



**UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL**

**SEGMENTACIÓN DE PERFILES DE CONSUMO UTILIZANDO
VARIABLES LATENTES PARA EL MERCADO DE CLIENTES
REGULADOS DE CHILECTRA**

**MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL
INDUSTRIAL**

SEBASTIAN ANDRES ORELLANA MONTINI

**PROFESOR GUÍA:
RICHARD WEBER HAAS.**

**MIEMBROS DE LA COMISIÓN:
LUIS ABURTO LAFOURCADE.
OTTMAR CARDENAS VILLACURA.**

**SANTIAGO DE CHILE
2016**

SEGMENTACIÓN DE PERFILES DE CONSUMO UTILIZANDO VARIABLES LATENTES PARA EL MERCADO DE CLIENTES REGULADOS DE CHILECTRA

En esta memoria se presenta el diseño y aplicación de una metodología basada en la clusterización de series de tiempo utilizando modelos de variables latentes para descomponer y describir el mercado de clientes regulados de Chilectra durante un período de estudio que abarca los años 2013 y 2014.

Dicho mercado es de interés para la compañía pues en la actualidad solo se manejan tasas de crecimiento y análisis de situación a nivel agregado y se desea segmentar este mercado en segmentos de clientes con patrones de consumos distintivos entre ellos, para ello la compañía cuenta con una serie de data sets provenientes de mediciones realizadas por distintos equipos de medición a lo largo del anillo de concesión en distintos niveles jerárquicos

Para ello se propone utilizar la herramienta estadística de Análisis de Componentes Independientes, en particular se propone utilizar el algoritmo fastICA para encontrar distintas combinaciones de componentes independientes dependiendo de cada experimento y luego seleccionar aquel resultado que mejor permita explicar las señales de consumo observadas e interpretar dichos componentes como patrones de consumos asociados a grupos de clientes individuales.

Dicha metodología fue aplicada a dos data sets provenientes de equipos de medición asociados a transformadores y a subestaciones, para los cuales se aplicaron distintos experimentos con variaciones en la cantidad de componentes independientes a encontrar variando desde $k=2$ hasta $k=15$, posterior a esto se seleccionó el experimento $k=5$ como aquel que mejor combinaba los resultados obtenidos en términos de reconstrucción y la interpretabilidad, los 5 segmentos encontrados fueron denominados como “Industrial”, “Residencial Anómalo”, “Residencial Alto Consumo”, “Residencial Bajo Consumo” y “Pequeño Negocio”.

Como output del algoritmo se obtuvieron las señales asociadas a los componentes independientes los cuales fueron interpretados como segmentos de clientes distintos que poseen perfiles de carga diarios distintos, además se obtiene una matriz de mezcla que captura la forma en la cual estos segmentos se mezclan por cada equipo de medición para generar las señales observadas, las participaciones de mercado promedio estimadas a nivel de transformadores son de 20.9%, 17.7%, 27.5%, 14.4% y 19.6% mientras que a nivel de subestación son de 19.1%, 14.3%, 28.0%, 24,8% y 13,8% respectivamente. Con ellos se estimaron las tasas de crecimiento interanual asociadas a cada segmento encontrado, las cuales a nivel de transformador son de 2.0%, 7.9%, 7.6%, 5.5%, 4.3% respectivamente, mientras que a nivel de subestación son de 3.9%, -0.9%, 4.4%, 4.2% y 8.0% respectivamente.

Tabla de Contenido

Capítulo 1: Introducción.....	1
1.1 Descripción del problema	1
1.2 Justificación	3
1.2.1 Proceso de distribución de carga y operación de la red	4
1.2.2 Proceso de facturación y compra de energía a corto plazo	5
1.2.3 Proceso de facturación y compra de energía a largo plazo	5
1.2.4 Análisis de tarifas y tramos tarifarios.....	5
1.3 Objetivos	6
1.3.1 Objetivo General	6
1.3.2 Objetivo Especifico.....	6
1.4 Alcance	6
1.5 Resultados Esperados	7
Capítulo 2: Marco Conceptual	8
2.1 Segmentación de curvas de carga	8
2.2 Clusterización de Series de Tiempo	8
2.3 Análisis de Componentes Independientes	10
2.4 Cálculo de errores	14
2.5 Cálculo de correlaciones	14
Capítulo 3: Metodología.....	15
3.1 Obtención de la data	15
3.2 Pre-procesamiento	15
3.3 Selección de equipos de medición	15
3.4 Transformación de la data	17
3.5 Construcción del modelo.....	17
3.6 Validación	18
Capítulo 4: Análisis Descriptivo	20
4.1 Observaciones data set transformadores	20
4.1.1 Descripción Curva Sistema	21
4.1.2 Transformadores de respaldo	21
4.1.3 Transformadores menor consumo observado	23
4.1.4 Transformadores mayor consumo observado	24
4.1.5 Test de normalidad y comparación de los sub-conjuntos.....	25
4.2 Observaciones data set subestaciones	27

4.2.1 Subestaciones menor consumo observado	28
4.2.3 Subestaciones mayor consumo observado	30
4.2.4 Test de normalidad y comparación de los sub-conjuntos.....	30
Capítulo 5: Resultados y Análisis	31
5.1 Transformaciones data sets transformadores y subestaciones.....	31
5.2 Descripción iteraciones algoritmo fastICA	32
5.3 Resultados y análisis a nivel de transformador	33
5.3.1 Cálculo de errores de reconstrucción	33
5.3.2 Principales iteraciones.....	36
5.3.2.1 Experimento k=4	36
5.3.2.2 Experimento k=5	36
5.3.2.3 Experimento k=10	37
5.3.2.4 Experimento k=15	37
5.4 Principales segmentos encontrados	37
5.4.1 Caracterización segmentos encontrados.....	38
5.4.2 Market-shares y tasas de crecimiento por transformador.....	41
5.4.3 Ejemplo reconstrucción 5 transformadores tipo.....	41
5.5 Resultados y análisis a nivel de subestacion	44
5.5.1 Cálculo de errores de reconstrucción	44
5.5.2 Principales iteraciones.....	46
5.5.2.1 Experimento k=4	46
5.5.2.2 Experimento k=5	46
5.5.2.3 Experimento k=10	46
5.5.2.4 Experimento k=15	46
5.6 Principales segmentos encontrados	47
5.6.1 Market-shares y tasas de crecimiento por transformador.....	48
5.6.2 Ejemplo reconstrucción 5 subestaciones tipo.....	50
Capítulo 6: Conclusiones	51
6.1 Principales conclusiones del estudio	51
6.2 Principales resultados obtenidos	52
6.3 Principales limitaciones de los resultados obtenidos.....	53
6.4 Trabajos futuros.....	54
Capítulo 7: Bibliografía	55
Capítulo 8: Anexos	57
Capítulo 9: Apéndices	90

Índice de figuras

<i>Figura 1: Area de concesión de la compañía</i>	57
<i>Figura 2: Tasa crecimiento energía por tipo de cliente</i>	57
<i>Figura 3: Tasa crecimiento total de clientes y ventas físicas de energía</i>	58
<i>Figura 4: Anillo de concesión</i>	58
<i>Figura 5: Esquema anillo de concesión</i>	58
<i>Figura 6: Visualización data sets subestaciones Club Hípico y Alonso de Córdova</i>	60
<i>Figura 7: Mapa conceptual herramientas de segmentación</i>	60
<i>Figura 8: Tabla medidas de distancia y/o similitud</i>	61
<i>Figura 9: Mapa conceptual enfoques para segmentar series de tiempo</i>	61
<i>Figura 10: Consumo Total Sistema</i>	20
<i>Figura 11: Consumo total del sistema visualización horaria</i>	62
<i>Figura 12: Descomposición serie consumo total sistema</i>	21
<i>Figura 13: Distribución consumos observados transformadores de respaldo</i>	22
<i>Figura 14: Consumo total transformadores de respaldo</i>	62
<i>Figura 15: Descomposición serie de consumo transformadores de respaldo</i>	22
<i>Figura 16: Distribución consumos transformadores menor consumo</i>	23
<i>Figura 17: Consumo total transformadores menor consumo</i>	63
<i>Figura 18: Descomposición serie de consumo transformadores menor consumo</i>	24
<i>Figura 19: Distribución consumos transformadores mayor consumo</i>	24
<i>Figura 20: Consumo total transformadores mayor consumo</i>	63
<i>Figura 21: Descomposición serie de consumo transformadores mayor consumo</i>	25
<i>Figura 22: Comparación menor consumo, mayor consumo y total sistema</i>	26
<i>Figura 23: Distribución consumos subestaciones menor consumo</i>	27
<i>Figura 24: Consumo total subestaciones menor consumo</i>	64
<i>Figura 25: Descomposición serie de consumo subestaciones menor consumo</i>	28
<i>Figura 26: Distribución consumos subestaciones mayor consumo</i>	29
<i>Figura 27: Consumo total subestaciones mayor consumo</i>	64
<i>Figura 28: Descomposición serie de consumo subestaciones mayor consumo</i>	29
<i>Figura 29: Consumo total calculado raw data</i>	31
<i>Figura 30: Consumo total calculado data transformada</i>	31
<i>Figura 31: Varianza explicada vs cantidad de componentes independientes</i>	32
<i>Figura 32.a: MAPE por cada experimento</i>	33
<i>Figura 32.b: WMAPE por cada experimento</i>	33
<i>Figura 33: MAE para cada experimento</i>	34
<i>Figura 34: N-RMSE para cada experimento</i>	35
<i>Figura 35: Serie sistema observada vs serie sistema estimada</i>	36
<i>Figura 36: Visualización componentes independientes experimento $k = 15$ $IC = 1, \dots, 5$</i>	65
<i>Figura 37: Visualización componentes independientes $IC = 6, \dots, 10$</i>	65
<i>Figura 38: Visualización componentes independientes $IC = 11, \dots, 15$</i>	66
<i>Figura 39: Visualización perfiles carga experimento $k = 15$ con $IC = 1, \dots, 5$</i>	66
<i>Figura 40: Visualización perfiles carga experimento $k = 15$ $IC = 6, \dots, 10$</i>	67
<i>Figura 41: Visualización perfiles carga experimento $k = 15$ $IC = 11, \dots, 15$</i>	67
<i>Figura 42: Visualización componentes independientes experimento $k = 10$ $IC = 1, \dots, 5$</i>	68
<i>Figura 43: Visualización componentes independientes $IC = 6, \dots, 10$</i>	68

<i>Figura 44: Visualización perfiles carga experimento $k = 10$ con $IC=1, \dots, 5$</i>	69
<i>Figura 45: Visualización perfiles carga experimento $k=10$ $IC=6, \dots, 10$</i>	69
<i>Figura 46: Visualización componentes independientes experimento $k = 5$ $IC=1, \dots, 5$</i>	70
<i>Figura 47: Visualización perfiles carga experimento $k = 5$ con $IC=1, \dots, 5$</i>	70
<i>Figura 48: Visualización componentes experimento independientes $k = 4$ $IC=1, \dots, 4$</i>	71
<i>Figura 49: Visualización perfiles carga experimento $k = 5$ con $IC=1, \dots, 4$</i>	71
<i>Figura 50: Segmentos obtenidos experimento $k=5$</i>	38
<i>Figura 51: Máximo nivel de consumo mensual por segmento</i>	71
<i>Figura 52: Consumo medio mensual por segmento</i>	72
<i>Figura 53: Factor de carga mensual para los distintos segmentos encontrados</i>	72
<i>Figura 54: Máximo consumo semanal por segmento</i>	40
<i>Figura 55: Consumo medio semanal por segmento</i>	40
<i>Figura 56: Factor de carga semanal por segmento</i>	41
<i>Figura 57: Tasas de crecimiento 2013-2014 por segmento</i>	42
<i>Figura 58: Relación entre el IMACEC y los consumos por segmento</i>	49
<i>Figura 59: Perfil observado vs. Perfil reconstruido Chacabuco T1</i>	77
<i>Figura 60: Perfil observado vs. Perfil reconstruido Alonso de Cordova T2</i>	78
<i>Figura 61: Perfil observado vs. Perfil reconstruido Apoquindo T1</i>	78
<i>Figura 62: Perfil observado vs. Perfil reconstruido Batuco T3</i>	79
<i>Figura 63: Perfil observado vs. Perfil reconstruido San Cristóbal T3</i>	79
<i>Figura 64: MAPE por cada iteración a nivel de subestación</i>	45
<i>Figura 65: Visualización componentes independientes experimento $k = 15$ $IC=1, \dots, 5$</i>	80
<i>Figura 66: Visualización componentes independientes $IC=6, \dots, 10$</i>	80
<i>Figura 67: Visualización componentes independientes $IC=11, \dots, 15$</i>	81
<i>Figura 68: Visualización perfiles carga experimento $k = 15$ con $IC=1, \dots, 5$</i>	81
<i>Figura 69: Visualización perfiles carga experimento $k=15$ $IC=6, \dots, 10$</i>	82
<i>Figura 70: Visualización perfiles carga experimento $k=15$ $IC=11, \dots, 15$</i>	82
<i>Figura 71: Visualización componentes independientes experimento $k = 10$ $IC=1, \dots, 5$</i>	83
<i>Figura 72: Visualización componentes independientes $IC=6, \dots, 10$</i>	83
<i>Figura 73: Visualización perfiles carga experimento $k = 10$ con $IC=1, \dots, 5$</i>	84
<i>Figura 74: Visualización perfiles carga experimento $k=10$ $IC=6, \dots, 10$</i>	84
<i>Figura 75: Visualización componentes experimento independientes $k = 5$ $IC=1, \dots, 5$</i>	85
<i>Figura 76: Visualización perfiles carga experimento $k = 5$ con $IC=1, \dots, 5$</i>	85
<i>Figura 77: Visualización componentes experimento independientes $k = 4$ $IC=1, \dots, 4$</i>	86
<i>Figura 78: Visualización perfiles carga experimento $k = 5$ con $IC=1, \dots, 4$</i>	86
<i>Figura 79: Perfiles de carga para la iteración $k=5$ aplicada al data set a nivel de subestación</i>	47
<i>Figura 80: Tasas de crecimiento para los segmentos encontrados durante el período de estudio</i>	48
<i>Figura 81: Relación entre el IMACEC y los consumos por segmento</i>	50
<i>Figura 82: Perfil observado vs Perfil reconstruido Lo Boza</i>	88
<i>Figura 83: Perfil observado vs Perfil reconstruido San Joaquín</i>	89
<i>Figura 84: Perfil observado vs Perfil reconstruido Chacabuco</i>	89
<i>Figura 85: Perfil observado vs Perfil reconstruido Apoquindo</i>	90
<i>Figura 86: Perfil observado vs Perfil reconstruido Alonso de Cordova</i>	90

Índice de tablas

<i>Tabla 1: Estadísticos descriptivos para la curva total del sistema</i>	<i>23</i>
<i>Tabla 2: Test de Anscombe-Glynn para los distintos grupos de transformadores</i>	<i>30</i>
<i>Tabla 3: Test de Anscombe-Glynn para los distintos grupos de subestaciones.....</i>	<i>34</i>
<i>Tabla 4: MAE, MAPE y N-RMSE de la curva total reconstruida para cada experimento..</i>	<i>39</i>
<i>Tabla 5: Market-Share por segmento para los primeros 11 transformadores</i>	<i>73</i>
<i>Tabla 6: Market-Share por segmento para los siguientes 11 transformadores.....</i>	<i>73</i>
<i>Tabla 7: Market-Share por segmento para los siguientes 11 transformadores.....</i>	<i>74</i>
<i>Tabla 8: Market-Share por segmento para los siguientes 11 transformadores.....</i>	<i>74</i>
<i>Tabla 9: Market-Share por segmento para los siguientes 11 transformadores.....</i>	<i>74</i>
<i>Tabla 10: Market-Share por segmento para los siguientes 11 transformadores.....</i>	<i>75</i>
<i>Tabla 11: Market-Share por segmento para los siguientes 11 transformadores.....</i>	<i>75</i>
<i>Tabla 12: Market-Share por segmento para los siguientes 11 transformadores.....</i>	<i>75</i>
<i>Tabla 13: Market-Share por segmento para los siguientes 11 transformadores.....</i>	<i>76</i>
<i>Tabla 14: Market-Share por segmento para los siguientes 11 transformadores.....</i>	<i>76</i>
<i>Tabla 15: Tasas de crecimiento por segmento encontrado a nivel de transformador.....</i>	<i>49</i>
<i>Tabla 16: Correlación tasa de crecimiento del IMACEC y los consumos mensuales</i>	<i>49</i>
<i>Tabla 17: Market-shares para los transformadores mencionados anteriormente.</i>	<i>50</i>
<i>Tabla 18: MAPE curva sistema por cada experimento.....</i>	<i>51</i>
<i>Tabla 19: Market-share por segmento para las primeras 11 subestaciones</i>	<i>86</i>
<i>Tabla 20: Market-share por segmento para las primeras 11 subestaciones</i>	<i>86</i>
<i>Tabla 21: Market-share por segmento para las primeras 11 subestaciones</i>	<i>86</i>
<i>Tabla 22: Market-share por segmento para las primeras 11 subestaciones</i>	<i>88</i>
<i>Tabla 23: Tasas de crecimiento por segmento encontrado</i>	<i>55</i>
<i>Tabla 24: Correlación tasa de crecimiento del IMACEC y los consumos mensuales</i>	<i>56</i>
<i>Tabla 25: Market-shares para los transformadores mencionados anteriormente.....</i>	<i>56</i>

Capítulo 1: Introducción

1.1 Descripción del problema

Chilectra corresponde a la empresa más grande de distribución eléctrica del país representando cerca de un 40% de la venta total neta a nivel nacional al año 2013. Desde un punto de vista operacional la industria de distribución eléctrica es una actividad altamente regulada y fiscalizada por distintos organismos estatales como lo son la CNE¹ quien actúa como ente regulador, el CDEC² quien actúa como operador del sistema y la SEC³ organismo que cumple el rol de fiscalizador de las distintas empresas de los distintos mercados energéticos en el país.

Tal como se ve en los anexos *figura 1* el área de concesión de la empresa es de 2.037[km^2] y abarca 33 comunas de la región metropolitana: Cerrillos, Cerro Navia, Conchalí, Estación Central, Independencia, La Cisterna, La Florida, La Granja, La Reina, Las Condes, Lo Espejo, Lo Prado, Macul, Maipú, Ñuñoa, Pedro Aguirre Cerda, Peñalolén, Pudahuel, Quinta Normal, Recoleta, Renca, San Joaquín, San Miguel, San Ramón, Vitacura, Santiago, Providencia, Huechuraba, Quilicura, Lo Barnechea, Colina, Lampa y Til Til. Si se considera también las áreas de concesión correspondientes a las filiales Empresa Eléctrica de Colina Ltda. y Luz Andes Ltda. el área total de concesión alcanza los 2.118[km^2].

La red administrada por Chilectra puede descomponerse en tres subconjuntos jerárquicos, Alta, Media y Baja tensión. En términos de kilómetros de red hay cerca de 355[km] de líneas asociadas a la alta tensión, 4.993[km] asociados a la media tensión y 10.476[km] asociados a las líneas de baja tensión, de acuerdo a la última memoria emitida por la empresa correspondiente al año 2013.

Asociados a cada línea de alta, media y baja tensión se encuentran distintos elementos que conforman el sistema eléctrico, para el desarrollo de esta memoria en particular nos interesan aquellos elementos del sistema eléctrico que tienen equipos de tele-medidas asociados, mediante los cuales se puede registrar el consumo medido como potencia cada ciertos intervalos de tiempo.

Ordenados jerárquicamente desde aguas arriba a aguas abajo en el anillo de concesión los elementos relevantes son los siguientes:

- Subestaciones
- Transformadores de poder
- Alimentadores

En total y considerando las distintas medidas de tensión al año 2013 el sistema se encontraba compuesto por 54 subestaciones, 183 transformadores de poder y 563 alimentadores, considerando equipos de respaldo y equipos no operacionales actualmente.

¹ Comisión Nacional de Energía.

² Centro de Despacho de Carga.

³ Superintendencia de Energía y Combustibles.

La empresa por ende posee tres tipos de medidas de consumo horario, las cuales van desde un menor nivel de agregación como son las medidas asociadas a los alimentadores hasta un mayor nivel de agregación como son las medidas asociadas a las subestaciones.

Dentro del organigrama de la empresa asociada a la Gerencia Técnica se encuentra la Unidad de Movimiento de Energía⁴, la cual se encarga de administrar, monitorear y evaluar las medidas generadas por los distintos equipos de medición, esta unidad tiene el interés de desarrollar un modelo de segmentación de mercado para los clientes regulados de Chilectra utilizando los datos asociados a los consumos registrados en los distintos equipos de medición.

Dado el marco regulatorio actual existen dos tipos de clientes que consumen energía del sistema eléctrico, los clientes libres, quienes tienen un umbral de consumo horario mayor a los 2,5 [MWh] y los clientes regulados, quienes consumen por debajo de dicho umbral.

En la actualidad la Gerencia Comercial realiza un seguimiento del consumo asociado a cuatro tipos de clientes regulados, sin embargo dicha tipificación es realizada ex ante al momento en que los clientes son conectados a la red y declaran el giro comercial asociado a su actividad, por ende dicha segmentación no considera los patrones de consumo de los clientes, los cuales son registrados a nivel agregado por las medidas manejadas en la Unidad de Movimiento de Energía.

