

UNIVERSIDAD SIMÓN BOLÍVAR
DECANATO DE ESTUDIOS PROFESIONALES
COORDINACIÓN DE TECNOLOGÍA E INGENIERÍA ELÉCTRICA

DESARROLLO DE MODELO DE DEMANDA APLICABLE AL SISTEMA ELECTRICO NACIONAL

Por:
Luis Eduardo Sotillo Aguiar

PROYECTO DE GRADO
Presentado ante la Ilustre Universidad Simón Bolívar
como requisito parcial para optar al título de
Ingeniero Electricista

Sartenejas, Marzo 2017



UNIVERSIDAD SIMÓN BOLÍVAR
DECANATO DE ESTUDIOS PROFESIONALES
COORDINACIÓN DE TECNOLOGÍA E INGENIERÍA ELÉCTRICA

DESARROLLO DE MODELO DE DEMANDA APLICABLE AL SISTEMA ELECTRICO NACIONAL

Por:
Luis Eduardo Sotillo Aguiar

Realizado con la asesoría de:
Johnny Rengifo

PROYECTO DE GRADO
Presentado ante la Ilustre Universidad Simón Bolívar
como requisito parcial para optar al título de
Ingeniero Electricista

Sartenejas, Marzo 2017

UNIVERSIDAD SIMÓN BOLÍVAR
VICERRECTORADO ACADÉMICO
DECANATO DE ESTUDIOS PROFESIONALES
Coordinación de Tecnología e Ingeniería Eléctrica

ACTA DE EVALUACION DE PROYECTO DE GRADO

Código de la asignatura: EP3201 Fecha: 10-01-2017
Nombre del estudiante: Luis Eduardo Sotillo Aguiar Carnet: 07-41571
Título del proyecto: Desarrollo de un Modelo de Demanda Aplicable al Sistema Eléctrico Nacional
Tutor: Prof. Johnny Rengifo
Jurados: Profesores Carmen Quintero y Alexander Bueno

X	APROBADO		REPROBADO

Observaciones:

El jurado examinador, **por unanimidad**, considera el proyecto de grado merecedor de la mención especial SOBRESALIENTE:

X	SI		NO

El trabajo utiliza métodos estadísticos para el análisis de datos de demanda de energía y estratificación de los consumidores, siendo este problema de mucho interés para el país en la actualidad.

Prof. Alexander Bueno
Jurado
C.I. 10.511.239

Prof. Carmen Quintero
Jurado
C.I. 14.964.202

Prof. Johnny Rengifo
Tutor Académico
C.I. 17.758.221

Nota: Colocar los sellos de los respectivos Departamentos Académicos. Para jurados externos usar el sello de la Coordinación Docente. Este documento debe entregarse sin enmiendas.

UNIVERSIDAD SIMÓN BOLÍVAR
VICERRECTORADO ACADÉMICO
DECANATO DE ESTUDIOS PROFESIONALES
Coordinación de Tecnología e Ingeniería Eléctrica

ACTA DE EVALUACION DEL INFORME DE PROYECTO DE GRADO

Código de la asignatura: EP3201 (tres avances) Fecha: 06-01-2017
Nombre del estudiante: Luis Eduardo Sotillo Aguiar Carnet: 07-41571
Título del proyecto: Desarrollo de un Modelo de Demanda Aplicable al Sistema Eléctrico Nacional
Tutor: Prof. Johnny Rengifo
Jurados: Profesores Carmen Quintero y Alexander Bueno

Nosotros, miembros del Jurado designado por la Coordinación Docente de Tecnología e Ingeniería Eléctrica, en la fecha indicada para la evaluación del Proyecto de Grado en cuestión, hemos evaluado el contenido del mismo y hacemos constar nuestra decisión de **aceptar el documento para su presentación en acto público**. Se convoca al estudiante, a la presentación pública del Proyecto de grado en el edificio "Energética" sala "Sala de Conferencias – Departamento de Conversión Y Transporte de Energía".

Fecha: 10/01/2017
Hora: 10:30 am
Lugar: Sede de Sartenejas

Prof. Alexander Bueno
Jurado (Presidente)
C.I. 10.511.239

Prof. Carmen Quintero
Jurado
C.I. 14.964.202

Prof. Johnny Rengifo
Tutor Académico
C.I. 17.758.221

Nota: Colocar los sellos de los respectivos Departamentos Académicos. Para jurados externos usar el sello de la Coordinación Docente. Este documento debe entregarse en original y sin enmiendas.

DESARROLLO DE MODELO DE DEMANDA APLICABLE AL SISTEMA ELECTRICO NACIONAL

RESUMEN

Realizado por: Luis Eduardo Sotillo Aguiar

En el presente trabajo se desarrolló una herramienta computacional en Python 3.5 para la identificación y estratificación de consumo de energía de clientes del sistema eléctrico. Se utiliza un algoritmo de aprendizaje no supervisado llamado K-medianas, de uso común en el proceso de minería de datos, para la identificación de patrones de consumo de usuarios del sistema, mientras el criterio de similitud se basó en la distancia Manhattan. Se procedió a comparar cuatro algoritmos: Jerárquico, K-medias, K-medianas y Sigue al líder modificado para decidir el más conveniente, mediante las medidas de eficiencia MIA (Índice de Adecuación de la Media) y CDI (Indicador de Dispersión del Conglomerado). La metodología desarrollada consta de cuatro fases principalmente: a) Recopilación de la data. b) Pre-procesamiento. c) Procesamiento. d) Visualización. La data utilizada para la evaluación del código constó de una combinación de data residencial y no residencial, de la cual, la primera estuvo compuesta por data real publicada para el estudio de análisis de datos y desagregación de carga, mientras que la segunda fue generada artificialmente mediante modelos probabilísticos de HMM (Modelos Ocultos de Markov). La metodología y la herramienta computacional para su aplicación, resultó viable para el reconocimiento de patrones y estimación de demanda, presentando esta última un error entre 2 y 6 %; permitiendo también la identificación de estratos de mayor impacto al sistema en horas pico.

Palabras clave: **Caracterización De Carga, Minería De Datos, Conglomerados, Software Libre, HMM, Curvas De Carga, Demanda De Energía, Demanda Horaria, Redes Eléctricas Inteligentes.**

Para crecer,
Hay que dar un paso a lo desconocido.

AGRADECIMIENTOS

A Dios.

A mi familia, que me ha acompañado y apoyado desde que dí mis primeros pasos en el mundo, y que jamás me dejo caminar cabizbajo.

A mis tutores, Francisco Ochoa, René Ortiz y Johnny Rengifo, quienes me introdijeron en el tema y me guiaron a lo largo de este proyecto.

A Giannina, quien ha sido más que mi compañera, mi amiga, mi equipo. Que compartió conmigo cada gota de sudor a lo largo de estos años, y fue esa patada que me mandaba hacia adelante cuando me quería detener. Es infinita mi gratitud.

A Explotion People, quienes han estado a mi lado en cada trajin de la vida, me han visto madurar y celebrar cada momento. Más que amigos son mi familia, mis hermanos, parte de mi personalidad y mi conciencia. Quien soy hoy, posee un granito de cada uno de ellos. Los amo con pasión y locura.

A Betha y Eizu, por su apoyo y ayuda, que hicieron posible que la última versión estuviese lista a tiempo.

A la coordinación, a Benicia y María Teresa quienes me prestaron su ayuda una y otra vez durante todos estos años de carrera.

A Samelis y Lili, que en tantas ocasiones me brindaron su ayuda desde el departamento.

ÍNDICE GENERAL

PORTADA I

ACTA DE EVALUACIÓN II

ACTA DE EVALUACIÓN DE INFORME III

RESUMEN IV

DEDICATORIA V

AGRADECIMIENTOS VI

ÍNDICE GENERAL VII

ÍNDICE DE TABLAS X

ÍNDICE DE FIGURAS XI

LISTA DE ABREVIATURAS XIV

INTRODUCCIÓN 1

1. MODELACIÓN DE CARGA 4

1.1. Modelos matemáticos 5

1.1.1. Modelo estático. 5

1.1.2. Modelo Dinámico 6

1.1.3. White Box o Modelo Físico 6

1.1.4. Black Box y Técnicas de Identificación 7

1.2. Modelos Heurísticos 7

1.3. Redes Neurales Artificiales (ANN) 7

1.4. Método de Uso Final 8

1.5. Tendencias recientes en la modelación de la carga 8

1.5.1. Identificacion de Caja Gris 9

1.5.2. Técnicas de Inteligencia Artificial con modelación Top-down 9

1.5.3. Modelación Bottom-Up con modelos individuales detallados 9

2. DEMANDA ELÉCTRICA 11

2.1. Factores característicos de la carga 12

2.2. Medidor de Energía 13

2.3. Gestión de la Demanda 13
2.4. Caracterización de la carga 15

3. MINERÍA DE DATOS 17
3.1. Análisis de Clustering 18
3.1.1. Medidas de similitud 18
3.1.2. Métodos de Clustering 19
3.1.2.1. Jerárquico 19
3.1.2.2. K-medias 20
3.1.2.3. K-medianas 21
3.1.2.4. Sigue al Líder Modificado 22
3.1.3. Validación de los clusters 23
3.2. Selección del algoritmo 25

4. METODOLOGÍA Y DESCRIPCIÓN DE LA PLATAFORMA PARA CARACTERIZACIÓN DE CARGA 33
4.1. Descripción general 33
4.2. Recopilación de la data 37
4.3. Pre-procesamiento 38
4.3.1. Sincronización 38
4.3.2. Depuración 38
4.4. Procesamiento y Visualización 40
4.4.1. Estratificación 41
4.4.2. Identificación 42
4.4.3. Potencia Agregada 43
4.4.4. Contribución en picos horarios 44
4.4.5. Visualización de un consumidor específico 46
4.4.6. Reducción 48

5. GENERACIÓN Y PREPARACION DE LA DATA 52
5.1. HMM 53
5.1.1. Elementos de un HMM 53
5.1.2. Definición formal del HMM para una observación discreta 54
5.1.3. Problemas fundamentales 55
5.2. Aplicando HMM y la librería *hmmlearn* para generación de data artificial . . 55

6. Caso de Estudio 60

7. RESULTADO **62**
7.1. Caso 1: Conjunto de datos PSEG (1reg/h) 62
7.1.1. Pre-procesamiento 62
7.1.2. Procesamiento y Visualización 62
7.1.3. Identificación 69
7.1.4. Potencia Agregada 70
7.2. Caso 2: 1reg/15min 71
7.2.1. Pre-procesamiento 72
7.2.2. Procesamiento y Visualización 72
7.2.3. Identificación 77
7.2.4. Potencia Agregada 78

CONCLUSIONES **81**

REFERENCIAS BIBLIOGRÁFICAS **83**

A. CAPTURAS DE PANTALLA **86**

ÍNDICE DE TABLAS

3.1. Indice Mia, de 2 a 15 grupos 26
3.2. Índice CDI, de 2 a 15 grupos 28
3.3. Índice MIA, de 2 a 15 grupos 29
3.4. Índice CDI, de 2 a 15 grupos 31

6.1. Descripción de data a estratificar. 60

7.1. Identificación de los estratos 69
7.2. Identificación de los estratos 77

ÍNDICE DE FIGURAS

3.1. Dendograma resultante de un proceso de clustering jerárquico [19] 20
3.2. Flujograma K-medianas 21
3.3. Flujograma Modified Follow de Leader [5] 23
3.4. Curva del codo evaluando de 2 a 30 grupos 25
3.5. Histograma Data usada para la prueba de los algoritmos. Eje x: Potencia máxima normalizada. Eje y: Frecuencia 26
3.6. Curva representativa de MIA para distintos algoritmos de clustering 27
3.7. Curva representativa de CDI para distintos algoritmos de clustering 28
3.8. Curva representativa de MIA para distintos algoritmos de clustering 30
3.9. Curva representativa de CDI para distintos algoritmos de clustering 31

4.1. Esquema de la metodología por etapas 34
4.2. (a)Flujograma correspondiente a *Caracterizacion.py* 35
4.3. (b)Flujograma correspondiente a *Caracterizacion.py* 36
4.4. (c)Flujograma correspondiente a *Caracterizacion.py* 37
4.5. (a) Diagrama de flujo del Pre-procesamiento 39
4.6. (b) Diagrama de flujo del Pre-procesamiento 40
4.7. Primer gráfico resultante de la caracterización de los patrones de consumo . 41
4.8. Identificación de cada estrato 42
4.9. Estratos identificados según el uso de la energía. Eje x: No de Usuarios. Eje y: Estratos 43
4.10. Gráfica de potencia agregada. Eje x: tiempo - 24 horas en intervalos de 1 hora.. Eje y: Vatios (w) 44
4.11. Flujograma de la función *fact_imp* 45
4.12. Contribución en % de cada estrato al consumo agregado total 45
4.13. Resultado de la función *consumidor* 47
4.14. Flujograma de la función *consumidor* 48
4.15. Flujograma de la función *Reduccion()* 49
4.16. Flujograma de la función inserción 50

5.1. Ejemplo de curvas de carga residencial con frecuencia de adquisición de 1Hz. Eje y: Vatios (w). Eje x: tiempo de 1 día en intervalos de 1 s. 52
5.2. Histograma de la data Residencial REDD y ECO a tres horas distintas del día. Eje y: Frecuencia. Eje x: Vatios (w) 53
5.3. Curva de amalgamiento correspondiente a la secuencia de observación 56
5.4. Parámetros de modelo 58

5.5. Ejemplos aleatorios de curvas producidas por el modelo entrenado de HMM. Eje y: Vatios (kw). Eje x: tiempo de 1 día en intervalos de 5 min. 58
5.6. Histograma de la data aleatoria generada a tres horas distintas del día. Eje y: Frecuencia. Eje x: Vatios (w) 59

6.1. Histograma de la data aleatoria generada a tres horas distintas del día. Eje y: Frecuencia. Eje x: Vatios (w) 61

7.1. Curva de amalgamiento 63
7.2. Perfiles de demanda agrupados por cluster. Eje x: tiempo - 24 horas en intervalos de 1 hora. Eje y: Potencia máxima normalizada 64
7.3. Centro de Estrato 1 64
7.4. Centro de Estrato 2 65
7.5. Centro de Estrato 3 65
7.6. Centro de Estrato 4 66
7.7. Centro de Estrato 5 66
7.8. Centro de Estrato 6 67
7.9. Centro de Estrato 7 67
7.10. Centro de Estrato 8 68
7.11. Centro de Estrato 9 68
7.12. Factores que describen la demanda en los estratos 69
7.13. Distribución de antiguos consumidores en nuevos estratos. Eje x: No de Usuarios. Eje y: Estratos 70
7.15. (%) consumo de energía por bloques horarios. 71
7.14. Potencia Agregada. Eje x: tiempo - 24 horas en intervalos de 1 hora. Eje y: Vatios (w) 71
7.16. Curva de amalgamiento 72
7.17. Perfiles de demanda agrupados por cluster. Eje x: tiempo - 24 horas en intervalos de 15 min. Eje y: Potencia máxima normalizada 73
7.18. Centro de Estrato 1 73
7.19. Centro de Estrato 2 74
7.20. Centro de Estrato 3 74
7.21. Centro de Estrato 4 75
7.22. Centro de Estrato 5 75
7.23. Centro de Estrato 6 75
7.24. Centro de Estrato 7 76
7.25. Centro de Estrato 8 76
7.26. Factores que describen la demanda en los estratos 77
7.27. Identificación de estratos. Eje x: No de Usuarios. Eje y: Estratos 78

7.28. Potencia Agregada. Eje x: tiempo - 24 horas en intervalos de 15 min. Eje y: Vatios (w) 79
7.29. % de consumo 79
7.30. Nueva Estratificación 80

A.1. Captura de pantalla. Inicio del programa: Introduccion de la data, pre-procesamiento e inicio del caso de estudio. 86
A.2. Captura de pantalla del programa. Tiempo de normalización y asignación del número de estratos. 87
A.3. Captura de pantalla. Factores que describen los estratos y contribución porcentual de cada estrato en rangos horarios. 87
A.4. Captura de pantalla. Descripción de dos usuarios aleatorios. 88

LISTA DE ABREVIATURAS

DSM	Gestión de la Demanda (Demand Side Management)
ML	Aprendizaje automático (Machine Learning)
EV	Carro Eléctrico (Electric Vehicle)
MFL	Sigue al líder modificado (Modified Follow the Leader)
MIA	Indice de adecuación de la media (Mean Index Adequacy)
CDI	Indicador de dispersión del clustering (Cluster Dispersion Indicator)
NaN	No un número (Not a Number)
EM	Esperanza - Maximización (Expectation-Maximization)
HMM	Modelos Ocultos de Markov (Hidden Markov Models)
PSEG	Grupo Incorporado y Empresa del Servicio Público (Public Service Enterprise Group Incorporated)
AR	Autoregresivo (Autoregressive)
ARX	Autoregresivo con entradas exógenas (Autoregressive with exogenous inputs)
ARMAX	Movimiento promedio autoregresivo con entradas exógenas (Autoreggresive moving average with exogenous inputs)
ANN	Redes Neuronales Artificiales (Artificial Neural Networks)
SEP	Sistema Eléctrico de Potencia
MA	Movimiento promedio (Moving Average)
DR	Respuesta a la Demanda (Demand Response)
TOU	Tiempo de uso (Time Of Use)

INTRODUCCIÓN

El crecimiento demográfico y tecnológico es algo ineludible, por lo que los encargados del servicio eléctrico deben estar preparándose continuamente para la consecuente expansión de la red. Esto implica analizar la demanda eléctrica de parte de cada usuario las 24 horas del día, los cuales poseen un comportamiento estocástico. A esta expansión le sigue la introducción continua de novedosos dispositivos cuyo crecimiento a nivel tecnológico en las últimas décadas ha llevado a una tendencia hacia el control electrónico, que hace que las cargas no se puedan describir de manera constante. En este sentido, en el presente trabajo se modela la potencia en los dispositivos de consumo final, a nivel de clientes residenciales y no residenciales, considerando el patrón promedio al que se tiende a usar la energía a lo largo del día por grupos que tienen similitudes a nivel social y de poder adquisitivo. Esto es lo que se conoce como un estudio de caracterización de carga.

Este tipo de estudio es necesario periódicamente en la red y es de interés para toda la industria eléctrica por su impacto desde generación hasta el ciclo comercial.

En Venezuela el último estudio de caracterización de carga a nivel nacional data del año 2002, en el cual se realizó el estudio para lanzar el último pliegue tarifario publicado en gaceta nacional y vigente hasta hoy. En ese tiempo, la población y número de clientes del servicio eléctrico ha crecido y la economía ha cambiado de manera significativa. Actualmente la información de este y otros estudios de caracterización de carga locales no se encuentran disponibles. Por otro lado, el crecimiento del número de usuarios lleva a un estrés adicional durante las horas pico, con mayor cantidad de consumidores, esto trae como consecuencia que se sobrepase la capacidad del sistema de suplir las exigencias de estos, causando fallas en el servicio.

El trabajo desarrollado durante este proyecto permite el análisis de información y descripción de usuarios del sistema eléctrico, necesario para el desarrollo de pliegues tarifarios y planes de eficiencia energética.

Antecedentes

Entre los tópicos con mayor auge en los últimos años entre la investigación y aplicación en la ingeniería eléctrica se destaca la aparición de las redes eléctricas inteligentes, con especial interés en los medidores inteligentes y cómo manejar la información que proporcionan, para mantenimiento y mejora de la calidad de servicio. En [1] se describe como las características del medidor inteligente pueden ser utilizadas para aumentar la participación del consumidor y beneficios del servicio eléctrico; y presenta una herramienta para la gestión de políticas de tiempo de uso (TOU, por sus siglas en inglés Time Of Use). De la misma manera [2] muestra

el manejo de la data generada por las redes inteligentes (Smart Grids) para la caracterización de los perfiles de consumo como base para la elección de técnicas de Gestión de la Demanda (DSM, por sus siglas en inglés Demand Side Management).

Kleiminger et al [3], describen la caracterización del comportamiento de los consumidores del sector eléctrico español de zonas geográficas distintas tales como edificios universitarios, zonas residenciales y mediana industria, utilizando el método de redes neuronales conocido como Kohonen maps. La literatura presenta varias técnicas para la caracterización de los perfiles de carga a través de técnicas estadísticas como los algoritmos de clustering o conglomerados, y en [4] se realiza la prueba y comparación de varios de los más utilizados para esta aplicación (K-medias, K-medias difuso, dos tipos de jerárquicos, Sigue al Líder Modificado (MFL, por sus siglas en inglés Modified Follow the Leader), entre otros), utilizando curvas de carga pertenecientes a clientes conectados a un sistema de distribución en Media Tensión. Mahmoudi-Kohan [5] nos ofrece una comparación entre el uso del método de WFA K-medias y Sigue al Líder Modificado aplicado a los consumidores del servicio eléctrico Tehran. Los autores de [6] hacen uso de los métodos de minería de datos en la perfilación de videojugadores, estratificando la población de jugadores para mejorar la interacción de estos con futuros videojuegos. El Proyecto de Grado presentado en [7] se enfoca en la comparación de distintos algoritmos de agrupación de datos basados en metaheurísticas e inteligencia de población llegando a concluir como mejor alternativa el uso de algoritmos genéticos, complementados con el uso de K-medias para un resultado final. En [8] la autora propone una metodología para el cálculo de pérdidas no técnicas en el Sistema Eléctrico Venezolano utilizando un modelo de demanda basado en consumo final de la energía. Los últimos estudios de caracterización de carga realizados en el país datan de 1998, por parte de, en aquel momento, Ministerio de Energía y Minas con el fin de obtener información para el cálculo del factor de responsabilidad de costos. En 2002 se realizó el estudio para el desarrollo de los pliegues tarifarios, los cuales siguen vigentes hasta la actualidad y en 2009 se realizó una campaña de medición a una cantidad de 6200 usuarios a través de una empresa consultora extranjera.

Objetivo General

Desarrollar una herramienta computacional de caracterización de carga que permita el modelo de demanda basado en el patrón de consumo promedio (curvas de carga) y estratificación de demanda del Sistema Eléctrico

Objetivos Específicos

1. Seleccionar un método para la simulación de cargas de acuerdo a un patrón de consumo promedio.

2. Caracterizar y estratificar la demanda según diferentes criterios: a) Patrón de consumo, b) Ubicación Geográfica, c) Nivel de tensión, Uso final de la energía.

3. Determinar la contribución de los estratos al consumo agregado, en su totalidad y en horas pico.

4. Desarrollar de algoritmo para agregación de nuevos usuarios a la base de datos según los criterios establecidos en la estratificación.

5. Desarrollar una metodología para la generación de curvas artificiales de registro horario aptas para la evaluación de los algoritmos de estratificación implementados en este proyecto.

CAPÍTULO 1

MODELACIÓN DE CARGA

La modelación de carga juega un papel muy importante en el estudio y operación de los sistemas eléctricos de potencia (S.E.P.), ya que el resultado de los mismos depende principalmente del modelo que se asume. La mayoría de los estudios que se realizan como análisis de estabilidad de voltaje, estimación de estado, flujo de carga, estudio de corto-cirtuito, entre otros, depende de los parámetros de la carga para llegar a un resultado lo más parecido posible a la realidad. Por ejemplo: Luego de un blackout (apagón general), si se usa un modelo equivocado para representar el comportamiento de la carga esto puede acarrear la salida de otros componentes del sistema en el momento del suceso real. Por otro lado, al momento de hacer estimación o pronóstico de la carga y planes de expansión el uso de un modelo errado puede significar la pérdida de millones de dólares [9].

Se presta muy poca atención al área de pronóstico y modelación de la carga. En gran parte, debido a la dificultad de realizar un modelo de carga preciso, y la falta de información acerca de los componentes de la carga. En los últimos años, con la aparición y aplicación de los medidores inteligentes e infraestructura de medición [10], se ha retomado la atención hacia este tema e incluso se ha profundizado, disponiéndose ahora de gran cantidad de data la cual antes era escasa.

Se pueden mencionar dos enfoques principales al modelar la carga:

Bottom-up: Se construye el modelo basado en la información particular de cada componente de la carga por individual en una barra específica [9].

Top-down: Se forma el modelo basado en la información relacionada con la sumatoria de los comportamientos de los componentes de la barra [9].

Estas dos vías pueden ser desarrolladas para describir, ya sea el comportamiento dinámico de los componentes o el comportamiento de una parte del sistema visto como el punto de medición de un usuario o un alimentador.

1.1. Modelos matemáticos

Un modelo matemático de la carga es una representación de la misma mediante una ecuación que relaciona la potencia y la tensión de la carga en una barra. Dentro de los enmarcados como modelos matemáticos se pueden encontrar:

1.1.1. Modelo estático.

Es un modelo utilizado para representar cargas estáticas, como resistencias, o hacer aproximaciones de cargas dinámicas, como motores de inducción. La potencia se define en función de la magnitud de la tensión y/o la frecuencia del sistema en un mismo instante [11].

De este tipo de modelos los mas conocidos son:

Impedancia Constante, en el cual la potencia varía en función del cuadrado de la tensión.

Corriente constante, en el cual la potencia varía proporcionalmente a la magnitud de la tensión.

Potencia constante, el cual es ampliamente usado en los análisis. La potencia no varía en función de la tensión.

Modelo Polinomial, a través del cual la potencia y tensión de la carga están relacionados por las siguientes ecuaciones:

$$P = P_o[a + b(\frac{V}{V_o}) + c(\frac{V}{V_o})^2] \tag{1.1}$$

$$Q = Q_o[d + e(\frac{V}{V_o}) + f(\frac{V}{V_o})^2] \tag{1.2}$$

Las constantes *a, b, c, d,* y *f* son parámetros del modelo. Po, Qo y Vo son valores iniciales (en caso de modelar un dispositivo en específico, se toman los valores nominales). También es conocido como modelo "ZIP" ya que los componentes del polinomio representan los modelos de impedancia, corriente y potencia constantes [9].

Modelo exponencial, la potencia y tensión de la carga están relacionadas por una ecuación exponencial [11]:

$$P = P_o(\frac{V}{V_o})^a \tag{1.3}$$

$$Q = Q_o(\frac{V}{V_o})^b \tag{1.4}$$

Donde a y b son parámetros del modelo. Po, Qo y Vo son valores iniciales. Es posible describir el modelo polinomial como una suma de tres componentes exponenciales, con exponente 2 (P ctte), 1 (I ctte) y 0 (Z ctte).

Modelo dependiente de la frecuencia,

$$Factor = [1 + a_f(f - f_o)] \tag{1.5}$$

siendo a_f la sensibilidad de la frecuencia, f, la frecuencia de la tensión en la barra, f_o la frecuencia nominal. Este modelo relaciona la potencia y frecuencia de la barra. El *Factor* se multiplica por el *modelo polinomial* o exponencial para así hacerlo dependiente de la frecuencia. Sin embargo al multiplicar el modelo ZIP por este factor pierde todo sentido físico.

Modelo EPRI LOADSYN, usado por el programa computacional del mismo nombre, y empleado por muchos servicios para calcular el comportamiento del S.E.P. Comprende las características descritas por los modelos ZIP y *dependiente de frecuencia* [9].

$$P = P_o\{P_{a1}(\frac{V}{V_o})^{KPV1}[1 + KPF1(f - f_o)] + (1 - P_{a1})(\frac{V}{V_o})^{KPV2}\} \tag{1.6}$$

$$Q = Q_o\{Q_{a1}(\frac{V}{V_o})^{KQV1}[1+KQF1(f-f_o)]+((\frac{Q_o}{P_o}-Q_{a1})(\frac{V}{V_o})^{KQV2}[1+KQF2(f-f_o)]\} \tag{1.7}$$

La potencia activa está representada por dos componentes: uno dependiente de la frecuencia de la tensión y otro que no lo és; y la potencia reactiva posee una primera componente que representa el consumo de reactivos en la carga y un segundo que está relacionado con el consumo de alimentadores y transformadores necesario para la inyección inicial de reactivos (Qo) [9].

1.1.2. Modelo Dinámico

Es un modelo que mediante ecuaciones diferenciales relaciona la potencia activa y reactiva de la carga con la magnitud del voltaje y la frecuencia de la red [11].

La potencia se representa como una función de la tensión y frecuencia de una barra definida para un instante de tiempo actual y pasado. Es la base de muchas técnicas de identificación usadas para modelación y pronóstico de comportamiento de carga [9].

1.1.3. White Box o Modelo Físico

Es también llamado modelo fenomenológico o conceptual. Estos modelos se basan en el conocimiento de las leyes físicas que rigen el comportamiento de la carga y las matemáticas que lo describen. Si se conoce bien el funcionamiento de la carga entonces son conocidos los parámetros del modelo [9].

1.1.4. Black Box y Técnicas de Identificación

Es también conocido como método de identificación o modelación empírica. Implica que no se necesita poseer un conocimiento detallado del funcionamiento interno de la carga para obtener un modelo matemático que la represente adecuadamente. No posee un significado físico [9]. Esta técnica sigue la filosofía de combinar observaciones y errores pasados en una serie de tiempo para así pronosticar su futuro [12]. Ejemplo de esto es, AR (Autoregresivo), ARX (Autoregresivo con entradas exógenas), ARMAX (Movimiento promedio autoregresivo con entradas exógenas), MA (Movimiento promedio) entre otros. Es necesario conocer las características de la serie de tiempo para seleccionar la especificación correcta del modelo [12].

