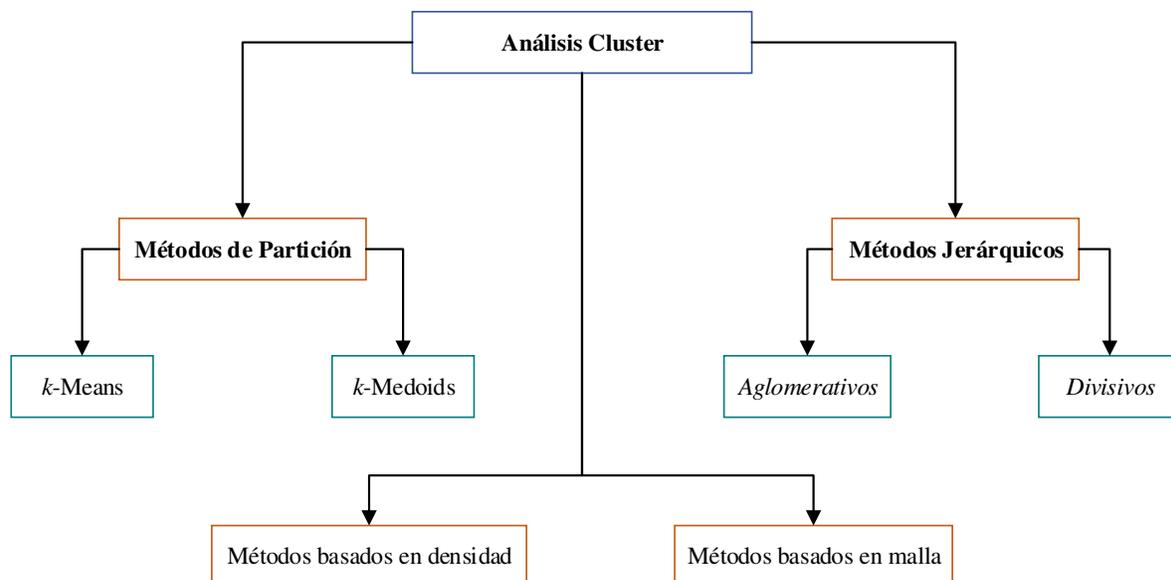


Figura 7. Clasificación de las técnicas de clustering.

2.1.4.1. Métodos de partición

Dado un conjunto de n objetos, este método construye k particiones de datos, donde cada partición representa un cluster y $k \leq n$. Divide la información en k grupos de manera que cada uno contenga al menos un objeto [6].

Los métodos de partición básicos adoptan típicamente la *separación exclusiva de cluster*. Es decir, cada objeto debe pertenecer exactamente a un grupo.

La mayoría de métodos de partición están basados en distancias. Dado k , el número de particiones a obtener, el método crea una partición inicial, después usa una **técnica de reubicación iterativa** la cuál intenta mejorar la partición moviendo objetos de un grupo a otro [6].

El criterio general de una buena partición es que los objetos en un mismo cluster son cercanos o relacionados entre sí con iguales atributos, mientras que los objetos en diferentes clusters están muy alejados o poseen diferentes atributos.

Alcanzar la optimalidad global en el agrupamiento basado en particiones es a menudo computacionalmente difícil, pudiendo necesitar una enumeración exhaustiva de todas las particiones posibles. En su lugar, la mayoría de las aplicaciones adoptan métodos heurísticos populares, como enfoques codiciosos como los algoritmos ***k-Means*** y ***k-Medoids***, que mejoran progresivamente la calidad de la agrupación y se aproximan a un óptimo local. Estos métodos de agrupación heurística funcionan bien para encontrar clusters de forma esférica en bases de datos de tamaño pequeño a mediano. Para encontrar clusters con formas complejas y para conjuntos de datos muy grandes, los métodos basados en particiones necesitan ser extendidos [6].

2.1.4.2. Métodos jerárquicos

Estos métodos crean una descomposición jerárquica del conjunto de objetos de la información dada. Un método jerárquico puede ser clasificado como aglomerativo o divisivo, basado en cómo el método esté formado.

La aproximación aglomerativa, también llamada aproximación ascendente, comienza con cada objeto formando un grupo por separado. A partir de esto, sucesivamente aparecen objetos cercanos unos de otros hasta que todos los grupos (clusters) resulten en uno solo (el mayor nivel jerárquico) [6].

La aproximación divisiva, también llamada como aproximación descendente, comienza con todos los objetos dentro de un mismo cluster. En cada iteración sucesiva cada cluster es separado en clusters más pequeños, hasta que eventualmente cada objeto esté en un cluster específico, o una condición de finalización se mantenga [6].

2.1.4.3. Métodos basados en densidad

La mayoría de métodos de partición agrupan objetos basados en la distancia entre dichos objetos. Otros métodos de clustering han sido desarrollados basados en un concepto de *densidad*. La idea general de estos métodos es hacer crecer un cluster dado hasta que su densidad (número de objetos)

en el grupo exceda algún límite. Estos métodos pueden ser usados para filtrar ruido u outliers y encontrar clusters de formas arbitrarias [6].

Los métodos basados en densidad pueden dividir un conjunto de objetos en múltiples clusters exclusivos, o una jerarquía de clusters. Normalmente, los métodos basados en densidad sólo consideran clusters exclusivos, y no consideran los clusters difusos.

2.1.4.4. Métodos basados en malla (Grid-Based Methods)

Estos métodos cuantifican el espacio de los objetos en un número finito de celdas que forman una estructura mallada. La ventaja principal de este enfoque es su rápido tiempo de procesamiento, que es típicamente independiente del número de objetos de datos y depende únicamente del número de células en cada dimensión en el espacio cuantificado. Y pueden ser combinados con otros métodos de clustering como los basados en densidad y los métodos jerárquicos [6].

2.1.5. Métodos de partición

La versión más simple y fundamental de un análisis cluster es la partición, la cual organiza los objetos de un conjunto en grupos o clusters exclusivos.

2.1.5.1. *k*-Means: Una técnica basada en los centroides

El método *k*-Means en el análisis clustering es una de las técnicas más simples y fáciles de implementar.

Supóngase un conjunto de datos, D , que contiene n objetos en un espacio Euclidiano. Los métodos de partición distribuyen los objetos en D en k clusters, C_1, \dots, C_k , esto es, $C_i \cap C_j = \emptyset$ para $(1 \leq i, j \leq k)$. Conceptualmente, el centroide de un cluster es su punto central. Esta técnica de partición usa el centroide de un cluster, C_i , para representar a dicho cluster [6].

El algoritmo de *k*-Means define el centroide de un cluster con el valor medio de los puntos dentro del cluster. Primero, el algoritmo selecciona de manera randómica k objetos en D , cada uno de los cuales representa la media de un cluster o un centro. Para el resto de objetos, el algoritmo asigna cada uno de ellos al cluster con el cual posee mayor similitud, basados en la distancia Euclidiana entre dichos objetos y el cluster principal.

El algoritmo ***k*-Means** entonces iterativamente mejora la variación intra-cluster. Para cada cluster, calcula la nueva media utilizando los objetos asignados al cluster en la iteración anterior.

A continuación, se reasignan todos los objetos utilizando las medias actualizadas a medida que se agrupan los nuevos centros de clusters. Las iteraciones continúan hasta que la asignación es estable, es decir, los clusters formados en la ronda actual son los mismos que los formados en la ronda anterior [6].

<p>Algoritmo <i>k</i>-Means. El algoritmo <i>k</i>-Means para partición, donde cada centro de los clusters está representado por el valor medio de los objetos en ese cluster.</p>

Entrada:

- k : el número de clusters,
- D : un conjunto de datos de n objetos.

Salida: Un conjunto de clusters.

Método:

- (1) arbitrariamente se elige k objetos de D como los centros de los clusters iniciales;
- (2) **repetir**
- (3) (re) asignar cada objeto al clúster al cual el objeto es el más similar, basado en el valor medio de los objetos del clúster;
- (4) Actualizar los medios del cluster, es decir, calcular el valor medio de los objetos para cada cluster;
- (5) **hasta** que no exista cambios;

2.1.5.2. Introducción al Clustering k -Means en MATLAB

La función **kmeans** de MATLAB divide los datos en k clústeres mutuamente exclusivos y devuelve el índice del clúster al que ha asignado cada observación [8].

A diferencia de los métodos jerárquicos para el análisis clustering, la técnica denominada como k -Means trabaja sobre los datos originales de la información, lo que ayuda a crear un solo nivel de clusters permitiendo que esta técnica sea más adecuada para su implementación cuando se trabaja con grandes cantidades de información [8].

La técnica **k -Means** trata a cada observación como un objeto con una localización en el espacio. Esta técnica encuentra una partición en la cual los objetos que se encuentran dentro de un cluster estén lo más cerca entre sí y lo más lejos posible de los otros objetos que este dentro de otros clusters [8].

Se puede elegir entre cinco medidas diferentes de distancia, lo cual va a depender del tipo de información con la que estemos trabajando.

A continuación, se presenta una tabla donde se explican las diferentes medidas de distancia con las que se puede trabajar.

Tabla 1. Medidas de distancia en MATLAB [8].

Medición de Distancia	Descripción	Fórmula
'sqeuclidean'	La distancia Euclidiana cuadrática. Cada centroide es el medio de todos los puntos en ese cluster.	<i>Ecuación 1 – Distancia Euclidiana</i> $d(x, c) = (x - c)(x - c)'$
'cityblock'	Suma de las diferencias absolutas. Cada centroide es el componente de la mediana de los puntos en ese cluster.	<i>Ecuación 2 – Distancia Manzana de Ciudad</i> $d(x, c) = \sum_{j=1}^p x_j - c_j $

'cosine'	Uno menos el coseno del ángulo incluido entre los puntos (tratados como vectores). Cada centroide es la media de los puntos en ese cluster, después de normalizar esos puntos a la unidad de longitud euclidiana.	<p><i>Ecuación 3 – Distancia de Coseno</i></p> $d(x, c) = 1 - \frac{xc'}{\sqrt{(xx')(cc')}}$
'correlation'	Una menos la correlación muestral entre puntos (tratados como secuencias de valores). Cada centroide es la media en componentes de los puntos en ese cluster, después de centrar y normalizar esos puntos a media cero y desviación estándar igual a uno.	<p><i>Ecuación 4 – Distancia de Correlación</i></p> $d(x, c) = 1 - \frac{(x - \bar{x})(c - \bar{c})'}{\sqrt{(x - \bar{x})(x - \bar{x})'}\sqrt{(c - \bar{c})(c - \bar{c})'}}$
'hamming'	Esta medida es solo adecuada para información binaria. Es la proporción de bits que difieren. Cada centroide es la mediana componente de los puntos en ese cluster.	<p><i>Ecuación 5 – Distancia de Hamming</i></p> $d(x, y) = \frac{1}{p} \sum_{j=1}^p I\{x_j \neq y_j\}$ <p>Donde I es la función indicador.</p>

2.1.5.3. *k*-Medoids: Una técnica representativa basada en los objetos

El algoritmo *k*-Means es sensible a los outliers (valores muy alejados) ya que, si los objetos se encuentran lejos de la mayoría de información, y, por lo tanto, cuando estos valores son asignados a un cluster, pueden distorsionar de manera significativa el valor medio del cluster. Esto, por lo tanto, afecta la asignación de otros valores a los clusters.

La pregunta que surge ante esta problemática es: **¿Cómo se puede modificar el algoritmo de *k*-Means para disminuir su sensibilidad ante los outliers?**

En lugar de tomar el valor medio de los objetos en un cluster como punto de referencia, podemos seleccionar objetos reales para representar los clusters, utilizando un objeto representativo por cada cluster. Cada objeto restante se asigna al cluster cuyo objeto representativo es el más similar. El método de partición se realiza entonces sobre la base del principio de minimizar la suma de las disimilitudes entre cada objeto *p* y su correspondiente objeto representativo.

Se usa el criterio del **error-absoluto**, definido como:

Ecuación 6 – Error absoluto

$$E = \sum_{i=1}^k \sum_{p \in C_i} dist(p, o_i),$$

donde, E es la suma del error absoluto para todos los objetos p en el conjunto de datos, y o_i es el objeto representativo de C_i . Esta es la base para el método ***k-medoids***, la cual agrupa n objetos en k clusters minimizando el error absoluto.

Algoritmo: *k-medoids*. Partición Alrededor de los Objetos Representativos (PAM), un algoritmo *k-medoids* basado en los objetos centrales.

Entrada:

- k : el número de clusters,
- D : un conjunto de datos de n objetos.

Salida: Un conjunto de clusters.

Método:

- (1) arbitrariamente se elige k objetos en D como los objetos representativos iniciales.
- (2) **repetir**
- (3) asignar cada objeto restante al cluster con el objeto representativo más cercano;
- (4) randómicamente se selecciona un objeto no representativo, o_{random} ;
- (5) calcular el costo total, S , de intercambiar el objeto representativo, o_j , con o_{random} ;
- (6) **If** $S < 0$ **then** intercambiar o_j , con o_{random} para formar el nuevo conjunto de k objetos representativos;
- (7) **hasta** que no exista cambios;

2.1.5.4. Introducción a *k-Medoids* en MATLAB

Esta técnica es comúnmente utilizada en dominios de información que requiere una alta robustez a los outliers de los datos, así como a mediciones arbitrarias de distancias, o simplemente porque medias y medianas en los clusters no están bien definidas [9].

Es una técnica muy similar al *k-Means*, es decir, el objetivo de ambos métodos es la de dividir un conjunto de mediciones u observaciones en k clusters. En la técnica denominada como *k-Means* el centro del conjunto está dado por la media de los datos que forman ese cluster, mientras que en la técnica denominada como *k-Medoids* el centro del conjunto es un miembro de ese cluster, denominado como medoid¹ [9].

El algoritmo *k-Medoids* devuelve los centros, los cuales son los datos originales de la información. Esto permite utilizar el algoritmo en las situaciones donde la media de la información no exista en ese conjunto de datos [9].

La diferencia principal entre *k-Medoids* y el *k-Means* es que los centroides proporcionados por *k-means* no deben estar necesariamente en el conjunto de datos. Mientras que la técnica *k-medoids* es útil cuando el centroide no está definido o es difícil de interpretar [9].

¹ En concepto es igual a la media o centroide, pero un medoid (centro) es siempre un punto de la información original de un cluster determinado.

La función de MATLAB **kmedoids** proporciona múltiples iteraciones que minimizan la suma de las distancias de cada objeto del cluster con su centro [9].

Uno de los algoritmos se denomina como partición alrededor de los centros (PAM), el cual trabaja en dos pasos [9]:

1. Paso de construcción (Build-step): Cada uno de los k clusters es asociado a un centro potencial.
2. Paso de intercambio (Swap-step): En cada cluster, cada punto es probado como un potencial centro verificando si la suma de las distancias de ese cluster es la menor usando ese punto como centro. Si eso sucede, ese punto es definido como un nuevo centro. Entonces cada punto es asignado al cluster con el centro más cercano.

2.1.6. Métodos jerárquicos

Estos métodos jerárquicos para análisis cluster, producen una serie de clusters anidados en lugar de un solo grupo de clusters como lo hacen los métodos de partición. Esencialmente los métodos jerárquicos intentan capturar la estructura de los datos construyendo un árbol de clusters. Con este enfoque se permite encontrar clusters en diferentes niveles de granularidad² [4].

Los métodos jerárquicos pueden encontrar dificultades en la selección de los puntos de unión o división. Así que esta decisión es crítica, ya que una vez que un grupo es unido o separado, el proceso en el siguiente paso va a trabajar en los nuevos clusters generados.

2.1.6.1. Comparación entre el Clustering Jerárquico Aglomerativo y Divisivo

Un clustering jerárquico puede ser aglomerativo o divisivo, dependiendo si la descomposición jerárquica es formada de abajo hacia arriba (aglomerativo) o si es de arriba hacia abajo (divisivo) [6].

Un método jerárquico de clustering aglomerativo típicamente comienza dejando que cada objeto sea un cluster, y luego usando una medición de la distancia en cada paso del método, comienza a unir dos clusters que sean cercanos entre sí, lo que reduce iterativamente el número de clusters formando clusters cada vez más grandes hasta que todos los objetos estén en un único cluster o se cumplan ciertas condiciones de terminación (número de clusters deseados) [6].

El cluster único se convierte en la raíz de la jerarquía. Para el paso de fusión, encuentra los dos clusters que están más próximos entre sí (de acuerdo con alguna medida de similitud), y combina los dos para formar un grupo. Debido a que dos clústeres se combinan por iteración, donde cada clúster contiene al menos un objeto, un método aglomerado requiere al menos n iteraciones [6].

Un método jerárquico de clustering divisivo empieza colocando todos los objetos en un mismo cluster, el cual es la raíz de la jerarquía. Luego divide ese cluster en varios subclusters más pequeños, y recursivamente divide dichos clusters en otros más y más pequeños. El proceso de división continua hasta que cada cluster en el nivel más bajo contenga solamente un objeto, o los objetos dentro del cluster sean lo suficientemente similares entre sí [6].

² Es el grado de detalle con el cuál se pueden especificar uno o más clusters.

2.1.7. Máquinas de soporte vectorial para la detección de anomalías

2.1.7.1. Introducción a las máquinas de soporte vectorial

Las **MSV (Máquinas de Soporte Vectorial)** son un modelo para resolver problemas de aprendizaje automático. Trabajan seleccionando un pequeño número de instancias, las cuales son denominadas vectores de soporte, y construyen una función discriminante lineal que las separa a la mayor distancia posible.

2.1.7.2. Principios de las máquinas de soporte vectorial

- **Datos Separables**

Se puede ocupar las MSV cuando la base de datos posee únicamente dos clases, una MSV clasifica la información encontrando el mejor hiperplano que separa una clase de otra. El mejor hiperplano se refiere a aquel que posee el margen mayor entre las dos clases, refiriéndose al margen como la amplitud máxima comprendida entre los ejes paralelos al hiperplano que no poseen puntos en su interior [10].

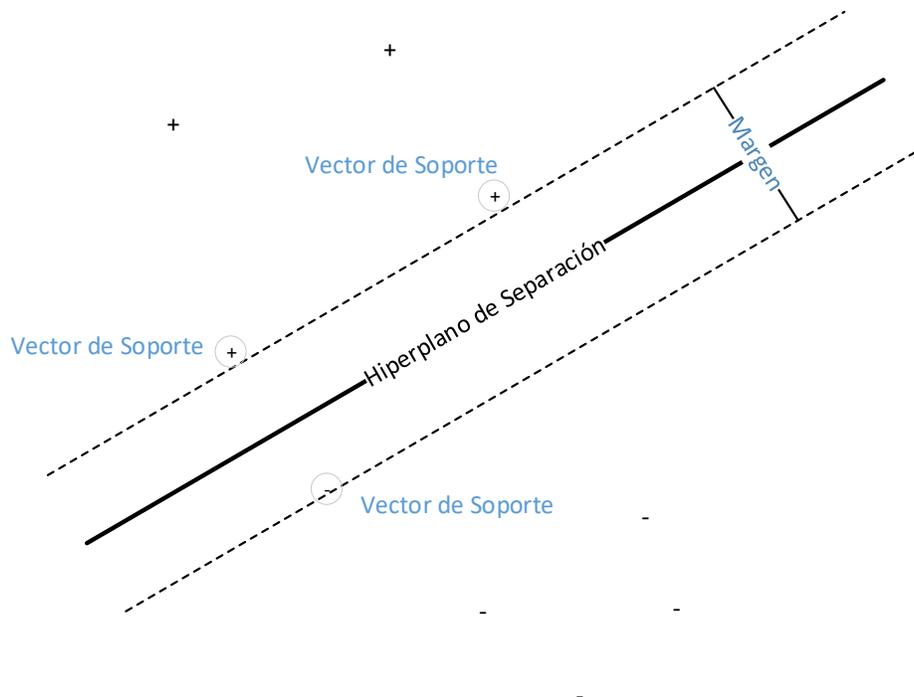


Figura 8. Ejemplo de hiperplano de separación [10].

- **Aprendizaje de una clase**

El aprendizaje de una clase o también denominadas MSV sin supervisión es un algoritmo altamente utilizado para la detección de VA. Básicamente este algoritmo separa todos los puntos del origen y maximiza la distancia desde el hiperplano al origen. Este tipo de aprendizaje intenta identificar objetos de una clase específica entre todos los objetos, aprendiendo de un conjunto de entrenamiento que contiene sólo los objetos de esa clase. En contraste con la clasificación tradicional es diferente y

más difícil ya que esta trata de distinguir entre dos o más clases con el conjunto de entrenamiento que contiene objetos de todas las clases [11].

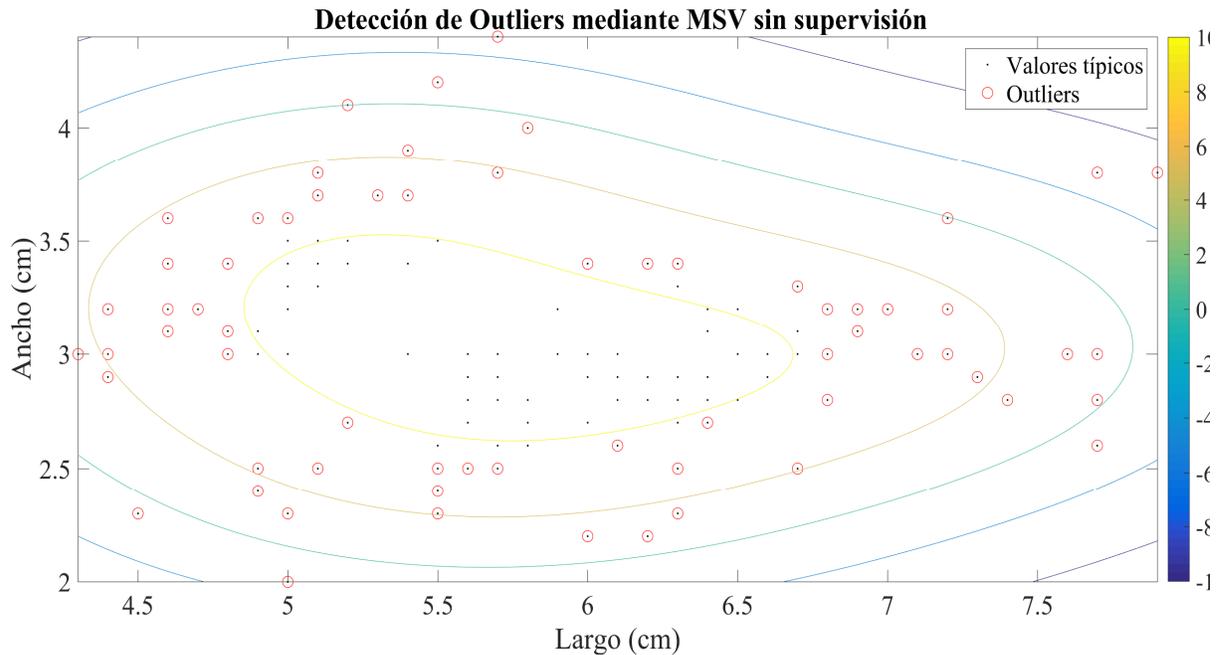


Figura 9. Ejemplo de MSV sin supervisión para la detección de outliers.

2.1.8. Distribuciones de Kernel

2.1.8.1. Introducción

Existen varios procedimientos estadísticos para mostrar la distribución de una base de datos univariada. De acuerdo a investigaciones de otros autores, se dice que es más fácil entender a una distribución como función de densidad que como una función de distribución acumulada [12].

Más adelante se explicará los estimadores clásicos de densidad: histogramas y polígonos de frecuencia para dar paso a estimadores más sofisticados, los cuáles se ocuparán en este documento para ajustar la función de probabilidad de densidad.

2.1.8.2. El histograma

Uno de los métodos más utilizados para representar la forma de una **FDP (Función de Densidad de Probabilidad)** es el histograma. Algunos autores realizar una revisión crítica de este procedimiento, y concluyen en que el histograma es útil para análisis descriptivos. Sin embargo, este procedimiento resulta con frecuencia en una estimación débil de la función de densidad de la población. No obstante, después de la correspondiente transformación de las unidades del eje de las abscisas, el histograma, proporciona una estimación cruda de la densidad [12].

Las principales limitaciones del histograma son:

- **El resultado es dependiente en el origen x_0**

Al momento de construir el histograma, se debe escoger la posición en la cual colocar el origen de los intervalos. Esta decisión se hace generalmente por conveniencia, con intervalos empezando en números centrados o redondeados [12].

- **El resultado depende de la amplitud y número de intervalos**

El número de intervalos debería determinarse con base en alguna función del tamaño de la muestra. Para lotes grandes, un gran número de intervalos debería dar una representación suave de la función de densidad desconocida. Al utilizar unos pocos intervalos, sin embargo, resultará en la pérdida de cualquier detalle de la distribución subyacente. Elegir un gran número de intervalos con un lote pequeño de datos producirá un diagrama univariado de dispersión. El considerar muy pocos intervalos producirá una imagen sin características. Frecuentemente, el número de intervalos y su amplitud se determinan arbitrariamente a pesar de su importancia, la cual radica en que la amplitud de intervalo determina el grado de suavidad del histograma resultante [12].

- **El histograma es discontinuo, con saltos al final de los intervalos**

Las discontinuidades del histograma están principalmente en función de la localización arbitraria de los intervalos y la naturaleza discreta de las datos más que en función de la población que fue muestreada. El hecho es que para el histograma, la densidad local se calcula sólo en cada centro de clase de los intervalos y entonces las barras se dibujan suponiendo una densidad constante a lo largo de cada intervalo [12].

- **La amplitud fija de intervalo resulta en una representación desproporcionada de la densidad en el centro y las de las colas de distribución**

Si los intervalos son lo suficientemente angostos para capturar detalles, típicamente en el centro de la distribución, estos pueden ser muy angostos para evitar ruido donde la densidad es baja, por lo común en las colas. En relación con este problema, el histograma puede generalizarse permitiendo que la amplitud de intervalo varíe. Esto se hace para el primer y el último intervalo del histograma, las cuales, generalmente se construyen para contener todos los puntos bajo un cierto valor y todos aquellos por encima de otro [12].

2.1.8.3. Polígonos de frecuencia

El polígono de frecuencia es un estimador de densidad continuo derivado por la interpolación lineal de los centros de clase del histograma. En contraste con los histogramas los polígonos de frecuencia:

- Aproxima mejor a densidades continuas por medio de interpolación lineal de intervalos más amplios.
- Es menos eficiente cuando la densidad subyacente es discontinua.
- Es más sensitiva con respecto a errores en la elección de la amplitud de intervalo.
- Requiere un error mayor en la amplitud de intervalo antes de que su error cuadrado medio integrado sea peor que el mejor error cuadrado medio integrado de histograma.

- Es más eficiente para los datos en relación con el histograma al crecer el tamaño de muestra.

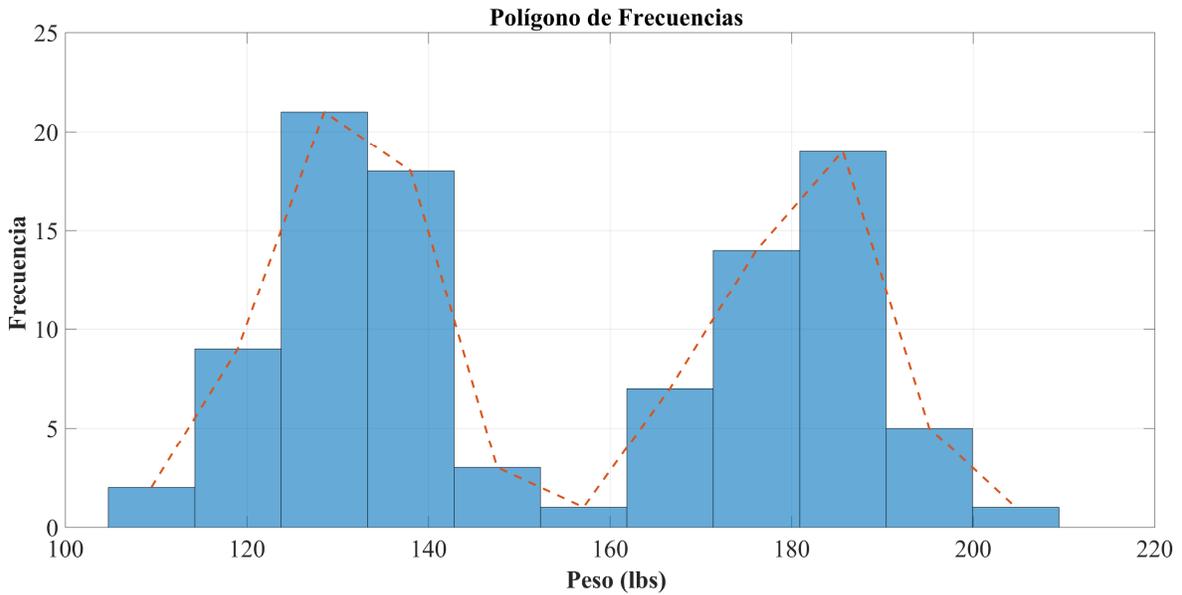


Figura 10. Polígonos de frecuencia.

2.1.8.4. Estimadores de densidad por Kernel

- **Estimador de densidad simple**

Varios autores describen a un estimador de densidad simple considerando que la densidad puede pensarse como el límite de un histograma de barras centradas en x , al tender la amplitud de intervalo a cero [12].

Ecuación 7 – Límite de un histograma de barras centradas en X

$$f(x) = \lim_{h \rightarrow 0} \left(\frac{1}{2h} \Pr(x - h < X < x + h) \right)$$

Un estimador de densidad simple substituye a la proporción de la muestra en una región pequeña (llamada ventana) alrededor de x por la probabilidad escalando la estimación de tal forma que el área total bajo $f(x)$ sea igual a la unidad.

Ecuación 8 – Estimador de densidad simple

$$\hat{f}(x) = \frac{1}{2h} \times \frac{\#[x - h < X_i < x + h]}{n}$$

Este estimador de densidad es semejante al histograma con intervalos de case igual a $2h$ pero con cada punto observado en el centro del intervalo, esto es, sin origen fijo.

La función ponderal $S(*)$ para este estimador simple viene dada por:

Ecuación 9 – Estimador de densidad simple con función ponderal $S(*)$

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n S\left(\frac{x - X_i}{h}\right)$$

- **Estimadores más sofisticados de densidad por kernel**

Para resolver el problema de la discontinuidad causada por las esquinas cuadradas de la función ponderal rectangular que utiliza el estimador simple presentado anteriormente, es conveniente considerar la generalización de este estimador de densidad simple [12]. Esto se consigue reemplazando la función ponderal $S(*)$ por otra produciendo figuras redondeadas en lugar de rectángulos. Para indicar esta distinción la nueva función ponderal se denomina función por kernel $K(*)$ [12].

Ecuación 10 – Estimador de densidad simple con función ponderal $K(*)$

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)$$

Algunas funciones por kernel se incluyen en el siguiente cuadro [12].

Tabla 2. Funciones por Kernel [12].

Funciones por Kernel	Ecuación	Eficiencia
<i>Ecuación 11 – Función por Kernel Epanechnikov</i>	$K(u) = \begin{cases} \frac{3}{4} \left(1 - \frac{1}{5}u^2\right) / \sqrt{5}, & \text{si } u < \sqrt{5} \\ 0, & \text{en cualquier otro caso.} \end{cases}$	1
<i>Ecuación 12 – Función por Kernel Bponderado</i>	$K(u) = \begin{cases} \frac{15}{16} (1 - u^2)^2 / \sqrt{5}, & \text{si } u < 1; \\ 0, & \text{en cualquier otro caso.} \end{cases}$	≈0.9939
<i>Ecuación 13 – Función por Kernel Triangular</i>	$K(u) = \begin{cases} 1 - u , & \text{si } u < 1; \\ 0, & \text{en cualquier otro caso.} \end{cases}$	≈0.9859
<i>Ecuación 14 – Función por Kernel Gaussiano</i>	$K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$	≈0.9512
<i>Ecuación 15 – Función por Kernel Rectangular</i>	$K(u) = \begin{cases} 1/2, & \text{si } u < 1; \\ 0, & \text{en cualquier otro caso.} \end{cases}$	≈0.9295

- **La elección de la amplitud de ventana**

La elección de la amplitud de ventana en las estimaciones por kernel es equivalente a la selección de amplitud de intervalo en histogramas. Esta elección determina las características cualitativas de la densidad por kernel. Un enfoque puede ser variar h hasta que resulte una figura satisfactoria [12].

La teoría estadística proporciona algunos lineamientos para la elección de la amplitud. Desafortunadamente, por lo general no es posible optimizar la amplitud de ventana sin el conocimiento previo de la forma verdadera de la densidad [12].

Según la guía de varios autores, puede emplearse a la distribución gaussiana como un estándar de referencia [12].

Ecuación 16 – Selección de amplitud de intervalo en histogramas

$$s = \min \left[\left(\frac{\sum (x_i - x)^2}{n - 1} \right)^{\frac{1}{2}}, \frac{H \text{ dispersión}}{1.349} \right]$$

Entonces, h puede escogerse como:

Ecuación 17 – Amplitud de ventana para estimadores de densidad por kernel

$$h = \frac{0.9s}{n^{\frac{1}{5}}}$$

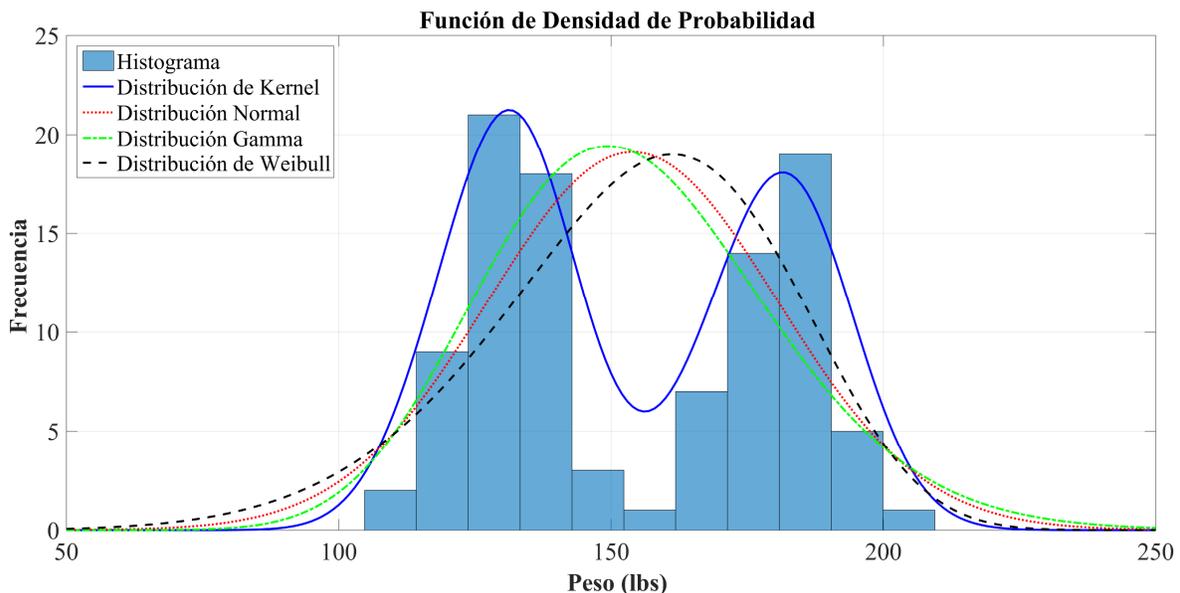


Figura 11. Comparación de las funciones de densidad de probabilidad ajustadas a un histograma.

2.1.9. Gráfica de silueta para la validación del cluster

La silueta es propuesta para la validación de los datos que pertenecen a un mismo cluster. Esta gráfica muestra qué objetos se encuentran dentro de su grupo, y cuáles están en algún lugar entre los grupos.

El clustering completo se muestra combinando las siluetas de cada muestra en un solo gráfico, lo que permite apreciar la calidad relativa de los clústeres y una visión general de la configuración de los datos. El ancho promedio de la silueta proporciona un análisis de la validez del clustering y puede utilizarse para seleccionar un número óptimo de clusters.

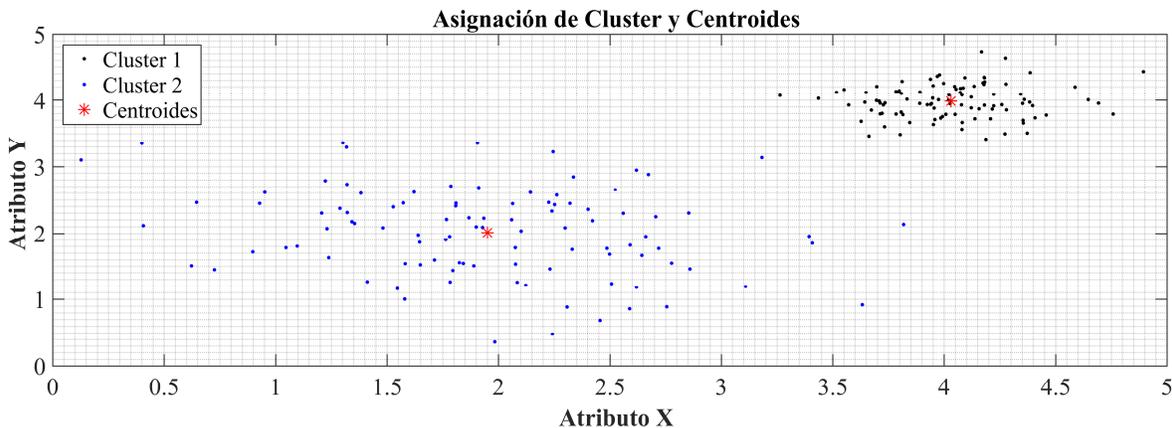


Figura 12. Clusters formados a partir de datos aleatorios.

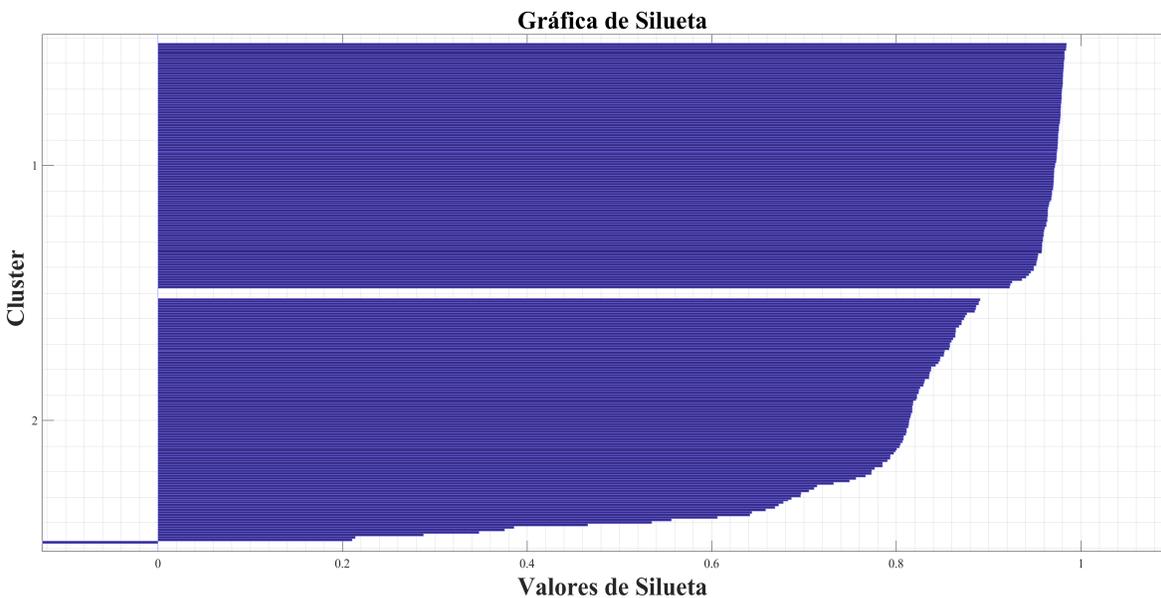


Figura 13. Valores de silueta para cada uno de los clusters.

CAPÍTULO III

3.1. Metodología para el desarrollo del programa en MATLAB

3.1.1. Descripción de metodologías previas y tipos de análisis

Las TC; conocidas también como análisis de segmentación o de taxonomía, son utilizadas para la segmentación de información, basadas en medidas de similitud. Los agrupamientos o clusters se construyen de tal manera que los datos en un mismo conjunto son muy similares y los datos en diferentes conjuntos son muy distintos [13].

Para el agrupamiento de datos se utiliza una amplia variedad de métodos; por ejemplo, el agrupamiento jerárquico que aglomera los datos mediante diversas escalas formando un árbol de cluster o dendrograma. En el dendrograma se observa una jerarquía multinivel, permitiendo decidir el nivel de agrupaciones que son más adecuado para su aplicación [14].

Por ejemplo, considere un conjunto de datos, X , compuesto por cinco objetos donde cada objeto es un conjunto de coordenadas (x, y) .

- Objeto 1: (1, 2)
- Objeto 2: (2,5, 4,5)
- Objeto 3: (2, 2)
- Objeto 4: (4, 1.5)
- Objeto 5: (4, 2.5)

Se calcula la distancia entre el objeto 1 y el objeto 2, el objeto 1 y el objeto 3, y así sucesivamente hasta que se hayan calculado las distancias entre todos los objetos. La figura 14 representa estos objetos en un gráfico. La distancia euclidiana entre el objeto 2 y el objeto 3 se muestra para ilustrar una interpretación de la distancia [14].

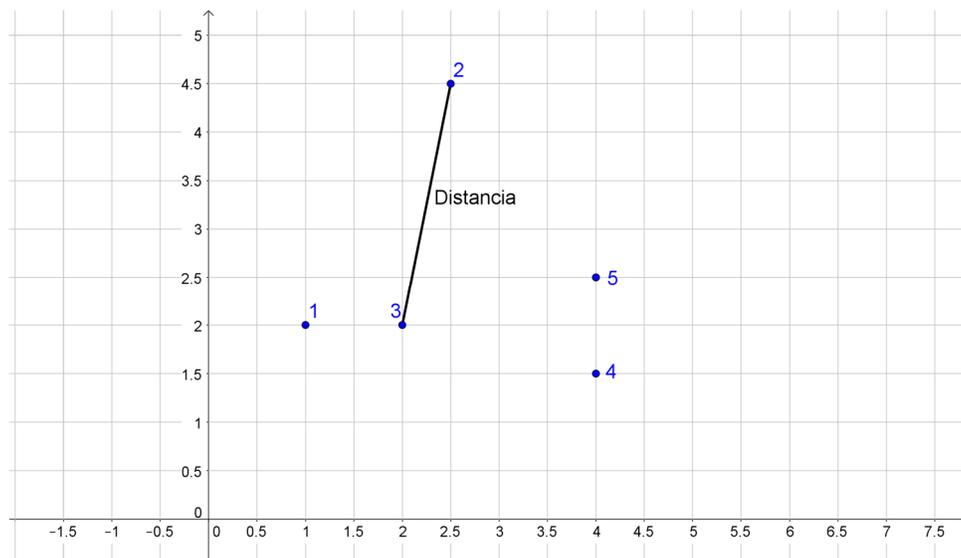


Figura 14. Objetos del conjunto de datos X.

Una vez que se ha calculado la proximidad entre objetos del conjunto de datos, se puede determinar cómo los objetos en el conjunto de datos deben aglomerarse en grupos, vinculando pares de objetos que se encuentran juntos en clusters binarios (clusters compuestos por dos objetos). Los clusters binarios se vinculan entre sí con otros objetos para crear clusters más grandes hasta que todos los objetos en el conjunto de datos original se enlazan en un árbol jerárquico [14].

La figura 15 ilustra gráficamente la forma en que los grupos vinculan los objetos en una jerarquía de clusters.

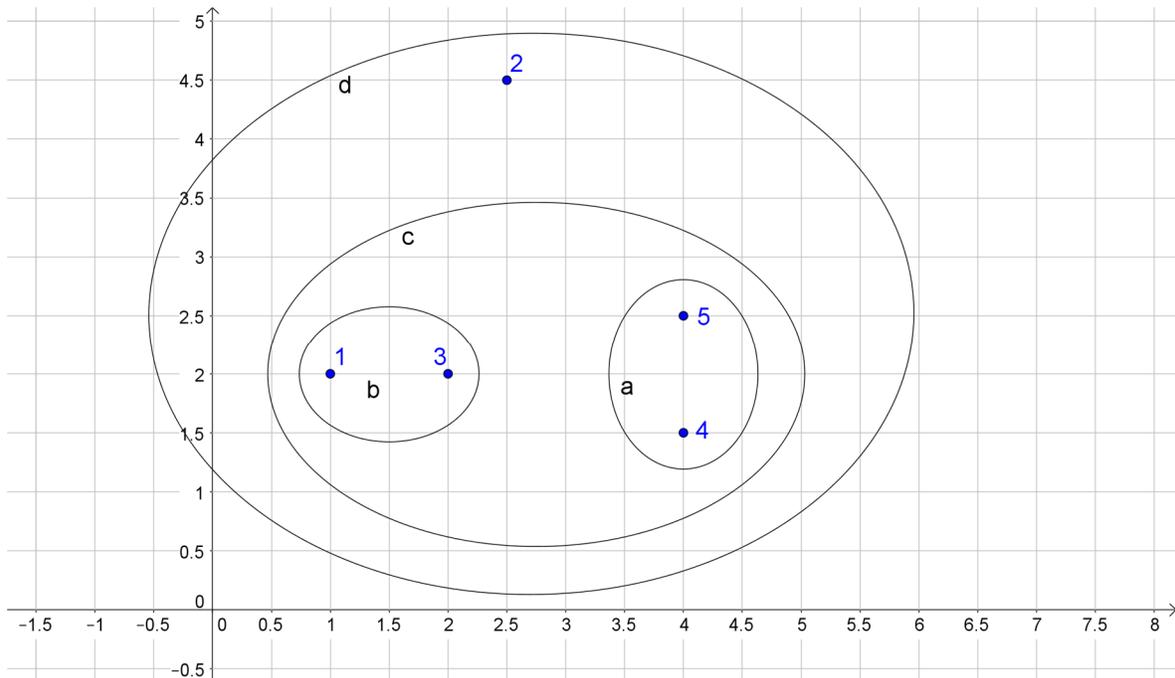


Figura 15. Objetos vinculados en grupos según la jerarquía de clusters.

El árbol jerárquico y binario del cluster creado se entiende más fácilmente cuando se visualiza gráficamente. El dendrograma representa el árbol de la siguiente manera.

En la figura 16, los números a lo largo del eje horizontal representan los índices de los objetos en el conjunto de datos original. Los enlaces entre los objetos se representan como líneas en forma de U invertidas. La altura de la U indica la distancia entre los objetos. Por ejemplo, el enlace que representa el cluster que contiene los objetos 4 y 5 tiene una altura de 1. El enlace que representa el cluster que agrupa el objeto 2 junto con los objetos 1, 3, 4 y 5 (que ya están agrupados como objeto d) tiene una altura de 2.5 [14].

Por otro lado, las TC denominadas k -Means y k -Medoids son los principales métodos utilizados para la partición de grandes bases de datos. Estos métodos están basados en la partición de n clusters mutuamente excluyentes, permitiendo identificar a que cluster fue asignado cada uno de los datos; y estableciendo un solo nivel de agrupaciones. Dichas técnicas asignan cada dato en análisis a un cluster, reduciendo al mínimo la distancia existente desde el dato en análisis al centroide

(conceptualmente el punto central del cluster) o medoide (punto representativo del cluster), respectivamente [15].

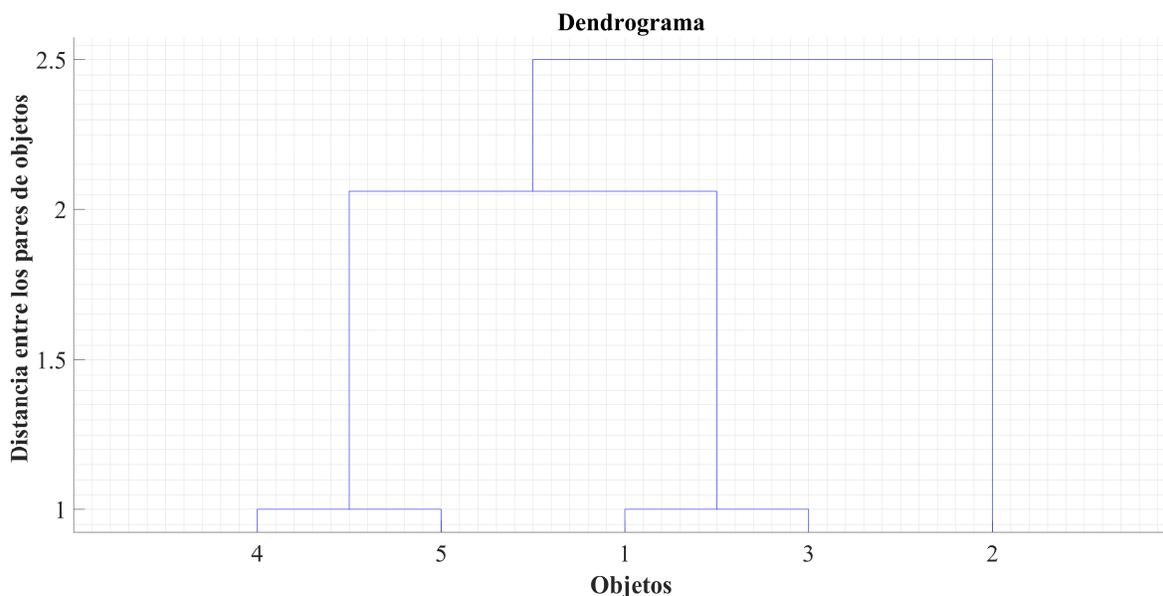


Figura 16. Dendrograma de los objetos según la jerarquía de clusters.

Las TC denominadas k -Means y k -Medoids tratan cada observación en sus datos como un objeto que tiene una ubicación en el espacio. Encuentra una partición en la que los objetos dentro de cada grupo están lo más cerca posible entre sí y lejos de los objetos de otros grupos como sea posible [15].

A diferencia de la agrupación jerárquica, la agrupación de k -Means o k -Medoids operan en observaciones reales (en lugar del conjunto más grande de medidas de disimilitud) y crea un único nivel de agrupaciones [15].

Se puede controlar los detalles de la minimización de la distancia utilizando varios parámetros de entrada opcionales para las TC denominadas k -Means o k -Medoids, incluidos los valores iniciales de los centroides del cluster y el número máximo de iteraciones. De manera predeterminada, las TC denominadas k -Means o k -Medoids utilizan el algoritmo k -means ++ para la inicialización del centro del cluster y la métrica euclidiana al cuadrado para determinar las distancias [15].

Las diferencias entre las TC mencionadas anteriormente permiten determinar que los métodos de partición k -Means o k -Medoids son a menudo más adecuados que la agrupación jerárquica para grandes cantidades de datos [15].

La aplicación de las TC a grandes cantidades de datos han sido evaluadas ampliamente en varios de estudios como en el documento investigativo [16]; donde se sugiere la utilización del algoritmo k -Means por su diferencia significativa en velocidad computacional para el procesamiento de datos con respecto al resto de algoritmos relacionados a las TC, además indica, “Sin embargo, la información más importante para evaluar la eficacia del algoritmo de agrupamiento que se utiliza es acerca de la adecuación de agrupamiento realizado” [16].

Para la detección de VA; después de haber obtenido la segmentación de los datos, se utiliza una amplia variedad de métodos conjuntamente; por ejemplo, análisis exploratorio y dispersión de datos mediante media aritmética, desviación estándar y rango intercuartil, estimadores de densidad por Kernel para funciones de densidad de probabilidad y MSV con aprendizaje de una clase o sin supervisión.

Lo anteriormente mencionado, permite obtener la respectiva clasificación entre **PCEE (Perfiles de Consumo de Energía Eléctrica)** diarios con datos típicos y atípicos para el **RPC (Reconocimiento de Patrones de Consumo)** del cliente o los clientes que se analizan.

Por lo tanto, este **PTEI (Proyecto Técnico de Enfoque Investigativo)** constituye un método de investigación descriptivo de las TC en base a los algoritmos *k*-Means y *k*-Medoids; además de un análisis exhaustivo para la detección de outliers y clasificación de los PCEE con el fin de obtener patrones de CEE del o los consumidores.

3.1.2. Bases de datos para el análisis clustering

Para la recogida de datos principales, se realizaron solicitudes de los PCEE diarios de diferentes tipos de clientes a la Empresa Eléctrica Regional CENTROSUR C.A., obteniendo información sobre el consumo energético de la **UPS (Universidad Politécnica Salesiana)** – Sede Cuenca.

No hubo información disponible de clientes residenciales; ya que, la Empresa Eléctrica Regional CENTROSUR C.A. obtiene las lecturas de CEE de manera mensual para este tipo de consumidores. Además, no se pudo obtener información por parte de esta institución de clientes comerciales e industriales por las políticas de privacidad implementadas en la misma (información no revelada).

Por lo tanto, se realizaron solicitudes de PCEE diarios a Continental Tire Andina y a la UPS – Sede Cuenca; obteniendo información de la planta y las instalaciones de dichas instituciones, respectivamente.

De igual manera, se obtuvo información del PCEE diarios de la **UPV (Universidad Politécnica de Valencia)** - España, la misma que fue proporcionada por el tutor del presente PTEI.

Los muestreos de CEE de diversos días recolectados se llevaron a cabo en intervalos de 15 minutos, como lo indica la REGULACIÓN Nro. ARCONEL – 001/16 de la Agencia de Regulación y Control de Electricidad.

La información recogida de CEE diario correspondiente a la acometida principal de la UPS – Sede Cuenca parte desde el 08 de marzo del 2017 hasta el 24 de Julio del 2017.

La información recogida de CEE diario correspondiente al Edificio Cornelio Merchán de la UPS – Sede Cuenca parte desde el 03 de marzo del 2017 hasta el 24 de Julio del 2017.

La información recogida de CEE diario correspondiente a las instalaciones de la UPV – España parte desde el 01 de enero del 2016 hasta el 27 de noviembre del 2016.

La información recogida de CEE diario correspondiente a Continental Tire Andina parte desde el 01 de enero del 2016 hasta el 31 de diciembre del 2016.

Los consumidores citados fueron seleccionados en base al número de PCEE diarios con los que cuentan; siendo necesario tener información de CEE diarios de 365 días que corresponde a un año para que los resultados a obtener sean certeros.

3.1.3. Enfoque investigativo

Para la realización de este PTEI se hizo uso de una combinación de enfoques investigativo - descriptivo; ya que, tiene como propósito principal realizar la segmentación de datos, **DA (Detección de Anomalías)** y clasificación de CEE diarios de determinado tipo de consumidor o consumidores; así como, describir sus patrones de CEE.

Las ventajas de usar esta metodología son:

- Organizar la información.
- Sintetizar la información.
- Ver sus características más relevantes.
- Presentar la información.

Esta vía permite identificar las relaciones que existen entre dos o más variables; presentando y resumiendo la información de manera cuidadosa y luego analizando minuciosamente los resultados que contribuyen al reconocimiento de patrones de CEE del cliente o diversos clientes.

3.2. Flujograma del programa

El programa desarrollado en el Software MATLAB denominado “TECNICA_CLUSTER_K_MEANS_Y_DETECCION_DE_ANOMALIAS” y “TECNICA_CLUSTER_K_MEDOIDS_Y_DETECCION_DE_ANOMALIAS” contiene varios procesos y subprocesos que se pueden visualizar con mayor facilidad en el flujograma que se muestra las figuras 17 y 18; siendo explicados detalladamente en el presente capítulo.

Los procesos tienen las opciones de selección para que el programa los ejecute mediante lo elegido por el usuario.

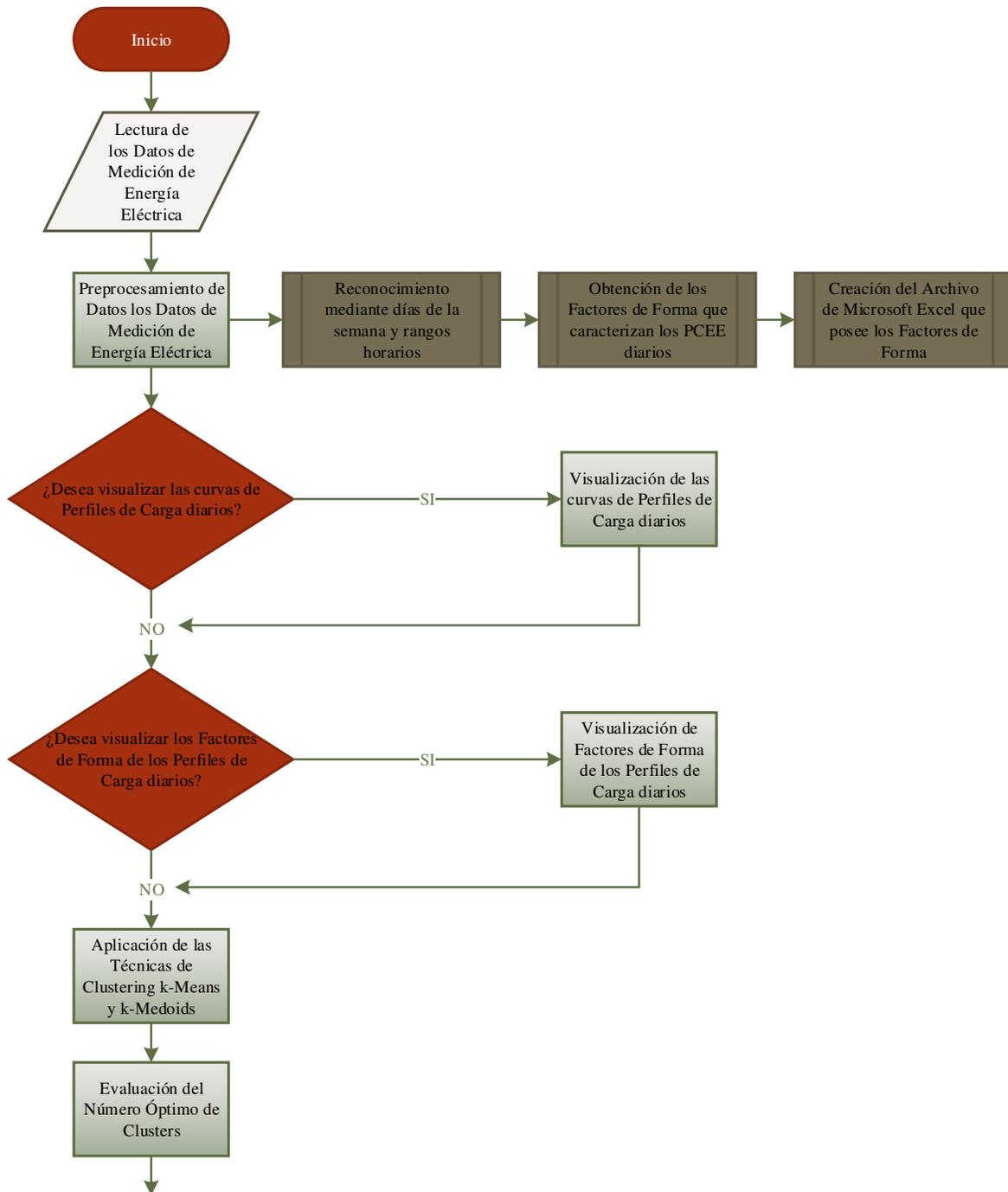


Figura 17. Flujograma del Programa desarrollado en el Software MATLAB.

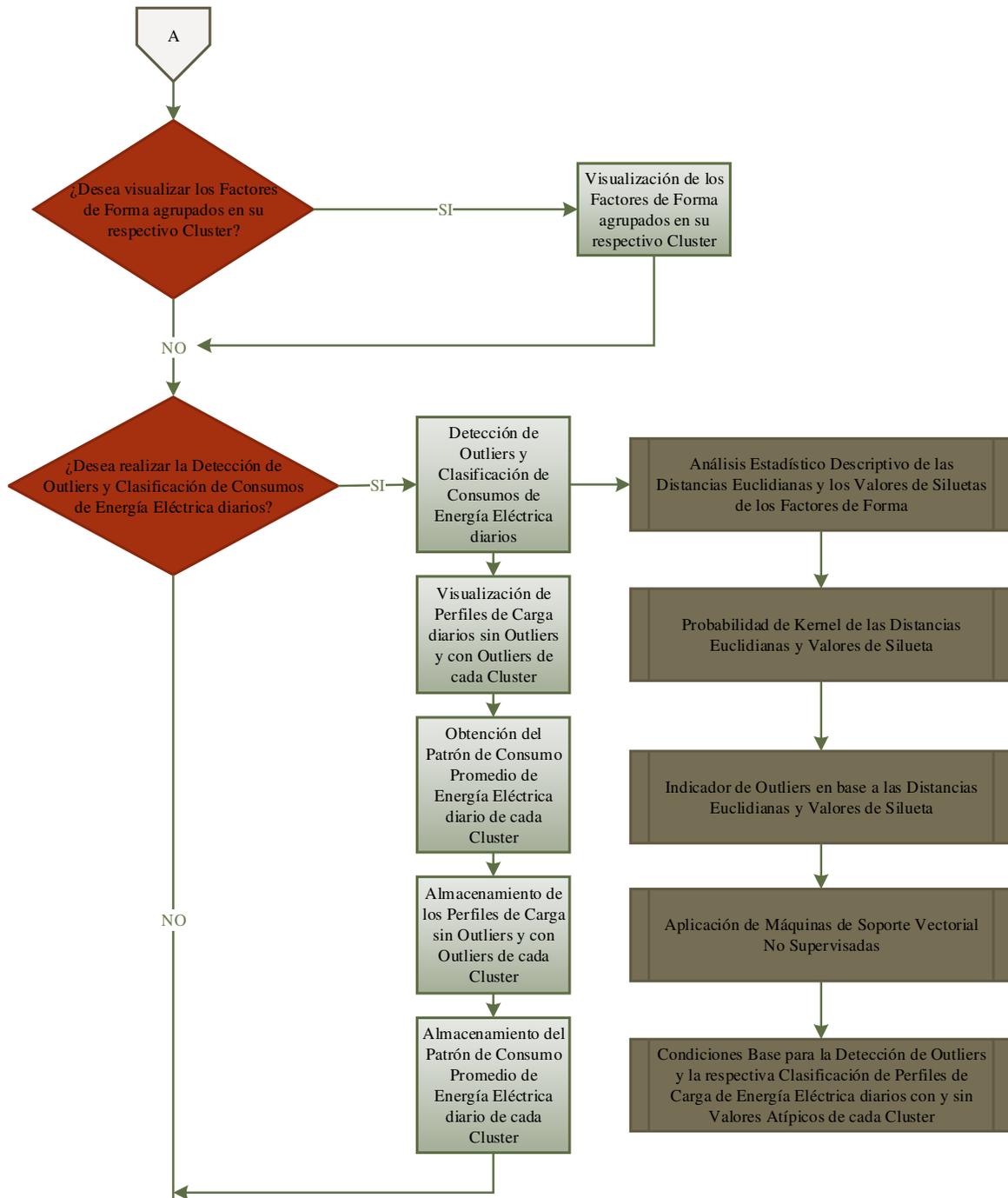


Figura 18. Flujograma del Programa desarrollado en al Software MATLAB.

3.3. Desarrollo del programa en el software MATLAB

El programa del presente PTEI contiene diversas etapas secuenciales respecto a su codificación que se explican en este capítulo. Inicialmente para cargar los datos de las mediciones cada 15 minutos por los contadores de energía eléctrica. Se debe tener en cuenta que en la Hoja de Microsoft Excel donde se encuentran los mismos tiene la siguiente configuración:

- La primera columna de la hoja debe contener las fechas de los datos en formato texto de la siguiente manera: (1/1/2016 0:00:00).
- La segunda columna debe contener los valores de energía en kWh.
- La primera fila de la Hoja contiene los nombres “Fecha” y “kWh” de cada columna, respectivamente.

3.3.1. Lectura de los datos de medición de energía eléctrica

Para el análisis de clustering, los datos de los PCEE diarios de un cliente o varios clientes a ser analizados se encuentran en un libro y hoja específica de un archivo de Microsoft Excel. Por lo tanto, el programa solicita ingresar el nombre del libro de Microsoft Excel donde se encuentran dichos datos.

```
NombreLibro = input('1. Ingrese el nombre del Libro de Microsoft Excel que contiene los Perfiles de Carga de Energía Eléctrica: \n','s');
```

A continuación, el programa solicita ingresar el nombre de la hoja de Microsoft Excel donde se encuentra los datos de medición de energía eléctrica de los PCEE diarios a ser analizados.

```
NombreHoja = input('2. Ingrese el nombre de la Hoja de Microsoft Excel que contiene los Perfiles de Carga de Energía Eléctrica: \n','s');
```

En el momento que los datos de fecha y medición de energía eléctrica de los PCEE diarios han sido cargados con éxito, dicha información es almacenada en una matriz de dimensiones $[Fx2]$, donde sus columnas son nominadas como “num” (kWh) con variables de tipo numérico y como “txt” (Fecha) con variables de tipo string; siendo eliminada la primera fila de la matriz que corresponde a los nombres de dichas columnas del archivo de Microsoft Excel ingresado en el programa. F representa el número de mediciones registrada por parte del contador de energía eléctrica.

```
[num,txt] = xlsread(NombreLibro,NombreHoja,");  
txt(1,:)=[];
```

A partir de la matriz anterior que contiene la información de PCEE diarios correspondiente al muestreo de mediciones realizadas en períodos de 15 minutos; se crea un vector de dimensiones $[Fx1]$ denominado como “DateString”, donde se encuentra la información de las fechas de los muestreos de las mediciones de energía eléctrica.

```
DateString = txt(:,1);
```

Para el vector anterior que contiene la información de las fechas de los muestreos de las mediciones de cada uno de los datos de consumo de energía; con el formato Día/Mes/Año/Hora/Minutos/Segundos, se realiza la transformación de los datos de tipo string en un vector con datos de tipo numérico que representan cada fecha de cada medición en el tiempo como un número. Dichos datos son almacenados en el vector denominado “DateNumber”.

```
DateNumber = datenum(DateString,'dd/mm/yyyy HH:MM:SS');
```

La función “datevec” convierte el vector “DateNumber” con datos de tipo numérico en valores que representan los componentes de las fechas en las cuales fueron realizadas cada una de las mediciones con el formato Año/Mes/Día/Hora/Minutos/Segundos [17].

```
DateVector = datevec(DateNumber(:,1));
```

Se especifica que la variable numérica para el reconocimiento del día de la semana al que corresponde cada medición de energía eléctrica es de tipo “long” para la variable “DayForm”. De esta manera el programa devuelve el nombre completo del día de la semana en el formato especificado.

```
DayForm = 'long';
```

Se obtiene el número de la semana al que corresponde el dato de medición de CEE según la fecha en que fue realizada mediante la función “weekday”. El valor que representa el día de la semana es devuelto como un vector denominado “DayNumber” de dimensiones $[F \times 1]$ de números enteros en el rango [1,7], donde 1 representa el día domingo y 7 representa el día sábado [18].

```
[DayNumber] = weekday(DateNumber,DayForm);
```

Se obtiene el tamaño del vector “num” que contiene los datos de las mediciones registradas a partir de los PCEE cargados; donde f corresponde al número total de mediciones de energía eléctrica.

```
[f, c] = size(num);
```

De esta forma, los datos de las mediciones de CEE diarios con sus respectivas fechas han sido leídos de manera exitosa por parte del programa; teniendo cada PCEE 96 datos de energía eléctrica correspondientes a las mediciones tomadas cada 15 minutos a lo largo de cada día y con su respectivo día de la semana al que corresponde.

3.3.2. Preprocesamiento de datos de medición de energía eléctrica

3.3.2.1. Reconocimiento mediante días de la semana y rangos horarios

Para el reconocimiento de los PCEE diarios se utiliza el método de peculiaridades de la forma directa, que se define fácilmente utilizando los perfiles de carga resultantes de las medidas realizadas en el dominio del tiempo adaptado en [19].

Cada perfil de carga se representa a continuación utilizando H peculiaridades, como $l = \{l_h, h = 1, \dots, H\}$, tal como el conjunto $L = \{l^{(n)}, n = 1, \dots, N\}$ contiene todos los PCEE diarios de los N clientes o N días en análisis [19].

El valor de H puede ser elegido según el número de datos medidos y con el detalle deseado para la representación del PCEE diario, típicamente, utilizando intervalos de tiempo por hora ($H = 24$), de media hora ($H = 48$) o de cuarto de hora ($H = 96$). En el presente PTEI se utiliza ($H = 96$); ya que, cada medición fue obtenida cada 15 minutos [19].

Con lo anteriormente indicado, se procede a la clasificación de las 96 mediciones correspondientes a cada uno de los PCEE diarios del cliente o los clientes por medio de rangos horarios. Dichos rangos horarios son necesarios posteriormente para el cálculo de los factores de forma que caracterizan a cada uno de los PCEE diarios que se encuentran en análisis [20].

A continuación, se incluyen los perfiles característicos de los principales sectores de consumo, correspondientes a cinco empresas distribuidoras que permiten determinar de una mejor manera los rangos horarios de CEE en el Ecuador [21]:

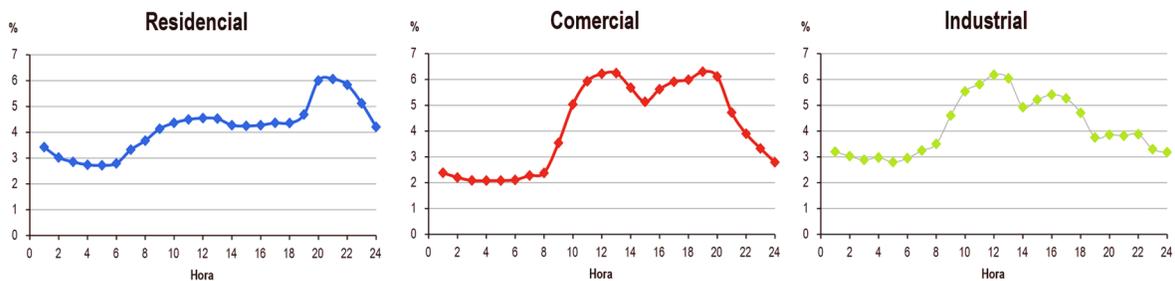


Figura 19. Curvas de carga - E.E. Ambato [21].

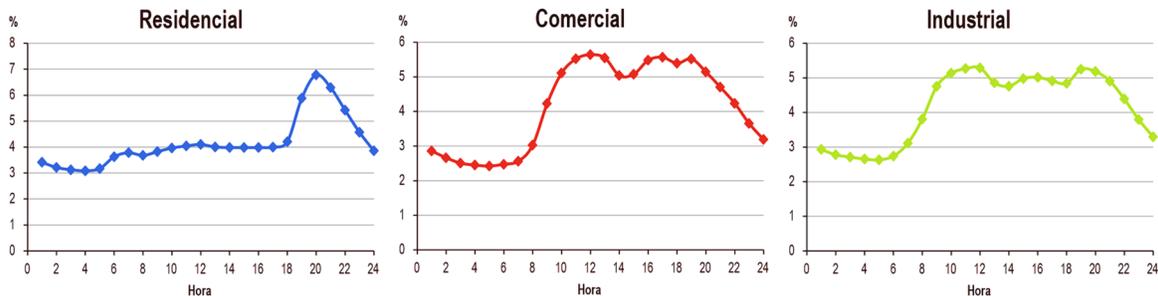


Figura 20. Curvas de carga - E.E. Centrosur [21].

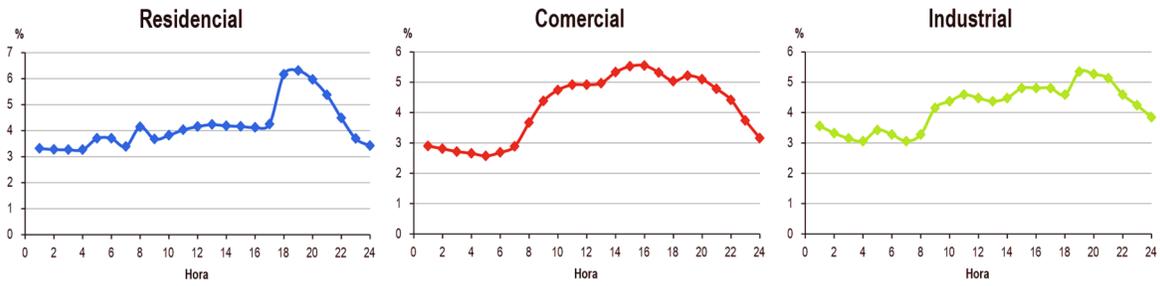


Figura 21. Curvas de carga - CNEC-Sucumbíos [21].

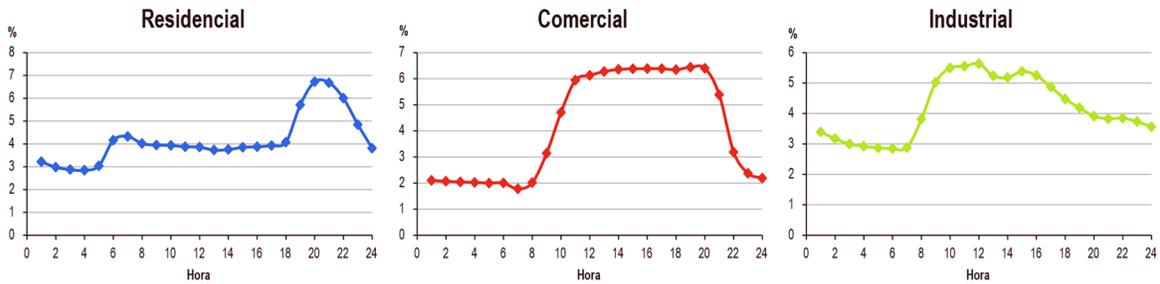


Figura 22. Curvas de Carga - E.E. Quito [21].

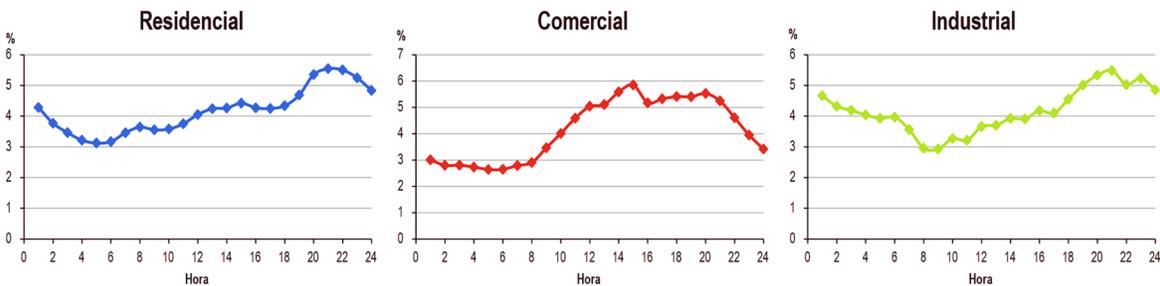


Figura 23. Curvas de Carga – Eléctrica de Guayaquil [21].

De esta manera, la clasificación de los rangos horarios para las 96 mediciones correspondientes a cada uno de los PCEE diarios utilizados para el desarrollo del presente PTEI se muestran en la siguiente tabla:

Tabla 3. Rangos Horarios para la Clasificación de las Mediciones de Energía Eléctrica.

Período de Tiempo	Descripción	Número de mediciones tomadas por parte del contador de energía cada 15 minutos
07h00 a 18h00	Rango horario de comercialización de Energía Eléctrica Ecuatoriano para consumidores Comerciales e Industriales.	44

18h00 a 22h00	Rango horario de comercialización de Energía Eléctrica Ecuatoriano para consumidores Comerciales e Industriales.	16
22h00 a 07h00	Rango horario de comercialización de Energía Eléctrica Ecuatoriano para consumidores Comerciales e Industriales.	36
06h00 a 12h00	Rango horario determinado a partir de las curvas de carga Residencial, Comercial e Industrial en el Ecuador.	24
12h00 a 15h00	Rango horario determinado a partir de las curvas de carga Residencial, Comercial e Industrial en el Ecuador.	12
15h00 a 18h00	Rango horario determinado a partir de las curvas de carga Residencial, Comercial e Industrial en el Ecuador.	12

Una vez que los rangos horarios se encuentran determinados; el programa procede a realizar la clasificación de cada una de las mediciones de los PCEE diarios según el rango horario al que corresponden. Dicha clasificación se hace conjuntamente según el día de la semana al que pertenece cada dato de medición.

Por ejemplo; para los PCEE correspondientes a los días lunes, se inicializa el contador de filas que corresponde al número de PCEE en análisis, y se inicializa el contador de columnas que corresponde al número de medición tomada por el contador de energía a lo largo del día para cada PCEE.

```
iflun=0;
iclun=1;
```

Se realiza la extracción de los datos de energía de las mediciones de cada lunes mediante un “bucle for” y según los rangos horarios especificados anteriormente, siendo almacenados en diversas matrices $[N \times M]$ para los diversos rangos horarios; donde N es el número de PCEE diarios correspondientes en el presente ejemplo a los días lunes y M es el número de mediciones obtenidas cada 15 minutos para cada rango horario.

Se debe tener presente que inicialmente los datos de las mediciones de los PCEE diarios son valores de energía (kWh); por lo tanto, en base a esa información se calcula la potencia media del intervalo de medida (kW).