Siguiendo esta segmentación realizada por la Gerencia Comercial se distinguen tres grandes tipos de clientes, los clientes residenciales, clientes comerciales y clientes industriales, además de otra categoría denominada “Otros Clientes” que agrupa a municipalidades, firmas de agrícolas en los sectores norte y sur de Santiago, distribuidoras, entre otras.

Tal como se puede ver en los anexos *figura 2*, para el año 2013 los clientes residenciales experimentaron un crecimiento inter-anual en términos de consumo medido como [GWh] de un 7.1%, los clientes comerciales un crecimiento cercano al 10.1% y los clientes industriales un crecimiento de -5.1%, mientras que la categoría denominada “Otros Clientes” experimento un crecimiento de cerca del 4.9%.

Tal como se puede ver en la *figura 3* de los anexos el total de clientes, considerando libre y regulados aumento en 2.1% en términos absolutos, pasando desde 1.659.000 hasta 1.694.000, mientras que en términos de ventas totales de energía hubo un aumento de un 4.9% pasando desde 14.445 [GWh] hasta 15.152[GWh].

Sin embargo pese a estos antecedentes la Gerencia Técnica no tiene confianza en la segmentación realizada por la Gerencia Comercial debido a que no considera los distintos patrones de consumo registrados por los equipos de medición del sistema, y solo considera los consumos registrados en el ciclo de facturación de la empresa, los cuales presentan un desfase natural y no poseen la exactitud y frecuencia que presentan las mediciones manejadas por la Unidad de Movimiento de Energía.

⁴ UME.

Es por esta razón que la Gerencia Técnica tiene el interés de desarrollar un modelo de segmentación de mercado que utilice como inputs las mediciones de consumo asociado a los distintos equipos de medición del sistema, con el fin de descomponer estas distintas medidas agregadas en distintos perfiles de consumo y asociar estos a distintos tipos de clientes.

1.2 Justificación

De acuerdo a los antecedentes mencionados anteriormente existen diversos elementos jerárquicos en los cuales se realizan medidas con una frecuencia de 15 minutos para capturar el consumo ya sea aguas arriba o aguas abajo en el sistema eléctrico.

En la *figura 4* de los anexos se puede ver una representación gráfica del anillo de concesión con las principales subestaciones asociadas a este. Asociada a cada subestación se encuentra un sub conjunto de transformadores y asociados a cada transformador se encuentra otro sub conjunto de alimentadores, tal como se observa en el esquema conceptual del anexo *figura 5*.

Siguiendo lo anterior los consumos de los clientes finales no son observables pues solo se tienen observaciones respecto a la suma de los consumos de distintos grupos de clientes asociados a distintos elementos de medición. La empresa desea poder generar una segmentación en base a variables latentes, en este caso el consumo individual de cada cliente que no es observado por la empresa.

Las mediciones que son observables por la empresa constan de medidas realizadas en intervalos de 15 minutos y pueden tomar cualquier valor en el dominio de los números reales. La Unidad de Movimiento de energía tiene data histórica desde el año 2011 pero para realizar el proyecto de descripción y descomposición del consumo observado se utilizara la data histórica que abarca desde el 1 de Enero del 2013 hasta el 31 de Diciembre del mismo año.

Asociado a cada transformador y a cada subestación del sistema se cuenta con un vector de medidas horarias cuyas dimensiones son 1×96 , luego considerando que se tienen datos de 365 días del año por un período de dos años y que son 115 transformadores asociados a clientes regulados se tienen 8.059.200 data points asociados a los consumos agregados por transformador, por otro lado a nivel de subestación se cuenta con un vector de medidas horarias cuyas dimensiones son de 1×24 , luego considerando que se tienen datos de 365 días del año por un período de dos años y que son 46 subestaciones, se tienen 805.920 data points asociados a mediciones realizadas a nivel de subestaciones del sistema.

Una forma natural de visualizar la data almacenada y con la cual se va a trabajar consiste en representarla como una serie de tiempo, agregando las medidas de acuerdo a distintos criterios así por ejemplo en la *figura 6* de los anexos se puede observar dos tipos de visualizaciones tanto a nivel de subestación como de transformador de distribución de los equipos asociados a las subestaciones Alonso de Córdova y sus transformadores además de la subestación Club Hípico y sus transformadores.

Una de las formas más comunes de visualizar la data y que da cuenta de los perfiles horarios de consumo consiste en agrupar las medidas de acuerdo al consumo acumulado asociado a una de las 24 horas del día generando un perfil de consumo promedio dentro de un intervalo temporal como puede ser una semana, un mes, un trimestre o un año.

Finalmente es con esta data que se plantea generar un modelo de segmentación de mercado, dado la naturaleza de esta se deben segmentar las distintas series de tiempo, considerando algunas de las posibles frecuencias mencionadas anteriormente u otras con el objetivo de poder asociar distintas series temporales a distintos segmentos del mercado de clientes regulados.

Como alternativa a esta solución propuesta y siguiendo lo propuesto por [1] la empresa podría eventualmente tener acceso a data contextual, demográfica, geográfica, entre otras, levantada desde distintas fuentes como por ejemplo el censo del año 2012, y en base a esta información generar distintos perfiles de consumo, sin embargo la frecuencia deseada para realizar el análisis propuesto y contrastar los resultados obtenidos hace incompatible poder utilizar data proveniente de fuentes externas como el censo del año 2012 con el fin de obtener los perfiles deseados.

Como output del modelo propuesto y tal como se ha mencionado anteriormente se pretende obtener una serie de perfiles asociados a los distintos tipos de consumo de los distintos tipos de clientes encontrados y contrastar la evolución de los distintos segmentos encontrados a través del período en el cual se realiza el estudio. En términos formales el objetivo que busca la Gerencia Técnica al realizar este estudio es poder describir el mercado de clientes regulados asociados al área de concesión utilizando variables asociadas a los patrones de consumo que poseen estos clientes que permitan poder segmentar a los distintos clientes de acuerdo a sus patrones de consumo observables a nivel agregado. Los resultados obtenidos con el modelo de segmentación servirán como input para apoyar la toma de decisiones de las gerencias técnica y comercial.

Desde el punto de vista de la Gerencia Técnica y en particular de la Unidad de Movimiento de Energía obtener información descriptiva respecto a la composición del mercado regulado les permite apoyar decisiones en distintos procesos operativos que son parte de las tareas llevadas a cabo por la Gerencia y en particular de la unidad.

1.2.1 Proceso de distribución de carga y operación de la red

Dentro de la Gerencia Técnica existe una unidad encargada de realizar estudios de la eficiencia operacional y de la planificación y mantenimiento de los diversos equipos de medidas con el fin de asegurar la continuidad operacional de las distintas componentes del sistema de distribución, denominada Unidad de Estudios de Operación⁵, una de las tareas que realiza esta unidad consiste en la distribución de cargas entre alimentadores cuando se realizan cortes programados de alimentadores o transformadores debido a razones técnicas. Actualmente esta distribución de carga se realiza siguiendo el conocimiento experto de los analistas de la unidad pero no considera la composición del consumo asociado a los distintos

⁵ UEO

alimentadores o transformadores al momento de transferir carga de un elemento a otro. Los resultados obtenidos por el modelo en términos de una segmentación del consumo asociado a los distintos transformadores o alimentadores permitirían apoyar la redistribución de carga incorporando criterios de composición de consumo a nivel de alimentadores y/o transformadores con el objetivo de verificar el conocimiento experto de los analistas y de generar redistribuciones de carga con una distribución de consumo más estable y con menor dispersión para tener mayor estabilidad operacional.

1.2.2 Proceso de facturación y compra de energía en el corto plazo

Una de las principales tareas de la UME consiste en operar el proceso de facturación mensual ligado a los retiros de energía en los distintos puntos de medición, dicho proceso genera como output un archivo en formato .xlsx que es auditado por el ente regulador y utilizado para distintas tareas dentro de otras gerencias asociadas a la compañía, como por ejemplo control de gestión, análisis del total de compras, retiros en algún tramo específico de la red de distribución entre otros. Dicho proceso consume un porcentaje importante del total de horas hombres con las cuales cuenta la unidad debido a que requiere de una serie de chequeos internos para verificar que alzas o bajas en los consumos registrados en los distintos puntos de medición no se deban a errores en los equipos o en el procesamiento de los datos registrados y efectivamente capturen el comportamiento del mercado regulado.

Los resultados del modelo de segmentación propuesto apoyarían la auditoría de estos datos pues la información obtenida en cuando al mix de consumo registrado en los distintos equipos de medición complementarían el análisis del consumo en aquellos puntos de medición que presenten un mix de consumo asociado a un segmento de clientes con patrones de consumo más estables y por ende la presencia de variaciones importantes en los consumos registrados se encuentren más probablemente asociados a errores en el procesamiento de los datos o en los equipos de medición.

1.2.3 Proceso de facturación y compra de energía en el largo plazo

A diferencia del proceso de facturación y compras de energía con un horizonte de corto plazo, en el largo plazo los resultados del modelo de segmentación pueden apoyar decisiones respecto a la topología de la red, en particular información descriptiva respecto al mercado regulado puede servir como información de apoyo al momento de tomar decisiones en términos de aumentar la capacidad de distribución de carga en distintos tramos del anillo de concesión, asociado a la baja o alza sostenida en los consumos registrados asociado a algún segmento en específico descrito por el modelo.

1.2.4 Análisis de Tarifas y de Tramos Tarifarios

Una de las principales tareas de la Gerencia Comercial consiste en el análisis del consumo facturado asociado a cada uno de los distintos tramos tarifarios asociados a clientes individuales que históricamente han sido considerados como parte del mercado regulado. Dichos tramos pueden clasificarse BT-1, BT-2, BT-3 y BT-4, dependiendo del nivel de consumo y del punto de retiro de la carga asociada a dicho consumo.

Los resultados del modelo de segmentación pueden apoyar la auditoría de dichos datos de consumo facturado provenientes de la Gerencia Comercial, contrastándolos con los segmentos encontrados a nivel de mercado regulado.

Además los resultados obtenidos por el modelo de segmentación pueden complementarse con los datos provenientes del proceso de tarifación con el fin de asociar a cada equipo de medición un mix tarifario proveniente de asociar a algún tramo tarifario en particular los distintos segmentos de clientes encontrados por el modelo.

1.3 Objetivos

Para el presente estudio se plantean los siguientes objetivos:

1.3.1 Objetivo General

- Descomponer el consumo observado del mercado de clientes regulados de Chilectra en base a sus componentes principales utilizando sus perfiles de consumo.

1.3.2 Objetivo Especifico

- Generar y describir perfiles de consumo energético independientes utilizando un modelo de clusterización de variable latente.
- Estimar el market-share para cada segmento encontrado para cada equipo de medición asociado al data set de transformadores y al data set de subestaciones.
- Evaluar el ajuste del modelo propuesto utilizando diversas métricas y el conocimiento experto de la contraparte en Chilectra.
- Explorar la relación entre variables macro-económicas y los perfiles de consumo energético encontrados.

1.4 Alcance

Se definen los siguientes alcances para el estudio:

- Se realizara el análisis considerando solamente los datos correspondientes a los años 2013 y 2014, excluyéndose el resto de la data histórica que abarca el periodo correspondiente a los años 2011, 2012 y las mediciones hasta octubre del 2015.
- Para desarrollar el análisis se utilizaran paquetes estadísticos y herramientas ya desarrolladas en R y no se desarrollaran paquetes ni herramientas nuevas.
- Para la empresa se desarrollara un reporte escrito que contenga un análisis descriptivo, visualización y evolución dentro del período de estudio de los principales perfiles encontrados, sin embargo dicho reporte no incorporara ningún tipo de pronóstico y será más bien descriptivo de la situación actual correspondiente al mercado regulado.

- Para la empresa se desarrollara un Script en R, Matlab o proceso en Rapid Miner con el objetivo de automatizar la implementación de las herramientas que dan solución a la problemática propuesta, en particular en lo referente al modelo de segmentación.

1.5 Resultados Esperados

Se definen los siguientes resultados esperados para el estudio:

- Generar una descomposición del mercado regulado que tenga validez desde un punto de vista de las lógicas del negocio, el conocimiento experto de la contraparte, los perfiles de validación generados y los distintos criterios técnicos que puedan elaborarse para validar los segmentos encontrados.
- Elaborar un perfil de consumo asociado a cada segmento encontrado considerando distintas visualizaciones de la data, considerando perfiles de carga acumulados en horizontes de 24 horas mensuales y visualizaciones diarias, semanales y mensuales de los componentes encontrados.

Capítulo 2: Marco Conceptual

Siguiendo los objetivos expuestos anteriormente se plantea el siguiente marco conceptual bajo el cual se pretende desarrollar el modelo de segmentación propuesto.

2.1 Segmentación y Clusterización de Curvas de Carga

Siguiendo lo propuesto [2] existen distintos enfoques al momento de segmentar y/o clasificar curvas de consumo asociadas a sistemas eléctricos, en particular perfiles de consumo horario, en general se distinguen tres técnicas de análisis, las cuales a su vez se descomponen en distintas categorías de acuerdo a sus características individuales.

Dentro de las tres principales categorías se distinguen las redes neuronales tanto con enfoques supervisados, no supervisados y de aprendizaje híbrido, técnicas de lógica difusa y técnicas estadísticas más tradicionales tal como se puede ver en los anexos *figura 7*.

La literatura recoge distintos experimentos y trabajos utilizando distintos enfoques, por ejemplo en [3, 4], se combinan dos técnicas para segmentar y clasificar series de tiempo registradas de un parque industrial en España, Self Organizing Feature Maps para reducir la dimensionalidad de la data y posteriormente K-means para segmentar las señales transformadas, en dicho ejemplo el data set se compone de data histórica con medidas de frecuencia cada 15 [minutos] durante un período de un año. Por otro lado en [4], realizan un estudio similar utilizando data histórica respecto al consumo de electricidad en Sri Lanka, sin embargo solamente utilizan K-Means para distintas ventanas temporales y distintas visualizaciones de la data, ya sean agregadas en consumos promedios diarios, medidas con distinta frecuencia, entre otros enfoques.

Dado la importancia que se le da a K-means en las distintas aplicaciones y casos de estudios recopilados en la literatura en el *apéndice 9.1* se puede ver una caracterización de esta misma extraída de [5].

2.2 Clusterización de Series de Tiempo

Tal como se propone en [6] al segmentar un data set compuesto por series de tiempo se debe definir alguna noción de distancia o similitud, como las expuestas en la *figura 8* de los anexos, apropiada para el data set además de un algoritmo o técnica que permita segmentar dicho data set en forma adecuada, al igual que cualquier problema de clusterización en data sets convencionales, pero además el autor distingue ciertas características exclusivas asociadas a la segmentación de series de tiempo como por ejemplo transformar la data en series discretas o continuas, con sampleos uniformes o no uniformes en términos de la frecuencia de las observaciones, variadas o multivariadas, entre otras.

En general el autor distingue tres tipos de enfoques para clusterizar series de tiempo, aquel denominado Raw-Data Approach, que trabaja directamente sobre series de tiempo comparando la distancia entre ambas mediante distintos algoritmos de clusterización.

Por otro lado, Feature-based Approach consiste en transformar cada serie de tiempo en una matriz de atributos asociados a esta para luego aplicar alguna técnica de segmentación en particular y finalmente Model-based Approach convierte cada serie de tiempo en una serie de parámetros asociados a un modelo, asumiendo que la serie es generada por algún proceso en particular.

En la *figura 9* de los anexos se puede observar un esquema con los tres principales enfoques de clusterización.

Al igual que en la etapa de segmentación y clasificación de curvas de carga la literatura recoge distintos experimentos y trabajos que utilizan distintos enfoques para clusterizar series de tiempo.

Un ejemplo de segmentación de series de tiempo utilizando un enfoque de Raw-Data puede verse en [7] y en [1], los autores utilizan transformadas de Fourier sobre las series en el dominio temporal para descomponer las señales en sus frecuencias principales utilizando el algoritmo Fast Fourier Transformation el cual viene implementado en diversas herramientas como Matlab y R.

En el *apéndice 9.2* se puede observar una caracterización de la transformada de Fourier y un resumen del paquete Fast Fourier Transformation.

Por otro lado un ejemplo de Feature-Based Approach puede verse en [1], quien transforma las señales en matrices de datos sobre las cuales se pueden aplicar técnicas de clusterización. También propone extraer ciertos atributos de la serie antes de aplicar un algoritmo de FFT.

Entre los principales atributos que propone extraer de las señales se encuentran:

- Energía total consumida en un período calculada como:

$$E_{total} = \frac{1}{2} \sum_j^m \sum_i^n E_{ij}$$
con j índice de días, i índice de períodos en el día y E_{ij} consumo medio por período.
- Energía máxima promedio en un período. $E_{MD} = \frac{1}{m} \sum_{j=1}^m \max\{E_{ij}, 1 \leq i \leq n\}$.
- Factor diario de carga que caracteriza cuan pronunciado es el peak de consumo para un perfil dado calculado de la siguiente forma: $E_{LF} = \frac{1}{m} \sum_{j=1}^m \frac{\frac{1}{n} \sum_{i=1}^n E_{ij}}{\max\{E_{ij}, 1 \leq i \leq n\}}$.

2.3 Análisis de Componentes Independientes

Siguiendo a los autores clásicos en Análisis de Componentes Independientes como [8] y [10] se puede utilizar esta técnica de procesamiento de señales con el fin de descomponer observaciones “mixtas” que se generan por la mezcla de distintas señales que no son observables, las aplicaciones de la misma van desde el procesamiento de scanners y resonancias magnéticas, análisis de series temporales en finanzas, telecomunicaciones y reducción de ruidos en imágenes naturales.

Por otro lado de acuerdo [9] la validación de los componentes independientes desde un punto de vista técnico se puede realizar mediante técnicas de visualización de la data en 2-D y clusterización de los estimados encontrados para luego validar los clusters generados.

En general la idea central detrás del análisis de componentes independientes es resolver el sistema de ecuaciones $\mathbf{x} = \mathbf{A}\mathbf{s}$ con \mathbf{x} matriz de observaciones, \mathbf{A} matriz de mezcla aleatoria y \mathbf{s} vector de componentes independientes.

Este modelo puede ser reescrito en términos de las columnas de la matriz \mathbf{A} de la siguiente forma: $\mathbf{x} = \sum_{i=1}^n \mathbf{a}_i s_i$ este modelo de variables latentes es denominado modelo de Análisis de Componentes Independientes (ICA).

El principal problema a resolver consiste en que se desconoce la composición de la matriz \mathbf{A} por ende tiene que ser estimada, para ello se realizan dos supuestos:

- Los componentes s_i del vector s son independientes
- Los componentes s_i del vector s siguen una distribución no-Gaussiana.

La razón para que estos componentes sigan una distribución no-Gaussiana es que cualquier transformación ortogonal de un par de variables $\mathbf{x}_1, \mathbf{x}_2$ que siguen una distribución Gaussiana tiene la misma distribución que $\mathbf{x}_1, \mathbf{x}_2$ y estos dos son independientes.

Por ende al aplicar ICA solo se puede estimar \mathbf{A} hasta una transformación ortogonal de $\mathbf{x}_1, \mathbf{x}_2$ por ende la matriz \mathbf{A} se vuelve no identificable.

Recordando que decimos que un modelo estadístico es identificable si cumple la siguiente condición:

$\wp = \{P_\theta: \theta \in \Theta\}$ modelo estadístico es identificable si y solo si $P(\theta_1) = P(\theta_2) \rightarrow \theta_1 = \theta_2$

En otras palabras valores distintivos de θ_1, θ_2 tienen que corresponder a valores distintos de sus probabilidades.

Para estimar los componentes independientes asumimos que cada uno de ellos viene dado por la siguiente combinación lineal: $\mathbf{y} = \mathbf{w}^t \mathbf{x} = \sum_{i=1}^M \mathbf{w}_i \mathbf{x}_i$

La única condición que falta para que esta combinación sea uno de los componentes independientes \mathbf{s}_i es que \mathbf{w} sea una de las filas de \mathbf{A}^{-1} dado que en la práctica no se puede determinar \mathbf{A} tampoco podemos determinar \mathbf{w} sin embargo se puede calcular el siguiente estimador $\mathbf{y} = \mathbf{w}^t \mathbf{x} = \mathbf{w}^t \mathbf{A} \mathbf{s} = \mathbf{z}^t \mathbf{s}$ dado que la suma de dos variables aleatorias independientes es “más Gaussiana” que las variables por si mismas $\mathbf{z}^t \mathbf{s}$ se vuelve lo menos Gaussiana posible cuando equivale a uno de los \mathbf{s}_i .

Por ende se concluye que se puede escoger \mathbf{w} tal que maximice la no-Gaussianidad de $\mathbf{w}^t \mathbf{x}$, el cual con la transformación anterior tendría un solo componente distinto de cero y equivaldría a uno de los componentes independientes.

Dado lo anterior el método presenta algunas ambigüedades las cuales hay que considerar al momento de analizar las soluciones encontradas para un problema dado.

- 1) No se puede determinar la varianza de los componentes independientes
 Dado que tanto \mathbf{A} como \mathbf{s} son variables desconocidas, cualquier multiplicador escalar μ_i puede cancelarse aplicando el multiplicador inverso $\frac{1}{\mu_i}$ a la columna \mathbf{a}_i de la matriz \mathbf{A} . Por ende se puede fijar arbitrariamente la magnitud de los componentes independientes mediante la restricción de que cada componente tenga varianza igual a 1, es decir $\mathbf{E}\{\mathbf{s}_i^2\} = \mathbf{1}$.