Los modelos descritos como dinámicos y estáticos corresponden a ecuaciones compuestas por variables y parámetros. Las variables se refieren a cantidades que pueden ser medidas o monitoreadas, mientras que los parámetros, correspondientes a constantes y exponentes de la ecuación, no pueden ser medidos directamente pero pueden ser estimados mediante técnicas de identificación. Por ejemplo la reducción de mínimos cuadrados [9].

1.2. Modelos Heurísticos

Este tipo de modelos están basados en reglas, asociadas a algunas condiciones y consecuencias, (antecedentes y consiguientes), dicho de otra manera están asociados a "patrones". El uso de este tipo de representación se basa en determinar esas reglas que marcan el comportamiento de la carga, y estas se obtienen de la experiencia; permitiendo el desarrollo de funciones no matemáticas para modelar la carga. Esta forma de modelación es usualmente implementada en sistemas expertos (inteligencia artificial) [9].

1.3. Redes Neurales Artificiales (ANN)

Las ANN tienen la habilidad de enfrentarse a problemas no lineales y sistemas complejos. Una ANN está compuesto por un conjunto de unidades de proceso llamadas 'neuronas artificiales', que simulan las funciones de las neuronas biológicas. Cada elemento está interconectado por "pesos" (parámetros matemáticos) que representan la sinapsis que ocurre en el cerebro humano. De manera general los ANN's son entrenados utilizando una serie de patrones de entrada/salida y el resultado de este entrenamiento son los valores finales de los pesos, que determinan el modelo [9]. Durante el proceso de aprendizaje, los pesos sufren modificaciones continuas, por lo que podemos decir que el proceso termina una vez que los pesos permanecen estables [13].

De forma general se consideran dos tipos de aprendizaje:

- Aprendizaje supervisado: Genera los valores de los pesos a través de la minimización de

una función que representa la diferencia entre la salida deseada y la salida obtenida por la ANN (tolerancia) [9]. Depende de la existencia de un agente externo que controla el proceso de aprendizaje y se caracteriza por arquitecturas en niveles y conexiones neuronales estrictamente hacia adelante [13]. Ejemplo: Hopfield Networks, Multilayer Perceptron Network [14], etc.

- Aprendizaje auto-organizado o no supervisado: El ANN no necesita de un agente externo pues se organiza a sí mismo en función de sus entradas. Son altamente utilizados en la clasificación de curvas de carga. Ejemplo: Herault and Jutten Networks y Kohonen Networks [3, 4].

Determinar el modelo de carga por ANN consiste en encontrar el conjunto de parámetros idóneos que permiten obtener los resultados más precisos, como N° de neuronas, N° de capas ocultas, valor de los pesos, etc. La topología usualmente se construye de forma empírica [9, 13]. Su capacidad de afrontar problemas de alta no-linealidad es una gran ventaja, sin embargo, es una debilidad la necesidad de una etapa de aprendizaje pues el resultado final depende en gran medida de la cantidad y calidad de la data utilizada para esto.

Debido a sus características ANN puede ser considerado como un modelo de Black Box.

1.4. Método de Uso Final

Es un método de ingeniería cuya idea responde al hecho de que la energía es usada a través de equipos que son los que realmente satisfacen las necesidades de los usuarios y que la energía total requerida será producto del número de equipos por estrato y la energía consumida por cada equipo. Se utiliza comúnmente en el sector residencial y poco en los sectores comercial e industrial, debido en su mayor parte a la homogeneidad de los equipos existentes en el sector residencial [12].

Este método proporciona información acerca de la forma en que es utilizada la energía eléctrica por el usuario, y es muy útil para evaluar el impacto de nuevos componentes y tecnologías que se agregan a la red; como por ejemplo computadores, consolas de vídeo juegos o un carro eléctrico. Permite también analizar el impacto de ciertos programas, políticas y respuestas de DSM y respuesta a la demanda (DR), dirigidos a manejar el uso de la energía de manera eficiente.

1.5. Tendencias recientes en la modelación de la carga

Los modelos que se han mencionado anteriormente pueden describir la carga en condiciones normales de operación, por lo que se puede decir que son adecuados para la forma en que son estudiados y analizados normalmente los S.E.P., como una foto de la red en un instante de tiempo dado, manteniendo valores constantes. Sin embargo, como ya

se ha mencionado antes el S.E.P. posee un comportamiento dinámico, además de ello, al presentarse una perturbación se espera que el modelo responda adecuadamente, y por cada tipo de perturbación hay un análisis pertinente, por lo que para cada análisis hay un modelo adecuado para su estudio. Eso implica que para analizar el S.E.P. con respecto a un fenómeno hacen falta distintos modelos, los cuales a pesar de adecuarse al análisis de la respuesta bajo perturbación, son homogéneos desde el punto de vista dinámico [9].

El problema radica en que los sistemas de potencia cada día se vuelven más complejos; cada día nuevos equipos son conectados para trabajar junto con los ya existentes y la electrónica se ha adueñado en gran parte de la carga del consumidor, por lo que la misma ya no se comporta de forma homogénea. De hecho, cada dispositivo tiene un comportamiento único dependiendo de la tecnología utilizada y del fabricante.

Por lo tanto, las nuevas técnicas y modelos se centran en representar el impacto de un gran número de dispositivos distintos en el comportamiento de la carga bajo condiciones de operación normal y anormal. De allí, tres grandes tendencias pueden ser resaltadas en post de esta tarea:

1.5.1. Identificacion de Caja Gris

Este método combina las ventajas del White Box (un significado físico con respecto a la carga) y Black Box (modelación sin conocimiento previo de la carga), por consiguiente está basado en un sistema de mediciones y también en información auxiliar, asociado a las leyes físicas que describen el funcionamiento de la carga [9].

Esta técnica de modelación corresponde a una nueva área de identificación y ya que provee un modelo más general incluso con un número de mediciones limitadas, tiende a ser usada para modelación y pronóstico.

1.5.2. Técnicas de Inteligencia Artificial con modelación Top-down

Top-down permite la modelación de las cargas asociadas vistas desde un alimentador, área o subestación. Este método combinado con técnicas de inteligencia artificial lleva a resultados interesantes ya que los parámetros del modelo pueden ser modificados a través de reglas heurísticas almacenadas en una base de conocimiento conocida. Ejemplo: ANN[4, 14], sistemas expertos, Lógica de asociacion difusa, etc [9].

1.5.3. Modelación Bottom-Up con modelos individuales detallados

Bottom-up está pensado para determinar el modelo de cada elemento del sistema. De tal forma, permite modelar elementos individuales como: lámparas, calentadores, aires acondicionados, etc. Con los avances en la programación y poder de computo este método

puede ser combinado con la programación orientada a objetos (o agentes) para representar cada modelo distinto como distintos casos del mismo contexto computacional [9].

Cada consumidor posee una cantidad de dispositivos conectados a la red que son desconocidos y muy variados en su comportamiento, por lo que no es una opción directa el aplicar técnicas de modelación que necesiten una entrada de los parámetros físicos específicos de las cargas.

Sin embargo, se posee las mediciones de consumo de potencia de dichos usuarios en ciertas muestras de tiempo, lo cual devuelve información acerca de cómo usan ese conjunto de cargas de las cuales disponen dentro de sus hogares, cuánto consumen y sus hábitos en el uso de ellos. Por lo tanto, basándonos en la data histórica de los usuarios de la red es viable utilizar métodos estadísticos para generar un modelo de demanda apropiado que pueda describir de forma general el comportamiento de grupos o estratos de consumidores, generalizando así un modelo no de la carga como dispositivo sino como la demanda ocasionada por el usuario.

CAPÍTULO 2

DEMANDA ELÉCTRICA

La Demanda Eléctrica, según la Ley del Servicio Eléctrico [15] se define como el requerimiento de potencia y energía eléctrica de un usuario, sector o sistema eléctrico.

La Energía Eléctrica, según la misma ley, está definida como "la potencia eléctrica producida, transmitida o consumida en un período determinado. Se mide y se expresa en vatio hora (Wh) o en sus múltiplos: kilovatio hora (kWh), Megavatio hora (MWh), Gigavatio hora (GWh) o Teravatio hora (TWh)".

Su análisis generalmente se realiza a través de la modelación de la energía demandada por los distintos tipos de consumidores, desagregando estos desde un punto de vista sectorial en residencial, comercial e industrial. Posteriormente la potencia es obtenida haciendo uso de variables asociadas a la forma temporal de la carga [12].

Existen múltiples factores vinculados directa o indirectamente con la demanda de energía eléctrica. Entre ellos se encuentran los socioeconómicos, demográficos, urbanos, climáticos, tecnológicos, entre otros. A corto y mediano plazo los factores que se usan para establecer la demanda de electricidad son el precio de la electricidad, el nivel de actividad económica y los climáticos [12].

A largo plazo se consideran además el precio del combustible, precio del equipamiento que usa la energía y factores socio-demográficos tales como el crecimiento de la población, grado de urbanización del país, etc.

El modus operandi clásico del sistema eléctrico de potencia es unidireccional y orientado de tal forma que la demanda total conocida por el servicio eléctrico es una suma de la demanda individual de un diverso conjunto de consumidores; cada uno con una cantidad desconocida y numerosa de dispositivos eléctricos vinculados a una barra del sistema. Tomando en cuenta el número de consumidores en el sistema de distribución, la deficiencia en la medición estacionaria, así como el registro de dispositivos en las barras del sistema, dificulta el análisis del comportamiento diario de la carga [2, 9].

Un número limitado de grandes plantas de generación alimentan la red para mantener satisfecha la demanda y el balance de energía en todo momento. La inestabilidad de las fuentes de energía renovables y dispositivos electro-móviles son nuevos retos para mantener este balance, lo que crea la necesidad de métodos de control más sofisticados.

2.1. Factores característicos de la carga

Para fines de operar y planificar el sistema de distribución de energía eléctrica, se requiere que el comportamiento de la demanda en el tiempo sea caracterizado mediante una serie de factores que la describen en cierto período muestreado. Un punto importante es que la demanda no se comporta de manera constante, se encuentra continuamente variando en el tiempo, y la representación de esta se hace mediante el conjunto de valores de la misma con respecto al tiempo, conocido como curva de carga y perfil de carga [8].

Curva de Carga: Es una representacion del comportamiento continuamente variable de la demanda, mediante el conjunto un valores de la misma con respecto al tiempo.

Factor de Demanda: Es la relación entre la demanda máxima de un consumidor en el tiempo muestreado y la carga total conectada. Este se calcula para cada consumidor de manera individual.

$$FD = \frac{P_{max}}{CTC} \tag{2.1}$$

Factor de carga: Corresponde a la relación entre la carga promedio y la carga máxima en el tiempo muestreado. Representa la porción de tiempo total de medición para la cual podría estimarse un pico de carga sostenido.

$$FC = \frac{P_{Prom}}{P_{max}} = \frac{\frac{1}{T}\int P(t)dt}{P_{max}} \tag{2.2}$$

Factor de Valle: Es la relación entre la carga mínima y la carga máxima dentro del tiempo evaluado. Mientras menor sea el valor mayor es la diferencia entre el valor mínimo y el valor máximo de la demanda.

$$F_{valle} = \frac{P_{min}}{\triangle P_{max}} \tag{2.3}$$

Factor de Pérdida: Relación entre las pérdidas de potencia activa promedio $\triangle P_{Prom}$ en la red en un período de tiempo y las pérdidas máximas en la red $\triangle P_{max}$.

$$F_{perd} = \frac{\triangle P_{Prom}}{\triangle P_{max}} \tag{2.4}$$

Factor de Potencia: Aplica a la relación entre la potencia activa y la potencia aparente, también se calcula como el coseno del ángulo entre tensión y corriente.

$$fp = \frac{P}{S} \tag{2.5}$$

Factor de Utilización: Es la relación entre la demanda máxima de potencia aparente S_{max} en un tiempo T y la capacidad nominal de un cierto equipo o sistema.

$$FU = \frac{S_{max}}{S_{nom}} \tag{2.6}$$

Factor de Diversidad: Relación entre la sumatoria de las demandas máximas de las cargas individuales y la demanda máxima del sistema.

$$Fdiv = \frac{\sum S_j^{max}}{S_{sistema}^{max}} \tag{2.7}$$

Factor de coincidencia: Se define como el inverso del factor de diversidad.

$$Fcoin = \frac{1}{Fdiv} \tag{2.8}$$

2.2. Medidor de Energía

El medidor es un dispositivo que permite la lectura de la energía consumida por cierto usuario en un período de tiempo (un día, un mes..) y le permite a la empresa el cobro del servicio por distribución de la energía, con lo cual esta puede mantener una calidad de servicio. Este dispositivo ha tenido una gran evolución con el tiempo y el avance de la tecnología, desde los primeros contadores de energía eléctrica de inducción cuyo funcionamiento se basaba en la reacción de la corriente de la carga sobre una bobina móvil, la cual al realizar su giro, permite la contabilización de energía activa de forma indirecta [10]. Hasta hoy en día con la aparición de los avanzados medidores inteligentes, pilar de las conocidas redes inteligentes.

Este último es un dispositivo, mayormente electrónico, que vincula múltiples funciones, permitiendo así la lectura de un conjunto de parámetros eléctricos, el almacenamiento de información, la comunicación alámbrica e inalámbrica de forma bidireccional, el corte y reconexión remotos, entre otros.

2.3. Gestión de la Demanda

El DSM es un conjunto de técnicas y políticas que pueden ayudar a proveer gran eficiencia a la hora de la planificación y operación de la red eléctrica, así como resolver cuestiones de sostenibilidad y conservación de energía [2]. DSM incluye todo aquello que es hecho del lado de la demanda en un sistema eléctrico. Comprende desde el cambio

de bombillos incandescentes por bombillos ahorradores, hasta la instalación de sistemas complejos para el manejo de cargas dinámicas [12]. En este ámbito se busca cada vez más, nuevas estrategias que involucren al consumidor en el manejo eficiente de la energía que consume en orden de ofrecer el mejor servicio a diferentes tipos de usuario o encontrar oportunidades en el mercado [2].

Su principal ventaja está en que es menos costoso influir en la demanda de forma inteligente, que construir una nueva infraestructura de generación o instalar muchos dispositivos de almacenamiento [12].

El manejo eficiente de la energía, ya sea en edificaciones o zonas industriales, comienza con la información y el entendimiento de los procesos involucrados [12]. Cada consumidor posee pequeños problemas ocultos que por si solos comparados con el sistema entero no son gran problema, pero al sumarse en gran número los pequeños problemas de cada uno aparecen coincidencias que pueden implicar despilfarro de energía que produce pérdidas para el servicio y disminución de la calidad para el propio y el resto de los usuarios.

Por ello, la adquisición de la información es sumamente importante, no se puede saber qué posee o no directamente el usuario dentro de su hogar sin preguntarle directamente, sin embargo, mediante una infraestructura de adquisición de la data, es decir, una adecuada infraestructura de medición y el posterior análisis de los datos generados por la misma, se puede conocer el patrón de consumo de dichos usuarios y ser agrupados por la similitud en su comportamiento.

Esta información permite a un especialista o un algoritmo inteligente interpretar el patrón de consumo y así el analista puede encontrar una forma de reducir los picos de carga, ya sea penalizando los picos o incentivando el consumo a ciertas horas a través de la implementación de una tarifa horaria [2].

Entre algunas de las técnicas más usadas en los programas de DSM están:

Recorte de picos: Es un programa de corte de carga, reducción de a demanda ante cargas pesadas. La reducción de la duración de los picos de demanda puede ser tratado a través del control directo de las cargas, apagando los equipos del consumidor o a través de la implementación de generación distribuida [2].

Conservación Estratégica: Este programa esta diseñado principalmente para la reducción de consumo de energía por temporada a través de un consumo eficiente. Incluye incentivos enfocados al cambio hacia tecnologías ahorradoras [2].

Carga del Edificio: Es un programa para controlar el incremento en el consumo de energía de un edificio según la temporada. El prestador de servicio emplea sistemas inteligentes,

procesos, equipos más eficientes y fuentes de energía más competentes para lograr una mayor eficiencia.

Llenado de Valles: Está enfocado al consumo fuera de los picos. Construye períodos de consumo fuera de los picos, los cual es particularmente deseable debido al bajo costo en generación. Esto causa un decremento en el precio promedio y mejora la eficiencia del sistema [2].

Cambio de carga: Consiste en transferir grandes bloques de demanda de los periodos pico a los de menor consumo sin cambiar el consumo total. Esto también es posible a través de generación distribuida [2].

Curva de Demanda Flexible: Es un conjunto de acciones y planes integrados entre el consumidor y el prestador de servicio, sujetoa la necesidad del momento. Es un trabajo en conjunto con el objetivo de moldear la carga del consumidor, sin afectar sus condiciones de seguridad, limitando la potencia y energía que el usuario puede consumir en ciertos instantes instalando dispositivos limitadores de carga [2].

2.4. Caracterización de la carga

La caracterización de la carga es un proceso en el cual se modela la misma como un patrón de consumo de potencia promedio en el tiempo, que representa el comportamiento del usuario en ese lapso. De este modelo es posible identificar factores que contribuyen en la evolución del sistema eléctrico, y así predecir el crecimiento de la demanda de una red existente o una nueva red de interconexión. Esto permite optimizar la planificación y manejo del sistema seleccionando las acciones mas apropiadas según el tipo de carga [8].

Según el artículo 61, de la ley para el servicio eléctrico vigente en la República Bolivariana de Venezuela, El Ministerio del Poder Popular con competencia en materia de energía eléctrica para la caracterización y clasificación de los usuarios, considerará los siguientes aspectos [15]:

- Uso de la energía eléctrica
- Nivel de consumo
- Ubicación geográfica
- Características técnicas del servicio suministrado

Este proceso se realiza a través de mediciones y encuestas que, a través de métodos estadísticos permiten conocer la forma en que cada usuario hace uso de la energía a través de distintos

dispositivos eléctricos y electrónicos asociados a su calidad de vida, trabajo, etc. [8] y además, conocer el patrón que más se asocia a cada estrato con el fin de mantener una clasificación de los usuarios según los aspectos anteriores, importante para la decisión de estratos tarifarios, planificación de operación, predicción de demanda, entre otros.

Existen múltiples métodos computacionales para la clasificación y modelación de estos patrones de demanda, entre las cuales se destacan en el estado del arte el uso modelos heurísticos, redes neuronales artificiales, sistemas expertos y técnicas de minería de datos. En el presente trabajo para este proceso nos centraremos en este último método.

CAPÍTULO 3

MINERÍA DE DATOS

La minería de datos es un campo de la estadística y la ciencia de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de datos [16], es el proceso estándar a la hora de recolectar información útil proveniente de un flujo de datos, como el consumo de usuarios, inyección de energía renovable, y estado de batería de un vehículo eléctrico, y transformar en algo entendible para su posterior uso. El proceso de minería de datos envuelve el uso de algoritmos para el descubrimiento de patrones entre datos siguiendo criterios de similitud [17].

En el problema de clasificación y caracterización de carga, un conjunto de puntos en una data pre clasificada alimenta al algoritmo de clasificación, el cual intenta descubrir una regla, que describe, tan preciso como sea posible, la clasificación observada. La clasificación de carga esta basada en un proceso de clustering, que es usado para definir grupos e identificar distribuciones en una data provista [17].

Definiciones importantes:

- **Media:** Es el valor obtenido al sumar todos los daros y dividir el resultado entre el numero de elementos.

$$\mu = \frac{1}{n}\sum_{i} x_i \tag{3.1}$$

- **Mediana:** Es el valor que ocupa el lugar central de todos los datos cuando estos se encuentran ordenados de menor a mayor.

- **Varianza:** Es el promedio del cuadrado de la desviación respecto a la media de una distribución estadística. La varianza intenta describir la dispersión de la data.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n} \tag{3.2}$$

- **Desviación típica:** Se define como la raíz cuadrada de la varianza. Es una medida de dispersión para variables cuantitativas y de intervalo.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}} \tag{3.3}$$

- **Dendograma:** Es una representación gráfica, en forma de árbol, en la cual se organiza la data en subcategorías que se van dividiendo en otras hasta llegar al nivel de detalle deseado.

3.1. Análisis de Clustering

Un cluster es un conjunto o grupo de data, cuyos elementos son similares entre sí, y disímiles entre ellos y otros clusters. El análisis de clusters incluye una colección de técnicas que permiten agrupar entidades multidimensionales de acuerdo a varios criterios de sus grados de homogeneidad y heterogeneidad, es decir, encontrar patrones de asociación formando grupos conformados por muestras de patrones más similares entre sí con respecto a otros grupos [18].

El problema central es repartir M individuos en K grupos de tal forma que en algún sentido los individuos de un grupo muestren homogeneidad y los grupos diferentes muestren heterogeneidad [18]. El análisis de clusters es distinto a una clasificación, ya que los elementos de la data no poseen etiquetas que los clasifiquen como partes de ciertos grupos; el algoritmo de clustering no parte de una hipótesis de pertenencia de estos objetos, sino que supone una la cual puede ir variando hasta identificar y agrupar cada patrón dentro de algún cluster.

3.1.1. Medidas de similitud

El proceso de clustering debe seguir un criterio, es allí donde el concepto de similitud es fundamental, pues es una medida de correspondencia entre los objetos que van a ser agrupados [5]. La similitud entre objetos puede medirse de varias formas, pero tres métodos dominan las aplicaciones en análisis cluster: medidas de correlación, medidas de distancia, y medidas de asociación [5].

Para el caso del análisis cluster en la identificación de consumidores del Sistema de Distribución de Energía Eléctrica, en el cual se emplea el algoritmo para identificar patrones de consumo, predominan las medidas de distancia.

- **Medidas de distancia**: Representan la similitud como la proximidad del objeto referente a otro objeto de la muestra. Son medidas de diferencia en donde los valores elevados indican una mayor dictancia entre los objetos [5]:

Considerando una muestra de $X = x_1, x_2, ..., X_n$ objetos a analizar, donde cada objeto consta de $m = 1, 2, \ldots, M$ elementos entonces:

1. Distancia Euclidiana:

$$d(x_i, x_j) = \sqrt{\sum_{m=1}^{M} (x_{im} - x_{jm})^2} \tag{3.4}$$

2. Distancia Manhattan:

$$d(x_i, x_j) = \sum_{m=1}^{M} (x_{im} - x_{jm}) \tag{3.5}$$

3. Distancia Minkowski:

$$d(x_i, x_j) = \sqrt[p]{\sum_{m=1}^{M} (x_{im} - x_{jm})^p} \tag{3.6}$$

Donde x_i,$x_j$son los dos objetos a comparar y $m = 1, 2, ..., M$ son los elementos dentro de cada objeto, siendo M el número total de elementos.

Al aplicar dichas distancias se obtiene una matriz triangular conocida como matriz de similaridad, cuyos elementos corresponden a la distancia entre cada objeto y el resto.

Propiedades de las medidas de distancia:

- $d(x_i, x_j) > 0$ si $i \neq j$, y $d(x_i, x_i) = 0$ ←Positividad
- $d(x_i, x_j) = d(x_j, x_i)$ ←Simetría
- $d(x_i, x_j) \leq d(x_i, x_k) + d(x_k, x_j)$ ←Desigualdad triangular

3.1.2. Métodos de Clustering

Existen muchos métodos de clustering en la literatura, y [4, 12, 14] muestran los más utilizados para la clasificación de consumidores en el sistema de distribución. Sin embargo, en este proyecto se compararon los métodos más relevantes dentro de la literatura consultada pera el objetivo a cumplir, los cuales son: Jerárquico, K-medias, K-medianas y Sigue al Líder Modificado (MFL).

3.1.2.1. Jerárquico

Este es un método en el cual se puede empezar desde una cantidad M de grupos o un solo grupo de M objetos. El primer caso se conoce como "aglomerativo", en el cual se posee M grupos que se van aglomerando para formar nuevos grupos según el criterio de distancia utilizado. El segundo caso se conoce como método "divisivo", en el cual el grupo M es desglosado hasta formar los K grupos pertinentes.

Se construye una matriz MxM usando alguna medida de distancia (la más utilizada es la distancia Euclidiana). A continuación las M muestras son agrupadas usando criterios de enlace basados en la matriz de similaridad calculada anteriormente. Este proceso se repite una y otra vez de forma iterativa ordenando los grupos de cada nivel: más grandes arriba; hasta que todas las muestras pertenecen a un grupo [4]. El historial del proceso se mantiene con la idea de formar una estructura de árbol binario que puede ser representada mediante un dendograma, como el mostrado en la figura 3.1.

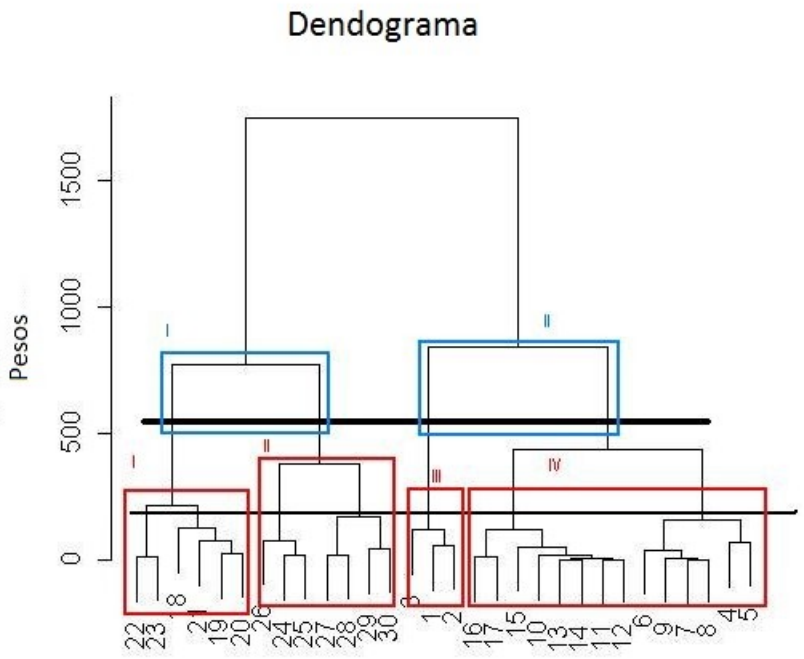


Figura 3.1: Dendograma resultante de un proceso de clustering jerárquico [19]

Para efectos de este proyecto se decidió optar por una variación de este algoritmo, en la cual se usa el criterio de similitud Ward. En dicho algoritmo, los grupos están formados de tal forma que se pueda minimizar la suma de los cuadrados de los grupos.

$$d_W(s,t) = \frac{n^{(s)} * n^{(t)}}{n^{(s)} + n^{(t)}} d^2(\bar{x}^{(s)}, \bar{x}^{(t)}) \tag{3.7}$$

La distancia $d_W(s,t)$ entre 2 grupos s y t es la medida del incremento de cuadrados si dos grupos se uniesen [4]: donde $n^{(s)}$ y $n^{(t)}$ son el n° de objetos en s y t, $\bar{x}^{(s)}$ y $\bar{x}^{(t)}$ son el centro de los 2 grupos, y $d^2(,)$ la distancia evaluada utilizando la norma Euclidiana.

3.1.2.2. K-medias

Es conocido también como método de partición. En este los K grupos son impuestos por el usuario, el número de grupos óptimo se define por ensayo y error.

En primer lugar se inicializa el algoritmo asignando al azar un grupo $k = 1, 2, ..., K$ para cada objeto de la muestra. El siguiente paso es calcular el centro $r = r_1, r_2, ..., R$ de cada grupo, computando este como la media entre los objetos pertenecientes al mismo grupo y se procede con el cálculo de la matriz de distancia entre cada objeto $m = 1, 2, ..., M$ y cada centro r_i. Los objetos luego serán reasignados al grupo más cercano. El proceso se repite iterativamente hasta que los grupos dejan de cambiar.

Este método es altamente usado, sin embargo, como el número de grupos es asignado a priori presenta la desventaja de ser sensible a los casos excepcionales en las muestras. Para resolver este detalle se han creado muchas variaciones: Fuzzy K-medias Difuso [4], WFA K-medias [5], entre otros.

3.1.2.3. K-medianas

Esta es una variación del conocido K-medias, en la cual se realiza el mismo procedimiento para la identificación de patrones en la muestra, sin embargo, el criterio de similitud utilizado en este algoritmo es la distancia Manhattan; por otro lado, se sustituye el cálculo de la media de los arreglos integrantes de un grupo para calcular los centros por el cálculo de la mediana. Esto permite que sea más robusto ante valores atípicos. En la figura 3.2 se muestra un flujograma del proceso de este algoritmo.