Ecuación 18 – Transformación de energía eléctrica a potencia eléctrica

$$Potencia (kW) = \frac{Energía (kWh)}{Tiempo (h)}$$

Como se conoce de antemano, las mediciones de energía eléctrica son tomadas cada 15 minutos y equivalen a ¼ de hora; por lo tanto:

Ecuación 19 – Transformación de energía eléctrica a potencia eléctrica

$$Potencia (kW) = \frac{Energía (kWh)}{\frac{1}{4} hora}$$

La codificación para el reconocimiento de los datos de medición mediante días de la semana y rangos horarios mencionado anteriormente se muestra a continuación:

```
for n = 1: f
    if (DayNumber(n,1)==2)
        iflun=iflun+1;
        if iflun>=26 & iflun<=49
            DLUNES6_12(iflun-25, iclun) = num(n,1)/0.25;
        end
        if iflun>=50 & iflun<=61
            DLUNES12_15(iflun-49, iclun) = num(n,1)/0.25;
        end
        if iflun>=62 & iflun<=73
            DLUNES15_18(iflun-61, iclun) = num(n,1)/0.25;
        end
        if iflun>=30 & iflun<=73
            DLUNES7_18(iflun-29, iclun) = num(n,1)/0.25;
        end
        if iflun>=74 & iflun<=89
            DLUNES18_22(iflun-73, iclun) = num(n,1)/0.25;
        end
        if iflun>=1 & iflun<=29
            DLUNES24_7(iflun, iclun) = num(n,1)/0.25;
        end
        if iflun>=90 & iflun<=96
            DLUNES22_24(iflun-89, iclun) = num(n,1)/0.25;
        end
        if iflun<96
            DLUNES(iflun, iclun) = num(n,1)/0.25;
        elseif iflun==96
            DLUNES(iflun, iclun) = num(n,1)/0.25;
            iclun=iclun+1;
            iflun=0;
        end
    end
end
```

```

end
end
DLUNES22_7 = [DLUNES24_7;DLUNES22_24];

```

Este procedimiento se realiza para los PCEE diarios correspondientes a los demás días de la semana que fueron obtenidos mediante la lectura de los datos de medición de energía eléctrica del programa.

3.3.2.2. Obtención de los Factores de Forma que caracterizan los PCEE diarios

Los factores de forma son valores de potencia activa normalizados, los cuales son obtenidos de los diferentes períodos de tiempo provenientes de los rangos horarios de las curvas de carga y del sistema de comercialización de energía eléctrica ecuatoriano [19].

Para los factores de forma se considera respectivamente $P^{MAX,\Delta t}$, $P^{MIN,\Delta t}$, $P^{AV,\Delta t}$ como potencia máxima, mínima y promedio en el período Δt , donde Δt puede representar el día entero (24 horas) u otros períodos de la mañana (6 horas, de 6h00 a 18h00), almuerzo (3 horas, de 12h00 a 15h00), de la tarde (3 horas, de 15h00 a 18h00), horas pico (4 horas, de 18h00 a 22h00), de la noche (9 horas, de 22h00 a 7h00); y de la mañana y tarde conjuntamente (11 horas, de 7h00 a 18h00) [19][22, p. 5].

Así; el programa realiza el cálculo de los valores máximos, mínimos y promedios de potencia activa de cada uno de los PCEE diarios en análisis con relación a los diversos rangos horarios especificados en la tabla 3.

Continuando con el ejemplo, para los PCEE de los días lunes se tiene la siguiente codificación para la obtención de los valores máximos, mínimos y promedios de potencia activa.

Valores máximos, mínimos y promedios de potencia activa del día completo:

```

PMAXLUNES = (max(DLUNES,[],1));
PMINLUNES = (min(DLUNES,[],1));
PMEDIALUNES = (mean(DLUNES,1));

```

Valores máximos, mínimos y promedios de potencia activa del rango horario de 07h00 a 18h00:

```

PMAX7_18LUNES = (max(DLUNES7_18,[],1));
PMIN7_18LUNES = (min(DLUNES7_18,[],1));
P7_18LUNES = (mean(DLUNES7_18,1));

```

Valores máximos, mínimos y promedios de potencia activa del rango horario de 18h00 a 22h00:

```

PMAX18_22LUNES = (max(DLUNES18_22,[],1));
PMIN18_22LUNES = (min(DLUNES18_22,[],1));
P18_22LUNES = (mean(DLUNES18_22,1));

```

Valores máximos, mínimos y promedios de potencia activa del rango horario de 22h00 a 07h00:

```

PMAX22_7LUNES = (max(DLUNES22_7,[],1));
PMIN22_7LUNES = (min(DLUNES22_7,[],1));
P22_7LUNES = (mean(DLUNES22_7,1));

```

Valores máximos, mínimos y promedios de potencia activa del rango horario de 06h00 a 12h00:

$$P_{MAX6_12LUNES} = (\max(DLUNES6_12, [], 1));$$

$$P_{MIN6_12LUNES} = (\min(DLUNES6_12, [], 1));$$

$$P_{6_12LUNES} = (\text{mean}(DLUNES6_12, 1));$$

Valores máximos, mínimos y promedios de potencia activa del rango horario de 12h00 a 15h00:

$$P_{MAX12_15LUNES} = (\max(DLUNES12_15, [], 1));$$

$$P_{MIN12_15LUNES} = (\min(DLUNES12_15, [], 1));$$

$$P_{12_15LUNES} = (\text{mean}(DLUNES12_15, 1));$$

Valores máximos, mínimos y promedios de potencia activa del rango horario de 15h00 a 18h00:

$$P_{MAX15_18LUNES} = (\max(DLUNES15_18, [], 1));$$

$$P_{MIN15_18LUNES} = (\min(DLUNES15_18, [], 1));$$

$$P_{15_18LUNES} = (\text{mean}(DLUNES15_18, 1));$$

Al obtener los valores máximos, mínimos y promedios de potencia activa de cada uno de los PCEE diarios en los diversos rangos horarios mencionados con anterioridad, se consideran varios factores de forma para la caracterización de las curvas de carga diarias.

Se utiliza como base un conjunto de factores de forma normalizados (tabla 4). La elección de los factores de forma requiere criterio e intuición y es la base para ejecutar el procedimiento de segmentación.

Bajo estas condiciones, se proponen varios factores de forma para definir el comportamiento de las curvas de carga de energía eléctrica diarias que se encuentran en análisis.

Tabla 4. Factores de Forma Normalizados para la Caracterización de las Curvas de Carga diarias.

Factor de Forma	Fórmula	Período
<i>Ecuación 20 – Factor de Carga (F1)</i>	$f1 = \frac{P_{\text{promedio del día}}}{P_{\text{máxima del día}}}$	24 horas del día
<i>Ecuación 21 – Coeficiente de no uniformidad (F2)</i>	$f2 = \frac{P_{\text{mínima del día}}}{P_{\text{promedio del día}}}$	24 horas del día
<i>Ecuación 22 – Impacto promedio en la mañana (F3)</i>	$f3 = \frac{P_{\text{promedio de la mañana}}}{P_{\text{promedio del día}}}$	Rango horario de 06h00 a 12h00
<i>Ecuación 23 – Impacto máximo en la mañana (F4)</i>	$f4 = \frac{P_{\text{máxima de la mañana}}}{P_{\text{promedio del día}}}$	Rango horario de 06h00 a 12h00
<i>Ecuación 24 – Impacto mínimo en la mañana (F5)</i>	$f5 = \frac{P_{\text{mínima de la mañana}}}{P_{\text{promedio del día}}}$	Rango horario de 06h00 a 12h00

<i>Ecuación 25 – Impacto promedio del período de almuerzo (F6)</i>	$f6 = \frac{P_{\text{promedio período almuerzo}}}{P_{\text{promedio del día}}}$	Rango horario de 12h00 a 15h00
<i>Ecuación 26 – Impacto máximo del período de almuerzo (F7)</i>	$f7 = \frac{P_{\text{máxima período almuerzo}}}{P_{\text{promedio del día}}}$	Rango horario de 12h00 a 15h00
<i>Ecuación 27 – Impacto mínimo del período de almuerzo (F8)</i>	$f8 = \frac{P_{\text{mínimo período almuerzo}}}{P_{\text{promedio del día}}}$	Rango horario de 12h00 a 15h00
<i>Ecuación 28 – Impacto promedio de la tarde (F9)</i>	$f9 = \frac{P_{\text{promedio de la tarde}}}{P_{\text{promedio del día}}}$	Rango horario de 15h00 a 18h00
<i>Ecuación 29 – Impacto máximo de la tarde (F10)</i>	$f10 = \frac{P_{\text{máxima de la tarde}}}{P_{\text{promedio del día}}}$	Rango horario de 15h00 a 18h00
<i>Ecuación 30 – Impacto mínimo de la tarde (F11)</i>	$f11 = \frac{P_{\text{mínima de la tarde}}}{P_{\text{promedio del día}}}$	Rango horario de 15h00 a 18h00
<i>Ecuación 31 – Impacto promedio del horario matutino y vespertino (F12)</i>	$f12 = \frac{P_{\text{promedio h. matutino y vespertino}}}{P_{\text{promedio del día}}}$	Rango horario de 07h00 a 18h00
<i>Ecuación 32 – Impacto máximo del horario matutino y vespertino (F13)</i>	$f13 = \frac{P_{\text{máximo h. matutino y vespertino}}}{P_{\text{promedio del día}}}$	Rango horario de 07h00 a 18h00
<i>Ecuación 33 – Impacto mínimo del horario matutino y vespertino (F14)</i>	$f14 = \frac{P_{\text{mínimo h. matutino y vespertino}}}{P_{\text{promedio del día}}}$	Rango horario de 07h00 a 18h00
<i>Ecuación 34 – Impacto promedio del atardecer (F15)</i>	$f15 = \frac{P_{\text{promedio del atardecer}}}{P_{\text{promedio del día}}}$	Rango horario de 18h00 a 22h00
<i>Ecuación 35 – Impacto máximo del atardecer (F16)</i>	$f16 = \frac{P_{\text{promedio del atardecer}}}{P_{\text{promedio del día}}}$	Rango horario de 18h00 a 22h00
<i>Ecuación 36 – Impacto mínimo del atardecer (F17)</i>	$f17 = \frac{P_{\text{promedio del atardecer}}}{P_{\text{promedio del día}}}$	Rango horario de 18h00 a 22h00
<i>Ecuación 37 – Impacto promedio de la noche (F18)</i>	$f18 = \frac{P_{\text{promedio de la noche}}}{P_{\text{promedio del día}}}$	Rango horario de 22h00 a 07h00
<i>Ecuación 38 – Impacto máximo de la noche (F19)</i>	$f19 = \frac{P_{\text{promedio de la noche}}}{P_{\text{promedio del día}}}$	Rango horario de 22h00 a 07h00
<i>Ecuación 39 – Impacto mínimo de la noche (F20)</i>	$f20 = \frac{P_{\text{promedio de la noche}}}{P_{\text{promedio del día}}}$	Rango horario de 22h00 a 07h00

<i>Ecuación 40 – Impacto máximo del día completo (F21)</i>	$f21 = P_{m\acute{a}xima\ del\ d\acute{a}a}$	24 horas del día
<i>Ecuación 41 – Impacto promedio del día completo (F22)</i>	$f22 = P_{promedio\ del\ d\acute{a}a}$	24 horas del día
<i>Ecuación 42 – Impacto mínimo del día completo (F23)</i>	$f23 = P_{m\acute{i}nima\ del\ d\acute{a}a}$	24 horas del día

Con lo anteriormente expuesto; el programa realiza el cálculo de los diversos factores de forma para definir el comportamiento de cada uno de los PCEE diarios en análisis.

Continuando con el ejemplo de los PCEE correspondientes a los días lunes se tiene la siguiente codificación:

```

F1LUNES = PMEDIALUNES./PMAXLUNES;
F2LUNES = PMINLUNES./PMEDIALUNES;
F3LUNES = P6_12LUNES./PMEDIALUNES;
F4LUNES = PMAX6_12LUNES./PMEDIALUNES;
F5LUNES = PMIN6_12LUNES./PMEDIALUNES;
F6LUNES = P12_15LUNES./PMEDIALUNES;
F7LUNES = PMAX12_15LUNES./PMEDIALUNES;
F8LUNES = PMIN12_15LUNES./PMEDIALUNES;
F9LUNES = P15_18LUNES./PMEDIALUNES;
F10LUNES = PMAX15_18LUNES./PMEDIALUNES;
F11LUNES = PMIN15_18LUNES./PMEDIALUNES;
F12LUNES = P7_18LUNES./PMEDIALUNES;
F13LUNES = PMAX7_18LUNES./PMEDIALUNES;
F14LUNES = PMIN7_18LUNES./PMEDIALUNES;
F15LUNES = P18_22LUNES./PMEDIALUNES;
F16LUNES = PMAX18_22LUNES./PMEDIALUNES;
F17LUNES = PMIN18_22LUNES./PMEDIALUNES;
F18LUNES = P22_7LUNES./PMEDIALUNES;
F19LUNES = PMAX22_7LUNES./PMEDIALUNES;
F20LUNES = PMIN22_7LUNES./PMEDIALUNES;
F21LUNES = PMAXLUNES;
F22LUNES = PMEDIALUNES;
F23LUNES = PMINLUNES;

```

El programa concatena los factores de forma de los PCEE de los días lunes en análisis, obteniendo una matriz $[N \times 23]$ denominada “FACTORFORMALUNES”; donde N es el número de PCEE correspondientes a los días lunes y 23 es el número de factores de forma que definen el comportamiento de cada uno de los PCEE diarios mencionados.

```

FACTORFORMALUNES = [F1LUNES F2LUNES F3LUNES F4LUNES F5LUNES F6LUNES...
                    F7LUNES F8LUNES F9LUNES F10LUNES F11LUNES F12LUNES...

```

F13LUNES F14LUNES F15LUNES F16LUNES F17LUNES...
F18LUNES F19LUNES F20LUNES F21LUNES F22LUNES...
F23LUNES];

El procedimiento de la obtención de los factores de forma se realiza para cada uno de los demás PCEE diarios que fueron obtenidos y agrupados mediante el reconocimiento de días de la semana y rangos horarios del programa; siendo almacenados en matrices $[N \times 23]$ denominadas “FACTORFORMALUNES”, “FACTORFORMAMARTES”, “FACTORFORMAMIERCOLES”, “FACTORFORMAJUEVES”, “FACTORFORMAVIERNES”, “FACTORFORMASABADOS” y “FACTORFORMADOMINGOS, donde N corresponde al número de PCEE diarios almacenados en cada matriz y 23 es el número de factores de forma que definen el comportamiento de cada uno de los PCEE diarios.

Una vez que se obtiene las matrices $[N \times 23]$ que contienen los factores de forma correspondientes a los PCEE diarios de todos los días de la semana, se concatenan dichas matrices en una sola matriz denominada “FACTORFORMADIASSN” con la finalidad de tener la información para el análisis de segmentación de una manera sistematizada.

FACTORFORMADIASSN = [FACTORFORMALUNES;...
FACTORFORMAMARTES;...
FACTORFORMAMIERCOLES;...
FACTORFORMAJUEVES;...
FACTORFORMAVIERNES;...
FACTORFORMASABADOS;...
FACTORFORMADOMINGOS];

De igual manera, se crea un matriz $[N \times 96]$ denominada “MATRIZDIASTOTAL”; donde N es el número de PCEE diarios agrupados y ordenados mediante el reconocimiento de días de la semana y 96 es el número de mediciones conseguidas cada 15 minutos. La codificación es la que se expresa a continuación:

MATRIZDIASTOTAL = [DLUNES';...
DMARTES';...
DMIERCOLES';...
DJUEVES';...
DVIERNES';...
DSABADOS';...
DDOMINGOS'];

3.3.2.3. Creación del archivo de Microsoft Excel que posee los factores de forma

El programa permite la creación de un archivo de Microsoft Excel en donde se almacena la matriz denominada “FACTORFORMADIASSN” que contiene los factores de forma correspondientes a todos los PCEE diarios agrupados por días de la semana que se encuentran en análisis.

La codificación para la creación del mencionado archivo se visualiza a continuación:

```
a=1;
while a==1
```

El programa consulta sobre la selección de si el usuario desea almacenar o no la matriz de los factores de forma en un archivo de Microsoft Excel.

```
prompt = '1. ¿Desea guardar los factores de forma de los perfiles de carga de energía eléctrica
diarios? Si/No: ';
str = 0;
```

El programa espera hasta que el usuario ingrese la selección.

```
while str ~= 'Si' | str ~= 'No'
    disp('...Ingrese correctamente su selección...')
    str = input(prompt,'s');
    if str == 'Si' | str == 'No'
        break
    end
end
```

Codificación del programa para la selección “Si” correspondiente al almacenamiento de los factores de forma en el Archivo de Microsoft Excel:

```
if str == 'Si'
    disp('-----')
    disp('...Su selección fue Sí, espere mientras se guarda correctamente los factores de forma...')
    disp('-----')
```

Creación de los nombres para las columnas del archivo de Microsoft Excel:

```
NOMBRESCOLUMNASFECHA = {'DÍA','MES','AÑO'};
NOMBRESCOLUMNASFACTORES = {'F1','F2','F3','F4','F5','F6','F7','F8',...
    'F9','F10','F11','F12','F13','F14','F15',...
    'F16','F17','F18','F19','F20','F21',...
    'F22','F23'};
NOMBRESCOLUMNASDIAS = {'LUNES','MARTES','MIERCOLES','JUEVES',...
    'VIERNES','SABADOS','DOMINGOS'};
```

Encadenamiento de los nombres de las columnas para el archivo de Microsoft Excel con la matriz denominada “FACTORFORMADIASSN” que contiene los factores de forma de los PCEE diarios que se encuentran en análisis:

```
FACTORFORMADIASCN2 = num2cell(FACTORFORMADIASCN2);
FACTORFORMADIASCN2 = [NOMBRESCOLUMNASFECHA
NOMBRESCOLUMNASFACTORES
FACTORFORMADIASCN2];
```

Escritura de información en la hoja de cálculo del archivo de Microsoft Excel denominado “FACTORES DE FORMA DE LOS PERFILES DE CARGA DIARIOS”:

```
[estado] = xlswrite('FACTORES DE FORMA DE LOS PERFILES DE CARGA
DIARIOS.xlsx', FACTORFORMADIASCN2,'Hoja1','A1')
```

Comprobación del almacenamiento de los factores de forma en el Archivo de Microsoft Excel:

```
if estado == 1
    disp('-----')
    disp('...Archivo de Microsoft Excel con los factores de forma guardado exitosamente..')
    disp('-----')
    a=2;
else
    disp('-----')
    disp('...Se produjo un error, vuelva a intentarlo...')
    disp('-----')
    a=1;
end
```

Codificación del programa para la selección de “No” almacenar los factores de forma en el Archivo de Microsoft Excel:

```
else str == 'No'
    disp('-----')
    disp('...Su selección fue No...')
    disp('-----')
    a=2;
end
end
```

El archivo de Microsoft Excel generado que contiene los factores de forma de los PCEE diarios en análisis es útil para realizar el procedimiento de segmentación en los Software Weka y SPSS con la finalidad de comprobar los resultados obtenidos posteriormente por parte del programa.

3.3.3. Visualización de las curvas de perfiles de carga diarios

El programa permite la visualización de las curvas de los PCEE diarios agrupados por días de la semana. Dicha visualización puede ser elegida por el usuario y la codificación es la siguiente:

El programa consulta sobre la selección de si se desea visualizar las curvas de los PCEE diarios agrupados por días de la semana.

```
prompt = '1. ¿Desea visualizar las curvas de perfiles de carga de energía eléctrica por días de la
semana? Si/No: ';
str = 0;
```

El programa espera hasta que el usuario ingrese la selección.

```
while str ~= 'Si' | str ~= 'No'
    disp('...Ingrese correctamente su selección...')
    str = input(prompt,'s');
    if str == 'Si' | str == 'No'
        break
    end
end
```

Codificación del programa para la selección de “Si” se desea visualizar las curvas de los PCEE diarios agrupados por días de la semana:

```
if str == 'Si'
    disp('-----')
    disp('...Su selección fue Sí, espere mientras se grafican las curvas de los perfiles de carga de energía eléctrica diarios...')
    disp('-----')
```

A continuación, se presenta la codificación correspondiente a la visualización de las curvas de los PCEE diario de los días lunes como modelo.

Se crea la figura donde se van a visualizar los PCEE de los días lunes.

```
figure;
```

Se obtiene el tamaño de la matriz denominada “DLUNES” que contiene los PCEE diarios correspondientes a los días lunes para poder ser extraídos a continuación.

```
[tfl, tcl] = size(DLUNES);
```

Mediante un “bucle for” se grafican los datos de las mediciones de CEE correspondientes a cada una de las curvas de los PCEE de los días lunes.

```
for i=1 : tcl
    plot(DLUNES(:,i),'-');
    hold on;
end
```

Se muestra la configuración de los títulos de la gráfica y de sus respectivos ejes; así como, la configuración del color que posee la gráfica y las propiedades de los ejes.

```
fig = gcf;
fig.Color = [1 1 0.75];
ax=gca;
ax.XTick = 1:96;
ax.XTickLabel = {'00:00','00:15','00:30','00:45',...
```

```

'01:00','01:15','01:30','01:45',...
'02:00','02:15','02:30','02:45',...
'03:00','03:15','03:30','03:45',...
'04:00','04:15','04:30','04:45',...
'05:00','05:15','05:30','05:45',...
'06:00','06:15','06:30','06:45',...
'07:00','07:15','07:30','07:45',...
'08:00','08:15','08:30','08:45',...
'09:00','09:15','09:30','09:45',...
'10:00','10:15','10:30','10:45',...
'11:00','11:15','11:30','11:45',...
'12:00','12:15','12:30','12:45',...
'13:00','13:15','13:30','13:45',...
'14:00','14:15','14:30','14:45',...
'15:00','15:15','15:30','15:45',...
'16:00','16:15','16:30','16:45',...
'17:00','17:15','17:30','17:45',...
'18:00','18:15','18:30','18:45',...
'19:00','19:15','19:30','19:45',...
'20:00','20:15','20:30','20:45',...
'21:00','21:15','21:30','21:45',...
'22:00','22:15','22:30','22:45',...
'23:00','23:15','23:30','23:45'};
ax.XTickLabelRotation = 90;
ax.FontSize = 12;
grid on;
grid minor;
xlabel '\bf \fontsize{24} Tiempo \fontsize{24} (Horas:minutos)';
ylabel '\bf \fontsize{24} Potencia \fontsize{24} (kW)';
title '\bf \color{blue} \fontsize{24} Perfiles de Carga de Energía Eléctrica de los días Lunes';

```

La codificación mencionada para la visualización de las curvas de los PCEE diario de los días lunes se realiza para cada uno de los PCEE diarios agrupados por días de la semana que fueron obtenidos mediante el reconocimiento de días de la semana y rangos horarios del programa.

Codificación del programa para la selección de “No” se desea visualizar las curvas de los PCEE diarios:

```

else str == 'No'
    disp('-----')
    disp('...Su selección fue No...')
    disp('-----')
end

```

3.3.4. Visualización de factores de forma de los perfiles de carga diarios

El programa permite la visualización de los factores de forma que caracterizan a cada una de las curvas de los PCEE diarios agrupados por días de la semana. La codificación para dicha visualización es la siguiente:

El programa consulta sobre la selección de si se desea visualizar los factores de forma que caracterizan a cada una de las curvas de los PCEE diarios.

```
prompt = '1. ¿Desea visualizar los factores de forma de los perfiles de carga de energía eléctrica diarios? Si/No: ';  
str = 0;
```

El programa espera hasta que el usuario ingrese la selección.

```
while str ~= 'Si' | str ~= 'No'  
    disp('...Ingrese correctamente su selección...')  
    str = input(prompt,'s');  
    if str == 'Si' | str == 'No'  
        break  
    end  
end
```

Codificación del programa para la selección de “Si” se desea visualizar los factores de forma que caracterizan a cada una de las curvas de los PCEE diarios:

```
if str == 'Si'  
    disp('-----')  
    disp('...Su selección fue Sí, espere mientras se grafican los factores de forma de los Perfiles de Carga de Energía Eléctrica diarios...')  
    disp('-----')
```

A continuación, se presenta la codificación correspondiente a la visualización de los factores de forma f_1 y f_2 que caracterizan a cada una de las curvas de los PCEE diarios.

Se crea la figura donde se van a visualizar los factores de forma f_1 y f_2 .

```
figure;
```

Al conocer anteriormente las dimensiones de cada una de las matrices $[N \times 23]$ denominadas “FACTORFORMALUNES”, “FACTORFORMAMARTES”, “FACTORFORMAMIERCOLES”, “FACTORFORMAJUEVES”, “FACTORFORMAVIERNES”, “FACTORFORMASABADOS”, “FACTORFORMADOMINGOS”; se procede a realizar la extracción de los factores de forma f_1 y f_2 a ser visualizados en la gráfica 2D.

Mediante diversos “bucles for” se grafican los factores de forma f_1 y f_2 de las curvas de los PCEE diarios que se encuentran agrupados por días de la semana.

Codificación para graficar los factores de forma f_1 y f_2 de los PCEE de los días lunes:

```
for n = 1: iflun
    plot(FACTORFORMALUNES(n,1),FACTORFORMALUNES(n,2),'LineStyle',...
        'none', 'Marker', 'none','Color', 'r','DisplayName','LUNES');
    text(FACTORFORMALUNES(n,1),FACTORFORMALUNES(n,2), '1','Color', 'r');
    axis([0 inf 0 inf]);
    hold on;
end
```

Codificación para graficar los factores de forma f_1 y f_2 de los PCEE de los días martes:

```
for n = 1: ifma
    plot(FACTORFORMAMARTES(n,1),FACTORFORMAMARTES(n,2),'LineStyle',...
        'none', 'Marker', 'none','Color', [0 0.5 0], 'DisplayName','MARTES');
    text(FACTORFORMAMARTES(n,1),FACTORFORMAMARTES(n,2), '2','Color',...
        [0 0.5 0]);
    axis([0 inf 0 inf]);
    %Se grafican los puntos que tienen índices cluster el valor de n
    hold on;
end
```

Codificación para graficar los factores de forma f_1 y f_2 de los PCEE de los días miércoles:

```
for n = 1: ifmie
    plot(FACTORFORMAMIERCOLES(n,1),FACTORFORMAMIERCOLES(n,2),'LineStyle',
        'none', 'Marker', 'none','Color', 'b','DisplayName','MÉRCOLES');
    text(FACTORFORMAMIERCOLES(n,1),FACTORFORMAMIERCOLES(n,2),...
        '3','Color','b');
    axis([0 inf 0 inf]);
    hold on;
end
```

Codificación para graficar los factores de forma f_1 y f_2 de los PCEE de los días jueves:

```
for n = 1: ifjue
    plot(FACTORFORMAJUEVES(n,1),FACTORFORMAJUEVES(n,2),'LineStyle',...
        'none', 'Marker', 'none', 'Color', 'm','DisplayName','JUEVES');
    text(FACTORFORMAJUEVES(n,1),FACTORFORMAJUEVES(n,2), '4','Color', 'm');
    axis([0 inf 0 inf]);
    hold on;
end
```

Codificación para graficar los factores de forma f_1 y f_2 de los PCEE de los días viernes:

```
for n = 1: ifvie
    plot(FACTORFORMAVIERNES(n,1),FACTORFORMAVIERNES(n,2),'LineStyle',...
        'none', 'Marker', 'none','Color', [1 0.88 0],'DisplayName','VIERNES');
    text(FACTORFORMAVIERNES(n,1),FACTORFORMAVIERNES(n,2), '5','Color',...
        [1 0.88 0]);
    axis([0 inf 0 inf])
    hold on;
end
```

Codificación para graficar los factores de forma f_1 y f_2 de los PCEE de los días sábados:

```
for n = 1: ifsab
    plot(FACTORFORMASABADOS(n,1),FACTORFORMASABADOS(n,2),'LineStyle',...
        'none', 'Marker', 'none','Color', [0 0.75 1],'DisplayName','SÁBADOS');
    text(FACTORFORMASABADOS(n,1),FACTORFORMASABADOS(n,2), '6','Color',...
        [0 0.75 1]);
    axis([0 inf 0 inf]);
    hold on;
end
```

Codificación para graficar los factores de forma f_1 y f_2 de los PCEE de los días domingos:

```
for n = 1: ifdom
    plot(FACTORFORMADOMINGOS(n,1),FACTORFORMADOMINGOS(n,2),'LineStyle',...
        'none', 'Marker', 'none','Color', 'k', 'DisplayName','DOMINGOS');
    text(FACTORFORMADOMINGOS(n,1),FACTORFORMADOMINGOS(n,2),...
        '7','Color', 'k');
    axis([0 inf 0 inf]);
    hold on;
end
```

Se muestra la configuración de los títulos de la gráfica y de sus respectivos ejes; así como, la configuración del color que posee la gráfica, las propiedades de los ejes y leyendas.

```
fig = gcf;
fig.Color = [1 1 0.75];
ax=gca;
ax.FontSize = 14;
grid on;
grid minor;
xlabel '\bf \fontsize{14} Potencia Media del día (kW)/Potencia Máxima del día (kW)'
ylabel '\bf \fontsize{14} Potencia Mínima del día (kW)/Potencia Media del día (kW)'
title '\bf \color{blue} \fontsize{14} Factores de Forma F1 vs F2';
legend('\color{red} \fontsize{14} 1. LUNES','\color{rgb}{0 0.5 0} \fontsize{14} 2. MARTES',...
    '\color{blue} \fontsize{14} 3. MIÉRCOLES','\color{magenta} \fontsize{14} 4. JUEVES',...
```

```

\color[rgb]{1 0.88 0} \fontsize{14} 5. VIERNES',...
\color[rgb]{0 0.75 1} \fontsize{14} 6. SÁBADOS',...
\fontsize{14} 7. DOMINGOS', 'Location', 'southoutside', 'Orientation', 'horizontal');

```

La codificación indicada para la visualización de los factores de forma f_1 y f_2 de cada una de las curvas de los PCEE diario, se realiza para los factores de forma restantes indicados en la tabla 4.

Los factores de forma restantes son representados mediante gráficas en 3D; siendo asociados de la siguiente manera:

Tabla 5. Asociación de los Factores de Forma normalizados para su representación mediante gráficas.

Número de Gráfica	Factores Asociados	Dimensiones	Período
1	$f_1 - f_2$	2D	24 horas del día
2	$f_3 - f_4 - f_5$	3D	Rango horario de 06h00 a 12h00
3	$f_6 - f_7 - f_8$	3D	Rango horario de 12h00 a 15h00
4	$f_9 - f_{10} - f_{11}$	3D	Rango horario de 15h00 a 18h00
5	$f_{12} - f_{13} - f_{14}$	3D	Rango horario de 07h00 a 18h00
6	$f_{15} - f_{16} - f_{17}$	3D	Rango horario de 18h00 a 22h00
7	$f_{18} - f_{19} - f_{20}$	3D	Rango horario de 22h00 a 07h00
8	$f_{21} - f_{22} - f_{23}$	3D	24 horas del día

Codificación del programa para la selección de “No” se desea visualizar las curvas de los PCEE diarios:

```

else str == 'No'
    disp('-----')
    disp('...Su selección fue No...')
    disp('-----')
end

```

3.3.5. Aplicación de las Técnicas de Clustering k -Means y k -Medoids

Para aplicar las TC denominadas k -Means o k -Medoids a los diversos factores de forma que caracterizan los PCEE diarios, previamente se obtiene un número de segmentaciones máximas que se pueden aplicar a dichos factores de forma basados en la Regla de Sturges.

3.3.5.1. Regla de Sturges

En estadística descriptiva existen diversos métodos para la recolección, agrupación y presentación de la información de una manera en la que sea fácil interpretarla [23].

La regla de Sturges es una regla muy utilizada al momento de calcular el número de clases (en el presente PTEI es utilizada para determinar el número de segmentaciones máximas que se pueden aplicar a los factores de forma) que son necesarios para representar exactamente los datos [23].

Ecuación 43 – Regla de Sturges

$$k = 1 + \log_2(N)$$

Donde, N es el número de muestras (PCEE diarios) y k representa el número máximo de clases (clusters) [23].

Por lo tanto, el programa realiza el conteo del total de PCEE diarios que se encuentran en análisis correspondientes a los días lunes, martes, miércoles, jueves, viernes, sábados y domingos.

```
f = iflun+ifma+ifmie+ifjue+ifvie+ifsab+ifdom;
```

Se aplica la regla de Sturges para obtener el número máximo de clusters que se pueden aplicar a los factores de forma.

```
numeromaximocluster = fix(1+log2(f))
disp('-----')
display(strcat(['EL NÚMERO MÁXIMO DE CLUSTER QUE PUEDE APLICAR BAJO LA
REGLA DE STURGES: ', num2str(numeromaximocluster)]));
disp('-----')
```

El programa permite ingresar al usuario el número de clusters n con los que se desea realizar las pruebas de segmentación de los factores de forma que caracterizan a cada uno de los PCEE diarios en análisis.

```
n=input('1. Ingrese el valor de número de clusters con los que se desea segmentar los factores de
forma en análisis: ');
```

Al ingresar el número de clusters con los que se desea realizar las pruebas de segmentación de los factores de forma, un “bucle for” permite ejecutar dichas segmentaciones en 2 clusters, luego en 3 clusters, y así sucesivamente hasta el número de clusters n que fue ingresado por el usuario.

```

for i=1:n
    if i==1
        disp('-----')
        display(strcat(['REALIZANDO SEGMENTACIÓN DE FACTORES DE FORMA CON ',
            num2str(i),' CLUSTER']));
        disp('-----')
    else
        disp('-----')
        display(strcat(['REALIZANDO SEGMENTACIÓN DE FACTORES DE FORMA CON ',
            num2str(i),' CLUSTERS']));
        disp('-----')
    end
end

```

En cada una de las pruebas de segmentación de los factores de forma se inicializa un cronómetro para medir el rendimiento computacional de las TC denominadas *k*-Means o *k*-Medoids. La función registra el tiempo interno en la ejecución del comando “tic” y muestra el tiempo transcurrido con la función “toc”.

```

tic;
timerVal(1,i) = tic;
elapsedTimekmeans(i,1) = i;
elapsedTimekmeans(i,2) = toc(timerVal(1,i));
NOMBRESTIEMPOEJECUCION = {'Número de Clusters','Tiempo de Ejecución'};
tiemposejecucion = num2cell(elapsedTimekmeans);
tiemposejecucion = [NOMBRESTIEMPOEJECUCION;tiemposejecucion];

```

3.3.5.2. Datos de entrada para las técnicas de clustering *k*-Means y *k*-Medoids

Los argumentos de entrada que se especifican para las TC denominadas *k*-Means o *k*-Medoids son los siguientes:

- Datos que en el presente caso son los factores de forma, especificados como una matriz numérica. Las filas corresponden a observaciones (Número de PCEE diarios), y las columnas corresponden a las variables (factores de forma de los PCEE diarios).
- Número de clusters a ser aplicados a los datos, especificado como un entero positivo.

3.3.5.3. Opciones de control para las técnicas de clustering *k*-Means y *k*-Medoids

Para la aplicación de la técnica de clustering *k*-Means o *k*-Medoids en cada una de las pruebas de segmentación de los factores de forma, se especifican las opciones de control del algoritmo iterativo para minimizar los criterios de ajuste desde el punto de vista de la optimización matemática.

Algunas de las opciones de control que se especifican para las TC mencionadas son las siguientes:

- **Medida de distancia**, en espacio *p*-dimensional, utilizada para la minimización, especificada como el par separado por comas que consiste en ‘Distancia’ y ‘squeclidean’, ‘cityblock’, ‘cosine’, ‘correlation’, o ‘hamming’. Las TC denominadas *k*-Means o *k*-Medoids calculan los

grupos de centroides o medoides de forma diferente para las diferentes medidas de distancia admitidas. La tabla 1 resume las medidas de distancia disponibles [8], [9].

Para el presente PTEI se utiliza como medida de distancia la distancia euclidiana cuadrática.

- **Número de veces que se repite el agrupamiento del cluster utilizando las nuevas posiciones iniciales del centroide o medoide del cluster**, que se especifica como el par separado por comas que consiste en ‘Réplicas’ y un número entero. La técnica de clustering *k*-Means o *k*-Medoids devuelven la solución en la variable “sumd” que se indica posteriormente [8], [9].

Para el presente PTEI se utiliza como número de veces que se repite el agrupamiento del cluster la cantidad de veinte y cinco repeticiones.

- **Nivel de salida para mostrar en la ventana de comandos**, especificado como el par separado por comas que consiste en ‘Pantalla’ y una de las siguientes opciones [8], [9]:
 - ‘final’ - Muestra los resultados de la iteración final.
 - ‘iter’: muestra los resultados de cada iteración.
 - ‘apagado’: no muestra nada.

Para el presente PTEI se utiliza el nivel de salida final.

- **Método para elegir las posiciones iniciales del centroide o medoide del grupo (o cluster)**, especificadas como el par separado por comas que consiste en ‘Inicio’ y ‘cluster’, ‘plus’, ‘sample’, ‘uniform’ o una matriz numérica [8], [9].

Por defecto es implementado el algoritmo *k*-means ++ para la inicialización del centroide de cada cluster [8], [9].

3.3.5.4. Distancia euclidiana

La distancia euclidiana es la distancia existente entre dos puntos de un espacio euclídeo, siendo su base el teorema de Pitágoras [24].

En general, la distancia euclidiana entre los puntos $P = (p_1, p_2, \dots, p_n)$ y $Q = (q_1, q_2, \dots, q_n)$, del espacio euclídeo *n*-dimensional, se define como [24]:

Ecuación 44 – Distancia Euclidiana

$$d_{Euclidiana}(P, Q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

De esta manera, la distancia euclidiana es la disimilitud más conocida y más sencilla de entender, pues su definición concuerda con el concepto más común de distancia [24].

3.3.5.5. Algoritmo *k*-means++

El algoritmo k -means++ usa una heurística para encontrar los centroides iniciales para la técnica de clustering k -Means o k -Medoids [8].

El algoritmo k -means++ escoge los centroides iniciales de la siguiente manera, teniendo en cuenta que el número de clusters es k [8]:

1. Se selecciona una muestra del conjunto de datos, X . Se denota como el primer centroide, c_1 .
2. Se calcula las distancias de cada observación a c_1 . Se denota la distancia entre c_j y la observación m como $d(x_m, c_j)$.
3. Se selecciona el siguiente centroide, c_2 al azar de X con probabilidad:

Ecuación 45 – Probabilidad de Selección del Centroide

$$\frac{d^2(x_m, c_1)}{\sum_{j=1}^n d^2(x_j, c_1)}$$

4. Para escoger el centro j :
 - a. Se calcula las distancias desde cada una de las observaciones hacia cada centroide y se asigna cada observación al centroide más cercano.
 - b. Para $m = 1, \dots, n$ y $p = 1, \dots, j - 1$, se selecciona aleatoriamente el centroide j de X con probabilidad:

Ecuación 46 – Probabilidad de Selección aleatoria del Centroide

$$\frac{d^2(x_m, c_p)}{\sum_{\{h; x_h \in C_p\}} d^2(x_h, c_p)}$$

Donde C_p es el conjunto de observaciones más cercanas al centroide c_p y x_m pertenece a C_p .

Es decir, se selecciona cada centro subsiguiente con una probabilidad proporcional a la distancia desde sí mismo al centro más cercano que ya se eligió.

5. Se repite el paso 4 hasta que los k clusters hayan sido elegidos.

La codificación para aplicar la técnica de clustering k -Means; bajo las opciones de control indicadas anteriormente, a los factores de forma que caracterizan a cada uno de los PCEE diarios es la siguiente:

```
opts = statset('Display','final');
[idx,C,sumd,D] = kmeans(FACTORFORMADIASSN,i,'Distance',...
    'squeclidean','Replicates',25,'Options',opts);
```

La codificación para aplicar la técnica de clustering k -Medoids; bajo las opciones de control indicadas anteriormente, a los factores de forma que caracterizan a cada uno de los PCEE diarios es la siguiente:

```
opts = statset('Display','final');
```

```
[idx,C,sumd,D] = kmedoids(FACTORFORMADIASSN,i,'Distance',...
    'sqeuclidean','Replicates',25,'Options',opts);
```

3.3.5.6. Datos de salida de las técnicas de clustering *k*-Means y *k*-Medoids

Los argumentos de salida que se obtienen al aplicar la técnica de clustering *k*-Means o *k*-Medoids son los siguientes:

- **Índices del cluster al cual pertenece cada PCEE diario**, devueltos como un vector de columna numérica. Dicho vector denominado “idx” tiene tantas filas como la cantidad total de PCEE diarios en análisis, y cada fila indica el número de cluster al que fue asignado cada PCEE diario [8], [9].
- **Ubicación de los centroides o medoides según la cantidad de clusters especificados**, devueltos en una matriz numérica denominada “C”; donde, la fila *j* indica la ubicación del centroide del cluster *j* [8], [9].
- **Suma de las distancias existentes desde cada uno de los puntos (factor de forma) perteneciente al cluster *j* a su centroide o medoide *j***. Dichas sumatorias son devueltas como un vector de columna numérica denominado “sumd” [8], [9].
- **Distancias desde cada punto (factor de forma) a todos los centroides o medoides existentes**, devuelto como una matriz numérica denominada “D”; donde, el elemento (*j*, *m*) es la distancia desde el punto *j* al centroide o medoide *m* [8], [9].

Para cada una de las pruebas de segmentación de los factores de forma que se ejecutan con 2 clusters, luego en 3 clusters, y así sucesivamente hasta el número de clusters *n* que fue ingresado por el usuario; se calcula los **VS (Valores de Silueta)** de cada uno de los puntos (factores de forma) de los diversos clusters.

3.3.5.7. Valor de silueta

El valor de silueta para cada uno de los puntos es la medida en cuan similar es ese punto en relación a los puntos que se encuentran en el mismo cluster, cuando se compara con los puntos que se encuentran en otros clusters [25].

El valor de silueta para un punto determinado *i* viene dado por [25]:

Ecuación 47 – Valor de Silueta

$$S_i = (b_i - a_i) / \max(a_i, b_i)$$

Donde, a_i es el valor de la distancia promedio desde el punto *i* a los otros puntos en el mismo cluster, y b_i es la medición mínima promedio de la distancia desde el punto *i* a los otros puntos de un cluster diferente [25].

El valor de la silueta varía de -1 a +1. Un alto valor de silueta indica que está bien adaptado a su propio cluster y que no coincide con los cluster vecinos. Si la mayoría de los puntos tienen un alto valor de silueta, entonces la solución de segmentación es apropiada [25].

La codificación para el cálculo de los VS para cada uno de los puntos (factores de forma) de los diversos clusters se muestra a continuación:

Se crea la figura donde se van a visualizar los VS de cada cluster.

```
figure;
```

La función “silhouette” devuelve los VS en un vector $[n \times 1]$, donde n es el número de PCEE diarios en análisis [26].

```
[silh,h] = silhouette(FACTORFORMADIASSN,idx,'sqeuclidean');
```

Se muestra la configuración de los títulos de la gráfica y de sus respectivos ejes; así como, la configuración del color que posee la gráfica y las propiedades de los ejes.

```
h = gca;  
h.FontSize = 24;  
h.Children.EdgeColor = [0.25 0.75 0.5];  
fig = gcf;  
fig.Color = [1 1 0.75];  
grid on;  
grid minor;  
ylabel '\bf \fontsize{24} Número de Cluster'  
xlabel '\bf \fontsize{24} Valor de Silueta'  
title '\bf \color{blue} \fontsize{24} Gráfica de Silueta';
```

La codificación mencionada en la sección 3.3.5.2 hasta la sección 3.3.5.7 es ejecutada para cada una de las pruebas de segmentación de los factores de forma hasta el número de clusters n que fue ingresado por el usuario.

3.3.6. Evaluación del número óptimo de clusters

Para seleccionar cuál es el número óptimo de clusters que se debe aplicar a los factores de forma que caracterizan a cada uno de los PCEE diarios, la evaluación se encuentra basada bajo 2 criterios:

1. Valor máximo promedio de los VS de las pruebas de segmentación.
2. Valor máximo de tiempo de ejecución de las pruebas de segmentación.

3.3.6.1. Análisis basado con respecto a los valores silueta

El valor máximo promedio de los VS de las pruebas de segmentación aplicadas a los factores de forma se lo obtiene mediante la función “evalclusters”. Dicha función evalúa las soluciones obtenidas por las TC denominadas k -Means o k -Medoids [27].

Los argumentos de entrada que se especifican para la función “evalclusters” son los siguientes [27]:

- Datos que en el presente caso son los factores de forma, especificados como una matriz numérica. Las filas corresponden a observaciones (Número de PCEE diarios), y las columnas corresponden a las variables (factores de forma de los PCEE diarios).
- Algoritmo de agrupamiento (Para el presente PTEI el algoritmo utilizado es “*k*-Means”).
- Criterio de evaluación de agrupamiento (Para el presente PTEI el criterio de evaluación son los VS).

```
disp('3. El número de Clusters más apropiados según el Valor de la Silueta:')
Ekmeans3 = evalclusters(FACTORFORMADIASSN,'kmeans','silhouette','klist',[1:f4])
```

El argumento de salida que entrega la función “evalclusters” es el número óptimo de clusters que se debe aplicar para la segmentación de los factores de forma.

3.3.6.2. Análisis basado con respecto a los tiempos de ejecución del programa

El valor máximo de tiempo de ejecución de las pruebas de segmentación aplicadas a los factores de forma se lo obtiene del vector denominado “tiemposejecucion”; siendo generalmente el valor en donde el procesamiento computacional para el resto de pruebas de segmentación es similar.

```
disp('1. Tiempos de Ejecución para el número de clusters analizados:')
tiemposejecucion
```

Se crea la figura donde se van a visualizar los tiempos de ejecución para cada una de las pruebas de segmentación.

```
figure;
```

Codificación para graficar los tiempos de ejecución para cada una de las pruebas de segmentación:

```
plot(elapsedTimekmeans(:,1),elapsedTimekmeans(:,2),...
'LineStyle','--','Marker','o','MarkerSize',5,'LineWidth',1,...
'DisplayName','Tiempos de Ejecución de las segmentaciones');
```

Se muestra la configuración de los títulos de la gráfica y de sus respectivos ejes; así como, la configuración del color que posee la gráfica y las propiedades de los ejes.

```
fig = gcf;
fig.Color = [1 1 0.75];
ax=gca;
ax.FontSize = 24;
ax.XTick = 0:1:n;
xlabel '\bf \fontsize{24} Número de Clusters'
ylabel '\bf \fontsize{24} Tiempo de Ejecución (segundos)'
xlim ([1 i]);
t1 = title(strcat(['\bf \color{blue} \fontsize{24} Tiempos de Ejecución según el número de cluster
k-Means']));
grid on;
```

```
grid minor;  
hold on;
```

3.3.7. Agrupamiento de Datos en base al número óptimo de clusters

Con los criterios de evaluación mencionados en la sección 3.3.6, el programa consulta al usuario el número de clusters n con los que se desea realizar finalmente la segmentación de los factores de forma que caracterizan a cada uno de los PCEE diarios en análisis.

```
NUMERO_CLUSTERS_APROPIADOS=input('1. Ingrese el valor del número de clusters k-  
Means óptimo a partir de los criterios de evaluación: ');
```

Se aplica la técnica de clustering k -Means o k -Medoids con el número de clusters n con los que se desea realizar finalmente la segmentación de los datos y con las opciones de control del algoritmo iterativo especificados anteriormente en las pruebas de segmentación.

```
opts = statset('Display','iter');  
disp('-----')  
disp('Procesando segmentación con el número de clusters k-Means óptimo a los datos en análisis')  
disp('-----')
```

```
[idx,C,sumd,D]=kmeans(FACTORFORMADIASSN,NUMERO_CLUSTERS_APROPIAD  
OS, 'Distance','sqeuclidean','Replicates',25, 'Options',opts);
```

```
opts = statset('Display','iter');  
disp('-----')  
disp('Procesando segmentación con el número de clusters k-Medoids óptimo a los datos en  
análisis')  
disp('-----')
```

```
[idx,C,sumd,D]=kmedoids(FACTORFORMADIASSN,NUMERO_CLUSTERS_APROPI  
ADOS, 'Distance','sqeuclidean','Replicates',25, 'Options',opts);
```

3.3.8. Visualización de los factores de forma agrupados en su respectivo cluster

El programa permite la visualización de los factores de forma que caracterizan a cada una de las curvas de los PCEE diarios agrupados en su respectivo cluster. La codificación para dicha visualización es la siguiente:

El programa consulta sobre la selección de si se desea visualizar los factores de forma que caracterizan a cada una de las curvas de los PCEE diarios agrupados en su respectivo cluster.

```
prompt = '1. ¿Desea visualizar los factores de forma en su respectivo cluster? Si/No: '  
str = 0;
```

El programa espera hasta que el usuario ingrese la selección.

```

while str ~= 'Si' | str ~= 'No'
    disp('...Ingrese correctamente su selección...')
    str = input(prompt,'s');
    if str == 'Si' | str == 'No'
        break
    end
end
end

```

Codificación del programa para la selección de “Si” se desea visualizar los factores de forma agrupados en su respectivo cluster:

```

if str == 'Si'

```

A continuación, se presenta la codificación correspondiente a la visualización de los factores de forma f_1 y f_2 que caracterizan a cada una de las curvas de los PCEE diarios agrupados en su respectivo cluster.

Se crea la figura donde se van a visualizar los factores de forma f_1 y f_2 .

```

marcadores='^o*.xsd+v><ph^o*.xsd+v><ph';
color='bmgrcykbmgrcykbmgrcyk';
figure;

```

Al conocer anteriormente las dimensiones de la matriz $[N \times 23]$ denominada “FACTORFORMADIASSN”; donde N corresponde al número de PCEE diarios en análisis y 23 es el número de factores de forma que definen el comportamiento de cada uno de los PCEE diarios, se procede a realizar la extracción de los factores de forma f_1 y f_2 agrupados en su respectivo cluster a ser visualizados en la gráfica 2D.

Mediante un “bucle for” se grafica los factores de forma f_1 y f_2 agrupados en su respectivo cluster.

Codificación para graficar los factores de forma f_1 y f_2 de los PCEE diarios agrupados en su respectivo cluster con su respectivo centroide o medoide:

```

for n = 1: NUMERO_CLUSTERS_APROPIADOS
    pos_marcadores=n; pos_color=n;
    plot(FACTORFORMADIASSN(idx==n,1),...
        FACTORFORMADIASSN(idx==n,2),...
        'LineStyle', 'none', 'Marker',...
        marcadores(pos_marcadores),'Color', color(pos_color),...
        'DisplayName', strcat(['Cluster ', num2str(n)]));
    hold on;
end
plot(C(:,1),C(:,2),'k*','MarkerSize',10,'LineWidth',3,...
    'DisplayName','Centroides');

```

Se muestra la configuración de los títulos de la gráfica y de sus respectivos ejes; así como, la configuración del color que posee la gráfica, las propiedades de los ejes y leyendas.

```
t1 = title(strcat(['\bf \color{blue} \fontsize{24}Asignación de ',...
    num2str(NUMERO_CLUSTERS_APROPIADOS),'\bf \color{blue} \fontsize{24}Clusters y
    ',(NUMERO_CLUSTERS_APROPIADOS), '\bf \color{blue} \fontsize{24}Centroides']));
fig =(gcf; % current figure handle
fig.Color = [1 1 0.75];
ax=gca;
ax.FontSize = 14;
legend('show')
grid on;
grid minor;
xlabel '\bf \fontsize{14} Potencia Media del día (kW)/Potencia Máxima del día (kW)'
ylabel '\bf \fontsize{14} Potencia Mínima del día (kW)/Potencia Media del día (kW)'
```

La codificación indicada para la visualización de los factores de forma f_1 y f_2 agrupados en su respectivo cluster con su respectivo centroide o medoide, se realiza para los factores de forma restantes indicados en la tabla 4.

Los factores de forma restantes son representados mediante gráficas en 3D; siendo asociados según lo mencionado en la tabla 5.

Codificación del programa para la selección de “No” se desea visualizar las curvas de los PCEE diarios:

```
% Su selección fue NO
else str == 'No' %#ok<STCMP>
    disp(' ') %Se muestra un espacio en blanco
    disp('-----')
    disp('...Su selección fue NO...')
    disp('-----')
end
pause(3);
```

3.3.9. Detección de outliers y clasificación de consumos de energía eléctrica diarios

El programa permite realizar la detección de outliers y clasificación de CEE diarios de cada cluster. Para dicha detección de datos atípicos y la respectiva clasificación de PCEE diarios sin outliers y con outliers existen varios condicionamientos basados en:

- Análisis estadístico descriptivo de las **DE (Distancias Euclidianas)** a sus respectivos centroides o medoides y los VS de los factores de forma de los datos de cada cluster.
- Evaluación de porcentajes de las funciones de distribución acumuladas inversas de las DE y VS de los datos de cada cluster.
- Aplicación de MSV No Supervisadas a los datos de cada cluster.

El programa consulta sobre selección de si se desea realizar la detección de outliers y clasificación de CEE diarios de cada cluster.

```
prompt = '1. ¿Desea agrupar las curvas de los Perfiles de Carga de Energía Eléctrica diarios según el cluster al que pertenecen con Outliers y sin Outliers? Si/No: ';  
str = 0;
```

El programa espera hasta que el usuario ingrese la selección.

```
while str ~= 'Si' | str ~= 'No'  
    disp('...Ingrese correctamente su selección...')  
    str = input(prompt,'s');  
    if str == 'Si' | str == 'No'        break  
end  
end
```

Codificación del programa para la selección de “Si” se desea realizar la detección de outliers y clasificación de CEE diarios de cada cluster:

```
if str == 'Si'
```

Se obtiene el tamaño del vector “idx” que contiene la numeración que indican a que cluster pertenece cada PCEE diario para poder extraer los datos de medición de dichos PCEE diarios posteriormente.

```
[tfC, tcC] = size(idx);
```

3.3.9.1. Análisis estadístico descriptivo de las distancias euclidianas y los valores de siluetas de los factores de forma

La estadística descriptiva ayuda a comprender la estructura de los datos, de manera de detectar tanto un patrón de comportamiento general como alejamientos del mismo. Esta permite organizar, resumir y presentar datos de manera informativa [28].

De esta manera, se trata de comprender la estructura de los VS y las distancias a los centroides o medoides de los factores de forma que pertenecen a cada uno de los clusters.

Mediante un “bucle for” el programa realiza la extracción de los VS y de las DE de cada punto (factores de forma) al centroide o medoide de cada cluster.

```
for j=1 : NUMERO_CLUSTERS_APROPIADOS  
    k = 1;  
    l = 1;  
    n=1;  
    for i=1 : tfC  
        if idx(i,tcC)==j  
            MATRIZDIASCLUSTER(n,:) = MATRIZDIASTOTAL(i,:);  
            MATRIZDIASCLUSTERCONFECCHA(n,:) = MATRIZDIASTOTAL2(i,:);
```

```

    FACTORFORMADIASCNCLUSTER(n,:) = FACTORFORMADIASCN(i,:);
    FACTORFORMADIASSNCLUSTER(n,:) = FACTORFORMADIASSN(i,:);
    DISTANCIACENTROIDE(n,:) = D(i,j);
    VALORSILUETA(n,:) = silh_final(i,:);
    indicadorclase(n,:) = idx(i,tcC);
    n=n+1;
end
end

```

Se obtiene el valor promedio de las DE existentes desde cada uno de los puntos que pertenecen a determinado cluster hacia su respectivo centroide o medoide.

Para determinar el valor promedio de las DE, se aplica la siguiente fórmula [28]:

Ecuación 48 – Valor Promedio de las Distancias Euclidianas

$$\mu_{DE} = \frac{\sum X_{DE}}{N_{DE}}$$

en donde:

- μ_{DE} representa el valor promedio de las DE; se trata de la letra minúscula griega mu.
- N_{DE} es el número de valores de las DE.
- X_{DE} representa el valor de cada distancia euclidiana.
- \sum es la letra mayúscula griega sigma e indica la operación de suma.
- $\sum X_{DE}$ es la suma de X valores de las DE.

La codificación del programa para determinar el valor promedio de las DE de cada cluster es la siguiente:

```
mediaaritmeticaDC = mean(DISTANCIACENTROIDE,1);
```

A continuación, el programa obtiene el valor de desviación estándar de las DE existentes desde cada uno de los puntos que pertenecen a determinado cluster hacia su centroide o medoide.

La desviación estándar se calcula mediante la fórmula [28]:

Ecuación 49 – Desviación Estándar de las Distancias Euclidianas

$$\sigma_{DE} = \sqrt{\frac{\sum (X_{DE} - \mu_{DE})^2}{N_{DE}}}$$

en donde:

- X_{DE} es el valor de cada distancia euclidiana.
- μ_{DE} representa el valor promedio de las DE; se trata de la letra minúscula griega mu.

- N_{DE} es el número de valores de las DE.
- σ_{DE} representa el valor de desviación estándar de las DE.

La codificación del programa para determinar la desviación estándar de las DE de cada cluster es la siguiente:

`desviacionestandarDC = std(DISTANCIACENTROIDE,1,1);`

Luego, se obtiene el valor del rango intercuartil de las DE existentes desde cada uno de los puntos que pertenecen a determinado cluster hacia su centroide o medoide. El rango intercuartílico es la distancia entre los percentiles 75% y 25% [28].

Los percentiles son los valores de un conjunto de datos ordenados (de mínimo a máximo) que dividen los datos en cien intervalos de frecuencias aproximadamente iguales. Por lo tanto, el rango intercuartil es el valor absoluto de la diferencia numérica entre los percentiles 75% y 25%. Cincuenta por ciento de los valores de una distribución se presentan en este rango [28].

La codificación del programa para determinar el rango intercuartil de las DE de cada cluster es la siguiente:

`rangointercuartilDC=iqr(DISTANCIACENTROIDE,1);`

Seguidamente, se obtiene el valor promedio de los VS de los factores de forma que pertenecen a determinado cluster.

Para determinar el valor promedio de los VS, se aplica la siguiente fórmula [28]:

Ecuación 50 – Valor Promedio de los Valores de Silueta

$$\mu_{VS} = \frac{\sum X_{VS}}{N_{VS}}$$

en donde:

- μ_{VS} representa el valor promedio de los VS; se trata de la letra minúscula griega mu.
- N_{VS} es el número de valores de los VS.
- X_{VS} representa el valor de cada valor de silueta.
- \sum es la letra mayúscula griega sigma e indica la operación de suma.
- $\sum X_{VS}$ es la suma de X valores de silueta.

La codificación del programa para determinar el valor promedio de los VS de cada cluster es la siguiente:

`mediaaritmeticaVS = mean(VALORSILUETA,1);`

A continuación, el programa obtiene el valor de desviación estándar de los VS de cada cluster.

La desviación estándar se calcula mediante la fórmula [28]:

Ecuación 51 – Desviación Estándar de los Valores de Silueta

$$\sigma_{VS} = \sqrt{\frac{\sum(X_{VS} - \mu_{VS})^2}{N_{VS}}}$$

en donde:

- X_{VS} es el valor de silueta.
- μ_{VS} representa el valor promedio de los VS; se trata de la letra minúscula griega mu.
- N_{VS} es el número de VS.
- σ_{VS} representa el valor de desviación estándar de los VS.

La codificación del programa para determinar la desviación estándar de los VS de cada cluster es la siguiente:

```
desviacionestandarVS = std(VALORSILUETA,1,1);
```

Se obtiene el valor del rango intercuartil de los VS de los factores de forma que caracterizan a cada PCEE diario que pertenecen a determinado cluster. El rango intercuartílico es la distancia entre los percentiles 75% y 25% [28].

Los percentiles son los valores de un conjunto de datos ordenados (de mínimo a máximo) que dividen los datos en cien intervalos de frecuencias aproximadamente iguales. Por lo tanto, el rango intercuartil es el valor absoluto de la diferencia numérica entre los percentiles 75% y 25%. Cincuenta por ciento de los valores de una distribución se presentan en este rango [28].

La codificación del programa para determinar el rango intercuartil de los VS de cada cluster es la siguiente:

```
rangointercuartilVS=iqr(VALORSILUETA,1);
```

Los valores obtenidos en la presente sección son utilizados posteriormente para determinar las funciones de densidad de probabilidad para las DE y los VS, además del condicionamiento base para la detección de outliers.

3.3.9.2. Probabilidad por Kernel de las distancias euclidianas y valores de silueta

El programa determina las funciones de densidad de probabilidad para las DE y los VS de cada cluster. En la sección 2.1.8 se puede observar la estimación de densidad de probabilidad por Kernel que es utilizada.

En primer lugar, para las DE de cada cluster se determina la amplitud de ventana.

```
minimoDC=min(desviacionestandarDC, rangointercuartilDC/1.34);  
hDC=0.9*minimoDC*length(DISTANCIACENTROIDE)^(-1/5);
```

A continuación, se realiza el ajuste de la distribución de probabilidad a los datos. La distribución de probabilidad utilizada se basa en el estimador de densidad por Kernel; ya que, se ajusta a cualquier conjunto de datos, siendo el estimador de densidad por Kernel en base a la función Epanechnikov que tiene una eficiencia de 1 [29].

```
pdDC = fitdist(DISTANCIACENTROIDE,'Kernel','Kernel','epanechnikov','Width',hDC);
xDC = min(DISTANCIACENTROIDE):1:max(DISTANCIACENTROIDE);
```

Finalmente, se determina la FDP para los datos correspondientes a las DE.

```
yDC = pdf(pdDC,xDC);
```

De esta manera, se obtiene un primer límite para determinar si algún PCEE diario de un determinado cluster contiene outliers, mediante el 90% de valores que contiene acumulados la función de distribución acumulada inversa en base al criterio de las DE.

```
limiteatipicos1 = icdf(pdDC,0.9)
```

En segundo lugar, para los VS de cada cluster se determina la amplitud de ventana.

```
minimoVS=min(desviacionestandarVS, rangointercuartilVS/1.34);
hVS=0.9*minimoVS*length(VALORSILUETA)^(-1/5);
```

A continuación, se realiza el ajuste de la distribución de probabilidad a los datos. La distribución de probabilidad utilizada se basa en el estimador de densidad por Kernel; ya que, se ajusta a cualquier conjunto de datos, siendo el estimador de densidad por Kernel en base a la función Epanechnikov que tiene una eficiencia de 1 [29].

```
pdVS = fitdist(VALORSILUETA,'Kernel','Kernel','epanechnikov','Width',hVS);
xVS = min(VALORSILUETA):0.001:max(VALORSILUETA);
```

Finalmente, se determina la FDP para los datos correspondientes a los VS.

```
yVS = pdf(pdVS,xVS);
```

De esta manera, se obtiene un segundo límite para determinar si algún PCEE diario de un determinado cluster contiene outliers, mediante el 10% de valores que contiene acumulados la función de distribución acumulada inversa en base al criterio de los VS.

```
limiteatipicos2 = icdf(pdVS,0.1)
```

3.3.9.3. Indicador de outliers en base a las distancias euclidianas y valores de silueta

Mediante los límites obtenidos en la sección 3.3.9.2, se obtiene indicadores fundamentales de outliers que permiten conocer si un PCEE diario de un determinado cluster contiene VA. Dichas condiciones complementan las condiciones bases para la detección de outliers final.

Las condiciones que se deben cumplir para determinar los PCEE diarios que contiene outliers son la siguientes:

- Si el valor promedio de los VS para los datos del cluster j es mayor a 0.7, y para que los PCEE diarios contengan outliers deben cumplir las siguientes condiciones [30]:
 - El valor de la distancia euclidiana de los factores de forma que caracterizan al PCEE diario i debe ser mayor o igual al primer límite o el valor de su silueta debe ser menor o igual al segundo límite.

```

if mediaaritmeticaVS >= 0.7
  for f=1 : k
    if DISTANCIACENTROIDE(f,:) >= limiteatipicos1 | VALORSILUETA(f,:) <=
      limiteatipicos2
      indicadoroutlier(f,1) = 1;
    else
      indicadoroutlier(f,1) = 0;
    end
  end
end

```

- Si el valor promedio de silueta para los datos del cluster j es menor a 0.7, y para que los PCEE diarios contengan outliers deben cumplir las siguientes condiciones [30]:
 - El valor de la distancia euclidiana de los factores de forma que caracterizan al PCEE diario i debe ser mayor o igual al primer límite o el valor de su silueta debe ser menor o igual a 0.4 [30].

```

elseif mediaaritmeticaVS < 0.7
  for f=1 : k
    if DISTANCIACENTROIDE(f,:) >= limiteatipicos1 | VALORSILUETA(f,:) <= 0.4
      indicadoroutlier(f,1) = 1;
    else
      indicadoroutlier(f,1) = 0;
    end
  end
end
end

```

3.3.9.4. Aplicación de máquinas de soporte vectorial no supervisadas

El programa aplica MSV no supervisadas a los factores de forma con la finalidad de determinar los PCEE diarios de cada cluster que contiene outliers. Esto se consigue mediante el hiperplano que separa la clase de datos típicos de la clase de datos atípicos indicada en la sección 2.1.7 [31].

```

SVMModel = fitsvm(FACTORFORMADIASSNCLUSTER,indicadorclase,...
  'KernelFunction','gaussian','BoxConstraint',1,...
  'KernelScale','auto','Standardize',true,'Solver',...
  'SMO','Alpha',0.5*ones(size(FACTORFORMADIASSNCLUSTER,1),1),...

```

'Verbose',1,'Weights',VALORSILUETA,'OutlierFraction',0.05);

Los argumentos de salida que se obtiene de la aplicación de MSV no supervisadas son:

- **Indicador de si el PCEE diario contiene outliers.** Si el indicador presenta un valor de 1 quiere decir que el PCEE diario contiene outliers, y si el indicador presenta un valor de 0 quiere decir que el PCEE diario no contiene outliers [31].

svInd = SVMModel.IsSupportVector;

- **Valor de gradiente de los factores de forma** que caracterizan a cada uno de los PCEE diario de cada cluster hacia la superficie del hiperplano creado por las MSV no supervisadas [31].

svGradiente = SVMModel.Gradient;

Los valores obtenidos en la presente sección son utilizados posteriormente para complementar las condiciones bases para la detección de outliers final y conseguir la clasificación de los PCEE diarios con y sin outliers de cada cluster.

3.3.9.5. Condiciones base para la detección de outliers y la respectiva clasificación de perfiles de carga de energía eléctrica diarios con y sin valores atípicos de cada cluster

Como se menciona con anterioridad; para la detección de datos atípicos y la respectiva clasificación de PCEE diarios con outliers y sin outliers de cada cluster, el programa aplica varios condicionamientos mencionados en las secciones 3.3.9.1, 3.3.9.2, 3.3.9.3 y 3.3.9.4.

Las condiciones finales que se deben cumplir para determinar si un PCEE diario contiene outliers son las siguientes:

- Si el indicador de MSV es igual a 1, y el indicador de outliers obtenido mediante los límites procedentes de las **FDAI (Función de Distribución Acumulativa Inversa)** por Kernel de las DE y VS es igual a 1, y valor del gradiente a la superficie del hiperplano es menor a 1; el PCEE diario contiene outliers.
- Si el indicador de MSV es igual a 1, y el indicador de outliers obtenido mediante los límites procedentes de las FDAI por Kernel de las DE y VS es igual a 0, y valor del gradiente a la superficie del hiperplano es menor a 1; el PCEE diario contiene outliers.
- Si el indicador de MSV es igual a 0, y el indicador de outliers obtenido mediante los límites procedentes de las FDAI por Kernel de las DE y VS es igual a 1, y valor del gradiente a la superficie del hiperplano es menor a 1; el PCEE diario contiene outliers.
- Si el indicador de MSV es igual a 1, y el indicador de outliers obtenido mediante los límites procedentes de las FDAI por Kernel de las DE y VS es igual a 1, y valor del gradiente a la superficie del hiperplano es menor a la diferencia entre la media de los valores de gradiente y la desviación estándar de los valores de gradiente; el PCEE diario contiene outliers.

```

for i=1 : k
    if (svInd(i,1)==1 & indicadoroutlier(i,1)==1 & svGradiente(i,1)<1) | ((svInd(i,1)==1 &
    indicadoroutlier(i,1)==0 & svGradiente(i,1)<1) | (svInd(i,1)==0 &
    indicadoroutlier(i,1)==1 & svGradiente(i,1)<1)) | ((svInd(i,1)==1 &
    indicadoroutlier(i,1)==1 & (svGradiente(i,1)<(mean(svGradiente)-std(svGradiente))))))
        PERFILES_CON_OUTLIERS(m,:) = MATRIZDIASCLUSTER(i,:);
        m=m+1;
    end
end

```

Las condiciones finales que se deben cumplir para determinar si un PCEE diario no contiene outliers son las siguientes:

- Si el indicador de MSV es igual a 0, y el indicador de outliers obtenido mediante los límites procedentes de las FDAI por Kernel de las DE y VS es igual a 0, y valor del gradiente a la superficie del hiperplano es mayor o igual a 1; el PCEE diario contiene outliers.
- Si el indicador de MSV es igual a 1, y el indicador de outliers obtenido mediante los límites procedentes de las FDAI por Kernel de las DE y VS es igual a 0, y valor del gradiente a la superficie del hiperplano es mayor o igual a 1; el PCEE diario contiene outliers.
- Si el indicador de MSV es igual a 0, y el indicador de outliers obtenido mediante los límites procedentes de las FDAI por Kernel de las DE y VS es igual a 1, y valor del gradiente a la superficie del hiperplano es mayor o igual a 1; el PCEE diario contiene outliers.
- Si el indicador de MSV es igual a 1, y el indicador de outliers obtenido mediante los límites procedentes de las FDAI por Kernel de las DE y VS es igual a 1, y valor del gradiente a la superficie del hiperplano es mayor o igual a la diferencia entre la media de los valores de gradiente y la desviación estándar de los valores de gradiente; el PCEE diario contiene outliers.

```

elseif (svInd(i,1)==0 & indicadoroutlier(i,1)==0 & svGradiente(i,1)>=1) |
((svInd(i,1)==1 & indicadoroutlier(i,1)==0 & svGradiente(i,1)>=1) | (svInd(i,1)==0 &
indicadoroutlier(i,1)==1 & svGradiente(i,1)>=1)) | ((svInd(i,1)==1 &
indicadoroutlier(i,1)==1 & (svGradiente(i,1)>=(mean(svGradiente)-std(svGradiente))))))
    PERFILES_SIN_OUTLIERS(n,:) = MATRIZDIASCLUSTER(i,:);
    n=n+1;
end
end

```

3.3.9.6. Visualización de perfiles de carga diarios sin outliers y con outliers de cada cluster

El programa permite la visualización los PCEE diarios con y sin outliers de cada uno de los clusters con los que se segmentaron los factores de forma y bajo los diversos criterios aplicados para la detección de dichos VA indicados entre las secciones 3.3.9.1 y 3.3.9.5. La codificación para dicha visualización es la siguiente:

- **Codificación para graficar los PCEE diarios sin outliers de cada cluster:**

Se obtiene el tamaño de la matriz denominada “PERFILES_SIN_OUTLIERS” que contiene los PCEE diarios sin outliers correspondientes a cada uno de los clusters.

```
[k, l] = size(PERFILES_SIN_OUTLIERS);
```

Se crea la figura donde se van a visualizar las curvas de los PCEE diarios sin outliers de cada cluster.

```
figure;
```

Mediante un “bucle for” se grafican los datos de las mediciones de CEE correspondientes a cada una de las curvas de los PCEE diarios sin outliers de cada cluster.

```
for i=1 : k
    subplot(2,1,1);
    plot(PERFILES_SIN_OUTLIERS(i,:),'-');
    hold on;
end
```

Se muestra la configuración de los títulos de la gráfica y de sus respectivos ejes; así como, la configuración del color que posee la gráfica y las propiedades de los ejes.

```
title(...
    strcat(['\bf \color{blue} \fontsize{14}Agrupación de Perfiles de Carga de Energía Eléctrica
    diarios sin Outliers del Cluster ', num2str(j)]));
fig = gcf; % current figure handle
fig.Color = [1 1 0.75];
ax=gca;
ax.XTick = 1:96;
ax.XTickLabel = {'00:00','00:15','00:30','00:45',...
    '01:00','01:15','01:30','01:45',...
    '02:00','02:15','02:30','02:45',...
    '03:00','03:15','03:30','03:45',...
    '04:00','04:15','04:30','04:45',...
    '05:00','05:15','05:30','05:45',...
    '06:00','06:15','06:30','06:45',...
    '07:00','07:15','07:30','07:45',...
    '08:00','08:15','08:30','08:45',...
    '09:00','09:15','09:30','09:45',...
    '10:00','10:15','10:30','10:45',...
    '11:00','11:15','11:30','11:45',...
    '12:00','12:15','12:30','12:45',...
    '13:00','13:15','13:30','13:45',...
    '14:00','14:15','14:30','14:45',...
    '15:00','15:15','15:30','15:45',...
    '16:00','16:15','16:30','16:45',...}
```

```

'17:00','17:15','17:30','17:45',...
'18:00','18:15','18:30','18:45',...
'19:00','19:15','19:30','19:45',...
'20:00','20:15','20:30','20:45',...
'21:00','21:15','21:30','21:45',...
'22:00','22:15','22:30','22:45',...
'23:00','23:15','23:30','23:45');
ax.XTickLabelRotation = 90;
ax.FontSize = 12;
grid on;
grid minor;
xlabel '\bf \fontsize{14} Tiempo \fontsize{14} (Horas:minutos)';
ylabel '\bf \fontsize{14} Potencia \fontsize{14} (kW)';
hold on;

```

- **Codificación para graficar los PCEE diarios con outliers de cada cluster:**

Se obtiene el tamaño de la matriz denominada “PERFILES_CON_OUTLIERS” que contiene los PCEE diarios con outliers correspondientes a cada uno de los clusters.

```
[k, l] = size(PERFILES_CON_OUTLIERS);
```

Mediante un “bucle for” se grafican los datos de las mediciones de CEE correspondientes a cada una de las curvas de los PCEE diarios con outliers de cada cluster.

```

for i=1 : k
    subplot(2,1,2);
    plot(PERFILES_CON_OUTLIERS(i,:),'-');
    hold on;
end

```

Se muestra la configuración de los títulos de la gráfica y de sus respectivos ejes; así como, la configuración del color que posee la gráfica y las propiedades de los ejes.

```

title(...
strcat(['\bf \color{blue} \fontsize{14} Agrupación de Perfiles de Carga de Energía Eléctrica
diarios con Outliers del Cluster ', num2str(j)]));
fig = gcf; % current figure handle
fig.Color = [1 1 0.75];
ax=gca;
ax.XTick = 1:96;
ax.XTickLabel = {'00:00','00:15','00:30','00:45',...
'01:00','01:15','01:30','01:45',...
'02:00','02:15','02:30','02:45',...
'03:00','03:15','03:30','03:45',...
'04:00','04:15','04:30','04:45',...
'05:00','05:15','05:30','05:45',...