Sin embargo dicha restricción deja implícita la restricción del signo del componente, no obstante tal como se menciona en [4] los autores argumentan que para corregir dicha ambigüedad basta ponderar cada componente encontrado por -1 para imponer que las soluciones tengan interpretación.

- 2) No se puede determinar el orden de los componentes independientes.
 Nuevamente dado que tanto \mathbf{A} como \mathbf{s} son desconocidas, se pueden intercambiar libremente los términos del modelo $\mathbf{x} = \sum_{i=1}^n \mathbf{a}_i \mathbf{s}_i$ dentro de la sumatoria.

Formalmente esto implica que el modelo $\mathbf{x} = \mathbf{A} \mathbf{s}$ puede reescribirse como $\mathbf{x} = \mathbf{A} \mathbf{P}^{-1} \mathbf{P} \mathbf{s}$ donde los elementos de la matriz de permutación \mathbf{P} corresponden a los elementos \mathbf{s}_i pero en distinto orden, mientras que $\mathbf{A} \mathbf{P}^{-1}$ puede interpretarse como una nueva matriz de mezcla que puede ser resuelta por las heurísticas para resolver el modelo ICA.

Sin embargo tal como se menciona en [11], en la mayoría de las aplicaciones se puede aproximar el orden de los elementos \mathbf{s}_i mediante la construcción de una lista semi-ordenada rankeando estos componentes en base a algún algoritmo como Búsqueda Exhaustiva, Heurísticas de orden sub-óptimo como TnA⁶ o ranking mediante L_∞ - norm.

⁶ Tenure-and-Acceptance, ver [11].

Por ende para estimar \mathbf{w} basta con maximizar la “no-Gaussianidad” para ello se puede tomar distintas medidas de esta como:

- Kurtosis:

$$\mathbf{kurt}(\mathbf{y}) = E\{\mathbf{y}^4\} - 3(E\{\mathbf{y}^2\})^2$$

- Negentropía:

$$J(\mathbf{y}) = H(\mathbf{y}_{gauss}) - H(\mathbf{y})$$

Donde la entropía se define de la siguiente forma:

$$H(\mathbf{y}) = - \int \mathbf{f}(\mathbf{y}) \log(\mathbf{f}(\mathbf{y})) \mathbf{d}\mathbf{y}.$$

Sin embargo dado que dicha integral en muchos casos no tiene una solución analítica se tiene la siguiente aproximación para la negentropía $J(\mathbf{y}) \approx \frac{1}{12} E\{\mathbf{y}^3\}^2 + \frac{1}{48} \mathbf{kurt}(\mathbf{y}^2)$

Dependiendo el tipo de aplicación esta aproximación puede no ser lo suficientemente estable debido a la poca robustez que presenta la kurtosis al momento de ser calculada empíricamente por lo cual generalmente se utiliza la siguiente aproximación para la negentropía

$J(\mathbf{y}) \approx \sum_{i=1}^p \mathbf{k}_i [E\{\mathbf{G}_i(\mathbf{y})\} - E\{\mathbf{G}_i(\mathbf{v})\}]^2$ con $\mathbf{k}_i > \mathbf{0}$ constante, $\mathbf{v} \sim N(\mathbf{0}, \mathbf{1})$, $\mathbf{y} \sim F(\mathbf{0}, \mathbf{1})$ con F una distribución cualquiera y \mathbf{G}_i una función no cuadrática.

En general se utilizan las siguientes aproximaciones para \mathbf{G}_i :

- $\mathbf{G}_1 = \frac{1}{a_1} \log \cosh(a_1 u)$

- $\mathbf{G}_2 = -e^{-u^2/2}$

Con $\mathbf{1} \leq a_1 \leq \mathbf{2}$ constante.

- Información mutua: $I(\mathbf{y}_1, \dots, \mathbf{y}_m) = \sum_{i=1}^m H(\mathbf{y}_i) - H(\mathbf{y})$ con $H(\mathbf{y}) = - \int \mathbf{f}(\mathbf{y}) \log(\mathbf{f}(\mathbf{y})) \mathbf{d}\mathbf{y}$ en caso de ser continua y reemplazando la integral por sumatoria y las densidades por probabilidades en caso de ser continua.

Además antes de poder aplicar un modelo de ICA sobre un set de observaciones que se asumen son mixtas estas deben sufrir dos pre-procesamientos:

- Centering

Se construye un nuevo set de observaciones transformado como sigue:

$$\tilde{\mathbf{x}} = \mathbf{x} - E\{\mathbf{x}\}.$$

- Whitening

Se construye un nuevo set de observaciones como sigue: $\tilde{\mathbf{x}} = \mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^t\tilde{\mathbf{x}}$ tal que se cumpla la siguiente condición $E\{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^t\} = \mathbf{I}$

Para lograr esto se utiliza la descomposición EVD⁷ o descomposición espectral de la matriz de covarianza $E\{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^t\} = \mathbf{E}\mathbf{D}\mathbf{E}^t$ con \mathbf{E} matriz ortogonal de los vectores propios de $E\{\tilde{\mathbf{x}}\tilde{\mathbf{x}}^t\}$ y \mathbf{D} matriz de sus valores propios.

Con el objetivo de encontrar esta estimación del vector \mathbf{w} se utiliza la herramienta fastICA, el cual implementa el algoritmo fastICA, desarrollado por los autores antes mencionados, en R.

Tal como se menciona en [13], para estimar el vector \mathbf{w} se debe resolver el siguiente problema de maximización:

$$\begin{aligned} \max \quad & \sum_{i=1}^N J(y_i) = \sum_{i=1}^N J(\mathbf{w}_i^t \mathbf{x}) \\ \text{sa: } & E\{(\mathbf{w}_k^t \mathbf{x})(\mathbf{w}_j^t \mathbf{x})\} = \delta_{jk} \end{aligned}$$

Donde la restricción corresponde a la restricción impuesta por la etapa de whitening del modelo, es decir una restricción que impone decorrelación entre las filas de la matriz \mathbf{x} de observaciones.

Siguiendo lo propuesto por [12] y se tiene el siguiente pseudo-código del algoritmo fastICA:

- 1) Se aplican los procesos de “centering” y “whitening” a la matriz \mathbf{x} de observaciones
- 2) Se computan las siguientes funciones:
 $\mathbf{g}_1 = \tanh(\mathbf{a}_1 \mathbf{u})$ y $\mathbf{g}_2 = \mathbf{u}e^{\frac{-\mathbf{u}^2}{2}}$ donde se observa que tanto \mathbf{g}_1 como \mathbf{g}_2 cumplen que $\frac{d}{du} \mathbf{G}_i(\mathbf{u}) = \mathbf{g}_i(\mathbf{u})$
 Con $\mathbf{1} \leq \mathbf{a}_1 \leq \mathbf{2}$
- 3) Se escoge N = cantidad componentes independientes, se inicializa el contador en $\mathbf{p} \rightarrow \mathbf{1}$
- 4) Se inicializa \mathbf{w}_p en forma aleatoria.
- 5) Se actualiza: $\mathbf{w}_p^+ = E\{\mathbf{x}\mathbf{g}(\mathbf{w}_p^t \mathbf{x})\} - E\{\mathbf{g}'(\mathbf{w}_p^t \mathbf{x})\}\mathbf{w}$
- 6) Se ortogonaliza $\mathbf{w}_p \leftarrow \mathbf{w}_p - \sum_{j=1}^{p-1} (\mathbf{w}_p^t \mathbf{w}_j) \mathbf{w}_j$
- 7) Se normaliza $\mathbf{w}_p^+ = \frac{\mathbf{w}_p^+}{\|\mathbf{w}_p^+\|}$
- 8) Si el algoritmo no converge, es decir $E\{\mathbf{x}\mathbf{g}(\mathbf{w}_p^t \mathbf{x})\} - E\{\mathbf{g}'(\mathbf{w}_p^t \mathbf{x})\}\mathbf{w} < \epsilon$ se retorna a la etapa de actualización de \mathbf{w}_p^+ , es decir a la etapa 5.
- 9) Se actualiza el contador $\mathbf{p} \rightarrow \mathbf{p} + \mathbf{1}$
- 10) Si $\mathbf{p} < N$ se retorna a la etapa 4.

⁷ Eigen-Value-Descomposition.

2.4 Cálculo de errores

Siguiendo lo propuesto por [15] se utilizan las siguientes métricas para el cálculo de los errores entre la reconstrucción realizada con los componentes independientes de la matriz de observaciones \tilde{x} y la matriz de las series originales x .

1. MAPE: $M = \frac{1}{n} \sum_{t=1}^n \frac{x_t - \tilde{x}_t}{x_t}$
2. RMSE: $RMSE = \sqrt{\frac{\sum_{t=1}^n (x_t - \tilde{x}_t)^2}{n}}$
3. N-RMSE: $N - RMSE = \frac{RMSE}{y_{max} - y_{min}}$
4. MAE: $MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|$

2.5 Cálculo de correlaciones

Siguiendo lo propuesto por [15] se tienen los siguientes coeficientes de correlación, los cuales serán utilizados para analizar la relación entre distintas variables aleatorias continuas:

1. **Coefficiente de Spearman**

$$\rho = 1 - \frac{6 \sum d_i^2}{n^2(n-1)}$$

Donde $d_i = x_i - y_i$ con x_i, y_i realizaciones de las variables aleatorias X, Y, además se tiene que $\rho \in [-1, 1]$.

2. **Coefficiente de Pearson**

$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$ donde $cov(X,Y) = E\{XY\} - E\{X\}E\{Y\}$ corresponde a la covarianza entre las variables aleatorias continuas X e Y y σ_X, σ_Y corresponden a las desviaciones estándares para ambas variables aleatorias. Además se tiene que $\rho_{X,Y} \in [-1, 1]$.

3. **Coefficiente de Ranking de Kendall**

Siguiendo a [16] sean (x_i, y_j) realizaciones de las variables aleatorias X, Y tales que que todos los valores (x_i, y_j) son únicos.

Para cualquier par de observaciones (x_i, y_j) con $i \neq j$ se dice que son *concordantes* si se cumple que: $(x_i > x_j \text{ y } y_i > y_j)$ o $(x_i < x_j \text{ y } y_i < y_j)$, en caso contrario se dice que son *no-concordantes*.

Entonces se define la tau de Kendall como sigue:

$$\tau = \frac{(\text{Cantidad de pares concordantes}) - (\text{Cantidad de pares no - concordantes})}{\frac{1}{2} n(n-1)}$$

Capítulo 3: Metodología

Metodológicamente se consideran las siguientes etapas en la elaboración del modelo, siguiendo lo propuesto por [15] además de basarse en las etapas clásicas de la metodología del proceso KDD.

3.1 Obtención de la data

En total la Unidad de Movimiento de Energía puso a disposición del alumno las bases de datos asociados a las mediciones tanto de los alimentadores, transformadores y subestaciones para la ventana temporal que abarca desde el 1 de Enero del 2013 hasta el 31 de Diciembre del 2104.

Dado que a nivel de alimentadores la data presenta un alto nivel de distorsión debido a las mediciones asociadas a clientes libres se consideran dos data sets, asociados a los transformadores y a las subestaciones del anillo de concesión.

A nivel de transformadores el data set cuenta con un universo de 183 equipos de medición lo que considerando vectores de observaciones diarias de 1×96 durante el período de estudio generan 12.824.640 data points.

Por otro lado a nivel de subestaciones el data set cuenta con un universo de 46 equipos de medición lo que considerando vectores de observaciones diarias de 1×96 durante el período de estudio generan 3.223.680 data points.

Sin embargo como se discutirá en secciones posteriores ambos data sets reducirán considerablemente su dimensionalidad producto de las funciones de agregación aplicadas y de los filtros utilizados para descartar sub-series de consumos asociadas a clientes libres o equipos de medición que violen alguna de las reglas de negocio establecidas por la contraparte experta.

Dicha data se encuentra almacenada en una base transaccional que registra las tele-medidas originadas por el sistema SCADA de medición denominada ION, las cuales son respaldadas en una serie de bases estáticas denominadas bases SNAPS.

Para obtener la data relevante se generan diversas consultas SQL con el fin de extraer la data referente a los transformadores que están asociados a los clientes regulados del mercado. Las consultas asociadas a los transformadores pueden verse en el *apéndice 9.3*.

3.2 Pre-procesamiento

A-priori cada data set considera vectores de potencia media de (1x96) medidas diarias asociadas a data punto de medición, siguiendo la metodología propuesta por [8] y [16].

Con el objetivo de reducir la dimensionalidad de la data se agregaron estas medidas con frecuencia de ¼ hora, en medidas horarias, por punto de medición, de esta forma se reduce la dimensionalidad del vector de observaciones diario desde 1x96 hasta un nuevo vector de 1x24.

Con el objetivo de hacer compatible la visualización de las series de tiempo asociadas al consumo observado en cada equipo de medición con la visualización generada del cómputo de los componentes independientes se genera una serie de vectores agregados mensuales de 1x24 dimensiones los cuales corresponden al perfil horario acumulado mensual mediante la siguiente función de visualización:

$PCHM_m^8 = \sum_{h=1}^M C_h \forall h \in \{1, \dots, 24\}$ con $M = |m|$ y $m \in \{1, \dots, 12\}$ y C_h el consumo registrado en la hora h.

Esta visualización permite “suavizar” eventos puntuales de traspasos de carga entre transformadores al momento de generar la visualización de los resultados.

Posteriormente con el objetivo de capturar patrones estacionales y tendenciales de consumo se genera una nueva visualización de la data utilizando como función de agregación:

$$CTD_t^9 = \sum_{i=1}^{24} C_i \forall t \in \{1, \dots, 365\}.$$

Esta visualización de la serie de consumos observados permite generar nuevas funciones de agregación como las siguientes:

$$1) PCHS_s^{10} = \sum_{h=1}^S C_h \forall h \in \{1, \dots, 24\} \text{ y } S = |s| \text{ y } s \in \{1, \dots, 52\}$$

$$2) CTS_t^{11} = \sum_{i=1}^7 CTD_i \forall s \in \{1, \dots, 52\}$$

Entre otras visualizaciones dependiendo de las distintas ventanas temporales seleccionadas.

3.3 Selección de equipos de medición

Tal como se menciona en la descripción del problema, de los 183 equipos de medición disponibles a nivel de transformadores de distribución actualmente instalados en el anillo de concesión el 37,1% de estos capturan medidas asociadas a clientes que no son del interés de este estudio.

De este sub-conjunto de equipos de medición 50 de ellos corresponden a transformadores de distribución asociados a clientes libres, es decir el 27,9% y 17 de ellos corresponden a equipos de medición asociados puntos de interconexión entre la compañía y la zona concesionaria aledaña administrada por CGE¹² y a equipos de medición fuera de operación.

⁸ Perfil de Consumo Horario Mensual

⁹ Consumo Total Diario

¹⁰ Perfil de Consumo Horario Semanal.

¹¹ Consumo Total Semanal.

¹² Compañía General de Electricidad.

Luego del total de equipos de medición se trabajara con un sub-conjunto de 115 transformadores de distribución, quienes representan las medidas del consumo asociadas al mercado de clientes regulados.

Por otro lado a nivel de subestación, el data set compuesto por 46 subestaciones considera en su totalidad a todo el mercado regulado del anillo de concesión sin considerar puntos de interconexión y consumos asociados a clientes libres, por ende se trabajara con el data set asociado al total de subestaciones presente en este.

3.4 Transformación de la data

Tal como se menciona en el marco conceptual del presente informe la técnica a utilizar para resolver el problema de variables latentes asociadas en última instancia al consumo individual de cada cliente requiere de la aplicación de dos transformaciones de la data:

- Centering: Consiste en centrar el vector \mathbf{x} restándole su media con el fin de transformarlo en un vector con media cero.
- Whitening: Consiste en transformar el vector \mathbf{x} linealmente con el fin de obtener un nuevo vector de observaciones cuyas componentes no están correlacionadas y poseen varianzas igual a la unidad.

3.5 Construcción del modelo

Una vez obtenido el data set pre-procesado, es decir centrado y blanqueado de acuerdo las transformaciones anteriores, se puede comenzar con la construcción del modelo mismo.

Para ello se utilizó la herramienta fastICA desarrollada [10] para R, además se utilizaron los siguientes paquetes adicionales: *Psych*, *Zoo*, *Signal* y *ggplot2 PerformanceAnalytics* del repositorio CRAN de R, con el objetivo de manejar los vectores de medición como series de tiempo, calcular indicadores de la performance y estadísticas descriptivas de las series.

Con esto se utilizaron los siguientes parámetros para el algoritmo fastICA:

- Fun= *logcosh*(\mathbf{u}).
- $\alpha = 1$ donde $\alpha \in [1, 2]$ en el caso que la función utilizada sea *logcosh*(\mathbf{u}).
- k , corresponde a la cantidad de componentes. En este trabajo se realizaron iteraciones del algoritmo con $k \leq 114$ la cual se asocia al peor escenario posible pues en el caso $k = 114$ se tiene aproximadamente un componente independiente por cada equipo de medición.

Para fijar la cantidad de k utilizada en el modelo final se utilizó el criterio de la varianza explicada, fijando el nivel de varianza en un 95% inicialmente, es decir se computaran los k componentes independientes tales que expliquen el 95% de la varianza asociada al modelo.

- Alg.type = “deflation”, es decir que el método de solución del algoritmo fuera deflacionario extrayendo un componente independiente a la vez, esto con el objetivo de asegurar la estabilidad de los componentes extraídos entre cada nuevo Experimento.
- Row.norm= FALSE, es decir debido a que ya se habían aplicado las transformaciones que son condiciones necesarias para el funcionamiento del algoritmo (centering y whitening) no se estandarizaron las filas de la matriz de observaciones $\tilde{\mathbf{x}}$.
- Maxit=100, es decir se impuso un máximo de 100 iteraciones antes de entregar resultados independiente de la convergencia del algoritmo.
- $\epsilon = 0,001$, es decir el valor para el cual el algoritmo determina si la matriz de “desmezcla” \mathbf{w} ha convergido.

Una vez aplicado el algoritmo fastICA bajo estos parámetros se computan los valores de la matriz de mezcla \mathbf{A} y de la matriz de variables no observables \mathbf{s} .

Finalmente para facilitar la reconstrucción y validación de los perfiles encontrados se separan ambas matrices en sus columnas.

Si suponemos que $\dim(\mathbf{A}) = n \times m$ y que $\dim(\mathbf{s}) = k \times m$ con k definido anteriormente como parámetro del algoritmo fastICA entonces computamos \mathbf{s}_j con $j \in \{1, \dots, k\}$ vector columna y \mathbf{a}_l con $l \in \{1, \dots, m\}$.

En el *apéndice 9.4* se observa el código utilizado en R para generar cada Experimento del algoritmo.

3.6 Validación

Una vez realizadas las diferentes iteraciones del modelo ajustando los parámetros mencionados anteriormente deben validarse los resultados obtenidos para ello se utilizan dos principales criterios:

- *Criterio de la contraparte experta*
Este criterio consiste básicamente en que los perfiles obtenidos anteriormente tengan sentido desde un punto de vista de las lógicas del negocio.

De acuerdo a las reuniones con la contraparte estas reglas de decisión para validar un set de componentes independientes obtenidos tienen que cumplir que las siguientes condiciones:

1. Sea $\tilde{\mathbf{x}} = \tilde{\mathbf{A}}\mathbf{s}$ entonces un set de componentes independientes obtenidos es validado por la contraparte si $\tilde{\mathbf{x}} > \mathbf{0}$, donde $\tilde{\mathbf{x}}$ corresponde a la reconstrucción de la matriz de observaciones \mathbf{x} .

2. Cada s_j obtenido debe capturar las principales variaciones asociadas al consumo dependiendo de la ventana temporal escogida.
 3. Al menos en el set de componentes independientes obtenidos debe aparecer una de las tres curvas teóricas de consumo conocidas por la contraparte experta. En los anexos figura 9 estas corresponden al transformador 1 o perfil horario típicamente industrial, al transformador 4 o perfil horario típicamente residencial y transformador 18 o perfil horario típicamente comercial.
- MAPE, RMSE y otros tipos de errores
- Este criterio consiste en comparar las curvas de consumo $\tilde{x} = \tilde{A}s$ reconstruidas con las señales observadas originalmente. Actualmente se han implementado dos tipos de errores:
5. MAPE: $M = \frac{1}{n} \sum_{t=1}^n \frac{x_t - \tilde{x}_t}{x_t}$
 6. RMSE: $RMSE = \sqrt{\frac{\sum_{t=1}^n (x_t - \tilde{x}_t)^2}{n}}$
 7. N-RMSE: $N - RMSE = \frac{RMSE}{y_{max} - y_{min}}$
 8. MAE: $MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|$

Dichas funciones fueron implementadas en un script de R junto al resto de los códigos necesarios para aplicar las distintas iteraciones de Análisis de Componentes Independientes y contrastar los resultados obtenidos, tal como se observa en el *apéndice 9.4*.

Capítulo 4: Análisis Descriptivo

4.1 Observaciones Data Set Transformadores

Tal como se menciona en la descripción del proyecto, el data set corresponde a distintas series de tiempo asociadas al consumo observado del mercado regulado en el anillo de concesión de la compañía.

4.1.1 Descripción Curva Sistema

En la *figura 10* se observa la serie de tiempo asociado al consumo total de energía del sistema a lo largo del período de estudio, considerando el perfil horario acumulado mensual.