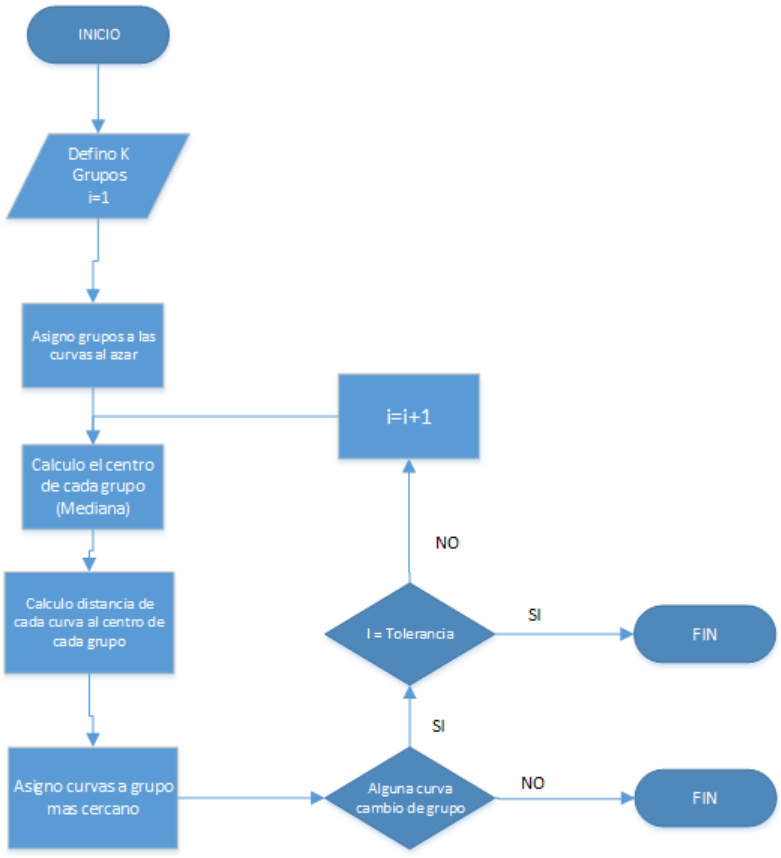


Figura 3.2: Flujograma K-medianas

3.1.2.4. Sigue al Líder Modificado

Una importante cualidad de este algoritmo es que no requiere establecer previamente los grupos, sino que los calcula sobre la marcha. Posee también una modificación en la distancia Euclidea implementada para calcular las distancia entre grupos; se le agrega un factor $\sigma_j^2/prom(\sigma^2)$, donde σ_j^2, es la varianza del j-ésimo objeto de la muestra total y $prom(\sigma^2)$ es el promedio entre las σ_j^2 [4, 14].

$$d(y^{(m)}, r_i^{(k)}) = \sqrt{\sum_{j=1}^{J} \frac{\sigma_j^2}{prom(\sigma^2)}(y_j^{(m)} - r_{ij}^{(k)})^2} \tag{3.8}$$

donde r_i^k es el centro k-ésimo durante la i-ésima iteración.

El procedimiento está compuesto por muchas iteraciones, en la primera iteración se asignan K clusters y el número $n^{(k)}$ de objetos pertenecientes a cada cluster, dependiendo de un umbral ρ. Todos los objetos son comparados secuencialmente, así, si la distancia euclidiana ponderada entre el objeto y un centro de alguno de los grupos existentes es la mínima entre los clusters y menor al umbral ρ, entonces ese objeto pasa a formar parte de dicho cluster y se recalculan los centros nuevamente. Así sucesivamente hasta que todos los objetos son asignados a su grupo, o cluster, más cercano q y se recalculan los centros $r^{(q)}$ [5].

Si la distancia al cluster más cercano no está dentro del umbral entonces ese objeto pasa a formar un grupo independiente. De esa forma los elementos se van moviendo de grupo en grupo según la distancia a los nuevos clusters hasta que el proceso termina en la i-ésima iteración cuando ya ningún objeto cambia de grupo o el algoritmo llega al límite de iteraciones [5]. Es posible observar el flujograma de este algoritmo en la figura 3.3.

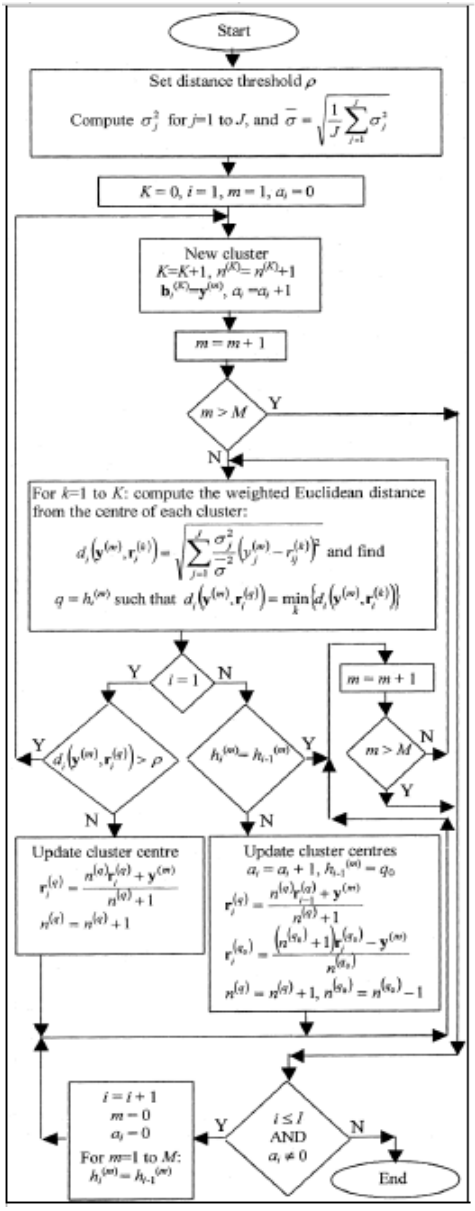


Figura 3.3: Flujograma Modified Follow de Leader [5]

Presenta la desventaja de una alta complejidad en su implementación, lo que lo hace además computacionalmente bastante lento. Es insensible a los casos excepcionales.

3.1.3. Validación de los clusters

Una vez realizado el proceso de clustering es conveniente validarlo a través de ciertas métricas, que permiten conocer la eficiencia del algoritmo al agrupar la data en K clusters. para ello definimos primero:

Distancia entre dos vectores $l^{(i)}$ y $l^{(j)}$ pertenecientes a un subconjunto $L^{(K)}$:

$$d(l^{(i)}, l^{(j)}) = \sqrt{\frac{1}{M} \sum_{m=1}^{M} (l_m^{(i)} - l_m^{(j)})^2} \tag{3.9}$$

donde M es el numero de elementos de cada vector.

Distancia del centro del grupo $r^{(k)}$ al subconjunto de objetos que lo constituyen $L^{(k)}$:

$$d(r^{(k)}, L^{(k)}) = \sqrt{\frac{1}{n^{(K)}} \sum_{m=1}^{n^{(K)}} d^2(r^{(k)}, l^{(m)})} \tag{3.10}$$

donde $n^{(k)}$ es el numero de objetos en un cluster.

Hecho esto, las medidas de eficiencia están definidas por :

1. Mean Index Adequacy (MIA): se define como la distancia promedio entre cada vector asignado a un cluster y su centro [5]:

$$MIA = \sqrt{\frac{1}{K} \sum_{k=1}^{K} d^2(r^{(k)}, l^{(k)})} \tag{3.11}$$

   esta medida está basada en la distancia Euclidiana entre los diagramas de carga y los centros de los grupos representados [4]; y muestra que tan distintos son entre sí los clusters. Algoritmos con mejores resultados en este índice implican mayor eficiencia para aplicaciones en las que se necesitan grupos bien diferenciados [5].

2. Clustering Dispersion Indicator (CDI): Depende de la distancia entre objetos en el mismo cluster, e inversamente, de la distancia entre centros de grupo [5].

$$CDI = \frac{\sqrt{\frac{1}{K} \sum_{k=1}^{K} [\frac{1}{2n^{(k)}} \sum_{n=1}^{n^{(k)}} d^2(l^{(n)}, L^{(k)})]}}{\sqrt{\frac{1}{2K} \sum_{k=1}^{K} d^2(r^{(k)}, R))}} \tag{3.12}$$

   Esta medida arroja información acerca de qué tan compactos son los grupos a los que pertenece cada curva [4]. Permite reconocer la homogeneidad de los grupos formados.

En ambos casos, MIA y CDI, mientras menor es el valor mejor; más eficiente es el algoritmo.

Dicho ya esto, una de las preguntas más importantes al utilizar cualquiera de los métodos de clustering, si no la más importante, es: ¿ En cuántos grupos debe clasificarse la muestra?

Para el caso de los algoritmos Jerárquicos y Sigue al Líder Modificado, es una pregunta que se responde a través del ensayo y error. En el primer caso observando el dendograma se realiza el corte en el árbol hasta obtener el número de grupos ideal, decidido por el analista;

mientras que en el segundo caso es necesario correr el algoritmo sucesivas veces variando el umbral hasta que se obtiene la cantidad de grupos más idónea [5]. Es importante destacar que ya que este algoritmo decide la cantidad de grupos a clasificar la muestra según el umbral, es posible que no pueda formar cierto número de grupos.

Otro método para la estimación de grupos a clasificar es el amalgamiento, que consiste en descubrir de forma visual el punto en donde la distancia entre grupos (MIA) empieza a variar de forma menos significativa. También se conoce como curva del codo y es frecuentemente utilizado para determinar el número de grupos a preestablecer en k-medias.

En la figura 3.4 se observa una curva de amalgamiento evaluada a través del algoritmo *K-medianas*, de 2 a 30 grupos, utilizando data publica proporcionada por la empresa de servicio eléctrico PSEG, de New Jersey, y ya se encuentra clasificada en 6 grupos.

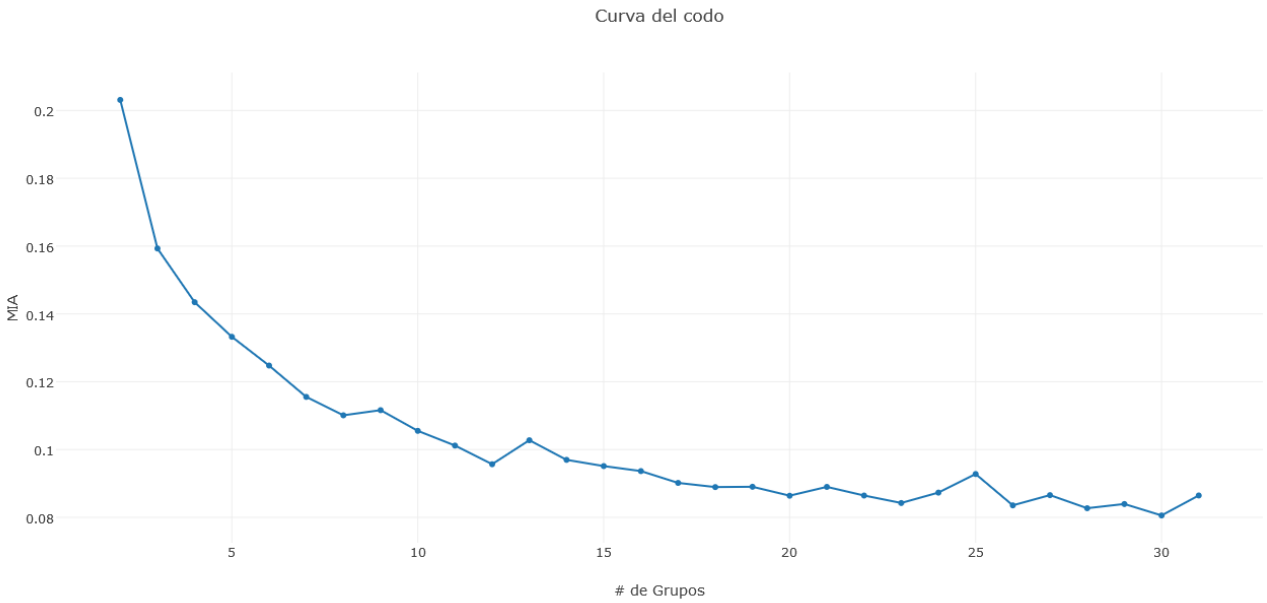
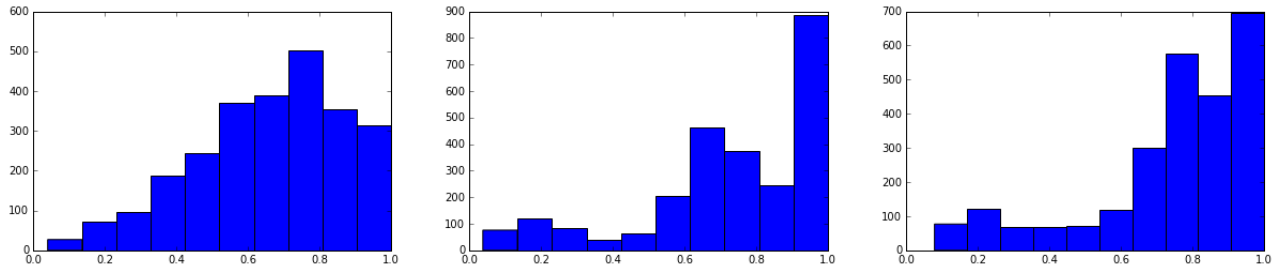


Figura 3.4: Curva del codo evaluando de 2 a 30 grupos

3.2. Selección del algoritmo

Tres de los algoritmos antes descritos K-medias, Kmedianas y Sigue al Líder Modificado fueron programados en la herramienta computacional de software libre Python 3.5, y se compararon entre si utilizando los índices de eficiencia MIA y CDI de 2 a 15 grupos para ver su comportamiento, además se plantea el algoritmo Jerárquico "ward" de la librería sklearn.cluster para comparar con un paquete comercial, utilizando el mismo set de curvas promedio PSEG, el cual es público y accesible a través de [20]. En la figura 3.5 se muestra

el histograma de dicha data a 3 horas distintas del día. El hardware usado para su corrida fue un ordenador con procesador intel core i7 de 2.5 GHz, con 8 Gb de memoria Ram, 1 Gb GPU. Los resultados obtenidos se muestran en las tablas 3.1 y 3.2:



(a) Histograma Data PEG 7 a.m. (b) Histograma Data PEG 12 p.m. (c) Histograma Data PEG 6 p.m.

Figura 3.5: Histograma Data usada para la prueba de los algoritmos. Eje x: Potencia máxima normalizada. Eje y: Frecuencia

Tabla 3.1: Indice Mia, de 2 a 15 grupos

grupo	Jerárquico	K-medias	K-medianas	MFL
2	0.175651993	0.173757669	0.180005957	NA
3	0.153962122	0.152728316	0.159316328	NA
4	0.142727541	0.136759913	0.143441820	NA
5	0.127441498	0.127901177	0.136158489	NA
6	0.116381046	0.117518763	0.127954101	NA
7	0.115418174	0.110955959	0.115556112	NA
8	0.110913108	0.106246326	0.109154706	NA
9	0.115550717	0.114232944	0.108559947	NA
10	0.110728389	0.114482003	0.103704044	NA
11	0.105920869	0.105342847	0.100349275	NA
12	0.103163933	0.09825277	0.102165693	NA
13	0.100605872	0.102386322	0.092529366	NA
14	0.097707268	0.108374018	0.092175372	NA
15	0.104914774	0.097468290	0.094748064	NA

MFL ofrece un resultado al converger todos los elementos hacia un estrato específico del cual no cambiará en una próxima iteración o al cumplir un número máximo de 1000 iteraciones (Ese número puede cambiar de manera subjetiva dependiendo del criterio de la persona que lo usa). Sin embargo, el MFL programado en este proyecto no logro converger ni

superar el numero máximo de iteraciones en un tiempo aceptable, pues se permitió la corrida de la simulación en un máximo de 12 horas en las cuales el algoritmo solo logro culminar 3 iteraciones. Por lo que no presenta entradas en la tabla comparativa.

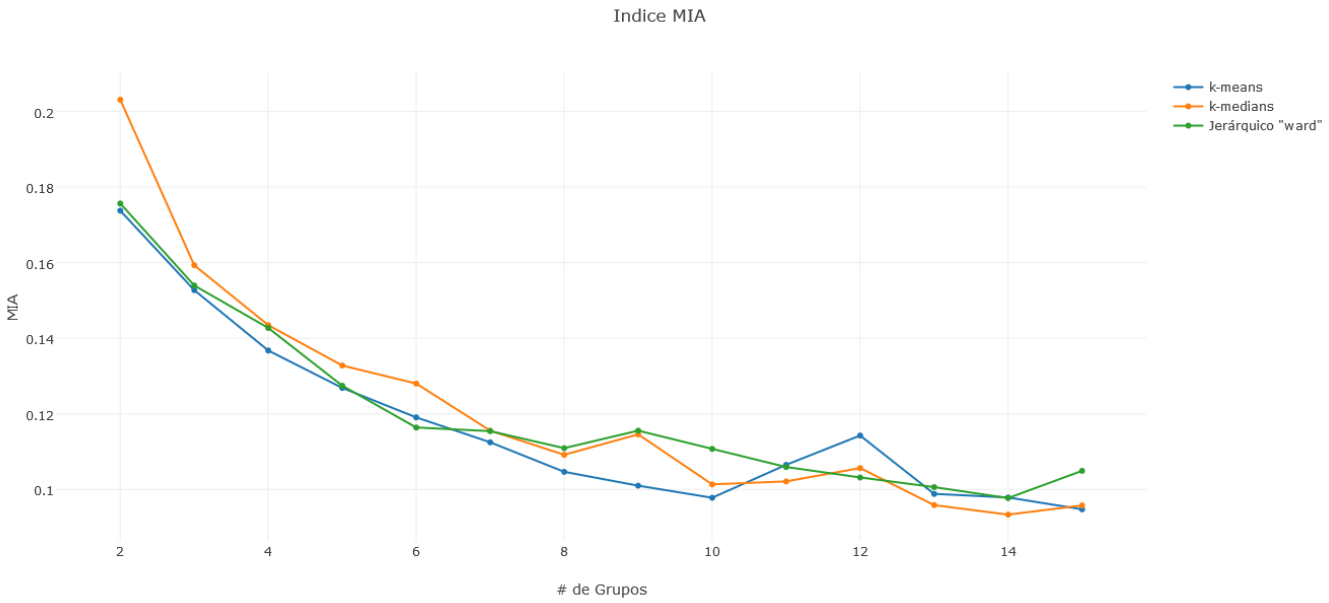


Figura 3.6: Curva representativa de MIA para distintos algoritmos de clustering

En la figura 3.6 se puede observar de forma gráfica la comparación entre los resultados del índice aplicado a los algoritmos Jerárquico, K-medias y K-medianas; no hay mucha diferencia entre la eficiencia de los algoritmos con respecto a su diferenciación de unos grupos con respecto a otros, sin embargo, k-medias a primera vista aparenta ser el más óptimo, hasta el décimo grupo, luego de ello se hace menos estable.

Tabla 3.2: Índice CDI, de 2 a 15 grupos

grupo	Jerárquico	K-medias	K-medianas	MFL
2	9.24024	9.13247	23.45842	NA
3	5.50856	5.41021	4.74890	NA
4	3.80866	3.68454	3.25705	NA
5	2.88589	2.82144	2.63025	NA
6	2.25302	2.18971	1.99397	NA
7	1.60490	1.84009	1.88328	NA
8	1.37942	1.36875	1.23221	NA
9	1.13799	1.14077	1.10786	NA
10	0.98859	1.14450	1.19427	NA
11	0.89554	0.94926	1.09867	NA
12	0.82358	0.87943	0.82253	NA
13	0.76000	0.91188	0.69652	NA
14	0.70705	0.83167	0.71308	NA
15	0.64697	0.65580	0.59225	NA

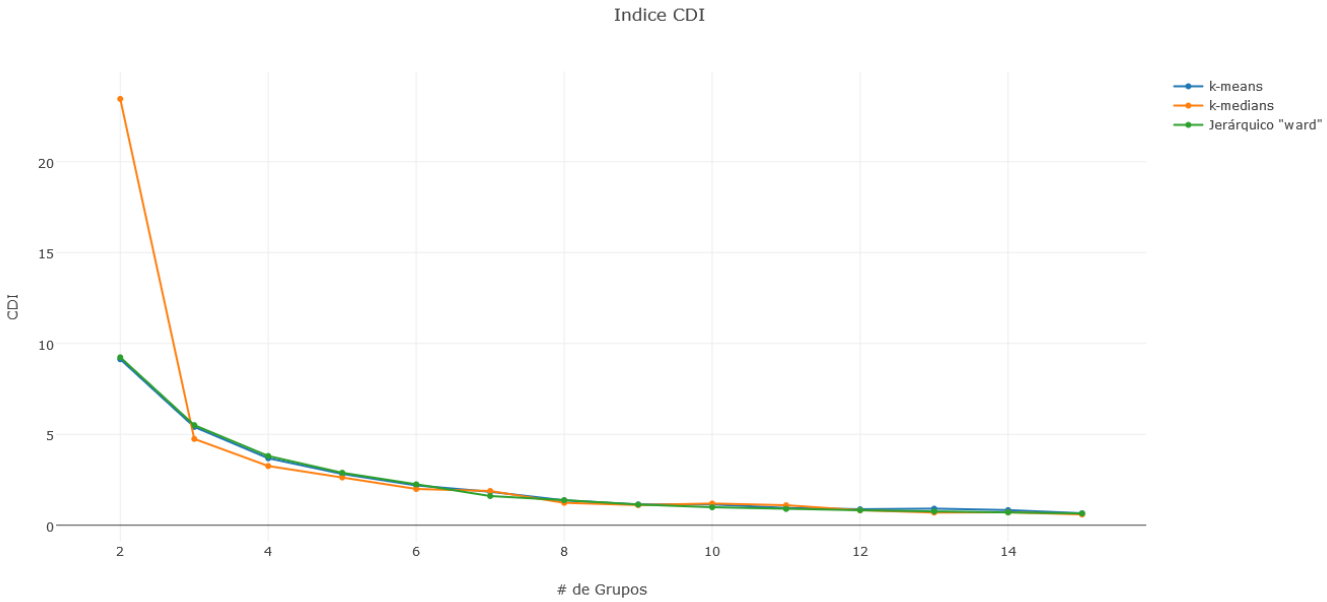


Figura 3.7: Curva representativa de CDI para distintos algoritmos de clustering

En figura 3.7 la curva resultado de la aplicación del índice CDI, desde 2 hasta 15 grupos, a los algoritmos evaluados. A partir del tercer grupo todos los algoritmos presentan un comportamiento estable en su clasificación con respecto a la similitud dentro de cada grupo.

Considerando que el objetivo es caracterizar el uso de la energía contratada por los consumidores, y esta información se obtiene a través de dispositivos, pueden surgir situaciones donde las mediciones de estos no sean precisas ya sea por una avería del medidor, a causa de una falla interna, daños ocasionados por un tercero, o simplemente el comportamiento normalmente estocástico del objetivo a medir. Es muy probable que se presenten mediciones atípicas, falta de datos o valores extremos que no se relacionan en nada con el comportamiento normalizado de la población, por lo que el algoritmo a escoger debe además ser robusto ante estos valores. Para poder simular esta situación se introdujeron valores extremadamente altos de manera aleatoria en el conjunto de datos evaluado y se calcularon los indices nuevamente para observar su comportamiento.

Tabla 3.3: Índice MIA, de 2 a 15 grupos

grupos	Jerárquico	K-medias	K-medianas	MFL
2	9.50174	0.38999	0.39063	NA
3	0.19050	0.34484	0.44160	NA
4	0.11126	0.40209	0.40791	NA
5	0.11114	0.35552	0.36858	NA
6	0.10868	4.64408	0.46991	NA
7	0.10734	4.29965	0.49653	NA
8	0.10057	0.30521	0.36886	NA
9	0.09464	3.79231	0.35306	NA
10	0.09623	3.59786	0.33389	NA
11	0.09353	3.43081	0.41272	NA
12	0.09915	3.28476	0.39809	NA
13	0.09599	3.15622	0.29874	NA
14	0.09386	3.04156	0.28886	NA
15	0.09203	2.93829	0.28116	NA

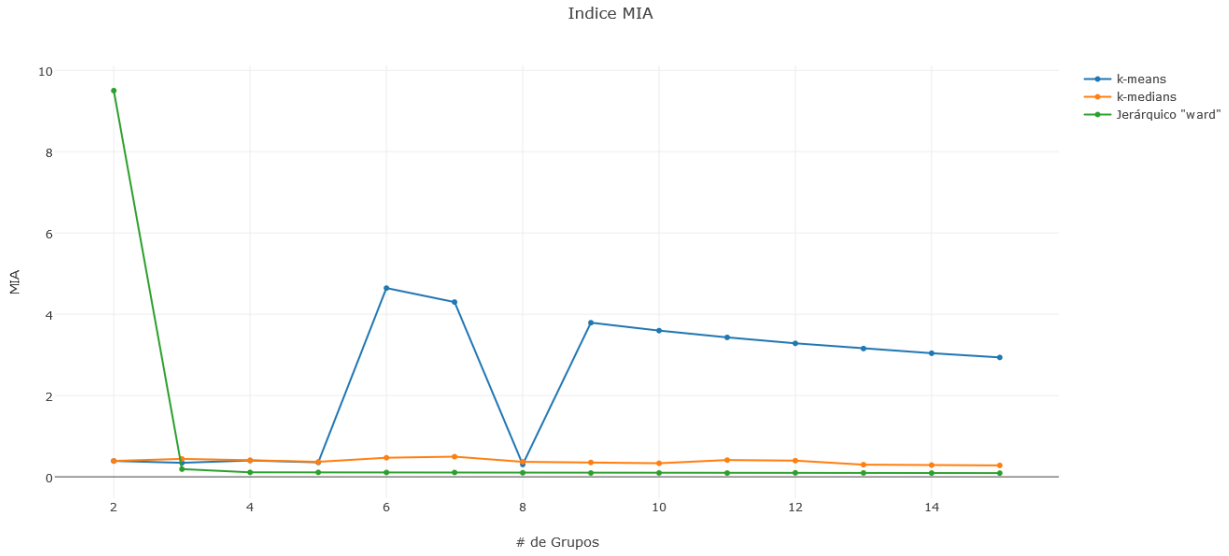


Figura 3.8: Curva representativa de MIA para distintos algoritmos de clustering

La tabla 3.3 y la figura 3.8 se muestran los resultados del índice MIA los cuales muestran un indice muy alto para el caso de dos grupos, sin embargo aparenta el mejor resultado para el resto de los casos evaluados, en segundo lugar se encuentra el K-medianas el cual obtiene un índice bastante bajo y se mantiene estable entre todos los casos. Por ultimo se observa que el K-medias posee un resultado contraproducente a partir del caso evaluado para 5 grupos y un comportamiento bastante inestable a partir de allí; esto se puede atribuir a que como el operador promedio usado en este método esta compuesto por la sumatoria todos los elementos entre la cantidad de los mismos el resultado puede ser muy influenciado por términos de magnitud muy alta o muy baja.

Tabla 3.4: Índice CDI, de 2 a 15 grupos

grupo	Jerárquico	K-medias	K-medianas	MFL
2	0.76291	0.68782	0.68797	NA
3	0.20637	0.46506	0.47754	NA
4	0.10511	0.37572	0.37571	NA
5	0.06469	0.32411	0.32428	NA
6	0.05076	0.16066	0.28730	NA
7	0.04168	0.14405	0.26089	NA
8	0.03604	0.24317	0.24160	NA
9	0.03134	0.12275	0.22576	NA
10	0.02811	0.11542	0.21276	NA
11	0.02543	0.10928	0.20069	NA
12	0.02308	0.10338	0.19124	NA
13	0.02160	0.09942	0.18399	NA
14	0.02033	0.09481	0.17674	NA
15	0.01923	0.09124	0.17023	NA

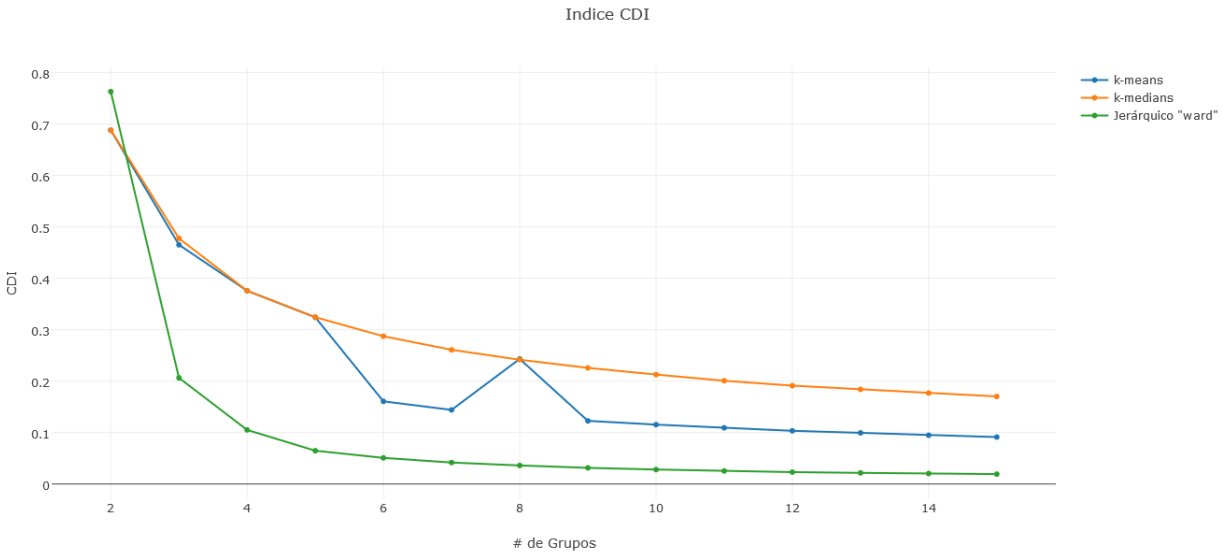


Figura 3.9: Curva representativa de CDI para distintos algoritmos de clustering

La tabla 3.4 y la figura 3.9 muestran los resultados al evaluar el caso con introducción de valores atípicos extremos en el conjunto de datos, en este caso todos los métodos muestran indices bajos lo que implica un buen ajuste para los propuestos, es decir, que los grupos formados por cada método son homogéneos en si mismos. El método que presenta

mayor homogeneidad es el Jerárquico, seguido por el K-medias, el cual presenta mejor comportamiento a partir de seis grupos.