```

```

'06:00','06:15','06:30','06:45',...
'07:00','07:15','07:30','07:45',...
'08:00','08:15','08:30','08:45',...
'09:00','09:15','09:30','09:45',...
'10:00','10:15','10:30','10:45',...
'11:00','11:15','11:30','11:45',...
'12:00','12:15','12:30','12:45',...
'13:00','13:15','13:30','13:45',...
'14:00','14:15','14:30','14:45',...
'15:00','15:15','15:30','15:45',...
'16:00','16:15','16:30','16:45',...
'17:00','17:15','17:30','17:45',...
'18:00','18:15','18:30','18:45',...
'19:00','19:15','19:30','19:45',...
'20:00','20:15','20:30','20:45',...
'21:00','21:15','21:30','21:45',...
'22:00','22:15','22:30','22:45',...
'23:00','23:15','23:30','23:45'};
ax.XTickLabelRotation = 90;
ax.FontSize = 12;
grid on;
grid minor;
xlabel '\bf \fontsize{ 14} Tiempo \fontsize{ 14} (Horas:minutos)';
ylabel '\bf \fontsize{ 14} Potencia \fontsize{ 14} (kW)';
hold on;

```

3.3.9.7. Obtención del patrón de consumo promedio de energía eléctrica diario de cada cluster

El programa obtiene el **PCPEE (Patrón de Consumo Promedio de Energía Eléctrica)** diario de cada cluster. Dicho patrón se obtiene mediante el cálculo del promedio de los datos de medición de los PCEE sin outliers diario de cada cluster. La codificación para el cálculo del patrón de consumo promedio de cada cluster es la siguiente:

```

MEDIA_PERFILES_SIN_OUTLIERS = (mean(PERFILES_SIN_OUTLIERS,1));
v=genvarname(['Patron de Consumo del Cluster ',num2str(j)]); %#ok<DEPGENAM>

```

El programa permite la visualización de dicho PCPEE diario de cada cluster mediante la siguiente codificación:

Se crea la figura donde se van a visualizar el patrón de CEE diario de cada cluster.

```
figure;
```

Se grafican los datos correspondientes al PCPEE diario de cada cluster.

```
plot(MEDIA_PERFILES_SIN_OUTLIERS,'-');
```

Se muestra la configuración de los títulos de la gráfica y de sus respectivos ejes; así como, la configuración del color que posee la gráfica y las propiedades de los ejes.

```

title(...
strcat(['\bf \color{blue} \fontsize{14}Patrón de Consumo Promedio de los Perfiles de Carga
de Energía Eléctrica diarios del Cluster ', num2str(j)]));
fig = gcf; % current figure handle
fig.Color = [1 1 0.75];
ax=gca;
ax.XTick = 1:96;
ax.XTickLabel = {'00:00','00:15','00:30','00:45',...
'01:00','01:15','01:30','01:45',...
'02:00','02:15','02:30','02:45',...
'03:00','03:15','03:30','03:45',...
'04:00','04:15','04:30','04:45',...
'05:00','05:15','05:30','05:45',...
'06:00','06:15','06:30','06:45',...
'07:00','07:15','07:30','07:45',...
'08:00','08:15','08:30','08:45',...
'09:00','09:15','09:30','09:45',...
'10:00','10:15','10:30','10:45',...
'11:00','11:15','11:30','11:45',...
'12:00','12:15','12:30','12:45',...
'13:00','13:15','13:30','13:45',...
'14:00','14:15','14:30','14:45',...
'15:00','15:15','15:30','15:45',...
'16:00','16:15','16:30','16:45',...
'17:00','17:15','17:30','17:45',...
'18:00','18:15','18:30','18:45',...
'19:00','19:15','19:30','19:45',...
'20:00','20:15','20:30','20:45',...
'21:00','21:15','21:30','21:45',...
'22:00','22:15','22:30','22:45',...
'23:00','23:15','23:30','23:45'};
ax.XTickLabelRotation = 90;
ax.FontSize = 12;
grid on;
grid minor;
xlabel '\bf \fontsize{14} Tiempo \fontsize{14} (Horas:minutos)';
ylabel '\bf \fontsize{14} Potencia \fontsize{14} (kW)';
hold on;

```

3.3.10. Almacenamiento de los perfiles de carga sin outliers y con outliers

El programa almacena automáticamente los PCEE diarios con y sin outliers de cada uno de los clusters en un archivo de Microsoft Excel denominado

“PERFILES_DE_CARGA_SEGMENTADOS_Y_CLASIFICADOS”. La codificación para dicha visualización es la siguiente:

- **Codificación para almacenar los PCEE diarios sin outliers de cada cluster:**

```
NOMBRESCOLUMNASFECHA = {'DÍA','MES','AÑO'};
NUMERODATOPOTENCIA = {'00:00','00:15','00:30','00:45',...
    '01:00','01:15','01:30','01:45',...
    '02:00','02:15','02:30','02:45',...
    '03:00','03:15','03:30','03:45',...
    '04:00','04:15','04:30','04:45',...
    '05:00','05:15','05:30','05:45',...
    '06:00','06:15','06:30','06:45',...
    '07:00','07:15','07:30','07:45',...
    '08:00','08:15','08:30','08:45',...
    '09:00','09:15','09:30','09:45',...
    '10:00','10:15','10:30','10:45',...
    '11:00','11:15','11:30','11:45',...
    '12:00','12:15','12:30','12:45',...
    '13:00','13:15','13:30','13:45',...
    '14:00','14:15','14:30','14:45',...
    '15:00','15:15','15:30','15:45',...
    '16:00','16:15','16:30','16:45',...
    '17:00','17:15','17:30','17:45',...
    '18:00','18:15','18:30','18:45',...
    '19:00','19:15','19:30','19:45',...
    '20:00','20:15','20:30','20:45',...
    '21:00','21:15','21:30','21:45',...
    '22:00','22:15','22:30','22:45',...
    '23:00','23:15','23:30','23:45'};
NOMBRESCOLUMNASFACTORES = {'F1','F2','F3','F4','F5','F6','F7','F8',...
    'F9','F10','F11','F12','F13','F14','F15',...
    'F16','F17','F18','F19','F20','F21',...
    'F22','F23'};
NOMBRESCOLUMNASDIAS = {'LUNES','MARTES','MIERCOLES','JUEVES',...
    'VIERNES','SABADOS','DOMINGOS'};
NOMBRESCOLUMNASTOTAL = [NOMBRESCOLUMNASFECHA...
    NUMERODATOPOTENCIA...
    NOMBRESCOLUMNASFACTORES...
    NOMBRESCOLUMNASDIAS];
patrones_consumo_cluster = [];
patrones_consumo_cluster = [PERFILES_CON_OUTLIERS_CON_FECHA
    FACTORFORMADIASCO];
patrones_consumo_cluster = num2cell(patrones_consumo_cluster);
patrones_consumo_cluster = [NOMBRESCOLUMNASTOTAL;...
    patrones_consumo_cluster];
```

La función “xlswrite” permite escribir datos en una hoja de cálculo de un archivo de Microsoft Excel.

```
xlswrite('PERFILES_DE_CARGA_SEGMENTADOS_Y_CLASIFICADOS.xlsx',patrones_consumo_cluster, strcat(['Perfiles con Outliers cluster ', num2str(j)], ''));
```

- **Codificación para almacenar los PCEE diarios con outliers de cada cluster:**

```
NOMBRESCOLUMNASFECHA = {'DÍA','MES','AÑO'};  
NUMERODATOPOTENCIA = {'00:00','00:15','00:30','00:45',...  
    '01:00','01:15','01:30','01:45',...  
    '02:00','02:15','02:30','02:45',...  
    '03:00','03:15','03:30','03:45',...  
    '04:00','04:15','04:30','04:45',...  
    '05:00','05:15','05:30','05:45',...  
    '06:00','06:15','06:30','06:45',...  
    '07:00','07:15','07:30','07:45',...  
    '08:00','08:15','08:30','08:45',...  
    '09:00','09:15','09:30','09:45',...  
    '10:00','10:15','10:30','10:45',...  
    '11:00','11:15','11:30','11:45',...  
    '12:00','12:15','12:30','12:45',...  
    '13:00','13:15','13:30','13:45',...  
    '14:00','14:15','14:30','14:45',...  
    '15:00','15:15','15:30','15:45',...  
    '16:00','16:15','16:30','16:45',...  
    '17:00','17:15','17:30','17:45',...  
    '18:00','18:15','18:30','18:45',...  
    '19:00','19:15','19:30','19:45',...  
    '20:00','20:15','20:30','20:45',...  
    '21:00','21:15','21:30','21:45',...  
    '22:00','22:15','22:30','22:45',...  
    '23:00','23:15','23:30','23:45'};  
NOMBRESCOLUMNASFACTORES = {'F1','F2','F3','F4','F5','F6','F7','F8',...  
    'F9','F10','F11','F12','F13','F14','F15',...  
    'F16','F17','F18','F19','F20','F21',...  
    'F22','F23'};  
NOMBRESCOLUMNASDIAS = {'LUNES','MARTES','MIERCOLES','JUEVES',...  
    'VIERNES','SABADOS','DOMINGOS'};  
NOMBRESCOLUMNASTOTAL = [NOMBRESCOLUMNASFECHA...  
    NUMERODATOPOTENCIA...  
    NOMBRESCOLUMNASFACTORES...  
    NOMBRESCOLUMNASDIAS];  
  
patrones_consumo_cluster = []; %#ok<NASGU>  
patrones_consumo_cluster = [PERFILES_SIN_OUTLIERS_CON_FECHA  
    FACTORFORMADIASSO];  
patrones_consumo_cluster = num2cell(patrones_consumo_cluster);
```

```
patrones_consumo_cluster = [NOMBRESCOLUMNASTOTAL;...
    patrones_consumo_cluster];
```

La función “xlswrite” permite escribir datos en una hoja de cálculo de un archivo de Microsoft Excel.

```
xlswrite('PERFILES_DE_CARGA_SEGMENTADOS_Y_CLASIFICADOS.xlsx',patrones
_consumo_cluster, strcat(['Perfiles sin Outliers cluster ', num2str(j)], ''));
display(strcat(['...Datos de los perfiles de carga de energía eléctrica diarios del cluster ',
num2str(j), ' sin Outliers guardados...']));
disp(' ') % Se muestra un espacio en blanco
pause(1);
```

3.3.11. Almacenamiento del patrón diario de consumo de energía eléctrica

El programa almacena automáticamente el PCPEE diario de cada uno de los clusters en un archivo de Microsoft Excel denominado “PERFILES_DE_CARGA_SEGMENTADOS_Y_CLASIFICADOS”. La codificación para dicha visualización es la siguiente:

- **Codificación para almacenar el patrón de consumo de energía eléctrica de cada cluster:**

```
NUMERODATOPOTENCIA = {'00:00','00:15','00:30','00:45',...
    '01:00','01:15','01:30','01:45',...
    '02:00','02:15','02:30','02:45',...
    '03:00','03:15','03:30','03:45',...
    '04:00','04:15','04:30','04:45',...
    '05:00','05:15','05:30','05:45',...
    '06:00','06:15','06:30','06:45',...
    '07:00','07:15','07:30','07:45',...
    '08:00','08:15','08:30','08:45',...
    '09:00','09:15','09:30','09:45',...
    '10:00','10:15','10:30','10:45',...
    '11:00','11:15','11:30','11:45',...
    '12:00','12:15','12:30','12:45',...
    '13:00','13:15','13:30','13:45',...
    '14:00','14:15','14:30','14:45',...
    '15:00','15:15','15:30','15:45',...
    '16:00','16:15','16:30','16:45',...
    '17:00','17:15','17:30','17:45',...
    '18:00','18:15','18:30','18:45',...
    '19:00','19:15','19:30','19:45',...
    '20:00','20:15','20:30','20:45',...
    '21:00','21:15','21:30','21:45',...
    '22:00','22:15','22:30','22:45',...
    '23:00','23:15','23:30','23:45'};
```

```
patrones_consumo_cluster = [];
```

```
patrones_consumo_cluster = MEDIA_PERFILES_SIN_OUTLIERS;  
patrones_consumo_cluster = num2cell(patrones_consumo_cluster);  
patrones_consumo_cluster = [NUMERODATOPOTENCIA;...  
    patrones_consumo_cluster];%#ok<AGROW>
```

La función “xlswrite” permite escribir datos en una hoja de cálculo de un archivo de Microsoft Excel.

```
xlswrite('PERFILES_DE_CARGA_SEGMENTADOS_Y_CLASIFICADOS.xlsx',patrones  
_consumo_cluster, strcat(['Patron Consumo cluster ', num2str(j)]),");
```

CAPÍTULO IV

4.1. Resultados obtenidos mediante el programa desarrollado en MATLAB

Los resultados obtenidos en base al algoritmo desarrollado en MATLAB pueden ser considerados como una forma de verificación a lo expuesto en [22], que refiere un trabajo desarrollado por el docente tutor del presente proyecto, en donde la base del proyecto es centrarse en una metodología estadística para evaluar los cambios en el perfil de consumo de los casos en estudio. En base a esta metodología estadística se obtiene un patrón de consumo a partir de una base de datos histórica de un intervalo de tiempo predefinido, de acuerdo con el tipo de día (día de la semana, día laborable o no laborable). Además, se propone un índice que evalúa el cambio en el perfil de consumo de electricidad. La evaluación de estos cambios permite asociar estos valores con posibles eventos en una instalación, lo que puede servir para generar alarmas en un sistema de administración de energía [22].

En el presente capítulo se presentan los resultados obtenidos durante la ejecución del programa realizado en MATLAB, obteniendo las diferentes gráficas en el análisis clustering a los datos de los diferentes clientes con los que se trabajó.

4.1.1. Consumo de energía eléctrica de la Universidad Politécnica Salesiana (Acometida Principal)

En primera instancia se presentan los resultados del proceso clustering a los datos de la acometida principal de la UPS. El análisis se realiza en base a las gráficas realizadas en el programa de MATLAB que se explica en el Capítulo III.

Se procede a graficar los perfiles de carga de cada uno de los días de la base de datos de la acometida principal de la UPS.

Perfiles de Carga de Energía Eléctrica de los días Lunes

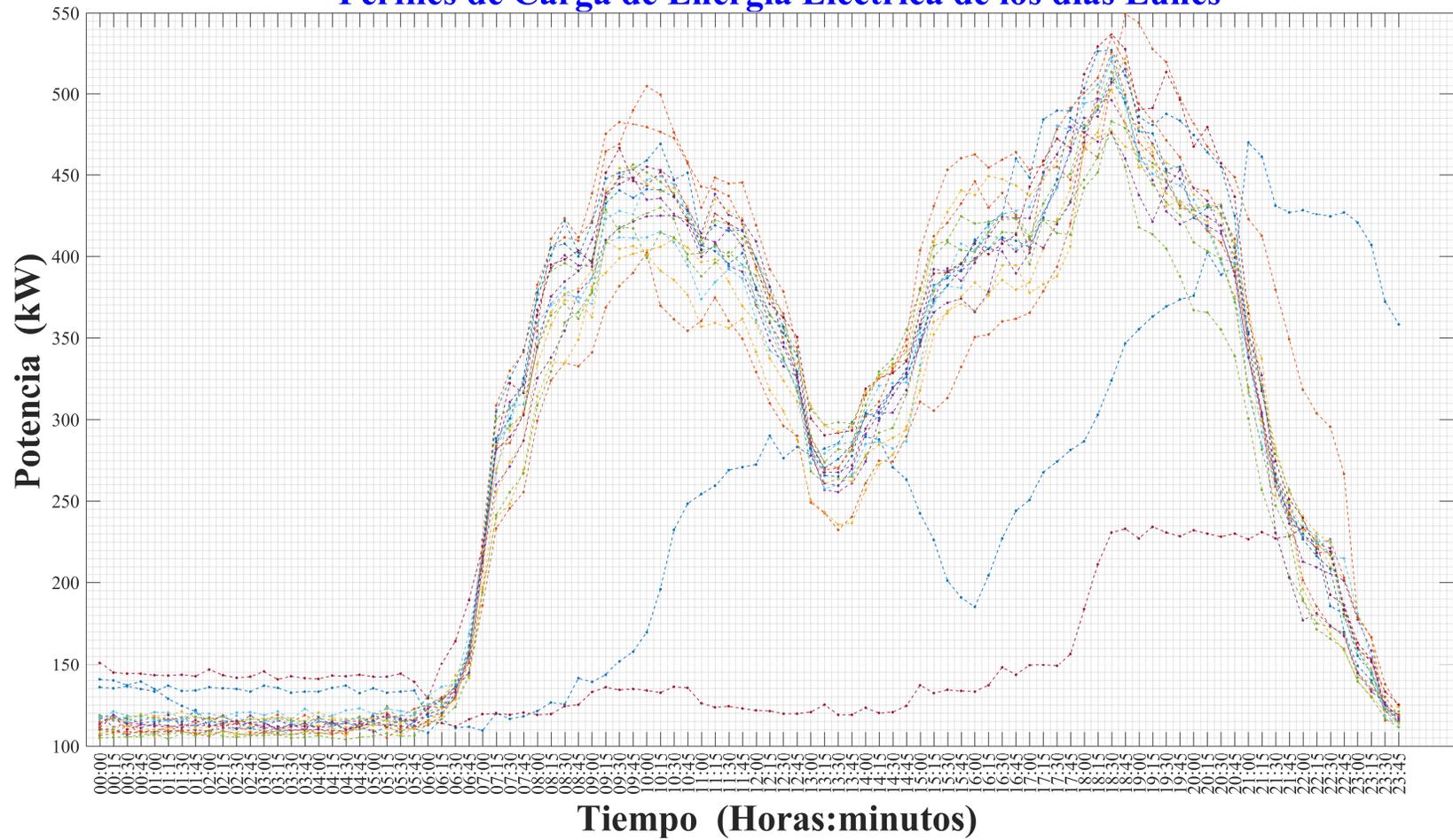


Figura 24. Perfiles de Carga de Energía Eléctrica (Lunes)

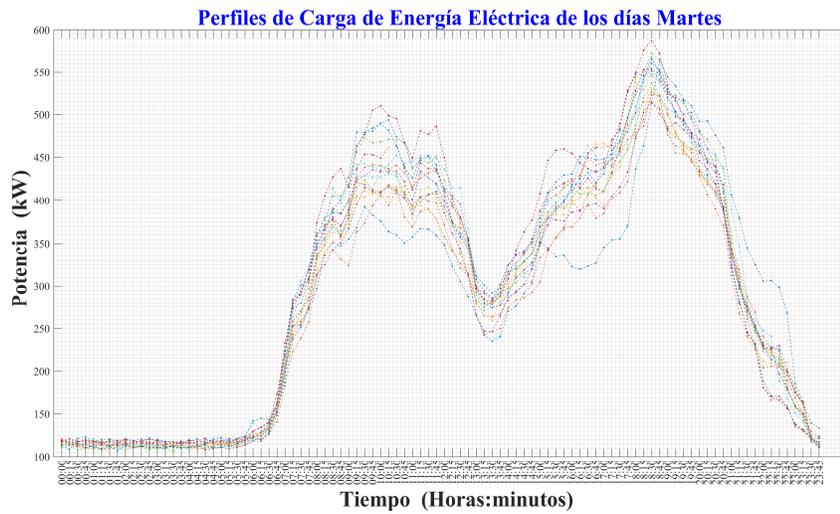


Figura 25. Perfiles de Carga de Energía Eléctrica (Martes)

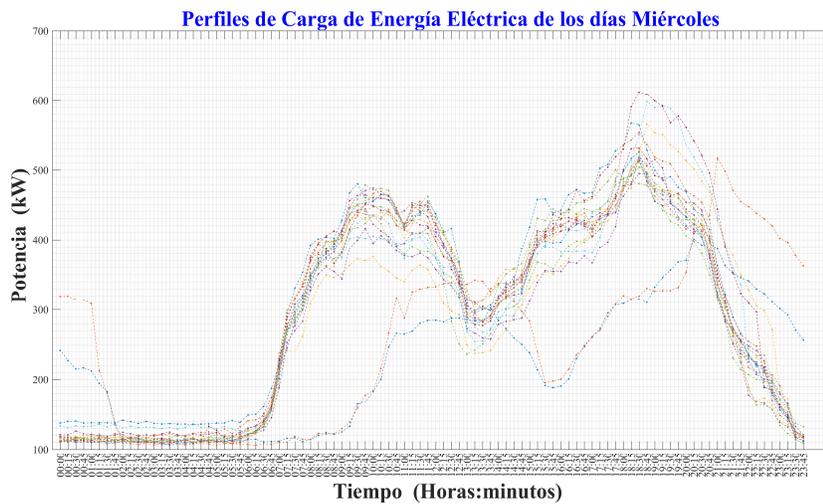


Figura 26. Perfiles de Carga de Energía Eléctrica (Miércoles)

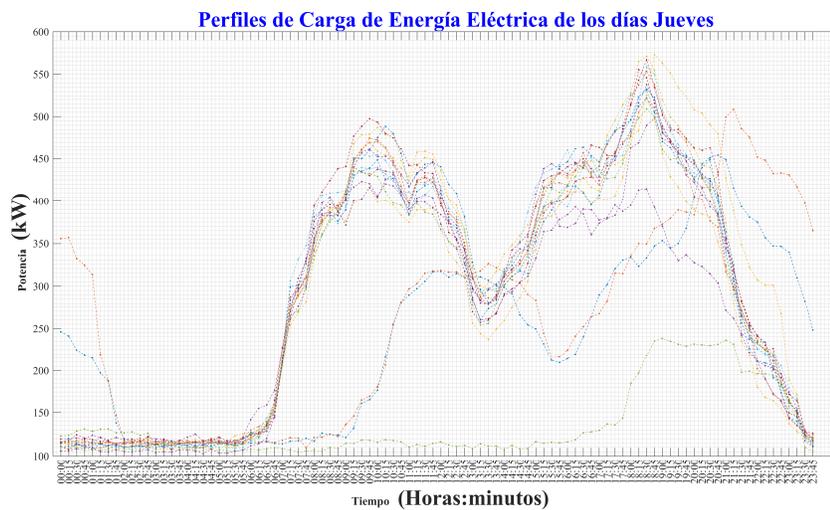


Figura 27. Perfiles de Carga de Energía Eléctrica (Jueves)

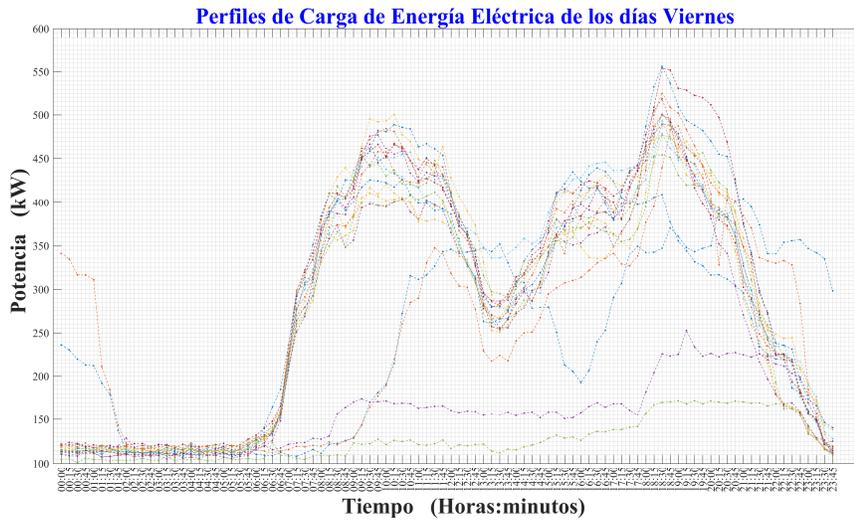


Figura 28. Perfiles de Carga de Energía Eléctrica (Viernes)

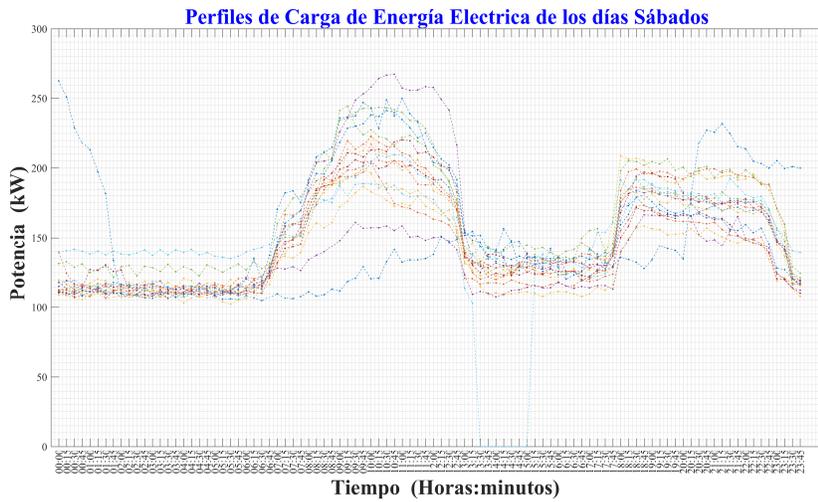


Figura 29. Perfiles de Carga de Energía Eléctrica (Sábados)

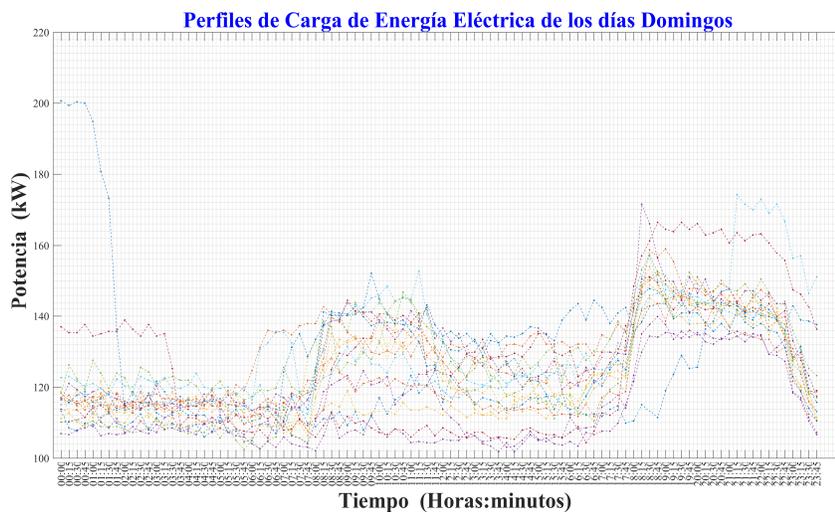


Figura 30. Perfiles de Carga de Energía Eléctrica (Domingos)

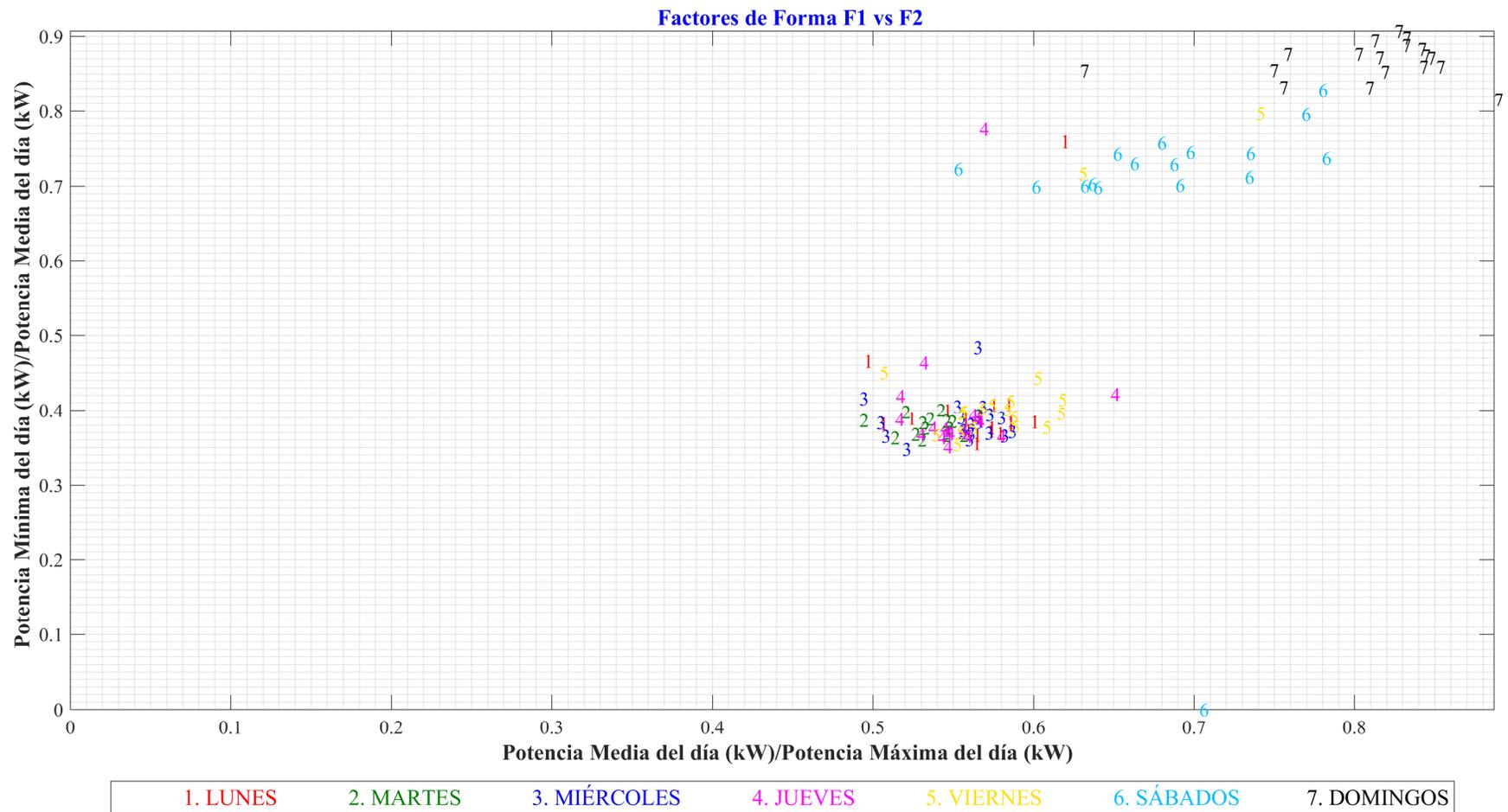


Figura 31. Factores de Forma F1 vs F2.

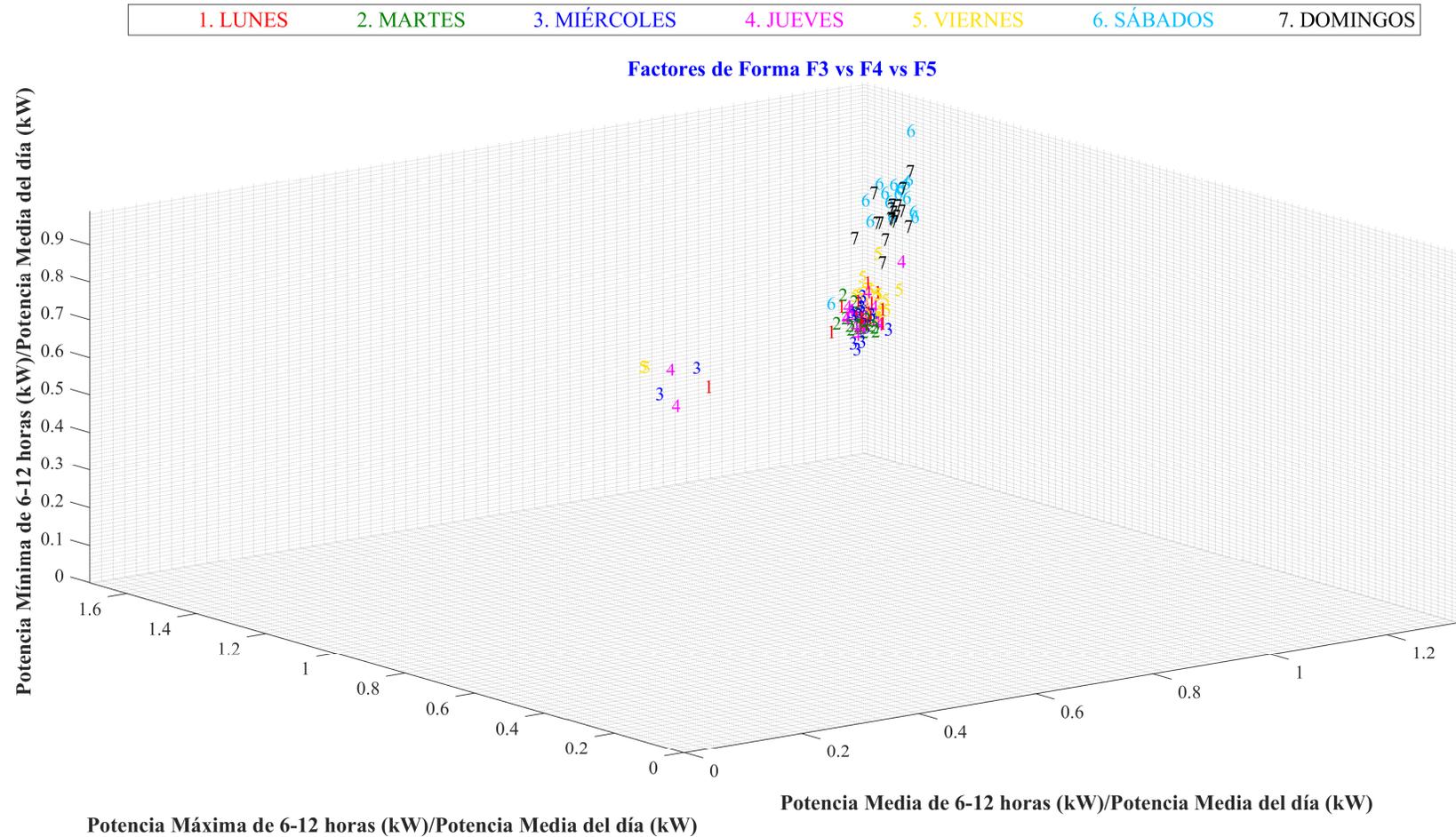


Figura 32. Factores de Forma F3 vs F4 vs F5.

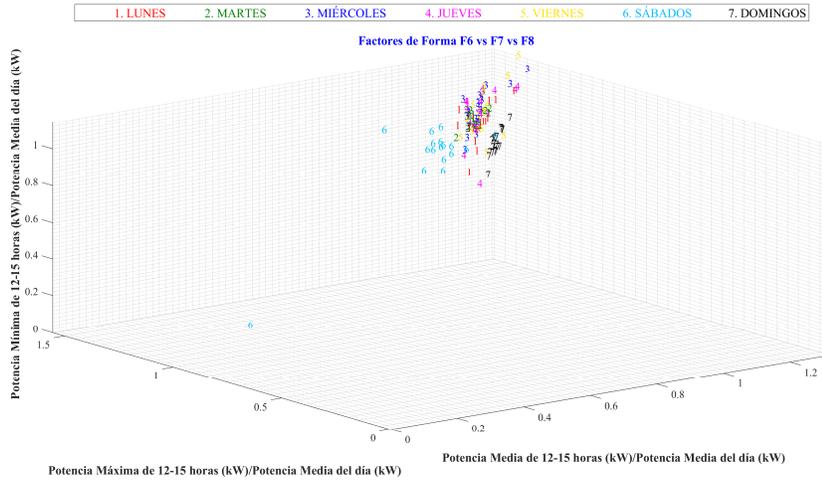


Figura 33. Factores de Forma F6 vs F7 vs F8.

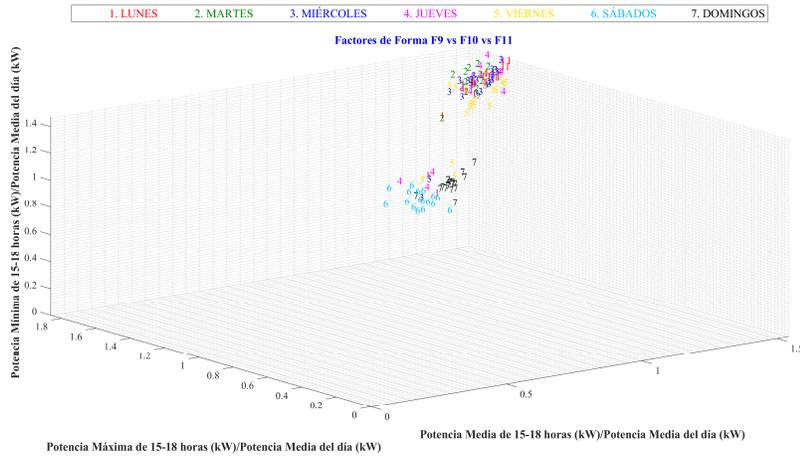


Figura 34. Factores de Forma F9 vs F10 vs F11.

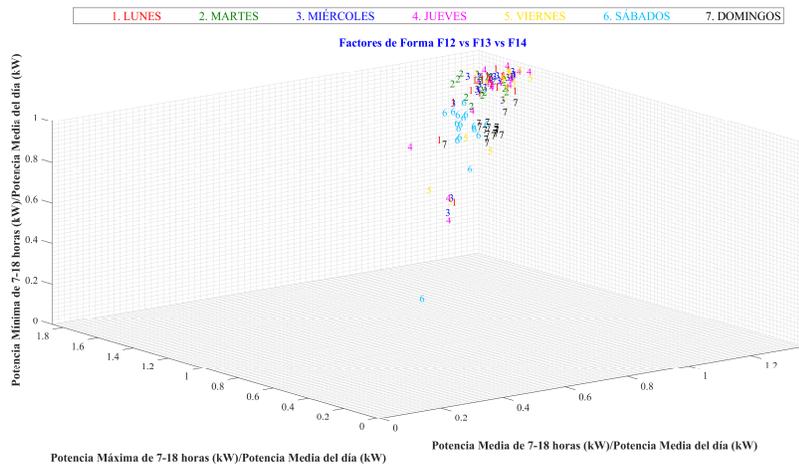


Figura 35. Factores de Forma F12 vs F13 vs F14.

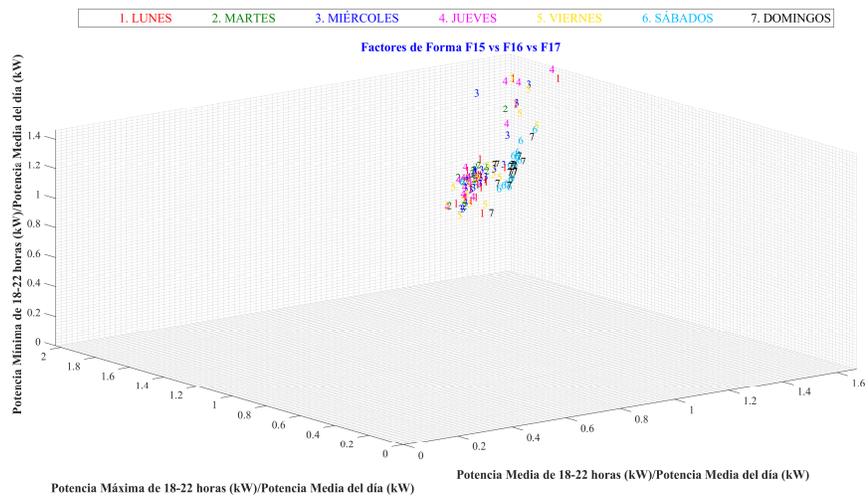


Figura 36. Factores de Forma F15 vs F16 vs F17.

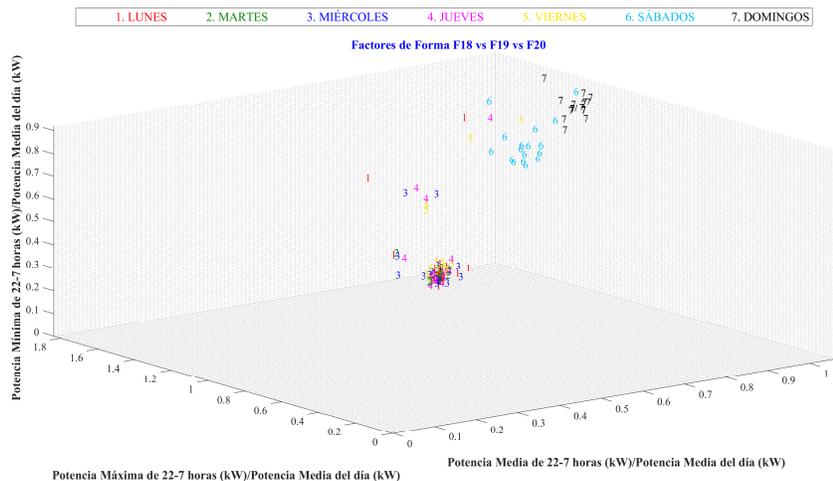


Figura 37. Factores de Forma F18 vs F19 vs F20.

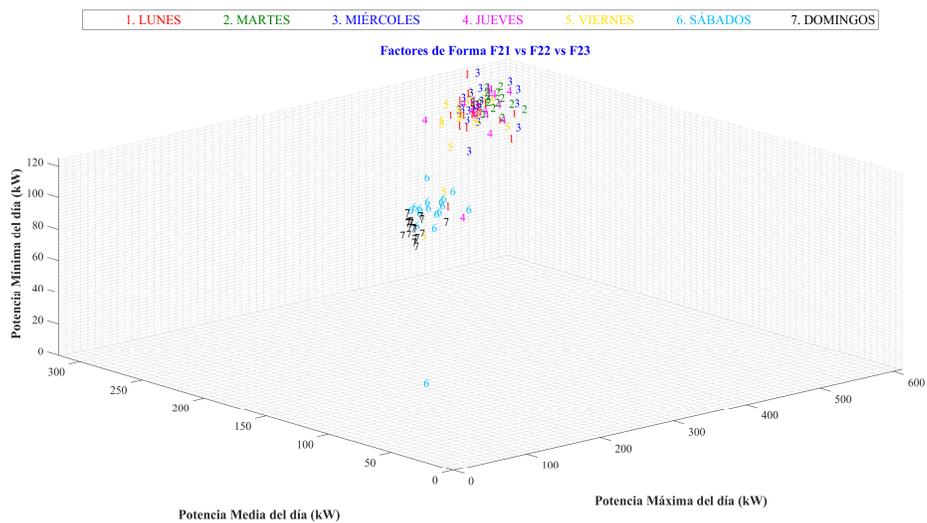


Figura 38. Factores de Forma F21 vs F22 vs F23.

4.1.1.1. Valores promedios de las siluetas según el número de clusters con los que se segmentaron los factores de forma

Tabla 6. Valores promedios de las Siluetas para el número de clusters analizados

Número de Clusters	Valor Promedio de Silueta
1.0000	-
2.0000	0.9663
3.0000	0.7435
4.0000	0.6768
5.0000	0.6789
6.0000	0.6618
7.0000	0.6847

4.1.1.2. Evaluaciones del número de clusters más óptimo para los datos en análisis

1. El número de Clusters más apropiados según el Valor de la Silueta:

Ekmeans3 =

SilhouetteEvaluation with properties:

NumObservations: 126

InspectedK: [1 2 3 4 5 6 7]

CriterionValues: [NaN 0.9663 0.7435 0.6768 0.7019 0.6618 0.6847]

OptimalK: 2

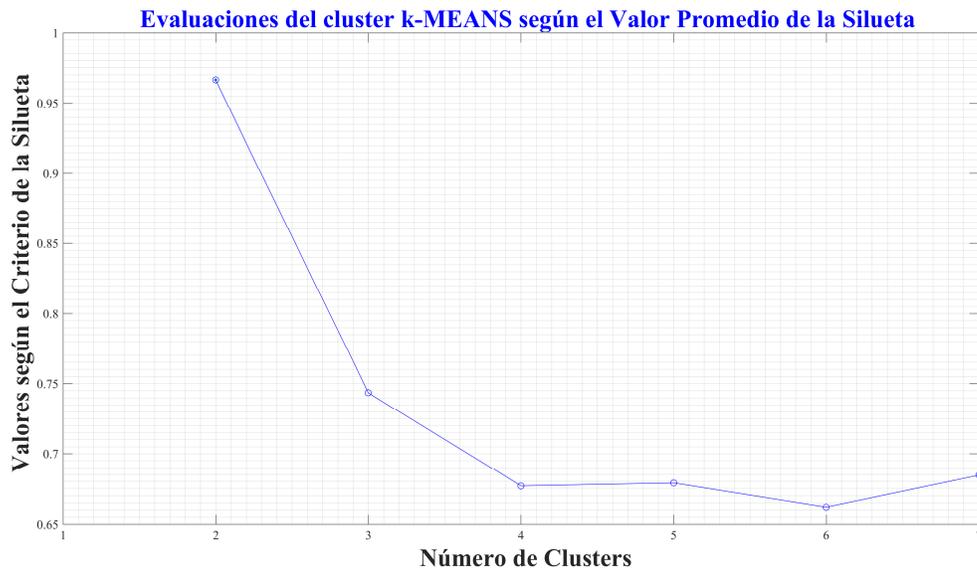


Figura 39. Evaluación según el valor promedio de Silueta del cluster k-Means.

4.1.1.3. Valores de los tiempos de ejecución para cada número de cluster con los que se segmentaron los factores de forma

Tabla 7. Tiempos de Ejecución para el número de clusters analizados

Número de Clusters	Tiempo de Ejecución (s)
[1]	[4.4199]
[2]	[9.1238]
[3]	[8.2149]
[4]	[8.1940]
[5]	[8.2074]
[6]	[8.2086]
[7]	[8.2275]

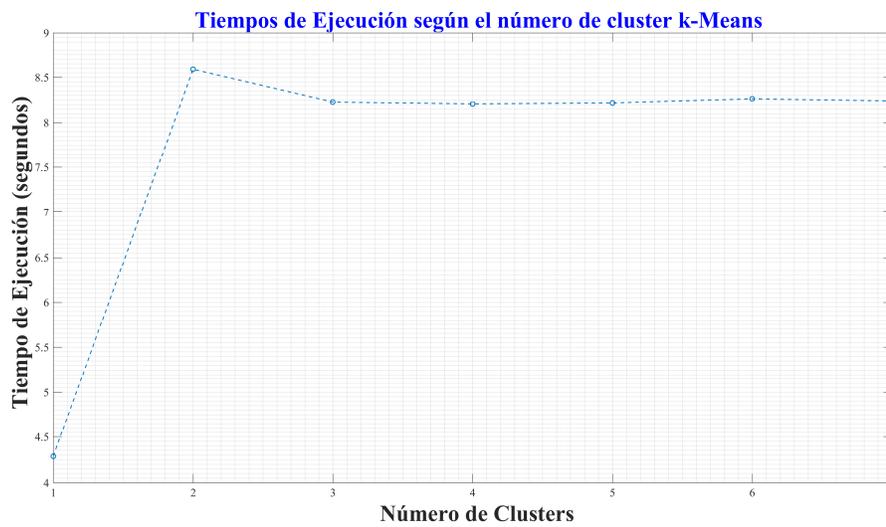


Figura 40. Tiempos de ejecución según el número de cluster k-Means.

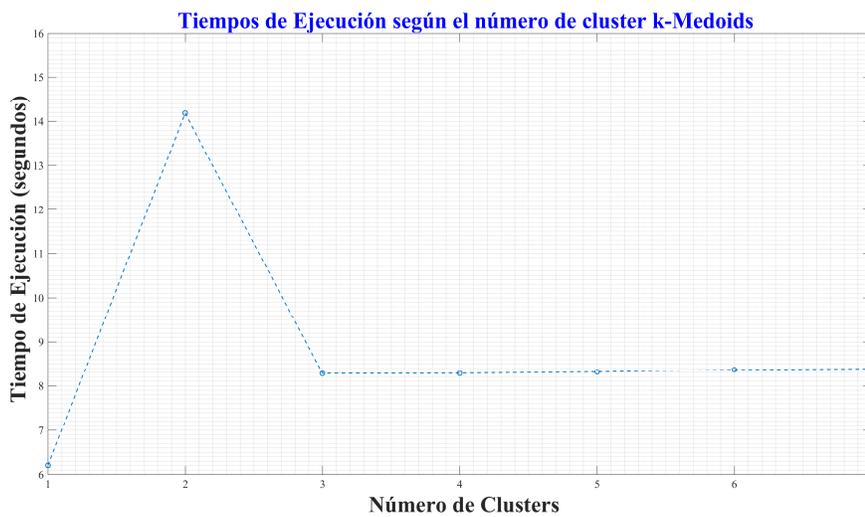


Figura 41. Tiempos de ejecución según el número de cluster k-Medoids.

Gráfica de Silueta con el número de clusters óptimo

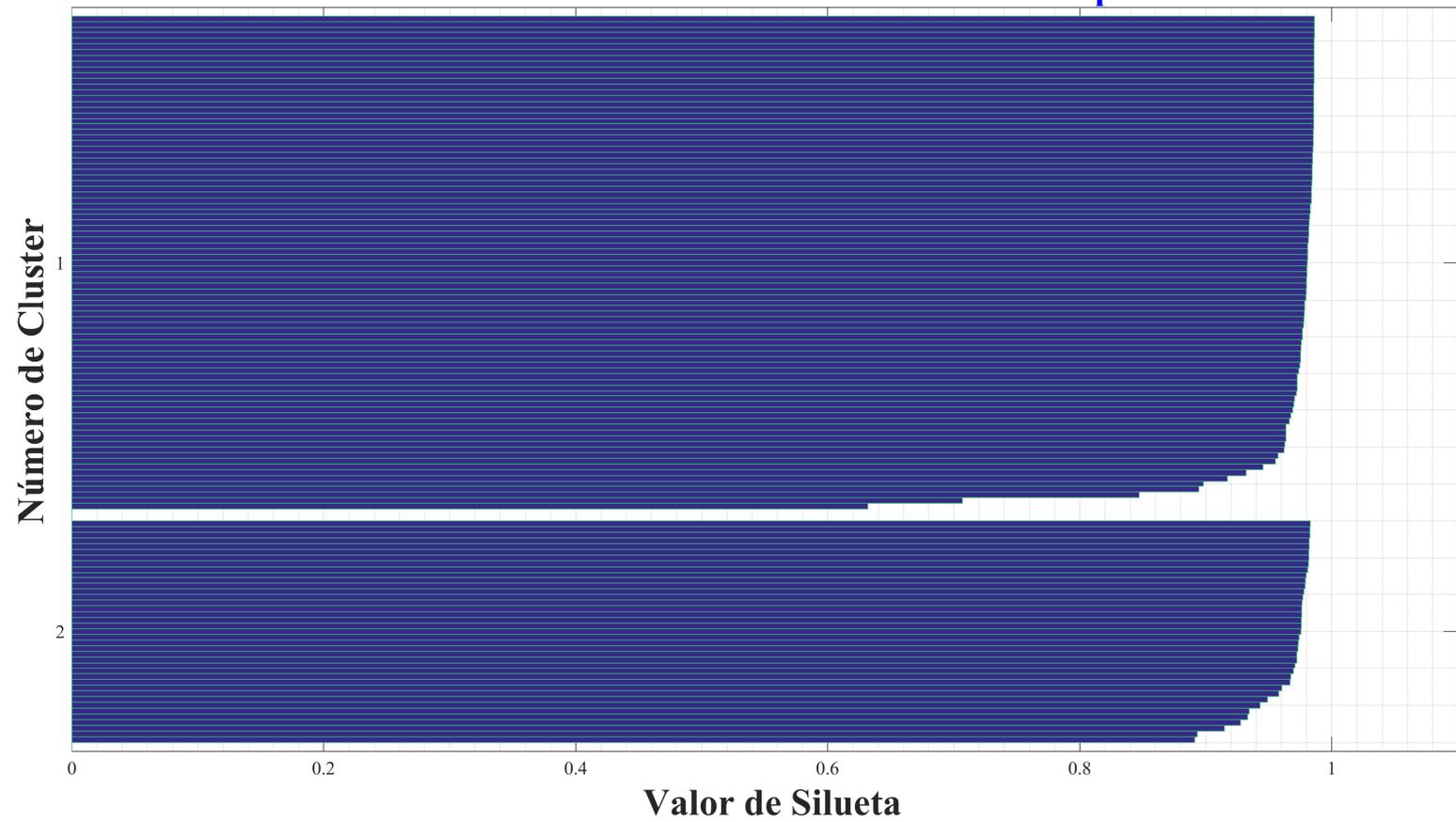


Figura 42. Gráfica de Silueta con el número de clusters óptimo.

4.1.1.4. Indicador del número de cluster a cuál pertenece cada PCEE diario

Tabla 8. Indicador del número de cluster al que pertenece cada PCEE.

Número de PCEE	Cluster	Número de PCEE	Cluster	Número de PCEE	Cluster
1	1	43	1	85	1
2	1	44	1	86	1
3	1	45	1	87	1
4	1	46	1	88	1
5	1	47	1	89	1
6	1	48	1	90	1
7	2	49	1	91	1
8	1	50	1	92	2
9	1	51	1	93	2
10	1	52	1	94	2
11	1	53	1	95	2
12	1	54	1	96	2
13	1	55	1	97	2
14	1	56	1	98	2
15	1	57	1	99	2
16	1	58	1	100	2
17	1	59	2	101	2
18	1	60	1	102	2
19	1	61	1	103	2
20	1	62	1	104	2
21	1	63	1	105	2
22	1	64	1	106	2
23	1	65	1	107	2
24	1	66	1	108	2
25	1	67	1	109	2
26	1	68	1	110	2
27	1	69	1	111	2
28	1	70	1	112	2
29	1	71	1	113	2
30	1	72	1	114	2
31	1	73	1	115	2
32	1	74	1	116	2
33	1	75	1	117	2
34	1	76	1	118	2
35	1	77	2	119	2
36	1	78	1	120	2
37	1	79	1	121	2
38	1	80	1	122	2
39	1	81	1	123	2
40	1	82	1	124	2
41	1	83	2	125	2
42	1	84	1	126	2

4.1.1.5. Ubicación de los centroides para cada cluster

Tabla 9. Centroides para cada cluster

CLUSTER 1	CLUSTER 2
0,553407162594901	0,734980252891419
0,386739444610642	0,779474650719651
1,17987609591013	1,06898102474674
1,52414260790732	1,22329504933557
0,440981145155340	0,828457151915435
1,10250397411706	0,937657406736642
1,29290803367973	1,10181121632593
0,947553351189216	0,856797694178242
1,41963568877389	0,942716125806155
1,63674466862711	1,12643435600511
1,28624640074748	0,883466699491562
1,28232755717841	1,01648869096183
1,65111372979213	1,27159104680930
0,873196398213098	0,826339796861185
1,40984063742933	1,20272440037017
1,80957783231333	1,27490805585043
0,822778224007343	1,13694779820888
0,472781591257794	0,889747421993246
0,850353176438735	1,20059810322050
0,386908324676329	0,804979306784955
519,411494252874	191,360000000000
286,760727969349	137,304957264957
110,547126436782	105,952820512821

4.1.1.6. Distancia de los factores de forma que caracterizan cada PCEE diario hasta el centroide de cada cluster

Tabla 10. Cuadro de distancias de cada Factor de Forma al centroide de cada cluster

CLUSTER 1	CLUSTER 2
5455,99826557454	86509,0380902585
1021,86081557131	147694,432472505
246,213434184465	127184,892223958
31,2236500208784	127213,634855305
2041,63274970487	100332,442590283
25,2102774604397	131050,225782276
101567,247549699	1905,52772077599
144,886443801324	137690,119594210
583,625123183739	146144,890856968
1910,79595128080	101159,380216627
550,416150145519	113764,068502315
55,9525712001631	126935,996282456
6,09726774883437	130681,021344876
927,539667120131	148456,738400246
217,920998530997	126472,677081923

140,315923976491	136570,573921366
311,727412057971	119981,918092004
1630,26836486479	105031,683686824
1470,11572747564	104412,612165816
2127,81230439092	158887,052341620
272,149946936538	127981,544725261
723,296819520603	148852,884933359
2253,17834105246	166014,712728318
2867,86017782667	169340,199469727
942,962976213287	152832,280376532
1361,22947998956	157918,378977205
2037,14039211829	164446,266840239
75,1904784707709	134382,842385997
159,491121786027	137015,779973988
68,4085361001832	124716,827546720
455,493957178137	145664,120691780
2442,41772527605	167233,306222705
5134,31314595233	186530,611379631
1181,93086813160	155902,046288132
41,8028010304344	128013,782935791
17780,5033793498	51731,8990526587
1114,42823161836	119302,716627370
2215,62925802275	162321,925488452
784,902660081630	110757,649313427
1388,51262113868	104736,993845811
6426,23180678617	192623,098052926
9597,78623653021	209781,750150547
97,7928423129507	135802,487393412
1663,75259146566	160808,828465934
155,628748531045	128492,536196721
112,893252346690	131730,879391572
288,451637660378	122659,486778228
28,6270983213439	128884,461526607
225,551354241164	140581,841193621
3341,35855449470	174096,729107027
70,2417044636955	128469,886331167
146,091929870250	138631,437650776
23,5768059930097	129212,980872543
154,285056050861	128163,647545233
6439,07038899831	79638,6144236856
775,688091870393	115765,095686199
2846,49976167170	169629,711076264
526,540904810654	114005,440203080
102263,136155776	2191,51008712731
1660,77635156212	160740,487552951
1323,97747376367	156497,205329291
204,986563425674	140075,611496391
1324,95105121390	157480,622981569
3043,05247409848	172776,396412126

9470,82626591100	72502,8552626086
112,678565416175	123175,087783165
223,831796795144	140543,159215816
2670,66239616087	169871,269944346
211,395597513521	135914,832352133
50,1137618289395	132687,831141210
14,8553101709733	131193,059765626
437,169708362939	144211,131430285
15168,2988900232	56516,1645167913
2683,38509152403	99607,9709688583
1908,45338752128	101350,164710494
1162,86021164849	106704,451100636
146447,375368465	512,495713710050
835,072069404260	117110,393448174
94,9375087774440	131181,758029128
6682,71237888647	79220,6779081050
403,055738495428	117521,501191273
1911,42001273601	101017,322587708
87707,9087656683	4244,83947192832
432,553770905025	115424,175579619
1701,98255169602	102969,107505615
1326,55643255305	157298,017516713
1712,99939831963	161426,324314793
53,1444014080931	134654,726167333
399,450779870088	119868,256591954
445,157431194020	115702,656334187
4328,12202150361	89130,6992081827
85924,3167782242	5196,43607025050
119012,285262572	296,275842404520
105338,492644546	1299,32865931056
148488,448269477	631,305671345033
90636,4619488590	3834,27645786008
125983,130790356	174,122901144457
116944,643413756	384,375136332660
89403,5398169672	3844,18744598121
109473,143250234	926,456260649243
130884,794958316	125,225151678132
79470,2574586769	6362,01601016622
92741,8785486104	3172,36603834185
127890,524516383	11647,0499504961
109329,461611915	915,935665634977
95276,9670948835	2739,30195975821
108220,071032364	1040,83441282911
125812,963184192	80,6433077749023
127404,064889345	213,328420219896
154499,669030731	1123,47821688989
166618,327941281	2311,47965379087
173301,226672775	3117,84680419646
155280,229964386	1312,72070435072

159587,492367271	1618,18504901063
161254,581269103	1743,91321012807
160810,182316257	1728,31930141926
160379,444714862	1690,73285998439
159715,551281067	1560,28080128586
145737,852114799	511,452366738968
161365,246882372	1712,83089578429
143521,746701939	366,348143219977
150720,174238532	769,447013385185
165610,362928373	2266,35190889616
162732,740834404	1861,66278535935
165530,416990484	2148,38535976351
177557,478049690	3712,92801247847

Asignación de 2 Clusters y 2 Centroides

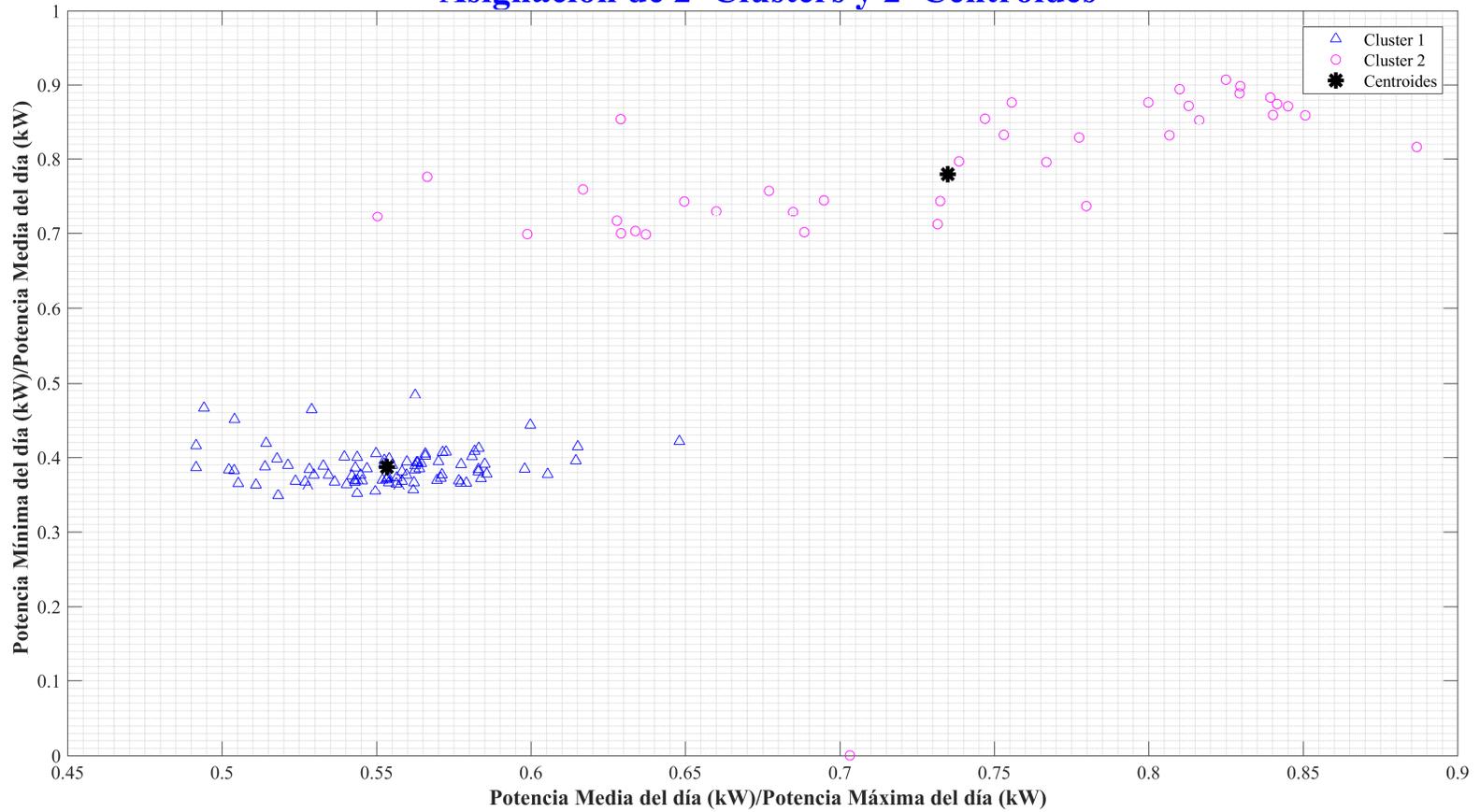


Figura 43. Segmentación de los datos en dos clusters.

Asignación de 2 Clusters y 2 Centroides

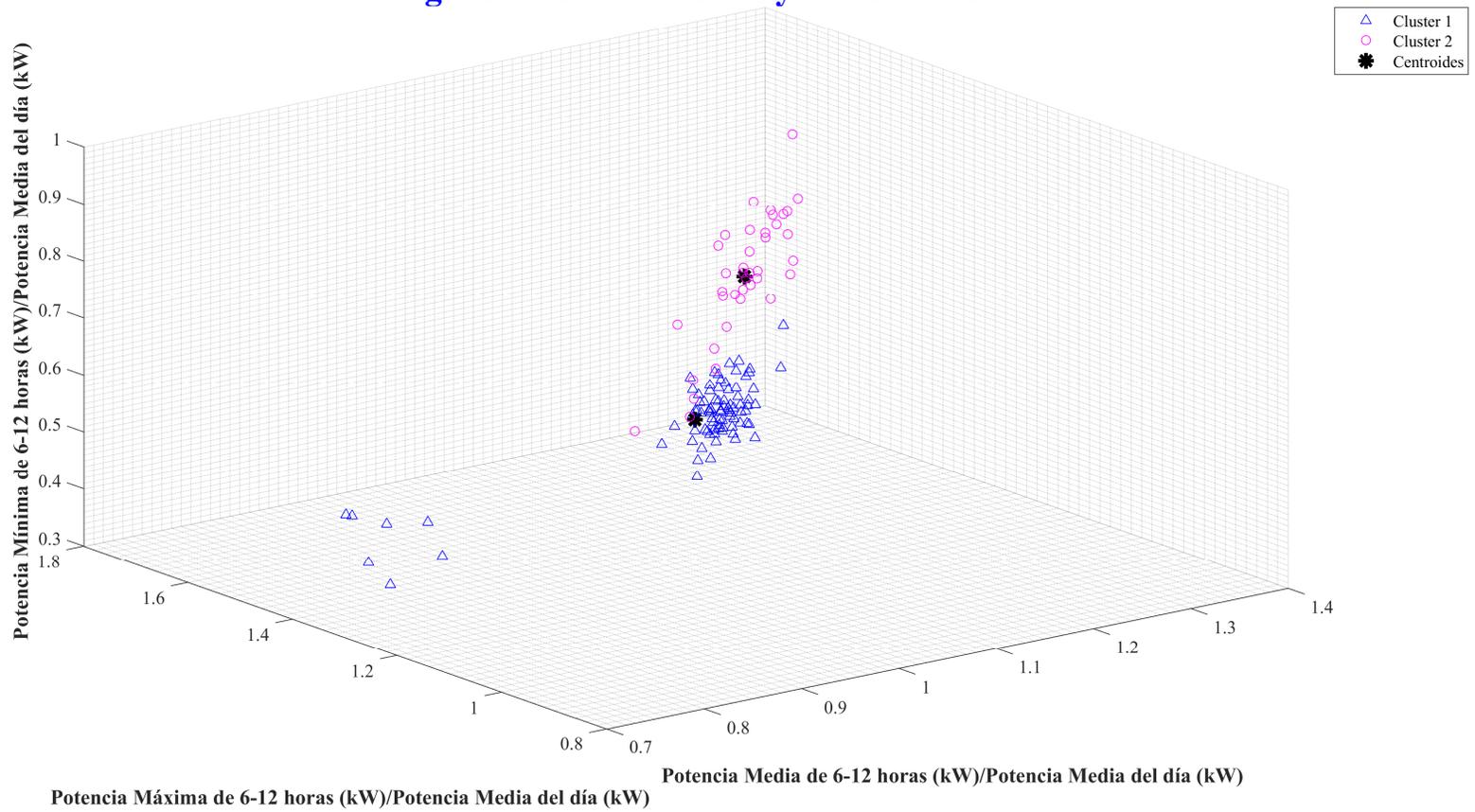


Figura 44. Segmentación de los datos en dos clusters.

Asignación de 2 Clusters y 2 Centroides

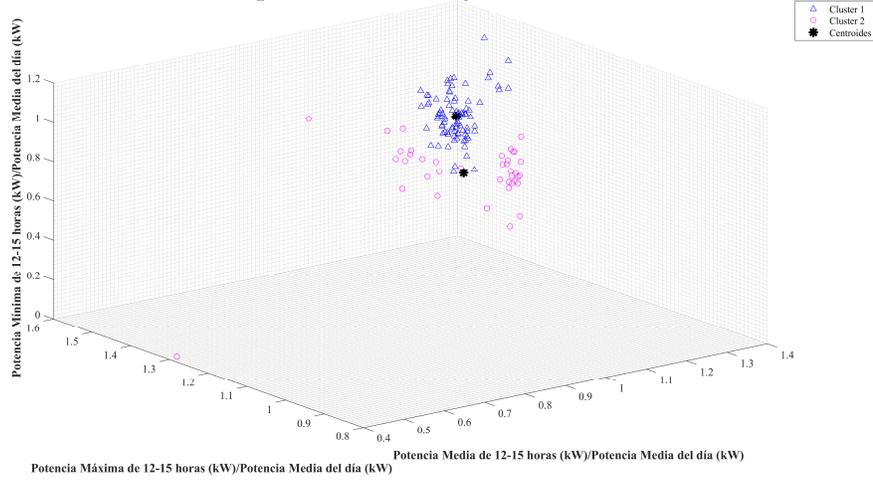


Figura 45. Segmentación de los datos en dos clusters.

Asignación de 2 Clusters y 2 Centroides

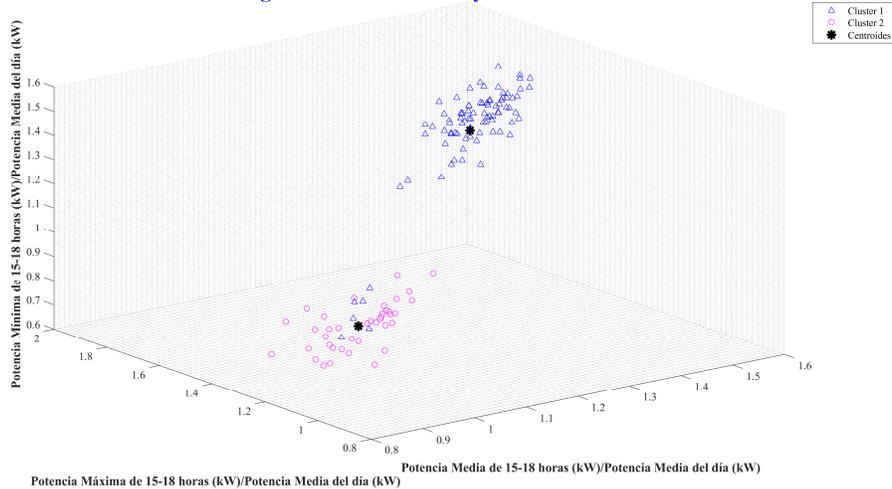


Figura 46. Segmentación de los datos en dos clusters.

Asignación de 2 Clusters y 2 Centroides

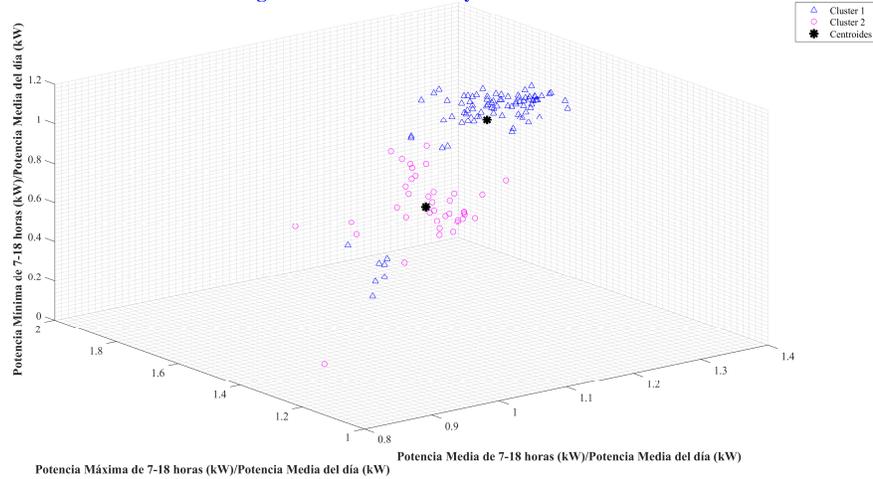


Figura 47. Segmentación de los datos en dos clusters.

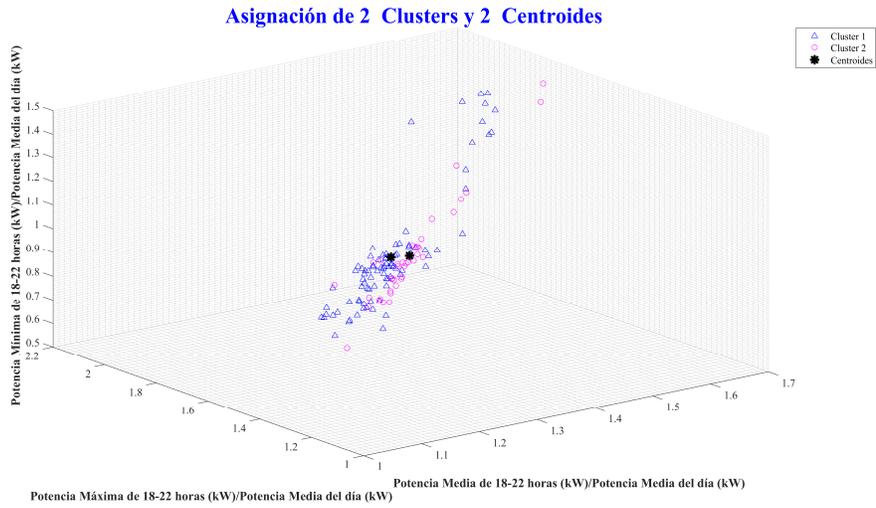


Figura 48. Segmentación de los datos en dos clusters.

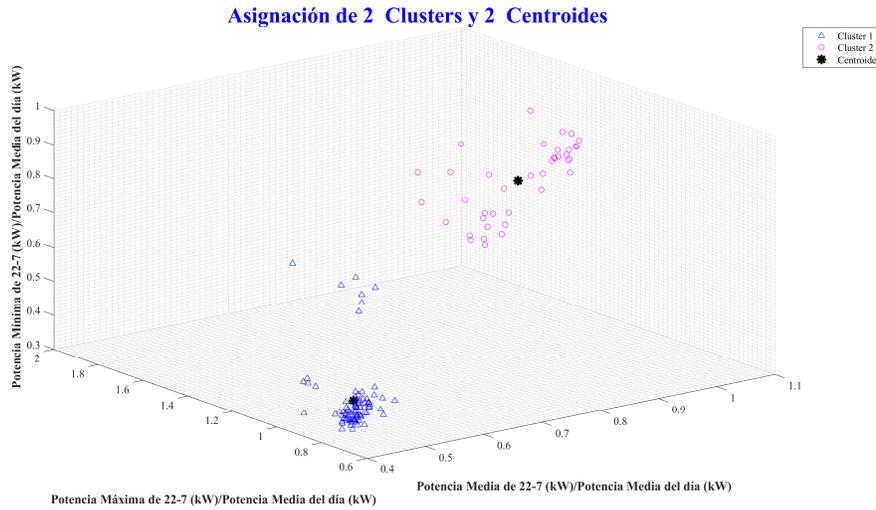


Figura 49. Segmentación de los datos en dos clusters.

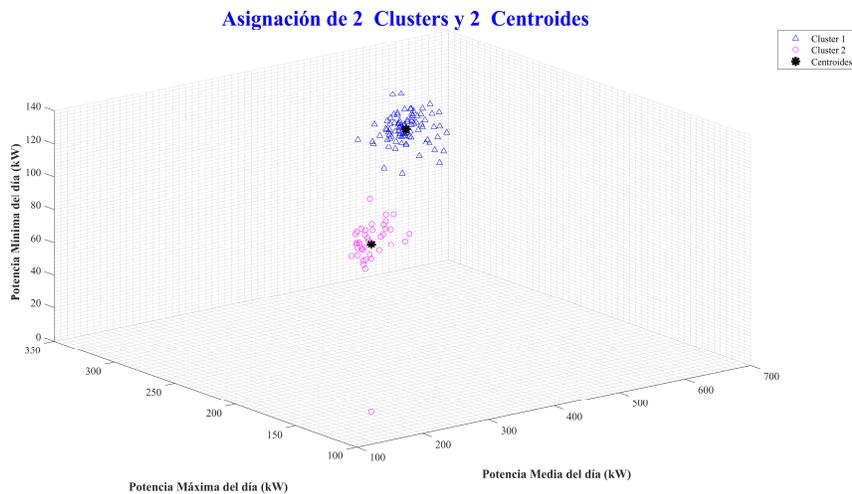


Figura 50. Segmentación de los datos en dos clusters.

4.1.1.7. Análisis estadístico de media aritmética y desviación estándar de las distancias a los centroides y de los valores de silueta de cada punto para la detección de perfiles de carga diarios con outliers y sin outliers de cada cluster

MediaAritméticaDCDelCluster1 =

1.8060e+03

DesviaciónEstándarDCDelCluster1 =

2.9839e+03

RangoIntercuartilDCDelCluster1 =

1.8491e+03

ValorMínimoDCDelCluster1 =

1.3799e+03

AnchoDeBandaDCDelCluster1 =

508.3912

limiteatipicos1 =

7.1791e+03

LimitesDeLosValoresAtipicosParaElCluster1medianteLaDistanciaAIC =

7.1791e+03

MediaAritméticaVSDelCluster1 =

0.9669

DesviaciónEstándarVSDelCluster1 =

0.0508

RangoIntercuartilVSDelCluster1 =

0.0130

ValorMínimoVSDelCluster1 =

0.0097

AnchoDeBandaVSDelCluster1 =

0.0036

limiteatipicos2 =

0.8982

LimitesDeLosValoresAtipicosParaElCluster1medianteLosValoresDeSi =

0.8982

MediaAritméticaDCDelCluster2 =

2.0845e+03

DesviaciónEstándarDCDelCluster2 =

2.1216e+03

RangoIntercuartilDCDelCluster2 =

1.9665e+03

ValorMínimoDCDelCluster2 =

1.4675e+03

AnchoDeBandaDCDelCluster2 =

634.7752

limiteatipicos1 =

5.8362e+03

LimitesDeLosValoresAtipicosParaElCluster2medianteLaDistanciaAIC =

5.8362e+03

MediaAritméticaVSDelCluster2 =

0.9649

DesviaciónEstándarVSDelCluster2 =

0.0237

RangoIntercuartilVSDelCluster2 =

0.0176

ValorMínimoVSDelCluster2 =

0.0131

AnchoDeBandaVSDelCluster2 =

0.0057

limiteatipicos2 =

0.9028

LimitesDeLosValoresAtipicosParaElCluster2medianteLosValoresDeSi =

0.9028

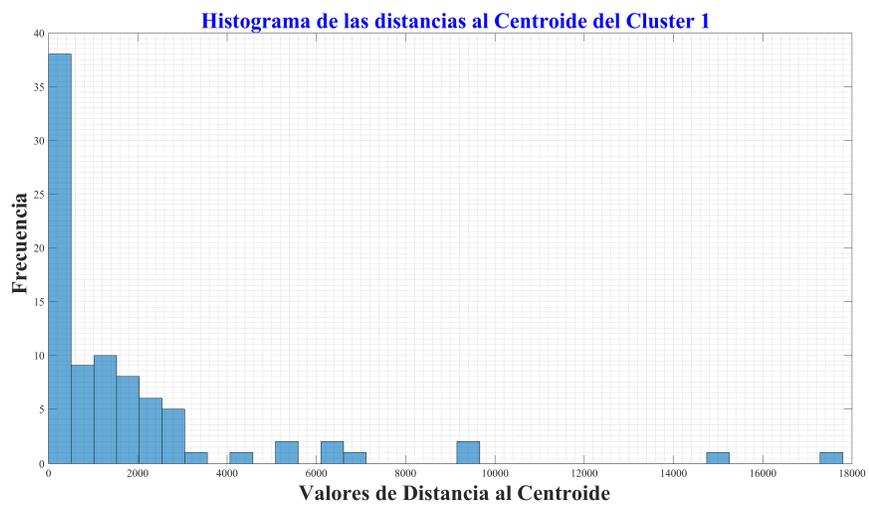


Figura 51. Histograma con las distancias al centroide del cluster 1.

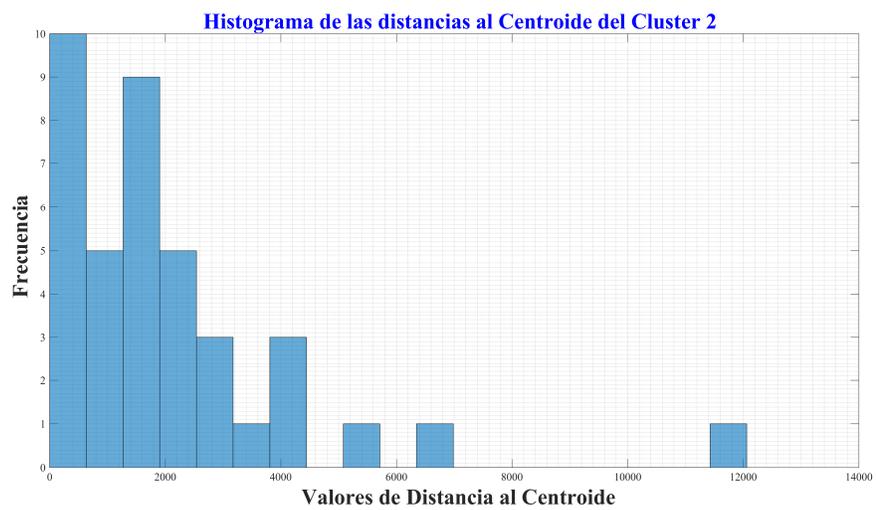


Figura 52. Histograma con las distancias al centroide del cluster 2.

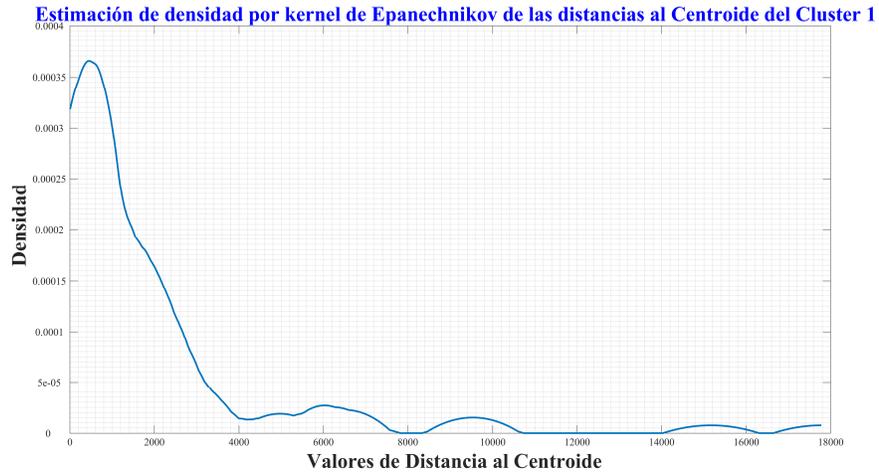


Figura 53. FDP de las distancias al centroide del cluster 1.