En los anexos *figura 11* se puede observar la serie de tiempo asociada a al consumo total del período de estudio considerando una frecuencia de observación horaria.

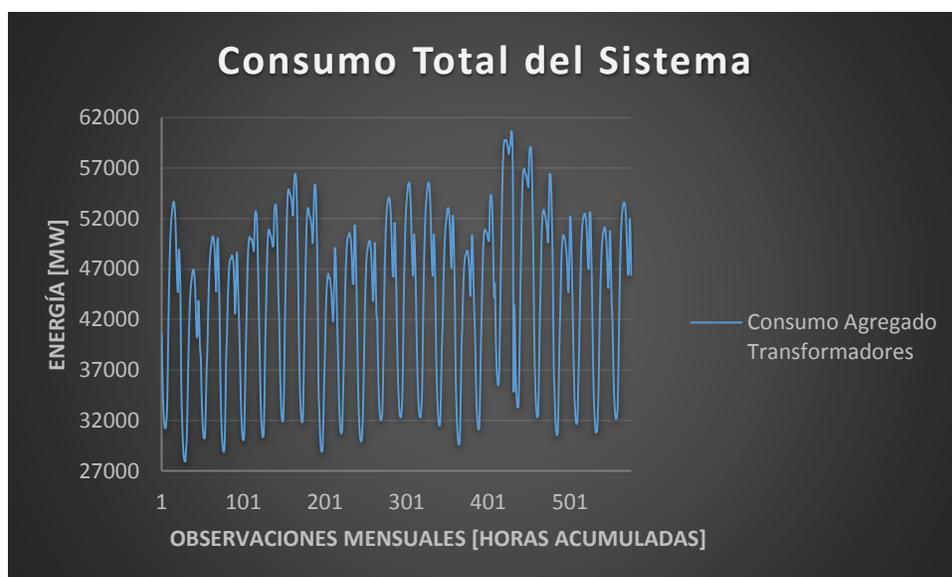


Figura 10: Consumo energético observado del sistema desde el 1-01-2013 hasta el 31-12-2014.

A continuación en la *tabla 1* se observa los principales estadísticos de resumen de la serie asociada al consumo total, la cual tiene un consumo medio 46.920[MW] y alcanza sus máximos niveles de consumo entre los meses {5, 6, 7, 8} y {16, 17, 18, 19}, respectivamente, es decir entre los meses Junio a Agosto para los dos años que dura el período de estudio.

<i>Min.</i>	<i>1st Qu.</i>	<i>Median</i>	<i>Mean</i>	<i>3rd Qu.</i>	<i>Max</i>
27930	36370	46290	44300	50590	60650

Tabla 1: Estadísticos descriptivos para la curva total del sistema.

En la *figura 12* se observa la descomposición de la serie asociada al consumo total del anillo observando la estacionalidad, tendencia y componente aleatoria de la serie.

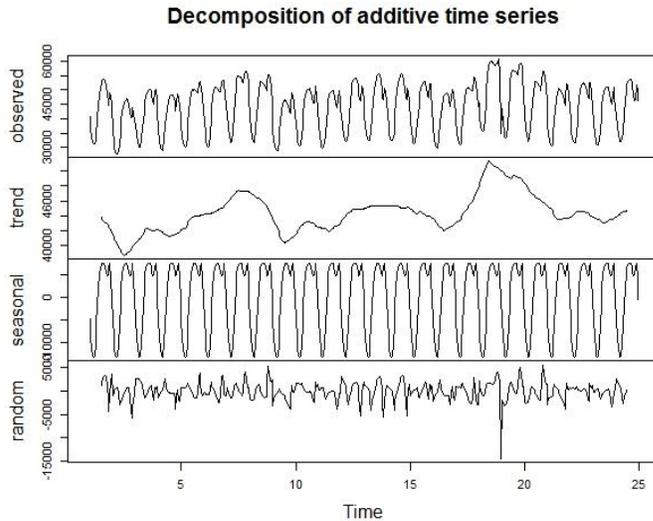


Figura 12: Descomposición de la serie observada de consumo total del sistema en sus componentes estacionales, tendenciales y aleatoria.

Tal como se menciona anteriormente el consumo total del sistema presenta una temporada de alta demanda, se observa en los meses entre junio y septiembre del período de estudio, esta observación es consistente con la política de precios de alta demanda de la compañía la cual determina arbitrariamente que el período de alta demanda transcurre desde abril hasta septiembre de cada año. Por otro lado en términos de la estacionalidad y dada la frecuencia utilizada en la representación de la serie y la visualización de la data en términos de perfiles acumulados de 24 horas se observa el típico patrón de consumo asociado al sistema en un horizonte de 24 horas.

Para facilitar el análisis descriptivo y la descripción de los data sets se separaron el total de 115 transformadores de distribución en tres grupos y las 46 subestaciones en dos grupos.

4.1.2 Transformadores de respaldo

Este primer grupo de transformadores operacionalmente cumplen el rol de respaldar al resto de los transformadores en caso de que estén dejen de estar en operación. Un transformador puede dejar de estar en régimen¹³ debido a diferentes razones, principalmente debido a fallas en dicho transformador o a operaciones de mantenimiento programadas, una tercera razón posible consiste en que la topología de la red cambie debido a transferencias de carga para mantener el balance de carga en caso de fallas en más de un transformador en otro punto de la red.

En la figura 13 se observa la distribución de los consumos registrados en los distintos transformadores de respaldo, como es de esperarse la gran mayoría de estos tiene un consumo registrado nulo durante el período de estudio sin embargo existe un subconjunto de estos que registran medidas de consumo energético positivo debido a que entraron en operación durante la ventana temporal considerada en el período de estudio.

¹³ En operación continua a lo largo de una ventana temporal t.

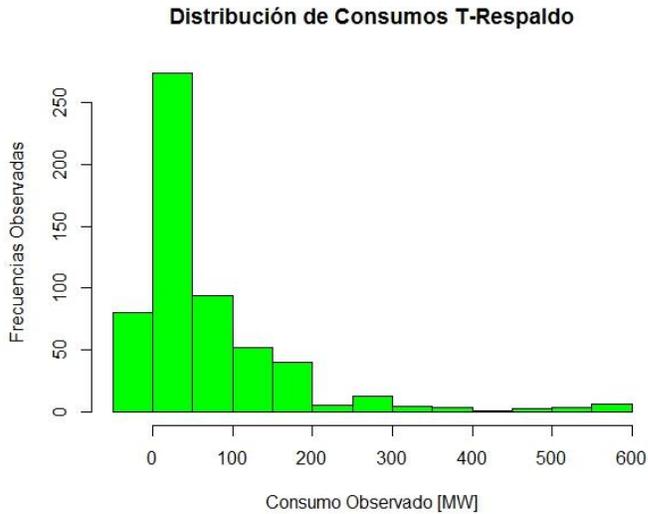


Figura 13: Histograma de los consumos observados en mega-watts de los transformadores de respaldo.

En los anexos *figura 14* se puede observar la serie de tiempo asociada al consumo total acumulado del subconjunto de transformadores de respaldo, al igual que en el gráfico anterior se observa visualmente la naturaleza de las medidas, las cuales debido a la función que cumplen los transformadores registran medidas de consumo en períodos muy acotados de tiempo.

Además en la *figura 15* se puede observar la descomposición de la serie de consumo agregado asociada a los transformadores de respaldo en términos de sus componentes estacionales, tendenciales y aleatorias, como es de esperarse en términos de la tendencia la serie captura solo eventos operacionales anómalos, en los cuales por razones técnicas este sub-conjunto de transformadores entran en régimen.

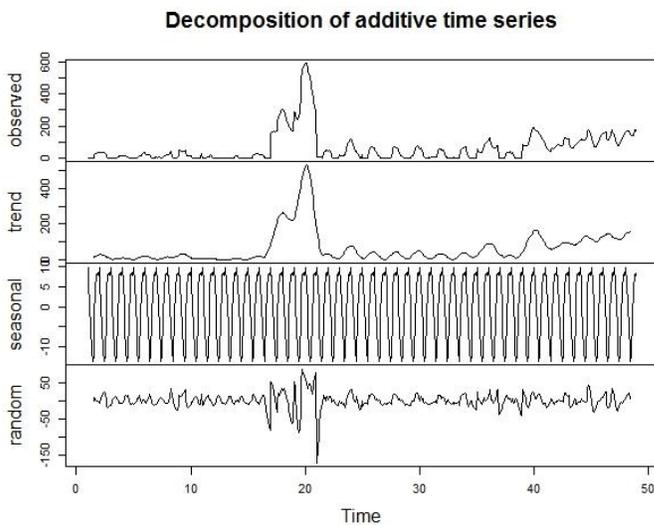


Figura 15: Descomposición de la serie de consumo asociada a los transformadores de respaldo.

4.1.3 Transformadores menor consumo observado

Un segundo grupo de transformadores son aquellos que operacionalmente se encuentran en régimen, sin embargo representan el menor nivel de consumo observado, es decir corresponden a aquellos equipos de medición que debido a variables externas presentan los menores niveles de consumo a pesar de encontrarse activos operacionalmente.

En la *figura 16* se observa la distribución de los consumos registrados en este tipo de transformadores, considerando los 20 transformadores que durante el período de estudio presentan un menor nivel de consumo acumulado.

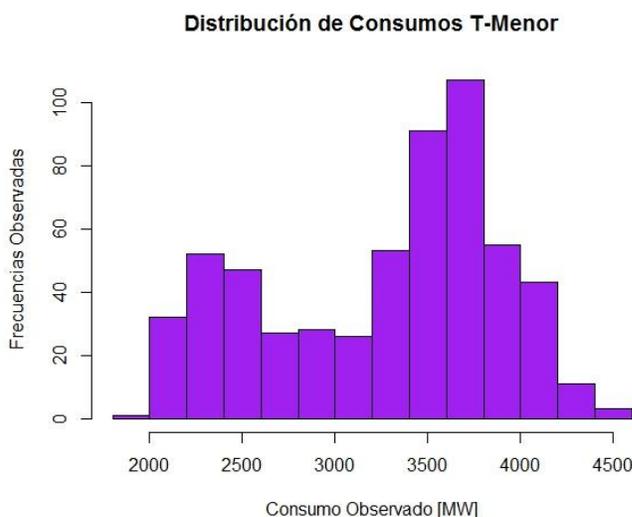


Figura 16: Histograma de los consumos observados en mega-watts de los 20 transformadores con menor consumo acumulado en el período de estudio.

En los anexos *figura 17* se puede observar la serie de tiempo asociada al consumo total acumulado del subconjunto de 20 transformadores con el menor nivel de consumo observado. En dicha visualización se observa como a pesar de tener un bajo nivel de consumo esta muestra de transformadores replica la curva de consumo total asociada al anillo de concesión.

Además en la *figura 18* se puede observar la descomposición en términos de los componentes estacionales, tendenciales y aleatorios de la serie de consumo agregado asociada a los 20 transformadores con un menor nivel de consumo.

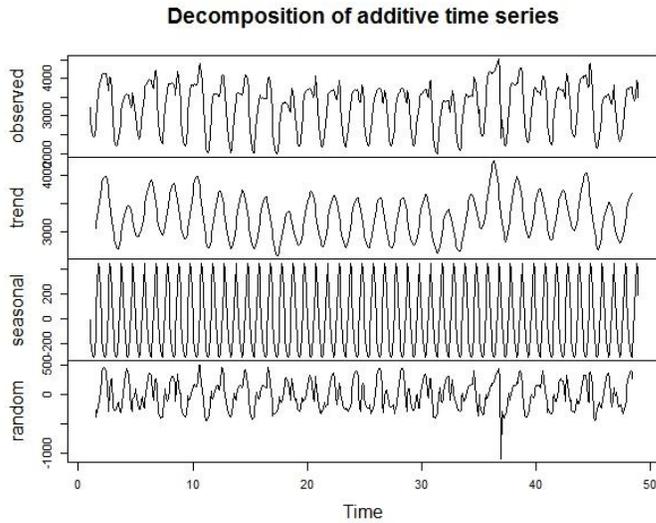


Figura 18: Descomposición de la serie de consumo asociada a los transformadores con un menor consumo observado.

4.1.4 Transformadores mayor consumo observado

Un tercer grupo de transformadores son aquellos que operacionalmente se encuentran activos y presentan el mayor consumo observado durante el período de estudio.

En la *figura 19* se observa la distribución de los consumos registrados para este tipo de transformadores, considerando los 20 transformadores que durante el período de estudio presentan un mayor nivel de consumo acumulado.

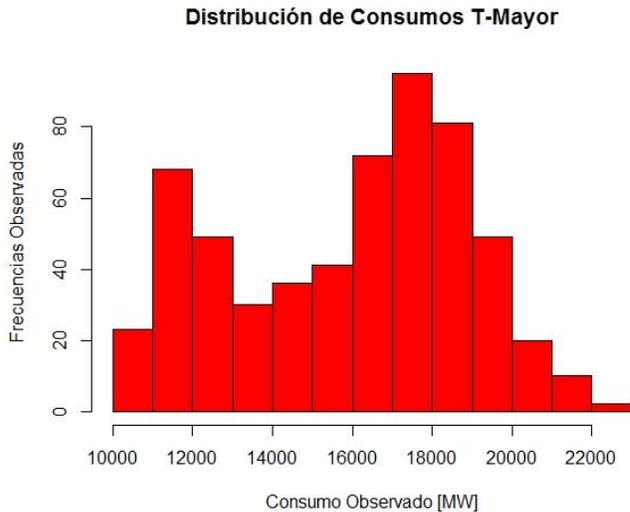


Figura 19: Histograma de los consumos observados en mega-watts de los 20 transformadores con mayor consumo acumulado en el período de estudio.

En los anexos *figura 20* se puede observar la serie de tiempo asociada al consumo total acumulado del subconjunto de 20 transformadores con el mayor nivel de consumo observado.

Al igual que en el caso anterior a pesar de tener un alto nivel de consumo esta muestra de transformadores presenta un comportamiento muy similar a la curva de consumo total asociada al anillo de concesión.

Además en la *figura 21* se puede observar la descomposición en términos de los componentes estacionales, tendenciales y aleatorios de la serie de consumo agregado asociada a los transformadores con un mayor nivel de consumo observado a lo largo del período de estudio.

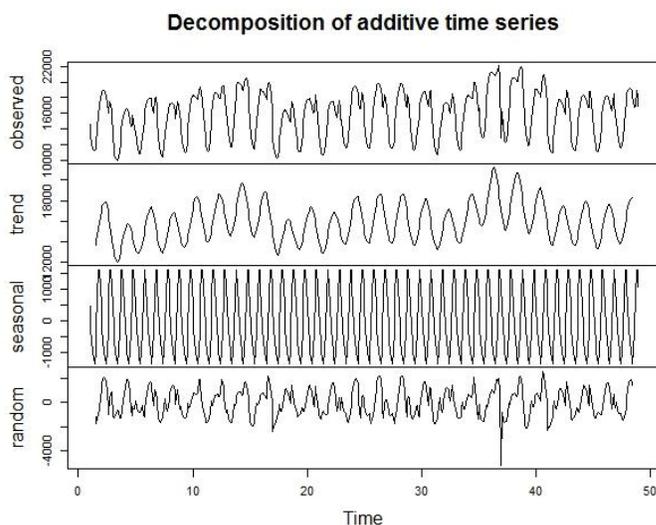


Figura 21: Descomposición de la serie de consumo asociada a los transformadores con un menor consumo observado.

4.1.5 Test de normalidad y comparación de los sub-conjuntos

En la *figura 22* se pueden observar tres series de consumo total acumulado las cuales corresponden a la curva del sistema, y las dos curvas de consumo total acumulado asociadas a los 20 transformadores con un menor y mayor consumo total observado.

En términos de composición el subconjunto de transformadores con mayor consumo observado representan en promedio el 36% del consumo total observado del sistema, mientras que el subconjunto de transformadores con el menor consumo observado representan en promedio el 7,4% del consumo total observado del sistema.

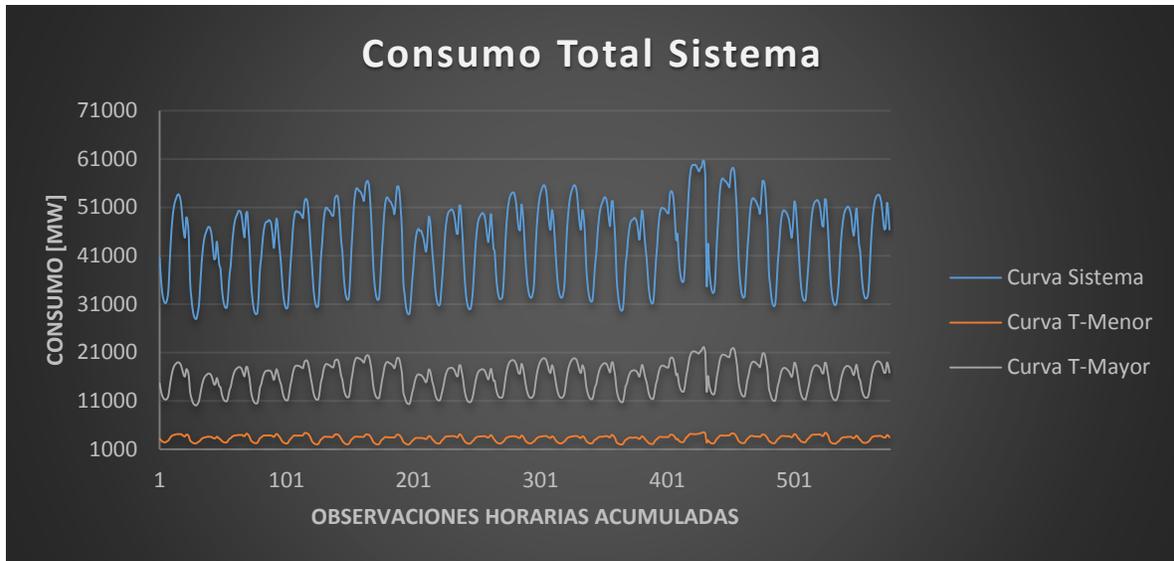


Figura 22: Series de tiempo del consumo total del sistema y de los subconjuntos de transformadores con menor y mayor consumo observado.

En la *tabla 2* se observan los resultados de aplicar el test Anscombe-Glynn [17] con el objetivo de testear si la kurtosis de la muestra corresponde a la kurtosis esperada para una distribución gaussiana.

En la versión canónica del test se tiene que la hipótesis nula corresponde a que la kurtosis de la muestra es igual a 3, mientras que la hipótesis alternativa corresponde a que la kurtosis de la muestra es estadísticamente distinta de 3.

Dado que la kurtosis de una distribución Gaussiana es exactamente igual a 3 este test permite chequear ex ante que se cumple uno de los supuestos básicos del Análisis de Componente Independientes, el cual tal como se menciona en el marco conceptual establece que las distribuciones tienen que ser lo más no gaussianas posible.

	Kurtosis	p-valor
Transformadores de respaldo	12,97	2.2e-16
Less k=20 Transformadores	2,04	2.2e-16
Top k=20 Transformadores	1,96	2.2e-16
Curva Sistema	1,92	2.2e-16

Tabla 2: Principales resultados test de Anscombe-Glynn para los distintos grupos de transformadores.

Tal como se ve en la tabla anterior, el p-valor para los tres grupos de transformadores y para la curva del sistema es $\ll 0.05$, que corresponde al nivel de significancia al cual se aplicó el test, por ende se rechaza la hipótesis nula en los cuatro casos y se acepta la hipótesis alternativa, es decir que la kurtosis es estadísticamente distinta que 3.

4.2 Observaciones Data Set Subestaciones

4.2.1 Subestaciones menor consumo observado

Al igual que en el data set compuesto por las mediciones registradas a nivel de transformadores en este sub conjunto de subestaciones se encuentran las 10 subestaciones que presentan el menor nivel de consumo durante el período de estudio, estas corresponden a las subestaciones de Alto Jahuel, Buin CGE, Chicureo, Lo Prado 044, Metro 110-2, Entel, Caleu, Rungue, Lo Prado 012 y Panamericana.

En la *figura 23* se observa la distribución de los consumos observados para estas subestaciones, la cual sigue una distribución distinta a aquella observada en los transformadores con el menor nivel de consumo, esto es debido a que dicha distribución viene dada por mediciones realizadas en equipos aguas arriba en comparación con las mediciones realizadas a nivel de transformador y no necesariamente agrupan las mediciones obtenidas con los transformadores con menor nivel de consumo.

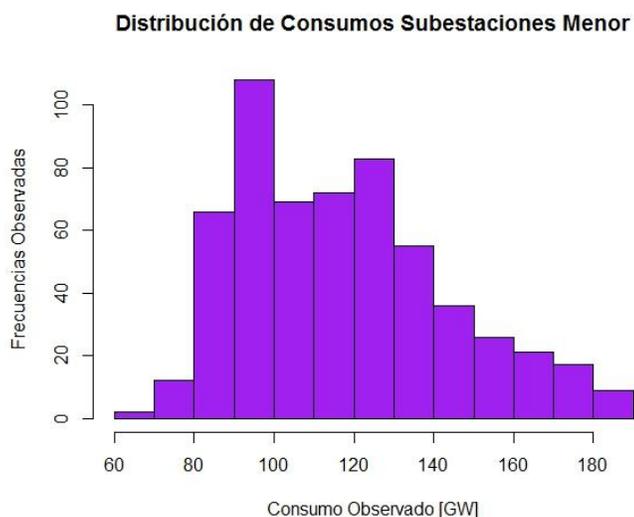


Figura 23: Histograma de los consumos observados en giga-watts de las 10 subestaciones con menor nivel de consumo.