Considerando los casos anteriores el primer algoritmo a tomar en cuenta es el Jerárquico, sin embargo la naturaleza aglomerativa con la que realiza la estratificación no permite considerarlo como primera opción para el objetivo de estratificación de consumidores de energía eléctrica. Esto se debe a que a pesar de que efectivamente separa los usuarios en estratos bien diferenciables entre si, al encontrar un consumidor con mediciones atípicas separa este usuario como un único estrato sin ningún otro usuario asociado.

Por esa razón, se decide tomar como algoritmo principal para la plataforma el algoritmo de K-medianas el cual ha mostrado buenos resultados tanto en las condiciones naturales del conjunto de datos evaluado como en el caso al cual se le introdujeron mediciones atípicas, observándose un mejor resultado al momento de realizar estratos diferenciados entre si en comparación con el K-medias.

CAPÍTULO 4

METODOLOGÍA Y DESCRIPCIÓN DE LA PLATAFORMA PARA CARACTERIZACIÓN DE CARGA

4.1. Descripción general

El presente capítulo esta guiado a explicar la metodología empleada para la caracterización de la carga y funcionamiento de la herramienta computacional desarrollada para la implementación de la misma. El software está compuesto por un programa principal llamado *Caracterizacion .py* y una serie de funciones enfocadas a la extracción y procesamiento de los datos, las cuales se encuentran empaquetadas en una librería de nombre *pyCyDes,* y han sido desarrollados usando el lenguaje de programación en software libre Python 3.5, debido a que es un lenguaje con una amplia gama de herramientas de análisis y desarrollo que permite la fácil manipulación de la data.

Otras librerías y funciones utilizadas en el desarrollo y corrida del algoritmo son:

- Numpy
- Pandas
- sklearn.metrics.pairwise: manhattan_distances, euclidean_distances
- plotly.graph_objs: Scatter, Pie, Box, Figure
- plotly.offline: plot
- time: time.time
- matplotlib.pyplot

La implementación de este trabajo permitirá al analista, utilizando técnicas de minería de datos, obtener información de miles de usuarios del sistema eléctrico que relaciona su consumo entre sí a modo de estratos que agrupan las tendencias actuales en el mismo, describir de manera muy cercana los patrones que sus vidas cotidianas reflejan en la demanda energética y la contribución de cada grupo en aquellas horas en que es más crítica para el sistema, cuando amenaza la capacidad del mismo; o en que el precio de generar esa energía es mayor.

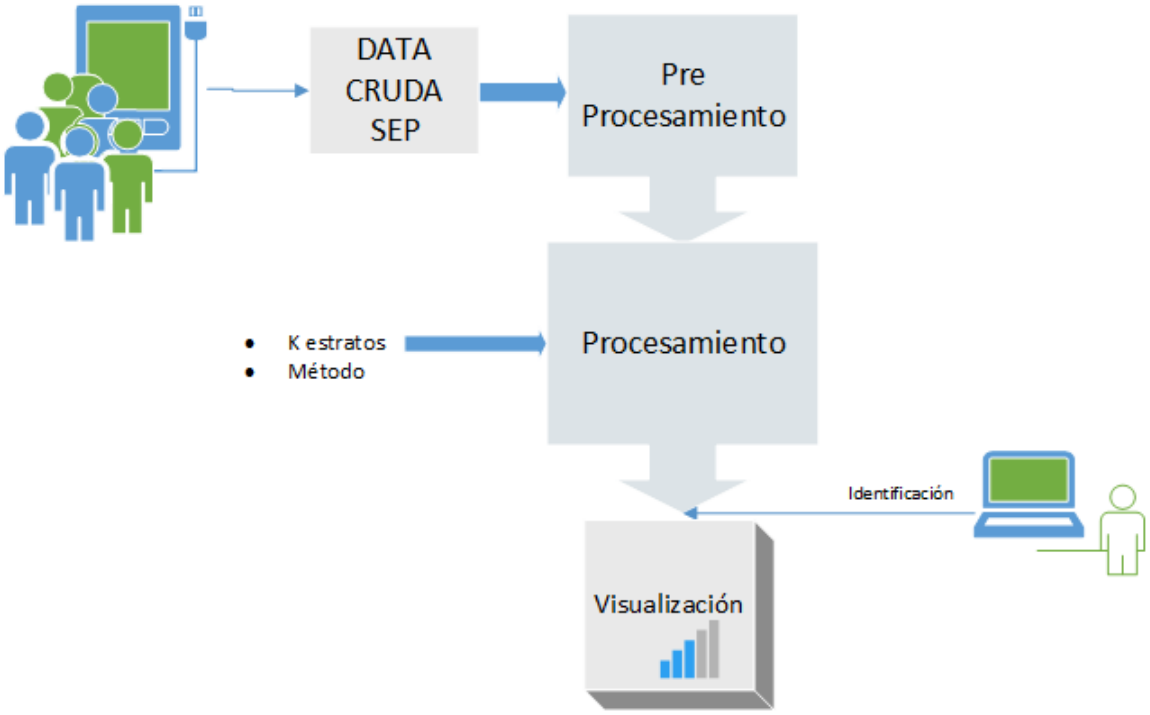


Figura 4.1: Esquema de la metodología por etapas

La metodología seguida es posible dividirla en 4 fases principales, como se muestra en la figura 4.1: Recopilación de la data, Pre-procesamiento, Procesamiento y Visualización.

La idea principal del software es la de tomar la data adquirida por el servicio eléctrico y realizar luego una fase de pre-procesamiento que la prepara y normaliza para así poder ser procesada correctamente en la siguiente fase, en la cual se obtienen los estratos para ser mostrada posteriormente al analista a través de la librería *pyplot* dedicada para gráficos interactivos. En la figura 4.2 se puede observar el flujograma que describe los procesos inherentes al programa principal.

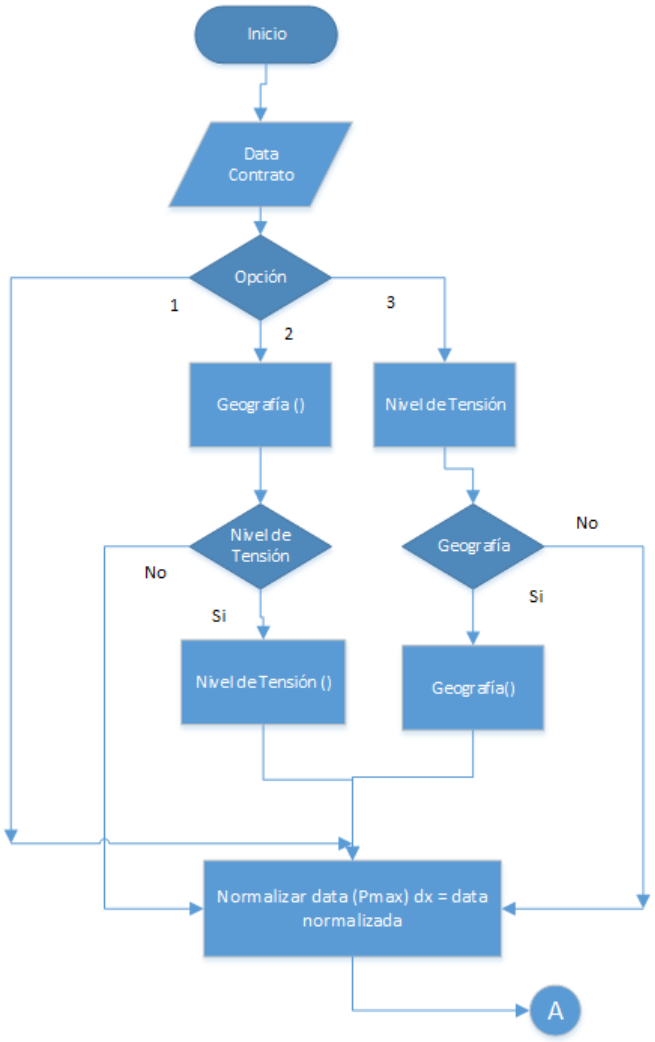


Figura 4.2: (a)Flujograma correspondiente a *Caracterizacion.py*

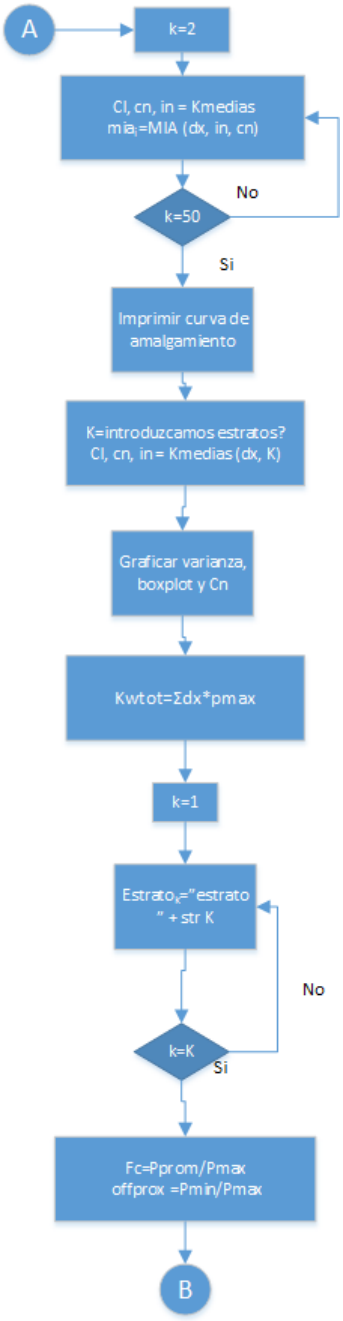


Figura 4.3: (b)Flujograma correspondiente a *Caracterizacion.py*

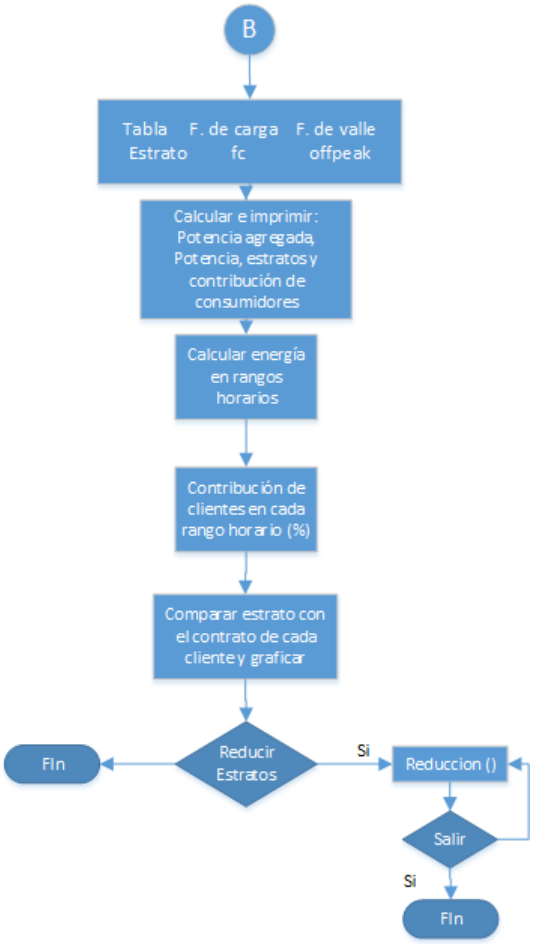


Figura 4.4: (c)Flujograma correspondiente a *Caracterizacion.py*

4.2. Recopilación de la data

En primera instancia es necesario conocer la población y establecer los criterios necesarios para seleccionar una muestra representativa. Una vez definido esto la fase de recopilación de la data continua con la toma de las mediciones de cada uno de los usuarios a estratificar. En un escenario en el cual existe una infraestructura de medición inteligente AMI esto es realizado por un medidor inteligente correspondiente a cada cliente del servicio, que toma la medición en cada intervalo de tiempo (15 min, 30 min o 1 hr) y envía la información a una estructura centralizada que recibe y procesa la data.

En el escenario actual del sistema venezolano no existe aun este tipo de infraestructura, por lo que la adquisición se lleva a cabo a través de una campaña de medición, en la cual se despliega por zona una cantidad de medidores (pertenecientes a la compaña eléctrica o contratados a una compañía externa) y luego es recolectada la información de vuelta para su procesamiento.

4.3. Pre-procesamiento

Luego de adquirida la data, esta debe pasar por un proceso de preparación previa para poder ser utilizada y manipulada correctamente en el proceso de estratificación. Este pre-procesamiento consiste en transformar los datos de las mediciones de consumo de energía de cada usuario a un formato de fácil manejo y luego de ello estos pasan por dos subprocesos importantes: sincronización y depuración.

Para llevar la data a una forma fácil de manipular dentro del programa la primera acción es extraer la información adquirida desde un formato .csv, el cual contiene las mediciones de potencia activa en intervalos de tiempo, más el contrato correspondiente de dicho cliente con el servicio eléctrico; y se convierte respectivamente en un objeto con características de tabla del tipo <pandas.DataFrame> que permite una fácil manipulación y búsqueda de los datos obtenidos a través de las funciones de las librerías Pandas y Numpy.

4.3.1. Sincronización

Ya sea el instante inicial de medición entre todos los medidores inteligentes o el momento de instalación e inicio de los medidores en una campaña de medición, ocurre que muy pocas de las mediciones empezarán el registro al mismo tiempo, y en algunos casos los intervalos de adquisición de la data entre medidores es distinto; algunos podrían estar registrando las mediciones cada minuto, mientras que otros cada hora o cada 15 minutos. Es por esto que el primer paso del pre-procesamiento es sincronizar las mediciones.

Este proceso comienza recorriendo cada uno de los consumidores desde el primer instante registrado por alguno de los medidores; revisando que exista o no mediciones vacías (expresadas en forma de NaN, utilizado para representar contenido dentro de la tabla que no son números) hasta que llega a una posición en la que todos los usuarios tienen un valor medido al mismo tiempo y envía un mensaje al analista mostrando a partir de qué día es posible hacer un estudio de la población completa. Seguidamente el analista procede a asignar una fecha en la cual se realizará la medición, y se obtiene una tabla con las posiciones de tiempo de cada medición para ese día específico en cada usuario.

Luego se sincronizan las mediciones de todos los usuarios al común denominador de las frecuencias de registro; por ejemplo, suponiendo que se poseen tres usuarios de los cuales el primero posee intervalos de medición de 1 s, el segundo a 5 min y el tercero a 15 min, se ajustan todos a intervalos de 15 min.

4.3.2. Depuración

El siguiente paso es identificar y eliminar, de las muestras obtenidas, las discontinuidades causadas por elementos NaN entre sus valores, y posteriormente los arreglos

con este tipo de defectos son sacados de la muestra, tanto de la tabla de medición de potencia como la correspondiente al contrato.

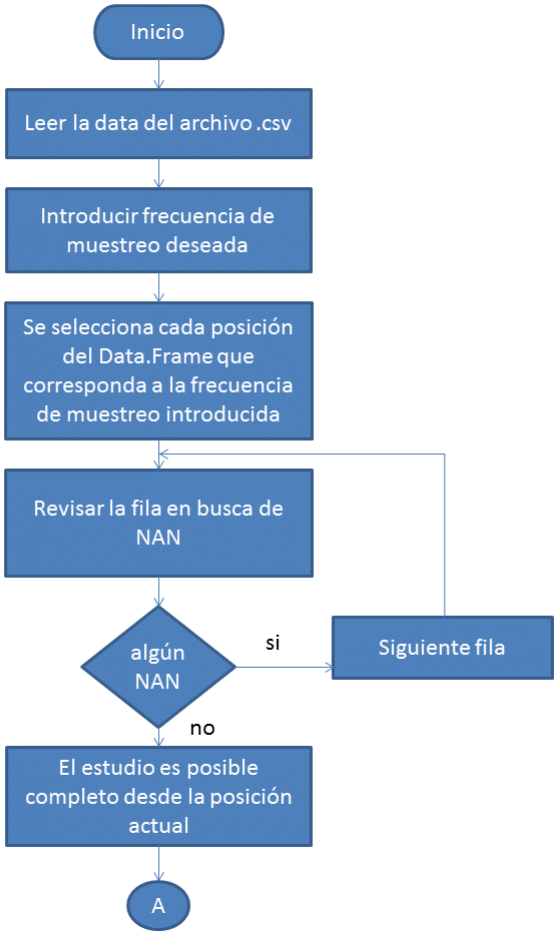


Figura 4.5: (a) Diagrama de flujo del Pre-procesamiento

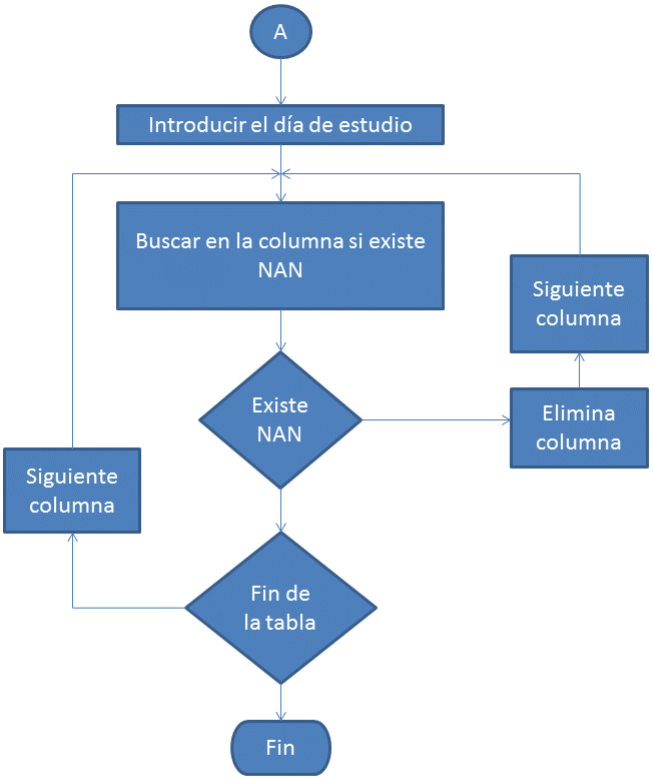


Figura 4.6: (b) Diagrama de flujo del Pre-procesamiento

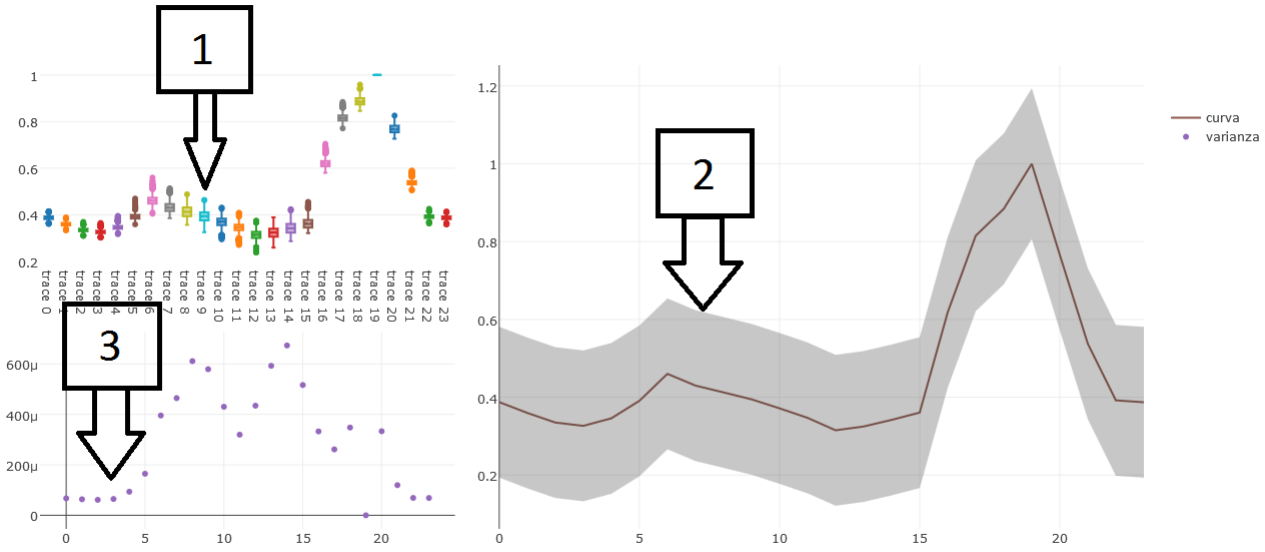
4.4. Procesamiento y Visualización

En esta fase se recibe la data resultante del pre-procesamiento, que consta de una matriz con los valores de medición de los consumidores. Luego de ello, se normaliza la data para trabajar con valores no mayores a 1, para esto se usa como valor de normalización la potencia máxima de cada usuario y se almacena en forma de arreglo en la variable *Pmax*. Este es un paso necesario para la comparación de los patrones, de otro modo, las medidas de similitud utilizadas medirían la diferencia de energía entre las curvas en vez de la distancia entre ellas, asociando a los usuarios en base a la energía demandada pero sin importar el perfil que se asocia a la actividad que realizan. Al igual que en la fase del pre-procesamiento, en esta etapa se tiene que seguir varios pasos como: estratificación, identificación, potencia agregada y contribución en picos horarios.

4.4.1. Estratificación

El siguiente paso es la estratificación de los usuarios, para lo cual se emplea el algoritmo escogido en el Capítulo 3. Utilizando el método de K-medianas se evalúa la data para obtener la curva representativa de cada estrato y la distribución de los usuarios que pertenecen a cada uno de ellos. Para escoger el número de estratos a generar se computa el índice MIA iterativamente para crear una curva de amalgamiento que le dé al analista un punto de partida.

Una vez se obtienen tantos modelos de curvas caracterizadas como grupos de clasificación estas se muestran al analista y se procede a calcular los factores de carga y de valle de cada uno de los estratos adquiridos, para así dar al analista un factor numérico que complemente la gráfica y le ayude en la identificación del tipo de actividad realizada o contrato asociado al patrón de carga estratificado, figura 4.7(b).



(a) Gráficos asociados al estrato definido

	F. de Carga	F. valle
Estrato 1	0.559443	0.110936
Estrato 2	0.474557	0.315024
Estrato 3	0.841612	0.536427

(b) Factores que definen la demanda en los estratos

Figura 4.7: Primer gráfico resultante de la caracterización de los patrones de consumo

En la figura 4.7(a) se muestra una gráfica generada al culminar el proceso de caracterización de las curvas. Esta está compuesta por tres gráficos, el gráfico 1 muestra

por cada hora un diagrama de caja o conocido también como boxplot que permite ver la distorsión en un conjunto de variables, con información referente a la media, mediana, valores máximo y mínimo, y muy útil para observar valores atípicos (aquellos que exceden de las líneas más lejanas de la caja). El gráfico 2 muestra una línea central que corresponde al patrón normalizado del conjunto de cargas caracterizado en el estrato actual, y alrededor de esta línea un sombreado gris que es producto de $\pm\sigma$(la desviación estándar). El gráfico 3 muestra cuanto se aleja o se acercan los valores, del set de datos estudiado, de la media utilizando como índice de medida la varianza. Como se puede notar las fases de Procesamiento y Visualización van actuando al mismo tiempo a medida que avanza el algoritmo.

4.4.2. Identificación

Luego de que han sido estratificados y graficados los grupos, inicia un proceso de identificación. Si la muestra pertenece a una población la cual esta establecida y posee un contrato vigente (una estratificación previa) se procede a comparar la cantidad de usuarios de cada estrato con su clasificación previa. De no poseer un contrato, es posible para el analista identificar los estratos asignando un nombre según el uso que le dan esos clientes a la energía; infiriendo esta información a partir de las curva media y los factores que la caracterizan los estratos obtenidos, figura 4.8.

```
define el tipo de consumidor asociado al estrato 1: com

define el tipo de consumidor asociado al estrato 2: ind

define el tipo de consumidor asociado al estrato 3: res
```

Figura 4.8: Identificación de cada estrato

De esta identificación es posible observar aquellos clientes cuyo patrón de consumo no corresponde al asociado con el patrimonio que posee. Un ejemplo de esta situación puede ser una persona que posee una tarifa residencial, sin embargo, monta en su casa un negocio, de soldadura, comida rápida, una lavandería, entre muchas posibilidades; y por lo tanto al final del día su consumo real tiende al comportamiento de un estrato no residencial comercial, y se debe poner atención en él para evitar irregularidades.

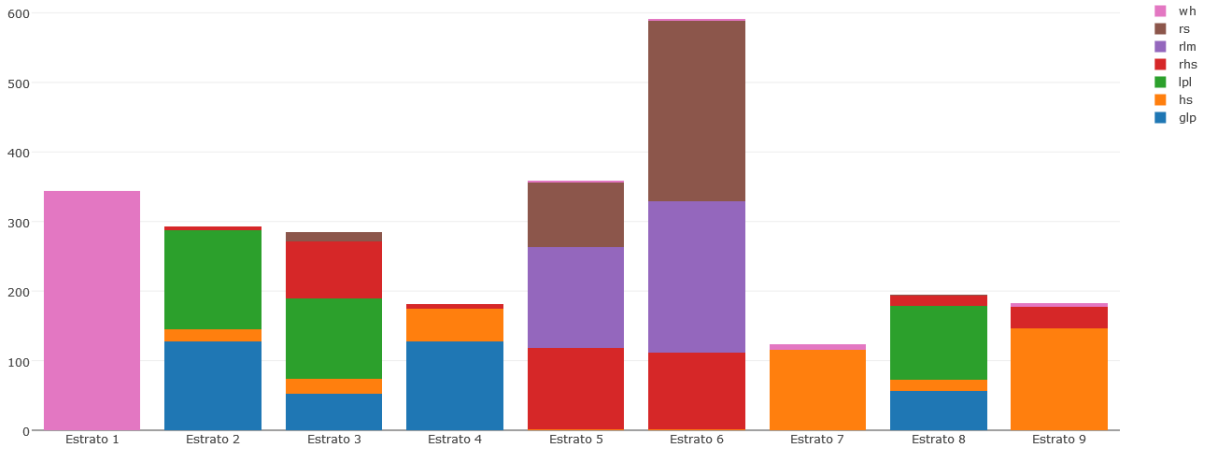


Figura 4.9: Estratos identificados según el uso de la energía. Eje x: No de Usuarios. Eje y: Estratos

4.4.3. Potencia Agregada

Luego de que se han definido los estratos se calcula la potencia agregada de la población estudiada, para esto, la variable principal normalizada es multiplicada por su potencia máxima correspondiente $Pmax_i$ y se suma la potencia de todos los usuarios que componen el estudio en una curva de potencia agregada; al mismo tiempo se toma el modelo de cada estrato $k = 1, ..., K$ y se multiplica por la sumatoria de las variables de normalización $\sum_{i=1}^{I} Pmax_i^k$ de cada grupo de usuarios perteneciente a dicho estrato, luego se suma la totalidad de las curvas para así obtener una curva de potencia estimada, que corresponde a la estimación del consumo de la totalidad de usuarios, basado en los modelos caracterizados.

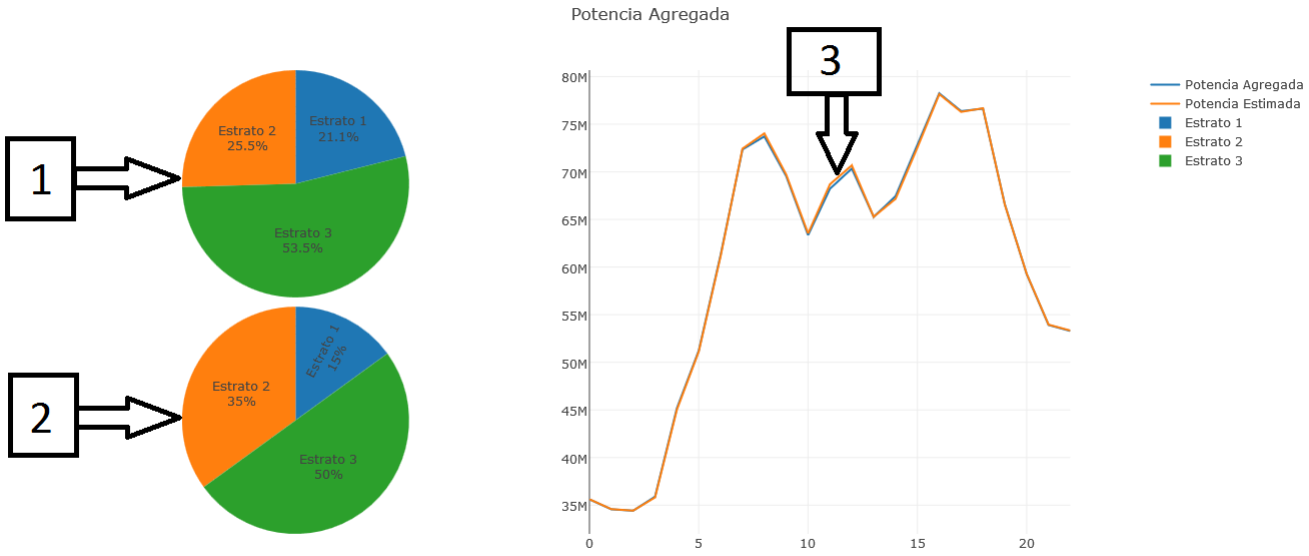


Figura 4.10: Gráfica de potencia agregada. Eje x: tiempo - 24 horas en intervalos de 1 hora.. Eje y: Vatios (w)

La figura 4.10 muestra la gráfica resultante del proceso antes descrito. La gráfica 1 corresponde a la distribución porcentual de la contribución de los consumidores identificados al total de potencia demandada al sistema; la gráfica 2 corresponde a la distribución numérica y porcentual de la cantidad de usuarios que componen cada estrato; y la gráfica 3 corresponde a la comparación de la potencia agregada y la potencia estimada con los modelos de demanda estratificados.