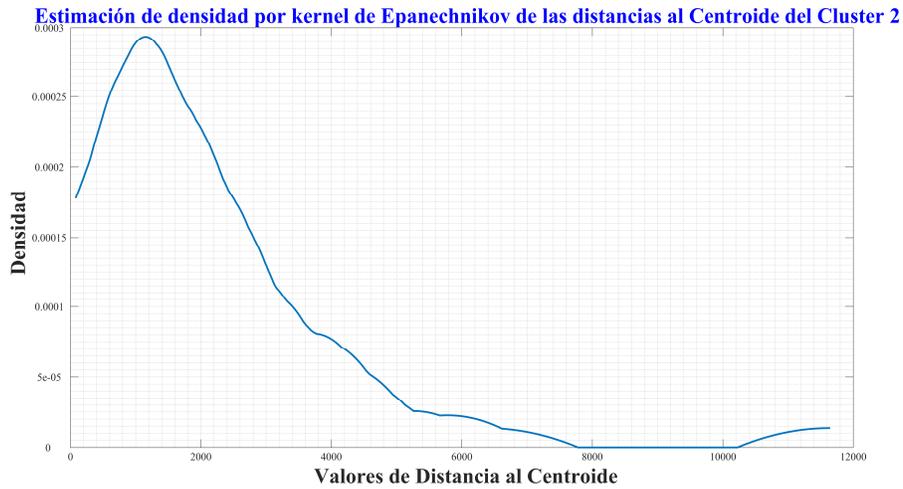


Figura 54. FDP de las distancias al centroide del cluster 2.

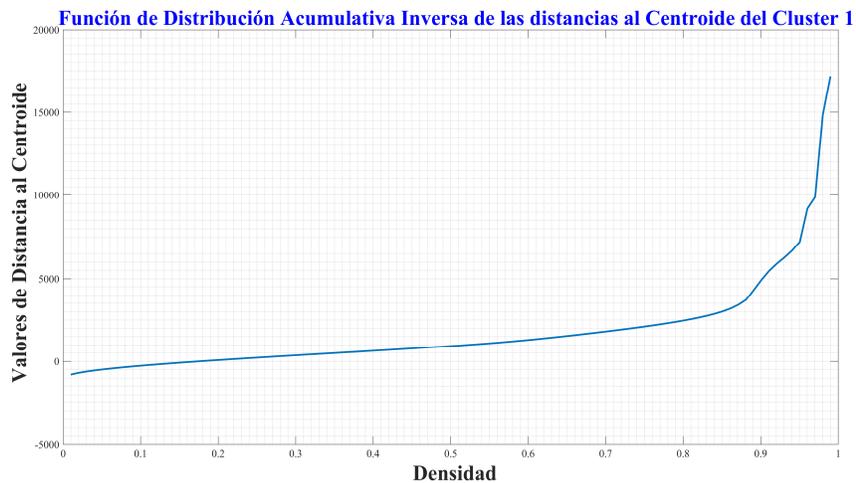


Figura 55. FDAI de las distancias al centroide del cluster 1.

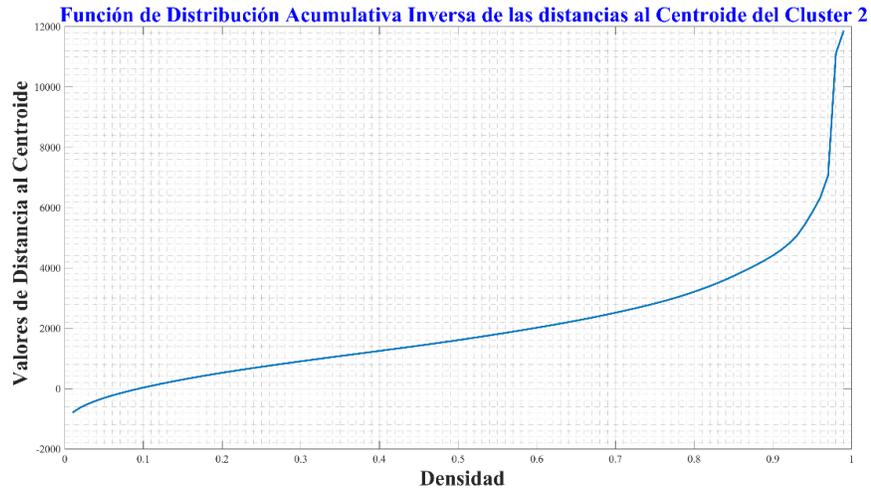


Figura 56. FDAI de las distancias al centroide del cluster 2.

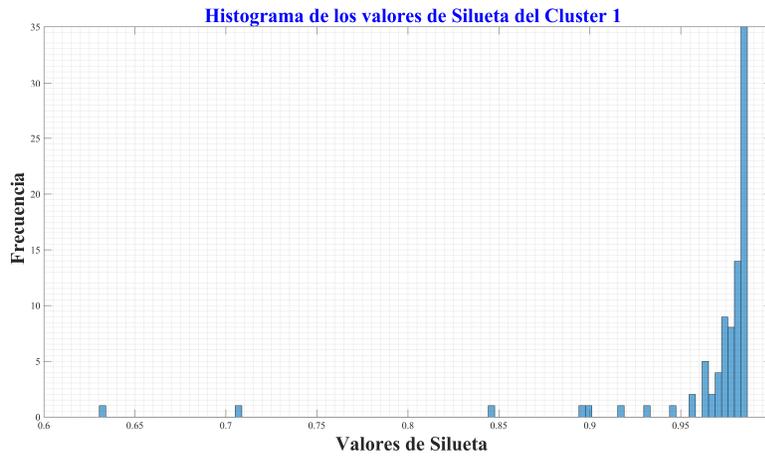


Figura 57. Histograma de los valores de Silueta del cluster 1.

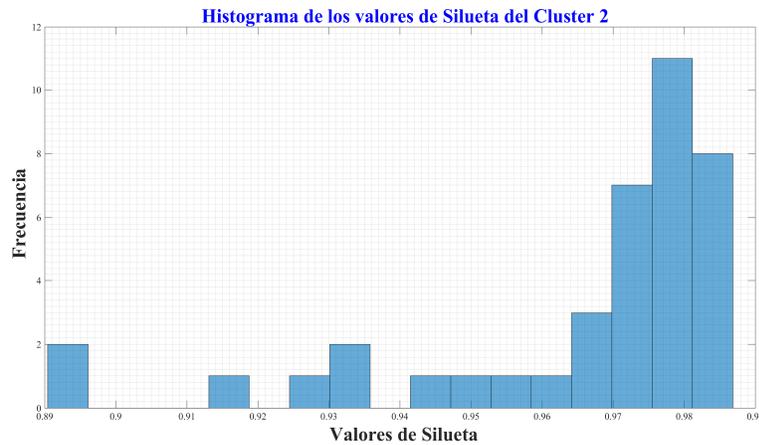


Figura 58. Histograma de los valores de Silueta del cluster 2.

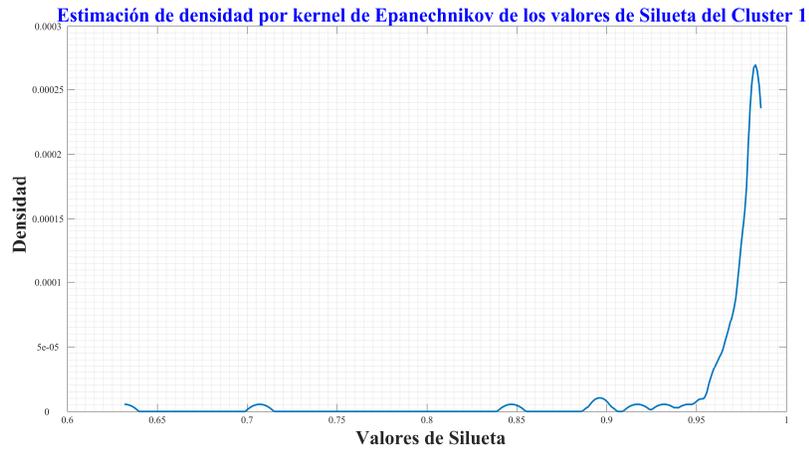


Figura 59. FDP de los valores de silueta del cluster 1.

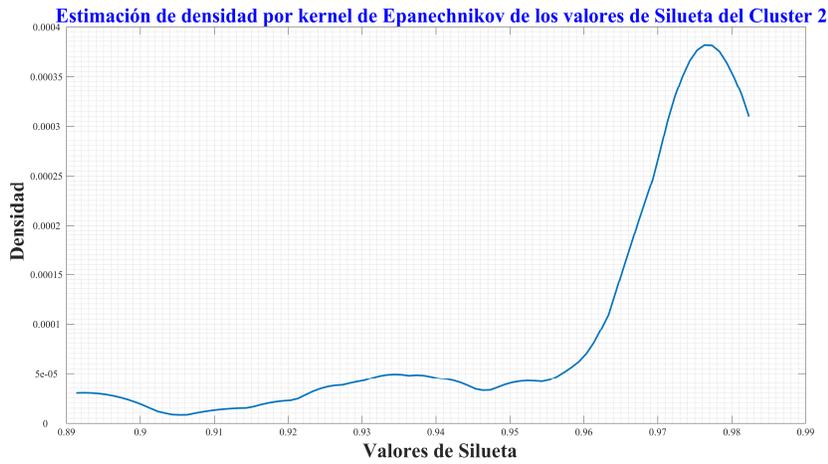


Figura 60. FDP de los valores de silueta del cluster 2.

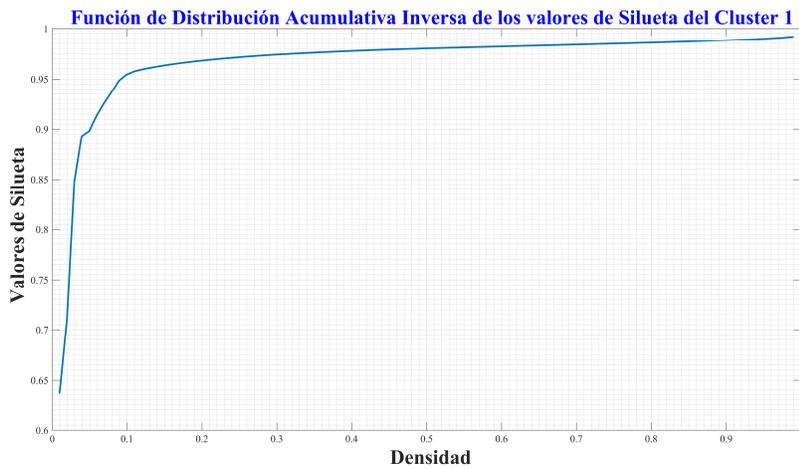


Figura 61. FDAI de los valores de silueta del cluster 1.

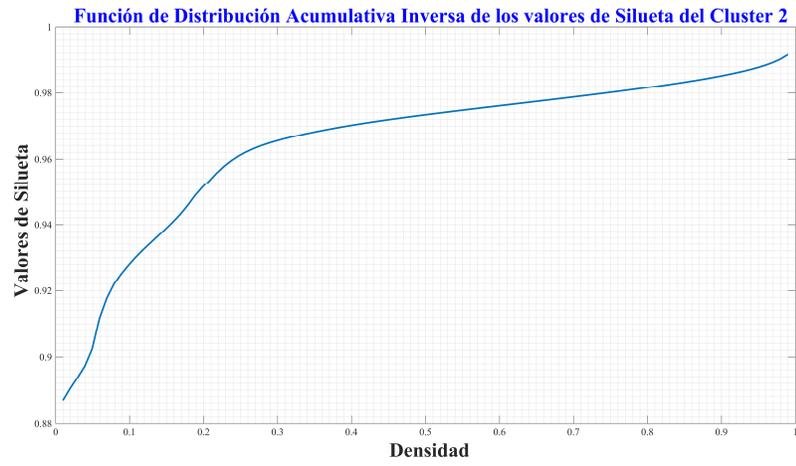


Figura 62. FDAI de los valores de silueta del cluster 2.

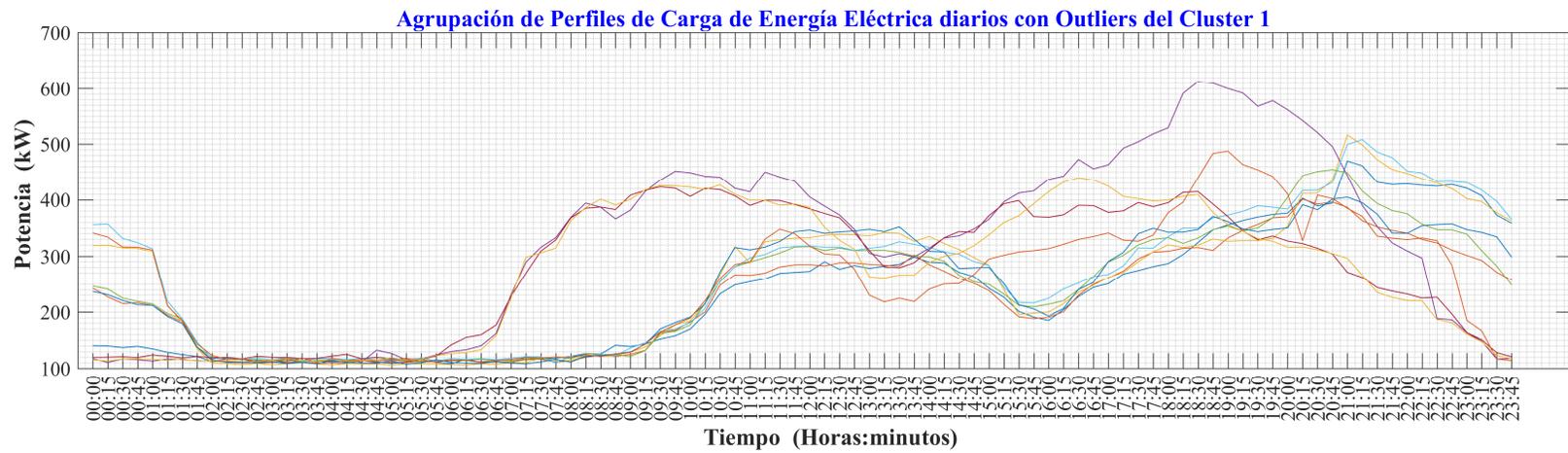
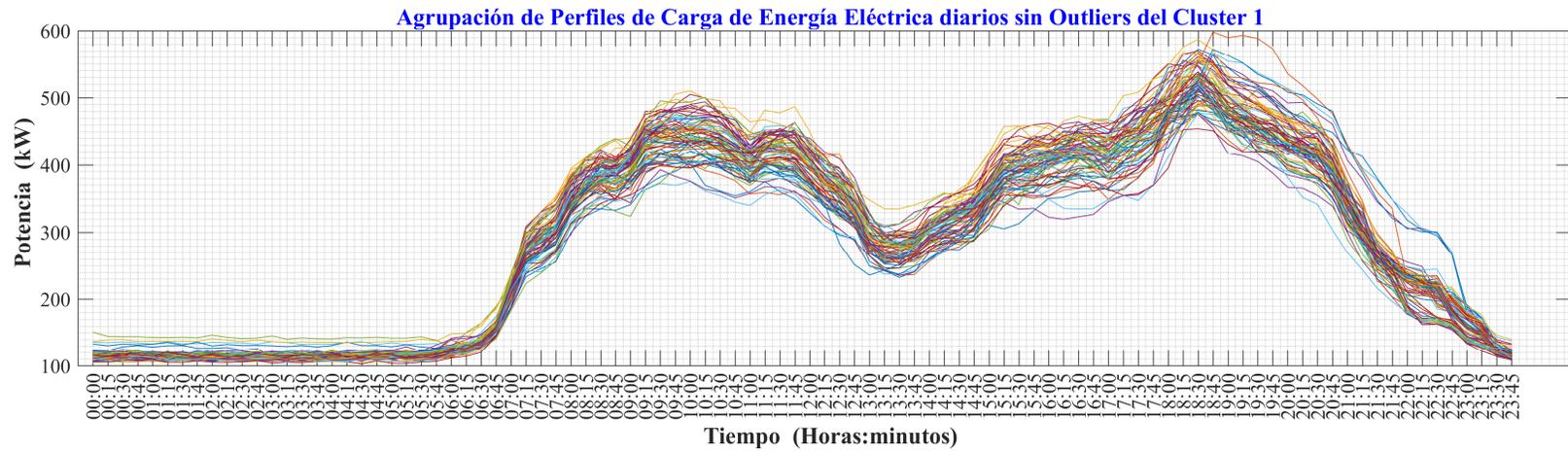


Figura 63. PCEE del cluster 1 (Perfiles sin outliers, 77 - Perfiles con outliers, 10).

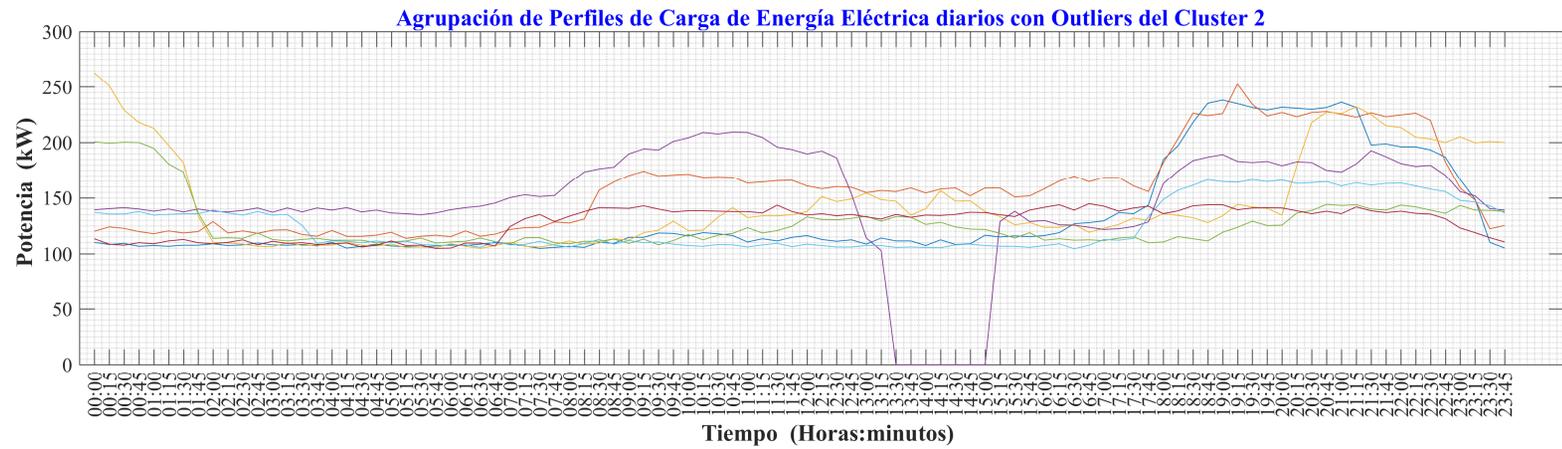
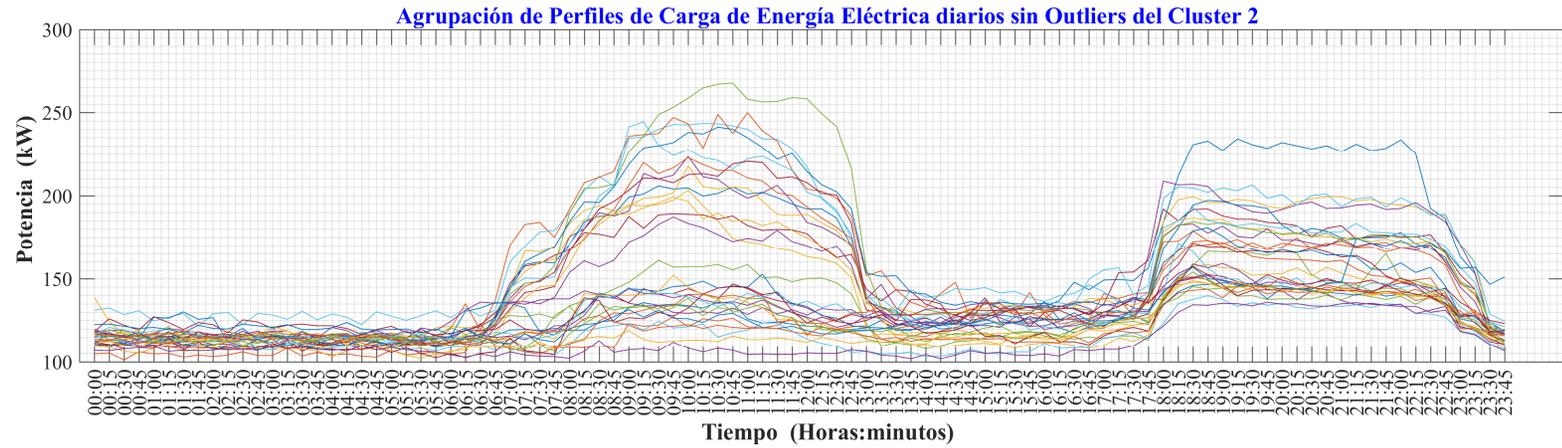


Figura 64. PCEE del cluster 2 (Perfiles sin outliers, 28 - Perfiles con outliers, 11).

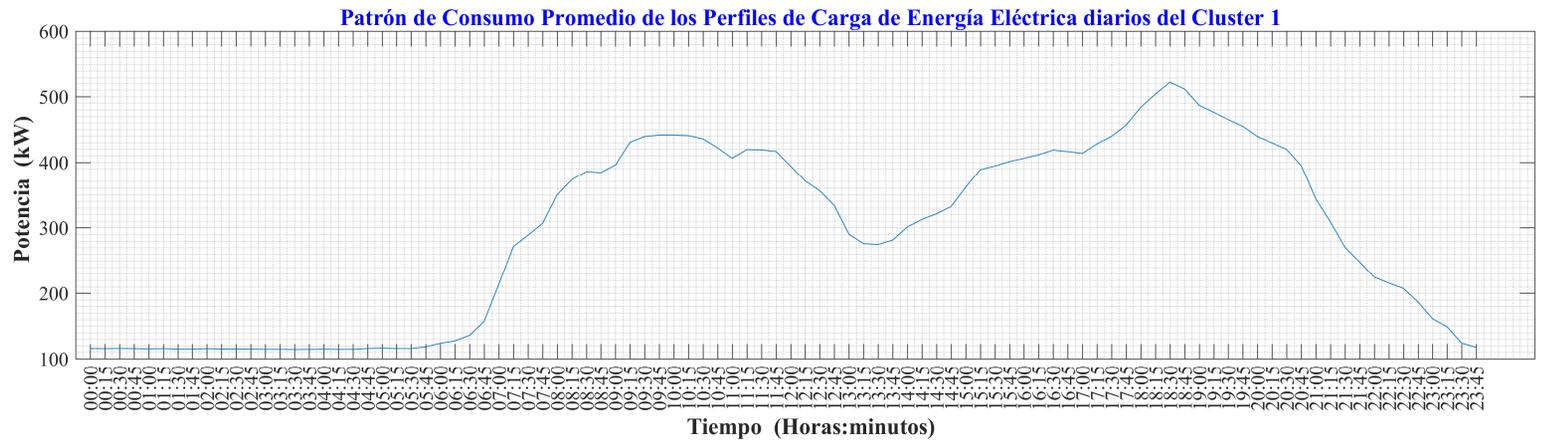
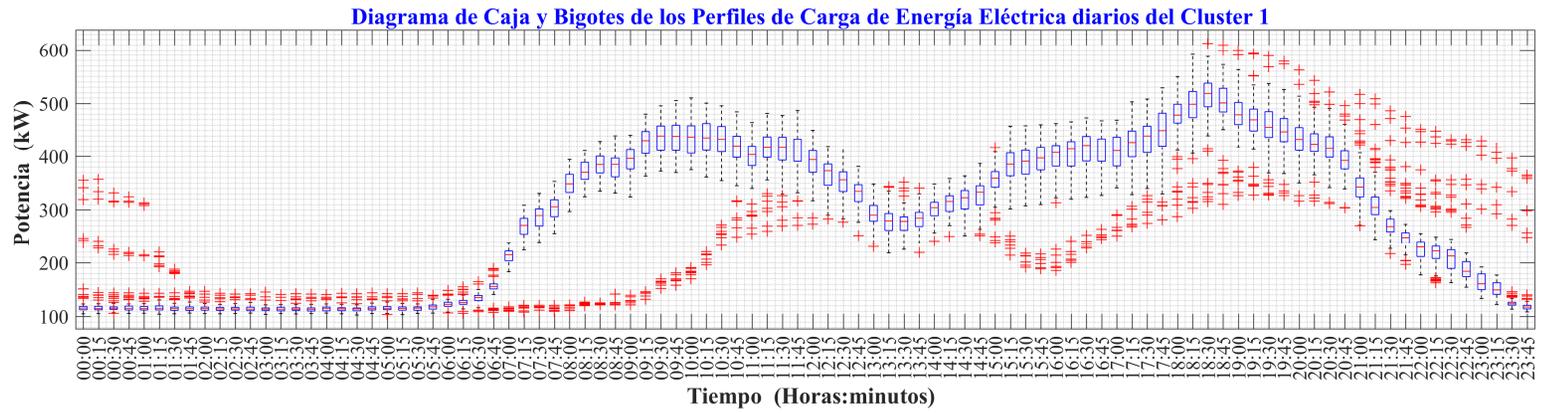


Figura 65. Patrón de consumo de PCEE diarios del cluster 1.

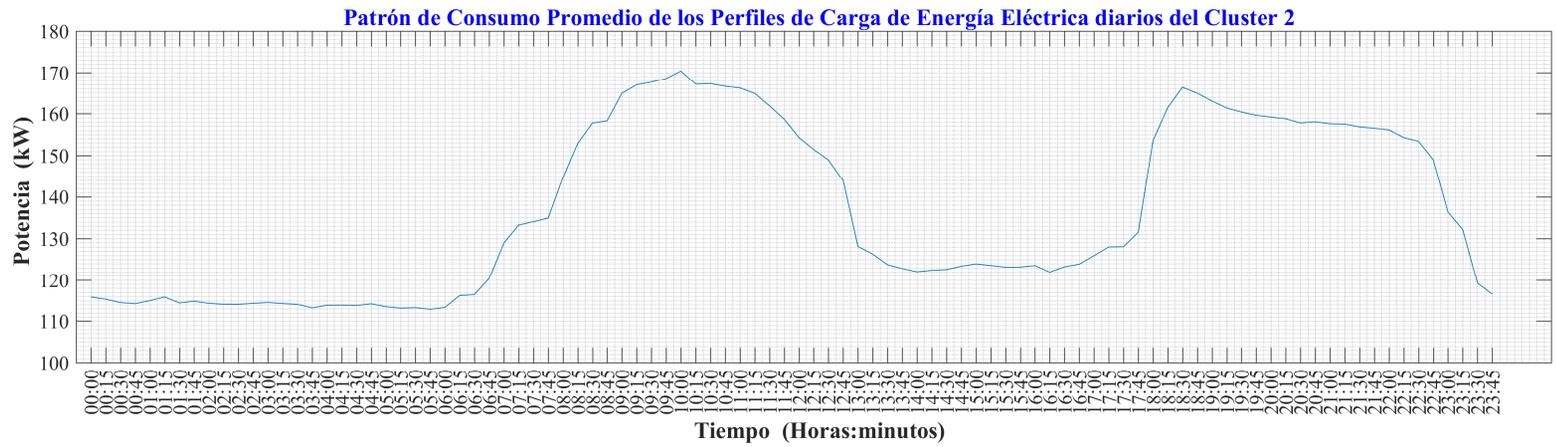
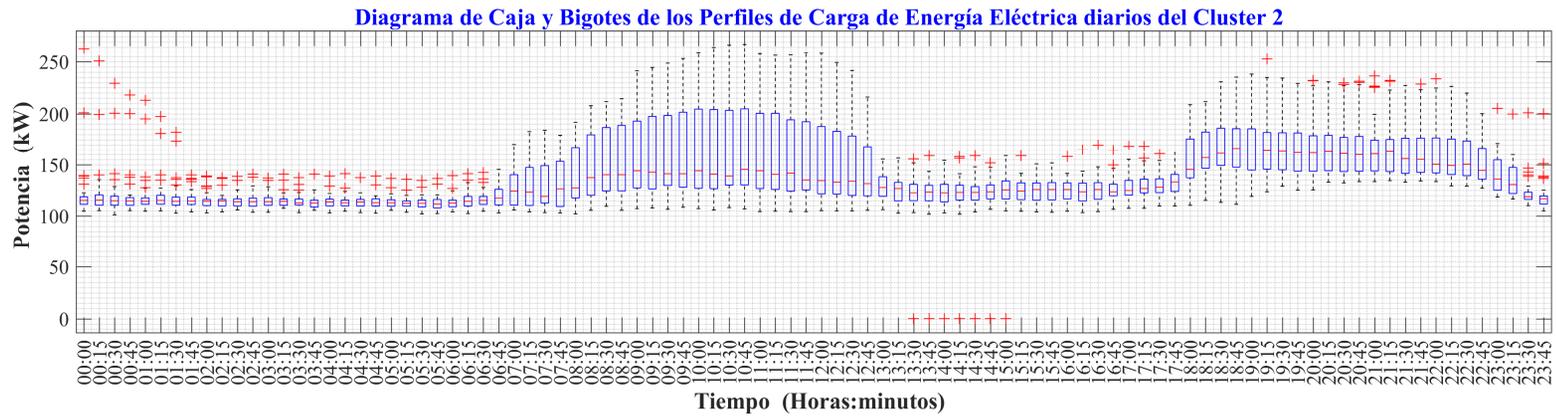


Figura 66. Patrón de consumo de PCEE diarios del cluster 2.

4.1.2. Consumo de energía eléctrica de la Universidad Politécnica Salesiana (Edificio Cornelio Merchán)

Se realiza el análisis de las curvas gráficas por el programa en MATLAB con los datos de CEE del Edificio Cornelio Merchán perteneciente a la UPS, en este marco se realiza el mismo estudio que en el caso anterior (acometida principal).

Perfiles de Carga de Energía Eléctrica de los días Lunes

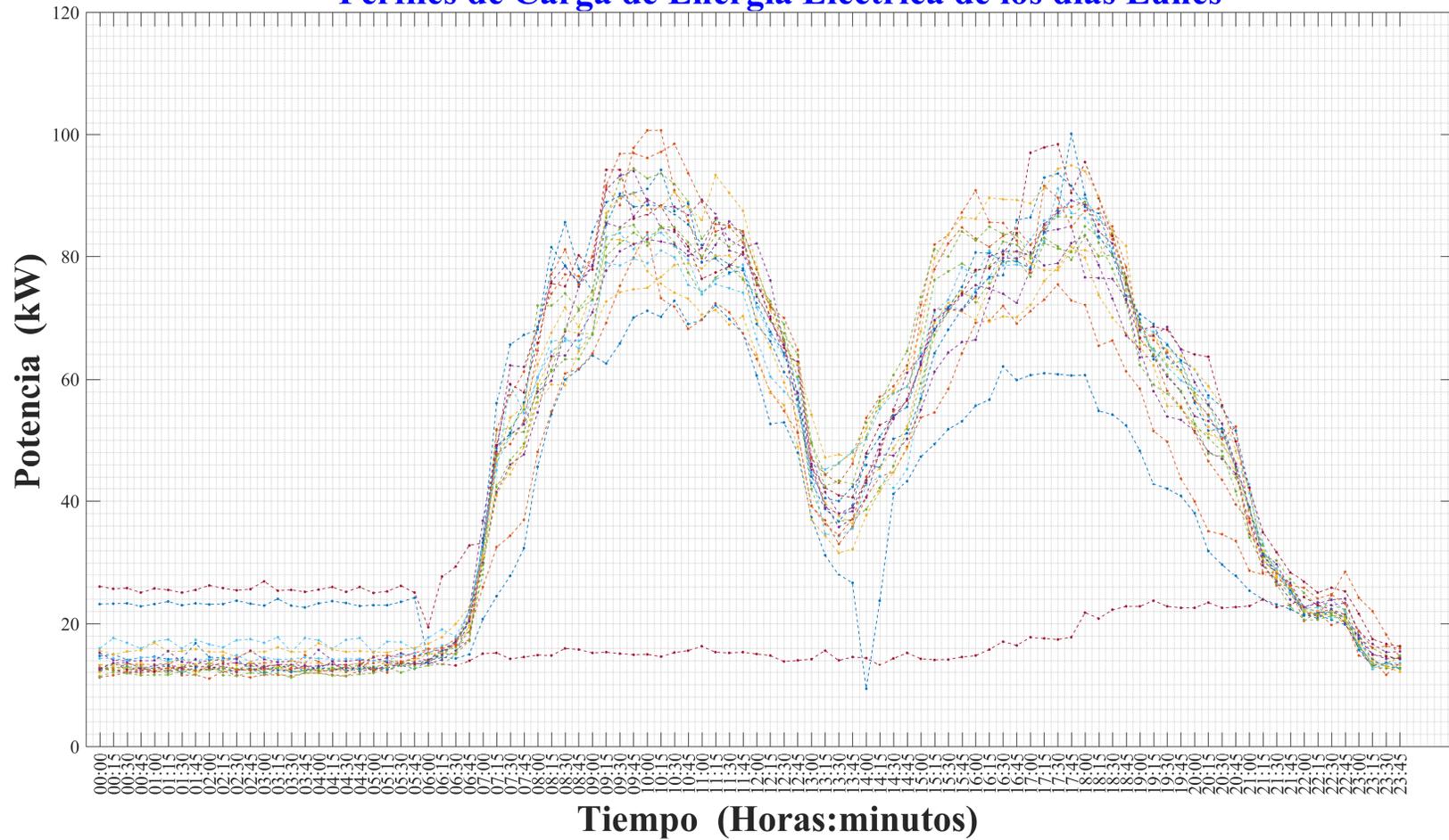


Figura 67. Perfiles de Carga de Energía Eléctrica (Lunes).

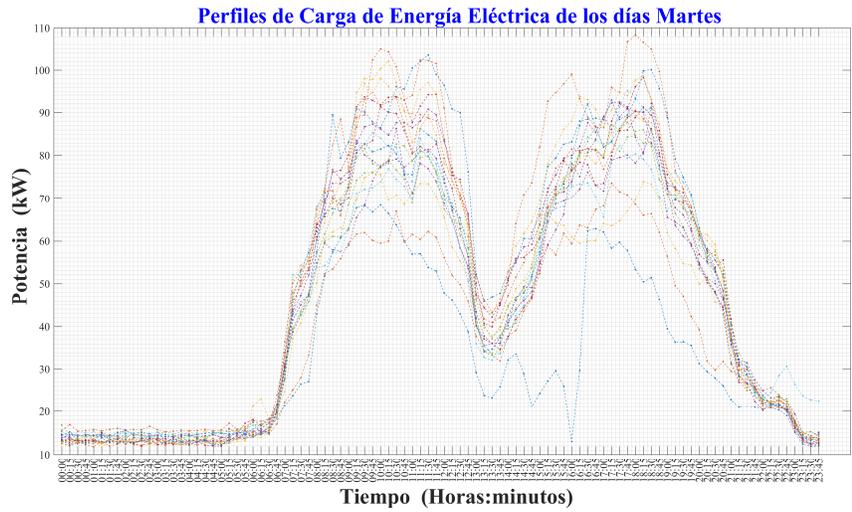


Figura 68. Perfiles de Carga de Energía Eléctrica (Martes).

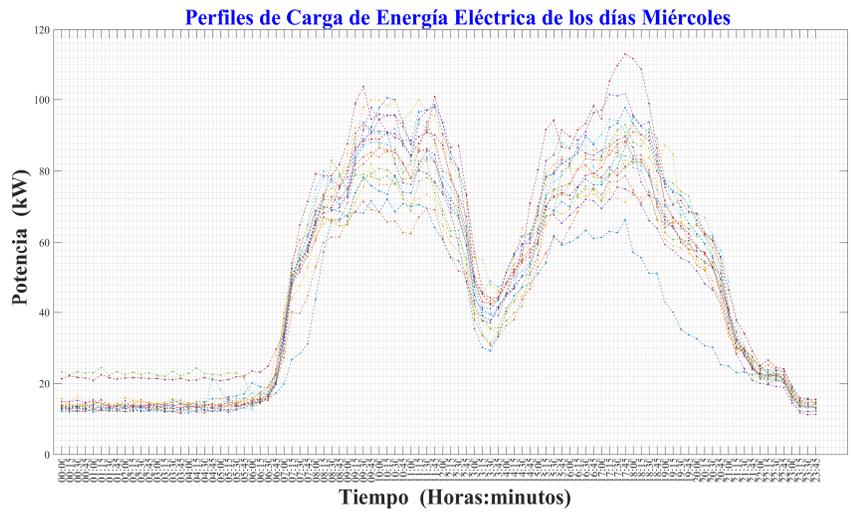


Figura 69. Perfiles de Carga de Energía Eléctrica (Miércoles).

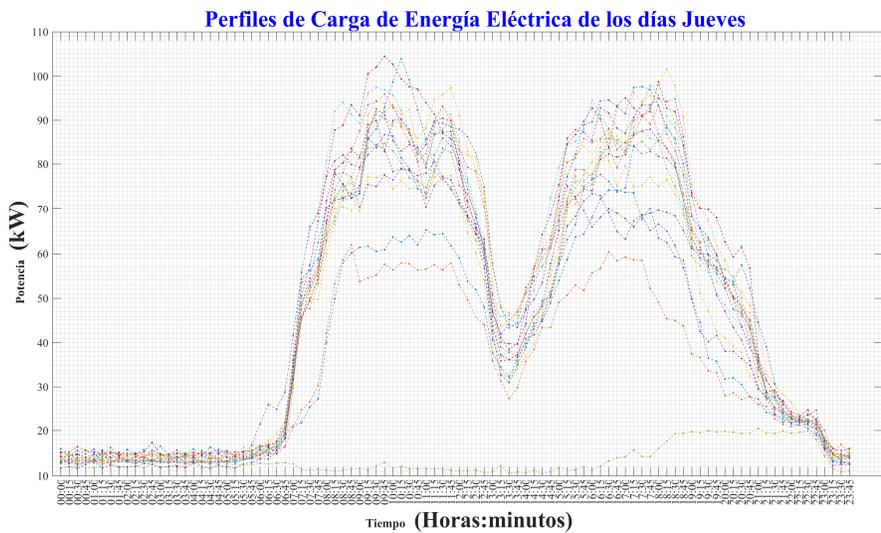


Figura 70. Perfiles de Carga de Energía Eléctrica (Jueves).

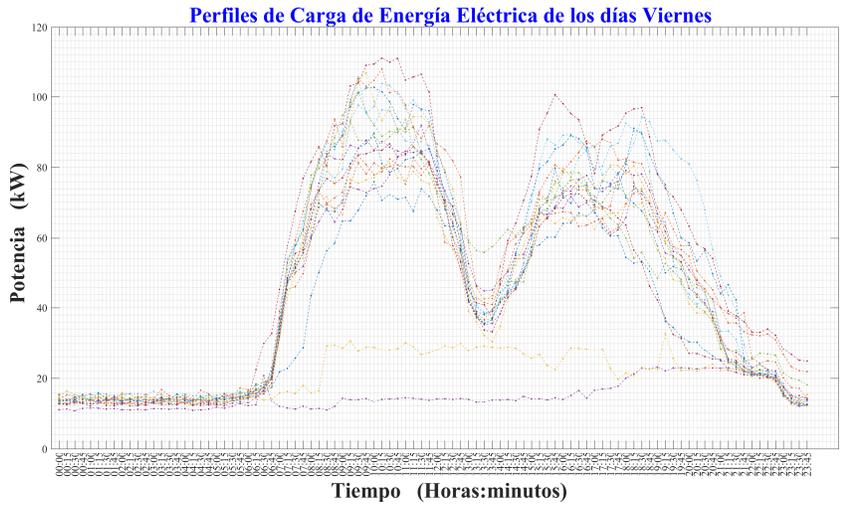


Figura 71. Perfiles de Carga de Energía Eléctrica (Viernes).

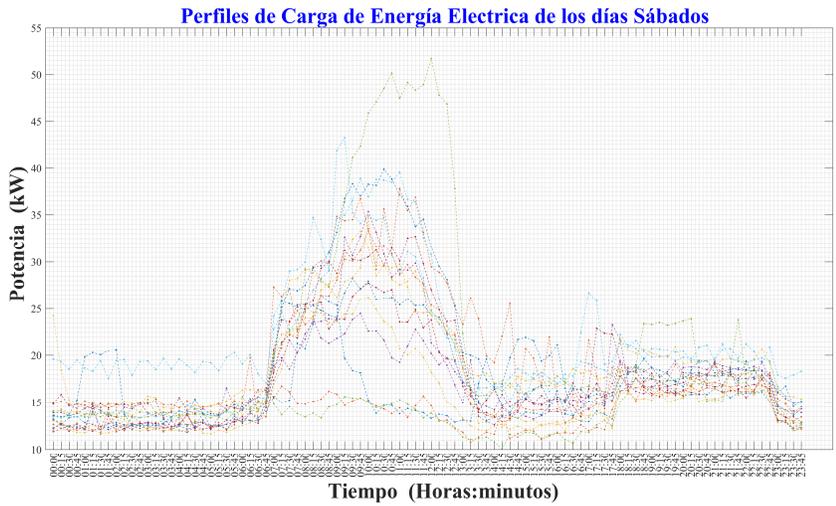


Figura 72. Perfiles de Carga de Energía Eléctrica (Sábados).

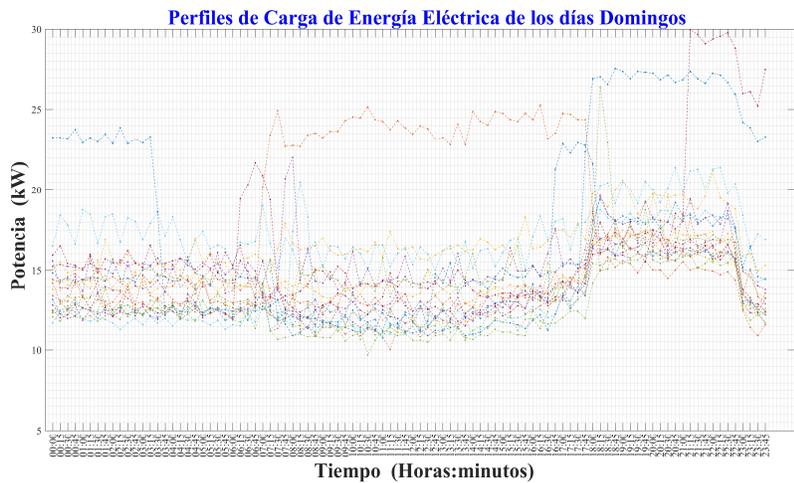


Figura 73. Perfiles de Carga de Energía Eléctrica (Domingos).

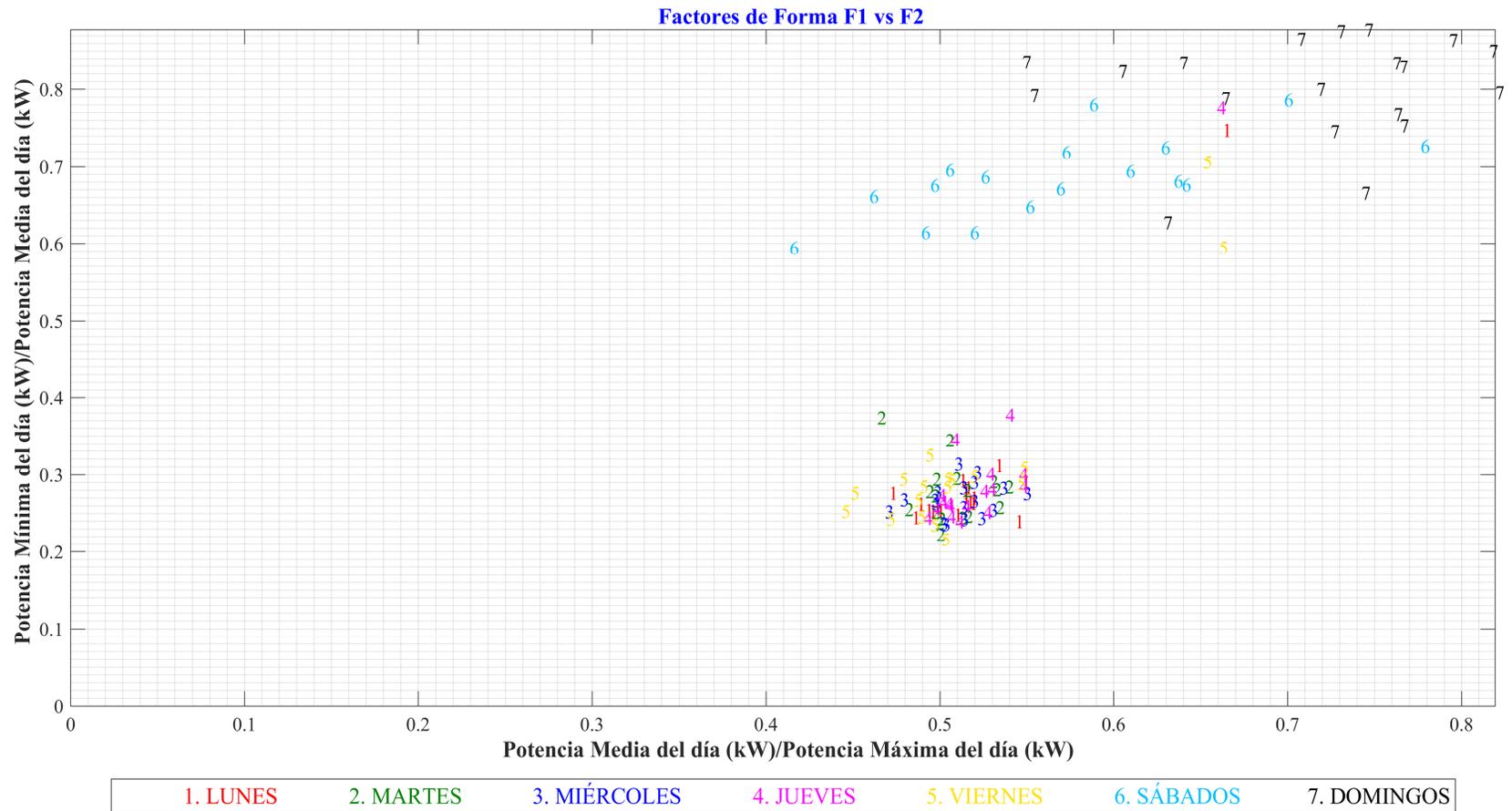


Figura 74. Factores de Forma F1 vs F2.

1. LUNES 2. MARTES 3. MIÉRCOLES 4. JUEVES 5. VIERNES 6. SÁBADOS 7. DOMINGOS

Factores de Forma F3 vs F4 vs F5

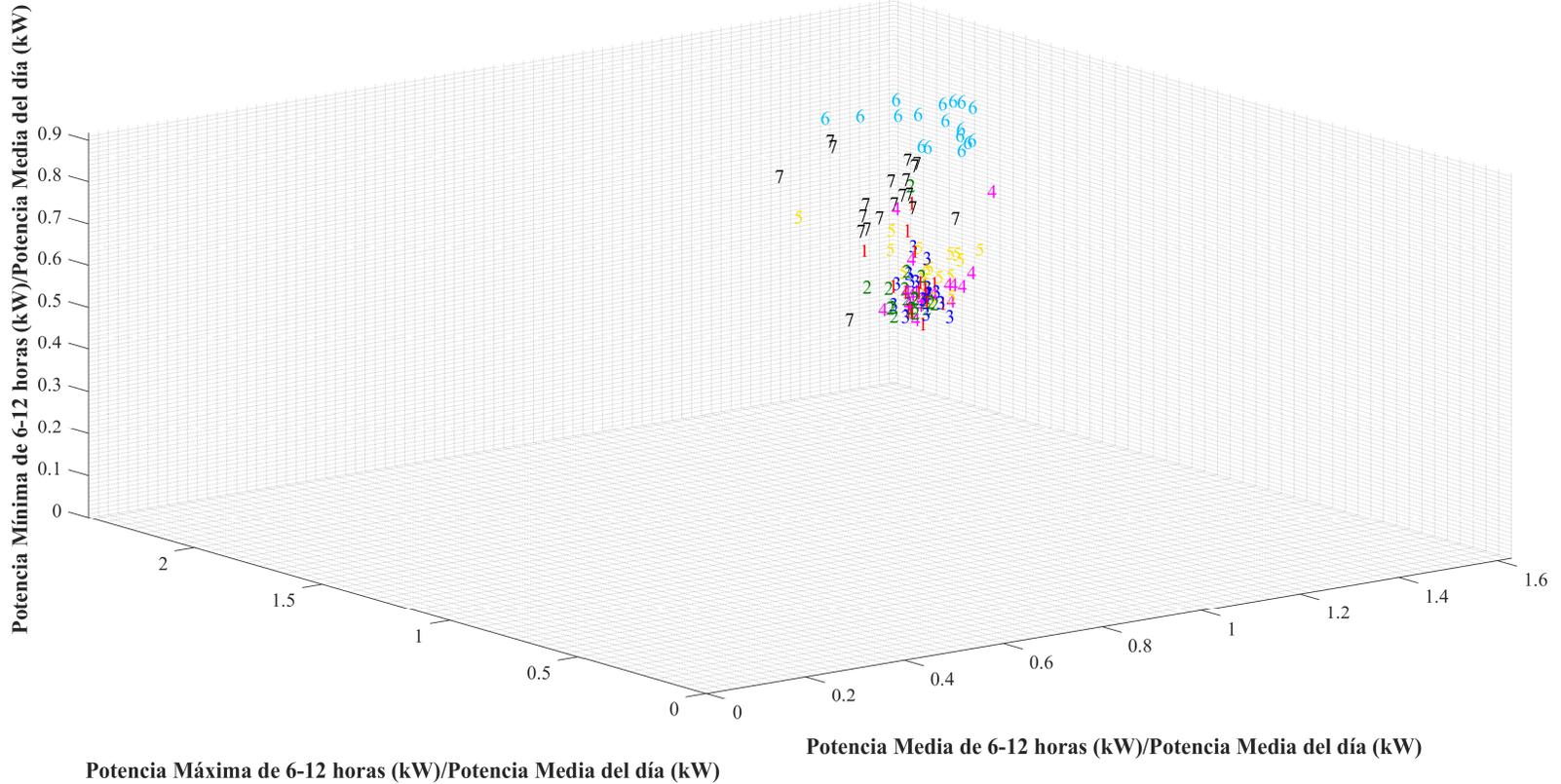


Figura 75. Factores de Forma F3 vs F4 vs F5.

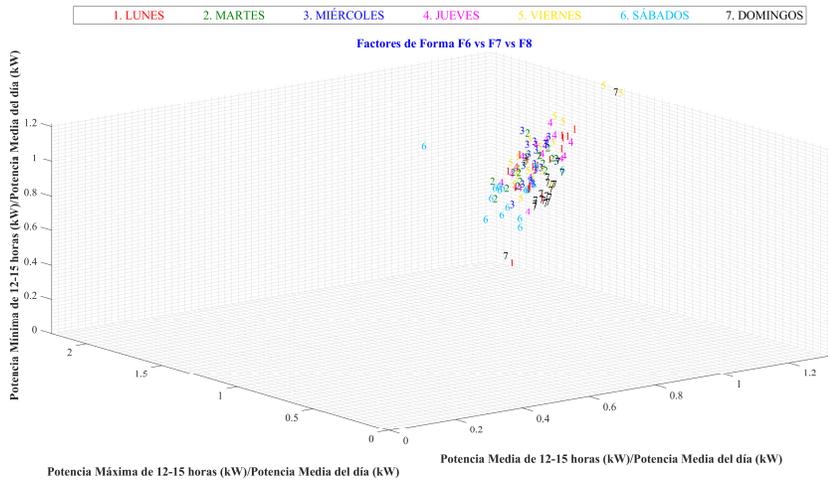


Figura 76. Factores de Forma F6 vs F7 vs F8.

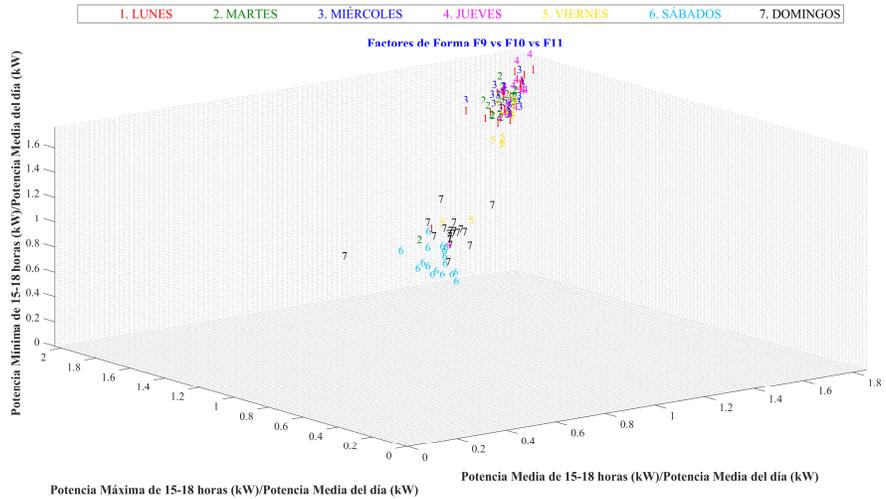


Figura 77. Factores de Forma F9 vs F10 vs F11.

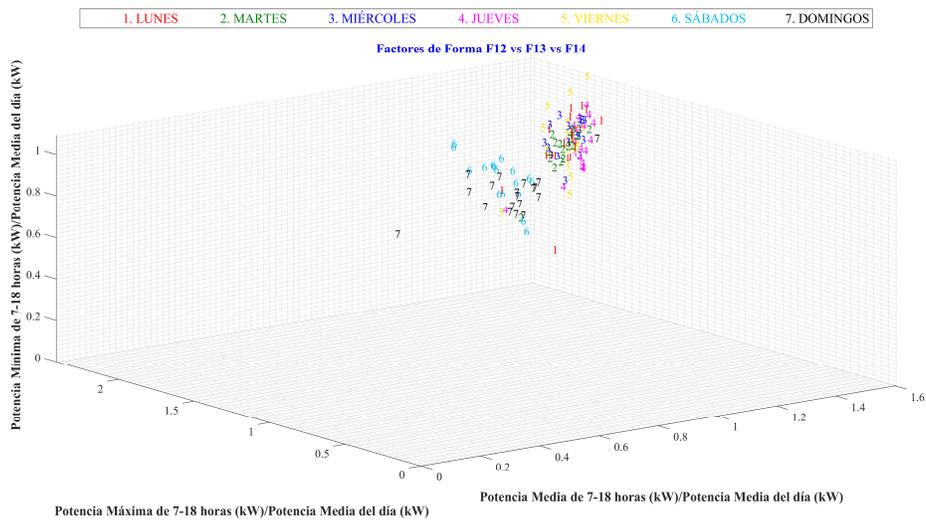


Figura 78. Factores de Forma F12 vs F13 vs F14.

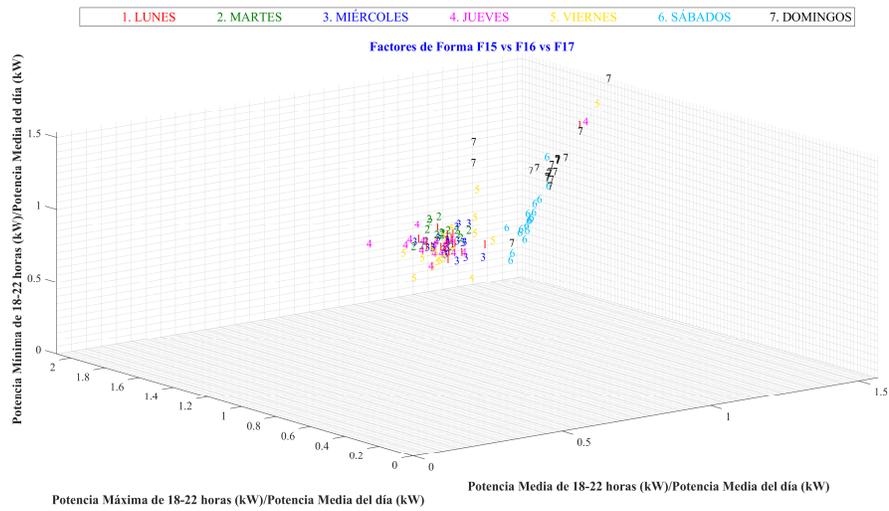


Figura 79. Factores de Forma F15 vs F16 vs F17.

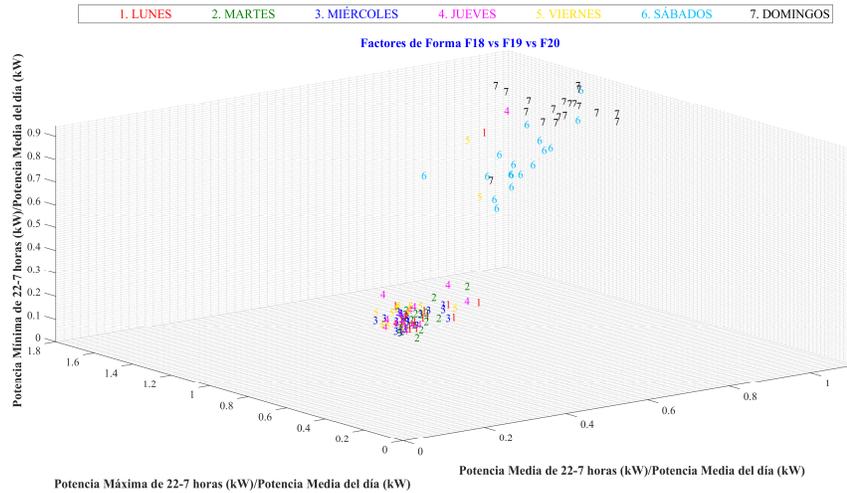


Figura 80. Factores de Forma F18 vs F19 vs F20.

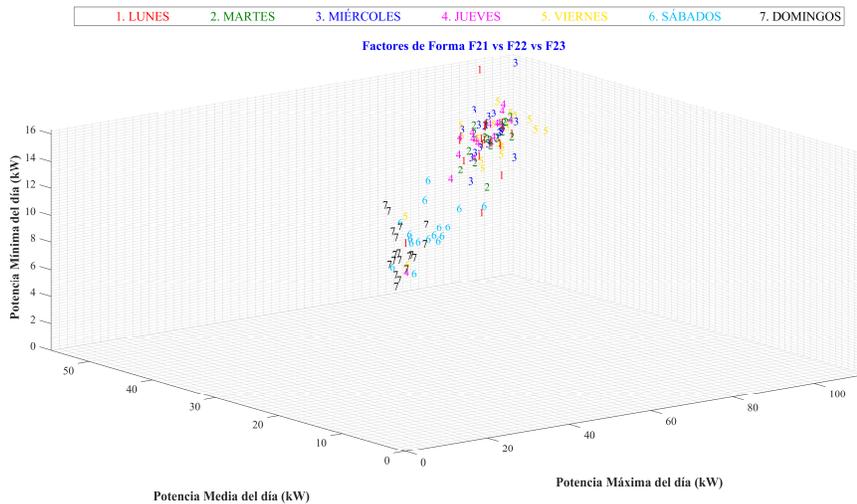


Figura 81. Factores de Forma F21 vs F22 vs F23.

4.1.2.1. Valores promedios de las siluetas según el número de clusters con los que se segmentaron los factores de forma

Tabla 11. Valores promedios de las Siluetas para el número de clusters analizados

Número de Cluster	Valor promedio de Silueta
1.0000	-
2.0000	0.9489
3.0000	0.7940
4.0000	0.7462
5.0000	0.6892
6.0000	0.7214
7.0000	0.6843

4.1.2.2. Evaluaciones del número de clusters más óptimo para los datos en análisis

El número de Clusters más apropiados según el Valor de la Silueta:

Ekmeans3 =

SilhouetteEvaluation with properties:

NumObservations: 127

InspectedK: [1 2 3 4 5 6 7]

CriterionValues: [NaN 0.9489 0.7940 0.7462 0.6892 0.7214 0.7101]

OptimalK: 2

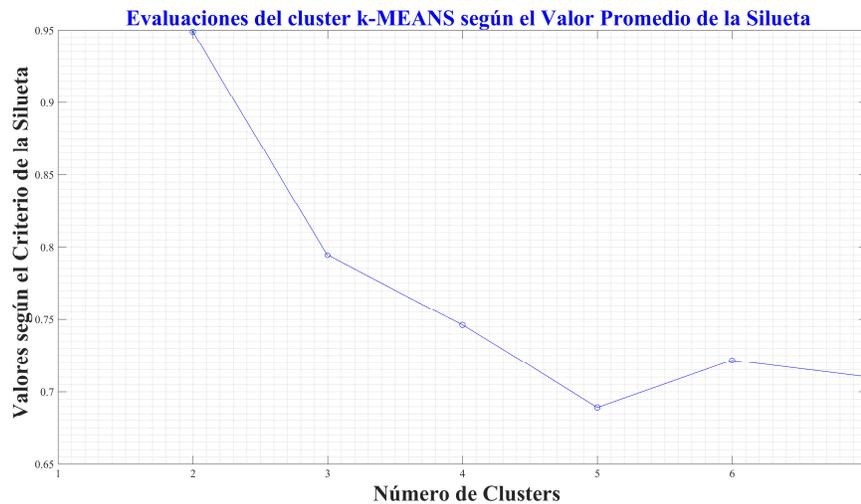


Figura 82. Evaluación según el valor promedio de Silueta del cluster k-Means.

4.1.2.3. Valores de los tiempos de ejecución para cada número de cluster con los que se segmentaron los factores de forma

Tabla 12. Tiempos de ejecución para el número de clusters analizados.

Número de Clusters	Tiempo de Ejecución (s)
[1]	[4.5245]
[2]	[8.6536]
[3]	[8.2071]
[4]	[8.1877]
[5]	[8.2165]
[6]	[8.2512]
[7]	[8.2425]

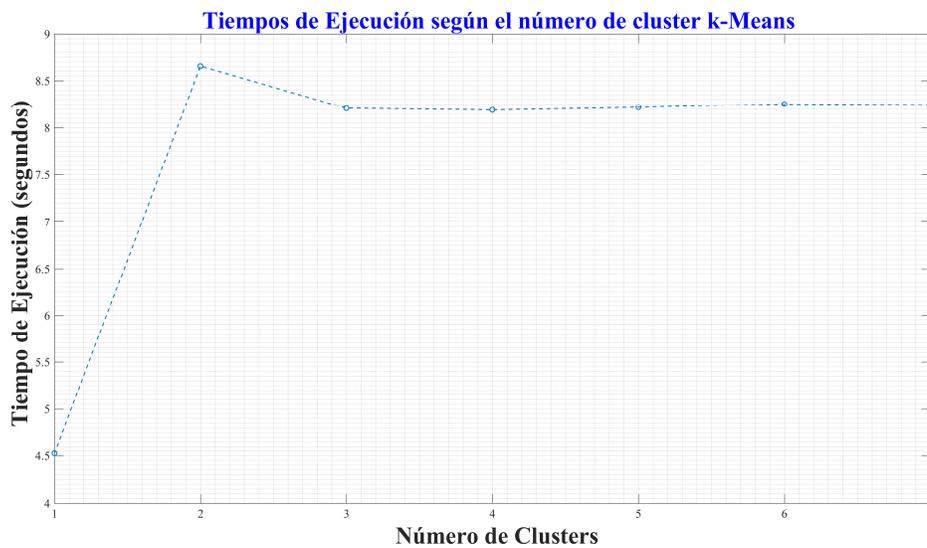


Figura 83. Tiempos de ejecución según el número de cluster k-Means.

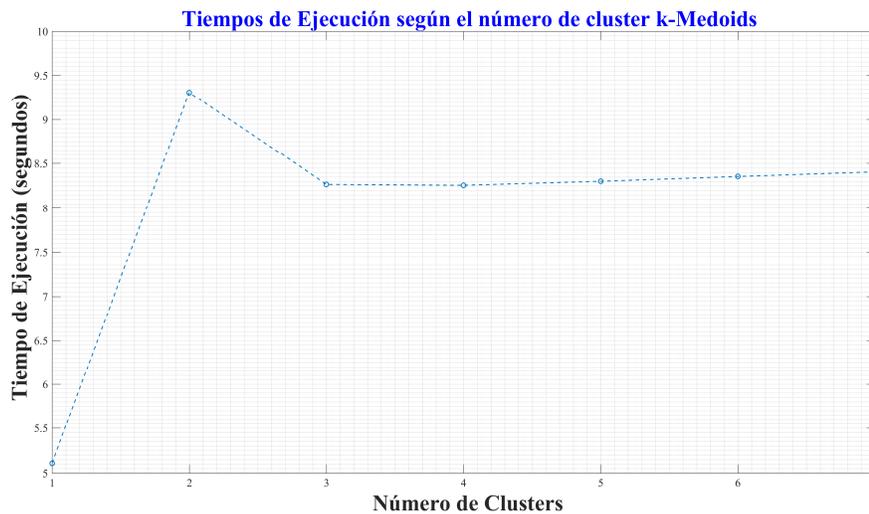


Figura 84. Tiempos de ejecución según el número de cluster k-Medoids.

Gráfica de Silueta con el número de clusters óptimo

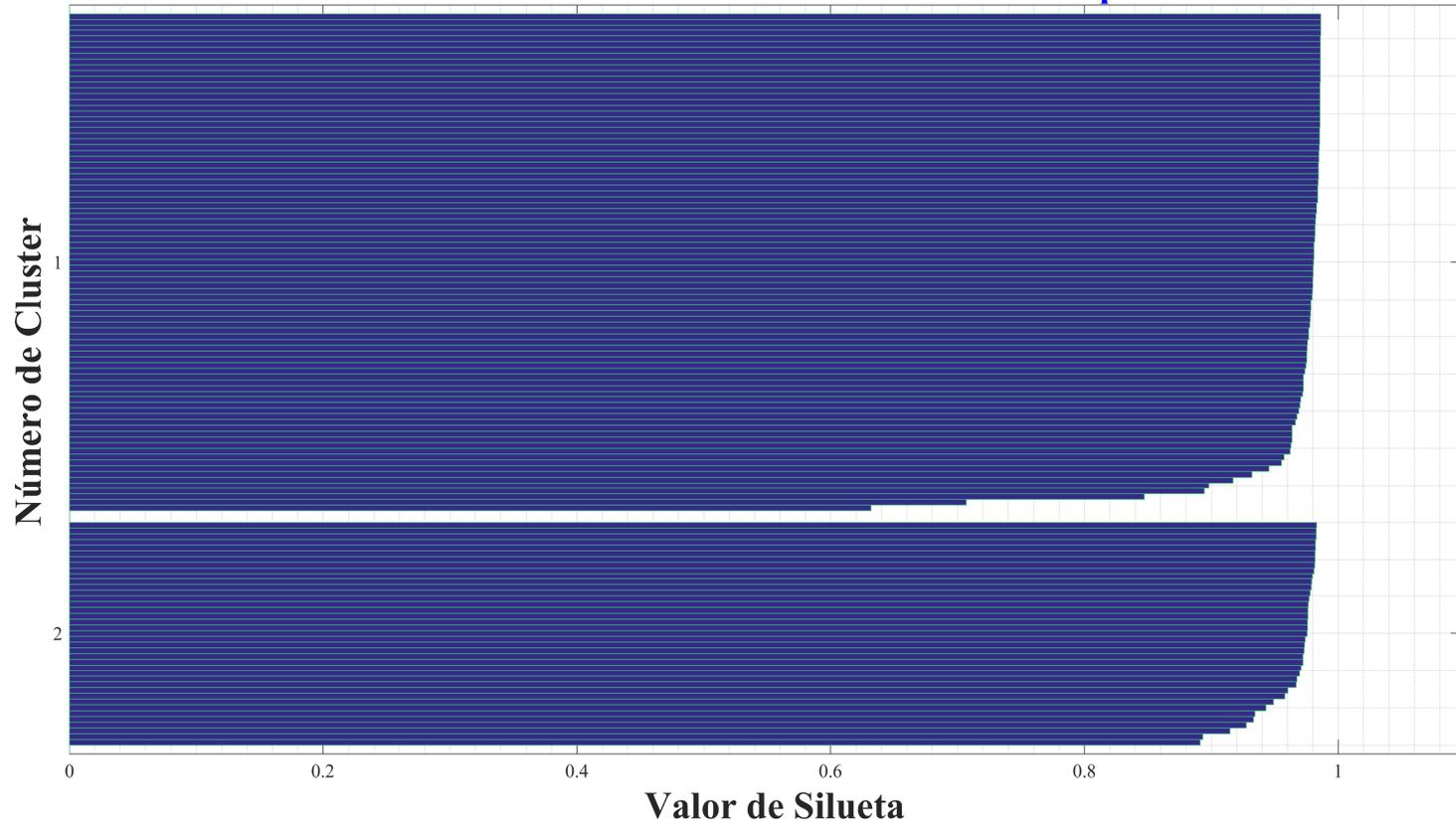


Figura 85. Gráfica de Silueta con el número de clusters óptimo.

4.1.2.4. Indicador del número de cluster a cuál pertenece cada PCEE diario

Tabla 13. Indicador del número de cluster de cada PCEE.

Número de PCEE	Cluster	Número de PCEE	Cluster	Número de PCEE	Cluster
1	2	44	2	87	2
2	2	45	2	88	2
3	2	46	2	89	2
4	2	47	2	90	2
5	2	48	2	91	2
6	2	49	2	92	1
7	1	50	2	93	1
8	2	51	2	94	1
9	2	52	2	95	1
10	2	53	2	96	1
11	2	54	2	97	1
12	2	55	2	98	1
13	2	56	2	99	1
14	2	57	2	100	1
15	2	58	2	101	1
16	2	59	2	102	1
17	2	60	1	103	1
18	2	61	2	104	1
19	2	62	2	105	1
20	2	63	2	106	1
21	2	64	2	107	1
22	2	65	2	108	1
23	2	66	2	109	1
24	2	67	2	110	1
25	2	68	2	111	1
26	2	69	2	112	1
27	2	70	2	113	1
28	2	71	2	114	1
29	2	72	2	115	1
30	2	73	2	116	1
31	2	74	2	117	1
32	2	75	2	118	1
33	2	76	2	119	1
34	2	77	1	120	1
35	2	78	2	121	1
36	2	79	2	122	1
37	2	80	2	123	1
38	2	81	2	124	1
39	2	82	2	125	1
40	2	83	1	126	1
41	2	84	2	127	1
42	2	85	2		
43	2	86	2		

4.1.2.5. Ubicación de los centroides para cada cluster

Tabla 14. Centroides para cada cluster

CLUSTER 1	CLUSTER 2
0,643984916633683	0,506359155437916
0,742875759492336	0,272679232361077
1,11079677032379	1,42659688709157
1,40071534951414	1,93421286247043
0,792673964073031	0,352021379620203
0,934189004519386	1,11618847415941
1,12487239067440	1,51982569932116
0,817859946284545	0,814329528848269
0,959420370244242	1,68612127135506
1,18363012416071	1,86011848524549
0,843987394310157	1,48103824120953
1,03809472164898	1,50003223458390
1,48622283274012	1,97689986085147
0,782793139078903	0,798297239781328
1,13603684313426	1,10433283539003
1,25297721738160	1,76106386767038
1,06082335808065	0,494982399726080
0,892978965480466	0,342479342001891
1,22061030760859	0,691570038989503
0,779689164662371	0,273515040375068
26,5510000000000	91,6591264367816
16,3650375000000	46,3271934865900
12,0064000000000	12,5324137931035

4.1.2.6. Distancia de los factores de forma que caracterizan cada PCEE diario hasta el centroide de cada cluster

Tabla 15. Cuadro de distancias de cada Factor de Forma al centroide de cada cluster

CLUSTER 1	CLUSTER 2
2514,05747993415	485,707633321234
3834,90554042363	108,829263606832
3993,74239451822	73,0000024794627
5446,97391497066	6,24481446315526
5532,36857020561	8,54462954592480
4853,08720415828	4,55020295332610
5,63610362638744	5455,99508981374
5477,18708001221	5,98348196013413
6229,53113494646	52,5228855786398
3824,73033455502	105,784247801473
4818,72772095521	6,15768875641531
4778,98394454244	7,36480437924887
5109,81201692366	1,14740086780781
6474,71646456433	95,4484849383869
6529,64110992166	83,1632475288694

6537,44828620039	86,4977526251116
5757,79834699837	18,8254781897091
4825,44623547620	5,81190787737887
4327,71026729312	45,0107000872361
1994,65360192023	754,004843489222
2650,30030118126	410,733002731090
3039,80849730424	278,766482426481
4280,18644331320	40,6841449242578
4400,25019964650	29,2281687815581
5654,72451240212	13,6075427705189
6096,94313979384	44,9294268536067
5233,39889691134	1,40235768074690
5680,34833637065	14,2763487253357
6161,91191186093	46,7681254655581
5203,34412024862	1,07052790632125
4651,38569730654	16,2786239036285
3874,02419565487	92,1717528840364
5605,49061088271	11,0874630635613
7167,97183047647	168,273677926354
8132,91048787260	341,909616455039
6902,30533005981	129,575199377675
5176,91586578216	4,81932958997713
2474,95360305354	484,241662797640
3320,53833653038	198,677062358530
3763,75584010767	107,599370849099
5377,24321056846	13,9503479274858
5320,19709788948	2,45905791498765
5762,31485688960	21,9177428480605
6525,05358770229	86,9723231261735
5695,24364310818	14,7731579016343
5555,26381714627	9,46701265663021
5479,61801127673	13,4290888022552
5542,98030958649	8,87070019707739
4581,47608228344	27,0822767860674
5535,88637633097	8,46037874241841
9071,05181986010	560,598400584590
6632,21685294748	95,2669038874110
5152,26090854770	2,70430817740865
6590,78499745721	90,8842654418834
6848,90491774416	122,553272917972
2731,95805057294	379,436739559168
1547,94528330155	1048,96196324466
3741,56819963856	119,365441728750
3457,07740570791	173,544388022554
48,7538375510225	6162,12049798387
6167,36055643393	47,0835804467018
5053,85495604556	3,19365082998992
4469,80023916140	25,1347031929874
5410,69291655969	3,54378124741036

6812,68291374695	118,870423588415
4156,26008630376	55,5468404883376
4373,94306790620	33,5614552138714
6118,50091623746	42,5759869897928
7378,46684824832	201,812693391290
7170,69359156465	169,188673913242
5915,57015193578	27,8734878368953
6074,08865568247	39,1413009336643
5747,91859625549	18,1122210343280
2889,80370419619	326,227158445279
3982,13315040366	84,5803117596363
3587,76571266935	140,071263929104
15,8283259600474	5682,46353461764
5333,36915171677	3,12890629016225
6134,98023868990	47,0955018343850
4227,68981010330	52,0073696417084
4771,46033751974	9,30619523623403
4060,43394965585	71,9757611617557
67,7488836826664	4081,68641554043
4136,25556978922	55,4610954140153
6497,44273541605	100,409387871020
7507,27508239007	224,821380105009
8684,13495182074	462,510670393291
7030,23243747437	148,211627801599
7821,03940569412	287,091501903145
7435,04194860578	235,905760213165
5095,82623335039	3,66186903682848
1,16205362732410	5208,91423144271
60,8660225749952	6277,97038091002
0,348719028505534	5182,10491137049
79,4308562745955	4008,33468884865
14,7756902681530	5613,85751134301
216,944498411419	3291,86091765953
3,11603772627990	4918,03358772460
5,98831132112518	4828,45313400979
139,536985119906	3613,94164053692
82,6889396388024	3960,38683679245
6,95259973316549	5374,76351913346
661,779413321036	2224,80842502879
295,240976391754	3042,42676550506
41,3128662843202	4283,98880022424
188,858793636983	3405,34989939411
47,9440158308134	4241,29275461168
46,4928778173286	4289,98562765078
20,9804299698101	5791,16024504345
77,0824076794273	6440,59479071679
52,6012626941575	6220,24137315618
96,4428113912236	6615,28774126074
119,377569188305	6788,52696471228

36,2494600978556	5782,86010730800
30,7635276220666	5961,46653611797
54,0043787211017	6214,65059184927
69,8235480826976	6392,31843535682
35,7304146434793	5851,53313460823
23,5148617980829	5806,07691274523
4,23598420917836	5275,00548164494
46,4946096115762	6146,17566360855
14,1388168515154	4702,33111674738
5,20182537480062	4962,33813494518
8,48144942011740	5173,11023525596
50,7835069003851	6184,57038895479
69,5361286500383	6398,81485917291
112,722857500413	6757,67949253453

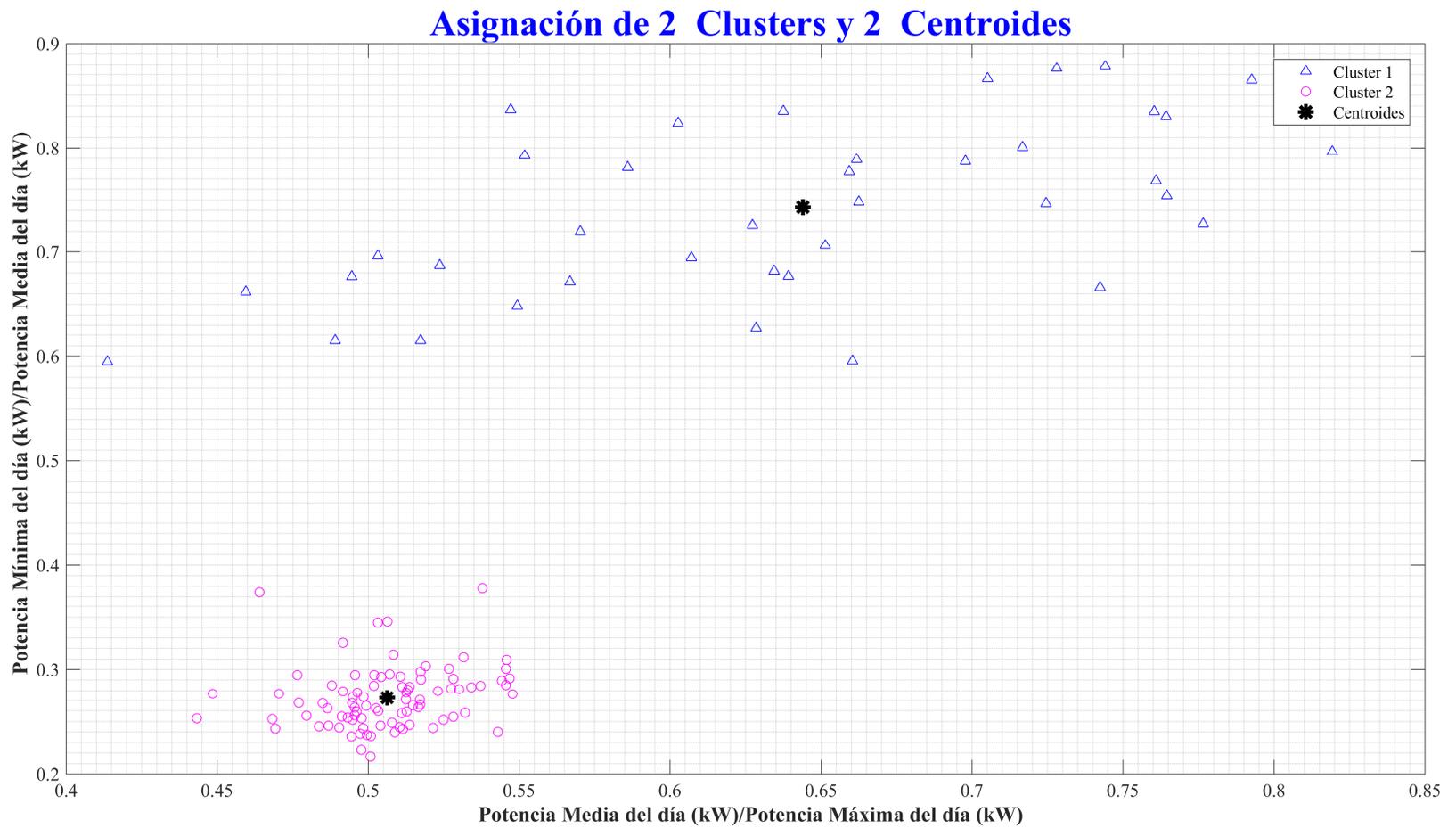


Figura 86. Segmentación de los datos en dos clusters.