En los anexos *figura 24* se observa la serie de consumo acumulado total para el subconjunto de subestaciones que presentan un menor nivel de consumo.

Además en la *figura 25* se observa la descomposición en términos de sus componentes tendenciales, estacionales y aleatorios de dicha serie de consumo.

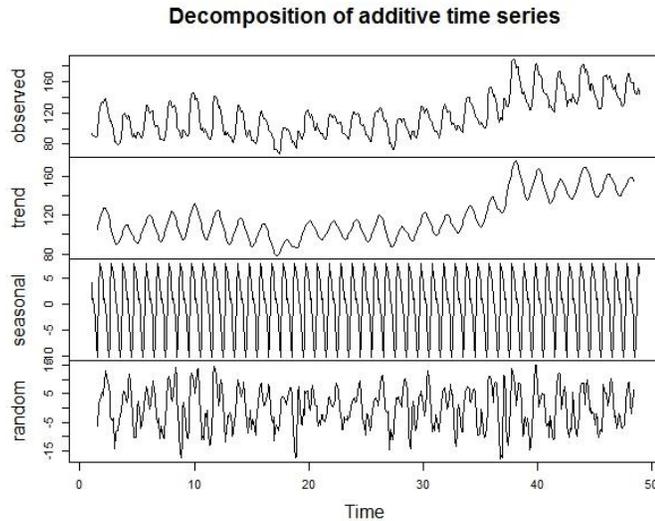


Figura 25: Descomposición de la serie de consumo a las subestaciones con menor consumo observado.

Se observan, tanto en la distribución de los consumos observados como en la serie temporal y en la descomposición de la misma diferencias respecto al subconjunto de transformadores de menor consumo, tal como se mencionaba anteriormente esto puede explicarse debido a que no necesariamente las subestaciones de menor consumo agrupan al subconjunto de transformadores de menor consumo y muchas de estas subestaciones tienen una composición heterogénea en términos de clientes, por ende la componente regulada de su consumo observado es significativamente baja, dicho lo anterior para efectos de este estudio se consideró solamente la componente regulada de su consumo observado y por lo mismo presenta un comportamiento diferente al observado en el grupo de transformadores de menor consumo observado.

4.2.2 Subestaciones mayor consumo observado

Siguiendo con lo mencionado anteriormente, se definió un subconjunto con las 10 subestaciones que presentan un mayor nivel de consumo durante el período de estudio, estas corresponden a las subestaciones Lo Boza, San Joaquín, Chacabuco, Apoquindo, Alonso de Córdova, Santa Elena, Pajaritos, San Cristóbal y Vitacura.

En la *figura 26* se observa la distribución de los consumos observados a nivel de subestaciones, estos a diferencia del subconjunto de subestaciones de menor consumo presentan una distribución bastante similar al subconjunto de transformadores con un menor consumo observado.

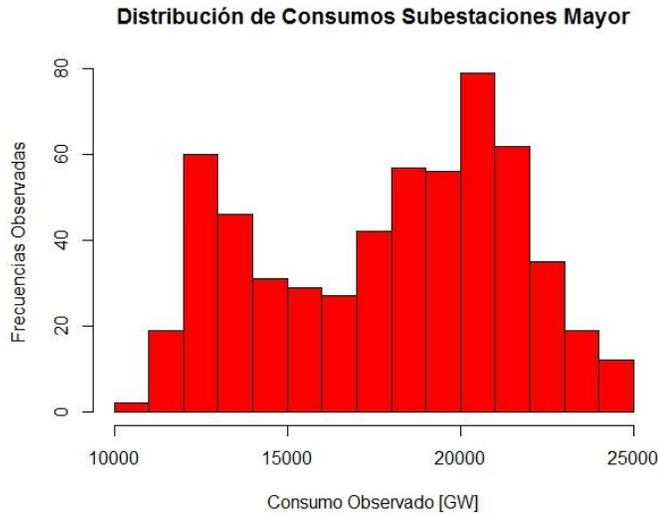


Figura 26: Histograma de los consumos observados en giga-watts de las 10 subestaciones con mayor nivel de consumo.

En los anexos *figura 27* se observa la serie de consumo acumulado total para el subconjunto de subestaciones que presentan un mayor nivel de consumo, además en la *figura 28* se observa la descomposición en términos de sus componentes tendenciales, estacionales y aleatorias de dicha serie de consumo.

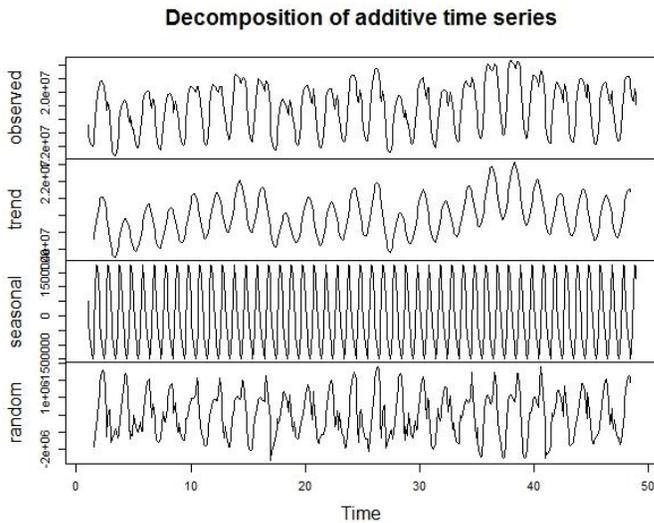


Figura 28: Descomposición de la serie de consumo a las subestaciones con mayor consumo observado.

A diferencia de las subestaciones con un menor nivel de consumo se observa tanto en la distribución de los consumos observados, en la serie de consumo acumulado y en la descomposición de la misma un comportamiento mucho más consistente con las observaciones asociadas al subconjunto de transformadores con un mayor nivel de consumo, esto puede explicarse a que en términos de la composición de la serie observada este subconjunto de subestaciones agrupa casi exclusivamente clientes regulados y dado que a nivel de transformador la gran mayoría de las observaciones corresponden a este tipo de clientes es esperable que su comportamiento sea similar a estos.

4.2.3 Test de normalidad y comparación de los sub-conjuntos

En la *tabla 3* se observan los resultados de aplicar el test Anscombe-Glynn a estos dos grupos de sub-estaciones, con el objetivo de testear que la kurtosis sea estadísticamente distinta a 3, y con ello validar la hipótesis de no-normalidad de las señales observadas.

	Kurtosis	p-valor
SubEst Menor Consumo	2,67	0,06
SubEst Mayor Consumo	1,87	2.2e-16

Tabla 3: Principales resultados test de Anscombe-Glynn para los distintos grupos de subestaciones.

Se puede observar que a diferencia del data set compuesto por observaciones registradas a nivel de transformadores, las subestaciones de menor consumo observado poseen una kurtosis cercana a 3 sin embargo es estadísticamente distinta con un nivel de significancia del 0.1, esto es debido a la naturaleza de las observaciones en dichos equipos de medición la cual se mencionó anteriormente, mientras que para las subestaciones con un mayor nivel de consumo observado la kurtosis es estadísticamente distinta de 3 con un nivel de significancia del 0.05.

Capítulo 5: Resultados y Análisis

5.1 Transformaciones Data Sets Transformadores y Subestaciones

Siguiendo lo propuesto en el marco conceptual se hace necesario aplicar las dos etapas de pre-procesamiento mencionadas anteriormente, es decir “centering” y “whitening”.

Inicialmente se aplicaron ambas transformaciones al data set consolidado de transformadores, en la *figura 29* se puede observar el consumo total calculado a partir del data set sin las transformaciones mencionadas anteriormente.

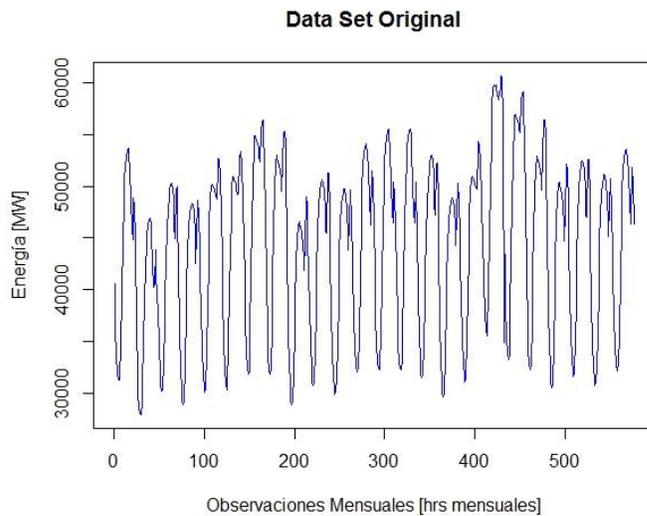


Figura 29: Visualización data set de transformadores sin transformar.

En la *figura 30* por otro lado se puede observar la misma serie de consumo total calculada a partir del data set ya transformado, claramente en este último se observa que las distintas instancias están de-correlacionadas y centradas en torno a su media.

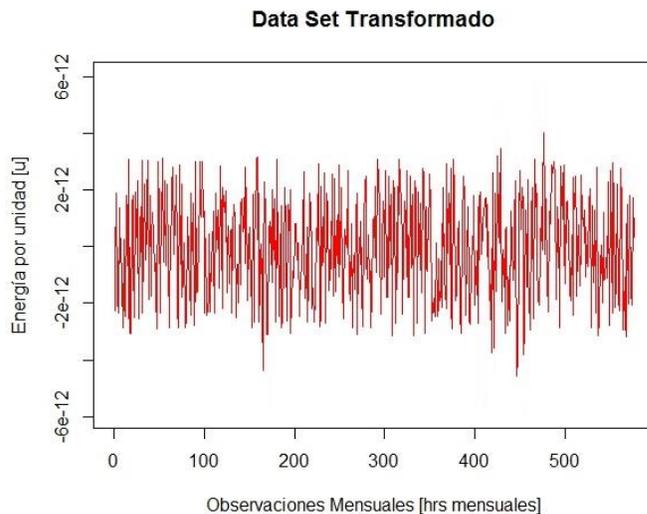


Figura 30: Visualización data set de transformadores transformado.

5.2 Descripción iteraciones algoritmo fastICA

Se realizaron 15 iteraciones del algoritmo fastICA variando el parámetro k asociado a la cantidad de componentes independientes utilizando $k \in \{2, \dots, 15\}$, los parámetros del algoritmo fastICA fueron los mismos para todas los experimentos, descritos en el marco conceptual.

Siguiendo lo propuesto por [18] y por [19] para cada experimento del algoritmo se computo la varianza explicada por el modelo utilizando la descomposición espectral de la matriz de covarianza de la matriz de observaciones \mathbf{x} asociada a cada Experimento es decir:

Sea $\mathbf{X} = \mathbf{Q}\mathbf{O}\mathbf{Q}^{-1}$ \mathbf{Q} matriz cuadrada de dimensiones $N \times N$, en este caso 115×115 , y donde cada q_i vector columna es un vector propio de la matriz \mathbf{X} , además \mathbf{O} es una matriz diagonal donde cada θ_i de la diagonal corresponde a un valor propio asociado a q_i , luego la varianza explicada por el modelo para cada Experimento viene dada por: $VAR_{explicada}(\mathbf{X}) = \sum_{i=1}^{dim(\mathbf{O})} \theta_i$

En la *figura 31* se puede observar la evolución de la varianza explicada por el modelo dependiendo de la cantidad de componentes independientes que se utilicen como inputs.



Figura 31: Varianza explicada vs. Cantidad de ICs estimados.

Del gráfico anterior se observa que no existen diferencias significativas en términos de la varianza explicada al considerar componentes independientes que pertenezcan al conjunto $\{15, 14, \dots, 6\}$ sin embargo hay variaciones importantes en términos de la varianza explicada al considerar 2, 3, 4 o 5 componentes independientes.

Dado que las transformaciones aplicadas al data set distorsionan la escala de los componentes encontrados y debido a las restricciones de interpretación en términos de la matriz de permutación mencionada en el marco conceptual, todos los componentes encontrados se encuentran normalizados mediante la siguiente transformación: $\hat{s}_i = \frac{s_i - \min(s)}{\max(s) - \min(s)}$

5.3 Resultados y Análisis a nivel de transformador

A continuación se describen los principales resultados obtenidos de aplicar el algoritmo fastICA para estimar el modelo ICA de variables latentes sobre el data set de medidas obtenidas a nivel de transformador.

5.3.1 Cálculo de errores de reconstrucción

Siguiendo lo propuesto por [18] dado que se tiene el siguiente modelo $\mathbf{x} = \sum_{i=1}^n \mathbf{a}_i s_i$ se puede calcular la reconstrucción de la matriz de señales observadas dada por $\tilde{\mathbf{x}} = \tilde{\mathbf{A}}\mathbf{s}$ de acuerdo a la siguiente ecuación:

$$\tilde{x}_i(t-j) = \sum_{k=1}^n a_{ik} s_k(t-j) \text{ con } j = 0, \dots, n-1$$

Donde $s_k(t-j)$ corresponde al valor estimado del k-esimo IC en el período t-j y a_{ik} corresponde al peso en la i-esima fila y k-esima columna de la matriz de mezcla \mathbf{A} .

En la *figura 32* se observa el error medio porcentual absoluto (MAPE) tal como fue definido en el marco conceptual, considerando el error entre la reconstrucción $\tilde{\mathbf{x}}$ y la matriz de señales original \mathbf{x} , para cada experimento del algoritmo por transformador.

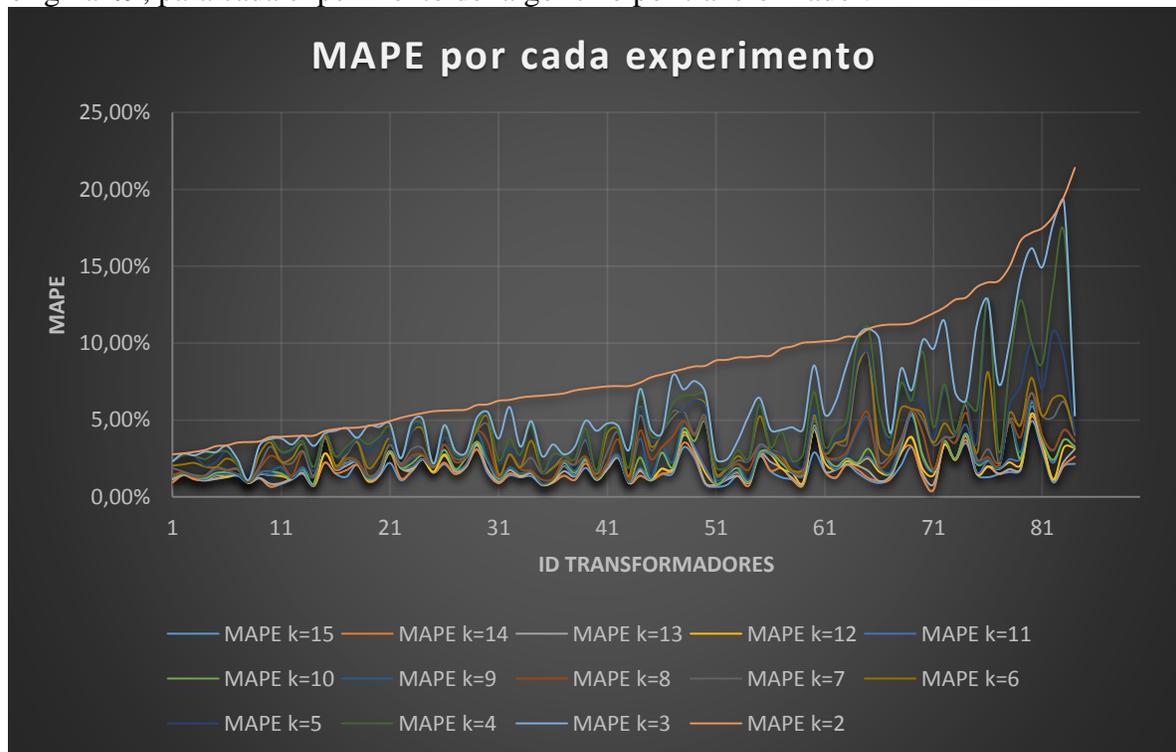


Figura 32.a: MAPE por cada escenario con cantidad de ICs distintos.

Se puede observar que consistentemente a medida que aumenta la cantidad de ICs que entran como input del algoritmo disminuye sistemáticamente el error de reconstrucción medido a través del MAPE.

En la figura 32.b se observa el error medio porcentual absoluto ponderado por el volumen total de energía asociada a dicho transformador (WMAPE) considerando el error entre la reconstrucción \tilde{x} y la matriz de señales original x , para cada experimento del algoritmo por transformador. Se observan los mismos comportamientos en términos de los errores y la cantidad de componentes que se observan al computar el MAPE tal como se observa en la figura 32.a.

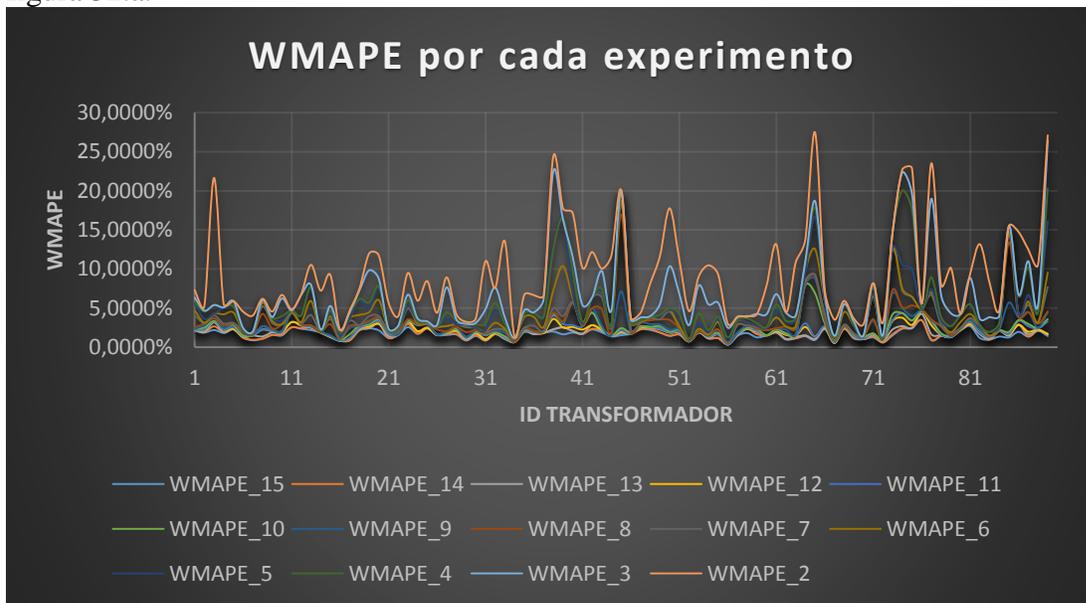


Figura 32.b: WMAPE por cada escenario con cantidad de ICs distintos.

Por otro lado en la *figura 33* se observa el error medio absoluto tal como se definió en el marco conceptual, nuevamente se observa que para un sub-conjunto de transformadores independiente de la cantidad de componentes con los cuales se cuenta los errores asociados a la reconstrucción de las series observadas es sistemáticamente alta.

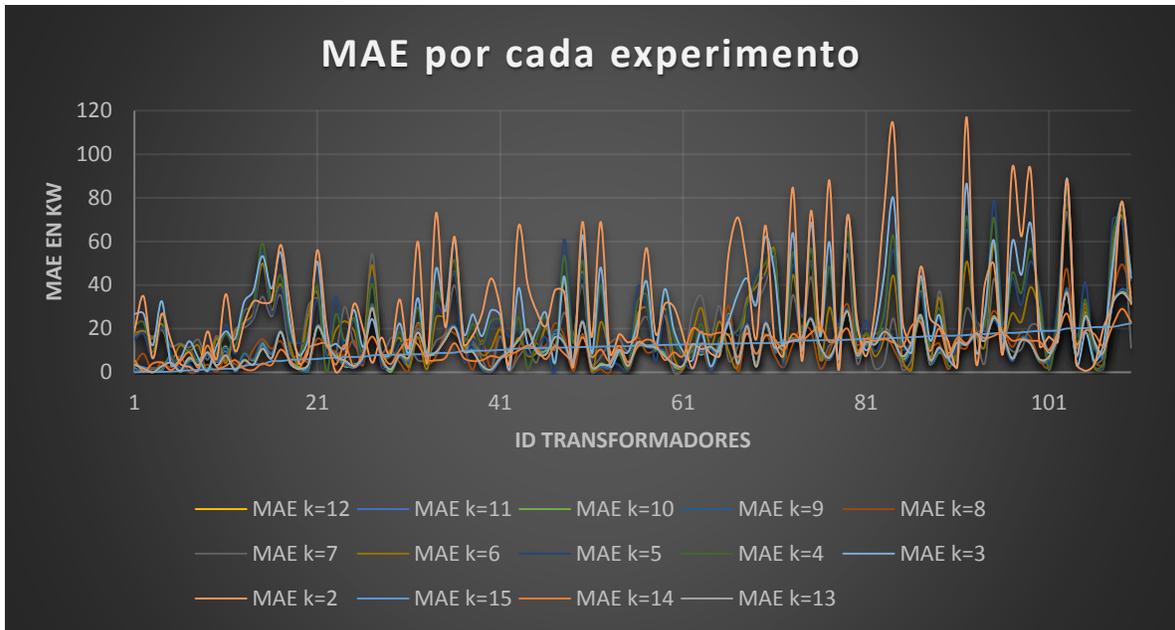


Figura 33: MAE por cada escenario con cantidad de ICs distintos.