4.4.4. Contribución en picos horarios

Hay momentos en el día en que el consumo de los usuarios tiende a coincidir en mayor escala, lo que atenta contra la capacidad de un sistema que no está preparado para el crecimiento poblacional que ha enfrentado, por lo que una vez calculada la potencia total agregada se calcula el consumo de energía en la curva de la misma, utilizando a función *fact_imp*, en 4 rangos horarios en los cuales se ha dividido el día: 5pm-11pm, 11pm-5am, 5am-11am, 11am-5pm. Según la energía consumida en estos rangos se han clasificado los mismos en 4 definiciones: hora pico de consumo, si la energía consumida en el rango estudiado es la máxima en el día; hora fuera de pico, si la energía consumida en dicho rango es la mínima en el día; pico moderado, si el consumo de energía se encuentra entre el promedio y el máximo; y por ultimo hora estándar, si el consumo en ese rango se encuentra entre el mínimo y el promedio. Estas definiciones se fijan según el consumo en cada rango horario de la curva agregada.

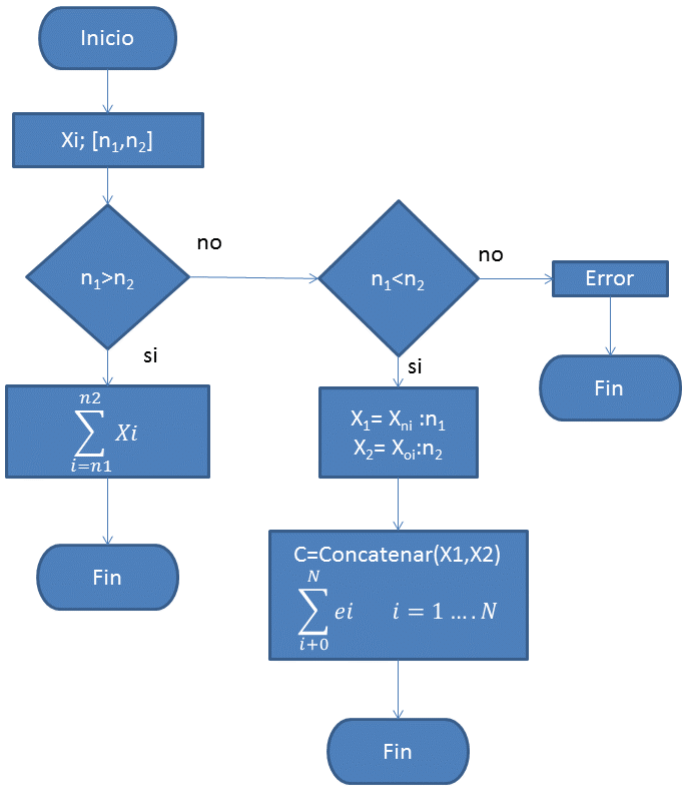


Figura 4.11: Flujograma de la función *fact_imp*

Todos los consumidores de la data en estudio pasan por el mismo proceso de cálculo de energía a través de la función *fact_imp* y esa información se guarda en un arreglo de tipo lista, el máximo consumo le es asignado la definición según el rango horario correspondiente en que ocurre y guardado en la variable tabla *Resultado*. La lista del consumo por rango horario de todos los usuarios luego es discriminada según el estrato al que pertenecen para así calcular de forma porcentual la contribución de cada uno de los estratos en el consumo total de la potencia agregada en cada uno de los rangos horarios antes definidos. Como se muestra en la figura 4.12.

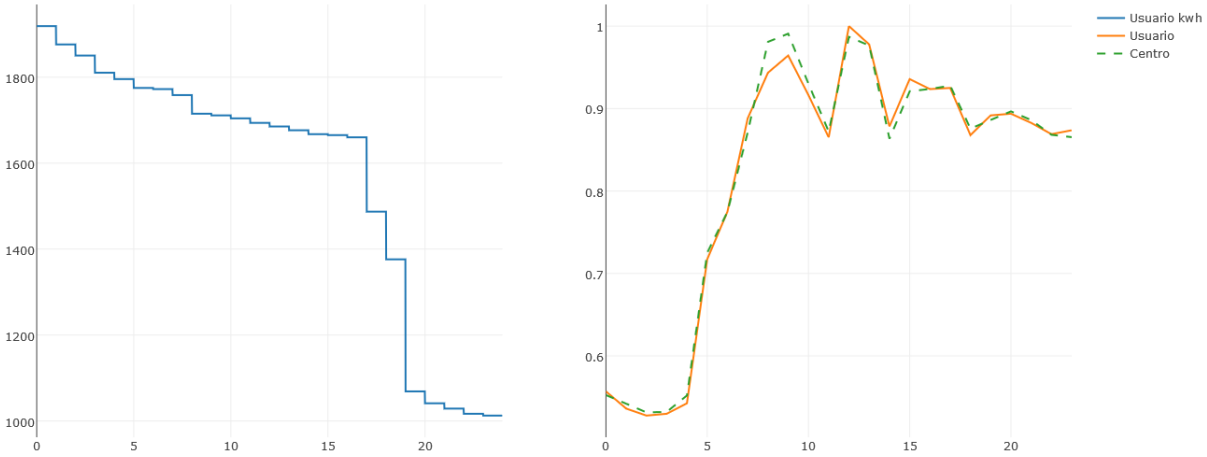
	h. Pico	h. fuera de Pico	Pico moderado	Pico moderado
Horas	de 17 a 22	de 23 a 4	de 5 a 10	de 11 a 16
res	25.1693	20.6818	15.4445	11.3044
com	15.9355	6.26751	17.3649	23.9229
ind	58.8951	73.0507	67.1906	64.7727

Figura 4.12: Contribución en % de cada estrato al consumo agregado total

4.4.5. Visualización de un consumidor específico

Después de la entrega del resultado de la contribución de cada consumidor por rango horario técnicamente ha concluido el estudio. Sin embargo, existe la opción de poner la lupa sobre alguno de los clientes estudiados. El código principal despliega un mensaje preguntando si se desea ver algún cliente en especifico, de ser negativa la respuesta el programa concluirá. De lo contrario, se corre la función *consumidor* la cual recibe como entrega el código del usuario, el set de datos, el contrato, el resultado y la potencia base. A cambio, muestra en pantalla el estrato asociado, contrato y zona horaria en la que posee su mayor consumo, así como factores que describen la demanda del usuario según su potencia activa, la potencia máxima y mínima, hora en la que ocurre cada una respectivamente, factor de carga y energía total consumida en el día por ese usuario, figura 4.10(b).

A la vez se despliega una gráfica compuesta por dos curvas, la primera corresponde a la frecuencia de demanda y representa, de mayor a menor, por cuanto tiempo el cliente a consumido cierto nivel de energía, la otra corresponde a la comparación de la curva de demanda normalizada con el centro del grupo al que pertenece, es decir, la curva caracterizada del estrato correspondiente, figura 4.14(a).



(a) Izquierda: curva de frecuencia de demanda; derecha: curva y patron de consumo en p.u.

```
Desea visualizar un usuario en especifico? (y/n): y

Ingrese codigo de usuario: U-5001
Estrato                1
Zona        Pico moderado
Contrato             com
Name: U-5001, dtype: object


             FC   P promedio   P Max  h de max   P Min  h de min  Energia
U-5001  0.820052  1573.516667  1918.8        12  1012.4         2  37764.4
```

(b) Ubicación por estrato y factores de descripción

Figura 4.13: Resultado de la función *consumidor*

La función consumidor se encuentra en un bucle, esperando la entrada del código de algún usuario a visualizar, hasta que sea introducida la palabra «exit», de ser así concluye el bucle y llega a su final el código principal. En la figura 4.13 se observa el flujograma correspondiente a esta función.

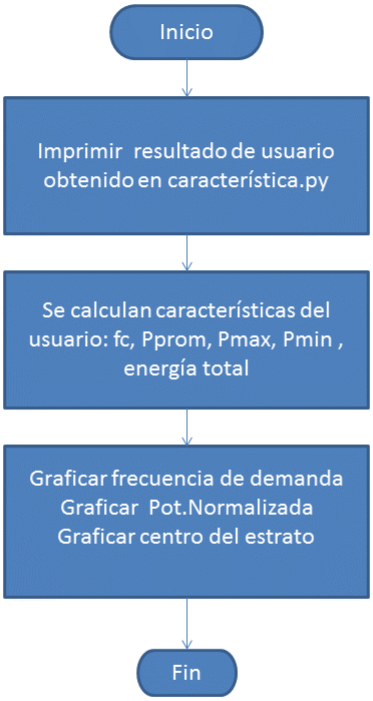


Figura 4.14: Flujograma de la función *consumidor*

4.4.6. Reducción

Reduccion() es una función diseñada para disminuir el número de estratos obtenidos en el proceso de clustering, debido a que en algunos casos la solución que mejor se ajusta diferenciando los estratos implica muchos grupos, lo cual es engorroso de manejar o es posible también que sea necesario utilizarla en casos en que surjan grupos que son muy poco significativos a nivel de población por lo que estos pequeños grupos se unifican a aquel estrato cuya tendencia más se asemeje.

Esta función posee dos modos de utilización los cuales se ajustan entre sus argumentos de entrada. La primera es usarla de forma manual (mode='man'), en este caso se introduce en conjunto el número de cada estrato a unir y se concatenan los clusters y centros de los estratos asociados mientras que se eliminan los estratos sobrantes para dar paso a un estrato nuevo, cuyo nombre corresponde al estrato de orden menor, que contiene los grupos y centros de todos los antes seleccionados. La segunda forma de usarla es de forma automática

(mode=’auto’), en este modo se introduce un estrato a reducir y la función realiza una comparación de su centro con el de cada uno de los demás estratos utilizando la distancia euclidea para encontrar el estrato cuyo patrón de consumo es mas similar y se procede a unificar estos dos estratos. Sin embargo, este último modo solo se puede utilizar con un estrato a la vez, por lo que para unificar mas estratos es necesario repetir la función.

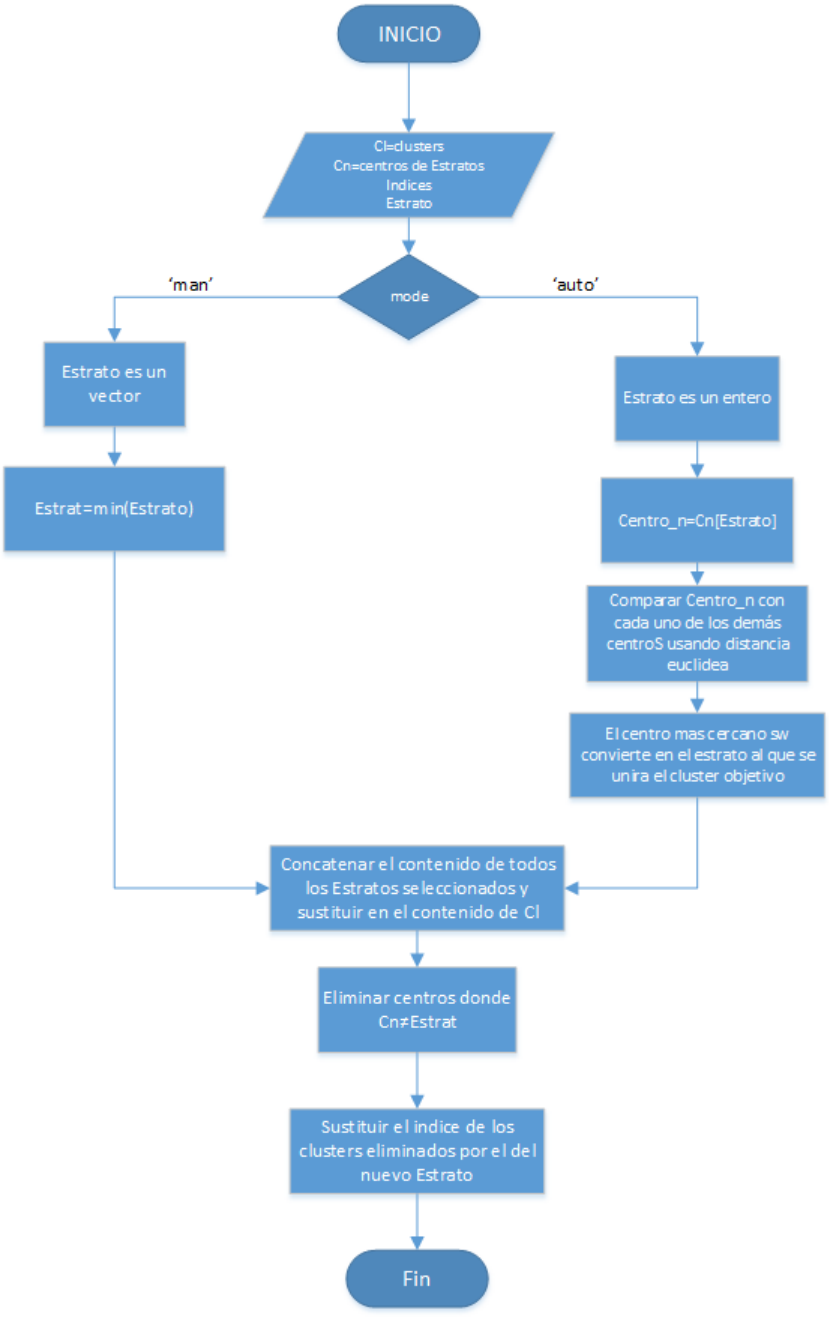


Figura 4.15: Flujograma de la función *Reduccion()*

Además del código y funciones antes mencionadas, el paquete desarrollado *pyCyDes* cuenta con otras funciones adicionales. Entre ellas se encuentran las funciones *MIA* y *CID* las cuales reciben la data, los centros y el indice de los grupos, y realizan el cálculo del índice de eficiencia respectivo. Con ellos es posible realizar la curva del codo para estimar el número de grupos a estratificar en la data.

Una última función a destacar es la función *insercion*(), la cual adquiere como entrada la curva de un nuevo usuario y, ya realizada la caracterización, la compara con los patrones de demanda ya definidos para así poder clasificarlo en el grupo más apropiado, sugiriendo su contrato entre los ya identificados y procediendo a concatenarlo en el resultado obtenido de la estratificación.

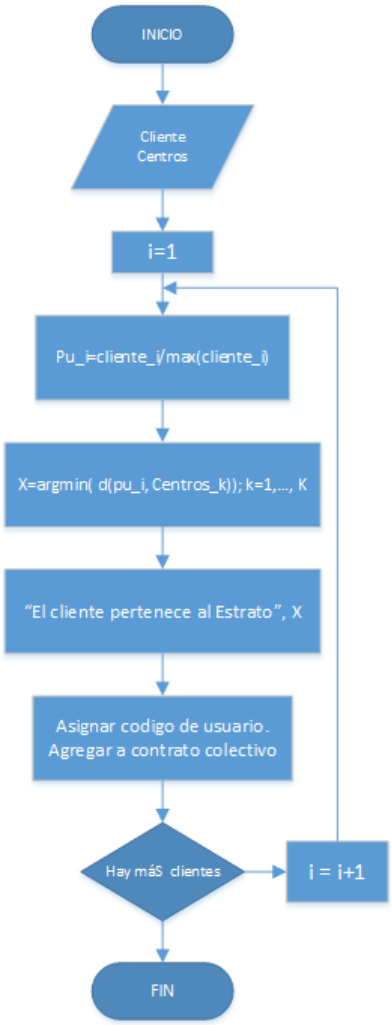


Figura 4.16: Flujograma de la función inserción

Un detalle importante de trabajar en el entorno de IPython es que a pesar de que ha

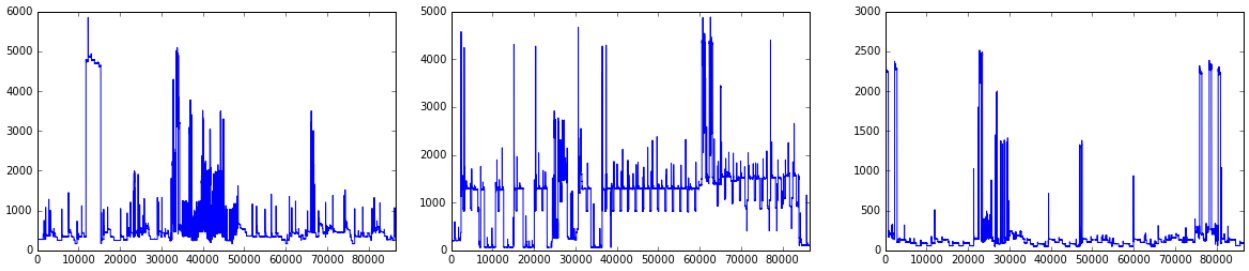
concluido el programa principal las funciones quedan cargadas, por lo que es posible utilizar de manera independiente las funciones contenidas en el paquete desarrollado.

CAPÍTULO 5

GENERACIÓN Y PREPARACION DE LA DATA

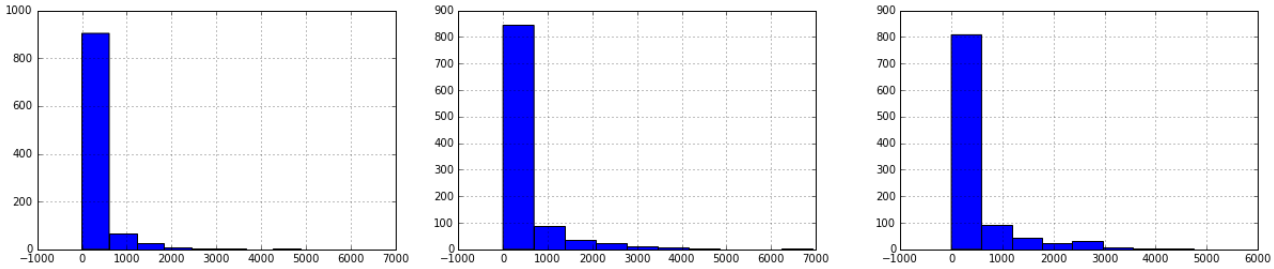
La información de entrada para probar la herramienta desarrollada debe al menos reflejar heterogenedad de los perfiles de carga de los consumidores quienes tienen dinámicas y patrones de vida muy distintos entre sí. La data que se ha preparado para realizar el caso de prueba está compuesta por mediciones que emulan los niveles residencial, no residencial e iluminación. En el caso de la data correspondeinte a un estrato residencial se cuenta con la combinación de las bases de datos de libre uso REDD [20] y ECO [21], medidas y publicadas para el estudio de análisis de datos y estudio de desagregación de carga. En el caso no residencial no se posee data real al alcance que pudiese emular los distintos estratos que la componen, por lo que se creó data artificial para este estudio a través de métodos de aprendizaje automático y modelos probabilísticos basados en cadenas de Markov, conocidos como Modelos Ocultos de Markov (HMM), y entrenados a partir de una curva base de la cual derivaron el resto de las curvas. Para llevar a cabo la generación de las curvas se recurrió a la librería de Python *hmmlearn,* especializado en Modelos Ocultos de Markov.

La figura 5.1 muestra el perfil de carga de un día de 3 consumidores al azar dentro de la data contenida en REDD y ECO, en un intervalo de medición de 1 s. La figura 5.2 muestra el histograma que representa el consumo de el set de datos en tres horas distintas del día (7 a.m., 12 p.m. y 6 p.m.) de un set compuesto por 1012 mediciones residenciales.



(a) (b) (c)

Figura 5.1: Ejemplo de curvas de carga residencial con frecuencia de adquisición de 1Hz. Eje y: Vatios (w). Eje x: tiempo de 1 día en intervalos de 1 s.



(a) 7 a.m. (b) 12 p.m. (c) 6 a. m.

Figura 5.2: Histograma de la data Residencial REDD y ECO a tres horas distintas del día. Eje y: Frecuencia. Eje x: Vatios (w)

5.1. HMM

Las cadenas de Markov o modelo de Markov, son un tipo de proceso estocástico en el cual la probabilidad de que ocurra un evento depende directamente de un evento ocurrido en el instante anterior; son señales observables representadas como un proceso aleatorio, en el cual los parámetros pueden determinarse con precisión. Dado un proceso que puede ser descrito por N estados S_1, S_2, ... S_N, cada cierto intervalo de tiempo incurre en un cambio de estado (incluyendo el quedarse en el estado actual) de acuerdo a la probabilidad asociada a cada estado. Se dice que en el tiempo t el sistema está en el estado q_t [22].

$$P[q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k .....] = P[q_t = S_j | q_{t-1} = S_i] \tag{5.1}$$

El Modelo Oculto de Markov tiene dos procesos estocásticos embebidos, de los cuales uno no es observable (estados ocultos), pero puede ser observado a través del otro proceso que produce una secuencia de observaciones [23]. En el caso a aplicar se puede tomar la potencia medida como el conjunto de observaciones, mientras que los estados ocultos corresponden a cada nivel de potencia que es posible de alcanzar por el elemento modelado en un instante de tiempo.

5.1.1. Elementos de un HMM

- Hay un número finito N de estados en el modelo; no se puede definir estrictamente lo que es un estado, sin embargo se puede decir que con la presencia de este la señal tiene algo que es medible y que la describe [22].
- Para cada muestra de tiempo t, hay un cambio hacia un nuevo estado basado en la distribución de probabilidad de transición, la cual depende del estado anterior (la propiedad de Markov).

- Después de cada transición se produce un símbolo observable a la salida, que es generado de acuerdo a la distribución de probabilidad de la que depende el estado actual. Esta se mantiene fija dependiendo de cuándo y cómo se introduce el estado. Hay N distribuciones de probabilidad de observación, las que representan aleatoriamente las variables o procesos estocásticos [23].

5.1.2. Definición formal del HMM para una observación discreta

- N: Número de estados de Markov en el modelo.
- M: Numero de símbolos observables, $V = \{v_1, ...v_M\}$.
- Distribución de probabilidad de transición entre estados,

$$A = \{a_{ij}\}, a_{ij} = P[q_t = S_j | q_{t-1} = S_i], i, j = 1....N \tag{5.2}$$

- Distribución de probabilidad de observación de símbolo en cada estado,

$$B = \{b_j(k)\}, b_j(k) = P[v_k | q_t = S_j], j = 1, ..., N; k = 1, ..., M \tag{5.3}$$

- Distribución de probabilidad del estado inicial,

$$\pi = \{\pi_i\}, \pi_i = P[q_i = S_i], i = 1, ...N \tag{5.4}$$

- Modelo: $\lambda = (A, B, \pi)$

Usando el modelo, una observación de secuencia $O = O_1, O_2, ...O_T$, se genera de la siguiente manera [22]:

1. Se elije un estado inicial $q_1 = S_i$ de acuerdo con la distribución de estado inicial π.
2. $t = 1$.
3. Se elige una observación $O_t = v_k$ de acuerdo a la distribución de probabilidad de observación del símbolo en el estado S_i. Ejemplo: $b_i(k)$.
4. Transición a un nuevo estado $q_{t+1} = S_j$ de acuerdo a la distribución de probabilidad de transición de estado S_i. Ejemplo: a_{ij}.
5. Reajuste $t = t+1$; si $t < T$ retornar al paso 3. De otro modo es el fin del procedimiento.

5.1.3. Problemas fundamentales

Dado el modelo HMM ya descrito, hay tres problemas básicos a solventar para hacer viable su uso en aplicaciones reales:

Problema 1: Evaluación. Dado el modelo $\lambda = (A, B, \pi)$ y la secuencia de observaciones $O = O_1, O_2, ...O_T$, calcular $P(O|\lambda)$, la probabilidad de que dados los parámetros del modelo se obtenga la secuencia de observación. Una solución a este problema es aplicar el algoritmo Backward-Forward [22, 23].

Problema 2: Dada la secuencia de observaciones O, cómo obtener la secuencia de estados $Q = q_1, q_2, ..., q_T$ que explica de forma óptima las observaciones. Dicho de otra forma, se intenta descubrir la parte escondida del modelo. Es un típico problema de estimación el cual se puede resolver computando el algoritmo de Viterbi. [22, 23]

Problema 3: Entrenamiento. Ajustar el modelo $\lambda = (A, B, \pi)$ para maximizar la probabilidad $P(O|\lambda)$. Se intenta optimizar los parámetros del modelo tal que se encuentre la mejor descripción del fenómeno. Se solucióna a través del algoritmo Baum-Welch [22, 23], o conocido también como algoritmo de Esperanza-Maximización (EM) [24].

5.2. Aplicando HMM y la librería *hmmlearn* para generación de data artificial

La librería *hmmlearn* posee todas las funciones necesarias para implementar los modelos de HMM, y por definición utiliza los algoritmos de EM y Viterbi. En el caso de este proyecto de utilizó un modelo Gaussiano definiendo el objeto *hmm.GaussianHMM.*

Para implementar el modelo, el primer paso es definir la variable como un objeto del tipo de modelo oculto de Markov, definiendo además el número de estados a estimar. En este caso se aplicó la misma metodología empleada anteriormente entre los algoritmos de clustering para encontrar el número óptimo de grupos a clasificar; se aplicó el algoritmo de K-medias y se le calculó el índice MIA, de forma iterativa a la data utilizada como base para entrenar el modelo, agrupando así los niveles de potencia que se repiten con mayor frecuencia. De esa forma a través de una curva de amalgamiento se estimó el número adecuado de estados; esta vez no se consideró un valor más a la derecha del codo de la curva, por lo que se tomarán en cuenta la media de niveles de potencia que no pertenecen directamente a los dispositivos o maquinaria detrás del consumo, sino al ruido inherente al sistema, de tal forma de formular una data lo más parecido posible a una muestra real. Para el caso de la secuencia utilizada se escogieron 90 estados.

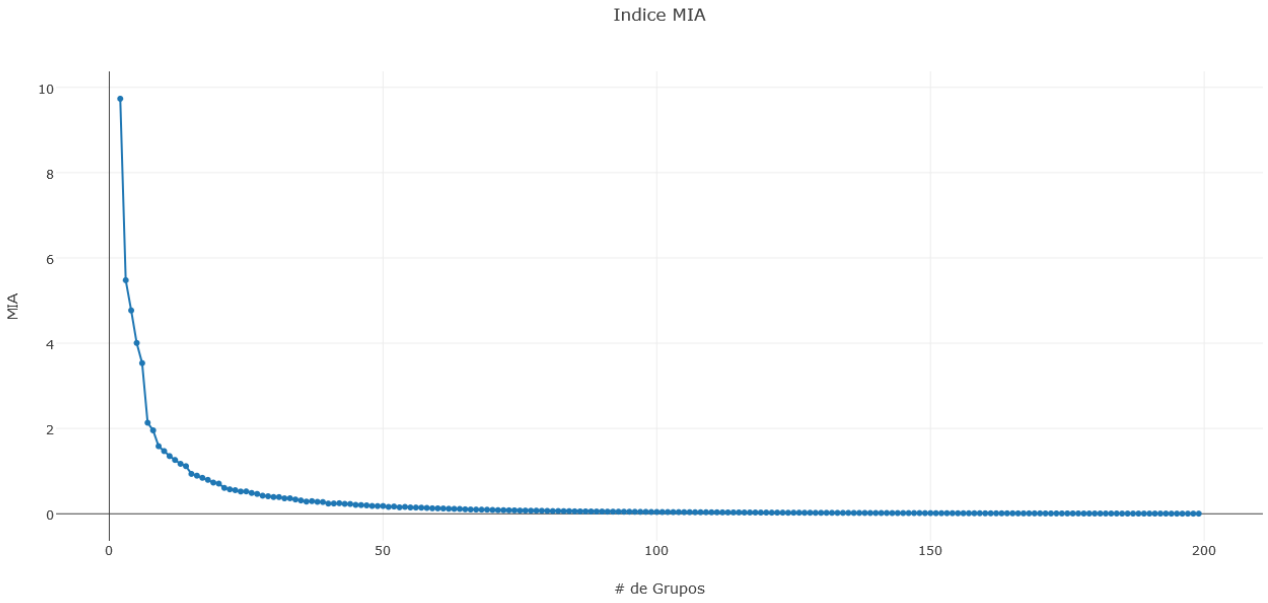


Figura 5.3: Curva de amalgamiento correspondiente a la secuencia de observación

Seguidamente se procede a entrenar el modelo usando la muestra como base, con el atributo *.fit(),* para estimar los parámetros λ de la muestra observada, la función aplica por definición el algoritmo EM (otros algoritmos compatibles, y métodos para uso de la función es posible encontrarlos en la documentación de la librería *hmmlearn*).