Asignación de 2 Clusters y 2 Centroides

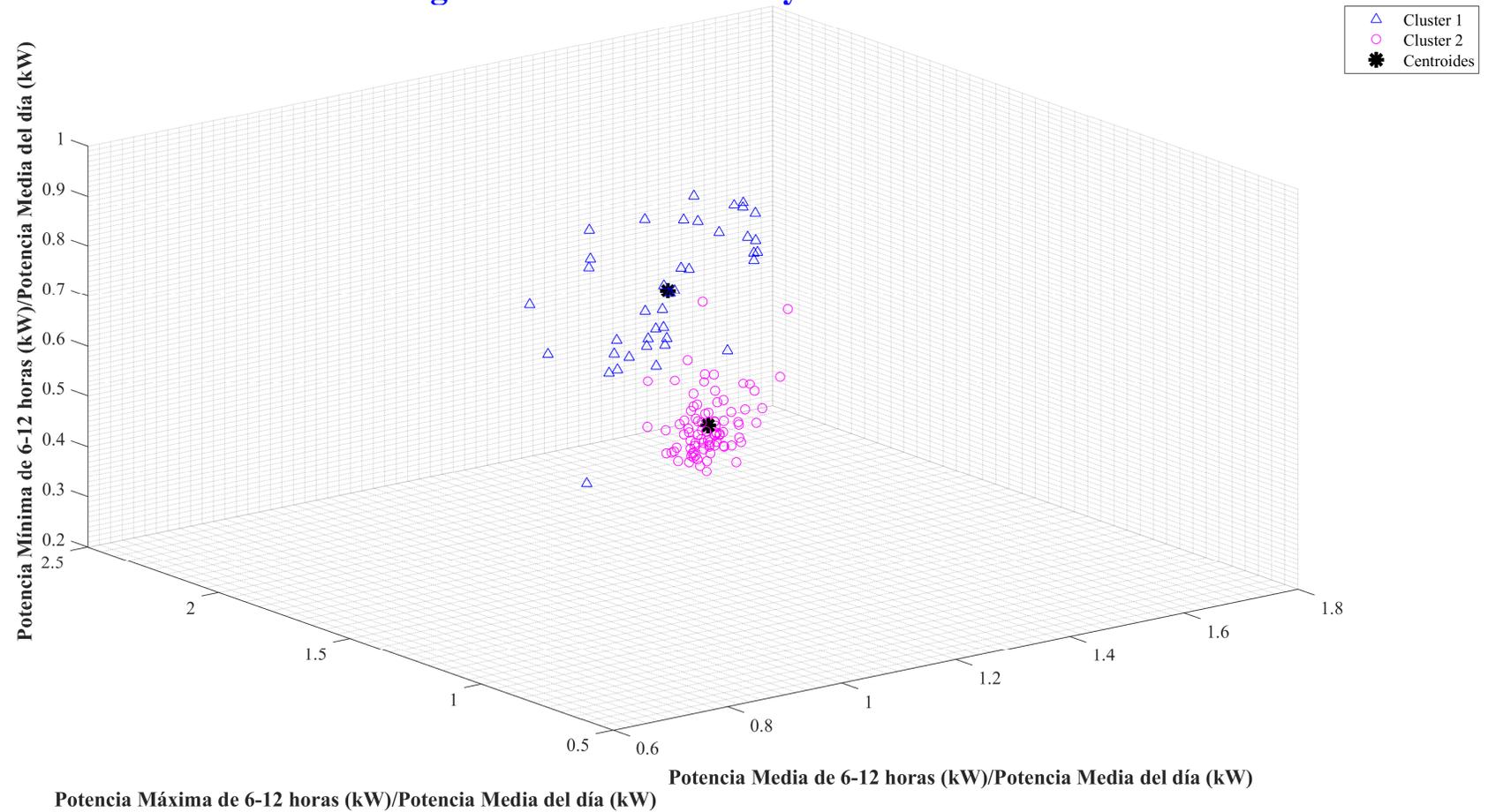
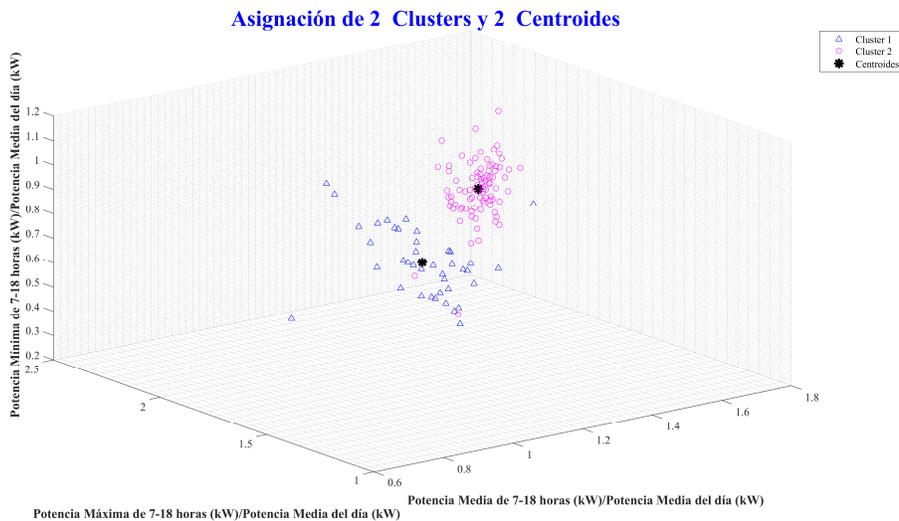
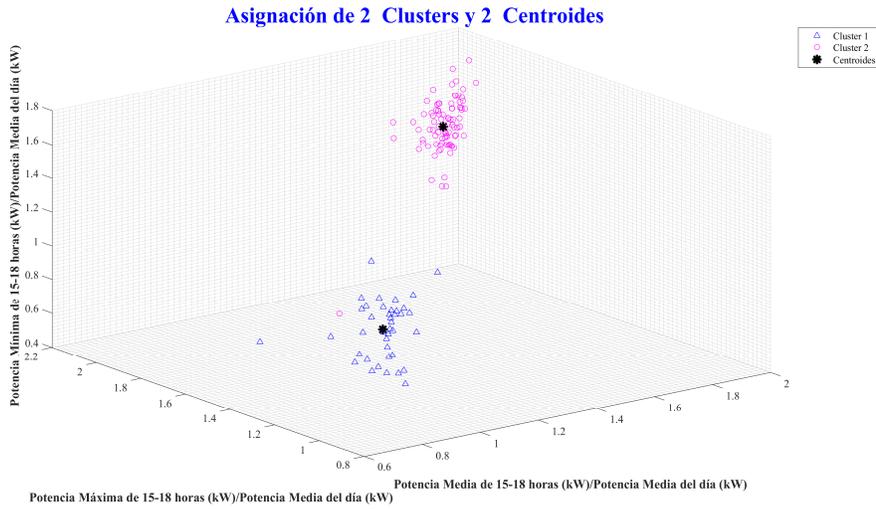
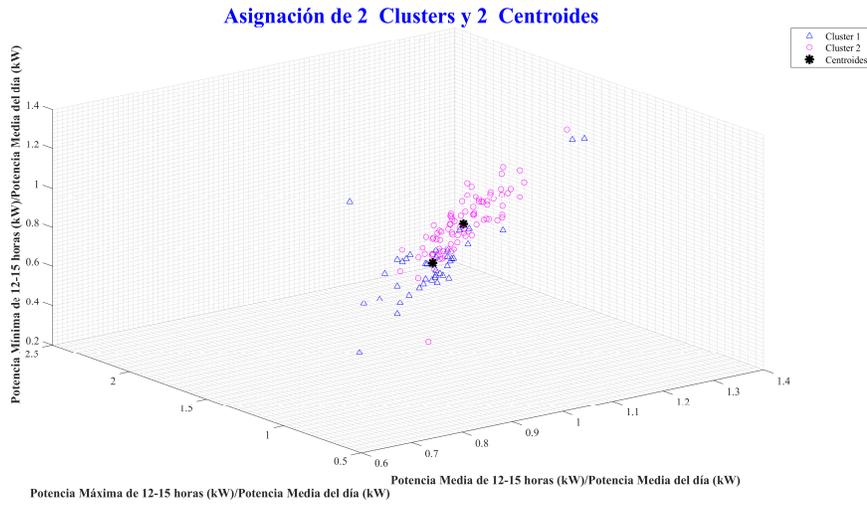


Figura 87. Segmentación de los datos en dos clusters.



Asignación de 2 Clusters y 2 Centroides

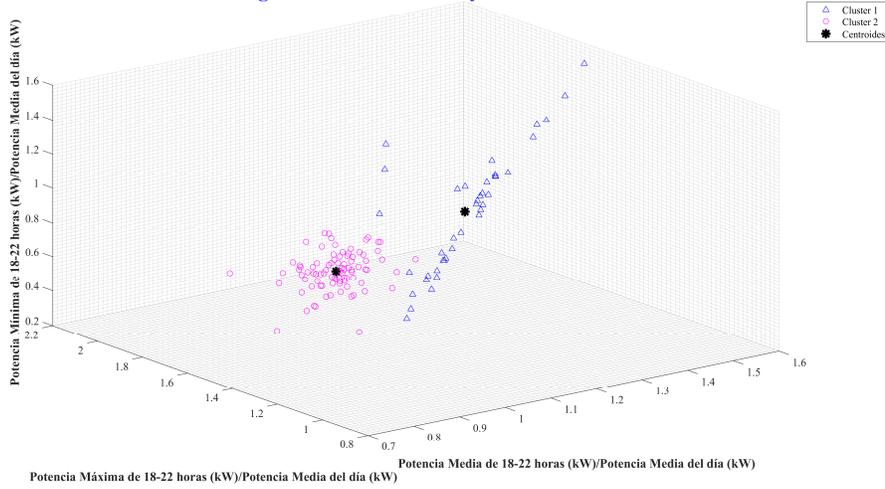


Figura 91. Segmentación de los datos en dos clusters.

Asignación de 2 Clusters y 2 Centroides

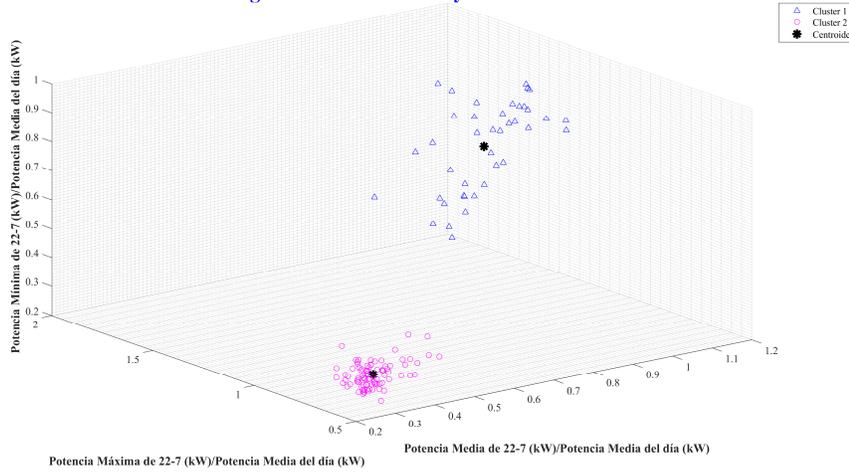


Figura 92. Segmentación de los datos en dos clusters.

Asignación de 2 Clusters y 2 Centroides

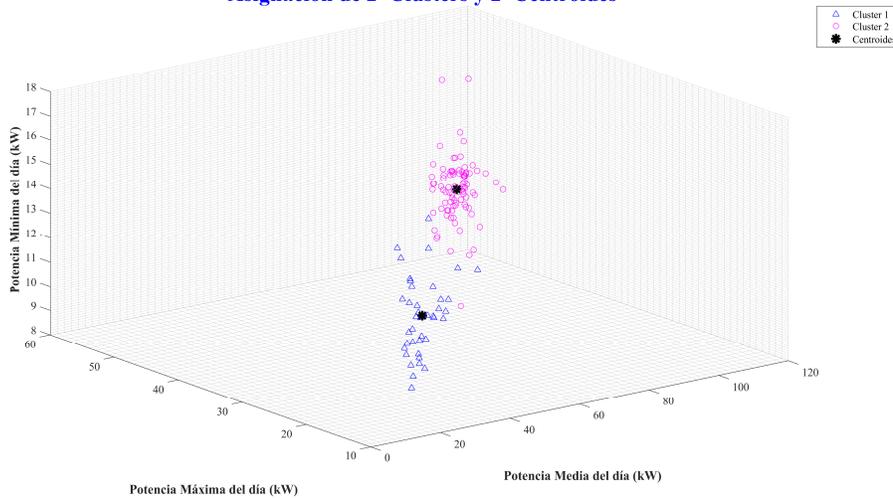


Figura 93. Segmentación de los datos en dos clusters.

4.1.2.7. Análisis estadístico de media aritmética y desviación estándar de las distancias a los centroides y de los valores de silueta de cada punto para la detección de perfiles de carga diarios con outliers y sin outliers de cada cluster

MediaAritméticaDCDelCluster1 =

73.9893

DesviaciónEstándarDCDelCluster1 =

112.4843

RangoIntercuartilDCDelCluster1 =

63.7994

ValorMínimoDCDelCluster1 =

47.6115

AnchoDeBandaDCDelCluster1 =

20.4900

limiteatipicos1 =

164.2126

LimitesDeLosValoresAtipicosParaElCluster1medianteLaDistanciaAIC =

164.2126

MediaAritméticaVSDelCluster1 =

0.9647

DesviaciónEstándarVSDelCluster1 =

0.0507

RangoIntercuartilVSDelCluster1 =

0.0121

ValorMínimoVSDelCluster1 =

0.0090

AnchoDeBandaVSDelCluster1 =

0.0039

limiteatipicos2 =

0.9322

LimitesDeLosValoresAtipicosParaElCluster1medianteLosValoresDeSi =

0.9322

MediaAritméticaDCDelCluster2 =

117.5805

DesviaciónEstándarDCDelCluster2 =

175.8370

RangoIntercuartilDCDelCluster2 =

116.1468

ValorMínimoDCDelCluster2 =

86.6768

AnchoDeBandaDCDelCluster2 =

31.9331

limiteatipicos1 =

343.0918

LimitesDeLosValoresAtipicosParaElCluster2medianteLaDistanciaAIC =

343.0918

MediaAritméticaVSDelCluster2 =

0.9416

DesviaciónEstándarVSDelCluster2 =

0.0932

RangoIntercuartilVSDelCluster2 =

0.0237

ValorM0xEDnimoVSDelCluster2 =

0.0177

AnchoDeBandaVSDelCluster2 =

0.0065

limiteatipicos2 =

0.9077

LimitesDeLosValoresAtipicosParaElCluster2medianteLosValoresDeSi =

0.9077

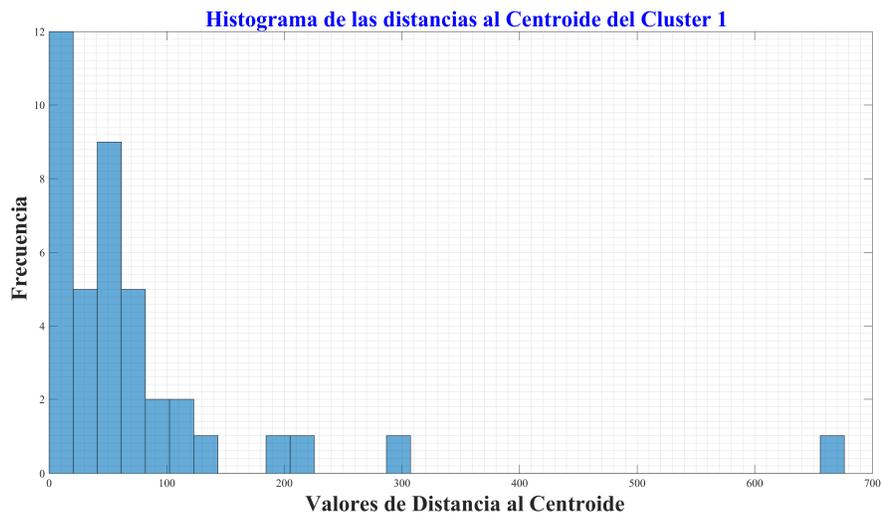


Figura 94. Histograma con las distancias al centroide del cluster 1.

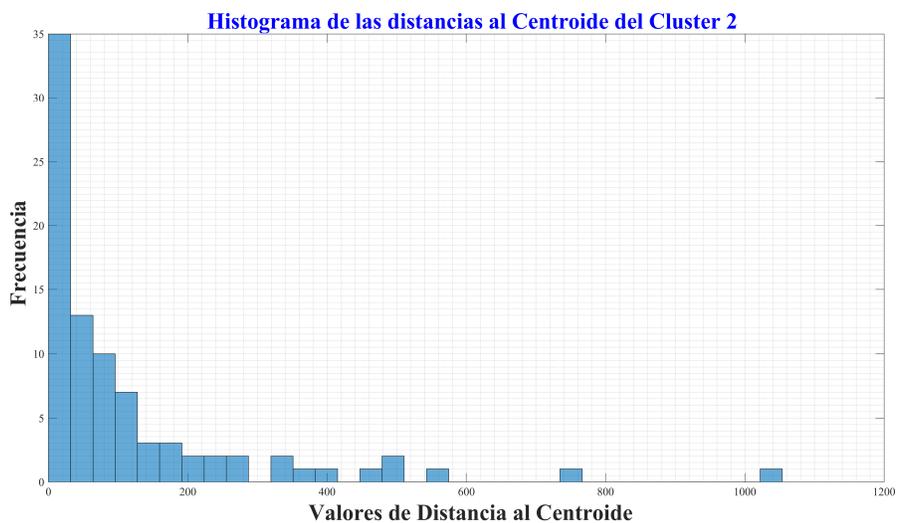


Figura 95. Histograma con las distancias al centroide del cluster 2.

Estimación de densidad por kernel de Epanechnikov de las distancias al Centroides del Cluster 1

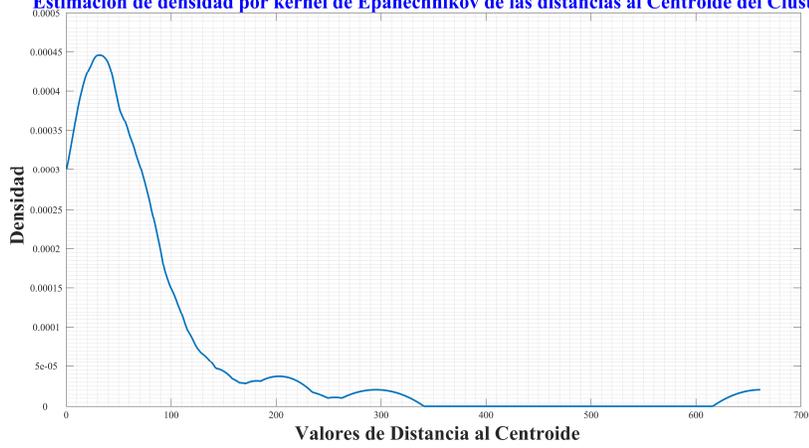


Figura 96. FDP de las distancias al centroide del cluster 1.

Estimación de densidad por kernel de Epanechnikov de las distancias al Centroides del Cluster 2

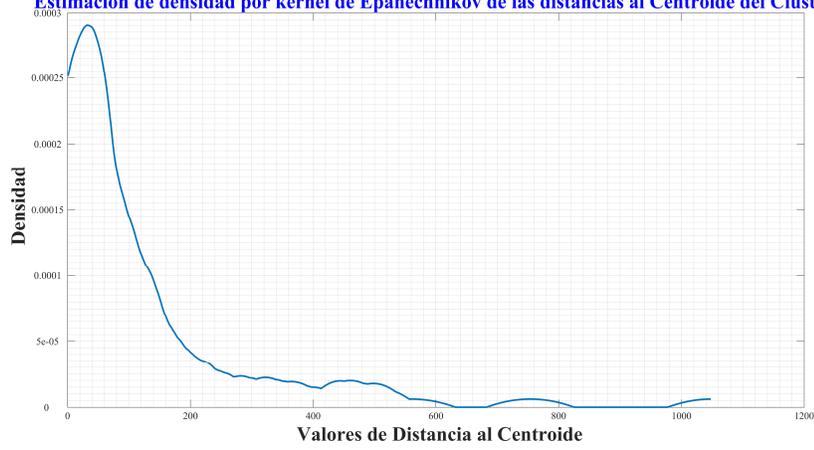


Figura 97. FDP de las distancias al centroide del cluster 2.

Función de Distribución Acumulativa Inversa de las distancias al Centroides del Cluster 1

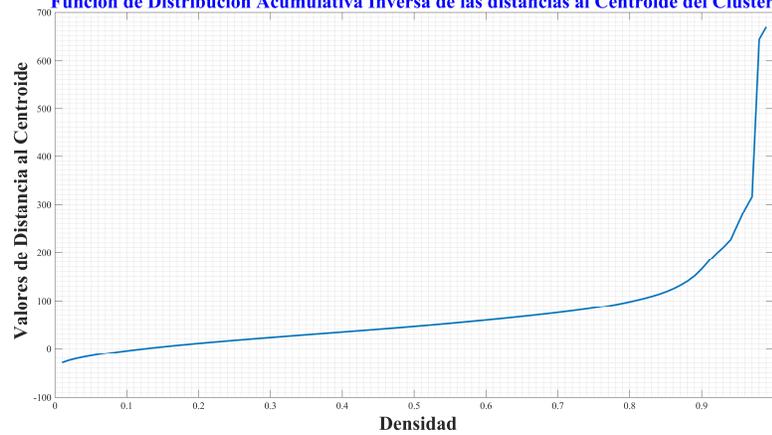


Figura 98. FDAI de las distancias al centroide del cluster 1.

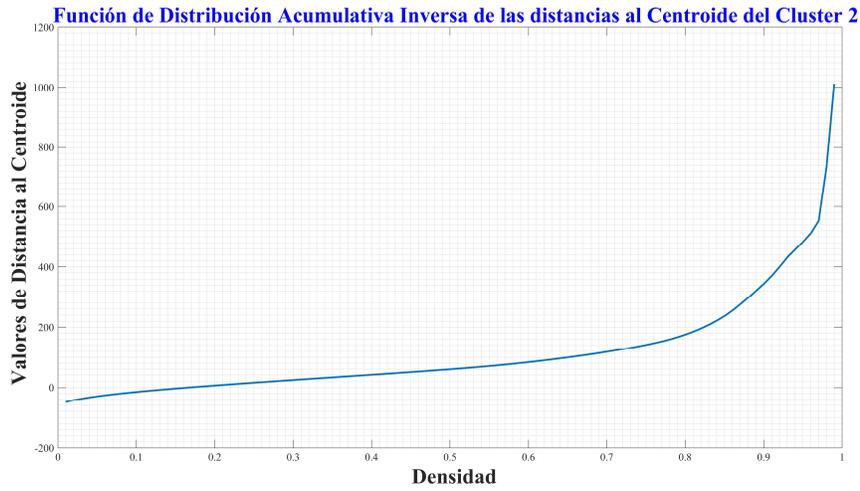


Figura 99. FDAI de las distancias al centroide del cluster 2.

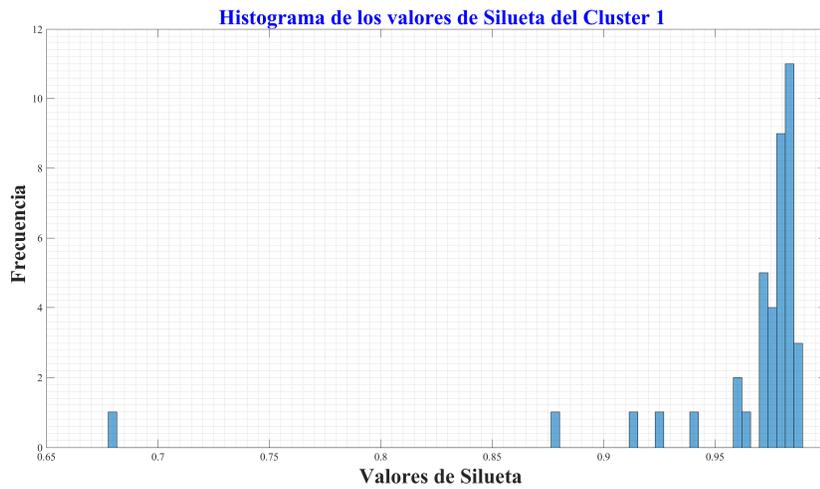


Figura 100. Histograma de los valores de Silueta del cluster 1.

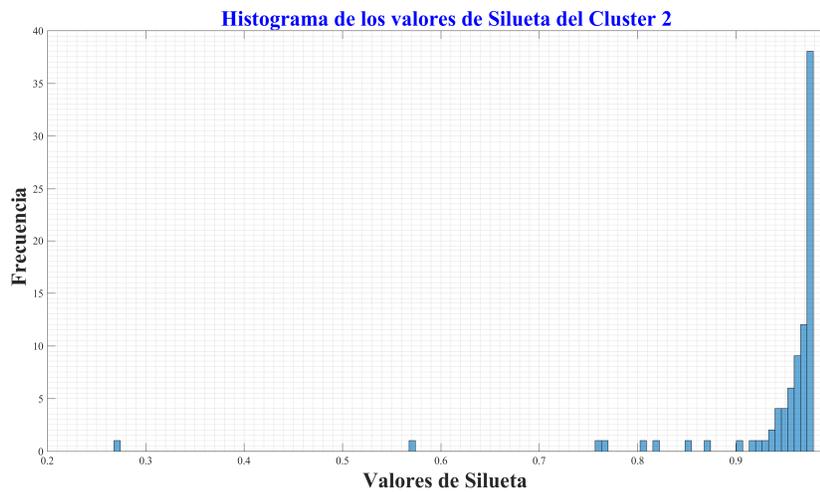


Figura 101. Histograma de los valores de Silueta del cluster 2.

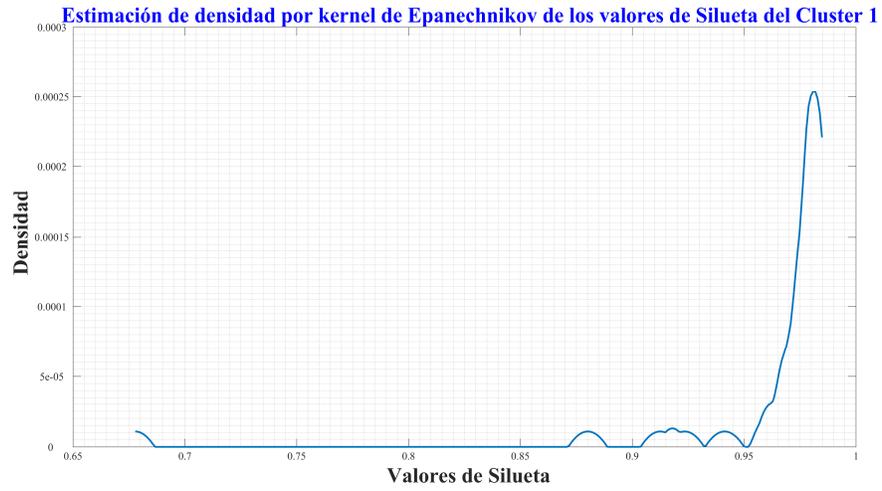


Figura 102. FDP de los valores de silueta del cluster 1.

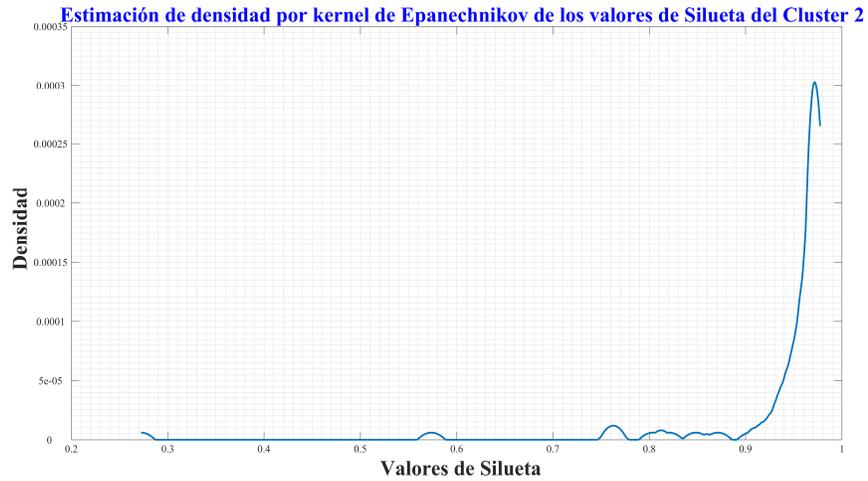


Figura 103. FDP de los valores de silueta del cluster 2.

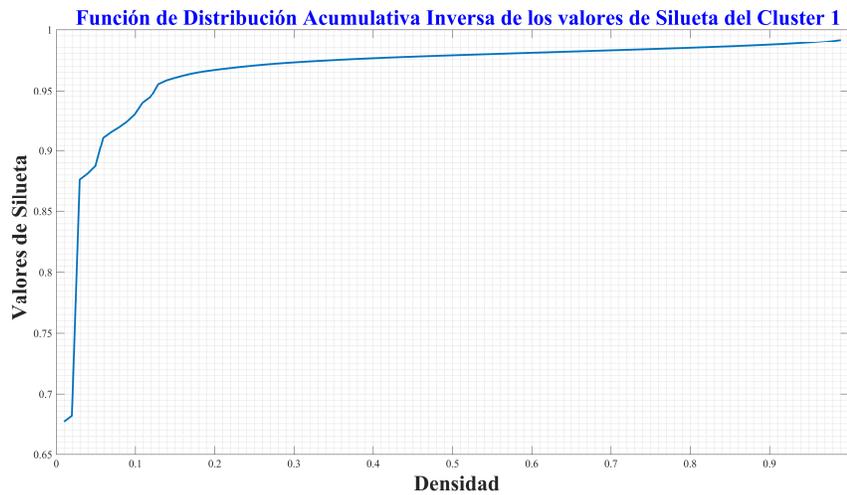


Figura 104. FDAI de los valores de silueta del cluster 1.



Figura 105. FDAI de los valores de silueta del cluster 2.

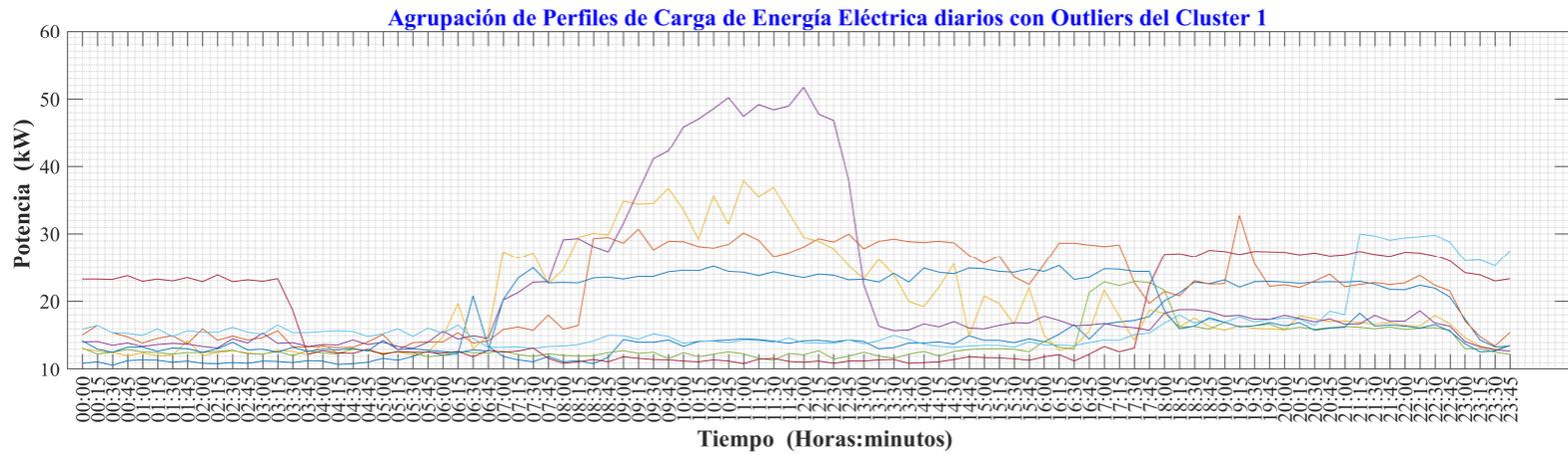
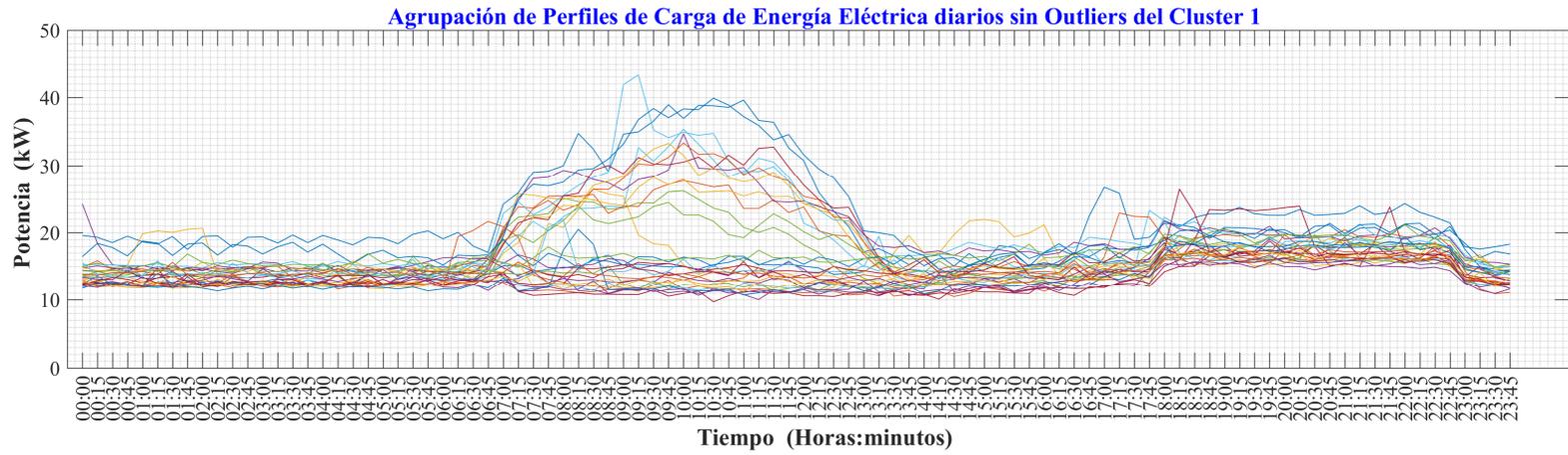


Figura 106. PCEE del cluster 1 (Perfiles sin outliers, 36 - Perfiles con outliers, 8).

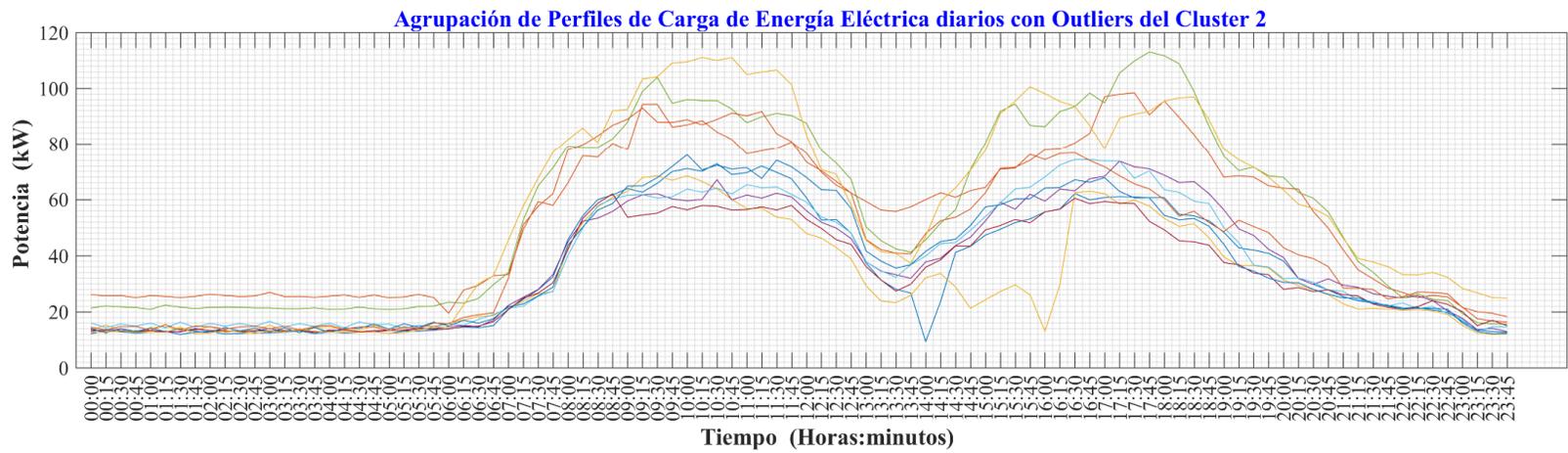
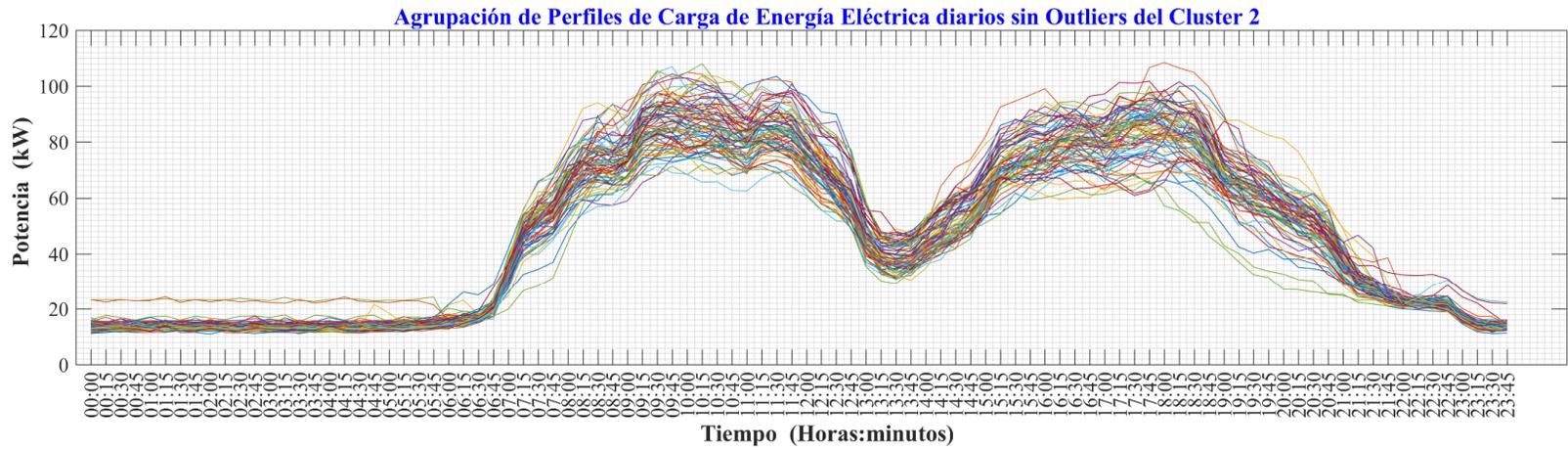


Figura 107. PCEE del cluster 2 (Perfiles sin outliers, 81 - Perfiles con outliers, 10).

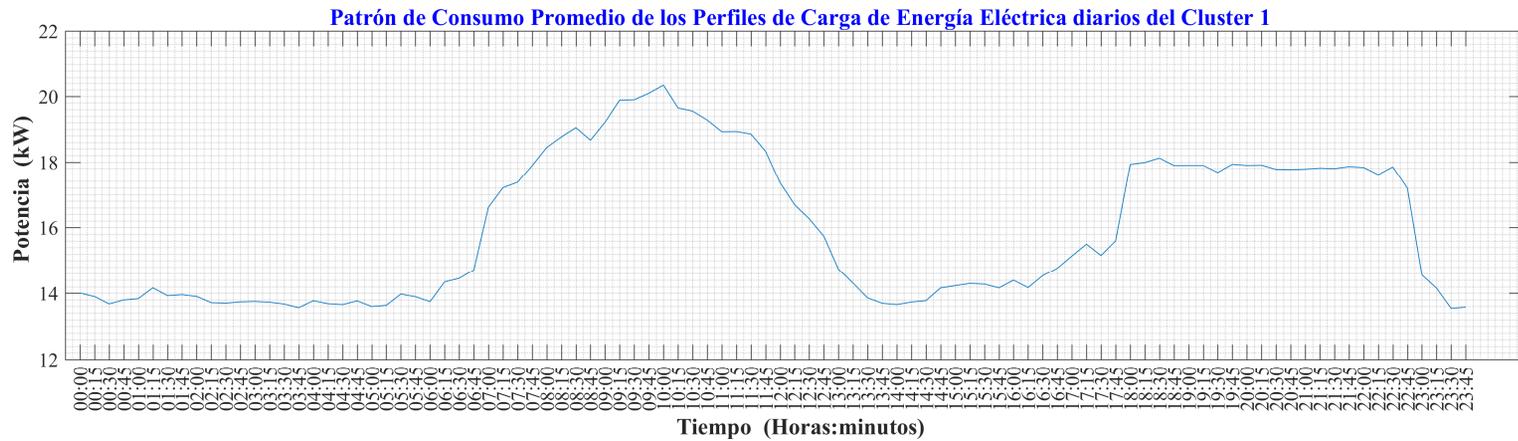
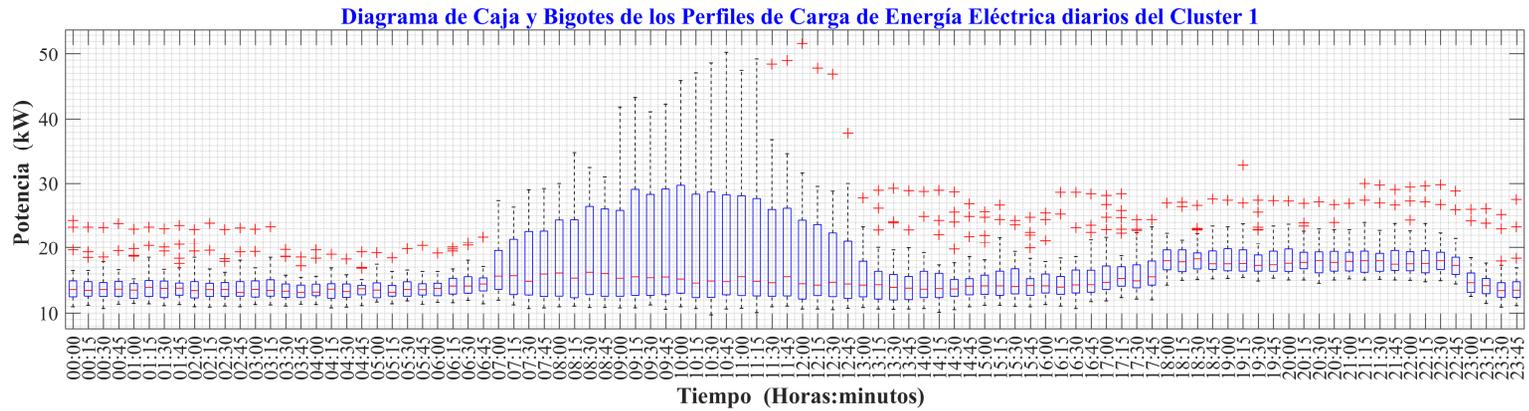


Figura 108. Patrón de consumo de PCEE diarios del cluster 1.

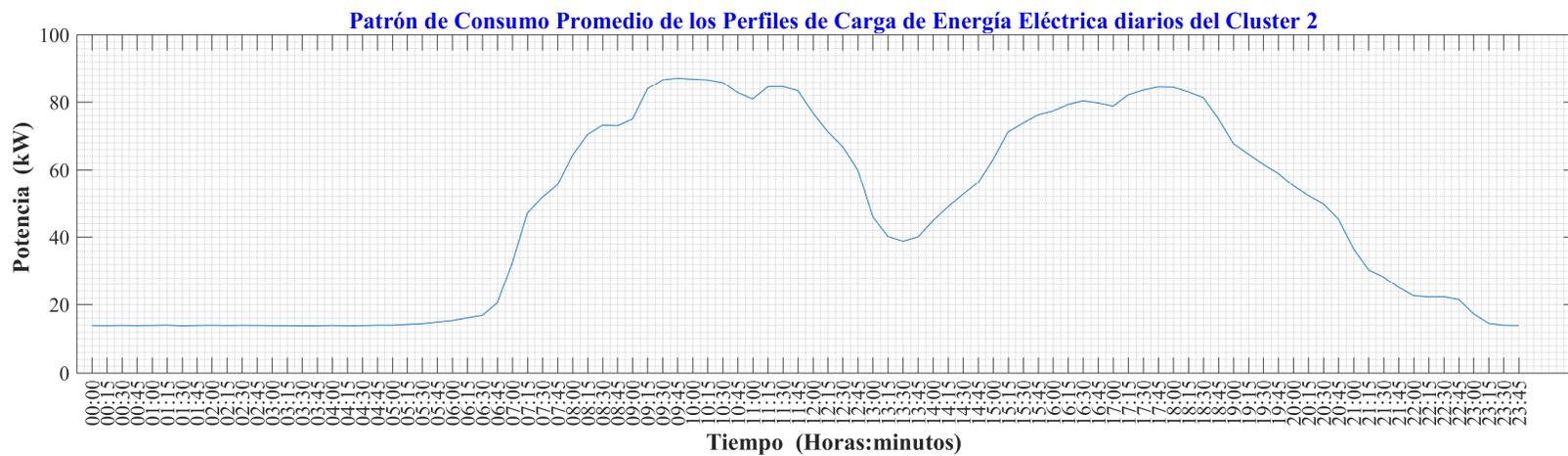
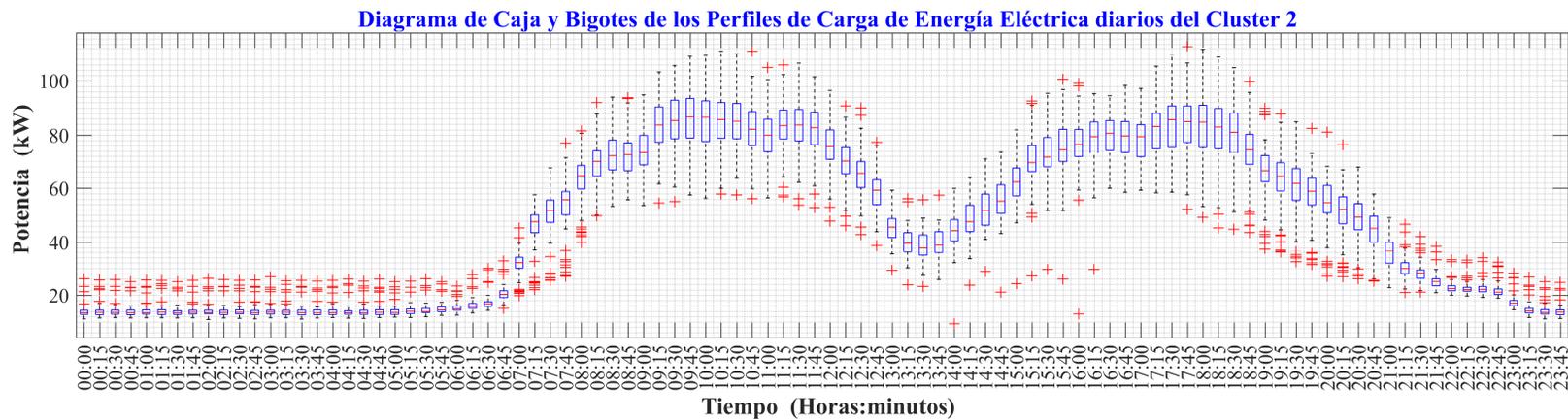


Figura 109. Patrón de consumo de PCEE diarios del cluster 2.

4.1.3. Consumo de Energía Eléctrica de CONTINENTAL TIRE ANDINA S.A.

En el marco del análisis de un cliente industrial como la llantera Continental Tire Andina S.A. se tiene la base de datos de CEE, con los cuáles se procede a realizar la ejecución del programa y observar los resultados del cómo se comporta este tipo de cliente con el CEE

Bajo este contexto se realiza el mismo análisis que en los casos anteriores.

Perfiles de Carga de Energía Eléctrica de los días Lunes

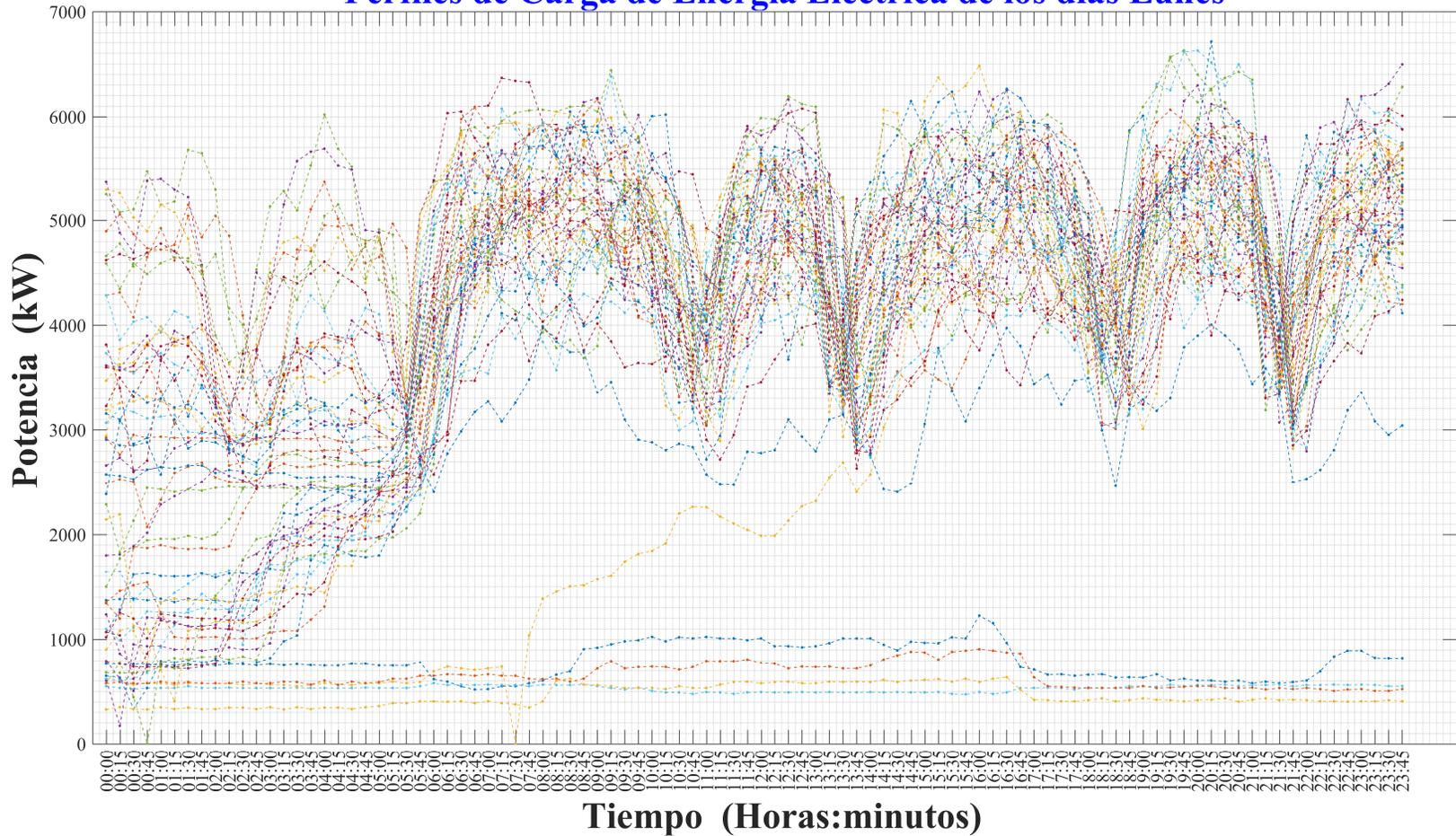


Figura 110. Perfiles de Carga de Energía Eléctrica (Lunes).

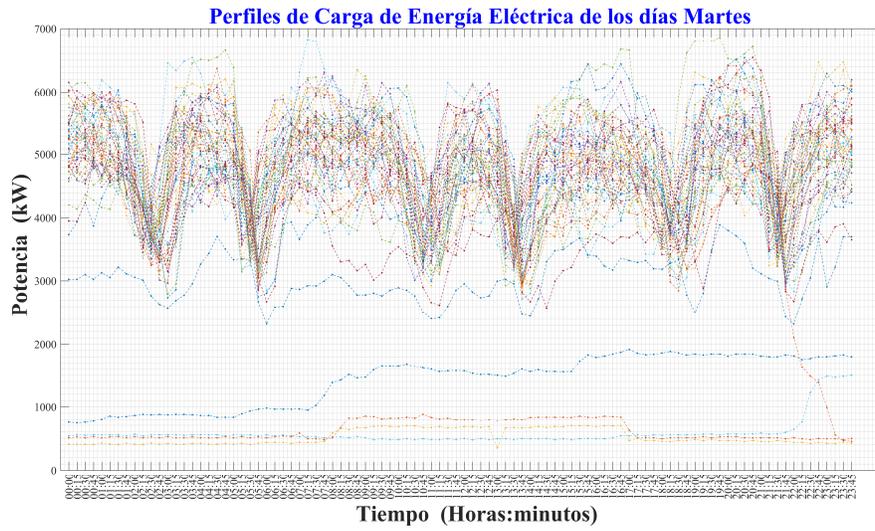


Figura 111. Perfiles de Carga de Energía Eléctrica (Martes).

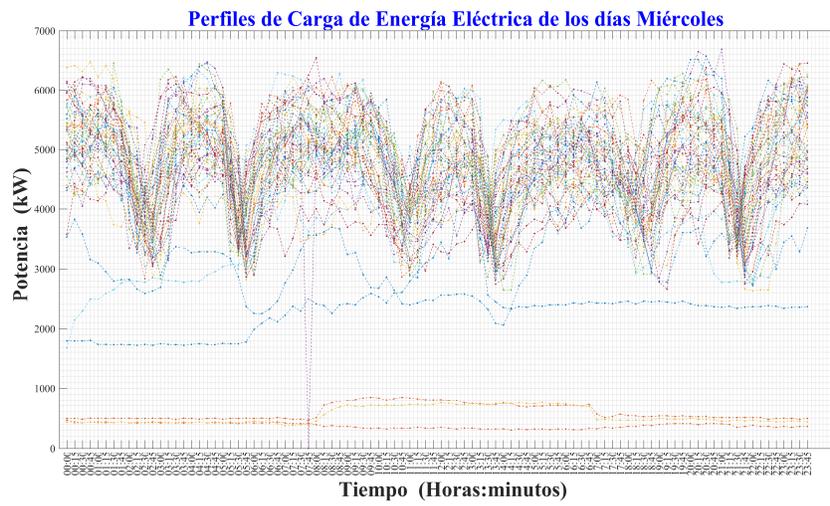


Figura 112. Perfiles de Carga de Energía Eléctrica (Miércoles).

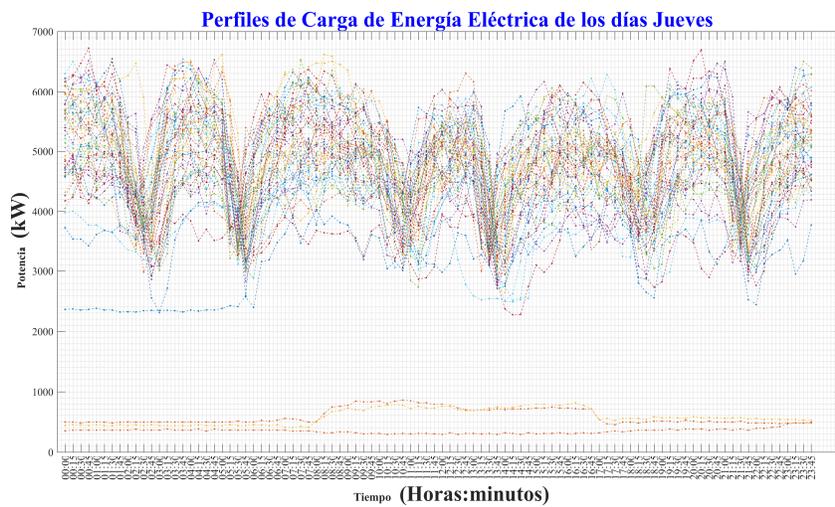


Figura 113. Perfiles de Carga de Energía Eléctrica (Jueves).

Perfiles de Carga de Energía Eléctrica de los días Viernes

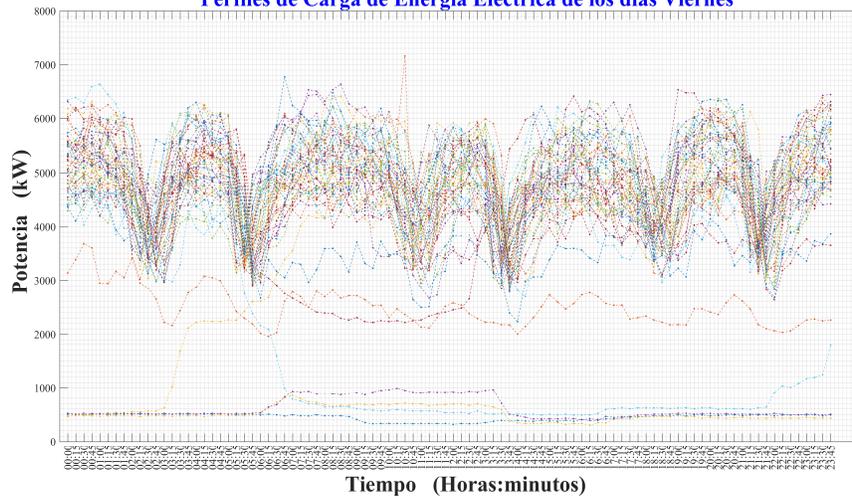


Figura 114. Perfiles de Carga de Energía Eléctrica (Viernes).

Perfiles de Carga de Energía Eléctrica de los días Sábados

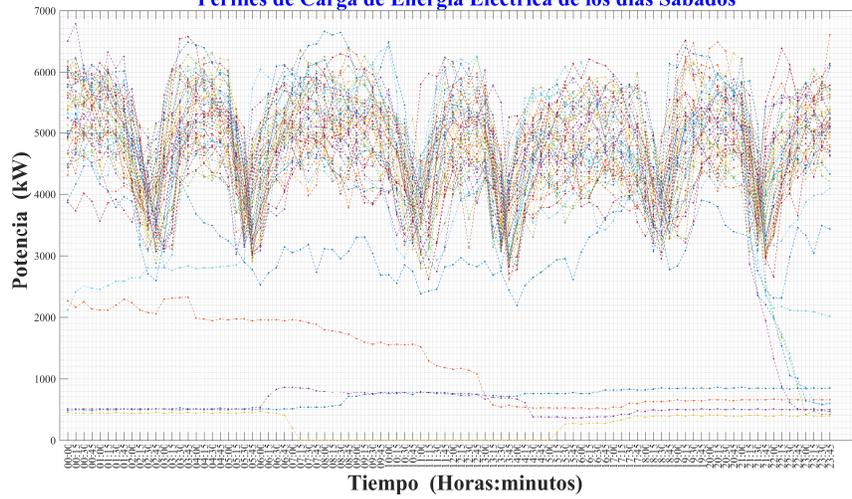


Figura 115. Perfiles de Carga de Energía Eléctrica (Sábados).

Perfiles de Carga de Energía Eléctrica de los días Domingos

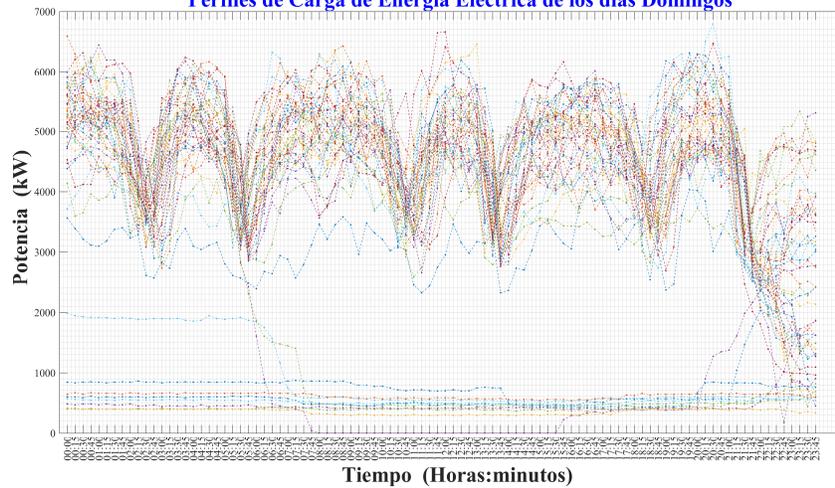


Figura 116. Perfiles de Carga de Energía Eléctrica (Domingos).

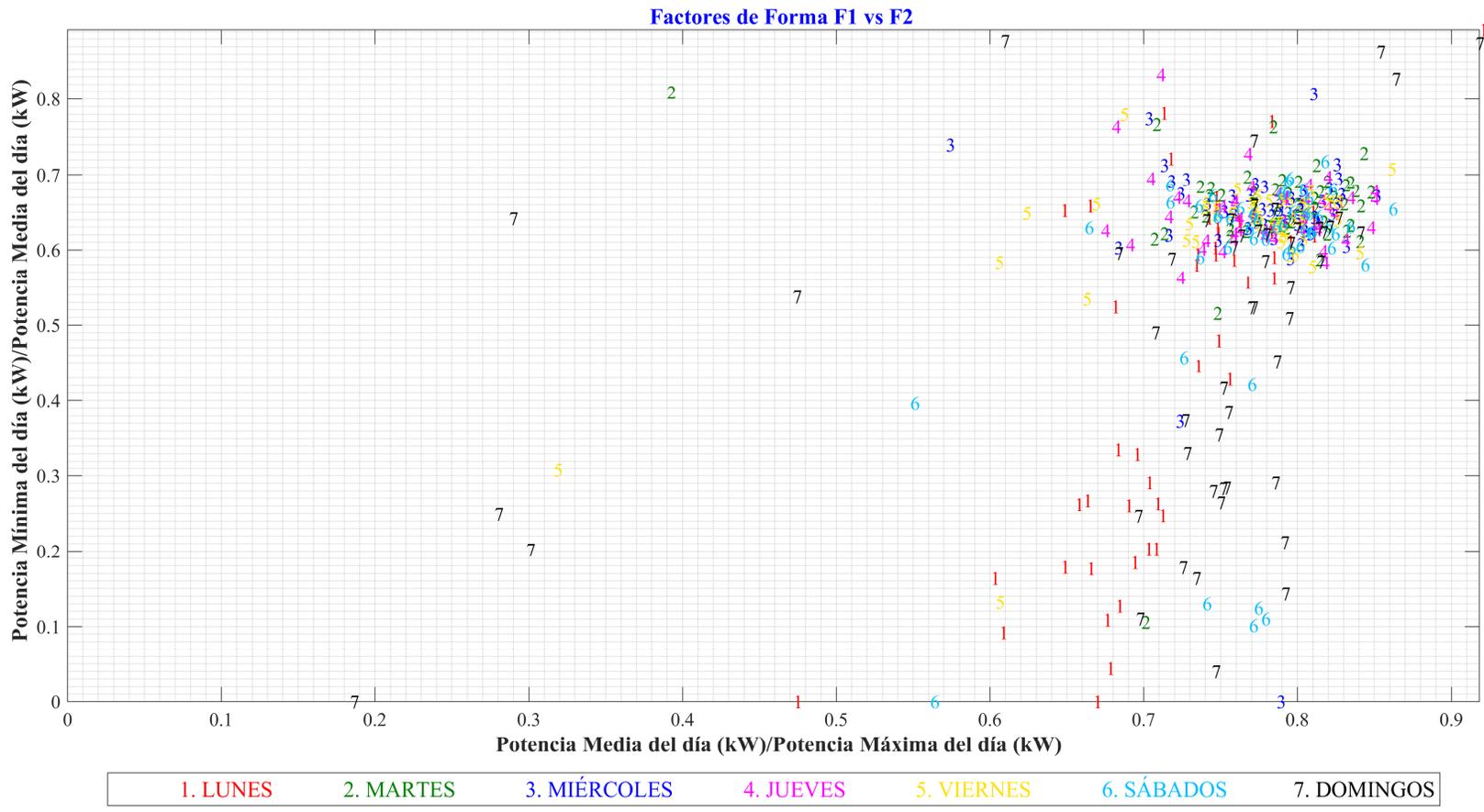


Figura 117. Factores de Forma F1 vs F2.

1. LUNES 2. MARTES 3. MIÉRCOLES 4. JUEVES 5. VIERNES 6. SÁBADOS 7. DOMINGOS

Factores de Forma F3 vs F4 vs F5

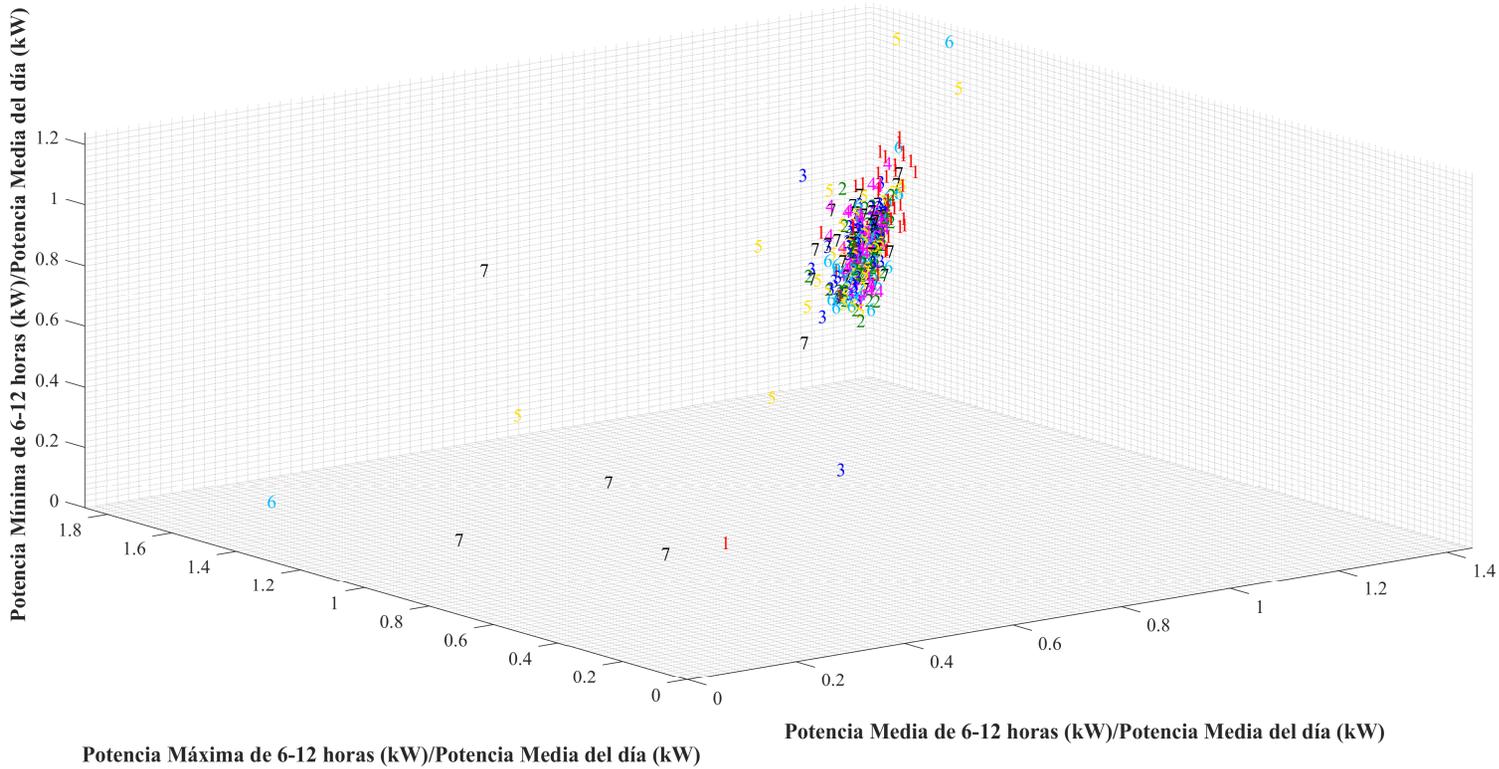


Figura 118. Factores de Forma F3 vs F4 vs F5.

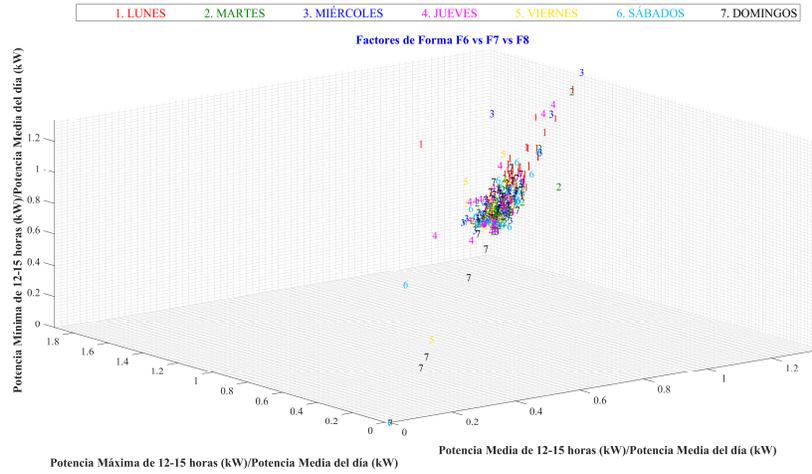


Figura 119. Factores de Forma F6 vs F7 vs F8.

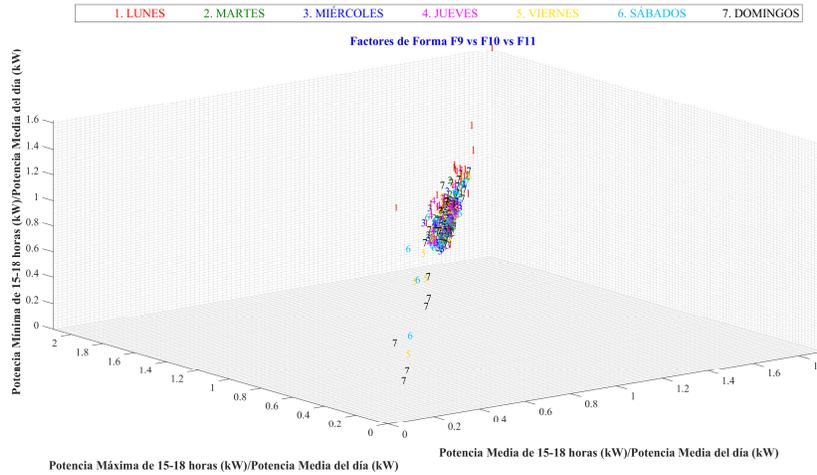


Figura 120. Factores de Forma F9 vs F10 vs F11.

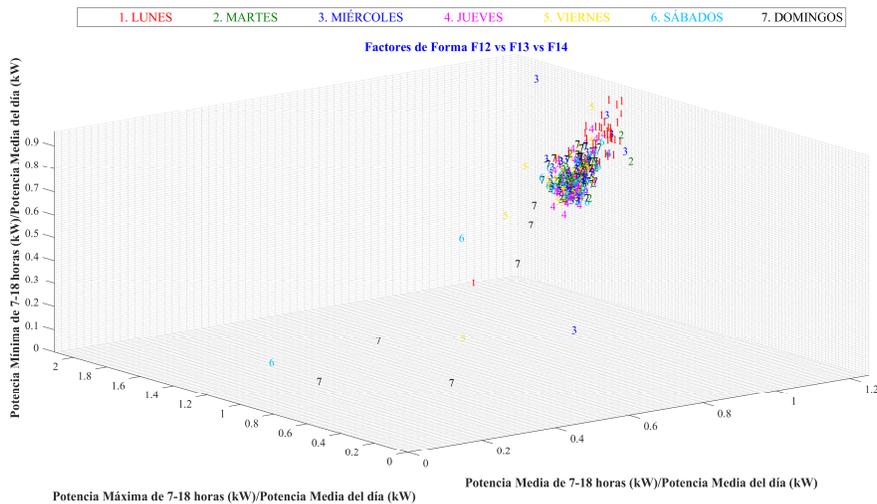


Figura 121. Factores de Forma F12 vs F13 vs F14.

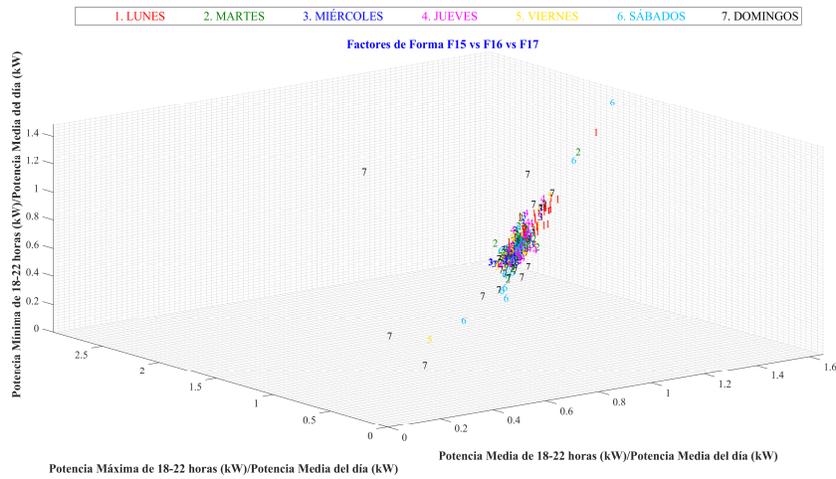


Figura 122. Factores de Forma F15 vs F16 vs F17.

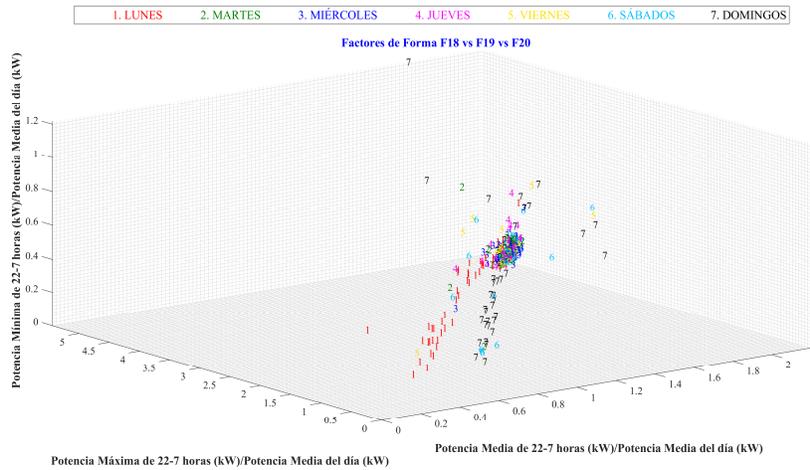


Figura 123. Factores de Forma F18 vs F19 vs F20.