En la *figura 34* se observa la raíz del error cuadrático medio normalizada para hacer comparables distintos tipos de transformadores, al igual que en los casos anteriores el error disminuye a medida que aumentan la cantidad de ICs, sin embargo existe un sub-conjunto de transformadores que consistentemente presentan altos errores en la reconstrucción de sus señales.

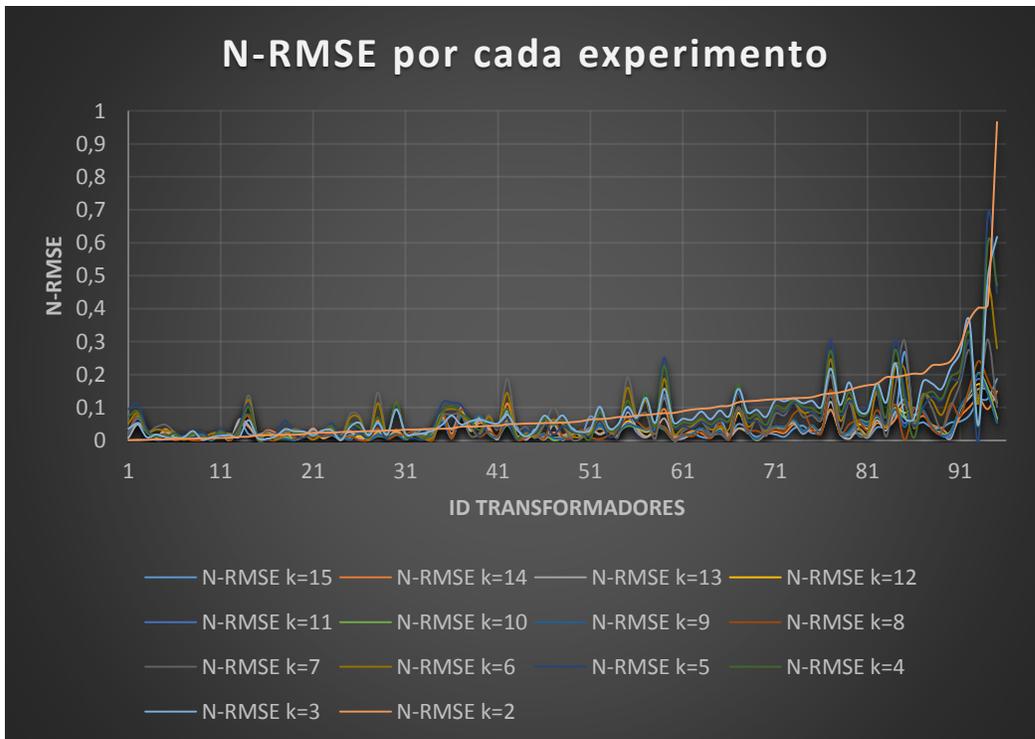


Figura 34: N-RMSE por cada escenario con cantidad de ICs distintos.

En términos de la reconstrucción de la curva de consumo total del sistema, de la *tabla 4* se puede observar como la tasa de los distintos tipos de errores disminuye consistentemente al aumentar la cantidad de componentes independientes, sin embargo no captura el efecto de que un sub-grupo de transformadores presenten problemas en términos de su reconstrucción debido principalmente a que la curva total del sistema suaviza estos errores pues se construye tomando la sumas de las curvas de consumo individuales por cada equipo de medición, es decir:

$$C_{sistema} = \sum_{i=1}^{115} C_i$$

IC	MAE	MAPE	N-RMSE	IC	MAE	MAPE	N-RMSE
15	34,4	0,08%	0,0015	8	88,19	0,21%	0,0039
14	38,8	0,09%	0,0017	7	88,11	0,21%	0,0039
13	48,8	0,11%	0,0021	6	246,36	0,57%	0,0111
12	51,56	0,12%	0,0022	5	286,18	0,67%	0,0124
11	51,69	0,12%	0,0022	4	286,94	0,67%	0,0125
10	52,31	0,12%	0,0023	3	404,82	0,94%	0,0174
9	52,56	0,12%	0,0023	2	418,63	0,97%	0,0178

Tabla 4: MAE, MAPE y N-RMSE de la curva total reconstruida para cada experimento con ICs distintos.

En la *figura 35* se puede observar la curva total del sistema contrastada contra tres reconstrucciones considerando las iteraciones con $k \in \{5, 10, 15\}$, tal como se muestra en la tabla anterior los errores en la reconstrucción son bastante pequeños debido a que por construcción la curva total del sistema tiende a suavizar estos errores, dado que suma sobre el total de las series de consumo por transformador.

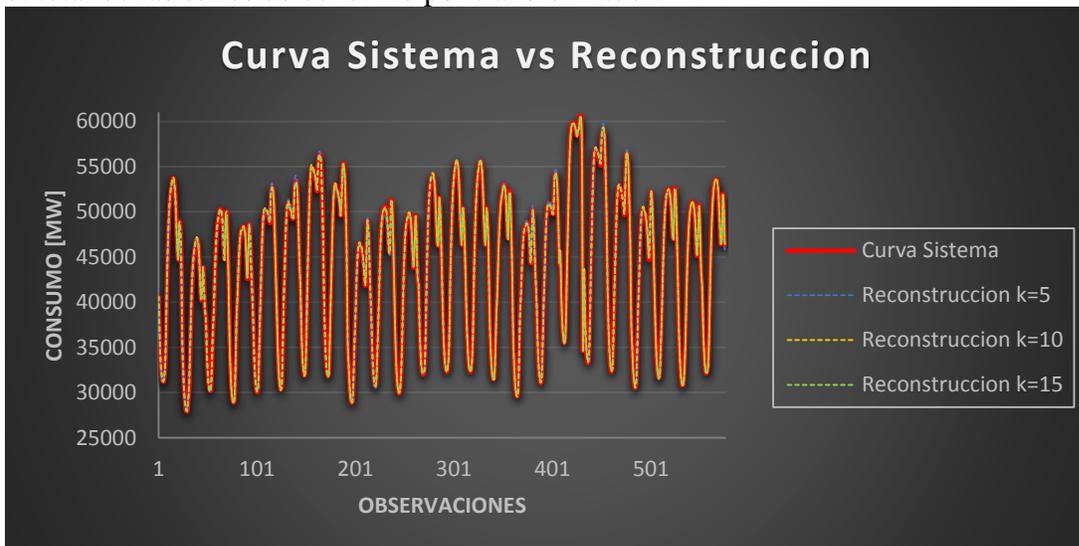


Figura 35: Curva sistema vs reconstrucción para $k \in \{5, 10, 15\}$

5.3.2 Principales experimentos

A continuación se muestran los resultados obtenidos los experimentos $k \in \{4, 5, 10, 15\}$.

5.3.2.1 Experimento $k=15$

Para este experimento se consideraron 15 componentes independientes. En los anexos *figuras 36, 37 y 38* se pueden observar las visualizaciones de los componentes independientes encontrados para este experimento.

Dado que la visualización más común para las curvas de carga consiste en generar perfiles de 24 horas y siguiendo lo mencionado en el marco conceptual se generaron perfiles horarios de 24 horas acumulados para los 15 componentes encontrados en este experimento.

En los anexos *figuras 39, 40 y 41* se pueden observar los perfiles de 24 horas estimados para los distintos componentes encontrados durante este experimento.

5.3.2.2 Experimento $k=10$

Para este experimento se consideraron 10 componentes independientes. En los anexos *figuras 42 y 43* se pueden observar las visualizaciones de los componentes encontrados para este experimento.

Además en los anexos *figuras 44 y 45* se pueden observar los perfiles de 24 horas estimados para los distintos componentes encontrados durante este experimento.

5.3.2.3 Experimento $k=5$

Para este experimento se consideraron 5 componentes independientes. En los anexos *figura 46* se pueden observar las visualizaciones de los componentes encontrados para este experimento.

Además en los anexos *figura 47* se pueden observar los perfiles de 24 horas estimados para los distintos componentes encontrados durante este experimento.

5.3.2.4 Experimento $k=4$

Para este experimento se consideraron 4 componentes independientes. En los anexos *figura 48* se pueden observar las visualizaciones de los componentes encontrados para esta Experimento.

Además en los anexos *figura 49* se pueden observar los perfiles de 24 horas estimados para los distintos componentes encontrados durante este experimento.

5.4 Principales segmentos encontrados

Una de las principales restricciones en términos de seleccionar una cantidad idónea de componentes independientes de tal forma de interpretar estos como segmentos distintos de clientes con patrones de consumo distintivos viene dada por la capacidad de los analistas de la compañía de interpretar los perfiles de carga obtenidos en términos de patrones de consumo descritos en la literatura y coherentes con las lógicas de negocio manejadas por la compañía.

Por ello del total de iteraciones realizadas se seleccionó la Experimento con $k=5$ componentes independientes, es decir con 5 segmentos distintivos de clientes que explican el comportamiento del mercado regulado.

Cada segmento fue nombrado de acuerdo a un patrón de consumo particular, los cuales corresponden a:

- Segmento “Industrial”
- Segmento “Residencial Anómalo”
- Segmento “Pequeños Negocios”
- Segmento “Residencial Bajo Consumo”
- Segmento “Residencial Alto Consumo”

En la *figura 50* se puede observar los perfiles de carga asociados a cada segmento encontrado.

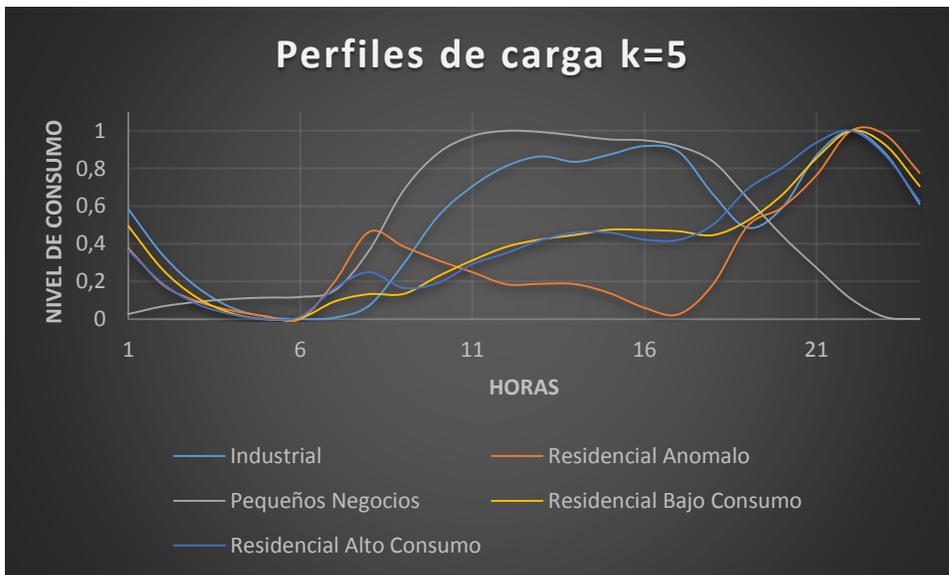


Figura 50: Perfiles de carga obtenidos para la Experimento $k = 5$

5.4.1 Caracterización segmentos encontrados

Tal como se menciona en secciones anteriores siguiendo lo propuesto [1] y por [20] se calcularon las siguientes variables para caracterizar los perfiles encontrados:

- Energía Máxima Consumida para una ventana temporal t :

$$\text{Max}_{\text{Consumo}_t} = \max\{E_{it} ; 1 \leq i \leq 5\}$$

- Energía Media Consumida para una ventana temporal t :

$$\text{Mean}_{\text{Consumo}_t} = \frac{1}{T} \sum_{t=1}^T E_{it} \text{ con } 1 \leq i \leq 5$$

- Factor de carga para una ventana temporal t :

$$E_{LF} = \frac{1}{m} \sum_{j=1}^m \frac{\frac{1}{n} \sum_{i=1}^n E_{ij}}{\max\{E_{ij}, 1 \leq i \leq n\}}$$

Dichas variables fueron calculadas en dos horizontes temporales, tanto a nivel mensual como a nivel semanal.

En los anexos *figuras 51, 52 y 53* se pueden observar la evolución de dichas variables a nivel mensual, se observa que tanto la Energía Máxima Consumida como la Energía Media Consumida en horizontes mensuales presentan un comportamiento bastante similar para todos los segmentos.

De la *figura 51* se observa como los segmentos residenciales de alto consumo y bajo consumo se encuentran altamente relacionados en ambas métricas y presentan un comportamiento bastante similar, por otro lado el segmento de pequeños negocios e industrial presentan un comportamiento independiente con medias de consumo que fluctúan y alcanzan sus máximos niveles en los primeros meses de los dos años del período de estudio.

Finalmente el segmento residencial anómalo presenta un comportamiento bastante similar a los dos residenciales pero durante el segundo año del período de estudio invierte su tendencia diferenciándose de ambos segmentos anteriores y bajando tanto sus niveles máximos como medios de consumo durante la segunda mitad del período de estudio.

Por otro lado el factor de carga mensual captura aquellos meses del período de estudio en los cuales se observa una mayor dispersión en los niveles de consumo observado, de la *figura 53* se observa como ambos segmentos residenciales presentan una alta dispersión durante la temporada de verano, es decir durante Octubre, Noviembre, Diciembre, Enero y Febrero de ambos años del período de estudio.

El segmento residencial anómalo por otro lado presenta pocas variaciones con excepción de los primeros meses del período de estudio. Finalmente y tal como es esperable los segmentos industriales y pequeños negocios presentan una alta variación en sus consumos durante la temporada de alta demanda.

A nivel de horizontes semanales se observa un mayor nivel de detalle en las distintas variables calculadas.

En la *figura 54* se observan como a nivel semanal los segmentos residenciales de bajo y alto consumo siguen el comportamiento descrito a nivel mensual, capturando con marcados peaks de consumo máximo durante las temporadas de alta demanda y durante las semanas 40-50 del segundo año del período de estudio, las cuales corresponden a los meses de Octubre-Noviembre-Diciembre, una posible hipótesis para explicar dicho comportamiento puede ser un alza en las temperaturas para dicha ventana temporal, lo cual debiese corroborado en futuros trabajos.

Por otro lado el segmento residencial anómalo presenta una disminución paulatina de sus máximos niveles de consumo a lo largo del período de estudio, mientras que los segmentos pequeños negocios e industrial replican el comportamiento observado con las variables mensuales.

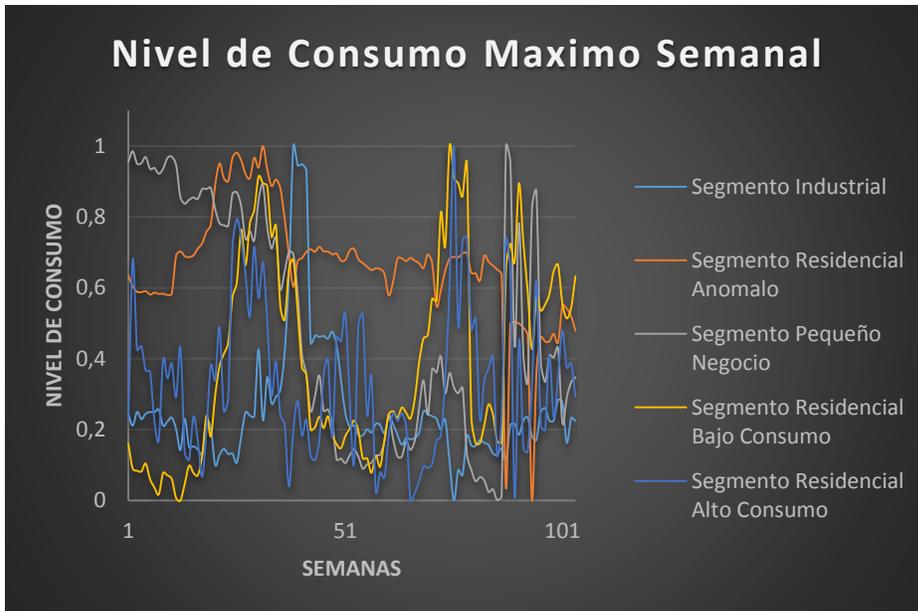


Figura 54: Máximo Consumo Semanal para los distintos segmento encontrados.

En la *figura 55* se observa el comportamiento del nivel de consumo medio semanal, al igual que para la variable anterior presentan marcados peaks y valles en las temporadas de alta y baja demanda respectivamente, el segmento residencial anómalo nuevamente muestra su tendencia a disminuir su nivel de consumo medio semanal mientras que los segmentos pequeños negocios e industrial replican lo observado a nivel mensual.

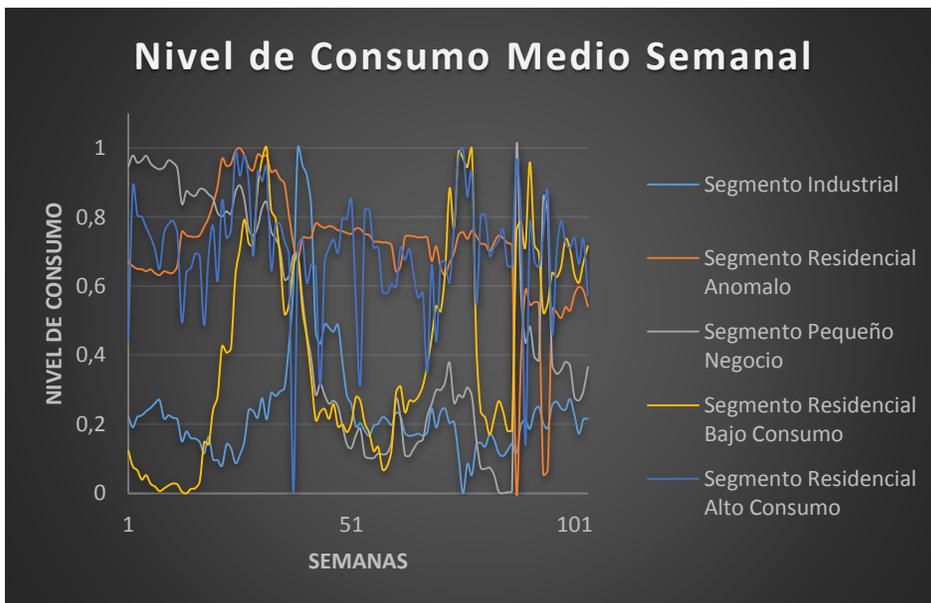


Figura 55: Consumo Semanal Medio para los distintos segmento encontrados.

Finalmente en términos de los factores de carga semanales se observan distintos niveles de dispersión para los distintos segmentos, se observan sin embargo diferencias de los comportamientos mensuales debido a que estas variables tienden a suavizar comportamientos particulares observables dentro de cada mes. Así por ejemplo el segmento industrial presenta

un bajo nivel de dispersión a lo largo del período de estudio y una gran diferencia entre sus consumos máximos y medios semanales lo cual se observa en la magnitud del factor de carga semanal¹⁴.

Por otro lado el segmento residencial de alto consumo presenta variaciones negativas durante las últimas semanas del período de estudio y variaciones positivas durante las semanas de alta demanda, el segmento residencial de bajo consumo en cambio presenta un rezago de un par de semanas manifestando el mismo comportamiento.

El segmento de pequeños negocios presenta un factor de carga relativamente estable durante la primera mitad del período de estudio, sin embargo presenta altas variaciones durante la segunda mitad del período de estudio, estas son positivas en el período de alta demanda y negativas en el período de baja demanda. Por otro lado el segmento residencial anómalo presenta altas variaciones en sus niveles de consumo al principio de la segunda mitad del período de estudio, todas ellas positivas.

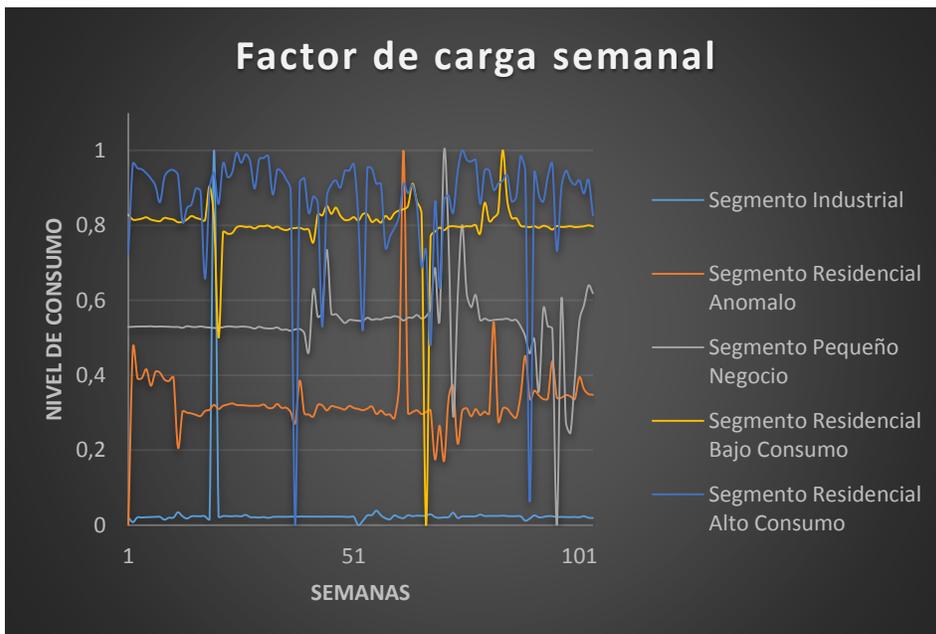


Figura 56: Factor de Carga Semanal.