Esta función actúa en su estructura de la misma forma que el algoritmo K-means, descrito en el Capitulo 3. La entrada en este caso es el número de estados, lo cual es el homólogo del número de grupos, sin embargo, los grupos no son decididos por la menor distancia entre ellos, sino la mayor probabilidad de pertenecer a alguno.

Cada elemento de la observación es iniciado con una probabilidad π de pertenecer a los grupos y una desviación típica ajustada a 1. Puede ser el mismo porcentaje de probabilidad para cada uno inicialmente; este porcentaje irá cambiando a medida que avanza el algoritmo. Por otro lado, para cada distribución de probabilidad es necesario calcular sus parámetros, y en este caso, se debe calcular la media y la desviación típica para cada una en función de los elementos observados. Dado que la probabilidad irá cambiando a medida que se ajustan estos valores el elemento pertenecerá al cluster cuya probabilidad sea mayor $\pi_i P_i(x|\mu,\sigma)$ [24].

Aplicando la densidad de distribución normal, se tiene que la suma de todas las probabilidades de pertenencia de una observación a los clusters es:

$$P(X|\pi,\overrightarrow{\mu},\overrightarrow{\sigma}) = \sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi}} \frac{\exp(-\frac{1}{2}(\frac{x-\mu_k}{\sigma_k})^2}{\sigma_k} \tag{5.5}$$

$$argmax \sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi}} \frac{\exp(-\frac{1}{2}(\frac{x-\mu_k}{\sigma_k})^2}{\sigma_k} \tag{5.6}$$

El algoritmo entonces alterna entre dos pasos, Esperanza y Maximización.

Esperanza: Dados los parámetros de la distribución, se refiere a la asignación de cada elemento al cluster de mayor probabilidad.

$$argmax \sum_{k=1}^{K} \pi_k P_k(x|\overrightarrow{\mu}, \overrightarrow{\sigma}) \tag{5.7}$$

Maximización: Con los elementos asignados actualmente a cada cluster, asignar el nuevo valor de los parámetros (π_k, μ_k, σ_k).

$$\pi_k = \frac{\#Elementos}{Total} \tag{5.8}$$

$$\mu_k = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{5.9}$$

$$\sigma_k = \sqrt{\frac{1}{n-1} \sum_{n=1}^{n} (x_i - \mu_k)^2} \tag{5.10}$$

Luego de obtener los parámetros finales de cada estado (cluster), la función de HMM pasa a obtener la mejor descripción de la observación entregada utilizando el algoritmo de Viterbi, y con ello el cálculo de la matriz de transición de estados.

La figura 5.4(a) a muestra el modelo definido en python usando la librería hmmlearn, y la figura 5.4(b) la matriz de transición de estado resultante de este modelo.

```
Out[113]:
GaussianHMM(algorithm='em', covariance_type='diag', covars_prior=0.01,
      covars_weight=1, init_params='stmc', means_prior=0, means_weight=0,
      min_covar=0.001, n_components=90, n_iter=10, params='stmc',
      random_state=None, startprob_prior=1.0, tol=0.01, transmat_prior=1.0,
      verbose=False)
```

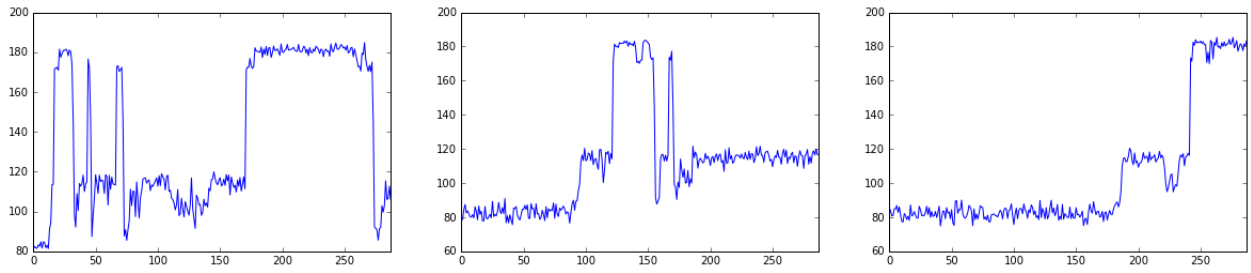
(a) Definicion del HMM con *hmm.GaussianHMM*

```
In [114]: modelos['ind'].transmat_
Out[114]:
array([[  2.58983009e-002,   1.03497385e-159,   1.08092340e-019, ...,
          1.35362772e-002,   5.48372154e-160,   3.85706422e-002],
       [  2.40689902e-137,   5.01544415e-002,   1.27838362e-069, ...,
          4.23702184e-128,   5.49828207e-002,   6.64695992e-140],
       [  3.06362479e-022,   3.24742999e-010,   6.33078343e-002, ...,
          2.28312719e-021,   1.73509833e-010,   1.80022439e-022],
       ...,
       [  2.52397116e-002,   5.45454258e-137,   5.27021602e-018, ...,
          1.50244259e-002,   2.89359501e-137,   3.54462027e-002],
       [  1.53765679e-140,   5.02895651e-002,   1.19739874e-070, ...,
          3.61437021e-131,   5.53686796e-002,   4.18931426e-143],
       [  2.59829370e-002,   6.11327246e-175,   1.94872142e-020, ...,
          1.27758904e-002,   3.23556526e-175,   3.99459102e-002]])
```

(b) Parte de la matriz de la distribución de probabilidad de transición de estado.

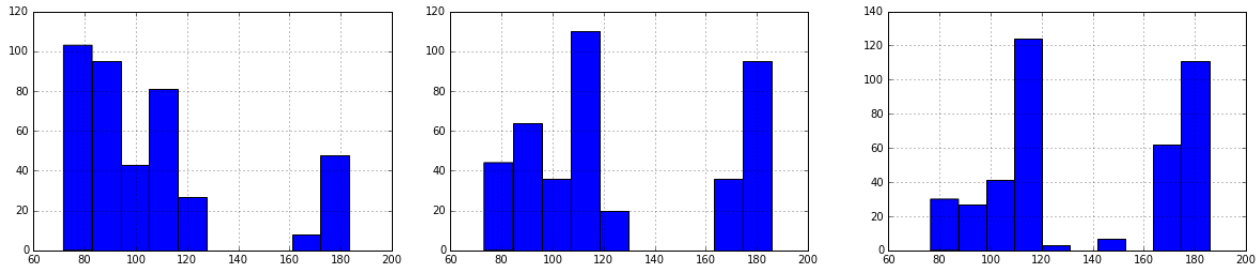
Figura 5.4: Parámetros de modelo

Con esto ya se encuentra definido y entrenado el modelo, por lo que para obtener las curvas solo es necesario hacer un muestreo aleatorio a través del atributo *.sample( ).* A este se le entrega como argumento de entrada el número de puntos a muestrear, que en nuestro caso correspondería a los intervalos de tiempo en que se simula la medición, y como resultado se obtiene una secuencia de observaciones acordes a la probabilidad de estados, es decir, a la probabilidad de que en cada intervalo de tiempo cambiase el nivel de potencia a lo largo de la muestra pedida. La figura 5.5 muestra 3 ejemplos de curvas generadas artificialmente con este método.



(a) (b) (c)

Figura 5.5: Ejemplos aleatorios de curvas producidas por el modelo entrenado de HMM. Eje y: Vatios (kw). Eje x: tiempo de 1 día en intervalos de 5 min.



(a) 7 a.m. (b) 12 p.m. (c) 6 a.m.

Figura 5.6: Histograma de la data aleatoria generada a tres horas distintas del día. Eje y: Frecuencia. Eje x: Vatios (w)

CAPÍTULO 6

Caso de Estudio

Para mostrar la eficiencia del código en general se hizo uso de dos set de datos compuesto por las mediciones a clientes reales, artificiales y promedio, mencionadas en capítulos anteriores. Para un primer set, se seleccionó las curvas PSEG compuestas por 2555 curvas promedio diarias de los usuarios del sistema eléctrico americano de Jersey en cada sector; a pesar de ser cada una de estas la curva promedio de un día, representando además distintas épocas del año (verano e invierno), se seleccionó el conjunto completo como una población de estudio para simular una mayor densidad de población, y agregar a la muestra patrones diferentes en los cuales estratificar. El segundo set de datos está compuesto por 1012 curvas residenciales reales de San Diego, USA; 366 curvas de iluminación reales y 405 curvas de demanda creadas artificialmente usando Modelos Ocultos de Markov, con el modelo definido en el Capítulo 5. Las tablas 5.1 (a) y 5.1(b) contienen una descripcion de la data a utilizar en la simulación.

Tabla 6.1: Descripción de data a estratificar.

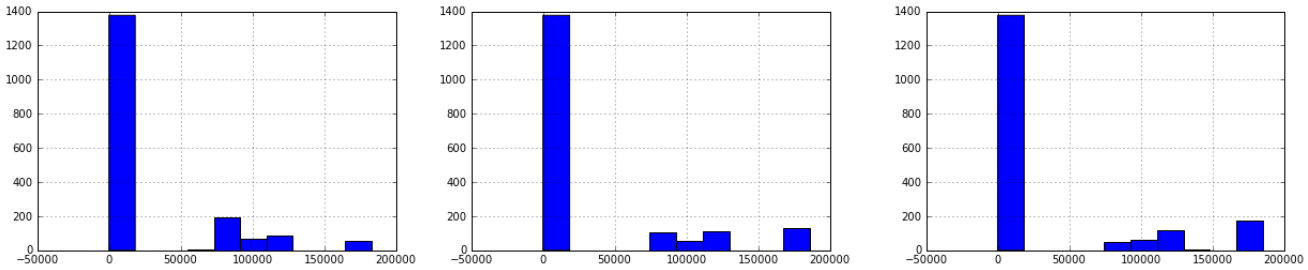
(a) Conjunto de Datos PSEG.

Data		Cantidad	Frec. Muestreo	Origen	Criterio
RS	Residencial	365	1/h	Curva Promedio	Verano
RHS	Residencial	365	1/h	Curva Promedio	Invierno
LPLS	Gran potencia iluminación	365	1/h	Curva Promedio	Verano
RLM	Industrial	365	1/h	Curva Promedio	Verano
GLP	Comercial	365	1/h	Curva Promedio	Verano
HS	Industrial	365	1/h	Curva Promedio	Invierno
WH	Industrial	365	1/h	Curva Promedio	Verano

(b) Conjunto de data real y artificial.

Data	Cantidad	Frec. Muestreo	Origen
Residencial	1012	1/seg	Real
No Residencial	405	1/5min	Sintética
Iluminación	366	1/5min	Promedio

Para ambos sets completos mencionados anteriormente, se simuló un contrato colectivo el cual contenía, nombres, apellidos, una ubicación geográfica aleatoria del cliente y un estrato comercial. De esta forma se pretende simular una muestra de clientes ya establecidos en el sistema los cuales poseen una factura de tipo residencial, no residencial o de iluminación respectivamente. En la figura 6.2 se muestra el histograma en 3 horas representativas del dia para el conjunto de data real y artificial completo.



(a) 7 a.m. (b) 12 p.m. (c) 6 a.m.

Figura 6.1: Histograma de la data aleatoria generada a tres horas distintas del día. Eje y: Frecuencia. Eje x: Vatios (w)

Para pruebas del algoritmo se estudiaran 2 casos en concreto:

1. Corrida del programa utilizando data PSEG, el set de datos descrito en la tabla 6.1(a) con un tiempo de muestreo de 1 registro por hora.

2. Corrida del programa utilizando el set de datos descrito en la tabla 6.1(b) con un tiempo de muestreo de 1 registro cada 15 min.

CAPÍTULO 7

RESULTADO

7.1. Caso 1: Conjunto de datos PSEG (1reg/h)

A continuación se presentan los resultados obtenidos al hacer la corrida el programa para el conjunto de datos PSEG de 2555 usuarios.

7.1.1. Pre-procesamiento

Todas las muestras en esta data se encuentran sincronizadas a un muestreo de 1 registro por hora, y no se detectó la presencia de ningún valor perdido por lo que el pre-procesamiento no era necesario. El estudio de la data completa es posible desde '2015-01-01 00:00:00' y el recorrido de reconocimiento se realizó en un tiempo $t = 8\ s$ incluyendo el tiempo que tomó introducir la fecha a evaluar.

7.1.2. Procesamiento y Visualización

Una vez realizado el pre-procesamiento inicia el código principal para el estudio de la población completa. El proceso de normalización de la data tomó $t = 16{,}17\ s$.

Con la data normalizada se procede a estimar el número de grupos a estratificar, tomando como indicio principal la curva de amalgamiento mostrada en la figura 7.1.

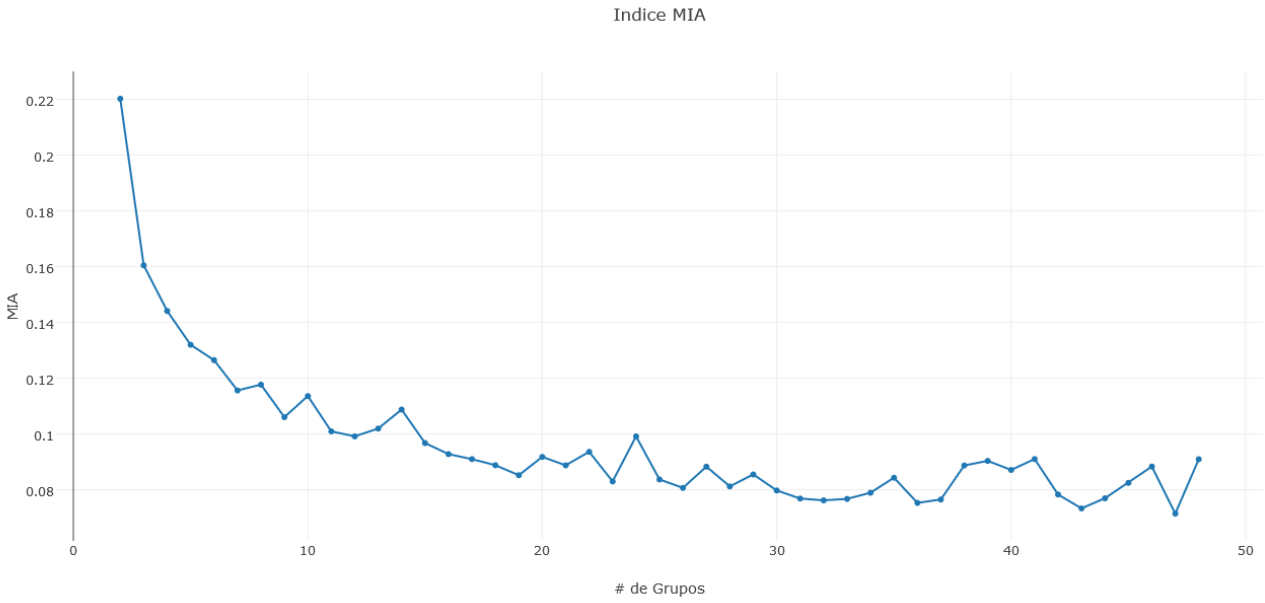


Figura 7.1: Curva de amalgamiento

Basado en esta curva, se optó por una estratificación en 8 grupos para inicializar el algoritmo *K-medianas*. Como consecuencia, el algoritmo convergió en $t = 0{,}058\ s$. En la figura 7.2 se muestra los grupos conformados por los consumidores de perfiles de carga similares según el criterio utilizado.

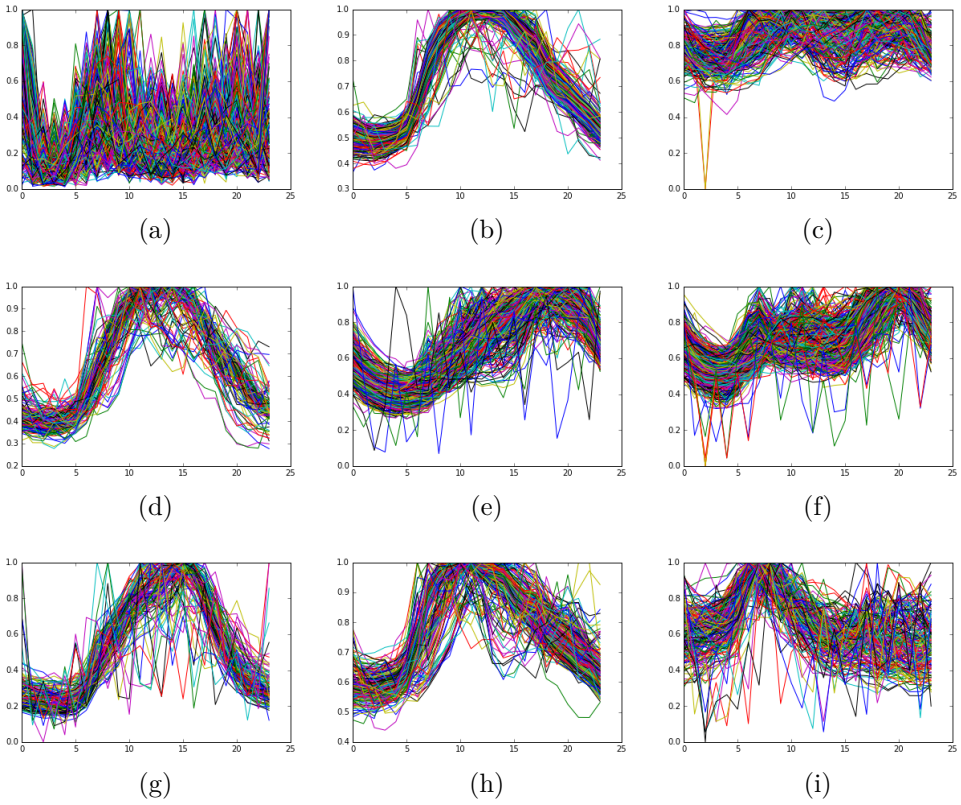


Figura 7.2: Perfiles de demanda agrupados por cluster. Eje x: tiempo - 24 horas en intervalos de 1 hora. Eje y: Potencia máxima normalizada

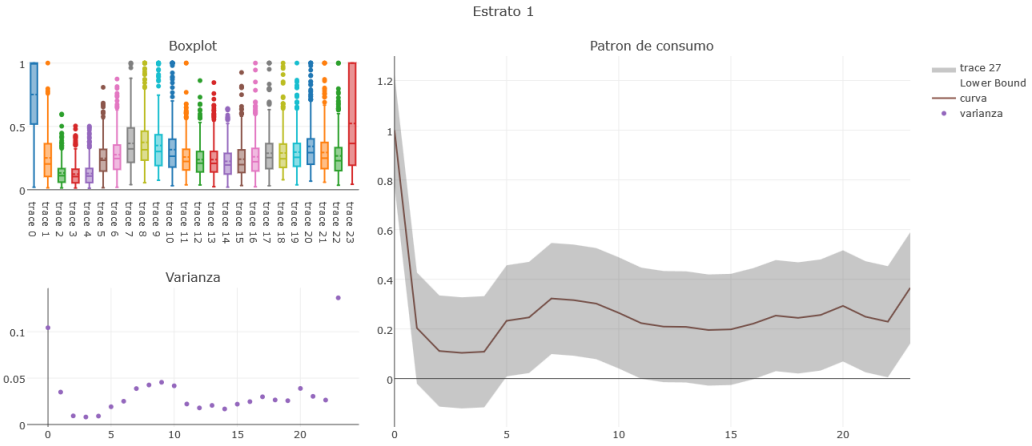


Figura 7.3: Centro de Estrato 1

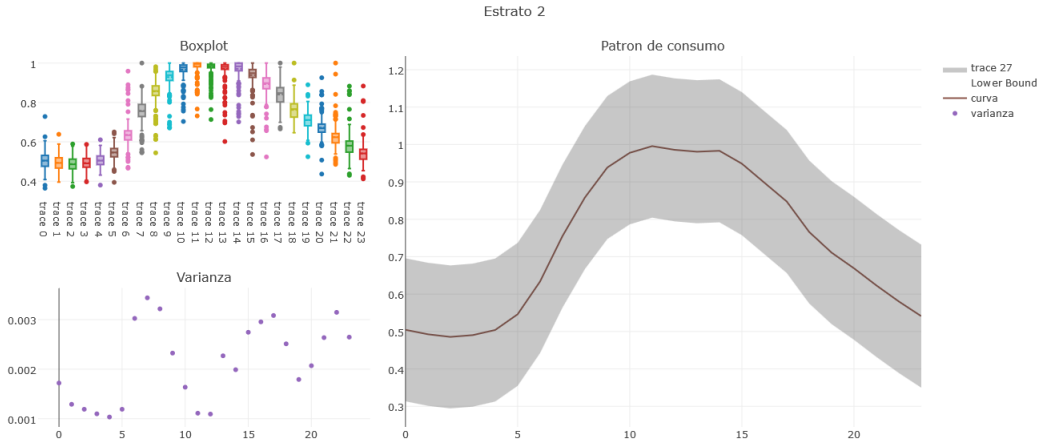


Figura 7.4: Centro de Estrato 2

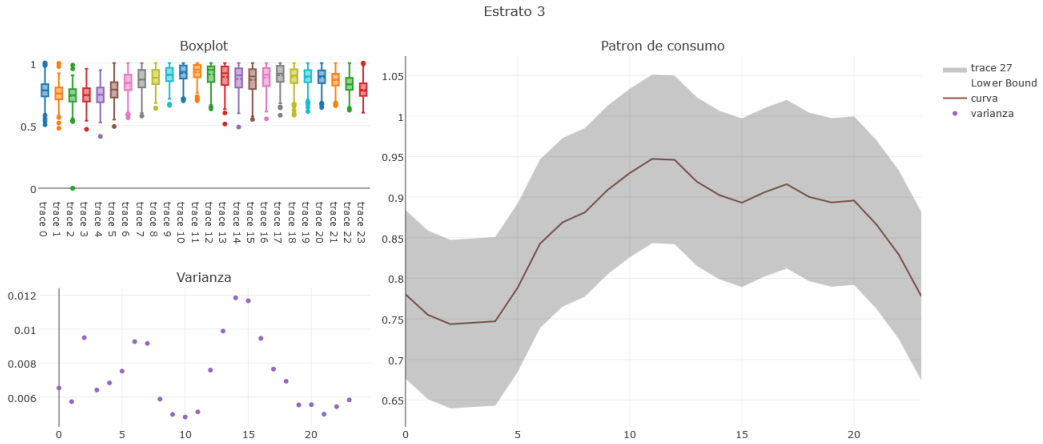


Figura 7.5: Centro de Estrato 3

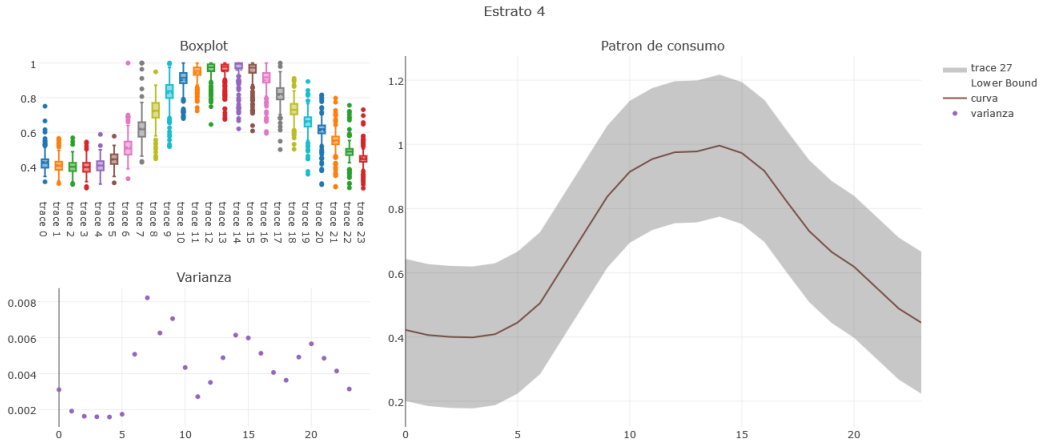


Figura 7.6: Centro de Estrato 4

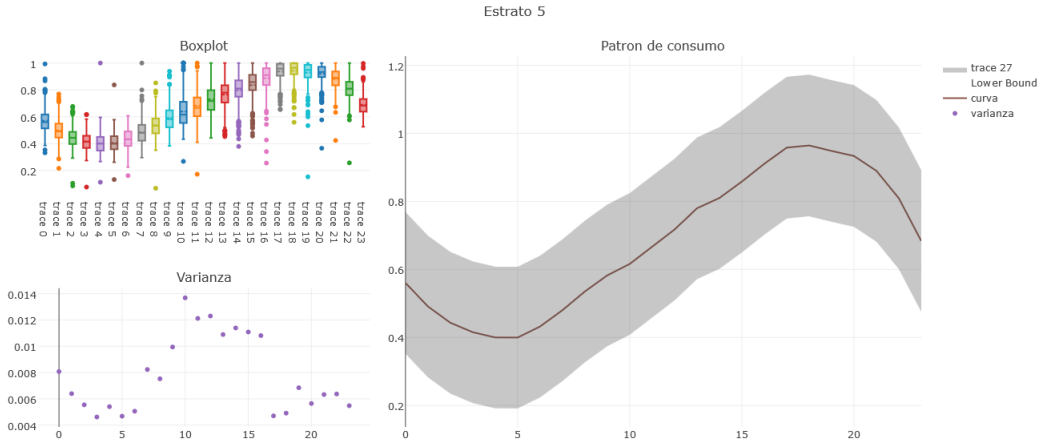


Figura 7.7: Centro de Estrato 5

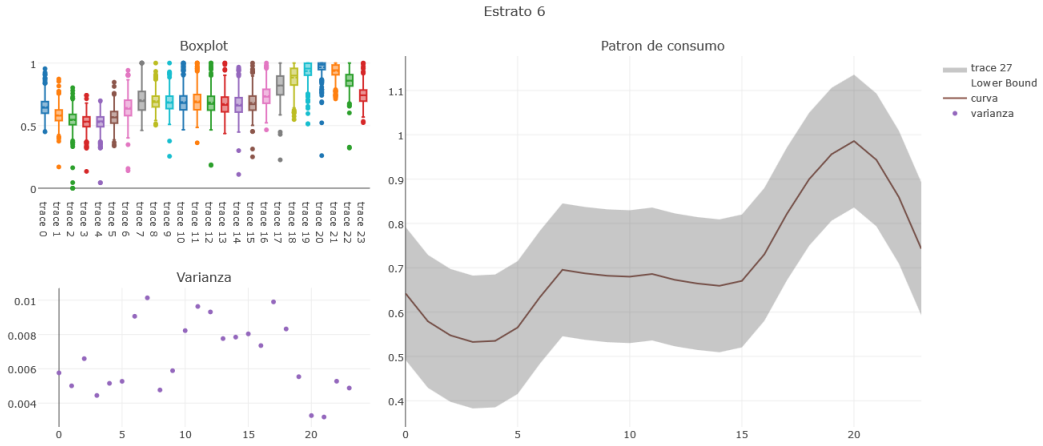


Figura 7.8: Centro de Estrato 6

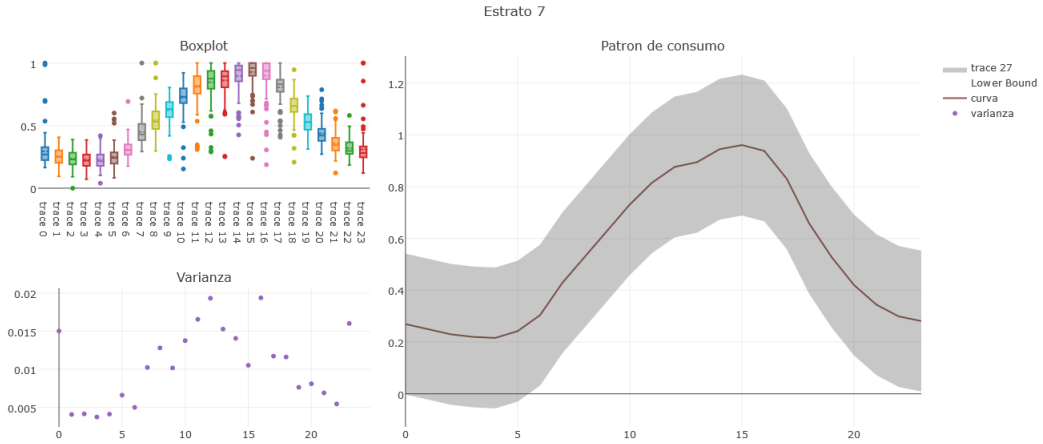


Figura 7.9: Centro de Estrato 7

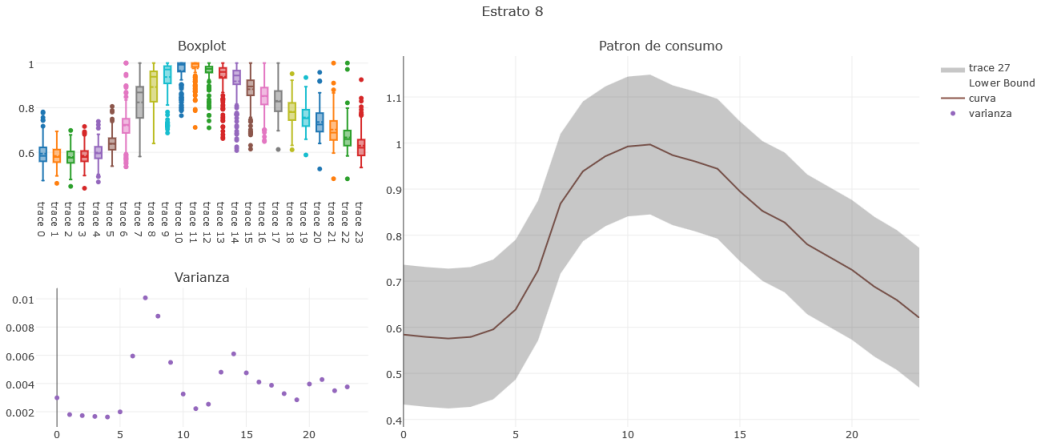


Figura 7.10: Centro de Estrato 8

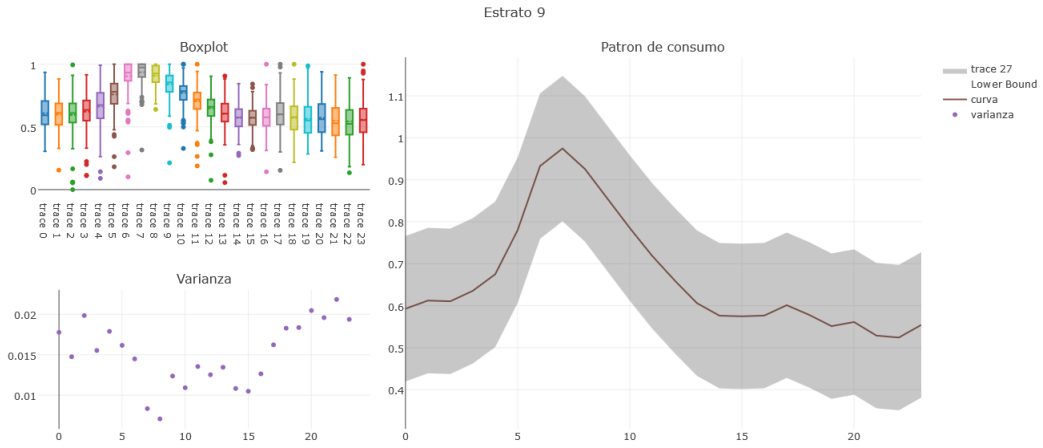
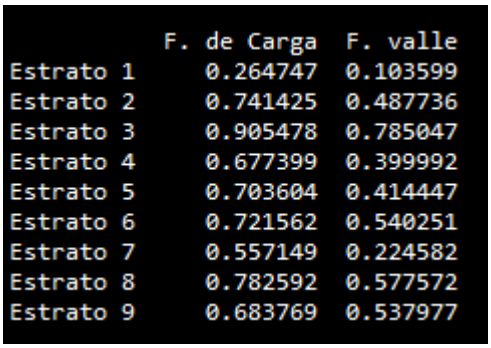


Figura 7.11: Centro de Estrato 9

Las figuras 7.3-11 presentan el centro que se ha calculado para cada uno grupo de los grupos estratificados, junto con un diagrama de boxplot que muestra la media, la mediana, y los valores extremos del conjunto de datos en cada intervalo de medición más un gráfico que señala la la dispersión del conjunto de datos en cada intervalo de tiempo a través de su varianza .Dada la normalización hecha a los perfiles de carga antes de estratificar, según la cual los valores pertenecen a un rango entre 0 y 1, la dispersión máxima posible respecto a la media es una varianza de 0.25. Según lo apreciable en las gráficas de varianza, los valores máximos en ciertos puntos de las mismas rondan entre 0.01 y 0.02 que representa

una dispersión de la data dentro del estrato entre 4 y 8 %, por lo que es visible una buena homogeneidad entre los usuarios agrupados.