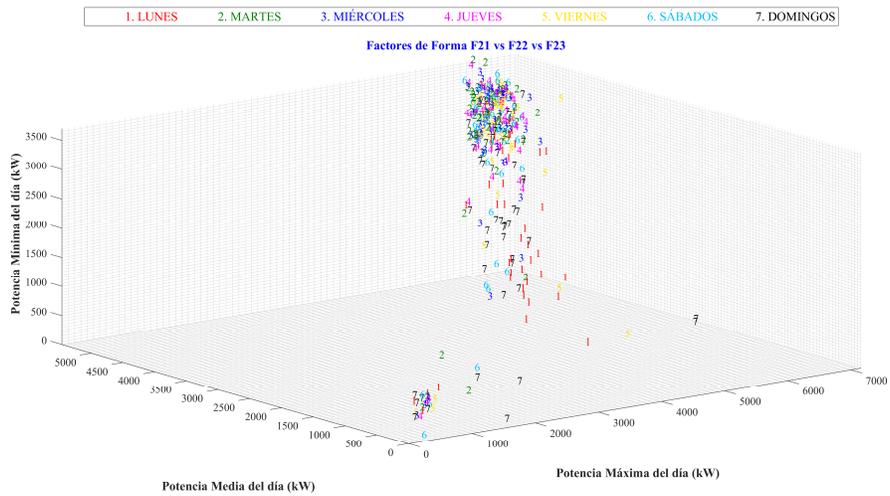


Figura 124. Factores de Forma F21 vs F22 vs F23.

4.1.3.1. Valores promedios de las siluetas según el número de clusters con los que se segmentaron los factores de forma

Tabla 16. Valores promedios de las Siluetas para el número de clusters analizados

Número de cluster	Valor promedio de Silueta
1.0000	-
2.0000	0.9369
3.0000	0.8201
4.0000	0.6795
5.0000	0.6180
6.0000	0.6333
7.0000	0.5954
8.0000	0.5936
9.0000	0.5858

4.1.3.2. Evaluaciones del número de clusters más óptimo para los datos en análisis

El número de Clusters más apropiados según el Valor de la Silueta:

Ekmeans3 =

SilhouetteEvaluation with properties:

NumObservations: 366

InspectedK: [1 2 3 4 5 6 7 8 9]

CriterionValues: [NaN 0.9369 0.8201 0.6795 0.6222 0.6333 0.6302 0.6008 0.5858]

OptimalK: 2

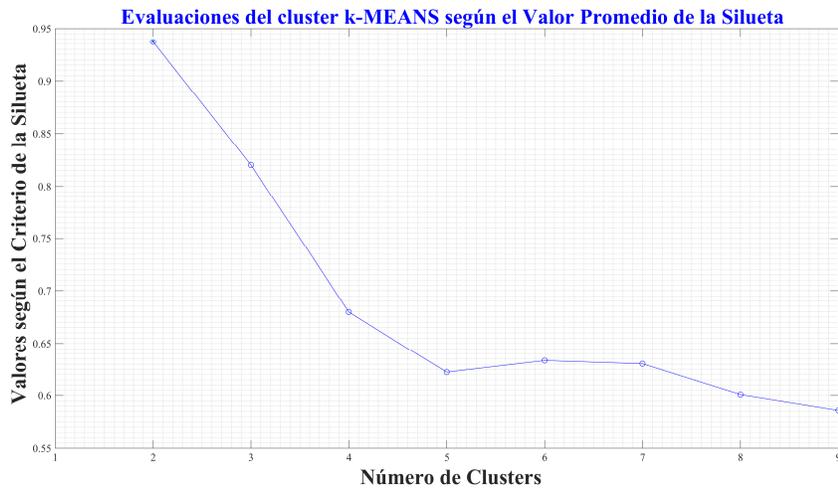


Figura 125. Evaluación según el valor promedio de Silueta del cluster k-Means.

4.1.3.3. Valores de los tiempos de ejecución para cada número de cluster con los que se segmentaron los factores de forma

Tabla 17. Tiempos de Ejecución para el número de clusters analizados

Número de Clusters	Tiempo de Ejecución (s)
[1]	[4.3158]
[2]	[8.9283]
[3]	[8.4715]
[4]	[8.5216]
[5]	[8.3490]
[6]	[8.3437]
[7]	[8.3852]
[8]	[8.4559]
[9]	[8.4680]

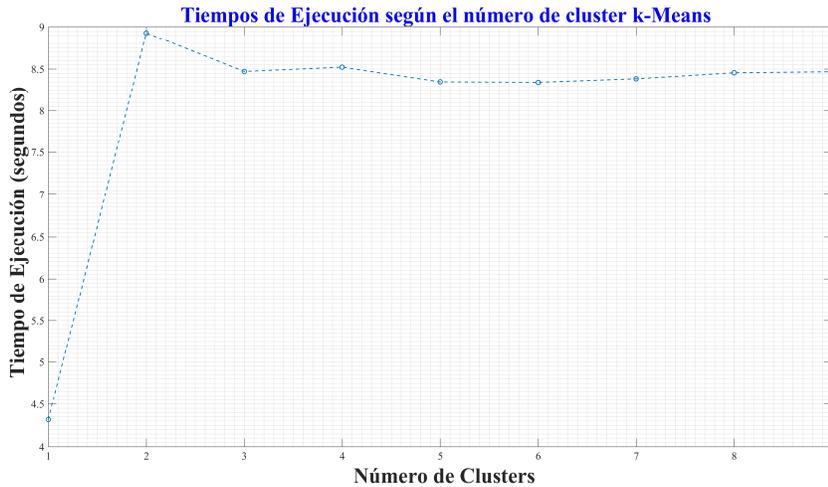


Figura 126. Tiempos de ejecución según el número de cluster k-Means.

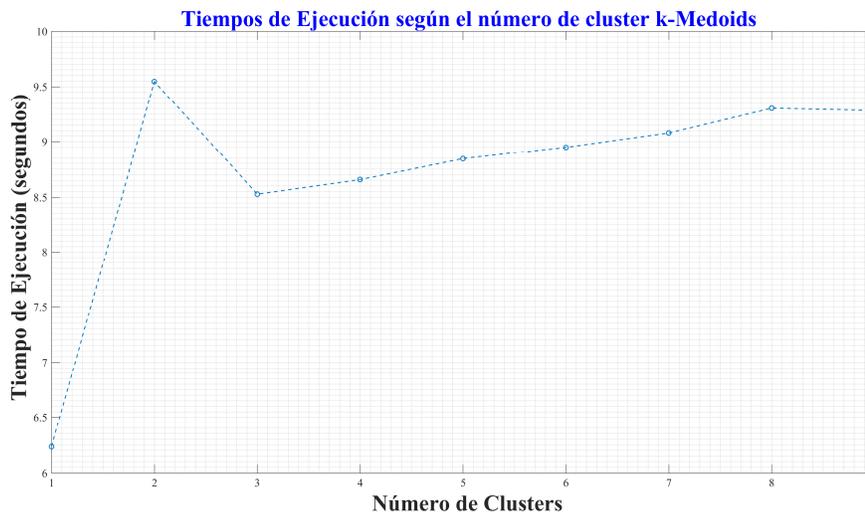


Figura 127. Tiempos de ejecución según el número de cluster k-Medoids.

Gráfica de Silueta con el número de clusters óptimo

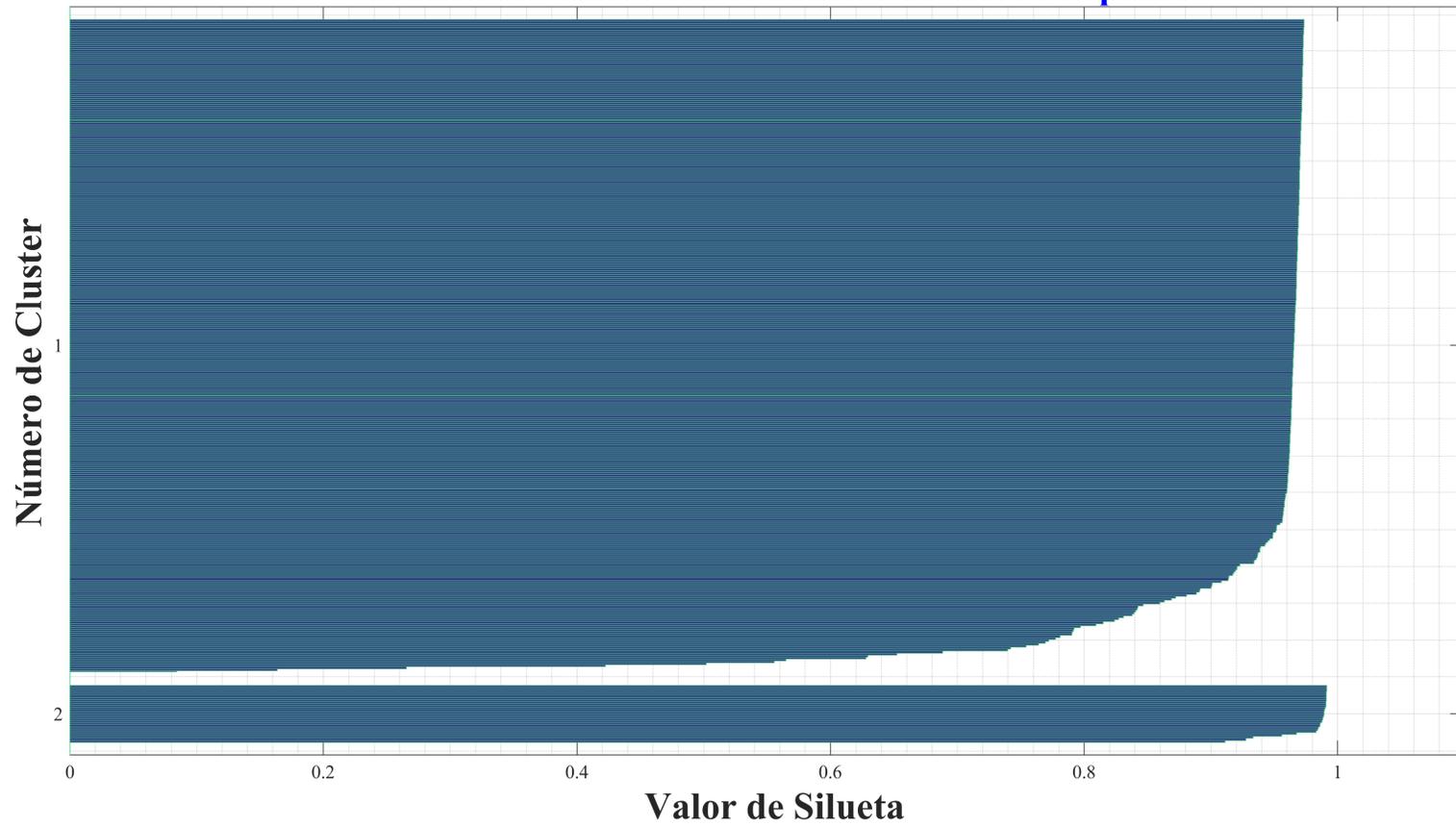


Figura 128. Gráfica de Silueta con el número de clusters óptimo.

4.1.3.4. Indicador del número de cluster a cuál pertenece cada PCEE diario

Tabla 18. Indicador del cluster al que pertenece cada PCEE

1	2	47	1	93	1	139	1	185	1	231	1	277	1	323	1
2	1	48	1	94	1	140	1	186	1	232	1	278	1	324	1
3	1	49	1	95	1	141	1	187	1	233	1	279	1	325	1
4	1	50	1	96	1	142	1	188	1	234	1	280	1	326	1
5	1	51	2	97	1	143	1	189	1	235	1	281	1	327	1
6	2	52	2	98	1	144	1	190	1	236	1	282	1	328	1
7	1	53	2	99	1	145	1	191	1	237	1	283	1	329	1
8	1	54	1	100	1	146	1	192	1	238	1	284	1	330	1
9	1	55	1	101	1	147	1	193	1	239	1	285	1	331	1
10	1	56	1	102	1	148	2	194	1	240	1	286	1	332	1
11	1	57	1	103	2	149	1	195	1	241	1	287	1	333	1
12	1	58	2	104	2	150	1	196	1	242	1	288	1	334	1
13	1	59	1	105	1	151	1	197	1	243	1	289	1	335	1
14	1	60	1	106	1	152	1	198	1	244	1	290	1	336	2
15	1	61	1	107	1	153	1	199	1	245	1	291	1	337	1
16	1	62	1	108	1	154	1	200	2	246	1	292	1	338	1
17	1	63	1	109	1	155	2	201	1	247	1	293	1	339	2
18	1	64	1	110	1	156	2	202	1	248	1	294	1	340	1
19	1	65	1	111	1	157	1	203	1	249	1	295	1	341	1
20	1	66	1	112	1	158	1	204	1	250	1	296	1	342	1
21	1	67	1	113	1	159	1	205	1	251	1	297	1	343	1
22	1	68	1	114	1	160	1	206	1	252	1	298	1	344	1
23	1	69	1	115	1	161	1	207	2	253	1	299	1	345	1
24	1	70	1	116	1	162	1	208	2	254	1	300	1	346	1
25	1	71	1	117	1	163	1	209	2	255	1	301	1	347	2
26	1	72	1	118	1	164	1	210	1	256	1	302	1	348	1
27	1	73	1	119	1	165	1	211	1	257	1	303	1	349	1
28	1	74	1	120	1	166	1	212	1	258	1	304	1	350	1
29	1	75	1	121	1	167	1	213	1	259	1	305	1	351	1
30	1	76	1	122	1	168	1	214	1	260	2	306	1	352	1
31	1	77	1	123	1	169	1	215	1	261	2	307	1	353	1
32	1	78	1	124	1	170	1	216	1	262	2	308	1	354	1
33	1	79	1	125	1	171	1	217	1	263	1	309	1	355	2
34	1	80	1	126	1	172	1	218	1	264	1	310	1	356	1
35	1	81	1	127	1	173	1	219	1	265	1	311	1	357	1
36	1	82	1	128	1	174	1	220	1	266	1	312	2	358	1
37	1	83	1	129	1	175	1	221	1	267	1	313	2	359	1
38	1	84	1	130	1	176	1	222	1	268	1	314	2	360	1
39	1	85	1	131	1	177	1	223	1	269	1	315	2	361	1
40	1	86	1	132	1	178	1	224	1	270	1	316	1	362	1
41	1	87	1	133	1	179	1	225	1	271	1	317	1	363	1
42	1	88	1	134	1	180	1	226	1	272	1	318	1	364	1
43	1	89	1	135	1	181	1	227	1	273	1	319	1	365	2
44	1	90	1	136	1	182	1	228	1	274	1	320	2	366	2
45	1	91	1	137	1	183	1	229	1	275	1	321	1		

46	1	92	1	138	1	184	1	230	1	276	1	322	1		
----	---	----	---	-----	---	-----	---	-----	---	-----	---	-----	---	--	--

4.1.3.5. Ubicación de los centroides para cada cluster

Tabla 19. Centroides para cada cluster

CLUSTER 1	CLUSTER 2
0,764413515517512	0,673368700543320
0,576976525906868	0,663651767034626
1,02779944211405	1,00280233093106
1,20491827102740	1,29329773936127
0,752974956342016	0,775475047589144
0,968298097163986	0,936111366300263
1,15335992192684	1,03168730939481
0,697219420949425	0,830755094062592
1,03967789698596	0,937485230712170
1,14390387613844	1,09266981126678
0,901094408056830	0,757653335918888
1,01435354936333	0,963535278086448
1,24538890237807	1,25194651088829
0,679244214842537	0,675956191462474
0,991528673718925	0,972477849248505
1,18244173654292	1,08999302258003
0,693625009347431	0,913718775204192
0,986221806903074	1,05680006045056
1,26650592976555	1,51811311826291
0,608894550353168	0,827286849742567
5985,38872403561	997,103448275862
4578,75222551929	603,735632183908
2679,00296735905	395,310344827586

4.1.3.6. Distancia de los factores de forma que caracterizan cada PCEE diario hasta el centroide de cada cluster

Tabla 20. Cuadro de distancias de cada Factor de Forma al centroide de cada cluster

CLUSTER 1	CLUSTER 2
41700260,1137947	100929,897529305
475450,384670555	39559119,3526793
104870,217140129	43310152,8619588
117608,453408918	50407827,4183712
493547,116946329	54200896,3206653
50533517,1775933	188807,822870913
12413,5386066471	45314967,8774680
16680,4753342387	46438536,4621105
319077,788342556	47797671,8611606
49823,2742797800	45628582,3156652
405047,826209307	54888882,9852088
1053386,23751077	60506230,4426597
896174,598212704	59193793,1549138

523902,186990884	55749318,7308650
607479,425576897	52187107,4968764
39164,1954118125	48156981,8399325
434836,485800719	54890368,6901279
43576,3680244957	47087113,4186637
1735085,81939253	46925033,1092290
11205,0272486500	45390987,5120166
218216,924623782	50771488,8577442
152490,863969053	41148774,0098541
2437413,58862575	36562896,7349879
247318,295443253	50789692,0803031
1084489,05025220	35865073,6473282
1115858,28040426	35319894,9799955
3895757,84967847	30116691,8650385
4232896,53296227	33824263,4632120
2827539,50446576	37289535,4596413
145913,770300902	41217579,4368583
11904364,3501256	22443466,5652603
5157199,88071157	34293853,3896322
4656861,15125941	33850706,0974765
3737230,79705887	31338864,8799621
3657403,66180296	30443649,7886453
1994876,37014411	37335247,1039770
3339838,64175127	34662013,6019745
4529256,85452918	29966393,9056103
6787015,30584507	33765720,0043794
131074,040589912	48937576,6657235
5128903,71298392	35939839,2564748
5856358,90580741	26979066,3685650
6459314,89713195	33518974,1749177
259789,953877630	41997909,4580759
6175777,27829697	29890118,9884897
4976394,89415609	28095434,8897530
8206563,46992197	30035230,8889098
315120,042469347	42595739,8925534
2175145,67695773	28418500,3435065
6113708,88382145	19447440,1867141
45984921,4795625	21456,5036615301
51151600,9697379	157161,497623797
30264263,5186972	1639372,29022335
403331,222350068	53414228,1966915
554263,950520485	51014843,0177934
446222,471982243	51112066,5480605
447790,388929238	50990849,6155353
40990179,6863022	255203,728890104
968294,138506424	59425349,8251800
1022146,67706771	59255113,9211492
636175,801717765	51840252,4769169
548236,278376490	55701728,2449470

342125,011846579	52866605,5279827
1415782,33654163	63219500,2564177
1897768,32099772	63934834,0781011
732316,370389404	54725967,2726021
890514,893569138	58948928,6440104
747667,939615111	57434616,9312726
2064480,81978069	64503833,9154757
710650,438575862	56326654,8224573
1148590,58978998	61140375,1244597
874816,117431432	56704289,8788152
747512,824277547	56644443,9734598
422878,265983694	51234050,3417611
424777,814876279	47737798,3704439
243438,143302388	49377983,9667057
181401,687295948	49158486,6542670
246172,645370581	43315218,1486551
236522,421713087	43480632,5464635
487417,353138805	48290130,3401333
868406,006658630	51014658,9820987
238829,316447857	48218255,5375119
59760,0426553442	46629724,0332382
137777,152984950	50719909,2317909
226505,988706006	45104118,8704686
829173,212463128	58143509,2804308
265848,386765823	41970683,5418804
152216,000958918	49199936,0256976
69405,4181531810	48672846,8920807
217946,050265048	42638951,3268812
302565,548765889	46328594,6224834
154834,889378197	49912073,9819032
244970,399718444	46272973,4977037
479154,481683568	40109060,8288962
77740,6745472152	46880565,8241627
5101433,96126866	38865325,1679429
176536,130544864	43225269,0012085
324735,314139632	41118052,2397786
87254,5926601922	43397510,5710928
562831,056327905	37173538,9602042
1410361,34883829	31803824,7585118
6899786,29897270	17990602,6999330
46582019,8092214	20523,2869499050
49948324,5736901	102785,527369699
9535363,97093314	14329722,8052737
207573,785391686	49385238,4988875
115565,575570220	46223897,2491450
7188306,36886318	41267278,9046063
884123,733494844	55109867,6631388
1069572,05784834	44903982,1392476
712595,389077666	55357499,1132249

994526,220563692	59366753,0807318
701923,993528944	54284009,2847925
786755,562930630	56275828,8639831
1176164,63372695	61318970,5858742
470842,156174776	52973563,9211645
1357410,41055920	59655385,8406837
738913,358393751	57302058,8897016
894210,684770947	59127364,2124586
842492,146522486	56586487,9332493
833772,649186588	58298063,1180156
525642,164858088	54108309,1438823
949454,746405841	58859554,0926108
1110090,22220422	58385256,5669305
841967,106729855	58269564,7618931
215962,983154448	47906883,4779736
236836,677271593	49639123,9795576
527878,246230148	44855227,7823724
396058,266850038	53188237,4831437
107523,140658879	44378532,5658974
238519,370893767	40323619,2300706
712914,149869806	55577419,9894703
643007,749979479	56125123,9885759
143001,946492761	50474713,3782576
537159,370943495	54805416,2596802
49007,8751563701	47498057,4266329
219911,099480774	51592263,1199723
154090,146749144	43908868,1386768
173269,846974909	50386624,1823659
128275,930571680	43943680,8718558
89600,7580635107	49503402,2340039
157209,523632024	45867253,3149658
199650,663457047	45438924,4589850
80514,3113531110	47489628,8766869
76637,5087713831	49305783,6333788
646157,807298031	36728813,6710517
118692,163521492	43697797,5578742
53889145,0302111	351420,435331776
777024,028619153	35460269,0055445
266561,799383890	41277180,4739757
253563,806533844	49181454,3078454
729807,547022187	36474216,9096344
1023640,47038442	34090356,3774152
5639040,67802237	19470354,6876441
47193992,0489122	26509,9787439982
49039310,1205055	64958,0988579065
906100,495001625	34691599,0800482
471566,229293459	54212599,1383191
440177,557387090	52157924,6690966
888789,603761798	57871403,4368524

762507,136346460	58108285,8190008
1108199,15413308	60702852,2870910
984243,667720344	60077807,7081457
963419,183848625	58227252,7010323
630764,581976085	56997536,1098699
1447366,65074490	62815230,4123582
729060,338003824	57477923,0759495
295295,188507156	50641244,8801208
700443,292457003	57627369,6281416
1770330,67605679	62408410,1390937
1032047,57193724	59656385,9414630
674835,828225615	55707705,4903288
1059291,93755276	60432085,6046383
1373403,99787745	62105403,7991710
899059,123945584	58874059,1942762
1012022,54064019	56702499,1235416
958250,909566315	60001363,9216100
873550,285216323	57115767,1027367
141952,795697032	49012283,7421548
407832,672174320	46445767,7101879
294413,813070065	51897616,8102296
321005,205242459	47549272,9690604
252496,081385774	39559179,7535603
262546,707081420	51186821,5010648
236915,009346761	51313824,6457816
315697,746233856	49103207,9815510
341948,783674095	54054352,4349868
284987,643206130	39898400,6426402
167300,163619126	50394472,3723462
346395,368668861	46189664,7952341
204621,323682506	42279293,1171718
130638,484197945	50481329,6486923
112400,246253451	47483912,2340966
337298,801761792	50530542,3406962
153338,482653776	41691821,8936384
121714,918614607	48010594,7945908
84997,3065933516	46677350,8673313
166280,042051971	51167671,8960269
436740,355910167	38865055,5761917
53856486,4522453	337258,714510748
173214,502812894	41512902,9830265
486626,084806922	40936845,4656597
361323,356902575	38895937,0328613
1476734,25390450	31190848,0234071
1407909,01890968	32067503,7133206
4115706,96177516	22917080,8362091
47141414,0006595	20779,3732122167
47900790,7863214	32183,5873173828
52530454,6892643	258667,466496877

145890,226403597	47759842,4775787
374466,561732113	52529529,6002068
234035,597585295	48649084,4547726
535998,073803604	52829176,3406008
813348,255740788	56298096,9527068
1008741,35671328	58185260,6802497
509110,266906715	55361623,2861690
565093,496971786	53528950,4209571
420434,137331068	54259579,3032276
553386,698346031	54927437,2489606
684757,211779140	56708479,9118581
14713037,0908526	17090056,6067974
488803,630769666	43768593,0515907
1558102,59555447	64058367,4617842
551891,010002882	55379730,3210798
534305,761734506	55698039,0108037
1061057,31635166	60654507,2869121
922535,114717656	58084829,7681027
1201061,40554032	61343361,6347573
1161623,86008727	60285212,7879289
297710,227948831	53185578,4314148
441022,040778449	53262449,6887280
295555,083282260	51773305,1044499
708546,213419946	56427086,3919925
70650,6058444657	47777262,1779002
260264,459982442	49572513,2312664
113790,697752426	43142386,9693409
470169,264242091	54735929,7646596
1630124,72486443	62877995,4012301
403282,462204144	52736058,8882579
82256,9088159542	47829419,0060663
72529,7199861433	47856076,7410333
69395,0526006078	46624730,5797597
41246,0770665521	47680582,1291796
63648,2363840975	47997481,1914360
180043,400877490	50568307,6677382
244454,103085002	51896517,5970741
369454,798603720	53571935,0970060
163693,646286736	42669521,5763498
191729,907734455	47513475,7100927
161873,910857812	49596498,6129494
86082,4518511602	43808940,0136970
92246,4559480383	44292564,5749931
5829041,68505865	33445372,4945948
50692,7377631148	43853343,7227945
257333,154406811	41364277,0131436
139128,695782071	41866155,3503657
730217,083504702	35531877,1718218
2001897,09005914	28819362,6243328

10133311,2323169	13251161,9609597
48533208,0949830	38184,9605350823
45829540,9240208	204,756594683084
46247370,0518853	30744,1422727935
278080,617771803	46439912,1945097
268794,961505398	52000334,9535636
264908,676106657	49161083,4970658
572005,211199882	54481242,4976951
545728,901190833	47572024,8151898
949061,497190640	59775902,4534051
770923,698133461	57680105,9712290
554968,516419969	55881909,5494037
714179,567550071	55698554,0509349
264139,064199067	51621451,5028837
794554,183204715	57529565,7406384
494234,747700490	48502441,4060485
580798,527553916	55728613,3767149
665378,141642047	56752769,9301601
1492431,58876241	62381361,9159353
903608,295770999	57211076,2193723
1502065,02314127	63570617,2661681
785139,535010612	55146596,6216365
1169161,66432718	58157542,0624474
1169615,78835940	60132940,8504758
4424732,02200532	41145842,2624679
506712,857733192	55384712,4180930
276195,609665539	50688075,3398872
5466448,54265264	31869578,5556035
178043,438747901	49342532,4390091
179838,885241744	44506587,3706798
507753,311185258	46325888,4771039
64928,2702125856	49209288,2391991
48480,7309676810	48239901,6714342
146692,433190774	46541499,6811218
218729,332051701	51105143,3138729
5754562,64968412	31172887,7476400
149767,187446143	50311145,7256374
97282,1154812330	45916431,5460813
575948,069782080	56444033,6064380
51919,3871157971	46415865,4996583
43437,8827427386	46809415,2914986
185884,164314313	47965395,7726676
137093,146438368	46438120,9505921
4553160,51715407	37526820,0992621
115104,255704723	48386265,5173495
40496,1791199827	46741316,3764573
206443,311820156	40288015,3355230
326895,634755964	40393542,9399495
67126,4371207229	43177677,5222141

99394,9099507138	42928963,1732299
130262,255895982	43675715,3280886
792939,567753388	35109348,2844710
4491638,34347462	21908300,5384957
29102254,8074937	2208058,84233167
56310819,9707729	558344,358202769
47663133,0123574	19992,5935223408
46182031,7203332	26726,0119240285
281835,623831953	39328399,3612631
101036,568871304	48298235,7007605
121118,305642091	44215649,8537175
267834,055352802	49078601,6691663
33966352,3213203	1105285,96108724
398330,530584809	54766047,8098704
182333,852757899	50522252,6467230
405676,077230817	54697219,7436975
55185,1155742420	47830886,5241010
150934,803991446	49374299,8499362
548075,348857977	54748938,4178121
1128223,52169152	61007604,6927403
1034816,28621929	60419038,3207438
239053,978868527	52725015,5367636
865099,643173812	52643548,3912889
363988,258054492	53481499,0404883
12515184,3630991	31316957,0038731
13213942,4520211	28721127,9727631
311133,548685072	52341729,9593300
542720,095965159	52340713,4962182
32902391,9768708	2000740,06268553
1473904,19376922	43040040,5857191
210481,062766508	46280856,6846084
42148455,1725743	969541,360024108
899595,787276232	35012010,6944428
1213042,37257611	39715446,5380266
2993186,18659008	38540843,6735695
1627278,21593033	36002723,6878598
131242,029165603	46562406,6687511
4393677,95344265	35384357,3036863
6407776,87698895	36534694,7932190
50943794,9230870	141791,260003141
2008950,55736364	39951522,6246917
2001575,82960069	40545430,3442889
2162114,85495465	43865698,6942427
2126812,75619338	36503207,7638071
430875,096308813	42238756,8541206
3870226,07452515	39713963,9589472
315856,179367638	40039921,1787624
50475100,5278946	164916,552057517
2773830,41923144	40180361,0585537

3893954,74259105	36371840,2650920
396651,623397215	39076619,6194267
2379768,32258789	36390561,0886435
1675199,52444433	33785394,1818332
5615469,95150239	32443119,1934292
933303,845771180	33996173,4042030
545153,643718820	36812741,7492894
6033598,19931979	19303999,6008660
48692680,8697945	129661,718379360
54768060,2623008	428688,318471294

Asignación de 2 Clusters y 2 Centroides

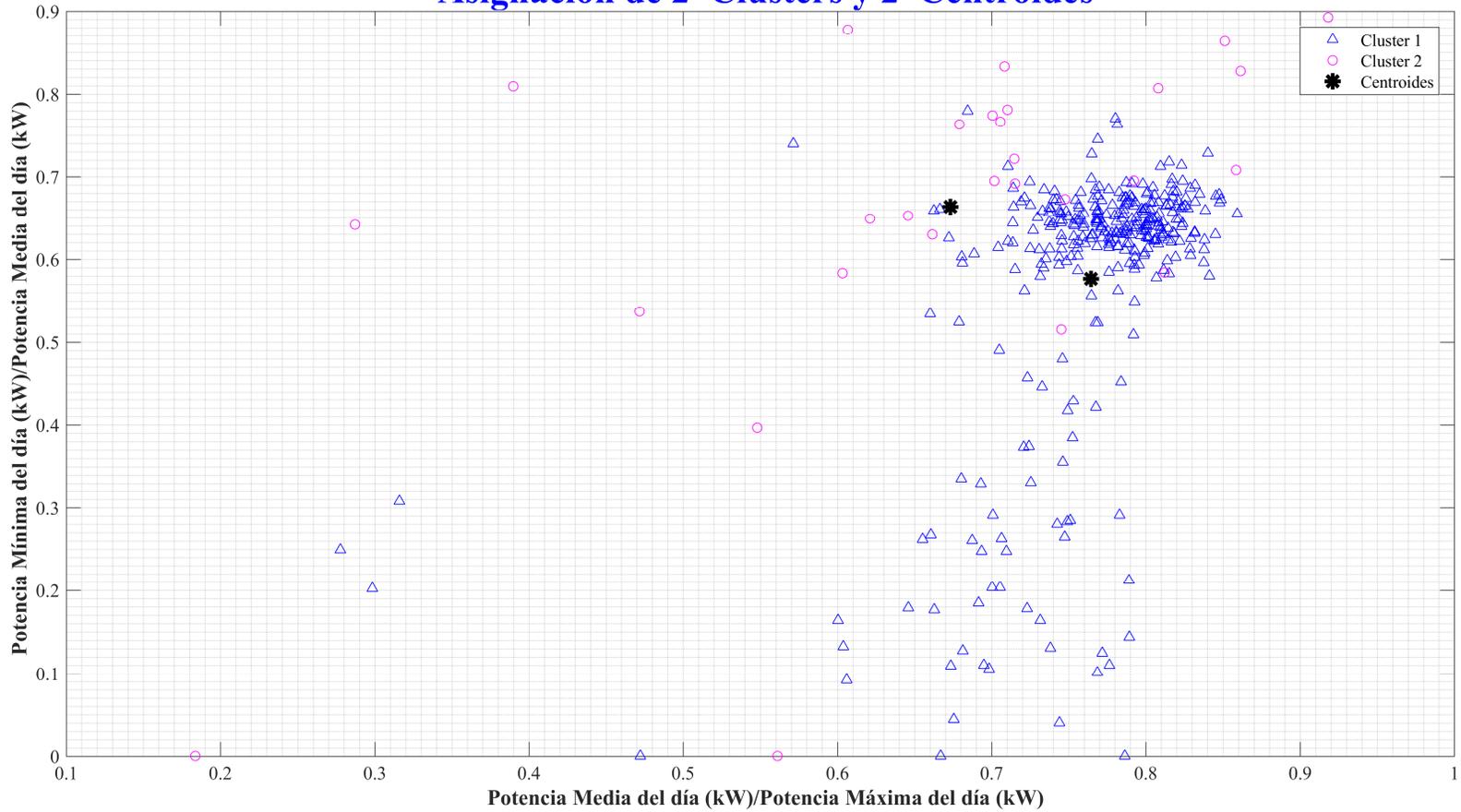


Figura 129. Segmentación de los datos en dos clusters.

Asignación de 2 Clusters y 2 Centroides

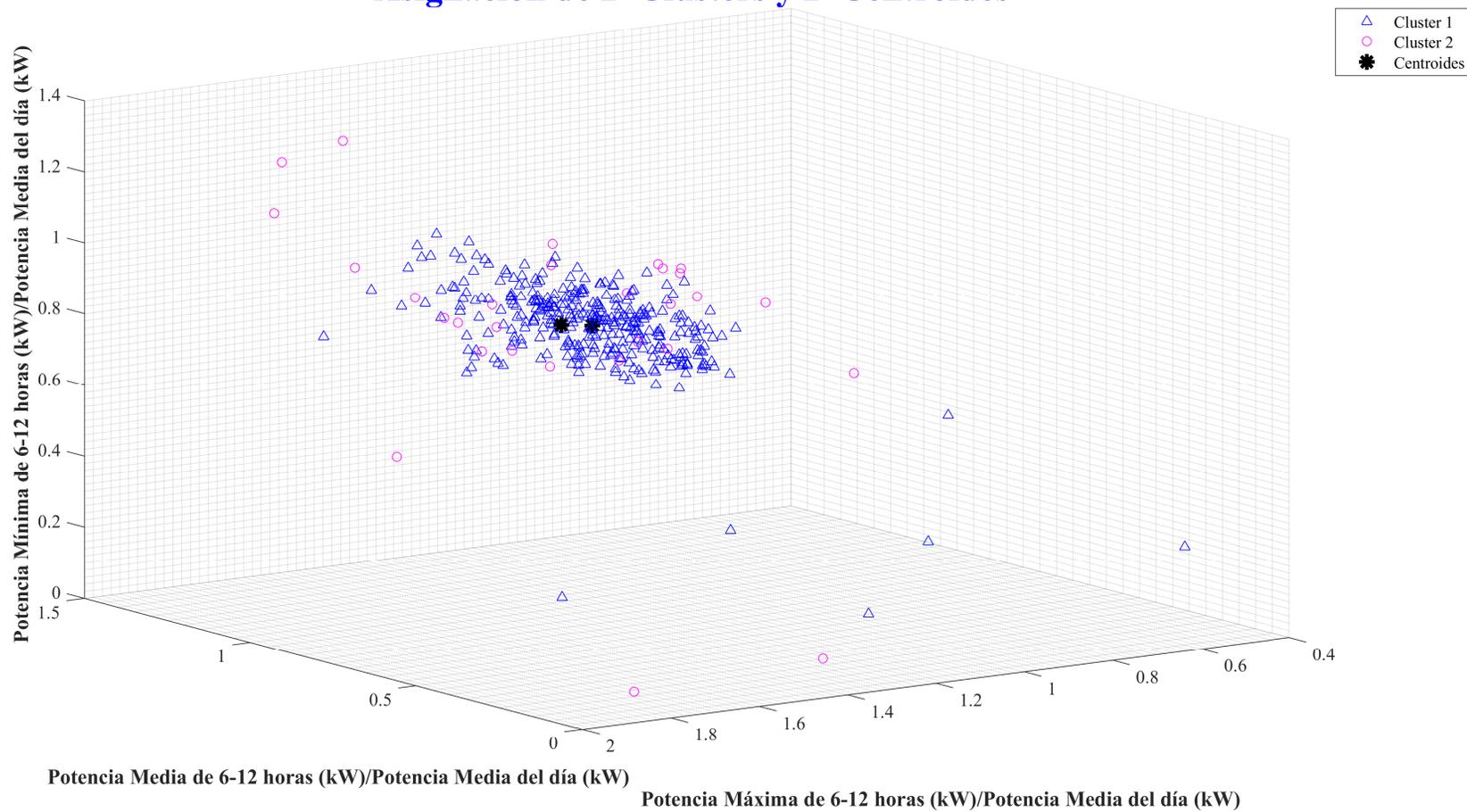


Figura 130. Segmentación de los datos en dos clusters.

Asignación de 2 Clusters y 2 Centroides

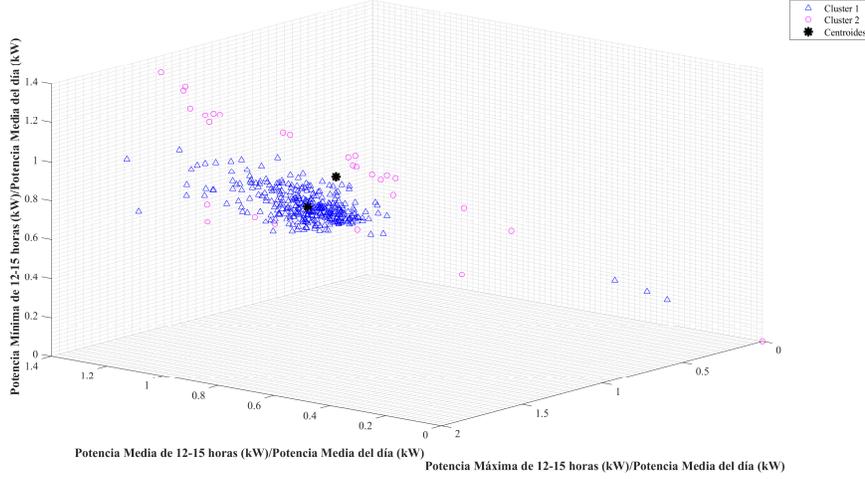


Figura 131. Segmentación de los datos en dos clusters.

Asignación de 2 Clusters y 2 Centroides

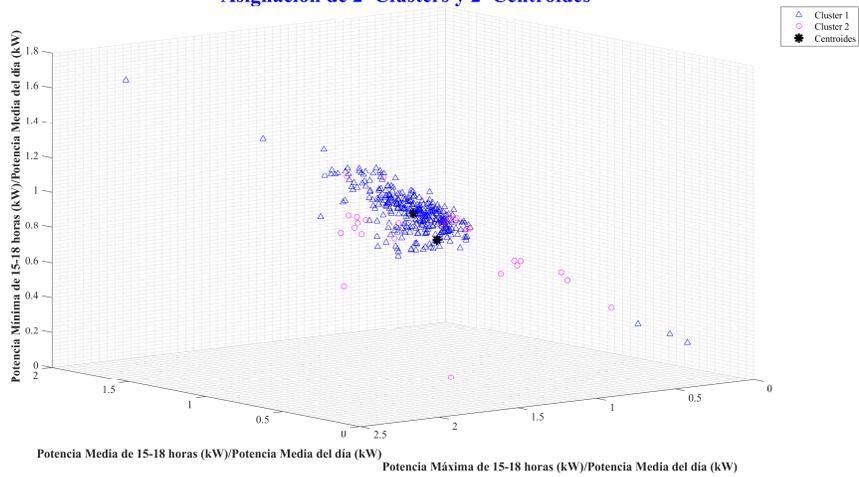


Figura 132. Segmentación de los datos en dos clusters.

Asignación de 2 Clusters y 2 Centroides

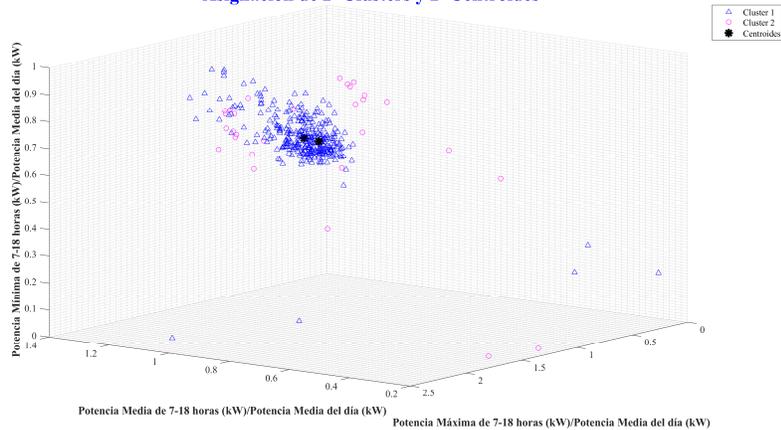


Figura 133. Segmentación de los datos en dos clusters.

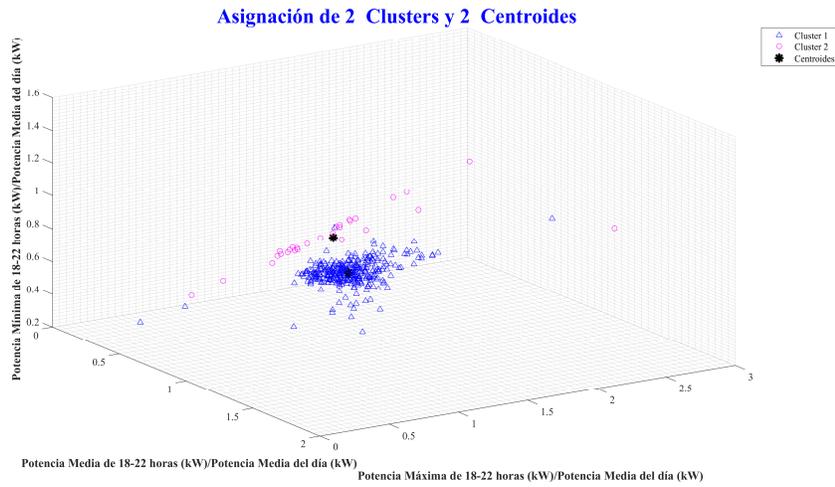


Figura 134. Segmentación de los datos en dos clusters.

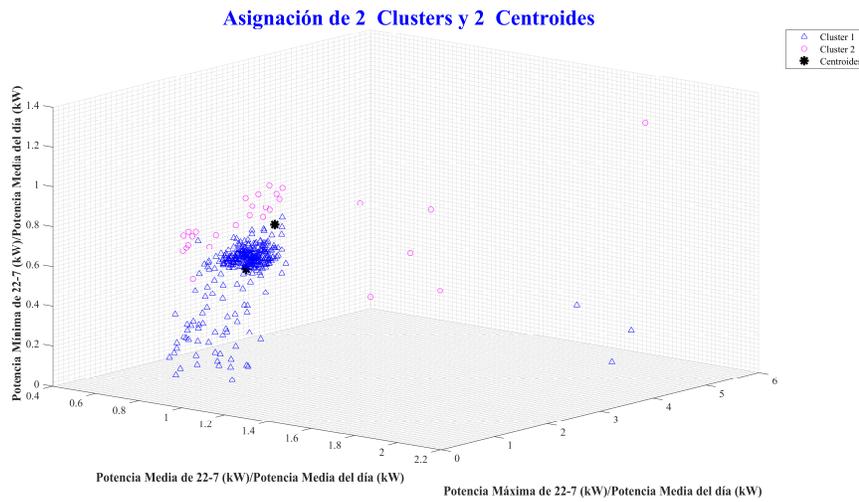


Figura 135. Segmentación de los datos en dos clusters.

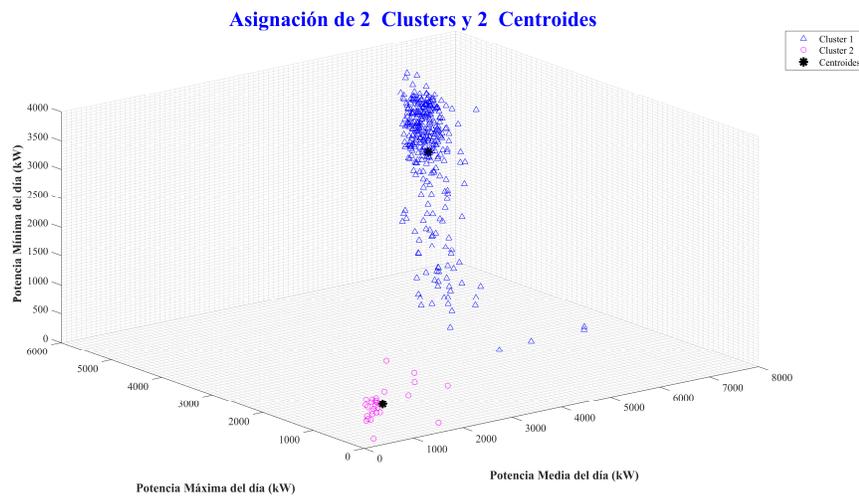


Figura 136. Segmentación de los datos en dos clusters.

4.1.3.7. Análisis estadístico de media aritmética y desviación estándar de las distancias de cada perfil de carga de energía eléctrica a su centroide y de los valores de silueta para la clasificación de dichos perfiles de carga con outliers y sin outliers

- **Análisis estadístico de media aritmética y desviación estándar para el cluster 1:**

MediaAritméticaDCDelCluster1 =

1.2466e+06

DesviaciónEstándarDCDelCluster1 =

2.1089e+06

RangoIntercuartilDCDelCluster1 =

8.4027e+05

ValorMínimoDCDelCluster1 =

6.2707e+05

AnchoDeBandaDCDelCluster1 =

1.7621e+05

limiteatipicos1 =

3.9861e+06

LimitesDeLosValoresAtipicosParaElCluster1medianteLaDistanciaAIC =

3.9861e+06

MediaAritméticaVSDelCluster1 =

0.9332

DesviaciónEstándarVSDelCluster1 =

0.1033

RangoIntercuartilVSDelCluster1 =

0.0125

ValorMínimoVSDelCluster1 =

0.0093

AnchoDeBandaVSDelCluster1 =

0.0026

limiteatipicos2 =

0.8444

LimitesDeLosValoresAtipicosParaElCluster1medianteLosValoresDeSi =

0.8444

- **Análisis estadístico de media aritmética y desviación estándar para el cluster 2:**

MediaAritméticaDCDelCluster2 =

3.9313e+05

DesviaciónEstándarDCDelCluster2 =

5.9499e+05

RangoIntercuartilDCDelCluster2 =

3.4100e+05

ValorMínimoDCDelCluster2 =

2.5448e+05

AnchoDeBandaDCDelCluster2 =

1.1679e+05

limiteatipicos1 =

1.4805e+06

LimitesDeLosValoresAtipicosParaElCluster2medianteLaDistanciaAIC =

1.4805e+06

MediaAritméticaVSDelCluster2 =

0.9804

DesviaciónEstándarVSDelCluster2 =

0.0208

RangoIntercuartilVSDelCluster2 =
0.0063

ValorMínimoVSDelCluster2 =
0.0047

AnchoDeBandaVSDelCluster2 =
0.0022

limiteatipicos2 =
0.9361

LimitesDeLosValoresAtipicosParaElCluster2medianteLosValoresDeSi =
0.9361

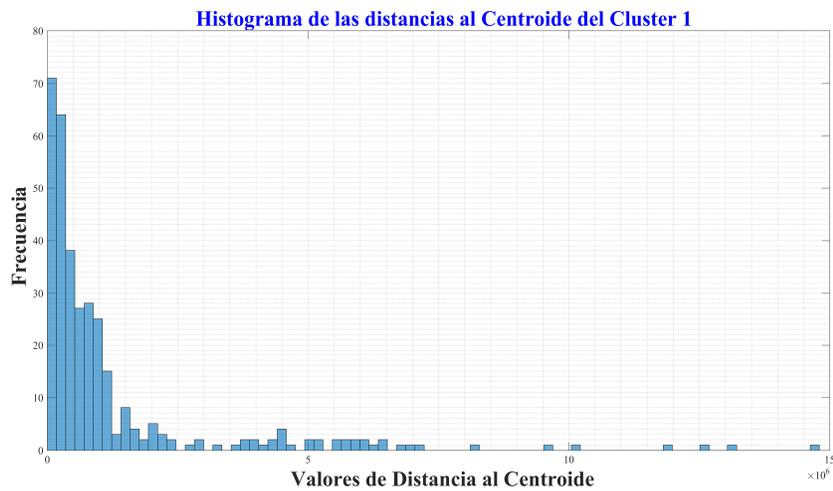


Figura 137. Histograma con las distancias al centroide del cluster 1.

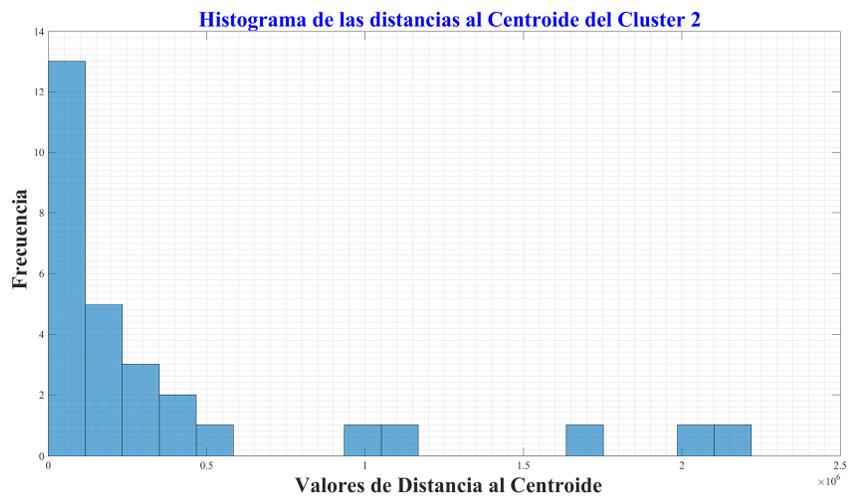


Figura 138. Histograma con las distancias al centroide del cluster 2.

Estimación de densidad por kernel de Epanechnikov de las distancias al Centroide del Cluster 1

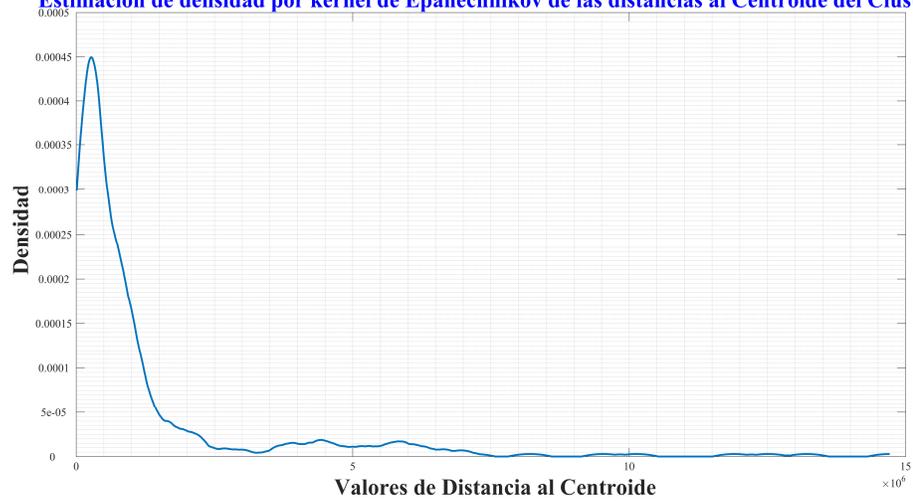


Figura 139. FDP de las distancias al centroide del cluster 1.

Estimación de densidad por kernel de Epanechnikov de las distancias al Centroide del Cluster 2

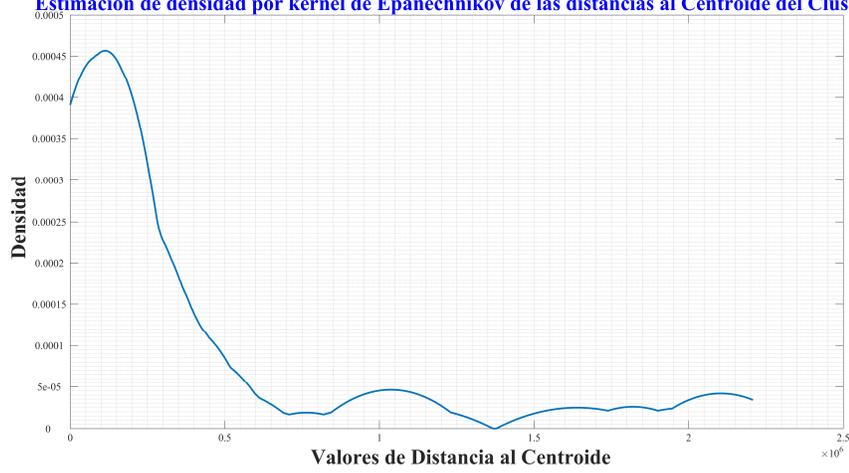


Figura 140. FDP de las distancias al centroide del cluster 2.

Función de Distribución Acumulativa Inversa de las distancias al Centroide del Cluster 1

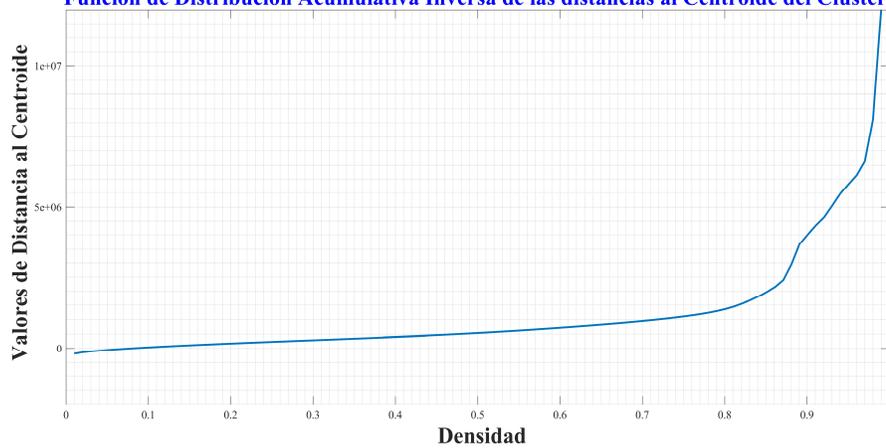


Figura 141. FDAI de las distancias al centroide del cluster 1.

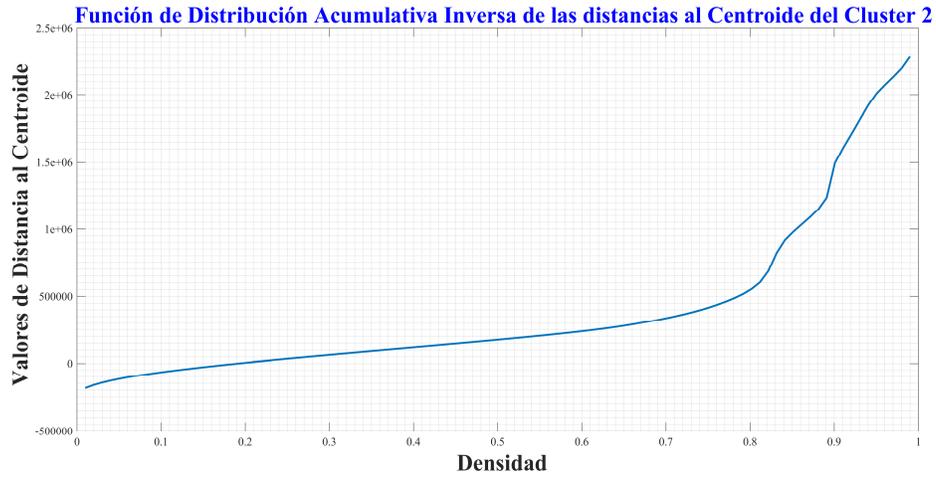


Figura 142. FDAI de las distancias al centroides del cluster 2.

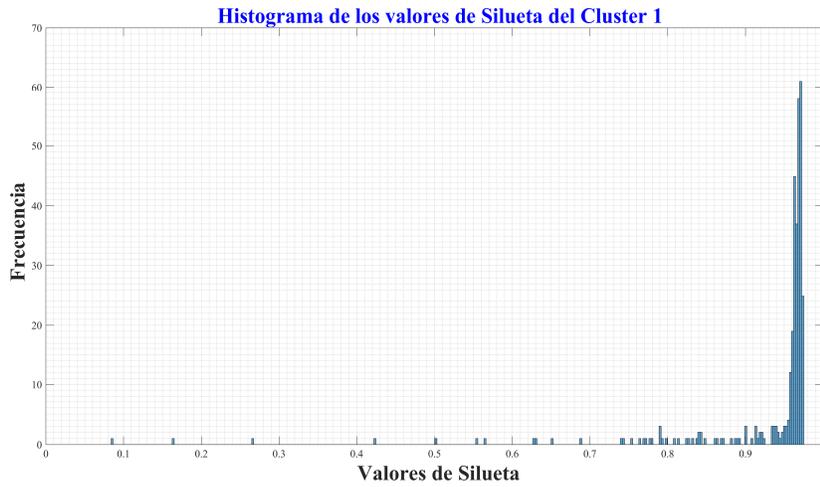


Figura 143. Histograma de los valores de Silueta del cluster 1.

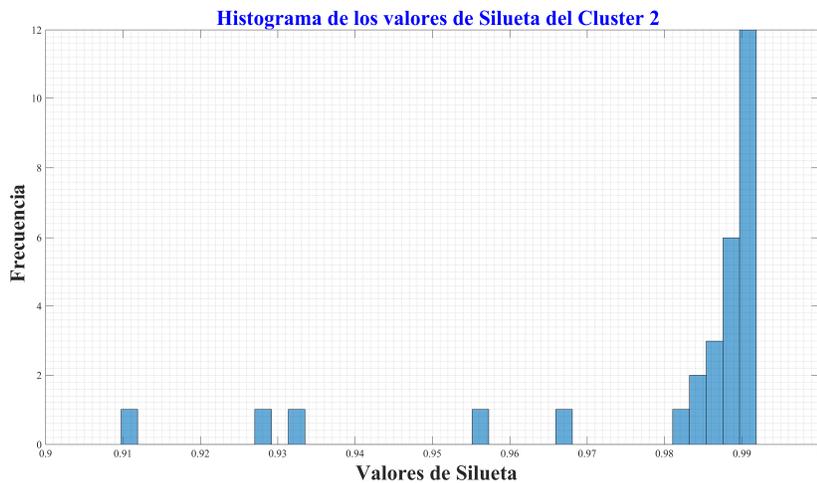


Figura 144. Histograma de los valores de Silueta del cluster 2.

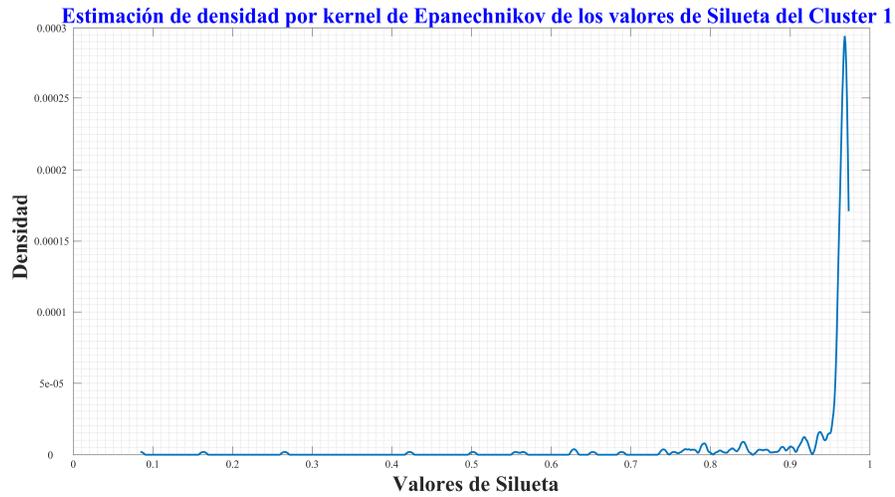


Figura 145. FDP de los valores de silueta del cluster 1.

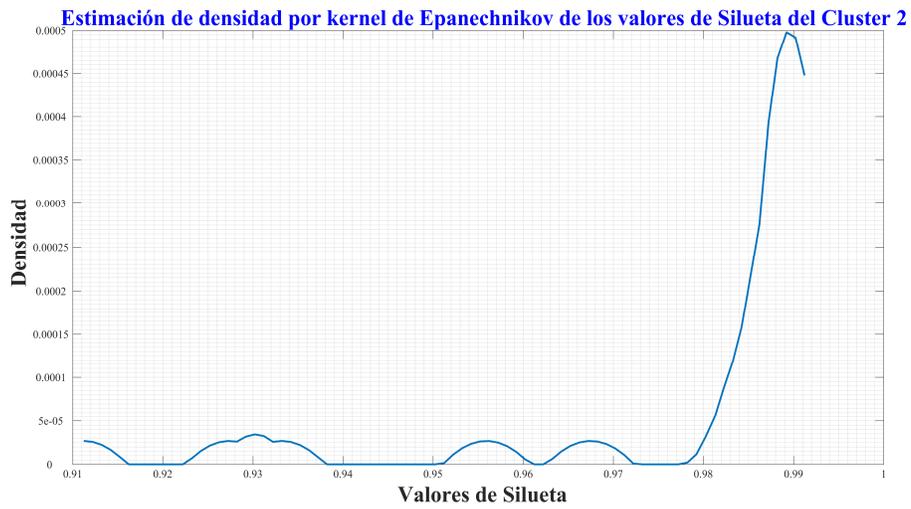


Figura 146. FDP de los valores de silueta del cluster 1.



Figura 147. FDAI de los valores de silueta del cluster 1.



Figura 148. FDAI de los valores de silueta del cluster 2.

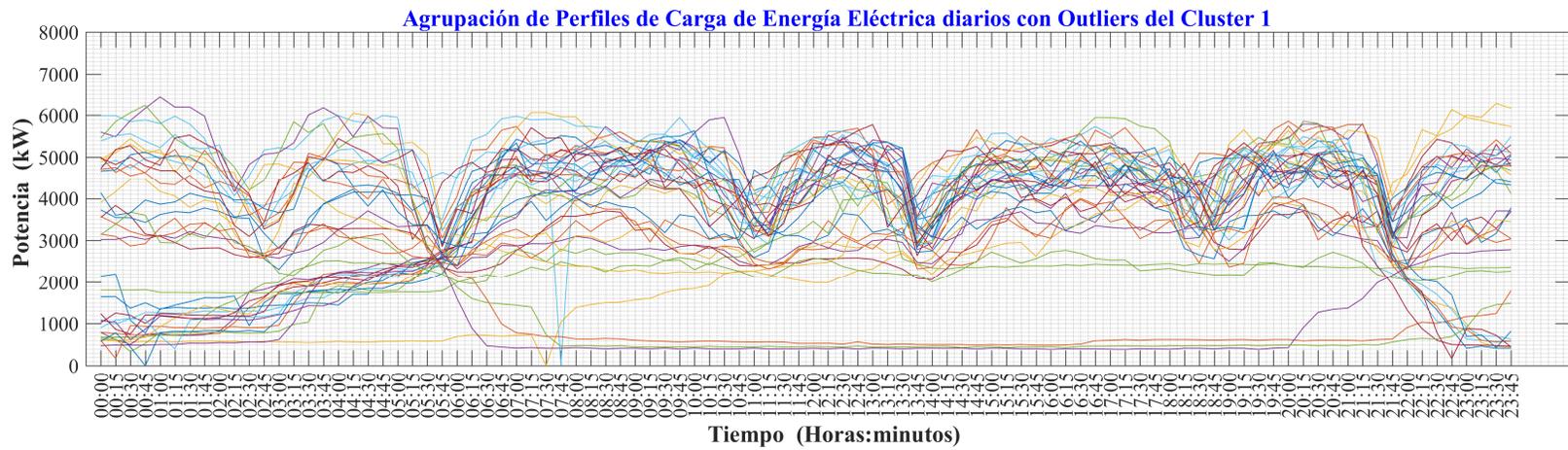
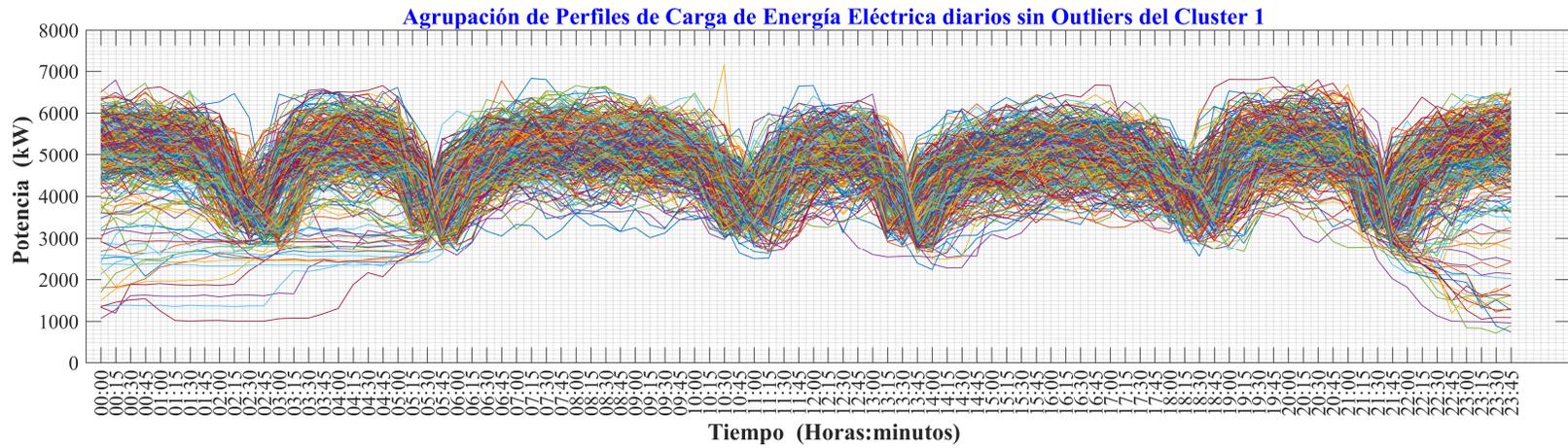


Figura 149. PCEE del cluster 1 (Perfiles sin outliers, 300 - Perfiles con outliers, 37).

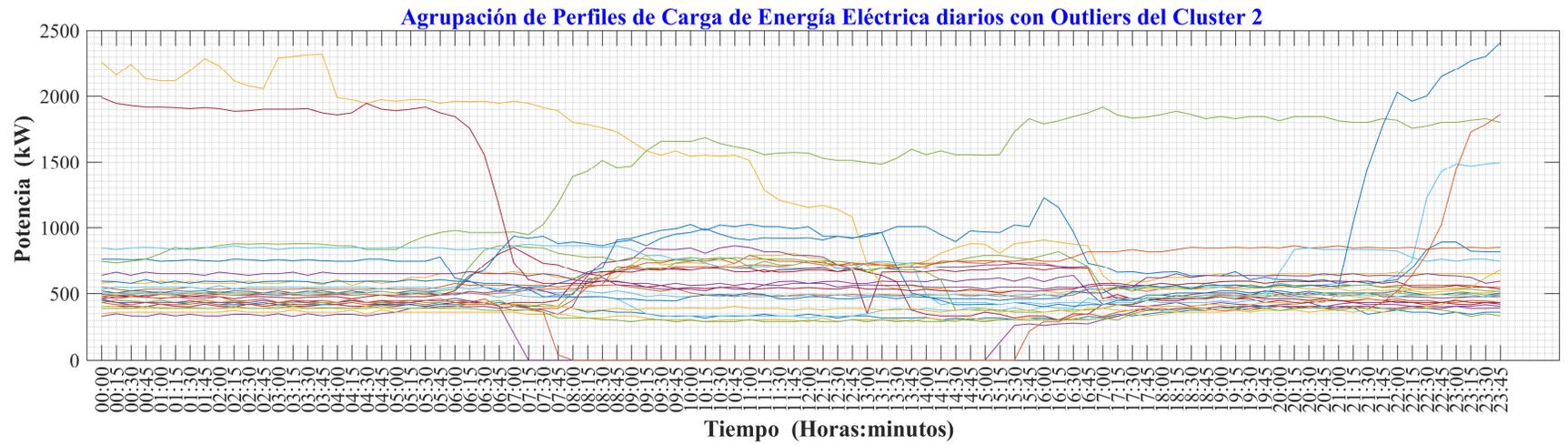


Figura 150. PCEE del cluster 2 (Perfiles con outliers, 26).

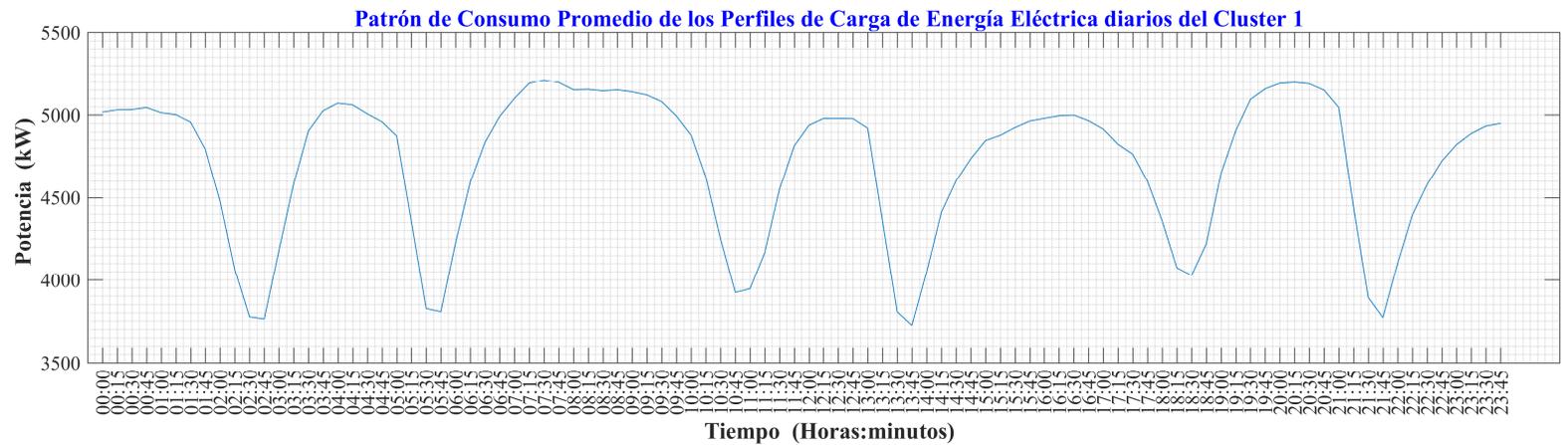
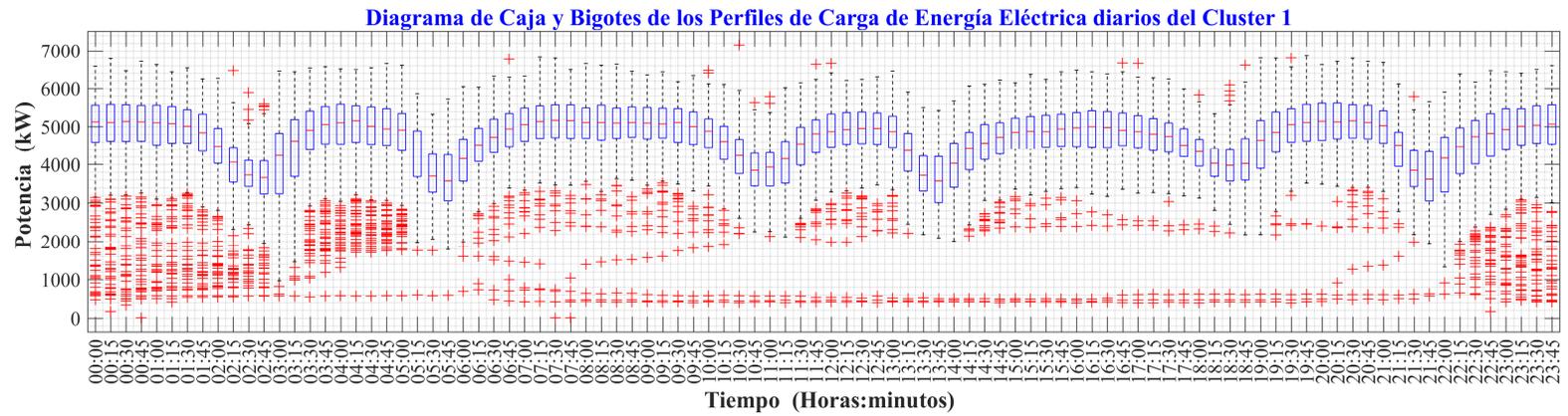


Figura 151. Patrón de consumo de PCEE diarios del cluster 1.

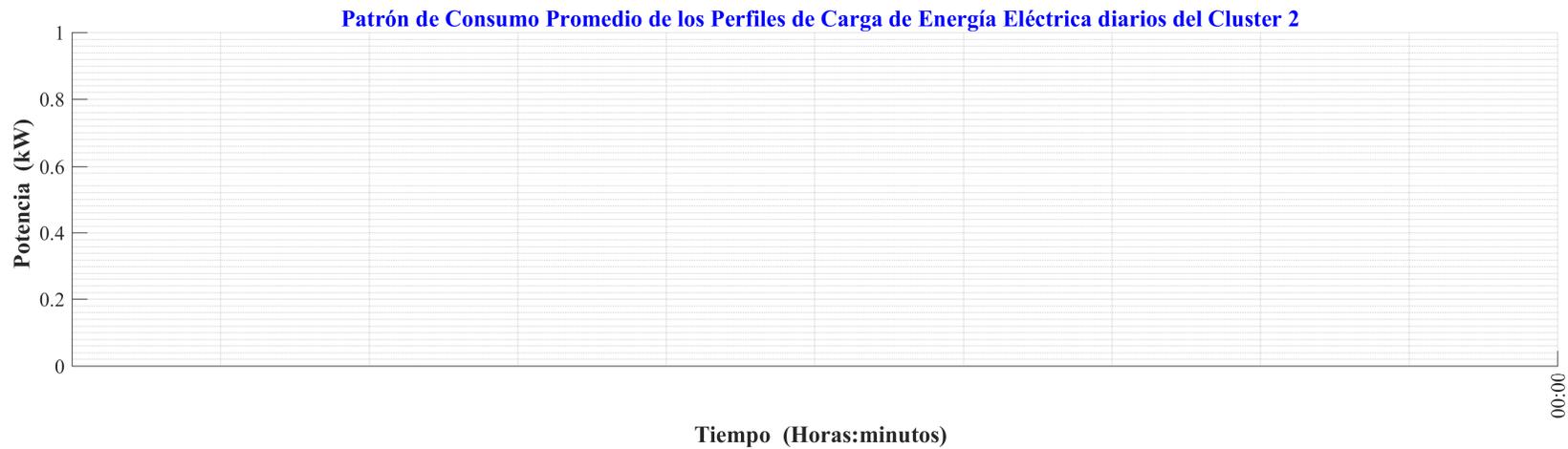
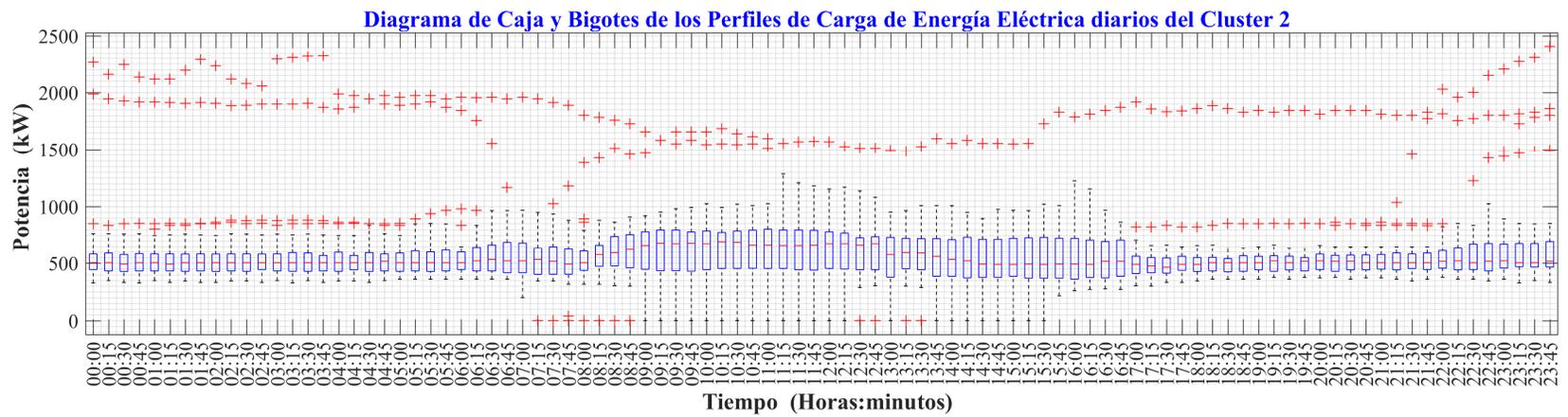


Figura 152. Patrón de consumo de PCEE diarios del cluster 2.