5.4.2 Market-shares y tasas de crecimiento por transformador

Dado que se tiene el siguiente modelo para capturar el proceso de mezcla asociado a cada equipo de medición:

$x_i(t) = \sum_{j=1}^n a_{ij} s_j(t)$ siguiendo lo propuesto por [21] se puede interpretar cada peso a_{ij} asociado a la i -ésima fila y j -ésima columna de la matriz de mezcla como un ponderador que representa el market-share de cada segmento de cliente, identificado mediante un componente independiente, por equipo de medición.

¹⁴ Muy cercano a cero lo que implica por construcción de la variable ELF una alta diferencia entre los máximos observados y el consumo medio observado.

Luego aplicando una normalización min-max : $\hat{a}_{ij} = \frac{a_{ij} - \min(a_i)}{\max(a_i) - \min(a_i)}$ por columna para cada componente \hat{a}_{ij} de la matriz estimada \mathbf{w} y posteriormente escalando cada $\tilde{a}_{ij} = \frac{\hat{a}_{ij}}{\sum_{i=1}^N \hat{a}_{ij}} \forall j = 1, \dots, \dim(\mathbf{x})$ es decir considerando el peso relativo de cada componente \hat{a}_{ij} normalizado por cada equipo de medición, se obtienen los market-shares estimados para cada equipo de medición asociado al data set.

En los anexos de la *tabla 5* a la *tabla 14* se puede observar el market-share para cada segmento de clientes por equipo de medición, donde **Ind** corresponde al segmento industrial, **Res_A** al segmento residencial anómalo, **Peq_neg** al segmento de pequeños negocios, **Res_BC** al segmento residencial de bajo consumo y **Res_AC** al segmento residencial de alto consumo.

En la *figura 57* se puede observar la tasa de crecimiento asociada a cada segmento de clientes encontrado, se observa que distintos segmentos de clientes presentan distintas tasas de crecimiento para el período en estudio, respondiendo en forma diferente a variables exógenas que puedan afectar sus patrones de consumo observado.

Por otro lado en la *tabla 15* se puede observar las tasas de crecimiento estimadas promedio por segmento encontrado para el período de estudio.

Ind	2,0%
Res_A	7,9%
Peq_Neg	7,6%
Res_BC	5,5%
Res_AC	4,3%

Tabla 15: Tasas de crecimiento por segmento encontrado a nivel de transformador.

Dentro del período en estudio los dos segmentos que manifestaron mayores de tasas de crecimiento corresponden a los segmentos de pequeños negocios y residencial anómalo, mientras que los dos segmentos que experimentaron menores tasas de consumo corresponden al residencial de alto consumo y al segmento industrial.

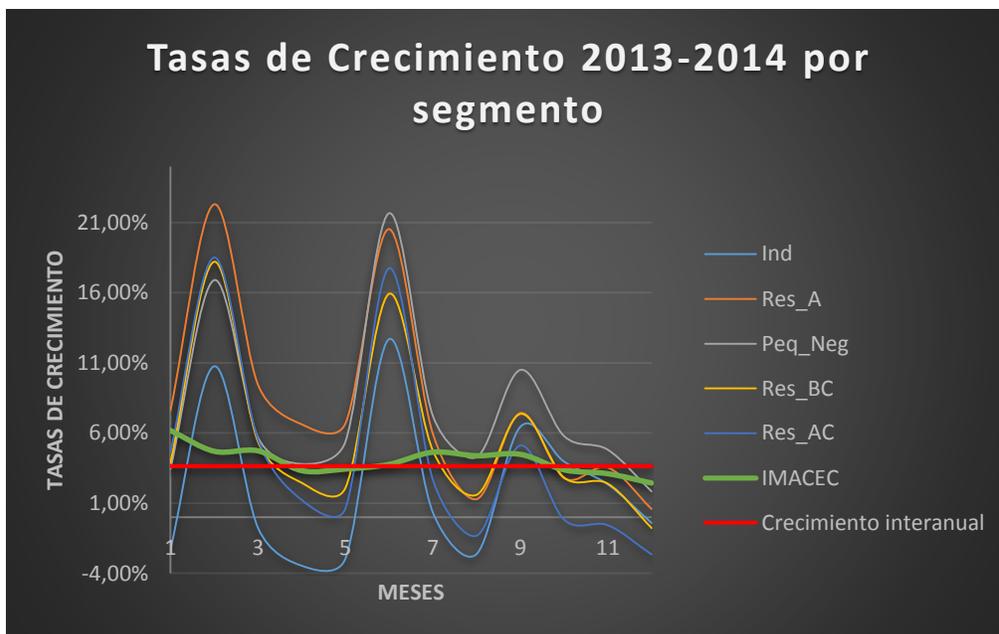


Figura 57: Tasas de crecimiento por segmento encontrado.

Finalmente con el objetivo de explorar la relación entre los consumos asociados a cada segmento de clientes encontrado y la variable macro-económica del IMACEC¹⁵ se calcularon las correlaciones de Pearson, Kendall y Spearman entre la serie compuesta por las tasas de crecimiento del IMACEC 2013-2014 y las tasas de crecimiento de los consumos de energía mensuales por segmento de clientes, tal como se muestra en la *tabla 15*, el script en R utilizado para dicho cálculo puede verse en apéndice 9.5.

	<i>Pearson</i>	<i>Kendall</i>	<i>Spearman</i>
ind	0,01	0,09	0,09
Res_A	0,31	0,48	0,66
Peq_Neg	0,14	0,24	0,33
Res_BC	0,32	0,42	0,64
Res_AC	0,36	0,52	0,70

Tabla 16: Distintas correlaciones entre las tasas de crecimiento del IMACEC y los consumos mensuales por tipo de cliente.

Se puede observar que el segmento que presenta una menor correlación independiente del estadístico utilizado para calcularla corresponde al segmento “Industrial” y “Pequeños Negocios”, por otro lado los segmentos de clientes que presenta una mayor correlación con el IMACEC corresponden a los segmentos “Residencial Alto Consumo”, “Residencial Bajo Consumo” y “Residencial Anómalo”, lo cual implica que poseen una alta sensibilidad a variaciones macro-económicas.

En la *figura 58*, se puede observar gráficamente la relación entre los consumos por segmentos y el IMACEC durante los años del período en estudio.

¹⁵ Indicador Mensual de Actividad Económica.

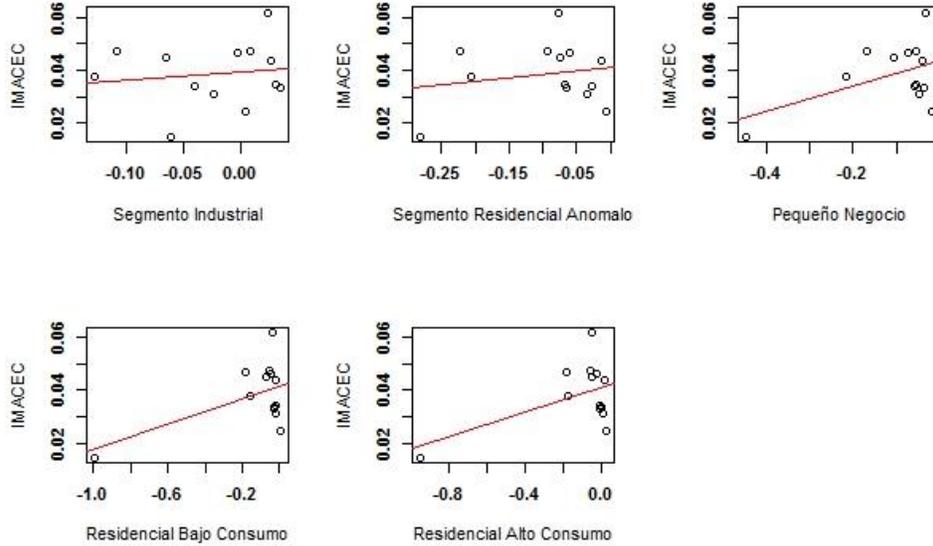


Figura 58: Relación entre el IMACEC y los consumos por segmento.

5.4.3 Ejemplo reconstrucción 5 transformadores tipo

Considerando el conjunto de transformadores con un mayor consumo total observado a lo largo del período de estudio se seleccionó un sub-conjunto de estos, los cuales corresponden a los transformadores de Apoquindo T1, Chacabuco T1, Batuco T3, San Cristóbal T3 y Alonso de Córdova T2.

En las *figuras 59, 60, 61, 62 y 63* de los anexos se observa la comparación entre el perfil de carga en un horizonte de 24 horas empírico versus el perfil de carga en el mismo horizonte calculado a partir de la reconstrucción de la matriz de observaciones utilizando la matriz de mezcla estimada y los componentes independientes encontrados.

A continuación en la *tabla 16* se observa la composición en términos de los market-share de cada segmento para los transformadores mencionados anteriormente.

	<i>Ind</i>	<i>Res_A</i>	<i>Peq_Neg</i>	<i>Res_BC</i>	<i>Res_AC</i>
CHACABUCOT1	24,25%	17,42%	26,28%	8,15%	23,90%
APOQUINDO T1	20,43%	25,20%	18,10%	27,26%	9,00%
ALONSO DE CORDOVA T2	28,31%	20,23%	8,79%	25,41%	17,26%
BATUCO-3	21,55%	14,50%	23,02%	17,70%	23,24%
SAN CRISTOBAL T3	21,65%	16,99%	21,69%	25,92%	13,75%

Tabla 17: Market-shares para los transformadores mencionados anteriormente.

5.5 Resultados y Análisis a nivel de subestación

A continuación se describen los principales resultados obtenidos de aplicar el algoritmo fastICA para estimar el modelo ICA de variables latentes sobre el data set de medidas obtenidas a nivel de subestaciones.

5.5.1 Cálculo de errores de reconstrucción

Siguiendo la misma metodología utilizada para el cálculo de errores computados en el data set de transformadores se computo el MAPE para el data set de subestaciones.

En la *figura 64* se observa el error medio porcentual absoluto (MAPE) tal como fue definido en el marco conceptual, considerando el error entre la reconstrucción \tilde{x} y la matriz de señales original x , para cada Experimento del algoritmo por transformador.

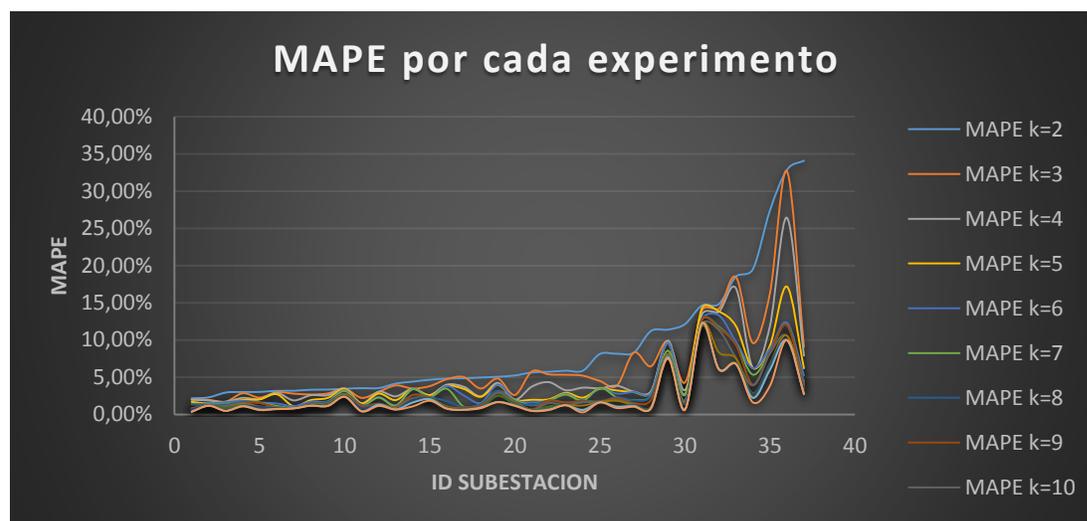


Figura 64: MAPE por cada experimento a nivel de subestación.

Al igual que en lo ocurrido a nivel de transformadores se observa un sub-conjunto de subestaciones que presentan consistentemente una mala reconstrucción independiente de la cantidad de componentes utilizados en la experimento, sin embargo a diferencia de lo ocurrido a nivel de transformadores las señales observadas en los equipos de medición asociados a las subestaciones no presentan distorsiones generadas por medidas de clientes libres, por ende para poder explicar los altos errores observados en la reconstrucción de dichas señales habría que intentar utilizar otros métodos de segmentación o contar con data contextual respecto a dichos equipos de medición para validar distintas hipótesis que expliquen dicho comportamiento anómalo, lo cual puede quedar propuesto para trabajos futuros.

En la *Tabla 17* se observa el MAPE asociado a la reconstrucción de la curva de consumo agregada del sistema a nivel de subestación, al igual que en el data set a nivel de transformador el error en la reconstrucción de la señal observada disminuye al aumentar la cantidad de componentes independientes que ingresan como input al algoritmo.

IC	MAPE	IC	MAPE
15	0,107%	8	0,154%
14	0,110%	7	0,161%
13	0,110%	6	0,161%
12	0,115%	5	0,172%
11	0,124%	4	0,392%
10	0,136%	3	0,520%
9	0,151%	2	1,106%

Tabla 18: MAPE curva sistema por experimento.

5.5.2 Principales iteraciones

A continuación se muestran los resultados obtenidos las iteraciones $k \in \{4, 5, 10, 15\}$.

5.5.2.1 Experimento $k=15$

Para esta Experimento se consideraron 15 componentes independientes. En los anexos *figuras 65, 66 y 67* se pueden observar las visualizaciones de los componentes independientes encontrados para esta Experimento.

En los anexos *figuras 68, 69 y 70* se pueden observar los perfiles de 24 horas estimados para los distintos componentes encontrados durante esta Experimento.

5.5.2.2 Experimento $k=10$

Para esta Experimento se consideraron 10 componentes independientes. En los anexos *figuras 71 y 72* se pueden observar las visualizaciones de los componentes independientes encontrados para esta Experimento.

En los anexos *figuras 73 y 74* se pueden observar los perfiles de 24 horas estimados para los distintos componentes encontrados durante esta Experimento.

5.5.2.3 Experimento $k=5$

Para esta Experimento se consideraron 10 componentes independientes. En los anexos *figura 75* se pueden observar las visualizaciones de los componentes independientes encontrados para esta Experimento.

En los anexos *figuras 76* se pueden observar los perfiles de 24 horas estimados para los distintos componentes encontrados durante esta Experimento.

5.5.2.4 Experimento $k=4$

Para esta Experimento se consideraron 10 componentes independientes. En los anexos *figura 77* se pueden observar las visualizaciones de los componentes independientes encontrados para esta Experimento.

En los anexos *figuras 78* se pueden observar los perfiles de 24 horas estimados para los distintos componentes encontrados durante esta Experimento.

5.6 Principales segmentos encontrados

Al igual que en el caso del data set compuesto por equipos de medición asociado a transformadores, existen restricciones en términos de la interpretabilidad de los perfiles de carga encontrados por parte de la contraparte experta en la compañía.

Por ello del total de iteraciones realizadas se seleccionó la Experimento con $k=5$ componentes independientes, es decir con 5 segmentos distintivos de clientes que explican el comportamiento del mercado regulado a nivel de subestación,

Cada segmento fue nombrado de acuerdo a un patrón de consumo particular, siguiendo la nomenclatura utilizada en los segmentos encontrados a nivel de transformador se tienen los siguientes segmentos:

- Segmento “Industrial”
- Segmento “Residencial Anómalo”
- Segmento “Pequeños Negocios”
- Segmento “Residencial Bajo Consumo”
- Segmento “Residencial Alto Consumo”

En la *figura 79* se puede observar los perfiles de carga asociados a cada segmento encontrado.

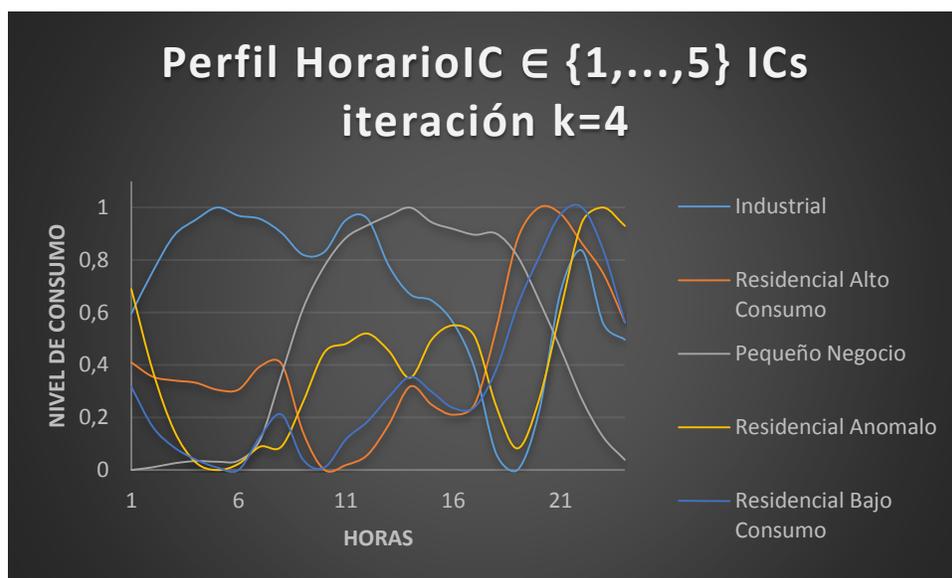


Figura 79: Perfiles de carga para la Experimento $k=5$ aplicada al data set a nivel de subestación.

5.6.1 Market-shares y tasas de crecimiento por sub-estación

Siguiendo la misma metodología utilizada para el cálculo de market shares a nivel de transformadores se obtienen los market-share por cada subestación asociados a cada segmento de clientes encontrados.

En los anexos de la *tabla 18* a la *tabla 21* se puede observar el market-share para cada segmento de clientes por equipo de medición, donde **Ind** corresponde al segmento industrial,

Res_A al segmento residencial anómalo, **Peq_neg** al segmento de pequeños negocios, **Res_BC** al segmento residencial de bajo consumo y **Res_AC** al segmento residencial de alto consumo.

En la *figura 80* se puede observar la tasa de crecimiento asociada a cada segmento de clientes encontrado, se observa que distintos segmentos de clientes presentan distintas tasas de crecimiento para el período en estudio, respondiendo en forma diferente a variables exógenas que puedan afectar sus patrones de consumo observado.

Por otro lado en la tabla 22 se observan la tasa de crecimiento promedio anual estimada por segmento encontrado, para el período de estudio aquel segmento que presenta una menor tasa de crecimiento corresponde al “Residencial de Alto Consumo”, mientras que el segmento que presenta una mayor tasa de crecimiento para el período en estudio corresponde al “Residencial de Bajo Consumo”.

Ind	3,9%
Res_AC	-0,9%
Peq_Neg	4,4%
Res_A	4,2%
Res_BC	8,0%

Tabla 22: Tasas de crecimiento por segmento encontrado a nivel de subestacion.

Dentro del período en estudio los dos segmentos que manifestaron mayores de tasas de crecimiento corresponden a los segmentos de pequeños negocios y residencial bajo consumo, mientras que los dos segmentos que experimentaron menores tasas de consumo corresponden al residencial de alto consumo y al segmento industrial.

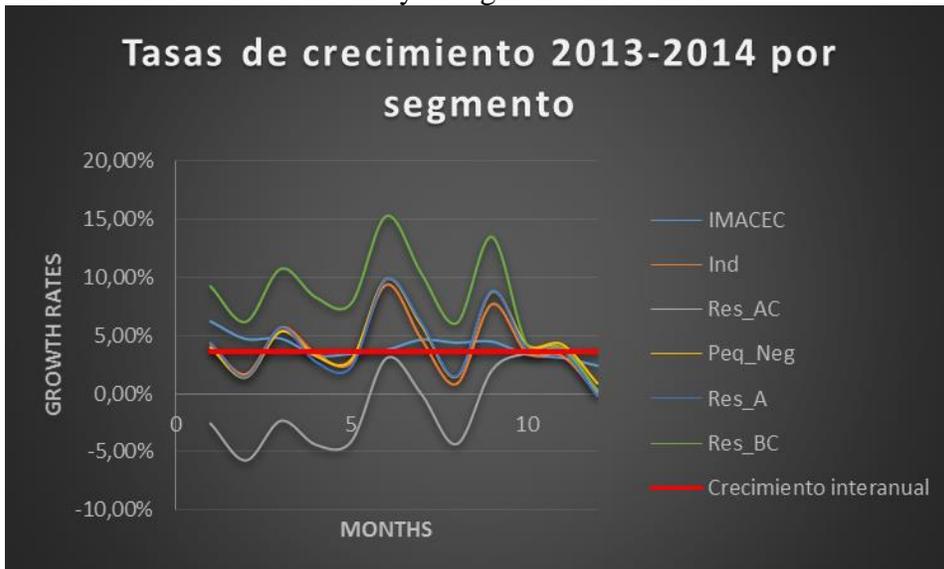


Figura 80: Tasas de crecimiento para los segmentos encontrados durante el período de estudio.