	F. de Carga	F. valle
Estrato 1	0.264747	0.103599
Estrato 2	0.741425	0.487736
Estrato 3	0.905478	0.785047
Estrato 4	0.677399	0.399992
Estrato 5	0.703604	0.414447
Estrato 6	0.721562	0.540251
Estrato 7	0.557149	0.224582
Estrato 8	0.782592	0.577572
Estrato 9	0.683769	0.537977

Figura 7.12: Factores que describen la demanda en los estratos

7.1.3. Identificación

Se posee los patrones de consumo de cada estrato y los factores que lo describen. Utilizando esa información en conjunto con el contrato ya anteriormente establecido de los usuarios se identifica en cada estrato los tipos de consumidores conocidos correspondientes a los patrones encontrados, Tabla 7.1 y figura 7.13.

Tabla 7.1: Identificación de los estratos

Estratos	RS	RHS	LPLS	RLM	GLP	HS	WH	Total
Estrato 1	0	0	0	0	0	0	344	344
Estrato 2	0	5	143	0	128	17	0	293
Estrato 3	13	82	115	1	52	22	0	285
Estrato 4	0	7	128	0	0	46	0	181
Estrato 5	92	117	0	146	0	1	3	359
Estrato 6	259	109	0	218	0	2	3	591
Estrato 7	0	0	0	0	0	115	9	124
Estrato 8	1	15	107	0	57	15	0	195
Estrato 9	0	30	0	0	0	147	6	183

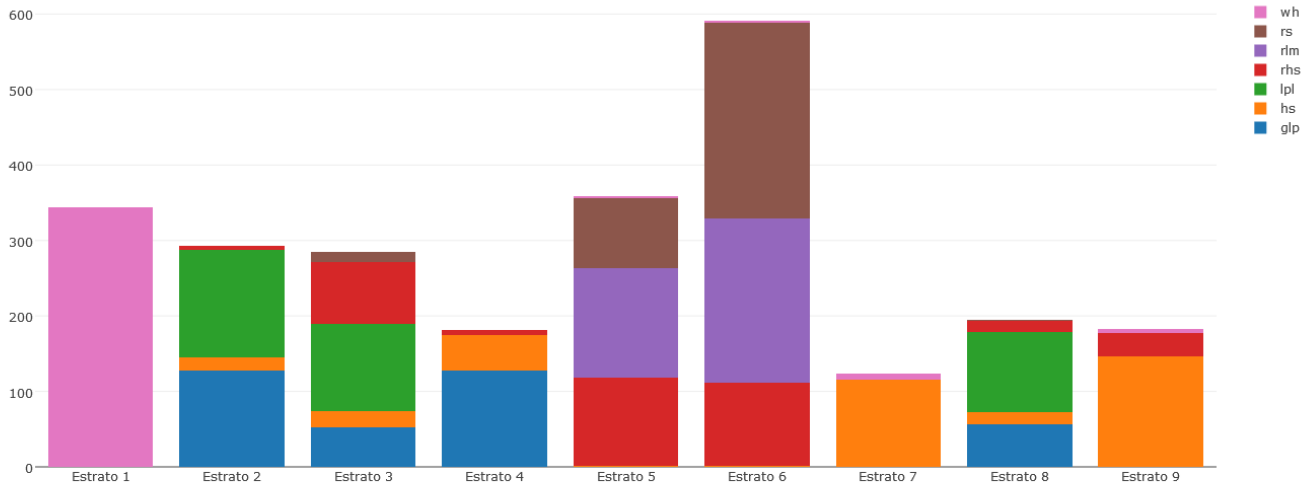


Figura 7.13: Distribución de antiguos consumidores en nuevos estratos. Eje x: No de Usuarios. Eje y: Estratos

Debido a que cada elemento en sí de la data representa el consumo promedio de un conjunto de clientes del sistema eléctrico al que pertenecen, algunas de las curvas aquí representadas presentan tendencias parecidas.

7.1.4. Potencia Agregada

A continuación se presenta el resultado de la agregación total de la carga y la estimación de la misma usando los modelos anteriormente definidos (figura 7.7) y la contribución por rango horario de cada estrato al consumo total (tabla 7.3).

Horas	Standard de 17 a 23	h. fuera de Pico de 23 a 5	Pico moderado de 5 a 11	h. Pico de 11 a 17
Estrato 3	24.9505	28.2394	23.1871	21.2469
Estrato 2	39.6709	36.426	40.0942	43.4082
Estrato 8	28.2568	29.0095	31.1042	29.6517
Estrato 4	1.51189	1.19338	1.34851	1.7439
Estrato 9	0.833056	1.16257	1.23693	0.732173
Estrato 6	2.21952	1.92947	1.5463	1.33028
Estrato 7	0.332103	0.191373	0.247381	0.424674
Estrato 5	1.98765	1.4472	0.98534	1.30814
Estrato 1	0.23755	0.401107	0.250089	0.154082

Figura 7.15: (%) consumo de energía por bloques horarios.

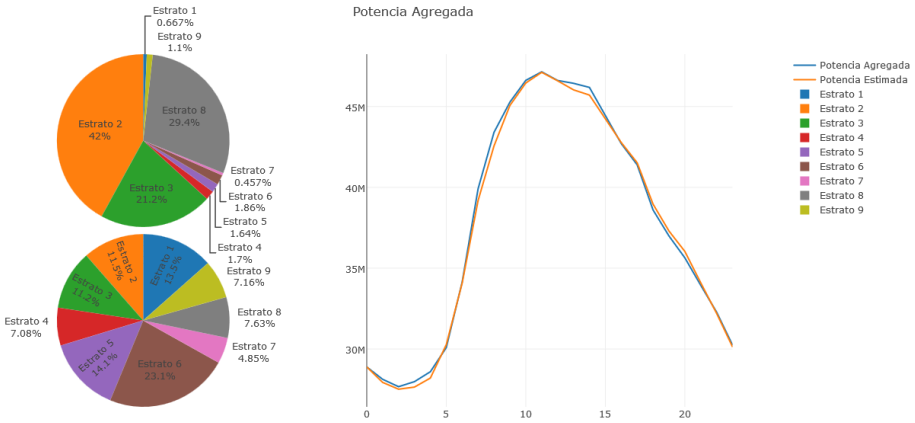


Figura 7.14: Potencia Agregada. Eje x: tiempo - 24 horas en intervalos de 1 hora. Eje y: Vatios (w)

Se observa entre la potencia agregada y la potencia estimada un error de $E = 0{,}85\,\%$.

Debido a la hora de mayor consumo de energía, el programa ha definido como rango de hora pico de 11am a 4pm, en la cual la mayor demanda de potencia se le atribuye a los estratos 2, 3 y 8, que están compuestos por usuarios comerciales y, de gran potencia e iluminación; que además representa a su vez el 71.8 % de la demanda total.

7.2. Caso 2: 1reg/15min

En este caso se realizó el estudio para el conjunto de datos real y artificial, pero con un muestreo de cada 15 min en la información. Esta frecuencia de muestreo es muy común de usar en los medidores inteligentes.

Se realizaron 4 corridas de este caso y se presenta a continuación el resultado para el mejor caso seleccionado.

7.2.1. Pre-procesamiento

Para este conjunto de datos existen varias muestras con frecuencias de muestreo distinto. Los usuarios residenciales poseen una frecuencia de muestreo de 1 registro por seg, mientras que los usuarios no residenciales del tipo industrial e iluminación poseen una frecuencia de muestreo de 1 registro cada 5 min; por lo que se realiza la sincronización para llevarlas todas a una misma frecuencia de muestreo de 1 registro cada 15 min, tomando como registro el valor máximo dentro del rango de tiempo entre una muestra y otra.

El tiempo de pre-procesamiento para este caso fue de $t = 302{,}12\ s.$

7.2.2. Procesamiento y Visualización

La adquisición y normalización de la data se logró en $t = 9{,}80\ s$. Basado en la curva de amalgamiento mostrada en la figura 7.16, se estimó 8 clusters inicialmente.

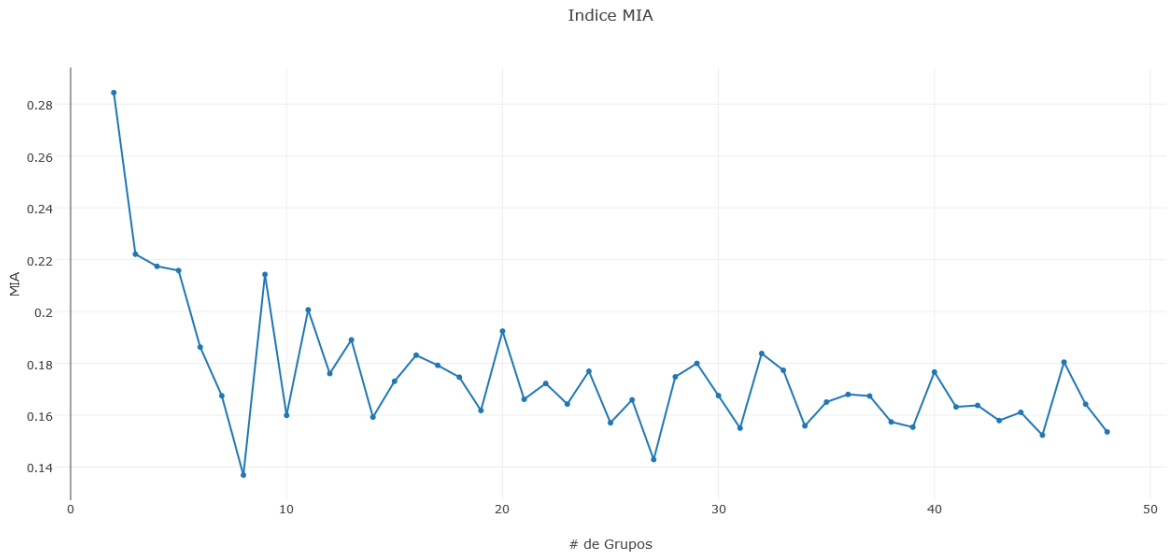


Figura 7.16: Curva de amalgamiento

Convergencia del algoritmo de K-medianas en $t = 0{,}07\ s$. En la figura 7.9 se muestran los grupos asociados por el algoritmo.

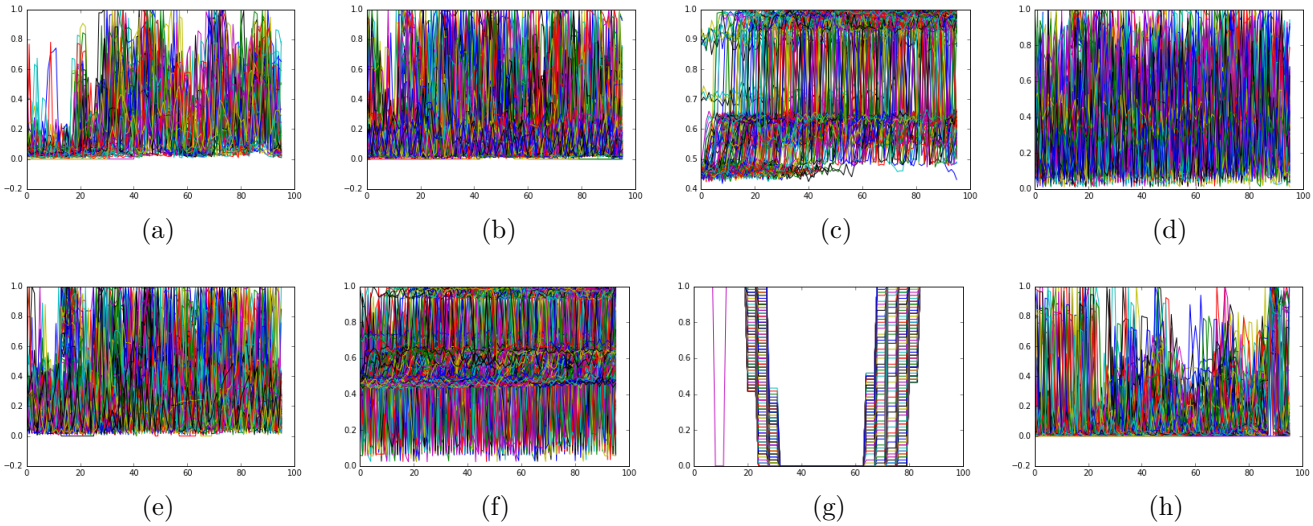


Figura 7.17: Perfiles de demanda agrupados por cluster. Eje x: tiempo - 24 horas en intervalos de 15 min. Eje y: Potencia máxima normalizada

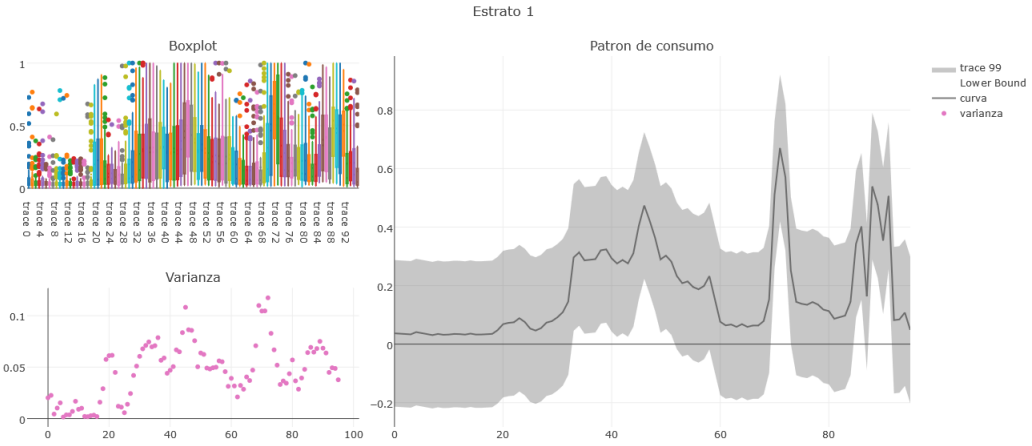


Figura 7.18: Centro de Estrato 1

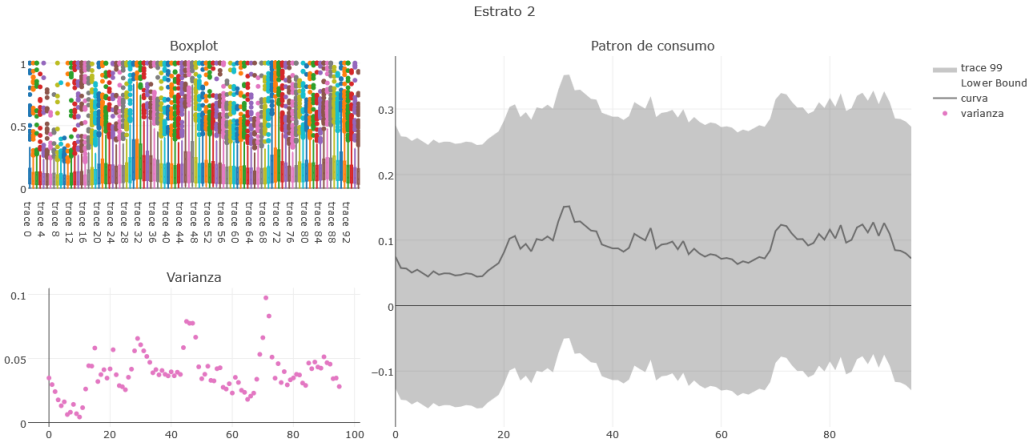


Figura 7.19: Centro de Estrato 2

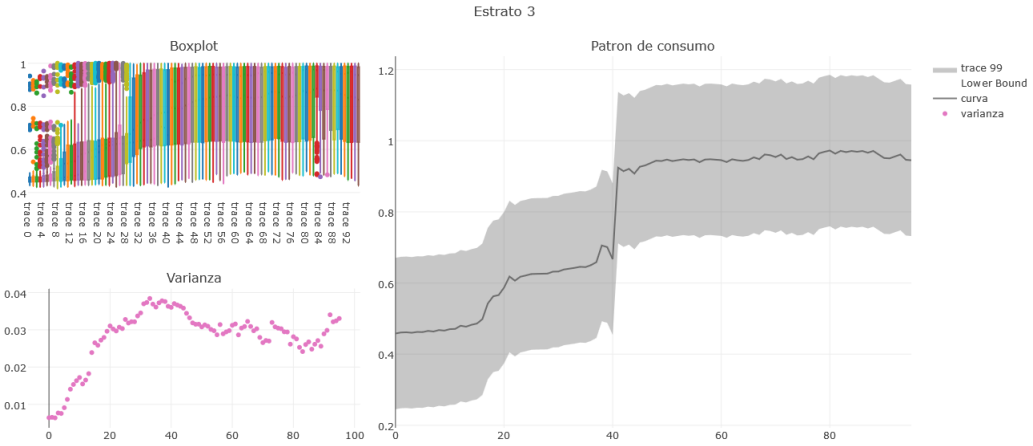


Figura 7.20: Centro de Estrato 3

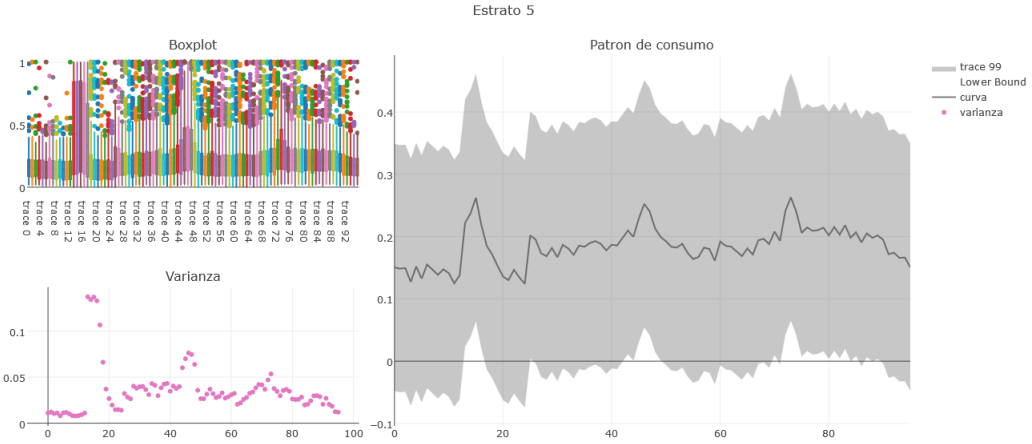


Figura 7.21: Centro de Estrato 4

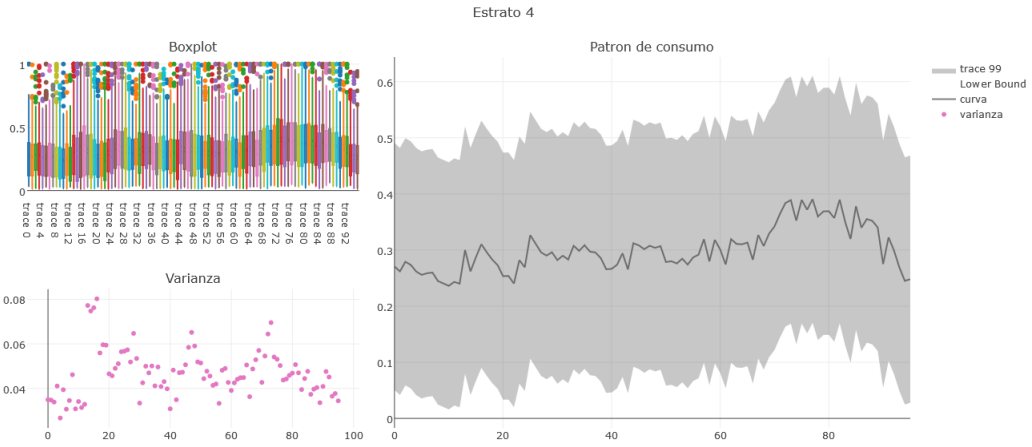


Figura 7.22: Centro de Estrato 5

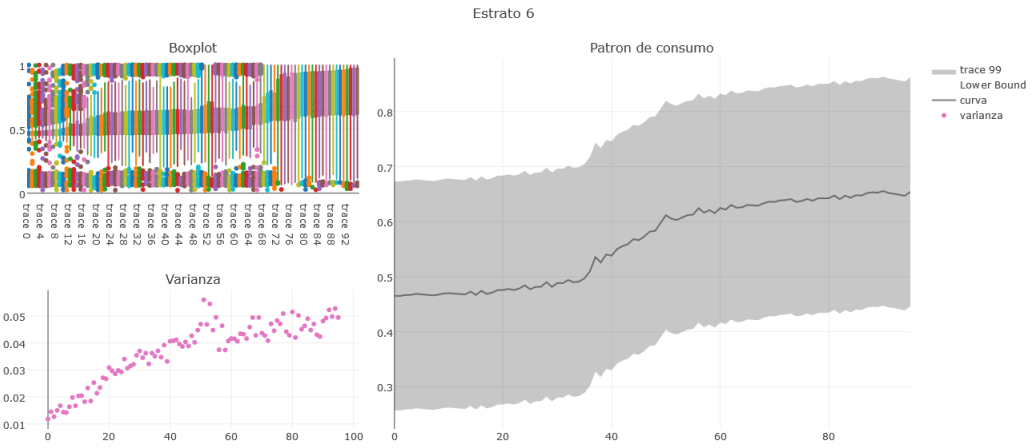


Figura 7.23: Centro de Estrato 6

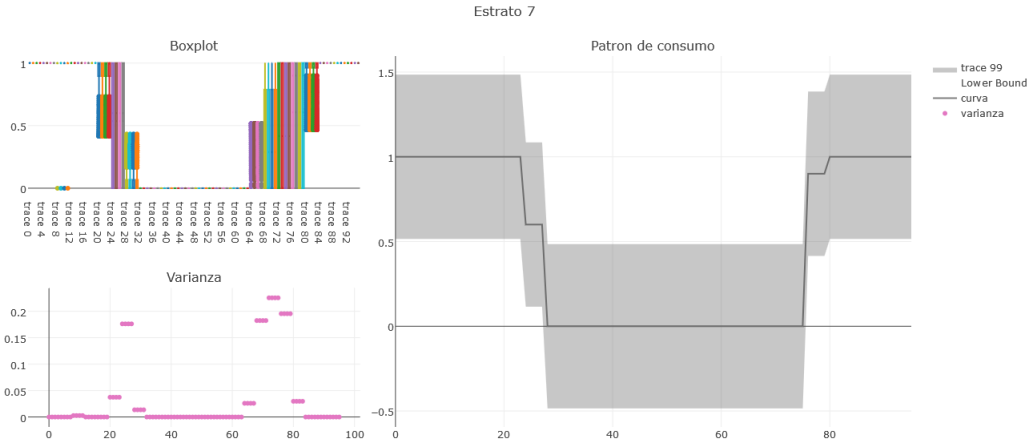


Figura 7.24: Centro de Estrato 7

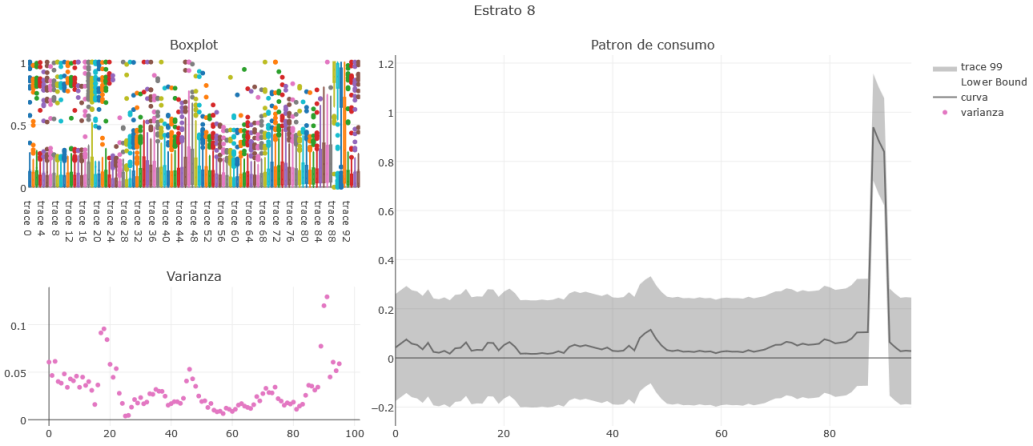
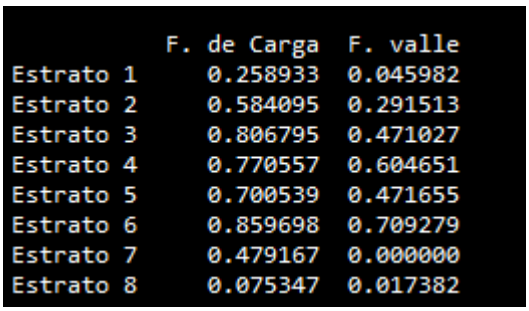


Figura 7.25: Centro de Estrato 8

En las figuras 18-25 se muestran los centros de grupo concebidos de la estratificación. Igualmente en los gráficos de boxplot y de varianza de los mismos se hace visible una dispersión baja de hasta 30 %, lo que indica grupos homogéneos, observando la dispersión mas alta entre los estratos en el Estrato 4 con un 40 %.Por otro lado, el boxplot muestra gran cantidad de valores atípicos, lo que puede atribuirse a la cantidad y diferencia de consumidores residenciales en este estrato. La figura 7.26 muestra el valor de los factores obtenidos de carga y de valle.