4.1.4. Consumo de Energía Eléctrica de la Universidad Politécnica de Valencia

Por último, se analiza el caso de la UPV, con los datos proporcionados por el docente tutor del proyecto.

Como en los casos anteriores se verán las gráficas de cómo es el comportamiento de este tipo de cliente y cuál es la mejor forma de agrupar los datos de acuerdo a los atributos analizados y de esta manera encontrar los patrones de consumo y detectar anomalías en esta base de datos, así como en las anteriores, para luego pasar al análisis en los softwares WEKA y SPSS y corroborar dicho análisis.

Perfiles de Carga de Energía Eléctrica de los días Lunes

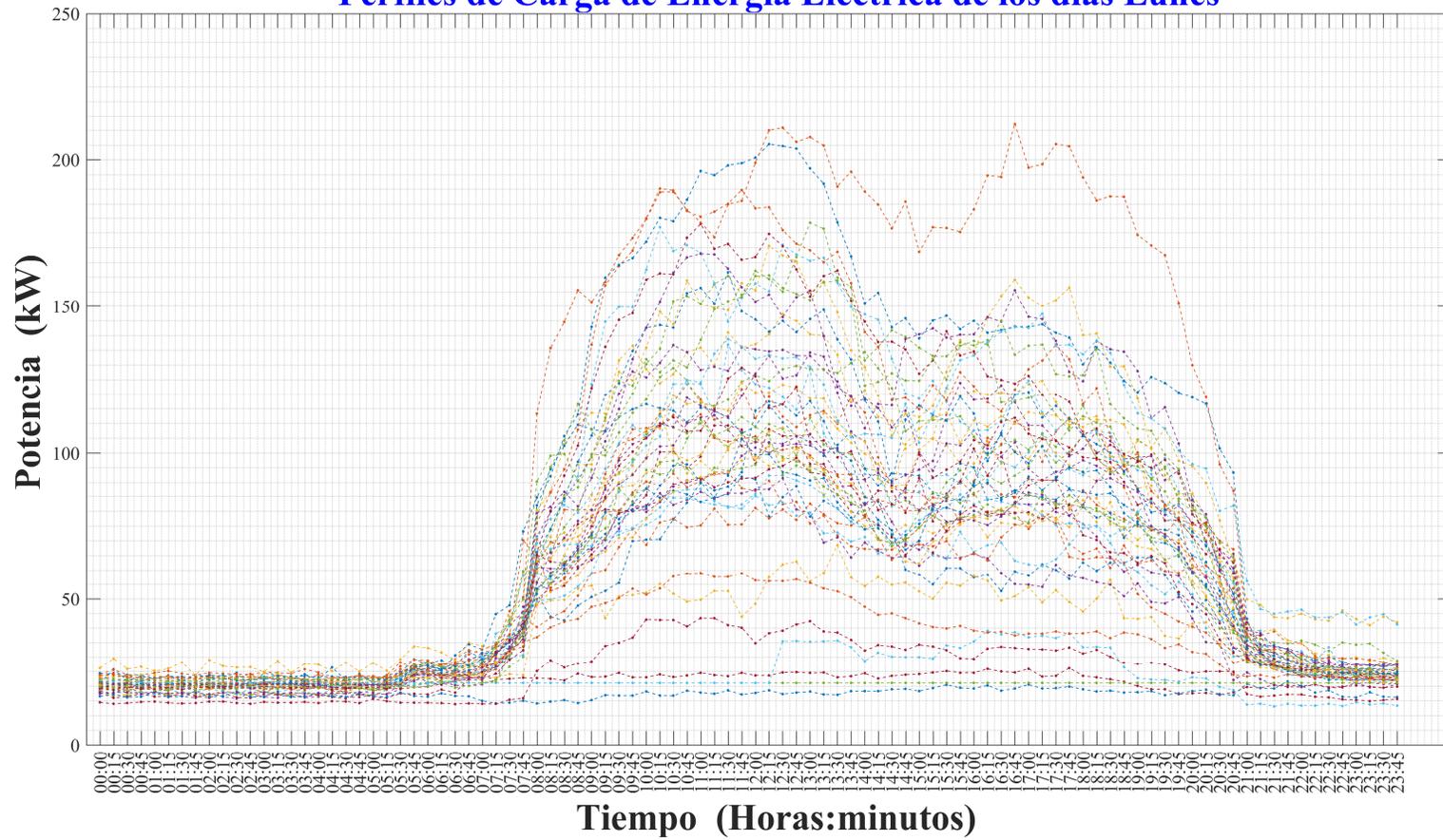


Figura 153. Perfiles de Carga de Energía Eléctrica (Lunes).

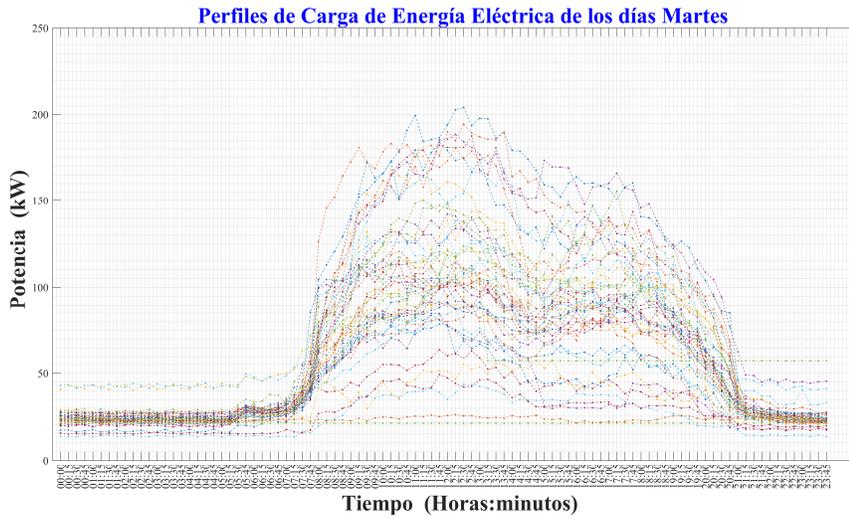


Figura 154. Perfiles de Carga de Energía Eléctrica (Martes).

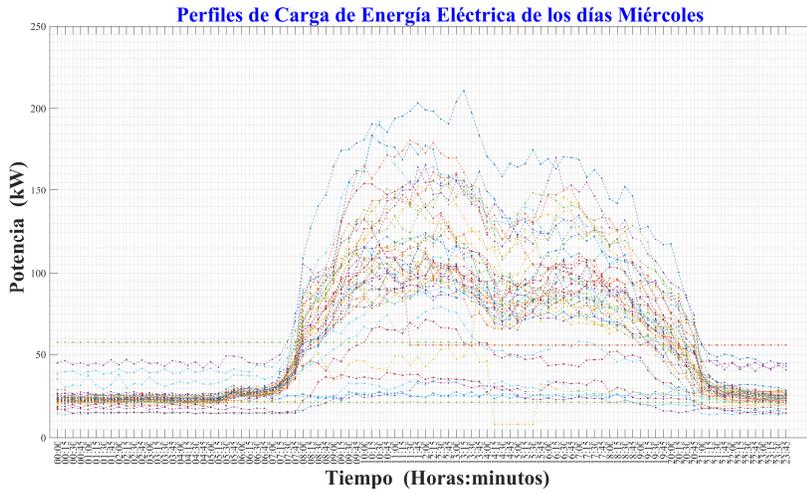


Figura 155. Perfiles de Carga de Energía Eléctrica (Miércoles).

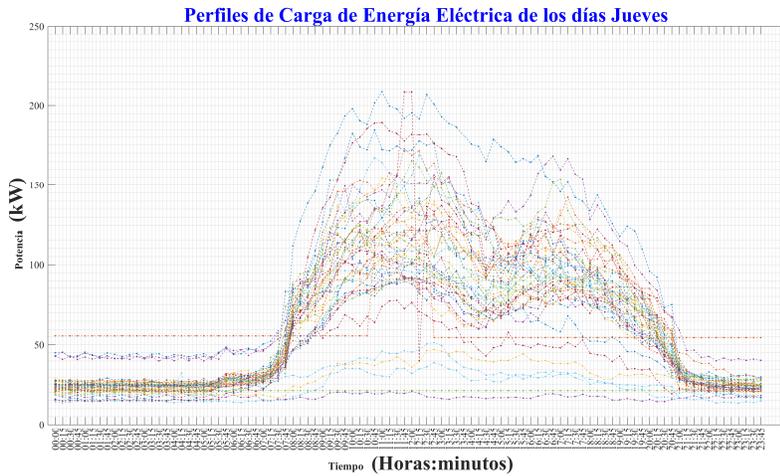


Figura 156. Perfiles de Carga de Energía Eléctrica (Jueves).

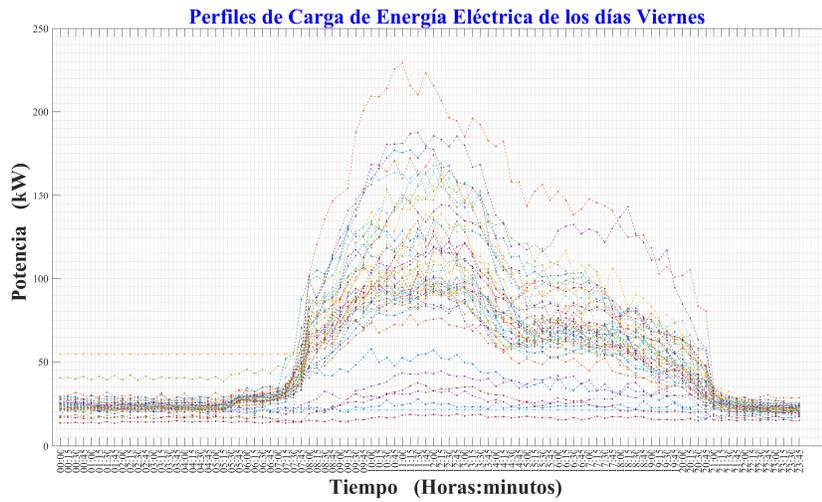


Figura 157. Perfiles de Carga de Energía Eléctrica (Viernes)

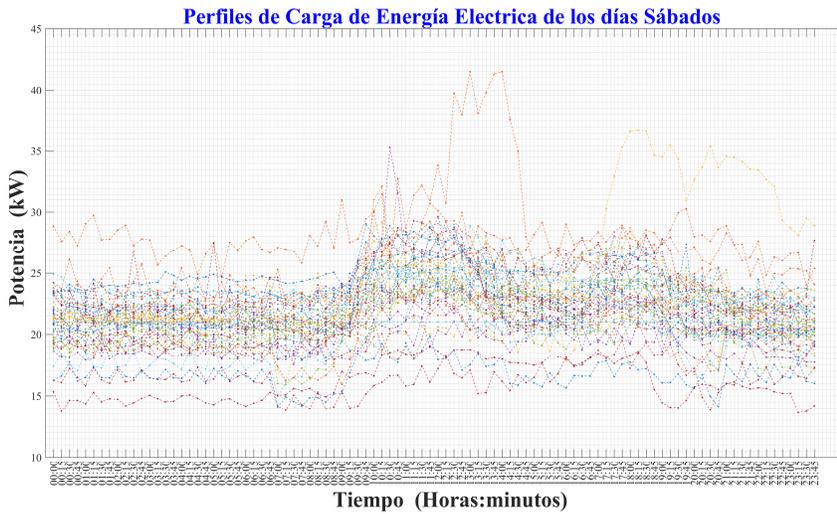


Figura 158. Perfiles de Carga de Energía Eléctrica (Sábados)

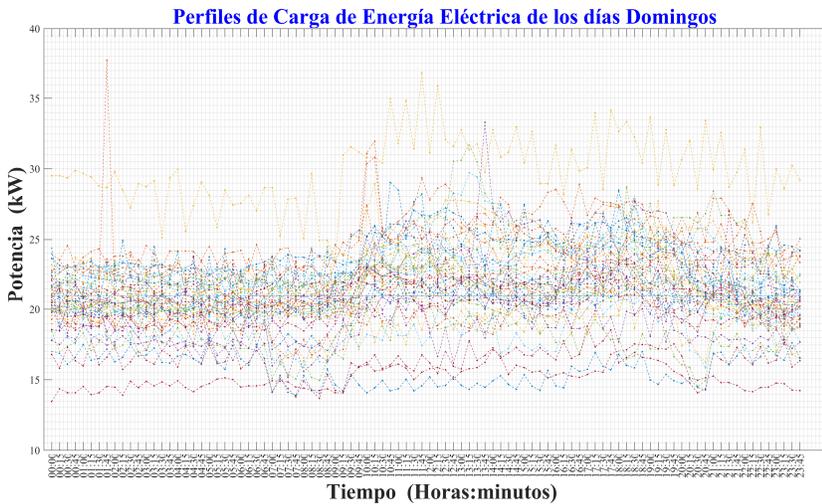


Figura 159. Perfiles de Carga de Energía Eléctrica (Domingos)

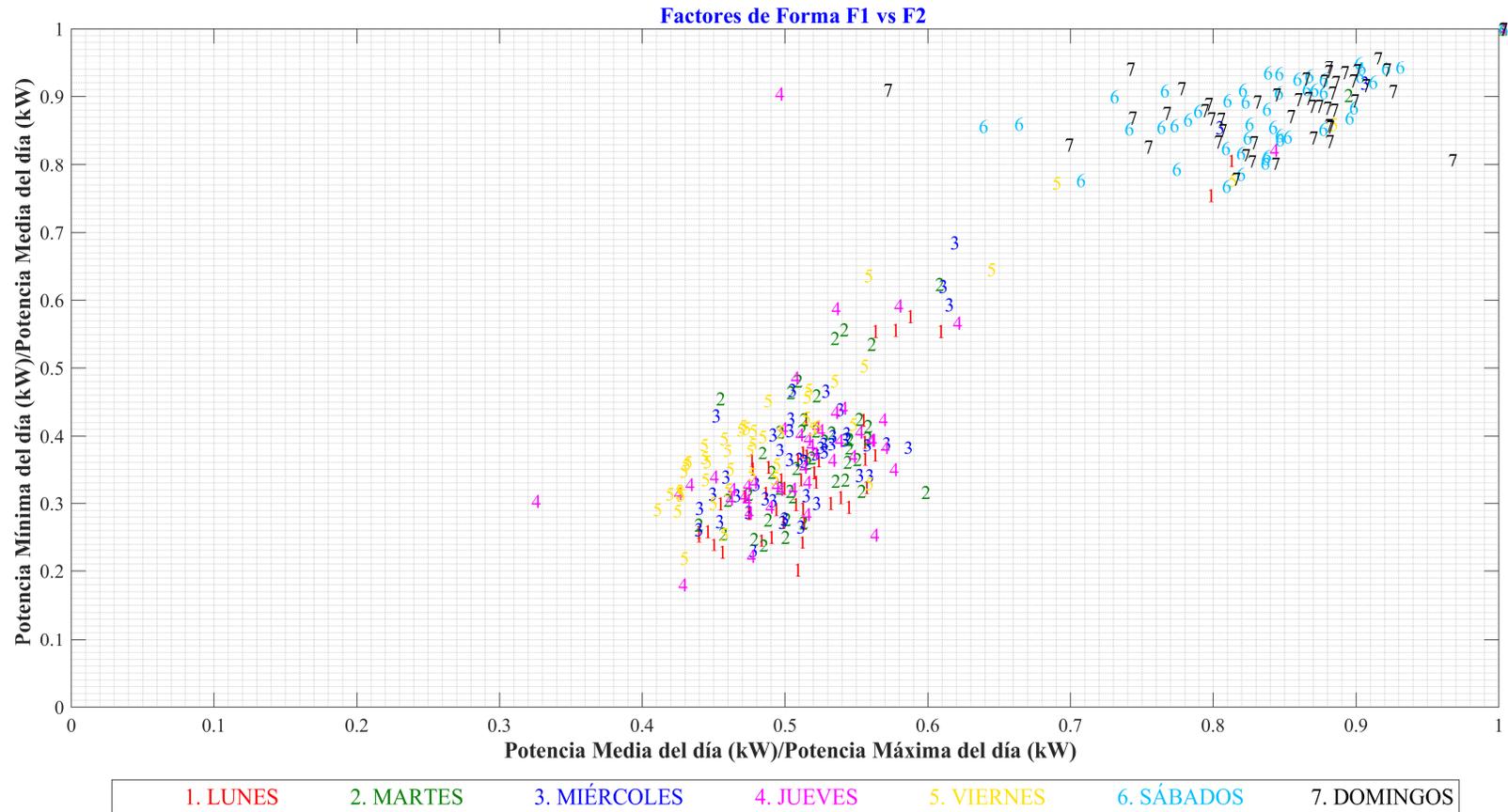


Figura 160. Factores de Forma F1 vs F2.

1. LUNES 2. MARTES 3. MIÉRCOLES 4. JUEVES 5. VIERNES 6. SÁBADOS 7. DOMINGOS

Factores de Forma F3 vs F4 vs F5

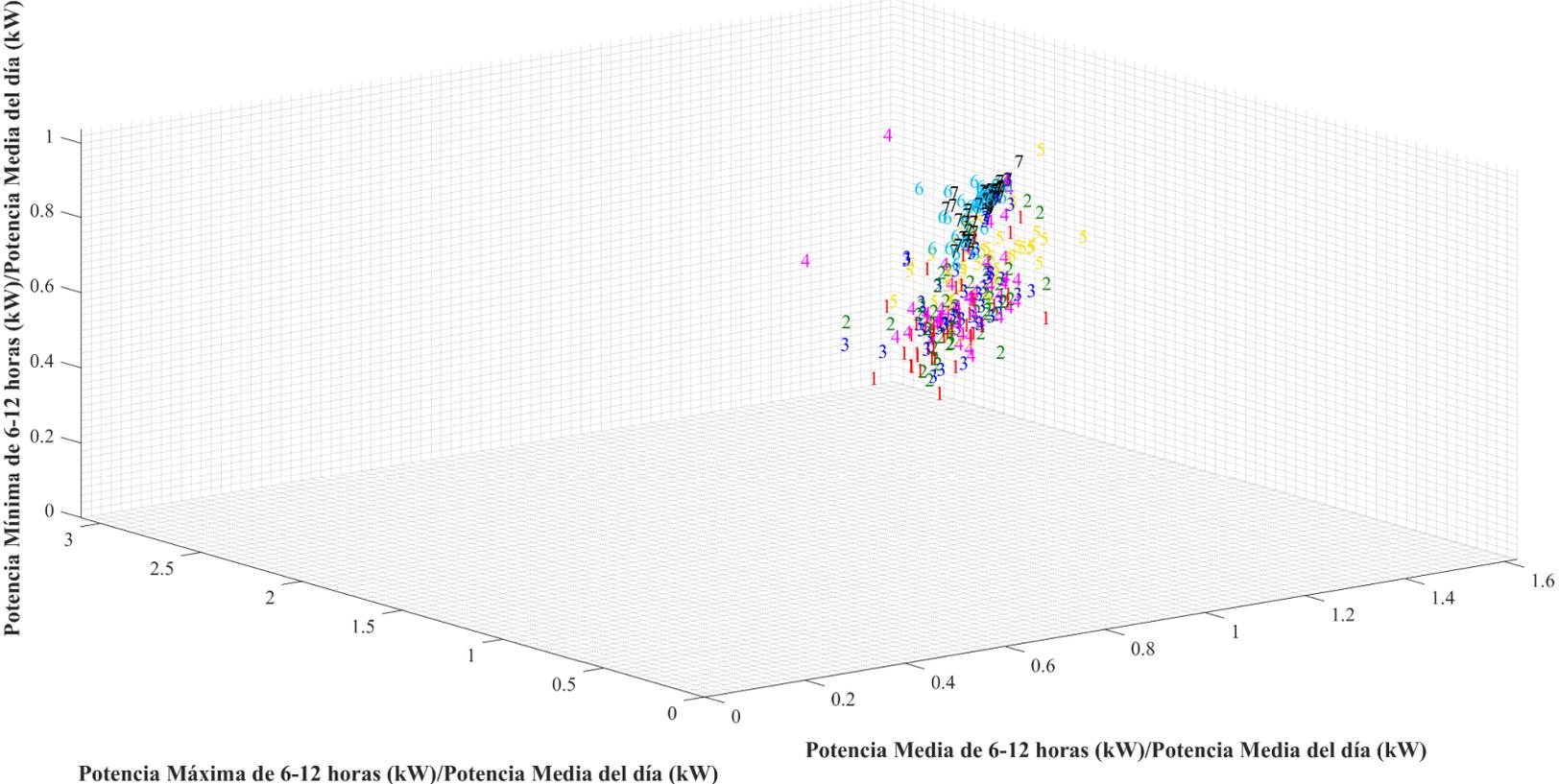


Figura 161. Factores de Forma F3 vs F4 vs F5.

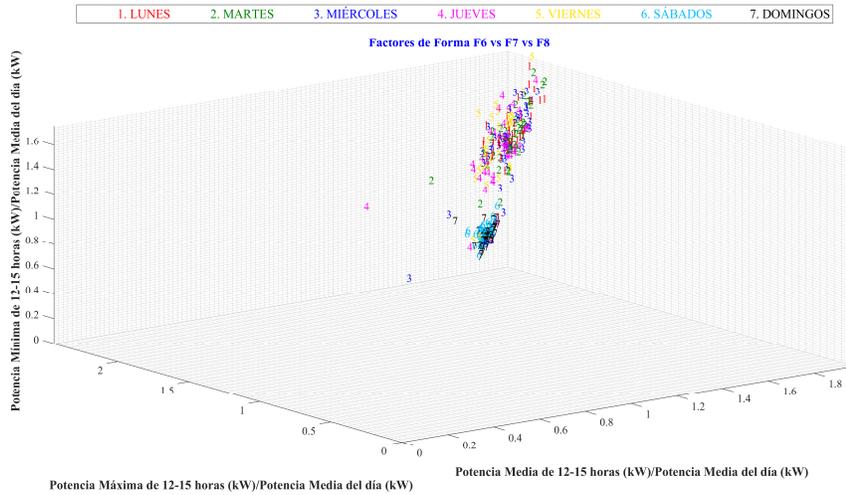


Figura 162. Factores de Forma F6 vs F7 vs F8.

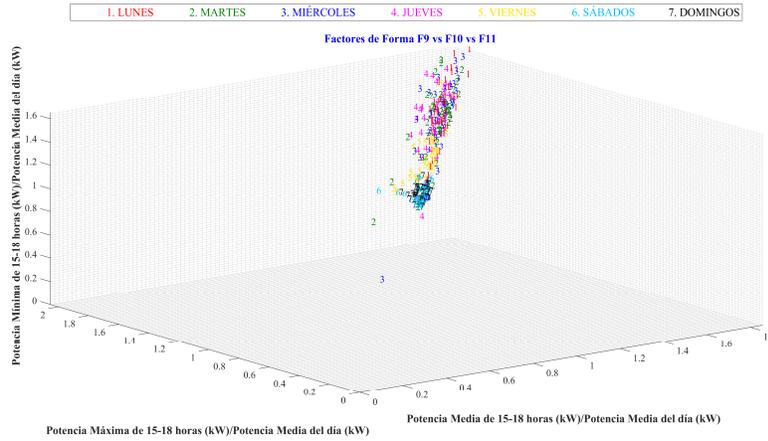


Figura 163. Factores de Forma F9 vs F10 vs F11.

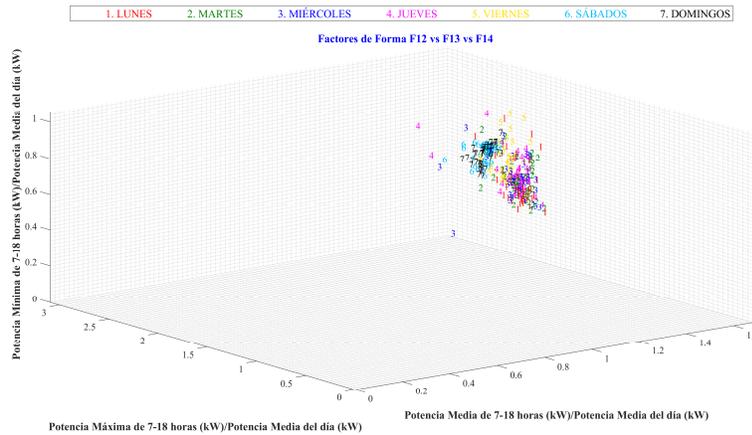


Figura 164. Factores de Forma F12 vs F13 vs F14.

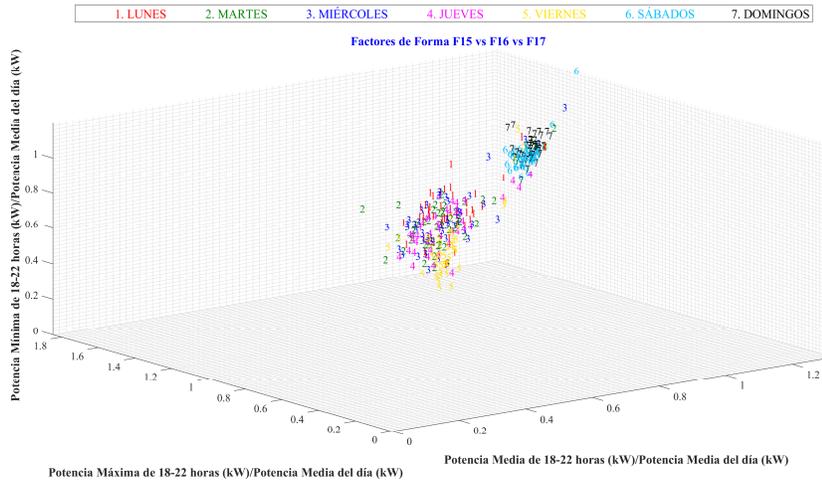


Figura 165. Factores de Forma F15 vs F16 vs F17.

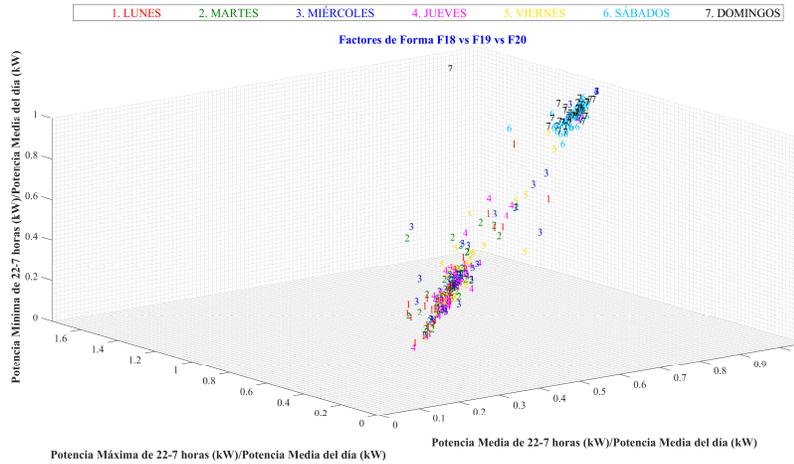


Figura 166. Factores de Forma F18 vs F19 vs F20.

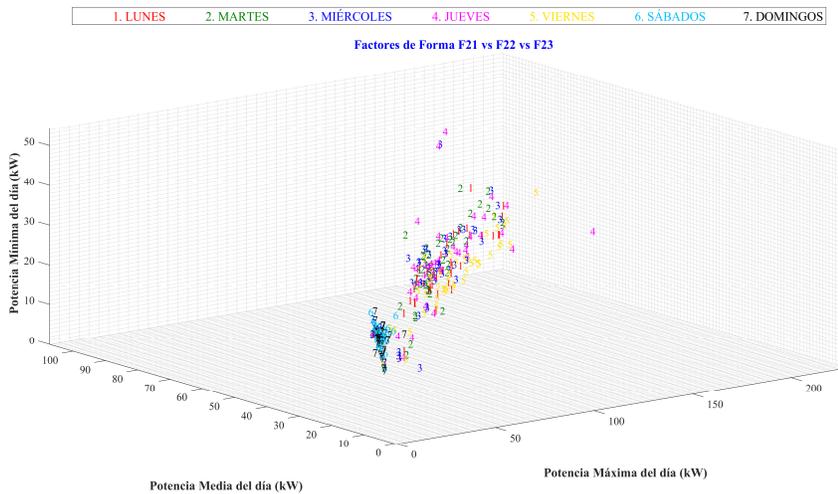


Figura 167. Factores de Forma F21 vs F22 vs F23.

4.1.4.1. Valores promedios de las siluetas según el número de clusters con los que se segmentaron los factores de forma

Tabla 21. Valores promedios de las Siluetas para el número de clusters analizados

Número de Cluster	Valor promedio de Silueta
1.0000	-
2.0000	0.8292
3.0000	0.8558
4.0000	0.7890
5.0000	0.7672
6.0000	0.7617
7.0000	0.7061
8.0000	0.7312
9.0000	0.6926

4.1.4.2. Evaluaciones del número de clusters más óptimo para los datos en análisis

El número de Clusters más apropiados según el Valor de la Silueta:

Ekmeans3 =

SilhouetteEvaluation with properties:

NumObservations: 332

InspectedK: [1 2 3 4 5 6 7 8 9]

CriterionValues: [NaN 0.8292 0.8558 0.7890 0.7647 0.7611 0.7031 0.7312 0.6904]

OptimalK: 3

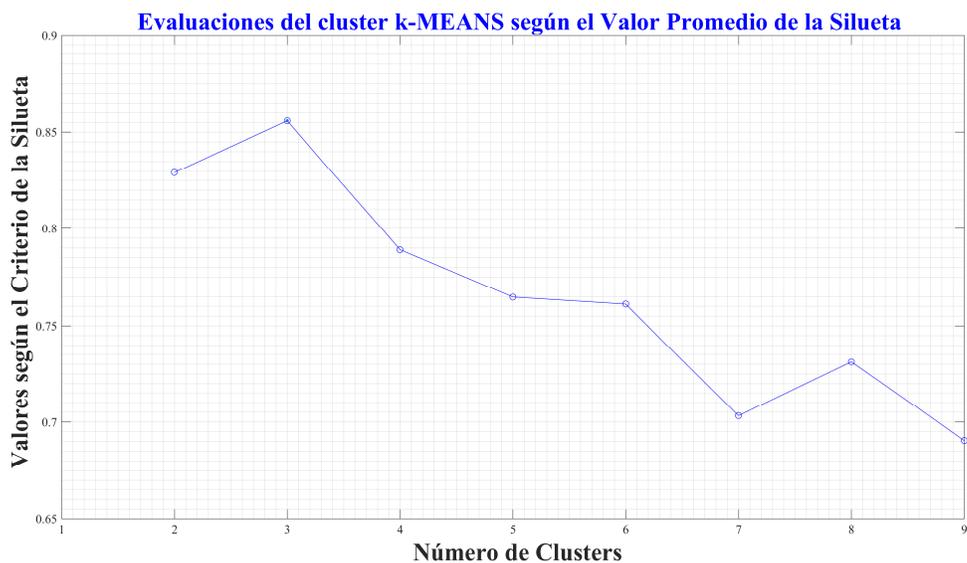


Figura 168. Evaluación según el valor promedio de Silueta del cluster k-Means.

4.1.4.3. Valores de los tiempos de ejecución para cada número de cluster con los que se segmentaron los factores de forma

Tabla 22. Tiempos de Ejecución para el número de clusters analizados

Número de Clusters	Tiempo de Ejecución (s)
[1]	[4.3163]
[2]	[8.6212]
[3]	[8.2472]
[4]	[8.2783]
[5]	[8.3454]
[6]	[8.3265]
[7]	[8.3738]
[8]	[8.3632]
[9]	[8.4061]

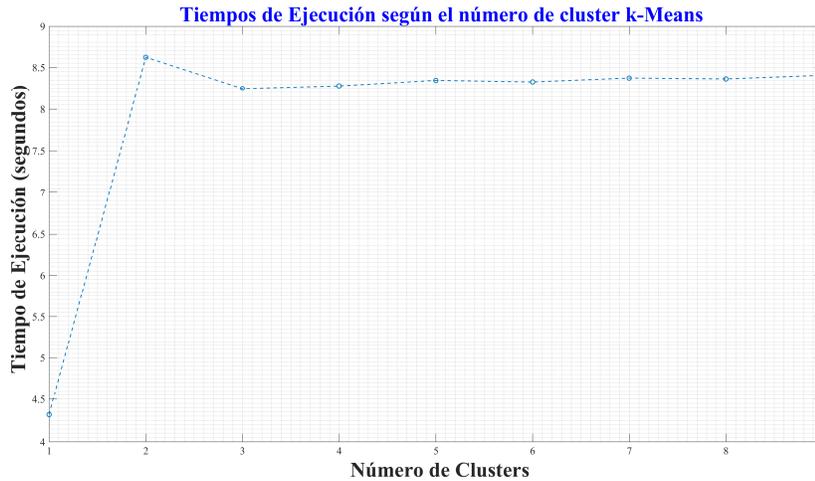


Figura 169. Tiempos de ejecución según el número de cluster k-Means.

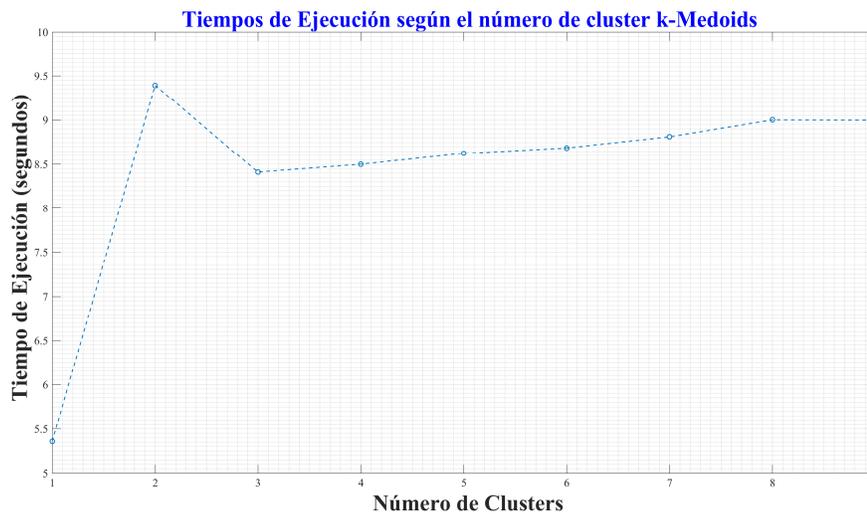


Figura 170. Tiempos de ejecución según el número de cluster k-Medoids.

Gráfica de Silueta con el número de clusters óptimo

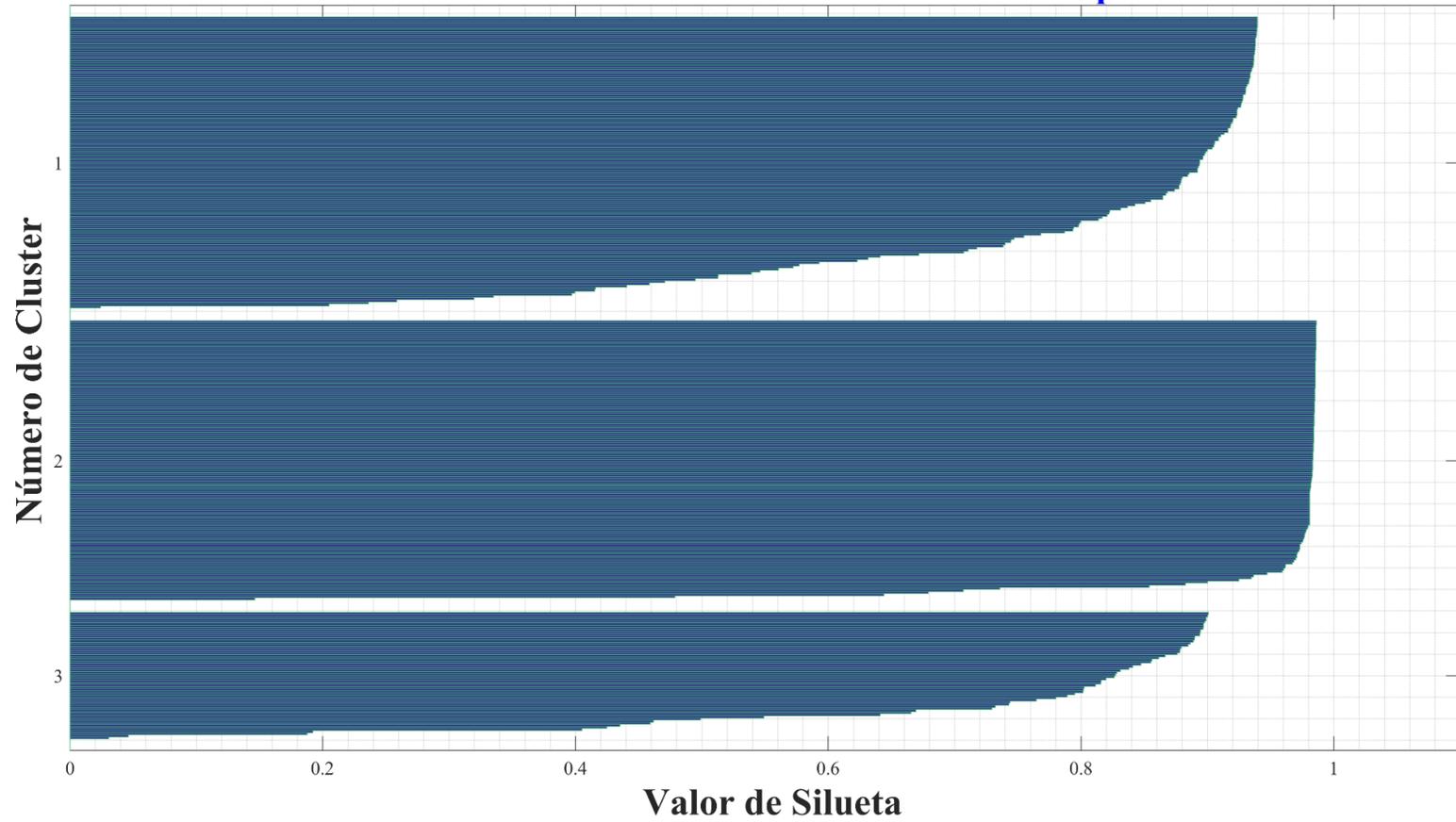


Figura 171. Gráfica de Silueta con el número de clusters óptimo.

4.1.4.4. Indicador del número de cluster a cuál pertenece cada PCEE diario

Tabla 23. Indicador del número de cluster al que pertenece cada PCEE

1	1	43	1	85	1	127	2	169	3	211	1	253	2	295	2
2	1	44	2	86	3	128	2	170	3	212	1	254	2	296	2
3	3	45	1	87	3	129	2	171	3	213	1	255	2	297	2
4	1	46	1	88	1	130	3	172	2	214	3	256	2	298	2
5	1	47	1	89	1	131	3	173	2	215	3	257	2	299	2
6	1	48	1	90	1	132	3	174	2	216	3	258	2	300	2
7	1	49	1	91	2	133	3	175	2	217	1	259	2	301	2
8	1	50	3	92	1	134	3	176	3	218	3	260	2	302	2
9	1	51	1	93	1	135	2	177	3	219	3	261	2	303	2
10	1	52	1	94	1	136	1	178	3	220	2	262	2	304	2
11	1	53	1	95	2	137	1	179	3	221	2	263	2	305	2
12	1	54	1	96	1	138	1	180	3	222	2	264	2	306	2
13	1	55	1	97	1	139	1	181	3	223	2	265	2	307	2
14	2	56	1	98	1	140	1	182	1	224	3	266	2	308	2
15	2	57	1	99	1	141	1	183	1	225	3	267	2	309	2
16	1	58	1	100	1	142	3	184	1	226	1	268	2	310	2
17	1	59	1	101	1	143	1	185	1	227	3	269	2	311	2
18	1	60	1	102	1	144	1	186	1	228	3	270	2	312	2
19	1	61	2	103	1	145	1	187	1	229	3	271	2	313	2
20	1	62	1	104	1	146	1	188	1	230	1	272	2	314	2
21	1	63	1	105	2	147	1	189	2	231	1	273	2	315	2
22	1	64	1	106	1	148	3	190	1	232	1	274	2	316	2
23	1	65	1	107	1	149	1	191	1	233	1	275	2	317	2
24	3	66	1	108	1	150	1	192	2	234	1	276	2	318	2
25	1	67	1	109	1	151	1	193	1	235	1	277	2	319	2
26	3	68	1	110	1	152	2	194	1	236	1	278	2	320	2
27	3	69	1	111	1	153	1	195	1	237	2	279	2	321	2
28	3	70	1	112	1	154	2	196	1	238	2	280	2	322	2
29	3	71	3	113	1	155	1	197	1	239	2	281	2	323	2
30	3	72	3	114	1	156	1	198	1	240	2	282	2	324	2
31	2	73	3	115	1	157	1	199	1	241	2	283	2	325	2
32	2	74	3	116	1	158	1	200	1	242	2	284	2	326	2
33	2	75	3	117	1	159	1	201	1	243	2	285	2	327	2
34	2	76	3	118	3	160	1	202	2	244	2	286	2	328	2
35	2	77	3	119	3	161	1	203	2	245	2	287	2	329	2
36	3	78	2	120	3	162	1	204	1	246	2	288	2	330	2
37	3	79	2	121	3	163	1	205	1	247	2	289	2	331	2
38	1	80	2	122	3	164	1	206	1	248	2	290	2	332	2
39	3	81	2	123	3	165	1	207	1	249	2	291	2		
40	3	82	2	124	3	166	3	208	1	250	2	292	2		
41	3	83	3	125	2	167	3	209	1	251	2	293	2		
42	1	84	3	126	2	168	3	210	1	252	2	294	2		

4.1.4.5. Ubicación de los centroides para cada cluster

Tabla 24. Centroides para cada cluster

CLUSTER 1	CLUSTER 2	CLUSTER 3
0,510435345109813	0,814962018884046	0,463009254861510
0,378903737966392	0,845380261380766	0,287082262044562
1,29833338716460	1,02234968273978	1,32155519302721
1,93010418721497	1,19807446472711	2,12342285496822
0,486865210011133	0,872746453079940	0,355125233337842
1,59132399641276	1,10301291143427	1,77486330721763
1,87582460744027	1,19617629486423	2,08137047574216
1,29786201341123	1,00958280020985	1,42038250593524
1,43697862460414	1,06534173815119	1,53132409043327
1,57683054511682	1,14094834872931	1,68919858645893
1,29469654679023	0,993557494112270	1,35186019153446
1,48762028377420	1,06530039897183	1,58854725054898
1,96809933026503	1,24298879496705	2,17430797715350
0,587977092585705	0,877429308519601	0,415710345310519
0,913848849963299	0,998348457903947	0,882306837453662
1,39568487771803	1,11800548976516	1,41902971676700
0,466869935138057	0,898021966848669	0,355371183073941
0,442309053181175	0,920922419966005	0,332972543794067
0,559029045538424	0,999694888339674	0,417257585601371
0,379058547167649	0,856135214006082	0,287082262044562
109,535420744309	28,7091858020324	171,875333333333
55,7072359639382	22,4074749099007	79,3319861111111
20,9046675456327	18,7018660734834	22,5926666666667

4.1.4.6. Distancia de los factores de forma que caracterizan cada PCEE diario hasta el centroide de cada cluster

Tabla 25. Cuadro de distancias de cada Factor de Forma al centroide de cada cluster

CLUSTER 1	CLUSTER 2	CLUSTER 3
386,694424646003	4606,50671862885	7444,07734600304
244,659037280925	10614,4686391879	2622,86900919448
1726,53646435489	16632,9963218568	649,493494838006
911,968763040695	13831,0654714852	1350,95354480880
46,1356672216021	8497,29611132297	3902,10955391370
829,262875018164	13470,2552836971	1464,67744416831
24,5057614968958	7850,09706454170	4351,56685471414
96,3472652042695	9313,42019727001	3363,38586757881
275,384352005388	5061,90098062308	6914,59381170368
46,0000488317352	8312,63139192698	4010,07298932117
485,732662591635	4424,17672037194	7749,71025390212
797,620153271393	13293,4163133121	1561,03322179023
681,263963660172	3804,24868893180	8585,15267037445
8208,57968518383	18,6216037551313	24726,1121562394
9253,93964738613	97,1256838451394	26502,8925155984

843,333578916934	3434,20236521152	9151,42655155306
291,665057037278	10876,0945992537	2508,71148794624
79,1340725421185	6353,24125240180	5569,19560497501
51,7230444644565	6513,04754844411	5407,11919033360
315,570181899912	4908,77557942838	7104,54719140806
163,008135990255	5610,72944487691	6287,34520997546
65,0240287750847	9009,81587607623	3514,84022292339
23,1988581015733	8427,99646957904	3901,41384607221
4469,54293501884	23803,4549339335	19,5453839465974
712,455004931630	13013,0141105806	1616,14727849352
5230,03892632666	25479,6611744205	52,9240701713905
4992,23420358905	24951,2599851073	36,3948720635580
5262,29224874029	25566,3115045230	50,3622748029839
3014,74505228467	20245,3824096611	145,591103185243
7334,94965878432	29944,9749398392	363,409719447762
1990,92023399814	1841,36944147316	12381,8960623153
9049,46487718887	67,0109990427283	26177,8934838163
9049,46487718887	67,0109990427283	26177,8934838163
6066,64793710807	140,021264979210	20851,4626198503
5370,77599734046	247,446755372679	19542,5340139179
10549,0130729830	36156,7906206539	1301,37723927921
13239,8461894152	40955,5663507595	2425,76877652429
303,769539188643	10921,8229623291	2503,51087568837
4032,37822488640	22779,1818469809	25,4833997503979
3173,63010001963	20666,7677726573	129,167213350527
1814,01352297022	16881,4199210167	621,763074159475
426,716973376658	11670,8149924400	2130,88366994498
13,6023427678212	7061,21450272451	4924,35009369030
3010,90955699545	1063,40459421446	14772,1341744396
87,4619865645187	6123,27694810490	5763,65921563139
15,5690609454548	8276,99945939333	3998,55335783894
18,4405572159026	7915,93866552759	4265,70883898481
677,376889629513	3795,80423078473	8570,21295156549
184,564446813553	10040,9720556036	2954,68064850840
1411,69339080783	15622,8041681556	865,820198900339
773,726084403687	13236,7540346744	1562,99576107197
87,3673840044609	9370,95998844000	3291,70731600135
1007,37595247425	14200,3707197066	1236,85608347765
57,8711309333620	8924,11969577356	3590,09628431788
31,2112484997143	8257,15541754504	4052,90455958480
252,751016975310	5169,03225772121	6794,44758136649
249,018966332301	5154,01894469739	6795,47350419166
935,602968876527	3390,93102639275	9317,71983858517
302,538931880215	10814,0613309416	2591,61768674416
653,775601333215	3856,19570015345	8482,02091099728
2559,74189401906	1375,57345576220	13730,9366519983
1152,04381470278	2915,83852046434	10087,2285777996
471,725052297088	4331,95622246074	7804,99369943633
7,56774348388671	7344,12486734376	4697,61913118793

344,523437304766	4774,65580741808	7255,66062957132
221,770933439260	10467,1611390898	2687,81804598407
29,8890015724515	8523,88851380283	3826,05274739994
64,6713469862451	6318,48197945770	5580,19438959637
51,7380615255989	8882,85171202940	3595,67078113175
83,9203615185497	9186,40674492826	3439,40086671443
5258,90144854159	25586,4458631660	50,7851961249833
1265,49524284780	15134,9510818955	976,587489090912
1849,45955565624	17017,6456434623	561,320157192015
4228,04452631656	23232,0770912804	34,7926774018682
6988,26357928870	29252,7907363082	287,182994543055
6207,03796322113	27602,1870311873	157,177287990190
5707,59324738769	26560,1149612590	83,2729610850531
3148,58160149709	991,652976637288	15072,7978142335
9049,46487718887	67,0109990427283	26177,8934838163
9049,46487718887	67,0109990427283	26177,8934838163
5395,20498627541	260,126572916384	19573,8595230141
4532,84567894762	454,168574659952	17905,3372796667
10635,0599559301	36319,4668858636	1344,63109585234
8680,20012859190	32618,5872791895	723,140969906965
266,475859520411	10613,1426070570	2676,94440293007
7916,97133202630	31068,5297368970	576,401059632102
2610,05114903411	19183,9377008110	265,959959530256
781,704569114419	12827,1090308075	1923,77581303187
18,9280463497848	7129,44055180839	4889,05275346430
165,354396221849	5640,70559918713	6280,26754584140
8095,06477976293	13,1803714906147	24537,5026159121
7,67250472335201	7194,31397681008	4810,87753227096
40,3327886429104	8540,73618294530	3849,32799987924
22,9248203162839	6884,01950852090	5069,77578233897
7174,74730075526	15,0252708616863	22912,9873968254
81,7594277231300	9155,46871170936	3455,00849349518
378,365219469499	11414,1164276232	2247,68352841658
105,393195946208	9295,19983671552	3393,97568703531
10,9944496338355	8234,58444603094	4023,56273406464
77,5559426194013	8940,39599092388	3621,19718815118
342,703870692867	11198,4054765865	2344,34140993314
257,580715784177	10691,9626854243	2581,49683035804
6,91445636402905	7395,34116989162	4646,55795357416
174,244443889252	5535,89375557056	6366,09301579811
6921,60586901245	51,1190648368040	22423,8133715979
390,266536092033	11468,9137202047	2217,13510273707
1176,81139249324	2858,75121600485	10161,8445139719
1791,51381293266	2060,61501328889	11856,7483997919
304,351072480270	5075,11026859447	6945,93450006227
49,7499465204437	7134,67703163448	4903,83501645959
9,92202366040091	8082,27302161739	4133,65624266994
282,034532355906	5066,27959173247	6928,28799136731
201,269802679679	5374,35376158976	6540,18191160577

234,770202099490	5218,98940658316	6720,35369830862
8,51586508037948	7199,59592815095	4810,71568253818
327,852345216400	10959,7027123848	2524,60294657073
657,858603385009	12727,4570991041	1720,99862509387
2908,69966687529	19983,8148318607	164,453540905373
3460,73957124586	21400,0640706728	63,0136344091758
2636,32184026608	19265,6089165717	242,121288148883
7796,30036084411	30882,0371386037	467,414322013472
2316,40032473226	18342,7926536312	361,463661606034
6178,05607247341	27538,4093554142	154,514065731485
5569,95919019039	26237,8393108705	74,1351786559757
3711,41689895392	944,537123328104	16078,4566551361
9049,46487718887	67,0109990427283	26177,8934838163
9049,46487718887	67,0109990427283	26177,8934838163
6652,96417121896	79,1369596644579	21929,9474929844
6020,07465586247	131,485987713061	20777,4653982578
12161,1123107083	39095,3480089367	1919,16141769612
3283,41027463396	20945,4448237945	110,221055584369
1419,29234391557	15629,8768520756	879,807483113521
5270,10575440258	25132,6381710690	444,552802084585
2629,76868666427	19219,8968697302	278,656035346015
7610,63962709204	28,3725922306541	23676,6513304477
24,1183878438375	7468,13378384181	4645,41167799020
4,17479852781439	7550,47925389529	4536,80392210156
43,2587632599048	6562,95282660804	5356,93525397502
14,0642983580785	7043,11630456980	4933,89539888141
7,09288131061302	7500,39970634455	4571,28832698758
59,6387554793240	7473,40604422746	4712,07104905335
1772,38612652974	16780,0597913906	605,684535808186
653,898839804327	12688,9586414097	1770,85419467302
838,265298567363	13543,4617641023	1426,34623643622
7,65269956192032	7466,26755174669	4604,79088328926
287,359687666743	10886,1536719828	2494,52730755801
85,4439253637966	6788,44084309368	5271,69949810908
2235,41750732199	18110,9745378840	395,831765074025
59,9961815335908	6493,43979626101	5440,89508374555
104,984610738276	6020,30446121665	5884,31145638145
349,956585104062	4821,50950608376	7235,29496715665
9450,60188811252	114,044451637656	26837,2605648697
148,359599362402	9906,40657366908	2990,90735969173
4215,61098518828	545,165976981916	17276,3789132707
1301,47266233852	2668,98760100447	10526,0062851246
447,514811558697	4575,77466449259	7569,76175778544
1291,35514079096	11372,0667787129	3880,96476112736
18,2390677370031	8274,47803022818	4013,25762166780
298,246786944782	4950,13336120140	7039,14335721193
143,859100223515	5890,74993617985	6057,95844196647
65,4397429696413	6508,01265244713	5448,45140947349
297,504243991734	4970,69526749341	7025,07618238349

767,494681471948	12867,0959866855	1856,24568976196
1031,61761057918	14299,1751585293	1199,21490551426
987,377085980264	14095,5259651696	1261,81988308219
2004,06791959261	17450,9990224243	523,810603948093
1734,52043687655	16657,6254579835	633,238433045065
3753,03620276998	22023,4102891572	75,1667378249983
9910,53142842747	34312,1487522144	1487,92872426990
6265,07559448515	27699,2353686788	180,691144354445
4548,30929801413	23984,4577312410	3,55442167834381
4656,00318534004	381,391196551191	18189,2862534421
9049,46487718887	67,0109990427283	26177,8934838163
9049,46487718887	67,0109990427283	26177,8934838163
6152,37672771302	135,311915382313	21002,8897272895
7052,95272648512	29260,7633641334	376,985582053953
11627,3651937104	38132,9918015138	1710,58537456686
1251,54549877015	15084,3737928445	986,877552424253
2307,87512110628	18355,8162776279	348,382911885399
4991,20554881067	24497,9758501419	437,288119818509
3620,11849930764	21781,0536502367	53,4046012608679
124,716163126127	9718,98073437487	3094,83976050789
1,99433866489723	7464,36422627859	4595,03893177650
66,0483302319933	6338,36110050944	5571,16245567985
43,4177703749976	6550,35262106045	5360,45560844991
22,9301613651412	6900,28900026510	5063,03450082770
5,23799239548415	7440,57302424094	4611,66674615123
14,4914546210465	8235,78148452458	4029,57790898670
7904,05823690068	4,31126711299573	24202,9753181589
94,0078605658187	6345,96087359535	5612,24555734455
202,367718079049	10281,5575103969	2791,26736521493
6086,04017095498	90,7479541887101	20931,6396418614
17,8209052035821	8125,60109711264	4104,56212048517
460,110671284819	11691,9461427618	2153,63581934124
36,7208254708570	6654,56618658089	5272,15949916357
573,467591560662	12348,0990810879	1861,74470667835
236,659856473282	5223,97023328583	6710,46234139767
145,946306199942	5868,22925605133	6045,56951339441
548,879838780766	4105,67275781185	8120,41925689170
312,988444342368	5028,30675413734	6987,71834040594
196,214131180839	5513,82641161737	6412,35303117969
9872,03627599536	160,555085742920	27544,1234191681
3285,04608227599	936,339365125511	15345,0324032333
1266,10442446613	2731,99988789652	10417,8572883891
80,4264409136858	8968,41462376108	3603,95222870581
104,895108633928	6163,93826969920	5750,55063167164
157,200482324838	5741,40806546102	6164,80186843578
142,120687254399	5865,62438234030	6042,04168489187
99,3060975873400	9024,31050742726	3551,63980329911
243,575097373441	5165,10213723500	6770,04702892754
638,865886330807	12498,7256070898	1849,02869767901

862,482655770890	13535,4023675269	1459,25781635685
162,853341798992	9719,77412291218	3132,70350451335
3096,44483080152	20362,1429442249	184,031688535992
3706,27053340898	21838,0108717090	125,358797148625
1750,98468158923	16569,9589590886	703,576039130162
662,006061130203	12593,3533519512	1817,27506707415
4527,51183308133	23865,7002519648	36,4343100403154
2899,75548168330	19829,9978394105	239,673910002960
4991,64108873902	284,200473750396	18856,4929502721
9049,46487718887	67,0109990427283	26177,8934838163
9049,46487718887	67,0109990427283	26177,8934838163
6468,71496459006	111,002487042022	21575,2056058763
5069,26113535491	25093,6072386523	54,8945096769796
16103,7262034521	45913,9362622559	3634,49485929879
929,910472564316	13863,7096894825	1335,88141995973
6986,57396026330	29249,6138556793	286,817183482971
3376,65754978003	21158,4006910889	84,1884959963434
3252,21521821815	20780,6394807178	141,514334437278
392,424329369815	11480,3254589504	2210,90347587994
755,731556681096	3609,18851062383	8852,78037670174
298,253210631933	4988,84128668432	6989,40071693460
43,0304589187219	6833,77515384283	5129,61467560352
41,5436873126549	7162,81484418679	4872,57552743433
45,8496898021737	7074,87771818555	4946,09479684402
9,06439635212226	7247,12474273782	4764,71333165853
7832,09233964239	3,02488191959081	24079,7119773099
6969,49166937343	61,4084498192329	22525,9804212429
6199,15293425186	78,6170322441130	21140,9892694553
6559,53101308394	46,6735690498127	21794,3393952747
7708,42031628416	1,25555993498547	23862,5491320937
7501,45869717095	2,40255823491788	23498,1242918730
7369,84343006545	9,83530451624190	23261,6437982013
7776,70920050873	1,63706776590240	23981,9630723781
7542,68488288679	2,94160814234616	23571,3130166753
8613,76014444994	34,7182447734292	25427,4430142387
8821,12176446680	42,1634436242946	25789,8597339964
8282,38846026319	21,0255627008134	24851,6971905151
8647,06066983520	43,8244870625676	25477,8182255737
9697,39058858800	143,847885760590	27248,1813587416
9616,24176040412	132,151907899578	27114,5501192140
8741,63702282915	43,2062855691885	25645,9668925809
7599,78821122528	0,736412609711289	23670,8941757529
8795,12014171668	40,9792295222895	25741,9600593748
8054,74702603494	5,50190972415812	24464,6789634071
8287,99584975310	12,8014522547139	24871,2059224257
7613,20883840461	1,16725999029648	23688,5935736500
7579,69834328530	0,997761533793332	23633,2134432996
7328,78106493097	6,18713735288774	23183,6842796421
8037,45676633400	9,77932960162313	24436,8915546819

8365,83241502320	26,0851885372481	25005,3209379924
8329,44574331954	15,4857419221149	24942,9548430036
8329,97445895216	16,0542204405593	24944,9870896290
8688,44035214383	34,2029220142910	25564,1436556415
7780,38226009620	25,1464485261960	23976,9595531985
6967,84211813060	21,0389019865100	22541,5286058331
7565,44847334274	0,622572311265400	23607,4194086904
8679,21235571560	33,6627611841078	25546,1754397812
9049,46487718887	67,0109990427283	26177,8934838163
9049,46487718887	67,0109990427283	26177,8934838163
9810,63392422844	158,566710957704	27436,4965796275
7778,14567985100	12,8485646108317	23980,1094380987
5497,99112839588	194,444495833022	19812,7076816251
8397,95084186958	19,0869452816662	25062,5491711696
8448,75069215385	23,4713091671983	25150,9457333862
8308,30444792254	23,3793332981345	24905,7359805245
7924,66009144598	22,2052072134311	24232,0721827913
7815,22373327716	10,4389303106572	24044,5634517429
7661,63256208572	1,86037832373263	23774,8783659130
8218,38240172876	12,9459607482199	24752,3225627873
7715,08849855175	3,04864825167283	23871,6039034323
8335,57062639625	14,8394596689203	24953,0018880265
8243,02560869489	11,1230224560961	24794,6738768870
7945,69033799396	13,0510179306938	24274,1598210348
7496,19350010338	2,78803368042001	23489,1827036870
7357,57957822158	25,5998058093264	23231,7165985545
5958,20217500154	144,319411309922	20680,6220503575
6872,82356753319	22,2974565251229	22367,1854848871
7076,37297724226	12,2203445616262	22737,9875649835
7402,89611054083	5,86808956300805	23320,3927322635
8010,73380700477	4,40062615409697	24388,1975346949
8702,62831893429	36,0364747154355	25585,6461010722
7745,14657671687	1,42129970267169	23926,8793070889
9112,43775269375	68,7922226163977	26280,1230290938
8971,94292197091	63,8608386181209	26034,7317236398
9184,20981854108	86,6966681064687	26389,0844849058
9261,20892293511	83,8466330789792	26529,7904039931
9845,68235298978	161,050585445446	27496,7196613131
10035,6063179523	183,186171676171	27813,2196138145
8506,20800392742	27,5734493822650	25241,7893016318
7881,22344960073	1,79769901488851	24164,2613751561
8553,80997380857	28,3176812418228	25325,2664248900
7728,14875825803	1,47185858762871	23889,8569553670
8385,61521725144	17,1776029098502	25041,9929389329
8737,73362854729	36,1887166374404	25647,3911060768
8181,28942783088	9,20825544527127	24687,1410433860
8374,33601789644	17,9289750779681	25021,8325123254
7720,39863648334	6,04257458794594	23882,3980794377
8583,01206287736	28,5118041895133	25383,0423484372

8635,32771594831	30,7488694155180	25471,2495815208
8597,83038502762	28,3020285769345	25408,8334874984
8683,03722947921	33,4741763364673	25553,6923124233
7770,35216407789	20,3679624460625	23962,5492210965
6992,20696934792	22,6032021698616	22583,4177071177
7567,89389463483	3,46004662351905	23611,5701896620
9121,44499528222	69,7612436685967	26300,4994198417
9049,46487718887	67,0109990427283	26177,8934838163
9049,46487718887	67,0109990427283	26177,8934838163
10159,6488393985	203,461778635846	28015,7076327510
8028,90860285203	13,6320874691562	24420,4873518384
7218,82828804014	15,8971238386081	22987,3194364255
7951,20751226255	5,37889364209227	24278,2675869560
8494,90753790077	24,0255909767206	25229,7862963635
8353,14157189893	24,0511713172072	24983,8596354164
8031,64025924738	9,67918377187637	24428,1015331418
8590,21731985091	29,7904207682040	25395,9322049317
8776,89401068378	40,6573448430730	25714,4071990565
6338,55520509937	83,4851491167232	21365,5457211387
8582,75211491710	29,0725502327928	25381,6996296745
8561,67407493209	25,7516852414331	25343,3427147613
8490,39845918466	22,2499936621953	25223,4791960395
8133,15011697756	13,2708968101453	24602,9285666275

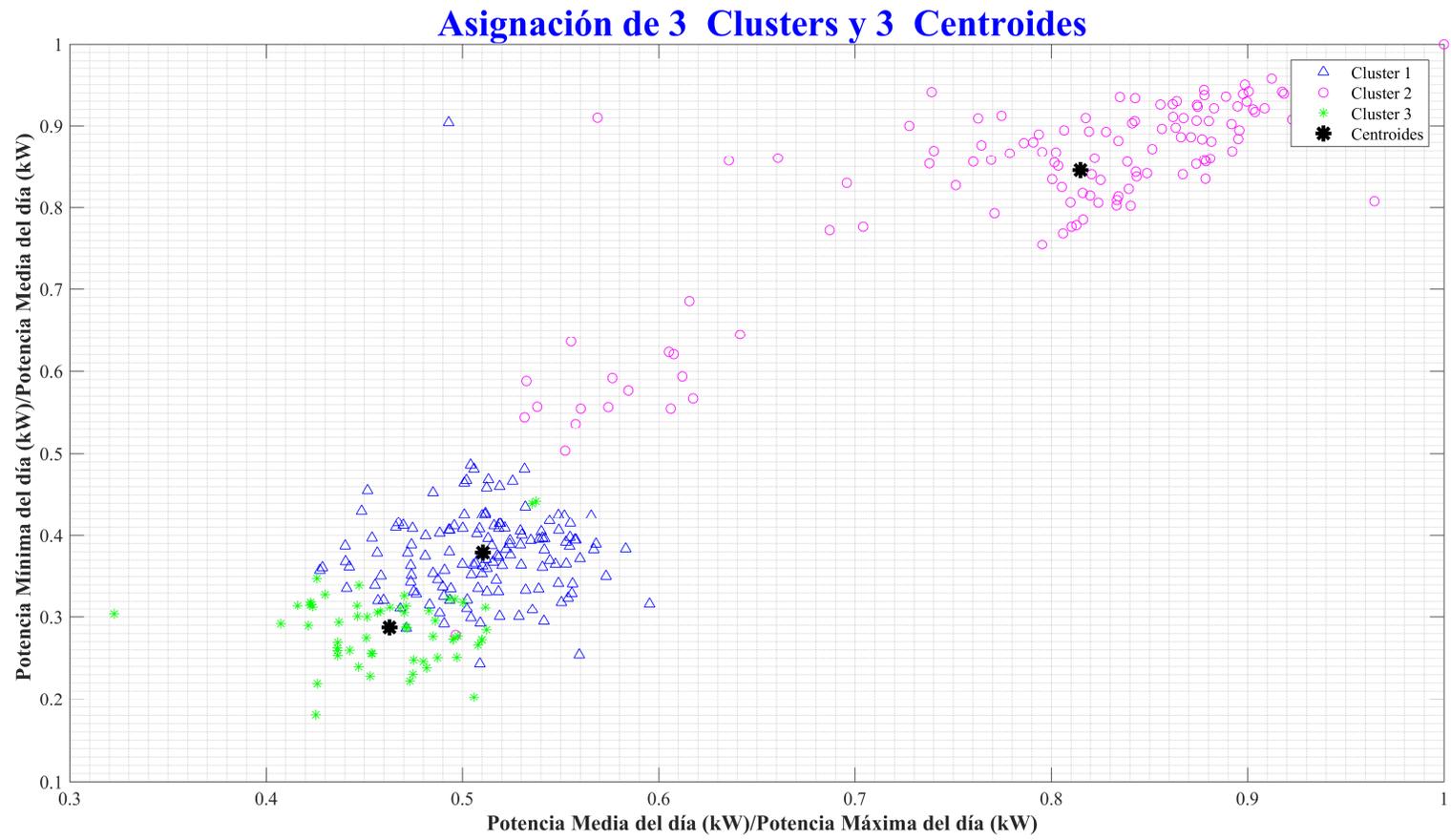


Figura 172. Segmentación de los datos en tres clusters.

Asignación de 3 Clusters y 3 Centroides

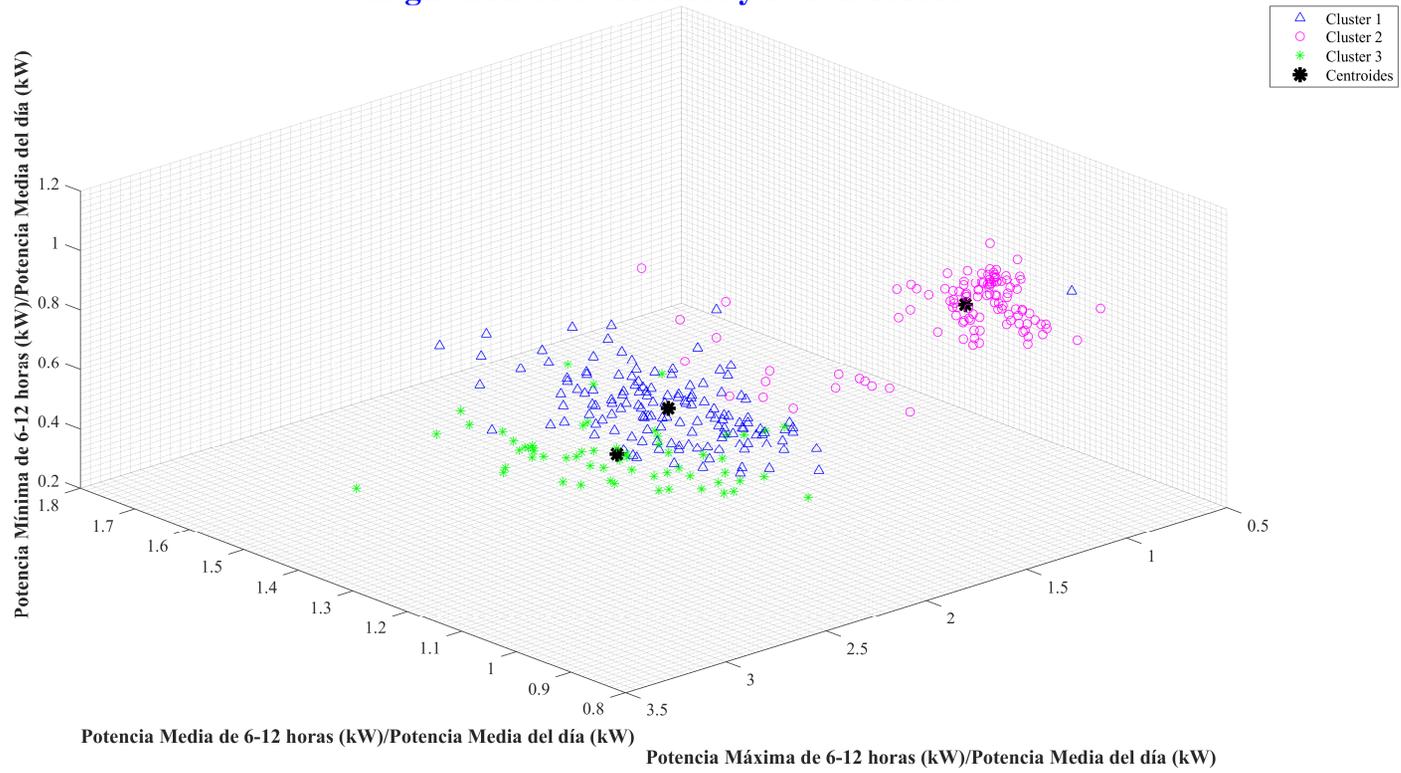


Figura 173. Segmentación de los datos en tres clusters.

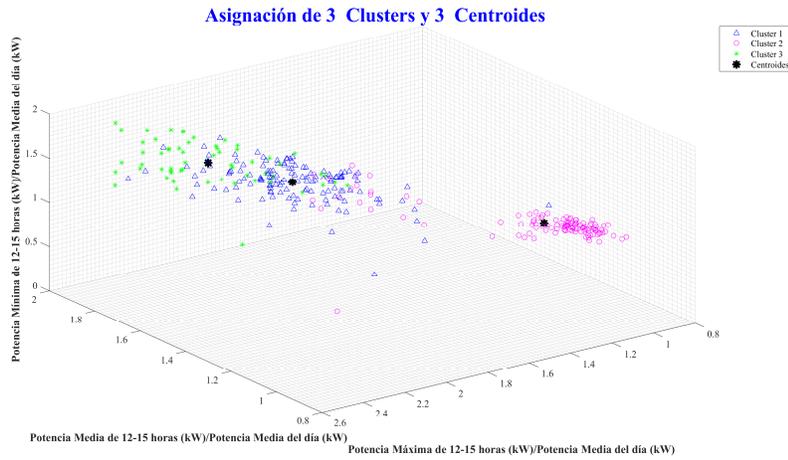


Figura 174. Segmentación de los datos en tres clusters.

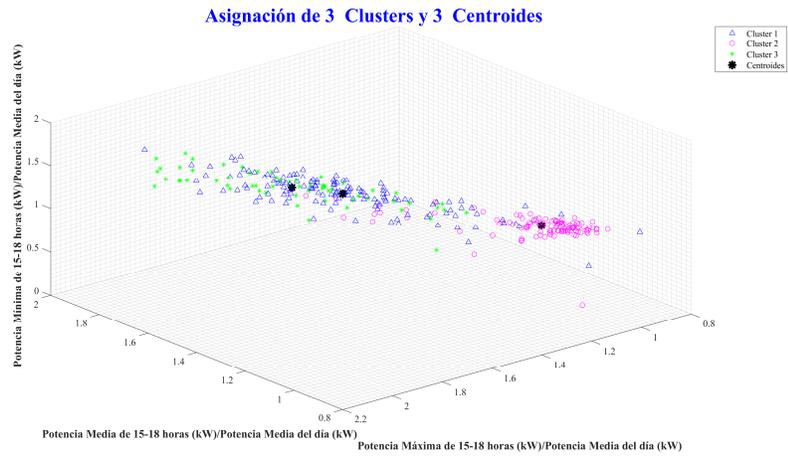


Figura 175. Segmentación de los datos en tres clusters.

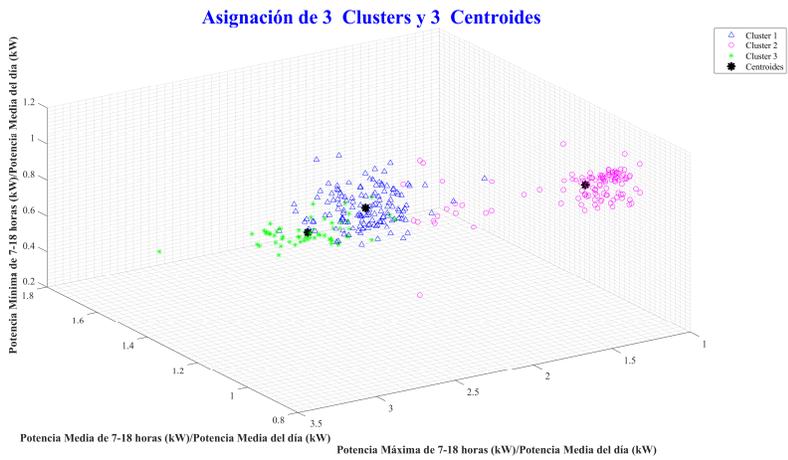


Figura 176. Segmentación de los datos en tres clusters.

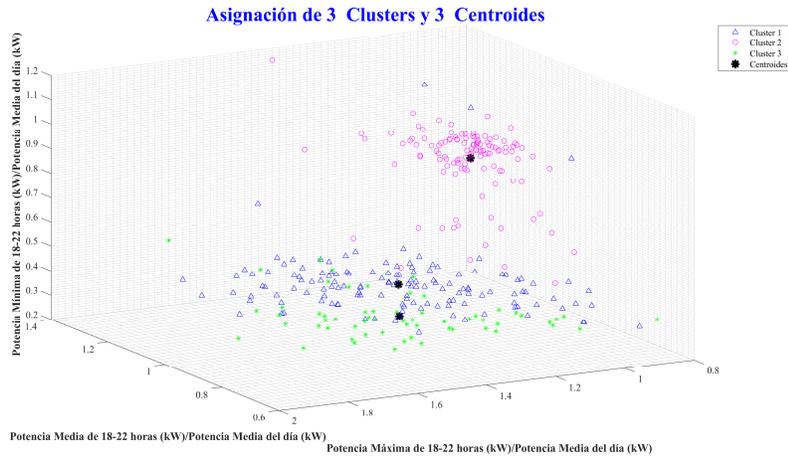


Figura 177. Segmentación de los datos en tres clusters.

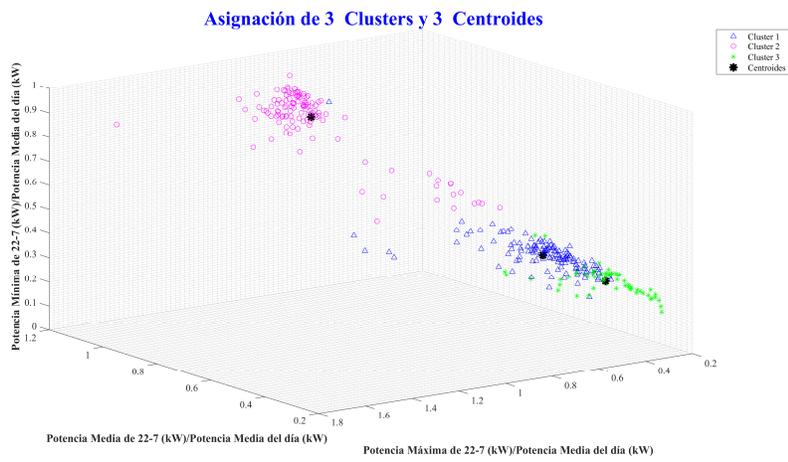


Figura 178. Segmentación de los datos en tres clusters.

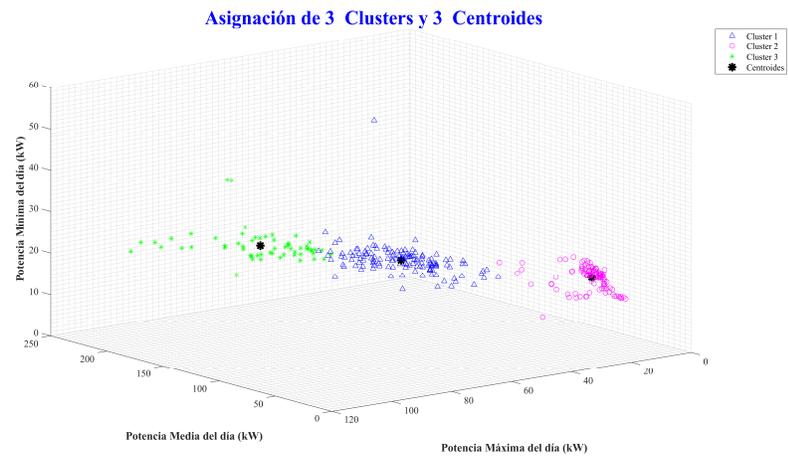


Figura 179. Segmentación de los datos en tres clusters.