	F. de Carga	F. valle
Estrato 1	0.258933	0.045982
Estrato 2	0.584095	0.291513
Estrato 3	0.806795	0.471027
Estrato 4	0.770557	0.604651
Estrato 5	0.700539	0.471655
Estrato 6	0.859698	0.709279
Estrato 7	0.479167	0.000000
Estrato 8	0.075347	0.017382

Figura 7.26: Factores que describen la demanda en los estratos

7.2.3. Identificación

Teniendo en cuenta el contrato previo de los usuarios el algoritmo procede a la identificación por cada estrato. En la tabla 7.2 se muestra la correspondencia de cada sector (Residencial, Industrial e Iluminación) con el estrato al que ha sido asociado el consumidor.

Tabla 7.2: Identificación de los estratos

Estratos	Residencial	Indsutrial	Iluminación	Total
Estrato 1	62	0	0	62
Estrato 2	253	0	0	253
Estrato 3	0	179	0	179
Estrato 4	156	0	0	156
Estrato 5	350	0	0	350
Estrato 6	47	226	0	293
Estrato 7	0	0	365	365
Estrato 8	144	0	0	144

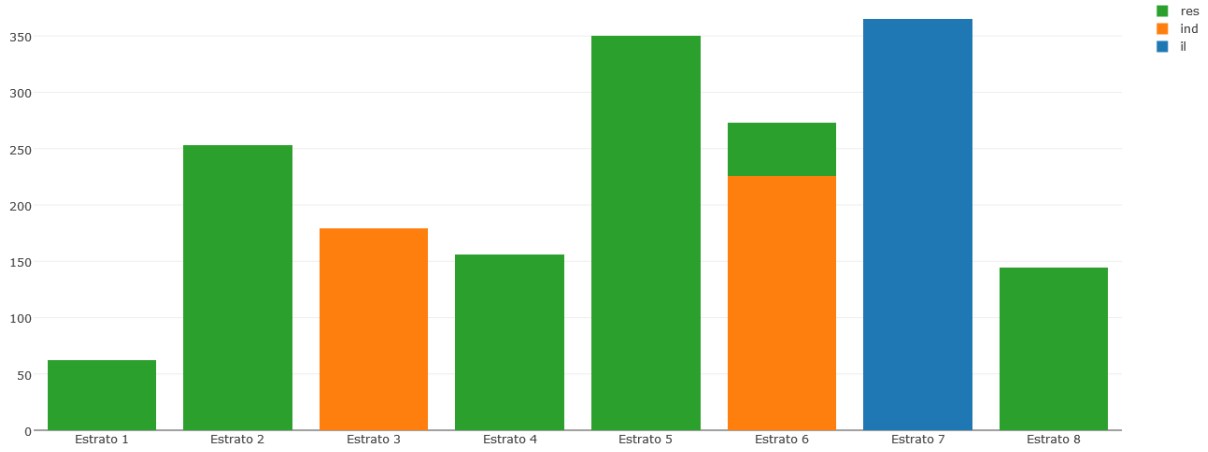


Figura 7.27: Identificación de estratos. Eje x: No de Usuarios. Eje y: Estratos

En la figura 7.27 se puede ver como se han distribuido efectivamente los estratos adquiriendo un patrón definido por sector; con excepción del estrato 6, el cual denota patrones compartidos tanto por usuarios industriales como residenciales. Dado que el porcentaje de usuarios residenciales dentro de este estrato es notablemente menor que el de usuarios industriales, se puede decir que son estos los que se asemejan en su tendencia de consumo a un usuario que se clasifique como industrial.

7.2.4. Potencia Agregada

La potencia agregada de la muestra, al igual que la estimación de la demanda se encuentran en la figura 7.28, con un error de esta última de $E = 5{,}81\,\%$ con respecto de la potencia agregada.

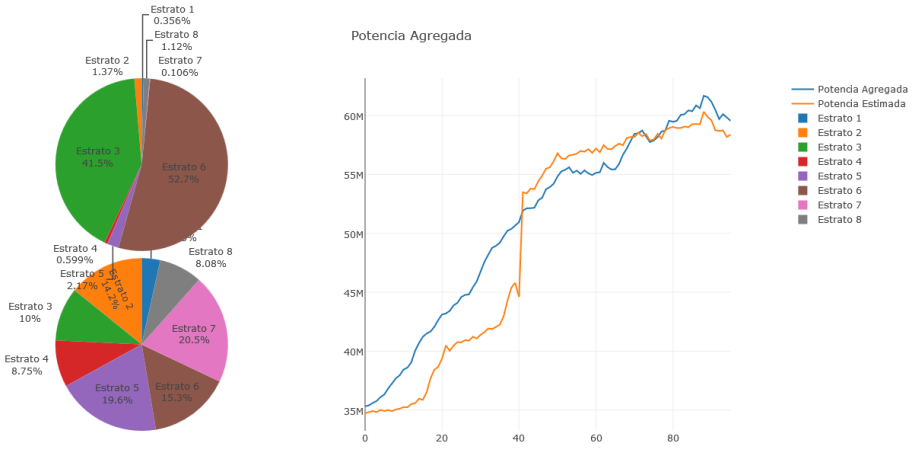


Figura 7.28: Potencia Agregada. Eje x: tiempo - 24 horas en intervalos de 15 min. Eje y: Vatios (w)

Horas	h. Pico de 17 a 23	h. fuera de Pico de 23 a 5	Standard de 5 a 11	Pico moderado de 11 a 17
Estrato 1	0.144835	0.0529861	0.128569	0.136358
Estrato 2	0.336036	0.289085	0.393407	0.324194
Estrato 8	0.315646	0.255485	0.181132	0.157021
Estrato 5	0.701414	0.910855	0.755803	0.743144
Estrato 4	0.290391	0.324	0.307415	0.278714
Estrato 6	50.8303	53.3534	50.0216	48.8078
Estrato 3	47.2845	44.6162	48.1663	49.5512
Estrato 7	0.096875	0.19799	0.0456964	0.00153626

Figura 7.29: % de consumo

La hora pico definida por el mayor consumo de energía, se encuentra entre las 5pm y las 11pm, y el cluster de mayor contribución corresponde a los Estrato 6 y 3, constituidos en su mayoría por usuarios industriales. El sector residencial, a pesar de ser mayor al doble a nivel poblacional no representa una carga considerable para el sistema en comparación a los estratos antes mencionados.

Dado la poca cantidad de usuarios alojados en el Estrato 1 se realizo utilizo la función *Reduccion()* para reducir los estratos y agrupar los usuarios en el Estrato 1 con el estrato cuyo patrón mas se asemejase, el cual en este caso resulta ser el Estrato 5. Dando como resultado un nuevo Estrato 1 compuesto por 412 usuarios de tipo residencial y la eliminación del anteriormente llamado Estrato5 (figura 7.30).

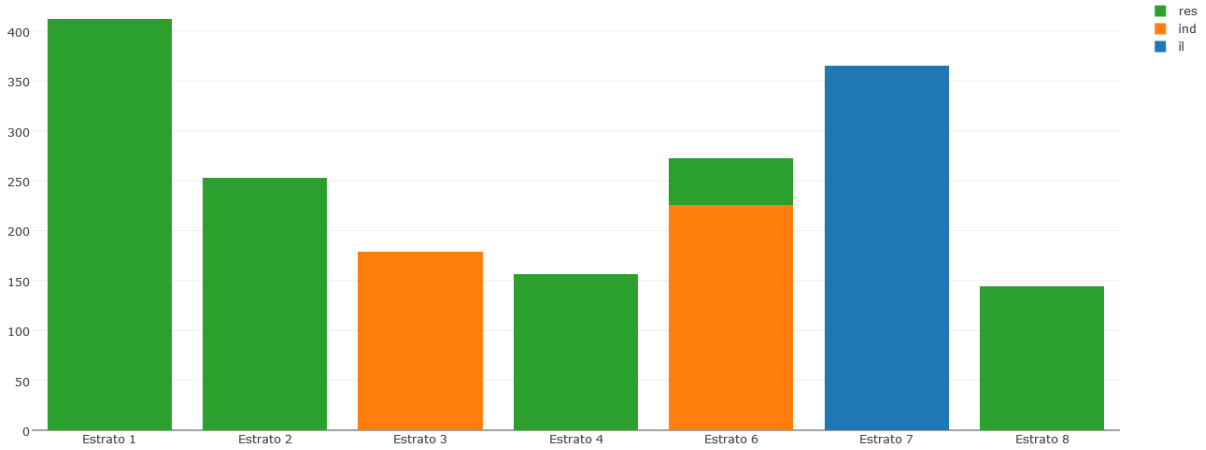


Figura 7.30: Nueva Estratificación

CONCLUSIONES Y RECOMENDACIONES

- En este proyecto se desarrolla una metodología implementada en software libre que permite modelar la demanda a través del reconocimiento de patrones de consumo en usuarios del sistema eléctrico. Este objetivo se concibió utilizando algoritmos de aprendizaje no supervisado, que lograron afianzarse a los patrones de la data de prueba con un errores entre 2 y 6 %. De la misma forma mostró un buen desempeño describiendo la contribución de los consumidores dentro de la demanda total de las muestras estudiadas.

- Se probaron distintos algoritmos de aprendizaje no supervisado para el reconocimiento de patrones en los perfiles de carga de usuarios del servicio eléctrico, entre los cuales destacó K-medianas por su simplicidad, efectividad al momento de agrupar muestras en clusters homogéneos y diferenciables, y su robustez ante valores atípicos. Lo que además lo convierte en un algoritmo potencialmente útil para el diseño de tarifas sustentables para el servicio.

- La curva de amalgamiento representa un punto de partida ayudando a estimar el numero de grupos al inicio del proceso de clustering. No obstante, solo es un punto de partida, es necesario llevar a cabo el proceso de estratificación repetidas veces en busca del resultado más satisfactorio y manejable para el analista, según los criterios de la empresa que lleva a cabo el estudio.

- En el sentido de lo expuesto en el punto anterior, se puede concluir también que la estratificación no es un proceso cien por ciento automatizable, por lo que es necesario la presencia del analista a cada paso del mismo para obtener un resultado satisfactorio.

- El programa desarrollado presenta ciertas limitantes en su interfaz gráfica, sin embargo el entorno de IPython permite la manipulación de la información con las funciones desarrolladas incluso cuando no se encuentra en ejecución el código principal, a través de lineas de comandos.

- Otra limitante para el desarrollo de estudios utilizando este programa es el hardware requerido, debido a que la data se aloja en la memoria RAM del computador a utilizar por lo que se recomienda la migración a librerías que no utilicen este método para cargar la información.

- Para dar mayor validez a los resultados se recomienda el uso de data real en su totalidad. La falta de acceso público a esta esta a nivel del sistema nacional merma el desarrollo de esta línea de investigación.

- El uso de modelos markovianos para generar data aleatoria, mostró muy buenos resultados ante el objetivo perseguido, permitiendo sortear el obstáculo de no contar con data real no residencial.

- Es posible continuar esta línea de investigación y extenderla hacia la desagregación de carga por equipos. En estos momentos la línea de investigación centrada en Modelos Ocultos de Markov para este propósito representa el estado del arte en desagregación de carga.

- Este programa puede ser mejor aprovechado para el análisis de información del sistema ante un escenario en el que se posea una infraestructura de medición bien consolidada.

- El país debería prepararse ante la inminencia y la necesidad de modernizar el sistema de medición, implementando y probando plataformas para el manejo de cantidades masivas de información.

REFERENCIAS BIBLIOGRÁFICAS

[1] L. R. Kulatunga N. Athula and K. S., "Smart meter based tools to enhance feedback oriented dsm with time of use rates," 2010.

[2] A. L. A. Macedo Maria N. Q., Galo Joaquin J.M. and L. A. C. C., "Typification of load curves for dsm in brazil for a smart grid enviroment," *ELSAVIER, Electrical Power and Energy Systems*, 2014.

[3] F. F. J. E. Valero Verdú Sergio Ortiz and G. L. Emilio, "Characterization and identification of electrical customers through the use of self-organizing maps and daily load parameters," 2004.

[4] N. R. Chicco Gianfranco and P. Federico, "Aplication of clustering algorithms and self organising maps to classify electricity customers," *IEEE Bologna PowerTech Conference*, 2003.

[5] M. P. Mahmoudi-Kohan, N. Moghaddam and S. M. Bidaki, "Evaluating performance of wfa k-means and modified follow the leader," 2009.

[6] D. V. F. Fernández, "Identificación de perfiles de jugadores con algoritmos de aprendizaje automático y técnicas de minería de datos," *Proyecto de Grado, Universidad Simón Bolívar*, 2005.

[7] A. D. Sousa and F. I. P. Betancourt, "Clustering de datos numéricos por medio de algoritmos basados en población e inteligencia colectiva," *Proyecto de Grado, Universidad Simón Bolívar*, 2011.

[8] C. V. Quintero García, "Estimación de la demanda para la localización de perdidas no-técnicas por punto de transformación," *Trabajo especial de grado Universidad Simón Bolívar*, 2012.

[9] L. T. Moreira Mota and A. de Assis Mota, "New trends on load modelling and forecasting: Models and techniques," *Electric Power: Generation, Transmission and Efficiency*, 2007.

[10] A. J. Pellicer Navarro, "Aspectos fundamentales de una infraestructura avanzada de medicion de energía eléctrica en venezuela," *Proyecto de Grado Universidad Simón Bolívar*, 2016.

[11] M. Q. Lucas, "Modelos de carga y sus efectos en estudios sistémicos: aplicaciones al sing," *Memoria para optar al título de ingeniero civil electricista, Universidad de Chile*, 2013.

[12] R. E. Villalobos, "Pronóstico de demanda eléctrica: Comparación de técnicas y una metodología de combinación.," *Trabajo de grado presentado ante la Universidad Central de Venezuela*, 1998.

[13] J. Nazarko and Z. A. Styczynsky, "Application of statistical and neural approaches to the daily load profiles modelling in power distribution systems," 1999.

[14] H. Tabares and J. Hernandez, "Mapeo curvas típicas demanda de energía elléctrica del sector residencial, comercial e industrial de la ciudad de medellin, usando redes neuronales artificiales y algoritmos de interpolación.," *Electric Power: Generation, Transmission and Efficiency*, 2008.

[15] M. de Energía y Minas and M. de Producción y Comercio, "Ley del servicio eléctrico nacional," *Gaceta No. 37.415*, 2002.

[16] D. J. Ramos Sergio, Duarte Joao M. and V. Zita, "A data-mining-based methodology to support mv electricity customers's characterization," *ELSAVIER, Energy and Building*, 2015.

[17] K. V. Diamantoulakis Panagiotis and K. George, "Big data analytics for dynamic energy management in smart grids," *ELSAVIER,Big Data Research*, 2015.

[18] M. E. Vázquez Villazón, Martha y de Bernard Cerezo, "El cluster analysis: su aplicación en el estudio de la validez de un instrumento psicodiagnóstico," *Revista cubana de Psicología*, 1987.

[19] $http://www.geocities.ws/r_vaquerizo/Manual_R10.html$.

[20] $http://www.pseg.com/business/energy_choice/third_party/load_profiles.jsp$, 2014.

[21] J. Z. Kolter and M. J. Johnson, "Redd: A public data set for energy disaggregation research," *SustKDD*, 2011.

[22] B. C. Kleiminger Wilhem and S. Silvia, "Household occupancy monitoring using electricity meters," *2015 ACM International Conference on Embedded Systems for Energy-Efficient Buildings*, 2015.

[23] L. R. Rabiner and B. H. Juang, "An introduction to hidden markov models," *IEEE ASSP MAGAZINE*, 1986.

[24] L. R. Rabiner and B. H. Juang, "A tutorial on hidden markov models and selected applications in speech recognition," *IEEE*, 1989.

[25] R. Moya, "Machine learning (en python), con ejemplos. ("work in progress")," *http://jarroba.com/expectation-maximization-python-scikit-learn-ejemplos/*, 2016.

[26] C. Bauckhage, "Numpy / scipy recipes for data science: k-medoids clustering," *ResearchGate*, 2015.

[27] T. Hair, Anderson and Black, "Análisis multivariante," *Prentice Hall, Quinta Edición.*

[28] M. J. Jang Dongsik Jiyong, Eorn Park and R. J. Jeung, "Variability of electricity load patterns and its effect on demand response: A critical peak pricing experiment on korean commercial and industrial customers," *ELSEVIER*, 2015.

[29] H. E. Ju P. and K. D., "Nonlinear dinamic load modelling: Model and parameter estimation," *IEEE Transaction on Power Systems*, 1996.

[30] B. Langmead, "Markov chains," *JOHNS HOPKINS Whiting School of Engineering, Department o Computer Science, www.langmead-lab.org/teaching-materials.*

[31] Z. L. Li Ya, Cheng Yu and H. Haitao, "Application of multivariable statistical analysis to classify electricity customers," *Ditribution Power Market*, 2008.

[32] M. A. Mota L.T.M. and M. A., "Load behaviour prediction under blackout conditions using a fuzzy expert system," *IET Gener. Transm. Distrib.*, 2007.

[33] P. Palensky and D. Dietrich, "Demand side management: Demand response, intelligent energy systems and smart loads," *IEEE Transacctions on Industrial Informatics, vol. 7*, 2011.

[34] O. Maimon and L. Rokach, "Data mining and knowledge discovery handbook," *Springer*, 2010.

[35] R. F. Rojas Moreno, "Caracterización del consumo de energía y potencial de ahorro del sector industrial venezolano," *Informe de Pasantía, Universidad Simón Bolívar*, 2012.

[36] M. R. Romero, "Posibilidad de discriminacion avo de arenas con fluidos y lutitas utilizando k-means y fuzzy k-means clustering," *Proyecto de Grado, Universidad Simón Bolívar*, 2005.

APÉNDICE A

CAPTURAS DE PANTALLA

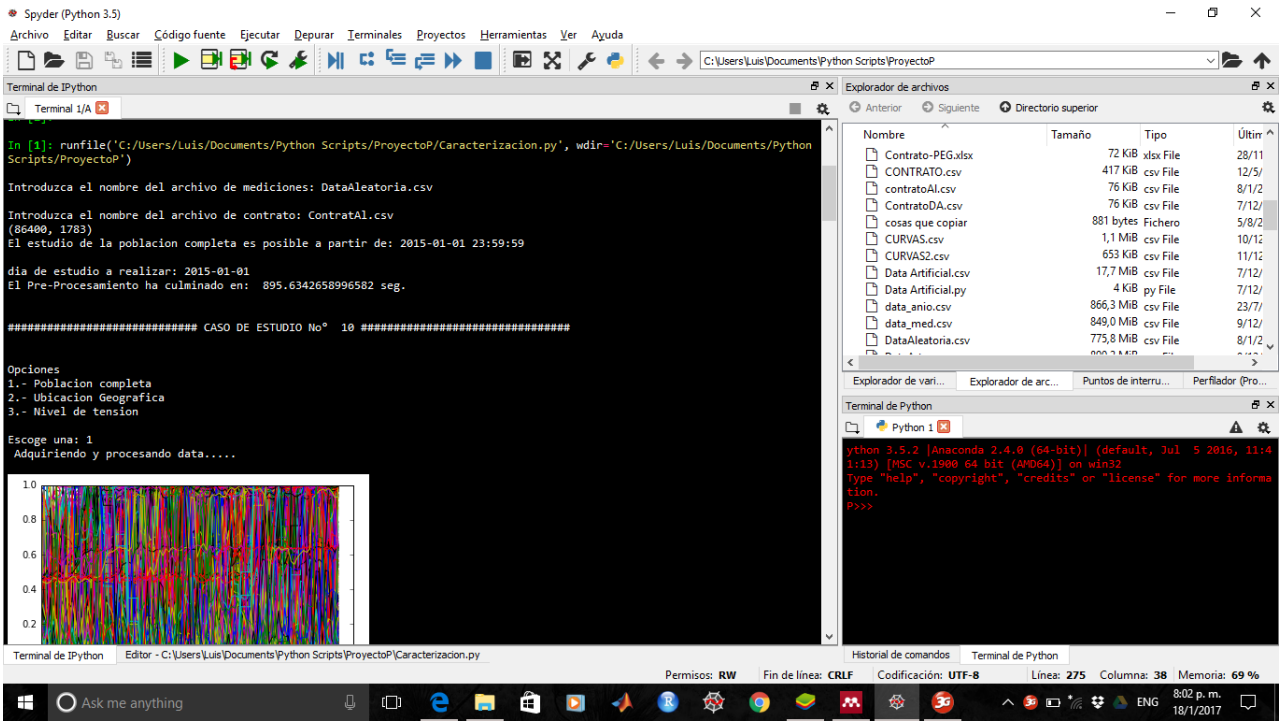


Figura A.1: Captura de pantalla. Inicio del programa: Introduccion de la data, pre-procesamiento e inicio del caso de estudio.

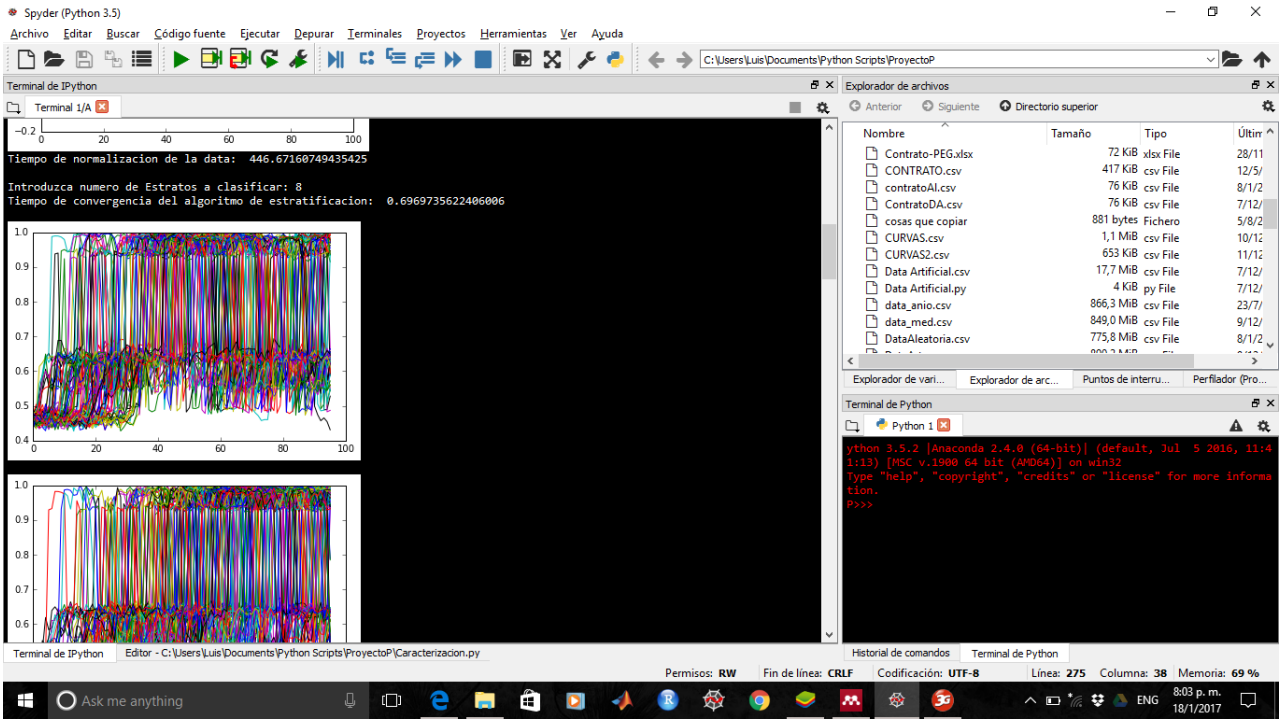


Figura A.2: Captura de pantalla del programa. Tiempo de normalización y asignación del número de estratos.

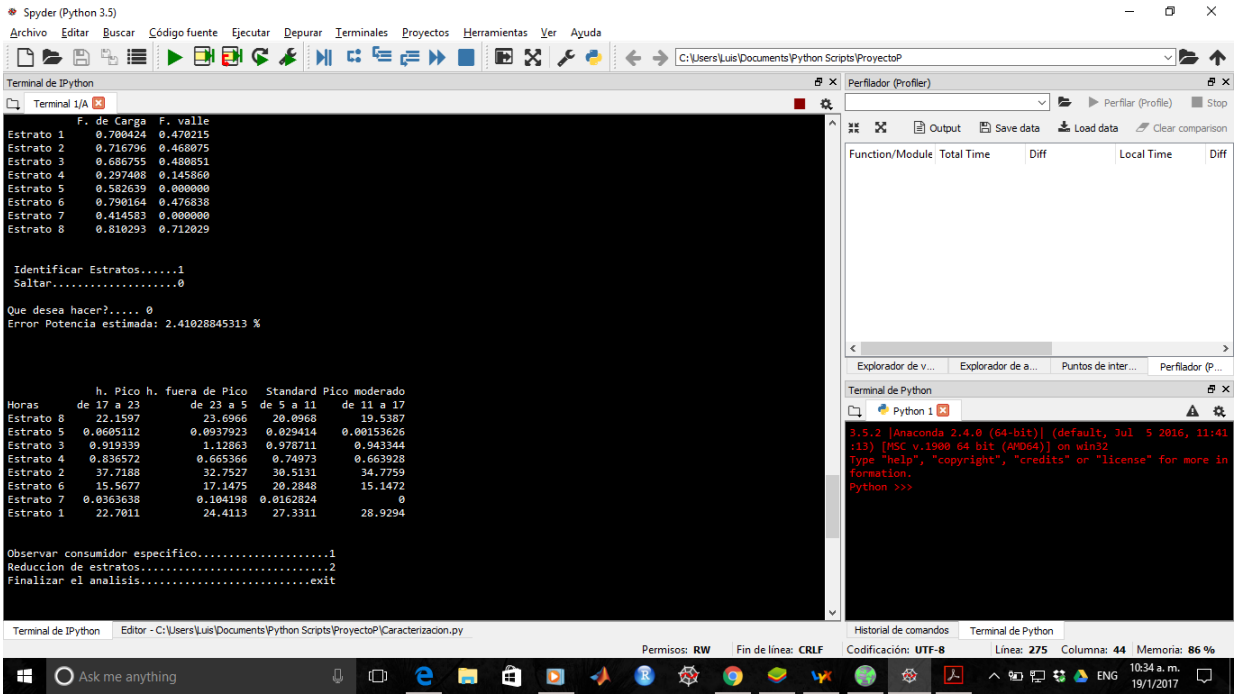


Figura A.3: Captura de pantalla. Factores que describen los estratos y contribución porcentual de cada estrato en rangos horarios.

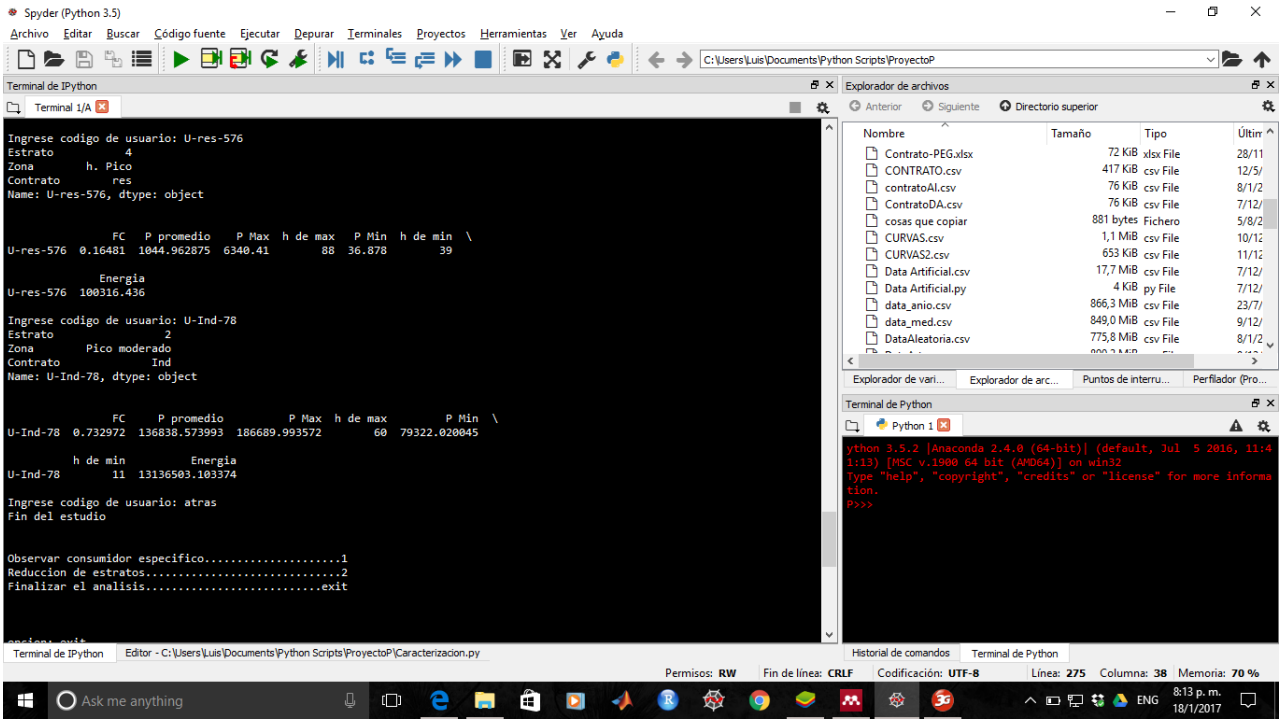


Figura A.4: Captura de pantalla. Descripción de dos usuarios aleatorios.