4.1.4.7. Análisis estadístico de media aritmética y desviación estándar de las distancias de cada perfil de carga de energía eléctrica a su centroide y de los valores de silueta para la clasificación de dichos perfiles de carga con outliers y sin outliers

- **Análisis estadístico de media aritmética y desviación estándar para el cluster 1:**

MediaAritméticaDCDelCluster1 =

312.4399

DesviaciónEstándarDCDelCluster1 =

352.5652

RangoIntercuartilDCDelCluster1 =

396.2814

ValorMínimoDCDelCluster1 =

295.7324

AnchoDeBandaDCDelCluster1 =

99.2061

limiteatipicos1 =

861.6941

LimitesDeLosValoresAtipicosParaElCluster1medianteLaDistanciaAIC =

861.6941

MediaAritméticaVSDelCluster1 =

0.8061

DesviaciónEstándarVSDelCluster1 =

0.1888

RangoIntercuartilVSDelCluster1 =

0.1720

ValorMínimoVSDelCluster1 =

0.1284

AnchoDeBandaVSDelCluster1 =

0.0431

limiteatipicos2 =

0.4942

LimitesDeLosValoresAtipicosParaElCluster1medianteLosValoresDeSi =

0.4942

- **Análisis estadístico de media aritmética y desviación estándar para el cluster 2:**

MediaAritméticaDCDelCluster2 =

110.6753

DesviaciónEstándarDCDelCluster2 =

258.6404

RangoIntercuartilDCDelCluster2 =

58.9504

ValorMínimoDCDelCluster2 =

43.9929

AnchoDeBandaDCDelCluster2 =

14.8886

limiteatipicos1 =

196.8876

LimitesDeLosValoresAtipicosParaElCluster2medianteLaDistanciaAIC =

196.8876

MediaAritméticaVSDelCluster2 =

0.9591

DesviaciónEstándarVSDelCluster2 =

0.0977

RangoIntercuartilVSDelCluster2 =

0.0072

ValorMínimoVSDelCluster2 =

0.0053

AnchoDeBandaVSDelCluster2 =

0.0018

limiteatipicos2 =

0.9576

LimitesDeLosValoresAtipicosParaElCluster2medianteLosValoresDeSi =

0.9576

- **Análisis estadístico de media aritmética y desviación estándar para el cluster 3:**

MediaAritméticaDCDelCluster3 =

503.6380

DesviaciónEstándarDCDelCluster3 =

647.5864

RangoIntercuartilDCDelCluster3 =

543.7700

ValorMínimoDCDelCluster3 =

405.7985

AnchoDeBandaDCDelCluster3 =

161.0359

limiteatipicos1 =

1.2956e+03

LimitesDeLosValoresAtipicosParaElCluster3medianteLaDistanciaAIC =

1.2956e+03

MediaAritméticaVSDelCluster3 =

0.7418

DesviaciónEstándarVSDelCluster3 =

0.2154

RangoIntercuartilVSDelCluster3 =

0.1547

ValorMínimoVSDelCluster3 =

0.1155

AnchoDeBandaVSDelCluster3 =

0.0458

limiteatipicos2 =

0.4216

LimitesDeLosValoresAtipicosParaElCluster3medianteLosValoresDeSi =

0.4216

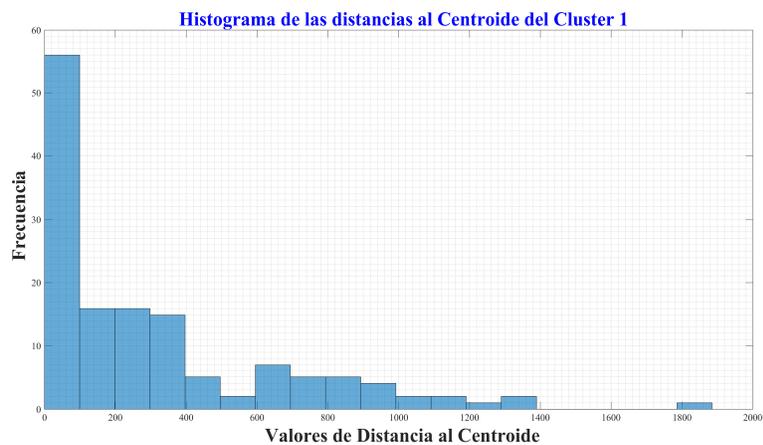


Figura 180. Histograma con las distancias al centroide del cluster 1.

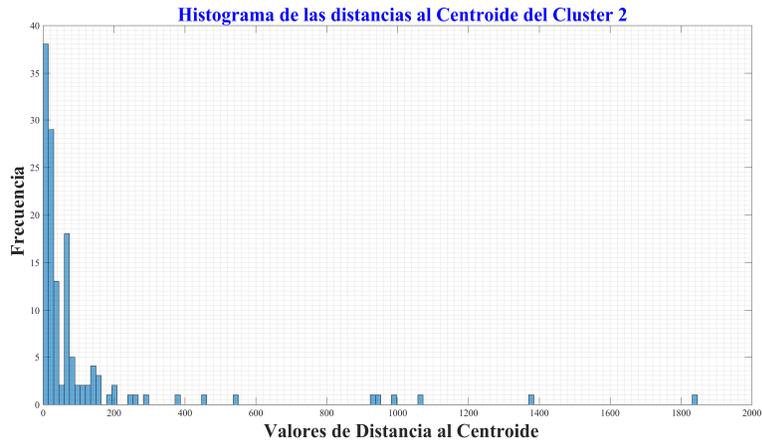


Figura 181. Histograma con las distancias al centroide del cluster 2.

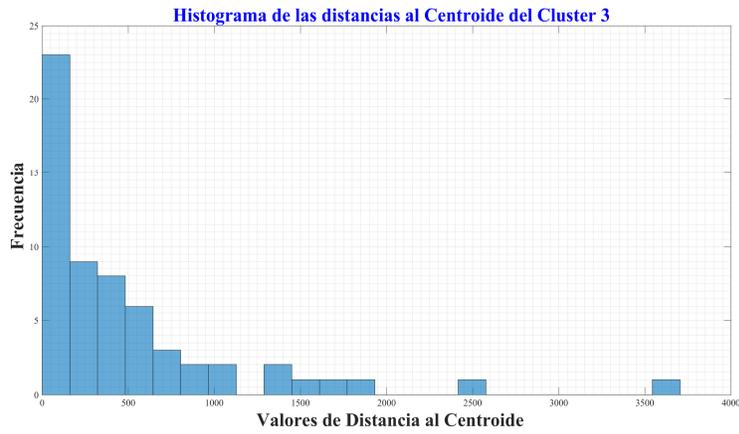


Figura 182. Histograma con las distancias al centroide del cluster 3.

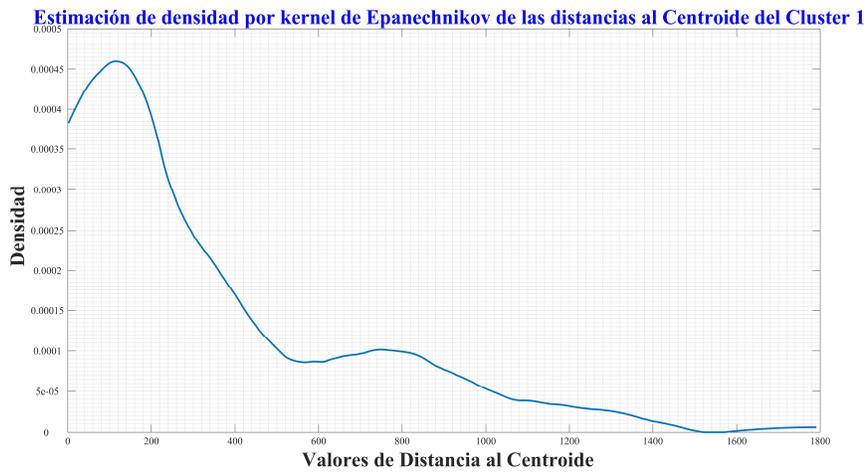


Figura 183. FDP de las distancias al centroide del cluster 1.

Estimación de densidad por kernel de Epanechnikov de las distancias al Centroides del Cluster 2



Figura 184. FDP de las distancias al centroides del cluster 2.

Estimación de densidad por kernel de Epanechnikov de las distancias al Centroides del Cluster 3

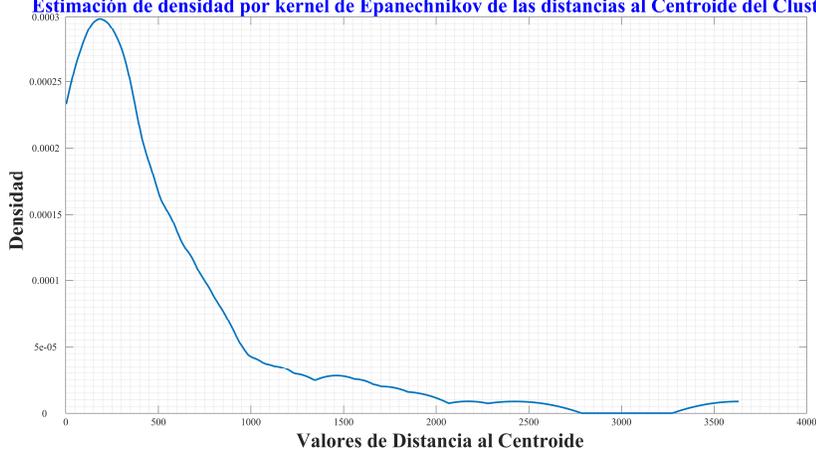


Figura 185. FDP de las distancias al centroides del cluster 3.

Función de Distribución Acumulativa Inversa de las distancias al Centroides del Cluster 1



Figura 186. FDAI de las distancias al centroides del cluster 1.



Figura 187. FDAI de las distancias al centroide del cluster 2.

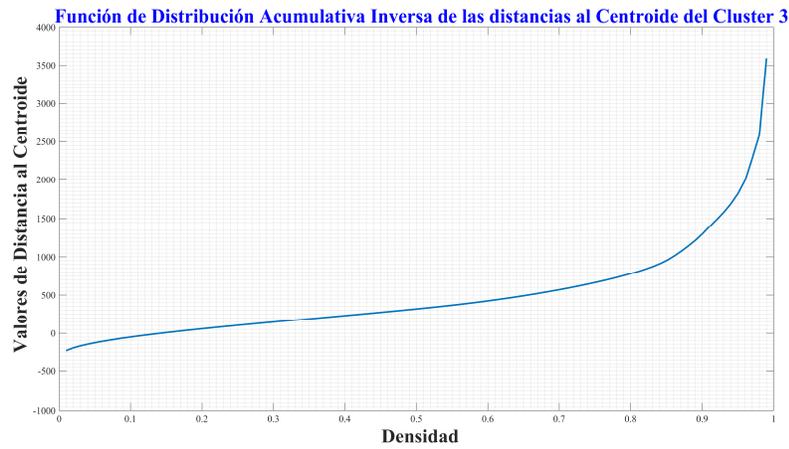


Figura 188. FDAI de las distancias al centroide del cluster 3.

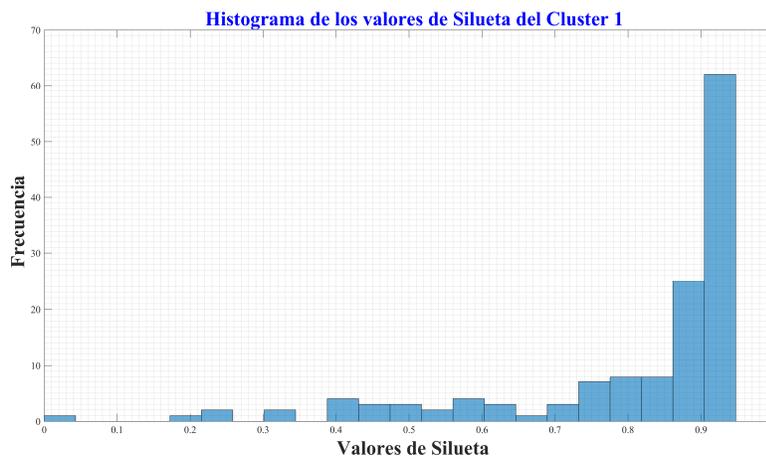


Figura 189. Histograma de los valores de Silueta del cluster 1.

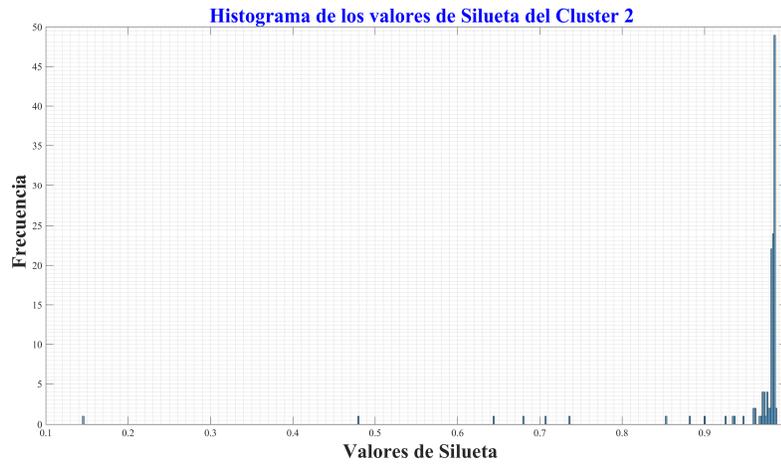


Figura 190. Histograma de los valores de Silueta del cluster 2.

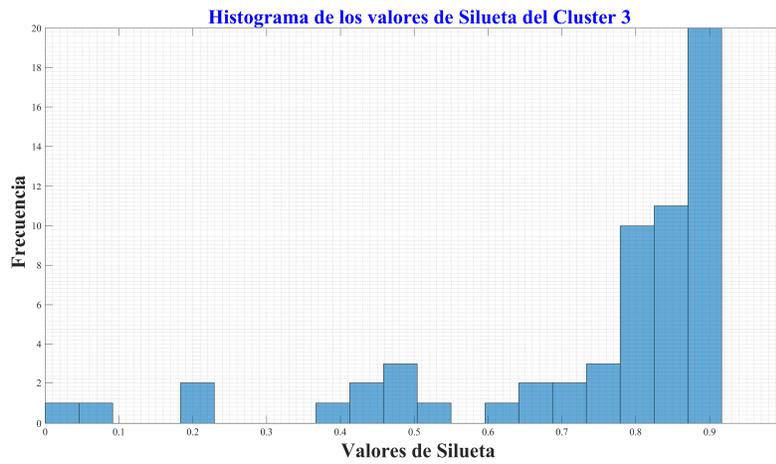


Figura 191. Histograma de los valores de Silueta del cluster 3.

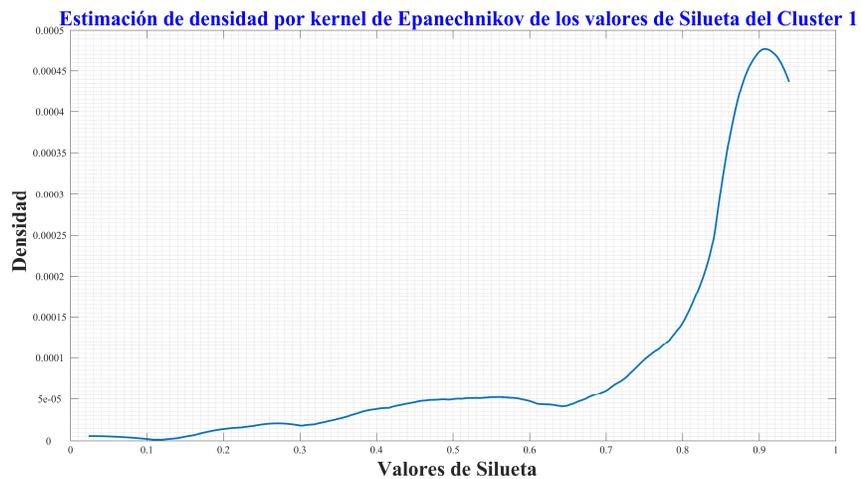


Figura 192. FDP de los valores de silueta del cluster 1.

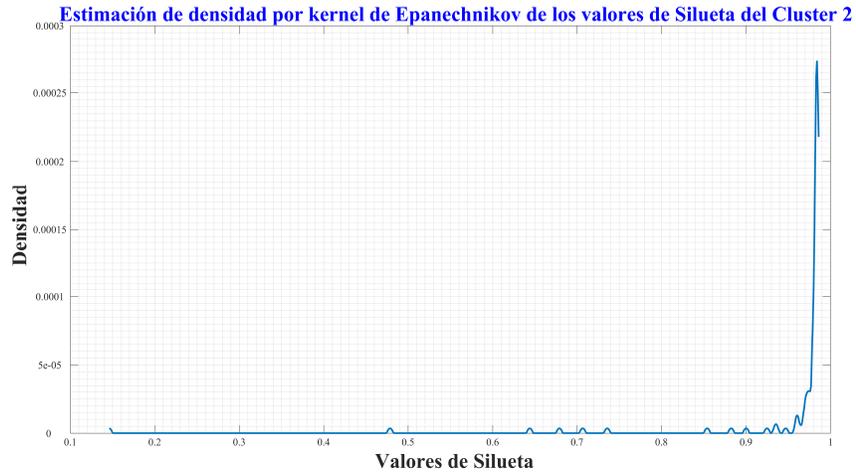


Figura 193. FDP de los valores de silueta del cluster 2.

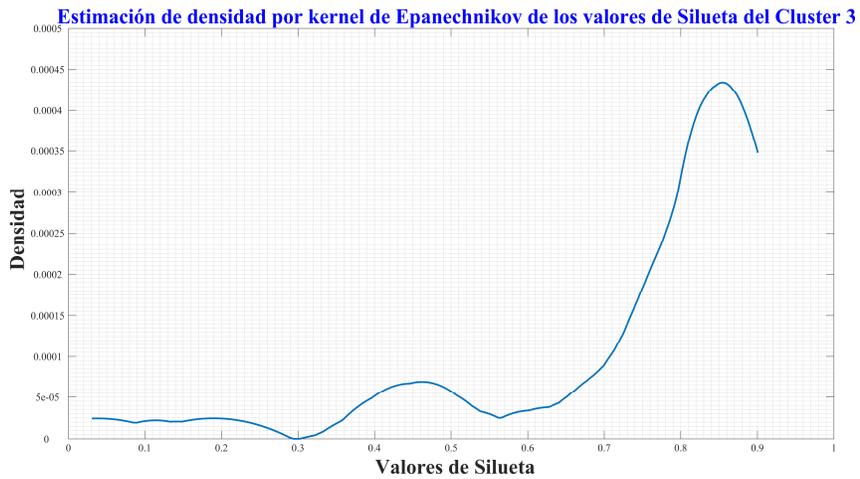


Figura 194. FDP de los valores de silueta del cluster 3.

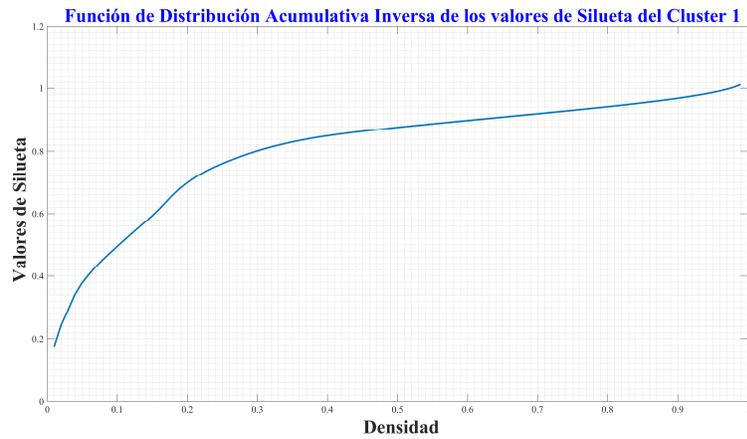


Figura 195. FDAI de los valores de silueta del cluster 1.



Figura 196. FDAI de los valores de silueta del cluster 2.



Figura 197. FDAI de los valores de silueta del cluster 3.

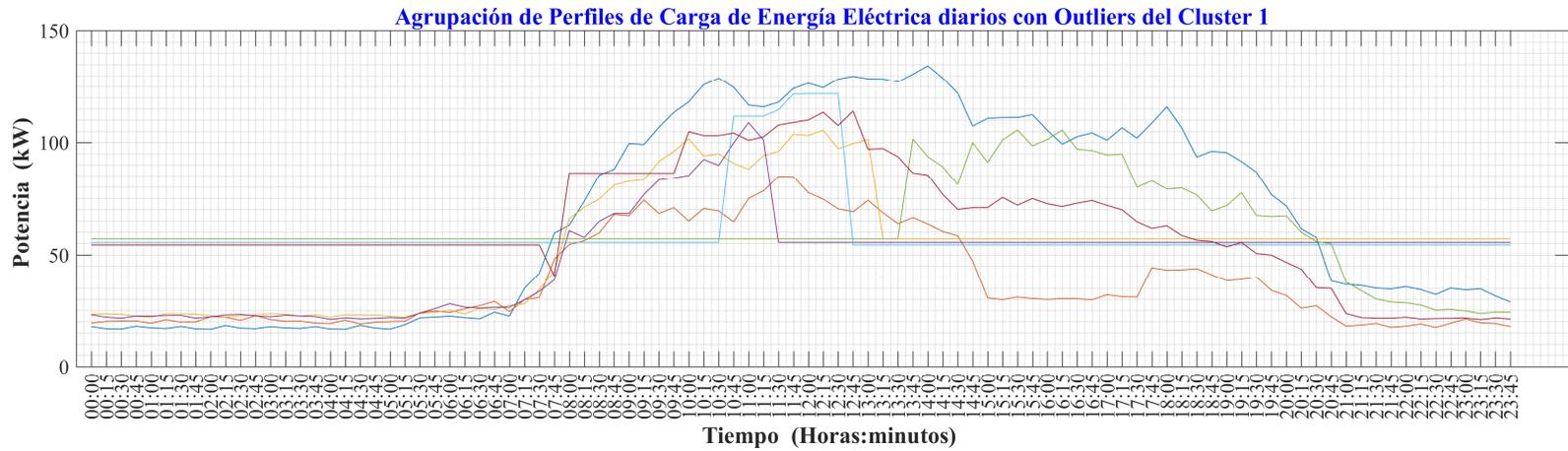
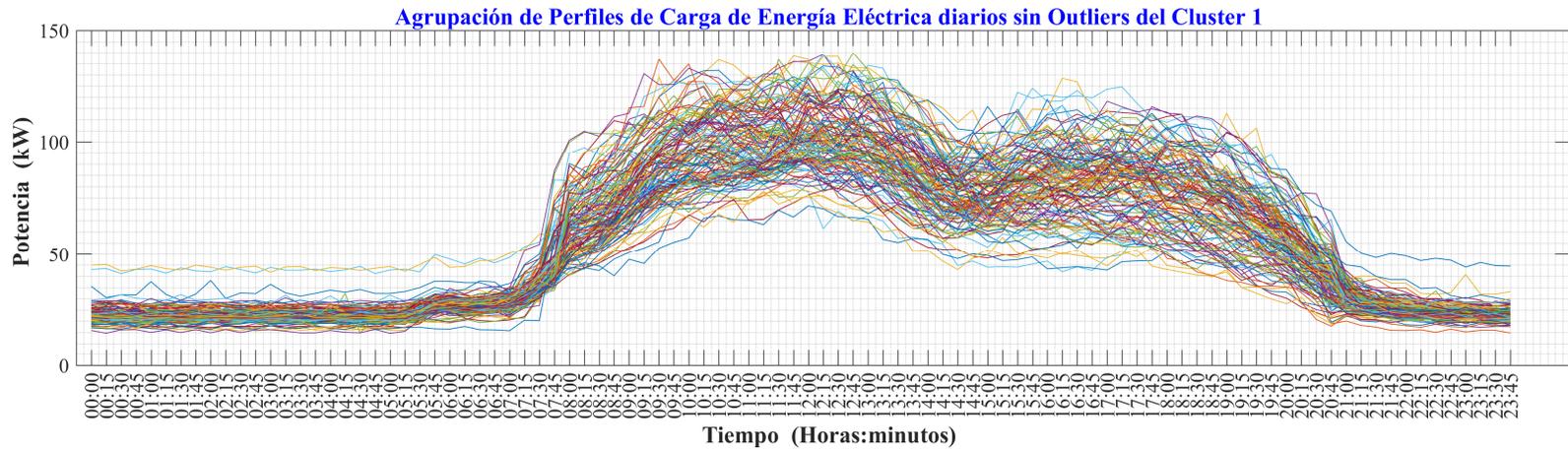


Figura 198. PCEE del cluster 1 (Perfiles sin outliers, 132 - Perfiles con outliers, 7).

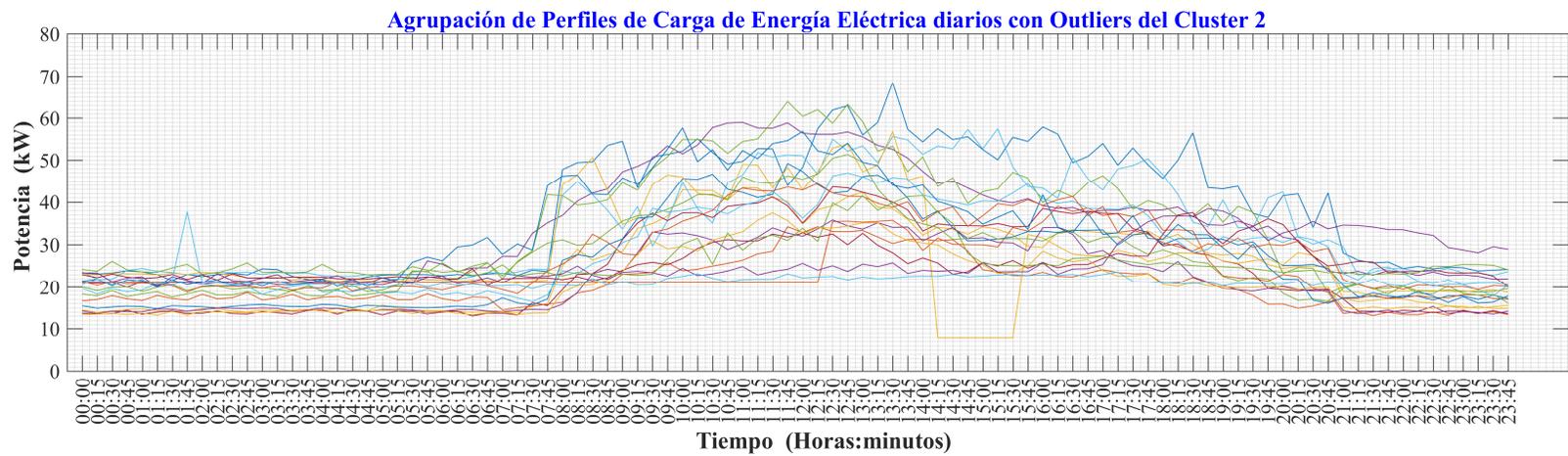
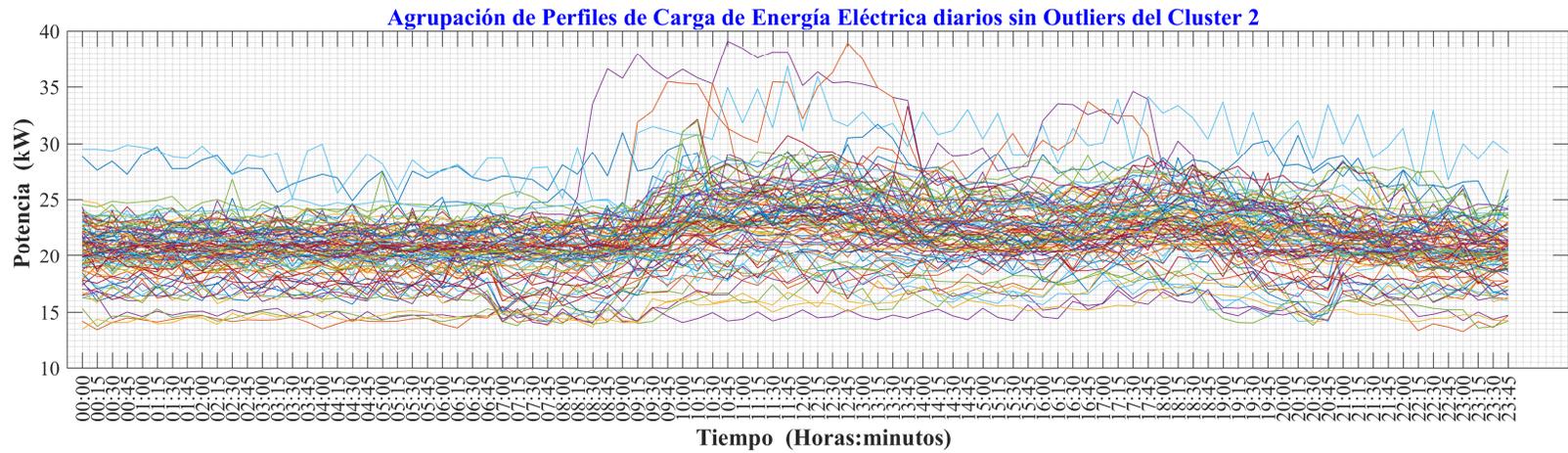


Figura 199. PCEE del cluster 2 (Perfiles sin outliers, 113 - Perfiles con outliers, 20).

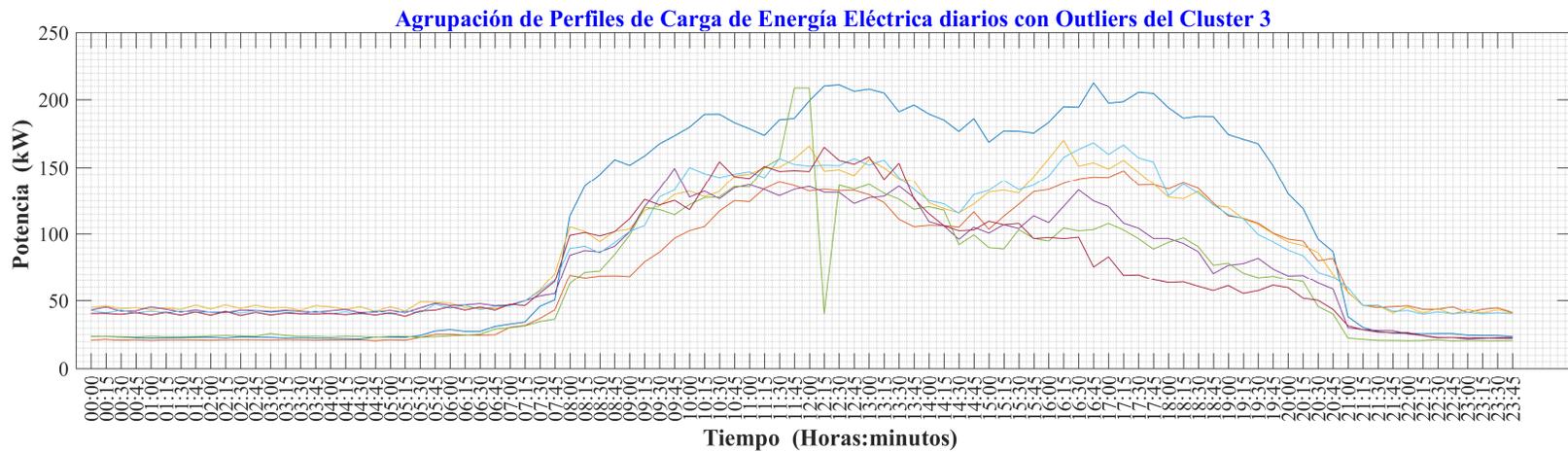
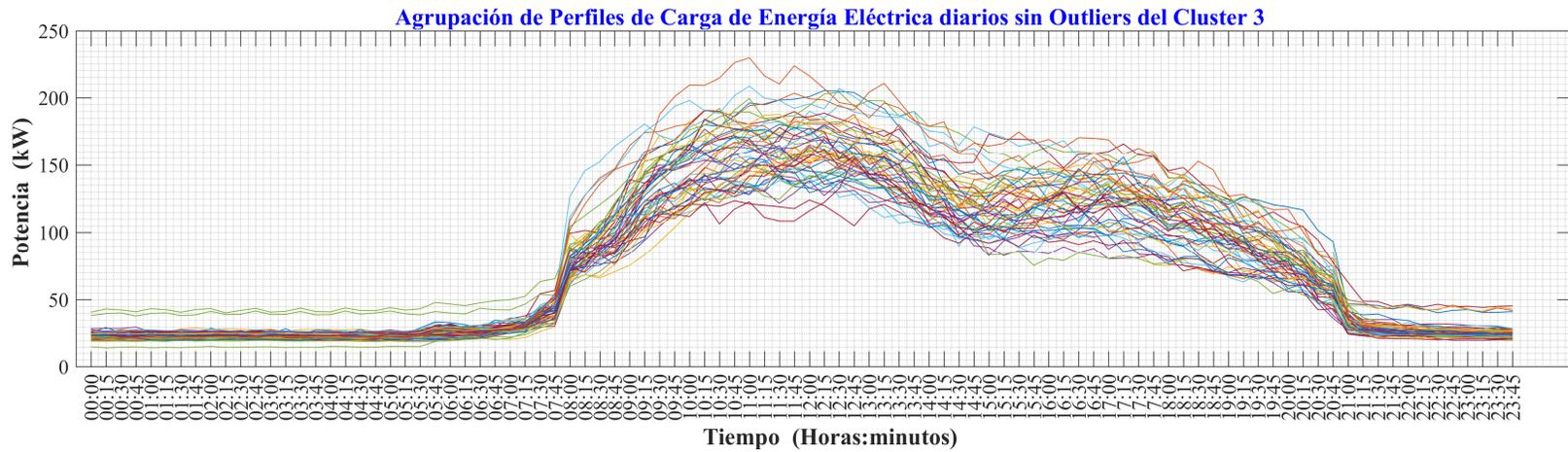


Figura 200. PCEE del cluster 3 (Perfiles sin outliers, 53 - Perfiles con outliers, 7).

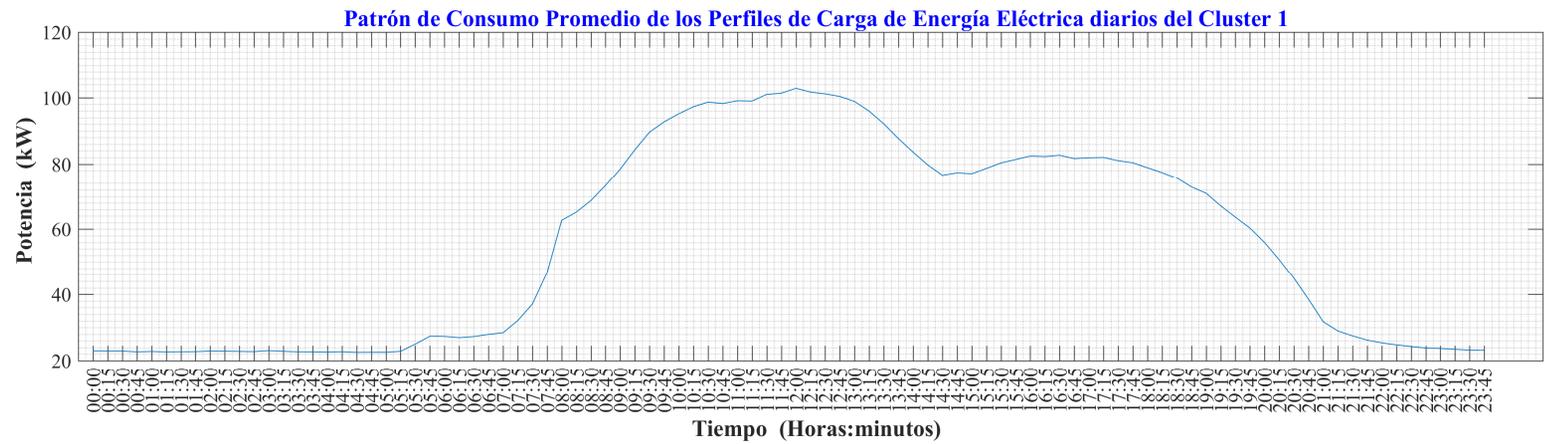
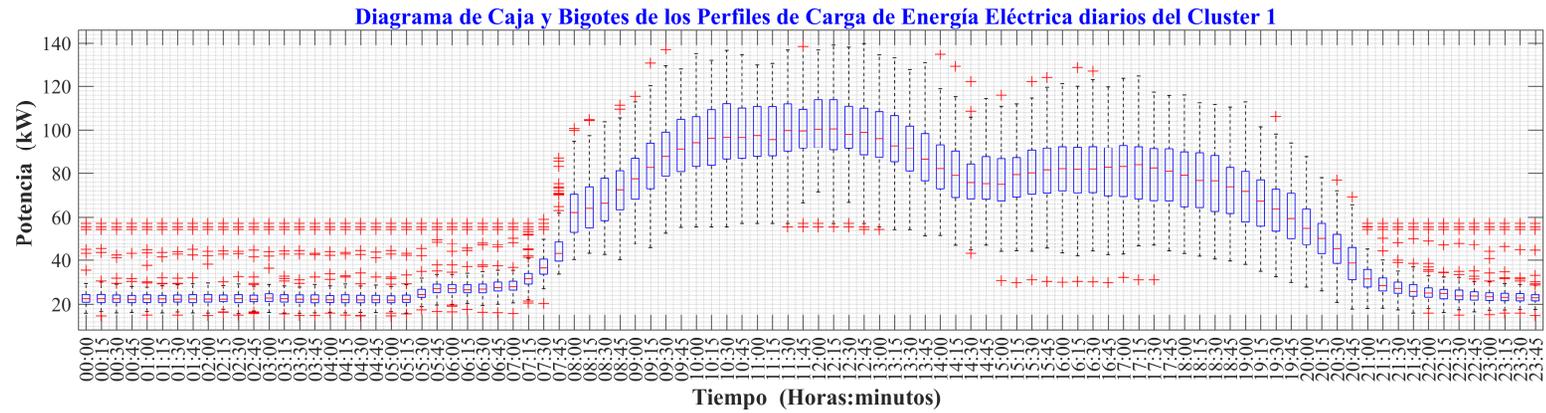


Figura 201. Patrón de consumo de PCEE diarios del cluster 1.

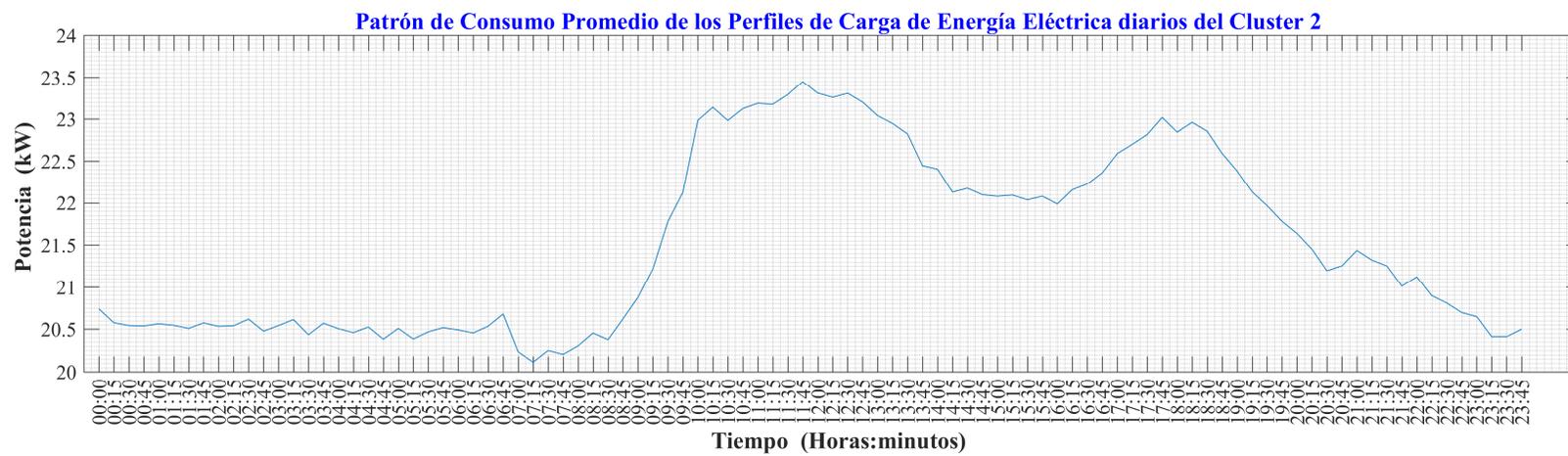
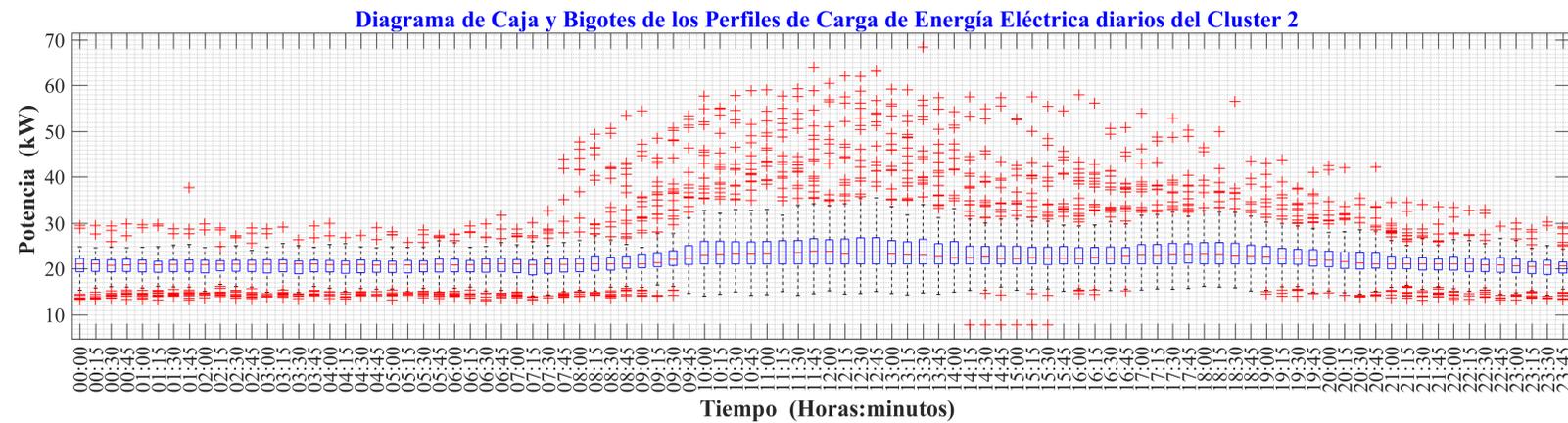


Figura 202. Patrón de consumo de PCEE diarios del cluster 2.

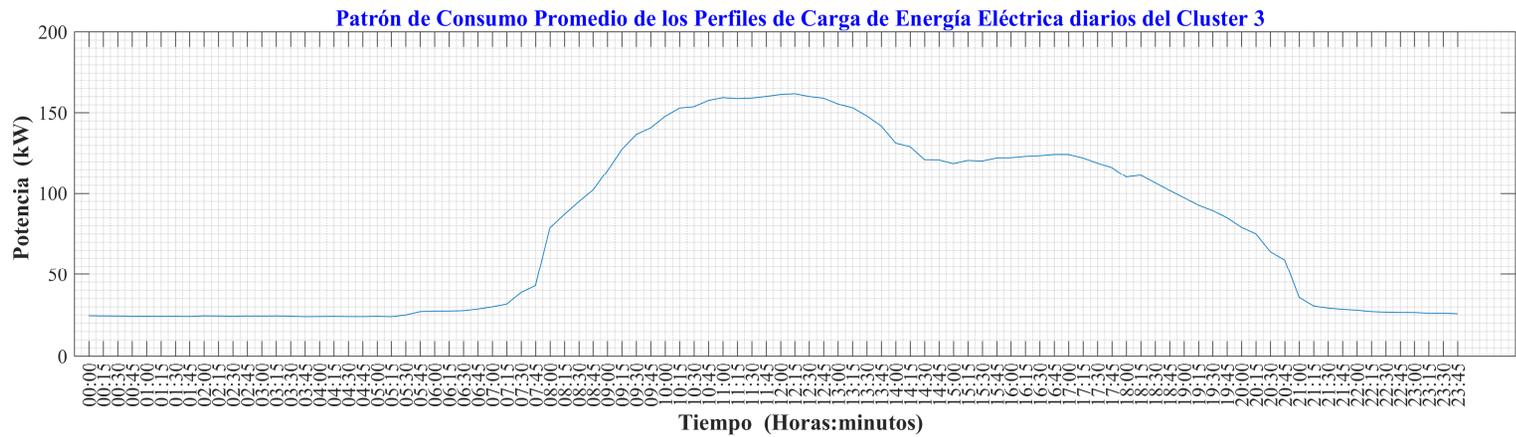
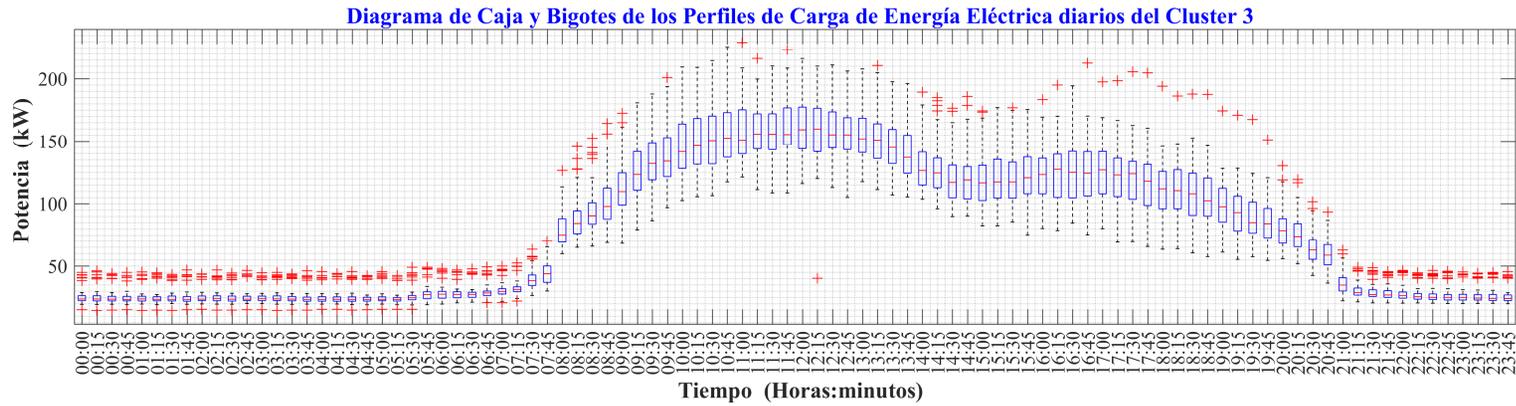


Figura 203. Patrón de consumo de PCEE diarios del cluster 3.

4.2. Análisis de resultados de segmentación mediante el software WEKA

El software WEKA es un conjunto de algoritmos de aprendizaje automático que son aplicados en procesos de minería de datos. WEKA posee herramientas que sirven para el preprocesamiento de datos, clasificación, agrupación, reglas de asociación y visualización.

Se utiliza WEKA para analizar los resultados obtenidos de aplicar las técnicas de minería de datos a las bases de datos de energía de tres clientes diferentes.

4.2.1. Análisis de resultados de la acometida principal de la UPS

Primero se muestran los resultados obtenidos de WEKA al aplicar la técnica **k-Means** a la base de datos de CEE de la acometida principal de la UPS:

=== **Run information** ===

Scheme: weka.clusterers.SimpleKMeans:

Relation:

FACTORES_DE_FORMA_DE_LOS_PERFILES_DE_CARGA_DIARIOS_ACOMETIDA_PRINCIPAL_UPS

Instances: 126

Attributes: 30

=== Clustering model (full training set) ===

kMeans

=====

Number of iterations: 2

Within cluster sum of squared errors: 240587.00295535714

Initial starting points (k-means++):

Cluster 0:

0.546888,0.384833,1.175184,1.48696,0.429237,1.064934,1.287712,0.876692,1.427755,1.681654,1.338947,1.272781,1.681654,0.876692,1.452661,1.828528,0.808378,0.465418,0.819764,0.384833,513.92,281.056667,108.16

Cluster 1:

0.81005,0.894483,0.979158,1.046179,0.923253,0.940689,0.980793,0.920638,0.950933,1.095872,0.912791,0.965041,1.095872,0.912791,1.178259,1.234491,1.132489,0.963502,1.132489,0.894483,151.04,122.35,109.44

Tabla 26. Centroides de los Clusters generados de la base de datos de la acometida principal de la UPS.

CLUSTER			
Atributo	Full Data	1	2
	126.0	87.0	39.0

F1	0.6096	0.5541	0.729
F2	0.5083	0.3858	0.7716
F3	1.1456	1.1851	1.0606
F4	1.431	1.5282	1.2221
F5	0.5609	0.4406	0.8195
F6	1.0515	1.1014	0.9443
F7	1.2338	1.293	1.1064
F8	0.9195	0.9464	0.8615
F9	1.272	1.4243	0.9446
F10	1.4788	1.6414	1.1292
F11	1.1616	1.2919	0.8813
F12	1.2	1.2861	1.0151
F13	1.5336	1.6554	1.2719
F14	0.8587	0.8775	0.8182
F15	1.3457	1.4068	1.2145
F16	1.6441	1.8071	1.2936
F17	0.92	0.8172	1.1411
F18	0.6018	0.4696	0.8862
F19	0.9588	0.8388	1.2166
F20	0.5163	0.386	0.7965
F21	417.8717	519.9888	198.32
F22	240.5006	287.3957	139.6761
F23	109.1251	110.5749	106.008

Time taken to build model (full training data): 0 seconds

Tabla 27. Instancias agrupadas para la base de datos de la acometida principal de la UPS en cada cluster.

Cluster	Instancias Agrupadas
1	87 (69%)
2	39 (31%)

4.2.2. Análisis de resultados de la del Edificio Cornelio Merchán - UPS

En el marco del análisis de resultados y como parte del proceso de implementación de medidores inteligentes en la UPS, existe la necesidad de ubicar uno de ellos en el edificio Cornelio Merchán donde se encuentran todos los laboratorios e instalaciones que consumen la mayor cantidad de energía eléctrica.

Por ello el siguiente análisis se refiere a la base de datos obtenidos del CEE de este edificio.

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans

Relation:FACTORES_DE_FORMA_DE_LOS_PERFILES_DE_CARGA_DIARIOS_CORNELIO_MERC
HAN

Instances: 127

Attributes: 30

=== Clustering model (full training set) ===

kMeans

=====

Number of iterations: 2

Within cluster sum of squared errors: 13189.074347320751

Initial starting points (k-means++):**Cluster 0:**0.498461,0.273167,1.364453,1.86572,0.333994,1.193678,1.61799,0.862633,1.721948,1.938713,1.
528409,1.500836,1.938713,0.862633,1.159232,2.006175,0.52311,0.317098,0.598313,0.273167,10
1.584,50.635667,13.832**Cluster 1:**0.744229,0.877967,0.950654,1.06883,0.900871,0.933317,1.02684,0.877967,0.98994,1.141358,0.9
42861,0.956134,1.141358,0.877967,1.1812,1.343672,1.13754,0.97308,1.17953,0.904688,19.712,1
4.67025,12.88**Tabla 28. Centroides de los Clusters generados de la base de datos del edificio CORNELIO
MERCHÁN de la UPS.**

CLUSTER			
Atributo	Full Data	1	2
	127.0	87.0	40.0
F1	0.5497	0.5064	0.644
F2	0.4208	0.2727	0.7429
F3	1.3271	1.4266	1.1108
F4	1.7662	1.9342	1.4007
F5	0.4908	0.352	0.7927
F6	1.0589	1.1162	0.9342
F7	1.3954	1.5198	1.1249
F8	0.8154	0.8143	0.8179
F9	1.4572	1.6861	0.9594
F10	1.6471	1.8601	1.1836
F11	1.2804	1.481	0.844
F12	1.3545	1.5	1.0381
F13	1.8224	1.9769	1.4862
F14	0.7934	0.7983	0.7828
F15	1.1143	1.1043	1.136
F16	1.601	1.7611	1.253
F17	0.6732	0.495	1.0608
F18	0.5159	0.3425	0.893

F19	0.8582	0.6916	1.2206
F20	0.4329	0.2735	0.7797
F21	71.1526	91.6591	26.551
F22	36.8903	46.3272	16.365
F23	12.3667	12.5324	12.0064

Time taken to build model (full training data): 0.01 seconds

Tabla 29. Instancias agrupadas para la base de datos del edificio Cornelio Merchán de la UPS en cada cluster.

Cluster	Instancias Agrupadas
1	87 (69%)
2	40 (31%)

4.2.3. Análisis de resultados de CONTINENTAL GENERAL TIRE ANDINA.

Se puede analizar un cliente del tipo industrial en donde el CEE es mayor a comparación a los otros tipos de clientes, los datos son mostrados al final en la parte de Anexos.

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans

Relation:

FACTORES_DE_FORMA_DE_LOS_PERFILES_DE_CARGA_DIARIOS_LLANTERA

Instances: 366

Attributes: 30

=== Clustering model (full training set) ===

kMeans

=====

Number of iterations: 6

Within cluster sum of squared errors: 4.314990462206023E8

Initial starting points (k-means++):

Cluster 0:

0.662819,0.176623,1.192209,1.423244,0.848932,1.181098,1.418686,0.791957,1.140361,1.304735,0.976557,1.179648,1.423244,0.791957,1.177894,1.333223,0.869443,0.701366,1.508707,0.176623,5296,3510.291667,620

Cluster 1:

0.789034,0.212109,1.090066,1.24537,0.817632,1.056364,1.267373,0.767465,1.051157,1.166159,0.814111,1.066386,1.267373,0.767465,0.94932,1.24185,0.398694,0.941386,1.248011,0.212109,5760,4544.833333,964

Tabla 30. Centroides de los Clusters generados de la base de datos de CONTINENTAL GENERAL TIRE ANDINA C.A.

CLUSTER			
Atributo	Full Data	1	2
	366.0	29.0	337.0
F1	0.7572	0.6734	0.7644
F2	0.5838	0.6637	0.577
F3	1.0258	1.0028	1.0278
F4	1.2119	1.2933	1.2049
F5	0.7548	0.7755	0.753
F6	0.9657	0.9361	0.9683
F7	1.1437	1.0317	1.1534
F8	0.7078	0.8308	0.6972
F9	1.0316	0.9375	1.0397
F10	1.1398	1.0927	1.1439
F11	0.8897	0.7577	0.9011
F12	1.0103	0.9635	1.0144
F13	1.2459	1.2519	1.2454
F14	0.679	0.676	0.6792
F15	0.99	0.9725	0.9915
F16	1.1751	1.09	1.1824
F17	0.7111	0.9137	0.6936
F18	0.9918	1.0568	0.9862
F19	1.2864	1.5181	1.2665
F20	0.6262	0.8273	0.6089
F21	5590.1421	997.1034	5985.3887
F22	4263.7919	603.7356	4578.7522
F23	2498.0546	395.3103	2679.003

Time taken to build model (full training data): 0.01 seconds

=== Model and evaluation on training set ===

Tabla 31. Instancias agrupadas para la base de datos del edificio Cornelio Merchán de la UPS en cada cluster.

Cluster	Instancias Agrupadas
1	29 (8%)
2	337 (92%)

4.2.4. Análisis de resultados de la UPV

Por último, podemos analizar la base de datos proporcionada por el tutor de nuestro proyecto, con el fin de comprobar de la efectividad de las TC se analizará los datos de la UPV.

=== **Run information** ===

Scheme: weka.clusterers.SimpleKMeans

Relation: FACTORES_DE_FORMA_DE_LOS_PERFILES_DE_CARGA_DIARIOS_UPV

Instances: 332

Attributes: 30

=== Clustering model (full training set) ===

kMeans

=====

Number of iterations: 13

Within cluster sum of squared errors: 88367.24130918454

Initial starting points (k-means++):

Cluster 0:

0.892231,0.869151,1.010093,1.120786,0.932783,1.01172,1.083185,0.958814,1.010033,1.052816,0.953029,1.013079,1.120786,0.932783,1.014311,1.094755,0.940014,0.977655,1.074508,0.869151,31,27.659167,24.04

Cluster 1:

0.456641,0.378403,1.422332,2.189906,0.61108,1.508043,1.876717,1.284154,1.297707,1.494289,1.088512,1.483841,2.189906,0.707694,0.806723,1.281739,0.458914,0.49454,0.648116,0.378403,108.8,49.6825,18.8

Cluster 2:

0.87688,0.883627,1.039788,1.140407,0.964815,1.045373,1.121526,0.981807,1.008555,1.070548,0.968591,1.037377,1.140407,0.968591,0.986056,1.072436,0.913836,0.960514,1.02901,0.883627,24.16,21.185417,18.72

Tabla 32. Centroides de los Clusters generados de la base de datos de la UPV.

CLUSTER				
Atributo	Full Data	1	2	3
	332.0	139.0	60.0	133.0
F1	0.6239	0.5104	0.463	0.815
F2	0.5492	0.3789	0.2871	0.8454
F3	1.192	1.2983	1.3216	1.0223
F4	1.6718	1.9301	2.1234	1.1981
F5	0.6176	0.4869	0.3551	0.8727
F6	1.4289	1.5913	1.7749	1.103

F7	1.6407	1.8758	2.0814	1.1962
F8	1.2045	1.2979	1.4204	1.0096
F9	1.3052	1.437	1.5313	1.0653
F10	1.4225	1.5768	1.6892	1.1409
F11	1.1844	1.2947	1.3519	0.9936
F12	1.3367	1.4876	1.5885	1.0653
F13	1.7149	1.9681	2.1743	1.243
F14	0.6728	0.588	0.4157	0.8774
F15	0.942	0.9138	0.8823	0.9983
F16	1.2887	1.3957	1.419	1.118
F17	0.6194	0.4669	0.3554	0.898
F18	0.6143	0.4423	0.333	0.9209
F19	0.7099	0.559	0.4173	0.9997
F20	0.5536	0.3791	0.2871	0.8561
F21	88.4225	109.5354	171.8753	28.7092
F22	46.6368	55.7072	79.332	22.4075
F23	20.3273	20.9047	22.5927	18.7019

Time taken to build model (full training data): 0.01 seconds

=== Model and evaluation on training set ===

Tabla 33. Instancias agrupadas para la base de datos del edificio Cornelio Merchán de la UPS en cada cluster.

Cluster	Instancias Agrupadas
1	139 (42%)
2	60 (18%)
3	133 (40%)

4.3. Análisis de resultados de segmentación mediante el software SPSS

4.3.1. Análisis de resultados de la acometida principal de la UPS

Tabla 34. Centros de clusters finales de la acometida principal de la UPS.

Centros de clusters finales		
	Cluster	
	1	2
F1	0.734980253000000	0.553407163000000
F2	0.779474651000000	0.386739445000000
F3	1.068981025000000	1.179876096000000
F4	1.223295049000000	1.524142608000000
F5	0.828457152000000	0.440981145000000
F6	0.937657407000000	1.102503974000000

F7	1.101811216000000	1.292908034000000
F8	0.856797694000000	0.947553351000000
F9	0.942716126000000	1.419635689000000
F10	1.126434356000000	1.636744669000000
F11	0.883466699000000	1.286246401000000
F12	1.016488691000000	1.282327557000000
F13	1.271591047000000	1.651113730000000
F14	0.826339797000000	0.873196398000000
F15	1.202724400000000	1.409840637000000
F16	1.274908056000000	1.809577832000000
F17	1.136947798000000	0.822778224000000
F18	0.889747422000000	0.472781591000000
F19	1.200598103000000	0.850353176000000
F20	0.804979307000000	0.386908325000000
F21	191.35999999999999	519.411494300000000
F22	137.30495729999998	286.760728000000000
F23	105.95	110.55

Tabla 35. Número de casos en cada cluster.

Número de casos en cada cluster		
Cluster	1	39.000
	2	87.000
Válidos		126.000
Perdidos		0.000

4.3.2. Análisis de resultados de la del Edificio Cornelio Merchán-UPS

Tabla 36. Centros de clusters finales de la base de datos del Edificio Cornelio Merchán.

Centros de clusters finales		
	Cluster	
	1	2
F1	0.643984917000000	0.506359155000000
F2	0.742875759000000	0.272679232000000
F3	1.110796770000000	1.426596887000000
F4	1.400715350000000	1.934212862000000
F5	0.792673964000000	0.352021380000000
F6	0.934189005000000	1.116188474000000
F7	1.124872391000000	1.519825699000000
F8	0.817859946000000	0.814329529000000

F9	0.959420370000000	1.686121271000000
F10	1.183630124000000	1.860118485000000
F11	0.843987394000000	1.481038241000000
F12	1.038094722000000	1.500032235000000
F13	1.486222833000000	1.976899861000000
F14	0.782793139000000	0.798297240000000
F15	1.136036843000000	1.104332835000000
F16	1.252977217000000	1.761063868000000
F17	1.060823358000000	0.494982400000000
F18	0.892978965000000	0.342479342000000
F19	1.220610308000000	0.691570039000000
F20	0.779689165000000	0.273515040000000
F21	26.551000000000002	91.659126440000010
F22	16.365037500000000	46.327193490000000
F23	12.006400000000001	12.532413790000000

Tabla 37. Número de casos en cada cluster.

Número de casos en cada cluster		
Cluster	1	40.000
	2	87.000
Válidos		127.000
Perdidos		0.000

4.3.3. Análisis de resultados de CONTINENTAL GENERAL TIRE ANDINA.

Tabla 38. Centros de clusters finales de la base de datos de CONTINENTAL GENERAL TIRE ANDINA.

Centros de clusters finales		
	Cluster	
	1	2
F1	0.673368701000000	0.764413516000000
F2	0.663651767000000	0.576976526000000
F3	1.002802331000000	1.027799442000000
F4	1.293297739000000	1.204918271000000
F5	0.775475048000000	0.752974956000000
F6	0.936111366000000	0.968298097000000
F7	1.031687309000000	1.153359922000000
F8	0.830755094000000	0.697219421000000
F9	0.937485231000000	1.039677897000000

F10	1.092669811000000	1.143903876000000
F11	0.757653336000000	0.901094408000000
F12	0.963535278000000	1.014353549000000
F13	1.251946511000000	1.245388902000000
F14	0.675956191000000	0.679244215000000
F15	0.972477849000000	0.991528674000000
F16	1.089993023000000	1.182441737000000
F17	0.913718775000000	0.693625009000000
F18	1.056800060000000	0.986221807000000
F19	1.518113118000000	1.266505930000000
F20	0.827286850000000	0.608894550000000
F21	997	5985
F22	603.735632199999990	4578.752226000000000
F23	395	2679

Tabla 39. Número de casos en cada cluster.

Número de casos en cada cluster		
Cluster	1	29.000
	2	337.000
Válidos		366.000
Perdidos		0.000

4.3.4. Análisis de resultados de la UPV

Tabla 40. Centros de clusters finales de la base de datos de la UPV.

Centros de clusters finales			
	Cluster		
	1	2	3
F1	0.510435345000000	0.814962019000000	0.463009255000000
F2	0.378903738000000	0.845380261000000	0.287082262000000
F3	1.298333387000000	1.022349683000000	1.321555193000000
F4	1.930104187000000	1.198074465000000	2.123422855000000
F5	0.486865210000000	0.872746453000000	0.355125233000000
F6	1.591323996000000	1.103012911000000	1.774863307000000
F7	1.875824607000000	1.196176295000000	2.081370476000000
F8	1.297862013000000	1.009582800000000	1.420382506000000
F9	1.436978625000000	1.065341738000000	1.531324090000000
F10	1.576830545000000	1.140948349000000	1.689198586000000

F11	1.294696547000000	0.993557494000000	1.351860192000000
F12	1.487620284000000	1.065300399000000	1.588547251000000
F13	1.968099330000000	1.242988795000000	2.174307977000000
F14	0.587977093000000	0.877429309000000	0.415710345000000
F15	0.913848850000000	0.998348458000000	0.882306837000000
F16	1.395684878000000	1.118005490000000	1.419029717000000
F17	0.466869935000000	0.898021967000000	0.355371183000000
F18	0.442309053000000	0.920922420000000	0.332972544000000
F19	0.559029046000000	0.999694888000000	0.417257586000000
F20	0.379058547000000	0.856135214000000	0.287082262000000
F21	109.535420699999990	28.709185800000000	171.875333300000000
F22	55.707235960000000	22.407474910000000	79.331986110000000
F23	20.904667550000000	18.701866070000000	22.592666670000000

Tabla 41. Número de casos en cada cluster.

Número de casos en cada cluster		
Cluster	1	139.000
	2	133.000
	3	60.000
Válidos		332.000
Péridos		0.000

CAPÍTULO V

5.1. CONCLUSIONES

- El algoritmo planteado mediante el programa diseñado en el software MATLAB permite conseguir en primera instancia la segmentación de los PCEE diarios analizados a partir de los datos de medición de energía eléctrica de uno o varios clientes por medio de las técnicas de clustering k -Means y k -Medoids aplicados a factores de forma que definen las características de comportamiento de los mencionados PCEE diarios, que en presente PTEI son valores de potencia activa normalizados, los cuales son obtenidos de los diferentes períodos de tiempo provenientes de los rangos horarios de las curvas de carga y del sistema de comercialización de energía eléctrica ecuatoriano.
- Las técnicas de clustering k -Means y k -Medoids encuentra una partición en la cual los factores de forma dentro de un cluster estén lo más cerca entre sí y lo más lejos posible de los otros factores de forma que este dentro de otros clusters; manejando medidas diferentes de distancia, siendo utilizada la distancia euclidiana por su sencillez de cálculo pero debiendo tener en cuenta que es una distancia sensible a las unidades de medida de las variables (factores de forma) que deben ser tipificados para conseguir valores de distancias euclidianas normalizadas en unidades similares (kW).
- Para la elección del número óptimo de clusters se realiza una evaluación bajo los criterios de valores de silueta que permiten minimizar el número de agrupamientos que comprenda todos los PCEE diarios y el rendimiento computacional de las técnicas de clustering mencionadas.
- Las técnicas de clustering k -Means y k -Medoids obtienen los mismos resultados de segmentación de datos correspondientes a los PCEE diarios de los consumidores mencionados en el presente PTEI; teniendo un mejor rendimiento computacional con respecto a los tiempos de ejecución la técnica de clustering k -Means en comparación a la técnica de clustering k -Medoids.
- Para la ratificación de la segmentación de PCEE diarios obtenidas mediante el programa desarrollado en el software MATLAB se utiliza el software WEKA y SPSS, respectivamente. Los datos de entrada utilizados en los programas mencionados provienen de los factores de forma que caracterizan a dichos PCEE diarios del cliente o los clientes que se encuentran en análisis y que son generados a partir del algoritmo desarrollado en MATLAB; debiendo realizar las configuraciones respectivas en el software WEKA y SPSS para obtener los resultados que corroboran la metodología planteada en el presente PTEI.
- En segunda instancia se ha buscado el procedimiento adecuado para la clasificación de los PCEE diarios con valores típicos y atípicos de cada uno de los clusters; de tal manera, que la detección de anomalías en el CEE del cliente o clientes sea eficiente y eficaz.

- El análisis estadístico descriptivo que se realiza a las distancias euclidianas y los valores de siluetas correspondientes a los factores de forma que definen las características de comportamiento de los PCEE ayudan a comprender la estructura que poseen, de manera de detectar tanto un patrón de comportamiento general como alejamientos del mismo.
- La determinación de la función de densidad de probabilidad que poseen las distancias euclidianas y los valores de siluetas; provenientes de los factores de forma que caracterizan a los PCEE diarios del cliente o los clientes que se encuentran en análisis, se ajusta de excelente manera mediante la función por Kernel que tiene un valor de eficiencia de uno y que permite determinar satisfactoriamente al algoritmo los límites que forman parte del condicionamiento para la detección de anomalías.
- La aplicación de MSV sin supervisión es un algoritmo altamente utilizado para la detección de valores atípicos (outliers), el cual permite separar todos los factores de forma del origen y maximiza la distancia desde el hiperplano que generan hacia su origen; que en el presente PTEI corresponde a cada uno de los centroides o medoides de cada cluster que fue seleccionado de manera óptima a partir del promedio máximo de los valores de silueta analizados mediante diversos conjuntos de segmentación de los datos.
- En última instancia se obtiene el patrón de consumo promedio de energía eléctrica diario de cada uno de los clusters. Dicho patrón se obtiene a partir de los PCEE sin outliers de cada cluster y permite conformar o determinar a qué clase de clientes pertenece el consumidor; así como, definir tarifas que sean limpias, transparentes y de fácil entendimiento, pero lo suficientemente flexibles como para seguir las variaciones del patrón de carga del cliente inducidas por tarifas específicas. Además, se puede mostrar una ayuda con la cual el cliente o clientes obtengan una mejora en la administración del CEE, cambiando sus hábitos con la finalidad de alcanzar la eficiencia energética.
- Al analizar los resultados obtenidos de los consumidores que se presentan en los diversos casos expuestos, se puede concluir que no todas las anomalías detectadas en los PCEE diarios son producidas necesariamente por una falla de carácter eléctrico, sino algunas de estas anomalías son generadas a partir de un origen diferente; como por ejemplo, días laborables y no laborables, días festivos o en un caso particular donde se realiza mantenimientos preventivos y/o correctivos de la diferente maquinaria que posee Continental Tire Andina S.A. cada fin de año.
- Los resultados obtenidos de cada uno de los consumidores analizados son independientes entre sí, debido a que cada consumidor presenta características únicas con relación a su tipo de consumo de energía eléctrica.

5.2. RECOMENDACIONES

- Partiendo del proceso de recolección y recopilación de información correspondiente a las mediciones de consumo de energía eléctrica que son manipulados en la aplicación de las técnicas de segmentación de datos (clustering) y detección de anomalías para la clasificación de consumos de energía eléctrica del cliente o clientes por medio del programa desarrollado en el software MATLAB, se recomienda el uso de contadores de energía con períodos de medición para la toma de datos que establecen las normativas.

- Se requieren tecnologías de medición adecuadas para aplicar la metodología propuesta. La tecnología propone una nueva generación de medidores de intervalos, con capacidades extendidas de administración de datos y comunicación, y con la posibilidad de lectura remota desde un centro de control. En muchos países, esta nueva tecnología ingresa rápidamente al mercado de la electricidad y mejorará enormemente la interacción entre el consumidor y el proveedor.
- Antes de aplicar las técnicas de segmentación de datos (clustering) y detección de anomalías para la clasificación de consumos de energía eléctrica del cliente o clientes por medio del programa desarrollado en el software MATLAB, se debe realizar una revisión de los datos de medición recolectados, observando que dichos datos pertenecientes a los PCEE diarios se encuentren completos para que el programa se ejecute con normalidad.
- Antes de aplicar las técnicas de segmentación de datos (clustering) y detección de anomalías para la clasificación de consumos de energía eléctrica del cliente o clientes por medio del programa desarrollado en el software MATLAB, se recomienda tener información de consumo de energía eléctrica por lo menos de un año para que los resultados finales tengan un mayor nivel de fidelidad.
- Para la segmentación de datos, se recomienda el uso de la técnica de clustering *k*-Means por su mejor rendimiento computacional con respecto a tiempos de ejecución.
- Complementar la metodología sobre la detección de valores atípicos en los PCEE diarios del cliente o clientes mediante estudios posteriores que permitan desarrollar diferentes algoritmos para la identificación de los tipos de anomalías o por qué se originaron.
- El programa desarrollado en el software MATLAB y el archivo de Microsoft Excel que contiene la información correspondiente a las mediciones de consumo de energía eléctrica, deben encontrarse ubicados en una misma carpeta para que los archivos generados por parte del programa sean almacenados en dicha carpeta.
- Se recomienda que los factores de forma que caracterizan a los PCEE diarios del cliente o los clientes deben ser elegidos dependiendo de la localidad de la cual proceden los datos de medición de energía eléctrica; ya que, los rangos horarios de medición difieren en cada país alrededor del mundo.
- Para el análisis de consumidores de energía eléctrica tipo residencial se recomienda poseer el consumo de energía eléctrica mensual de un año o más, teniendo como número mínimo 12 valores que conformarían la curva de carga anual de cada cliente y los factores de forma que caracterizarían dichas curvas procederían de las estaciones del año que se presentan en la región a la cual pertenecen los consumidores.

BIBLIOGRAFÍA

- [1] G. Chicco, “Customer Behaviour and Data Analytics,” *2016 Int. Conf. Expo. Electr. Power Eng.*, no. Epe, pp. 771–779, 2016.
- [2] J. Molina and J. García, “Técnicas de Análisis de Datos,” p. 266, 2006.
- [3] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From Data Mining to Knowledge Discovery in Databases,” *AI Mag.*, pp. 37–54, 1996.
- [4] G. K. Gupta, *Introduction To Data Mining With Case Studies*. PHI Learning Pvt. Ltd., 2014.
- [5] A. A. De Armas, “Detección de Outliers En Grandes Bases de Datos,” 2015.
- [6] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.
- [7] S. M. H. Jansen, “Customer Segmentation and Customer Profiling for a Mobile Telecommunications Company Based on Usage Behavior,” *Proc. - 3rd Int. Conf. Data Min. Intell. Inf. Technol. Appl. ICMIA 2011*, no. July, pp. 308–313, 2007.
- [8] MathWorks, “k-Means.” [Online]. Available: <https://es.mathworks.com/help/stats/k-means-clustering.html>.
- [9] MathWorks, “k-Medoids.” [Online]. Available: <https://es.mathworks.com/help/stats/kmedoids.html>.
- [10] MathWorks, “Support Vector Machines for Binary Classification.” [Online]. Available: <https://www.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html#bsr5b42>.
- [11] D. M. J. Tax, “One-class classification,” 2001.
- [12] I. H. Salgado, “Métodos Estadísticos No Paramétricos de Suavización para Análisis de Datos Biológicos,” 2000.
- [13] MathWorks, “Introduction to Cluster Analysis,” 2017. [Online]. Available: <https://www.mathworks.com/help/stats/introduction-to-cluster-analysis.html>.
- [14] MathWorks, “Hierarchical Clustering,” 2017. [Online]. Available: <https://www.mathworks.com/help/stats/hierarchical-clustering.html>.
- [15] MathWorks, “k-Means Clustering,” 2017. [Online]. Available: <https://www.mathworks.com/help/stats/k-means-clustering.html>.
- [16] G. Chicco, R. Napoli, and F. Piglione, “Application of clustering algorithms and self organising maps to classify electricity customers,” *2003 IEEE Bol. PowerTech - Conf. Proc.*, vol. 1, pp. 373–379, 2003.
- [17] MathWorks, “datevec,” 2017. [Online]. Available: https://www.mathworks.com/help/matlab/ref/datevec.html?s_tid=doc_ta.
- [18] MathWorks, “weekday,” 2017. [Online]. Available: https://www.mathworks.com/help/matlab/ref/weekday.html?s_tid=doc_ta.
- [19] G. Chicco, R. Napoli, and F. Piglione, “Application of clustering techniques to load pattern-based electricity customer classification,” no. June, pp. 6–9, 2005.
- [20] M. Piao, J. B. Lee, H. S. Shon, E. J. Cha, K. Ah Kim, and K. H. Ryu, “Identification of temporal interval relation of frequent patterns during incremental phase,” *Legacy*, pp. 497–502, 2011.
- [21] Consejo Nacional de Electricidad, “Estudio y gestión de la demanda eléctrica,” *Plan Maest. Electríf. 2013-2022*, vol. 2, pp. 41–42, 2013.
- [22] X. Serrano-Guerrero, G. Escrivá-Escrivá, and C. Roldán-Blay, “Statistical Methodology to Assess Changes in Buildings Electricity Consumption Profile,” *ENERGY AND BUILDINGS, International Journal, ELSEVIER*, 2018, pp. 1–24, 2018.
- [23] J. J. González, “El histograma con la TI-92: optimización de clases,” *Revista de Didáctica de las Matemáticas*, vol. 61, pp. 67–72, 2005.
- [24] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, “Customer Characterization Options for Improving the Tariff Offer,” *IEEE Power Eng. Rev.*, vol. 22, no. 11, p. 60, 2002.

- [25] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *J. Comput. Appl. Math.*, vol. 20, no. C, pp. 53–65, 1987.
- [26] MathWorks, “silhouette,” 2017. [Online]. Available: https://www.mathworks.com/help/stats/silhouette.html?s_tid=doc_ta.
- [27] MathWorks, “evalclusters,” 2017. [Online]. Available: https://www.mathworks.com/help/stats/evalclusters.html?s_tid=doc_ta.
- [28] D. A. Lind, W. G. Marchal, and S. A. Wathen, *Estadística aplicada a los negocios y la economía*. 2012.
- [29] MathWorks, “Kernel Distribution,” 2017. [Online]. Available: https://www.mathworks.com/help/stats/kernel-distribution.html?searchHighlight=epanechnikov&s_tid=doc_srchtile.
- [30] F. Scarlatache, G. Grigora, G. Chicco, and G. Câr, “Using k-Means Clustering Method in Determination of the Optimal Placement of Distributed Generation Sources in Electrical Distribution Systems,” *IEEE Optim. Electr. Electron. Equip. (OPTIM)*, 2012 13th Int. Conf., pp. 953–958, 2012.
- [31] MathWorks, “Detect outliers using SVM and One-Class Learning,” 2017. [Online]. Available: https://www.mathworks.com/help/stats/fitcsvm.html?s_tid=doc_ta.

ANEXOS

El programa desarrolla en el software MATLAB crea archivos de Microsoft Excel que contiene información con respecto a los diversos procesos ejecutados por parte del mismo. De esta manera, se da a conocer la información que contiene dichos archivos que se mencionan a continuación:

- **Archivo de Microsoft Excel denominado “FACTORES DE FORMA DE LOS PERFILES DE CARGA DIARIOS”:** Este archivo contiene los factores de forma obtenidos a partir de los PCEE diarios que fueron cargados en el programa. Este archivo es utilizado en el software WEKA y SPSS para realizar la verificación del proceso de segmentación de datos realizado por el programa.
- **Archivo de Microsoft Excel denominado “PERFILES_DE_CONSUMO_DE_ENERGÍA_ELECTRICA_SEGUN_SU_CLUSTER”:** Este archivo contiene los datos de las mediciones de energía eléctrica correspondientes a los PCEE diarios de cada uno de los clusters. Cada PCEE diario contiene el número de cluster al que pertenece, los factores de forma que lo caracterizan; así como, la fecha y el día de la semana al que corresponde.
- **Archivo de Microsoft Excel denominado “PERFILES_DE_CARGA_SEGMENTADOS_Y_CLASIFICADOS”:** Este archivo contiene los datos de las mediciones correspondientes a los PCEE diarios con outliers, PCEE diarios sin outliers y patrón promedio de consumo de energía eléctrica de cada cluster en hojas diferentes. Cada PCEE diario contiene el número de cluster al que pertenece, los factores de forma que lo caracterizan; así como, la fecha y el día de la semana al que corresponde.