Se observan diferencias en las tasas de crecimiento estimadas por segmentos de clientes en relación al data set de medidas asociadas a transformadores, una posible explicación radica en que dicho data set presenta distorsiones por series de consumo asociadas a clientes libres

que no fueron limpiadas debido a restricciones en términos de los puntos de conexión de dichos clientes.

Lo anterior puede explicar porque el segmento “Residencial Anómalo” varía sus tasas de crecimiento, pues una posible hipótesis es que dicho segmento de clientes capture ciertos patrones de consumo subyacentes asociados a clientes libres presentes en el data set asociado a los transformadores.

Finalmente siguiendo lo realizado a nivel de transformador se calcularon las correlaciones de Pearson, Kendall y Spearman entre la serie del IMACEC 2013-2014 y los consumos de energía mensuales por segmento de clientes, tal como se muestra en la *tabla 22*.

	<i>Pearson</i>	<i>Kendall</i>	<i>Spearman</i>
ind	0,57	0,43	0,61
Res_A	0,57	0,39	0,57
Peq_Neg	0,34	0,26	0,40
Res_BC	0,42	0,30	0,43
Res_AC	0,68	0,58	0,77

Tabla 23: Distintas correlaciones entre el IMACEC y los consumos mensuales por tipo de cliente.

El segmento que presenta una menor correlación en términos de sus tasas de crecimiento interanual con las tasas de crecimiento interanual observadas del IMACEC corresponde al segmento identificado como “Pequeños Negocios”, el cual presenta una correlación que varía entre un 0,4 y un 0,25 dependiendo del estadístico utilizado para medirla, por otro lado los segmentos “Residencial Alto Consumo”, “Residencial Anómalo” e “Industrial” presentan altas correlaciones medidas en los tres estadísticos mencionados anteriormente con la tasa de crecimiento observada del IMACEC.

Finalmente el segmento “Residencial Bajo Consumo” también presenta bajas correlaciones con la tasa de crecimiento observada del IMACEC para el período de estudio.

Se puede inferir que aquellos segmentos que son más sensibles a variaciones en atributos macro-económicos corresponden a los dos segmentos residenciales de alto consumo y anómalo además del perfil de consumo asociado a un segmento industrial, mientras que aquellos que presentan una menor sensibilidad a variaciones macro-económicas puede asociarse los segmentos etiquetados como residencial bajo consumo y pequeños negocios.

En la *figura 81*, se puede observar gráficamente la relación entre los consumos por segmentos y el IMACEC durante los años del período en estudio.

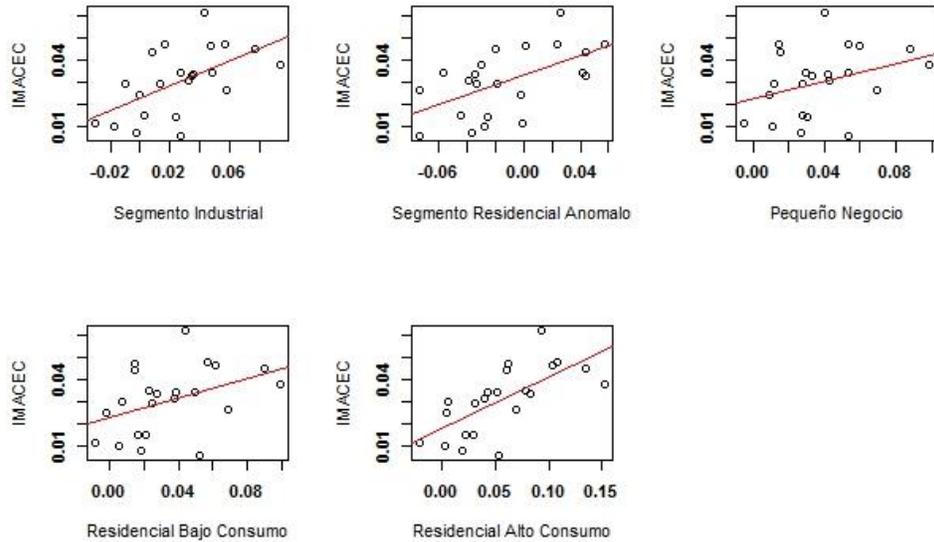


Figura 81: Relación entre el IMACEC y los consumos por segmento.

5.6.2 Ejemplo reconstrucción 5 sub-estaciones tipo

Considerando el conjunto subestaciones con un mayor consumo total observado a lo largo del período de estudio se seleccionó un sub-conjunto de estos, los cuales corresponden a las subestaciones de Lo Boza, San Joaquin, Chacabuco, Apoquindo y Alonso de Córdoba.

En las *figuras 82, 83, 84, 85 y 86* de los anexos se observa la comparación entre el perfil de carga en un horizonte de 24 horas empírico versus el perfil de carga en el mismo horizonte calculado a partir de la reconstrucción de la matriz de observaciones utilizando la matriz de mezcla estimada y los componentes independientes encontrados.

A continuación en la *tabla 23* se observa la composición en términos de los market-share de cada segmento para las subestaciones mencionadas anteriormente.

	Ind	Res_AC	Peq_Neg	Res_A	Res_BC
ALONSO DE CORDOVA	30,6%	7,3%	15,8%	26,7%	19,5%
APOQUINDO	11,8%	20,4%	18,9%	17,8%	31,0%
CHACABUCO	0,0%	16,0%	35,2%	33,5%	15,3%
LO BOZA	9,8%	23,8%	34,5%	24,7%	7,2%
SAN JOAQUIN	17,3%	27,2%	22,6%	16,0%	16,9%

Tabla 24: Market-shares para los transformadores mencionados anteriormente.

Capítulo 6: Conclusiones

A continuación se detallan las principales conclusiones del estudio realizado además de un balance del trabajo realizado, evaluando el grado de cumplimiento de los objetivos definidos en la sección de objetivos general y específico. Además se resumen las principales conclusiones extraídas de los resultados obtenidos y una serie de recomendaciones para trabajos futuros.

6.1 Principales conclusiones del estudio

Uno de los principales desafíos de la Unidad de Movimiento de Energía consiste en administrar los sistemas de bases de datos en los cuales se registran las observaciones capturadas por diversos equipos de medición en distintos puntos del anillo de concesión de Chilectra.

De dichos sistemas la unidad extrae una serie de medidas las cuales son utilizadas para generar reportes de facturación y compra de energía, los cuales son vitales para la operación de la Gerencia Técnica, dentro de la cual se encuentra la unidad mencionada anteriormente.

Sin embargo a pesar de contar data sets de gran tamaño que capturan observaciones con un alto nivel de frecuencia y que pueden ser enriquecidos con data sets provenientes de otros sistemas de información dentro de la compañía, la unidad realiza pocos estudios de minería de datos y búsquedas de patrones para intentar describir, predecir y prescribir el comportamiento capturado por dichas observaciones, asociados a patrones de consumo eléctrico.

El principal objetivo de esta memoria es intentar descomponer las señales de consumo asociadas al mercado de clientes regulados para describir dicha señal en término de sus componentes independientes y encontrar segmentos de clientes con perfiles de consumo similares. Considerando el punto anterior este trabajo puede entenderse como una primera aplicación de técnicas de data mining para extraer inteligencia y describir patrones dentro de los data sets con los cuales opera la unidad.

La metodología desarrollada se encuentra basada en las etapas clásicas del proceso KDD además y el framework propuesto [17], dentro de las diversas etapas que contemplan la metodología propuesta aquella que consumió una mayor cantidad de horas hombre dentro del proyecto corresponde a las etapas de extracción y pre-procesamiento de los data sets obtenidos, esto es debido a que paradójicamente a pesar de que la Unidad de Movimiento de Energía maneja grandes volúmenes de datos provenientes de diversas bases de datos transaccionales poseen sistemas de administración de bases de datos poco robustos, lo cual dificultó la extracción de las medidas asociadas tanto a nivel de transformadores como de subestaciones.

Uno de los principales problemas enfrentados durante el desarrollo del proyecto corresponde a la contaminación de las señales extraídas por parte de los clientes considerados libres, los cuales no pertenecen al consumo observado para el mercado regulado, debido a que operacionalmente la Unidad de Movimiento de Energía genera reportes de facturación y compra utilizando información consolidada de series de consumos extraídas a nivel de subestación, las medidas extraídas a nivel de transformador no se encuentran totalmente depuradas, y existen inconsistencias inter-temporales en términos de los puntos de conexión de estos clientes libres que no permiten depurar completamente las señales observadas para generar un data set que capture solamente las señales asociadas al mercado regulado a nivel de transformador.

Dada la naturaleza del problema, en el cual las señales de consumo individuales no son registradas por los equipos de medición y lo que se registran son señales de consumo generadas por un proceso de mezcla aleatorio, la herramienta seleccionada para clusterizar dichas señales es la adecuada considerando que estas son generadas por variables latentes asociados al consumo individual de cada cliente.

La metodología implementada permite cumplir exitosamente los diversos objetivos planteados inicialmente, distinguiendo un total de 5 segmentos de clientes, con perfiles de consumo particulares observados a lo largo de distintas ventanas temporales, además dichos perfiles fueron encontrados en dos data sets con niveles de mezcla distinto, uno “agua arriba” en el anillo de concesión como es el data set asociado a las subestaciones, el cual tiene la ventaja que considera solamente señales totalmente depuradas asociadas al mercado regulado y otro “aguas abajo” en el anillo de concesión asociado a señales que en su gran mayoría están asociadas al mercado regulado pero que presentan cierto grado de distorsión debido a algunos clientes libres cuyas curvas de consumo están capturadas en dichos equipos de medición.

6.2 Principales Resultados obtenidos

- En términos descriptivos ambos data sets presentan distribuciones de consumo que siguen un proceso aleatorio no-gaussiano, además se observa una tendencia anual asociada a los periodos de alta demanda, meses de Abril a Septiembre, y baja demanda, resto del año, dentro del período de estudio.
- Al incrementar la cantidad de componentes independientes que ingresan como parámetros al algoritmo fastICA van disminuyendo paulatinamente los errores en términos de la reconstrucción de la matriz de observaciones asociada a cada experimento para ambos data sets.
- A nivel de transformadores existe un subconjunto de estos que consistentemente presentan altos errores en la reconstrucción, medidos a través del MAPE, MAE y N-RMSE de las señales observadas por equipo de medición, independientemente de la cantidad de componentes independientemente utilizados. A nivel de la curva reconstruida del sistema el MAPE promedio es de un 0.35%, el MAE promedio de 153.45 [KW] y el N-RMSE es de un 0.67%.
- A nivel de subestaciones los errores medidos en términos del MAPE son bastante menores a los observados a nivel de transformador, el MAPE promedio para todos los experimentos a nivel de transformador es de un 0.36%, mientras que el MAPE promedio a nivel de subestación es de un 0.251%.
- Considerando la cantidad de segmentos escogidos para el experimento con $k=5$ componentes independientes el MAPE en la reconstrucción de la curva del sistema es de un 0.172% a nivel de subestación versus un 0.57% a nivel de transformador.
- A nivel de equipos de medición el MAPE medido también disminuye a nivel de subestación comparado con el nivel de transformador, pasando desde un 4.36% a un 6.44% en el caso de los transformadores.
- Considerando la curva total del sistema, los errores asociados a la reconstrucción disminuyen consistentemente con la cantidad de componentes independientes encontrados en ambos data sets, pues la forma funcional de esta tiende a suavizar errores puntuales en la reconstrucción asociados a diversos equipos de medición.
- A nivel de transformadores los 5 segmentos encontrados presentan diferencias en términos de los consumos máximos, medios y factores de carga para las ventanas temporales definidas, observándose máximos de consumos y medias de consumo distintas para los segmentos encontrados. Esto sugiere que los distintos segmentos tienen desfases en sus temporadas de alta y baja demanda y fluctúan sus patrones de consumo en forma distinta a lo largo del período de estudio.
- A nivel de transformadores se observa que los segmentos que presentan mayores tasas de crecimiento durante el período en estudio corresponden a los segmentos

“Pequeños Negocios” y “Residencial Anómalo” con tasas de crecimientos interanuales estimadas de 7.6% y 7.9% respectivamente, mientras que los segmentos que presentan menores tasas de crecimiento corresponden a los segmentos “Residencial Alto Consumo” e “Industrial” con tasas de crecimiento interanuales estimadas de 4.3% y 2.0% respectivamente.

- A nivel de subestación se observa que los segmentos que presentan mayores tasas de crecimiento durante el período en estudio corresponden a los segmentos “Pequeños Negocios” y “Residencial Bajo Consumo” con tasas de crecimiento interanuales estimadas de 4.4% y 8.0% respectivamente, mientras que los segmentos que presentan menores tasas de crecimiento corresponden a los segmentos “Residencial Alto Consumo” e “Industrial” con tasas de crecimiento estimadas de -0.9% y 3.9% respectivamente.
- A nivel de transformadores los segmentos que presentaron una menor correlación entre las tasas de crecimiento del IMACEC observadas y las tasas de crecimiento del consumo eléctrico observado corresponden a los segmentos de “Industrial” y “Pequeños Negocios”, con correlaciones de Pearson, Kendall y Spearman de {0.01; 0.09; 0.09} y {0.14; 0.24; 0.34} respectivamente. Por otro lado el segmento que presentaron una mayor correlación entre las tasas de crecimiento del IMACEC observadas y las tasas de crecimiento del consumo eléctrico observado corresponde al segmento “Residencial Alto Consumo” con correlaciones de Pearson, Kendall y Spearman de {0.36; 0.52; 0.70}.
- A nivel de subestación los segmentos que presentaron una menor correlación entre las tasas de crecimiento del IMACEC observadas y las tasas de crecimiento del consumo eléctrico observado corresponden a los segmentos de “Residencial Bajo Consumo” y “Pequeños Negocios”, con correlaciones de Pearson, Kendall y Spearman de {0.34; 0.26; 0.40} y {0.42, 0.30, 0.43} respectivamente. Por otro lado el segmento que presentaron una mayor correlación entre las tasas de crecimiento del IMACEC observadas y las tasas de crecimiento del consumo eléctrico observado corresponde al segmento “Residencial Alto Consumo” con correlaciones de Pearson, Kendall y Spearman de {0.68; 0.58; 0.77}.

6.3 Principales limitaciones de los resultados obtenidos

La principal limitación de los resultados obtenidos viene dada por la ausencia de un data set con variables contextuales, socio-demográficas, geográficas entre otras, que permitan validar distintas hipótesis de comportamiento para los segmentos encontrados, y con ellas explicar las distintas particularidades en los patrones de consumo hallados, refinando en cierta forma la etiqueta dada para cada segmento encontrado, así por ejemplo teniendo información muestral sobre una serie de clientes y sus puntos de conexión se podría establecer la verdadera naturaleza del segmento “Residencial Anómalo”, o calcular diversas elasticidades para los segmentos encontrados en base a indicadores económicos, serie de temperaturas medias, entre otros. Por otro lado las restricciones en términos de la distorsión de las señales en el data set de transformadores comparado con el data set a nivel de subestación no permiten comparar directamente los segmentos encontrados.

6.4 Trabajos futuros

La metodología planteada utilizarse para generar segmentaciones de clientes dentro del mercado de clientes regulados en diversos horizontes temporales, sin embargo sería interesante poder observar la evolución de estos segmentos en distintas escalas temporales capturando la dinámica de cómo estos segmentos asociados a patrones de consumo van evolucionando en base a las variaciones observadas en una serie de variables explicativas como indicadores económicos, temperatura, entre otros.

Una extensión interesante del presente trabajo sería comparar los resultados obtenidos por distintas herramientas en términos de los segmentos encontrados, dada la naturaleza de la herramienta ICA sería natural realizar una comparación en términos de la reconstrucción de las señales observadas mediante los segmentos encontrados utilizando ICA, PCA y alguna herramienta de clusterización estándar como K-medias con la idea de evaluar el ajuste de los distintos segmentos encontrados.

Otro de los puntos en los cuales se podrían desarrollar futuros trabajos y siguiendo lo propuesto por [1] dice relación con levantar un data set con variables contextuales asociadas a los clientes finales, considerando diversas muestras de clientes dentro del anillo de concesión, sus diversos puntos de conexión y atributos que caractericen sus consumos individuales como atributos que describen el tipo de cliente, el tipo de hogar, cantidad de aparatos eléctricos por cliente, tipos de aparatos eléctricos por cliente, zona geográfica, entre otros. Con lo anterior se podrían caracterizar de forma más certera los segmentos encontrados.

Un último punto en el cual se puede ahondar consiste en incorporar información respecto a la naturaleza temporal de las observaciones, separando efectos entre observaciones a nivel diario, semanal o mensual y efectos presentes dentro del día, es decir cruzando información para obtener segmentos “dentro de un horizonte de 24 horas” y “entre horizontes de 24 horas”, esto permitiría capturar patrones de consumo dentro del día o dentro de un horizonte temporal menor a 24 horas, asociados a cambios rápidos en los consumos observados y efectos de más largo plazo como aquellos patrones capturados por el presente estudio.

Capítulo 7: Bibliografía

- [1] F. McLoughlin, «“Characterising Domestic Electricity Demand for Customer Load Profile Segmentation”», Dublin Institute of Technology, Dublin, 2013.
- [2] M. O. G. C. S. A. G. M. Sergio Valero Verdu, «“Classification, Filtering, and Identification of Electrical Customers Load Patterns Through the Use of Self-Organizing Maps”», *IEEE Transactions on Power Systems*, vol. 21, pp. 885-895, 2006..
- [3] C. B. J. M. A. B. C. a. A. S.-E. I. 1.-1. 5. 5.-5. 2. d. M. 2. D. Luis Hernandez, «“Classification and Clustering of Electricity Demand Patterns in Industrial Parks”», *Energies ISSN*, vol. 5, pp. 5215-5228., Abril 2012..
- [4] P. G. Andrzej Cichocki, «Blind Source Separation Algorithms with Matrix Constraints», de *Blind Signal Processing: IEICE Transactions Fundamentals*, 2003.
- [5] M. A. L. H. A. K. K.A.D. Deshani, «“An exploratory analysis on half-hourly electricity load patterns leading to higher performances in neural network predictions”», *International Journal of Artificial Intelligence & Applications*, vol. 5, pp. 122-137, 2014.
- [6] D. W. I. Rui Xu, «"Survey of Clustering Algorithms"», *IEEE TRANSACTIONS ON NEURAL NETWORKS*, vol. 16, pp. 645-679, 2005.
- [7] T. W. Liao, «“Clustering of time series data-a survey”», *The Journal of the Pattern Recognition Society*, vol. 38, pp. 1857-1874, 2015.
- [8] S. F. C. a. N. Kourentzes, «“The impact of preprocessing on forecasting electrical load: an empirical evaluation of segmenting time series into subseries”», de *Proceedings of International Joint Conference on Neural Networks*, 2015.
- [9] E. O. A. Hyvärinen, «Independent Component Analysis: Algorithms and applications», *Neural Networks*, vol. 13, pp. 411-430, 2000.
- [10] J. K. a. E. O. Aapo Hyvärinen, de *Independent Component Analysis*, New York, John Wiley & Sons Inc., 2001, p. Chapter 7: "What is ICA?".
- [11] A. H. y. F. E. Johan Himberg, «Validating the independent components of neuroimaging time-series via clustering and visualization», *Neuro Imaging*, vol. 22, pp. 1214-1222, 2004.
- [12] L. X. Yiu-Ming Cheung, «Independent Component Ordering in ICA Time Series Analysis», *Neurocomputing*, vol. 41, pp. 145-152, 2001.
- [13] H. L. a. E. G. Dagmar Niebur, «Chapter 13: Independent Component Analysis Techniques for Power System Load Estimation», de *Applied Mathematics to Power Systems*, 2005, pp. 288-315.
- [14] Dagmar Niebur, Huaiwei Liao, «Load Profile Estimation in Electric Transmission Network Using Independent Component Analysis», *IEEE Transactions on Power Systems*, vol. 18, pp. 707-715, 2003.
- [15] Y. Zhao, *R and Data Mining: Examples and Case Studies*, Sydney: Australia: Academic Press., 2012.

- [16] A. Prokhorov, «Kendall coefficient of rank correlation,» de *Encyclopedia of Mathematics*, Springer, 2001, pp. 978-998.
- [17] S. K. P. a. P. M. Sushmita Mitra, «“Data Mining in a Soft Computing FrameWork: A Survey”»,» *IEEE Transactions on Neural Networks*, vol. 13, pp. 3-14, 2002.
- [18] E. C. I. H. Hyunchul Ahn, «Extracting underlying meaningful features and canceling noise using independent component analysis for direct marketing.,» *Expert Systems with Applications.*, vol. 33, pp. 181-191, 2007.
- [19] F. Anscombe y W. J. Glynn, «"Distribution of the kurtosis statistic b2 for normal statistics".,» *Biometrika*, vol. 70, n° 1, pp. 227-234, 1983.
- [20] A. S. W. Andrew D. Back, «A First Application of ICA to Extracting Structure from Stock Returns,» *International Journal of Neural Systems*, vol. 8, pp. 97-115, 1997.
- [21] K. K. E. O. Simona Malaroiu, «Time Series Prediction with ICA,» de *Proceeding of the International Conference on Neural Information Processing.*, Tokyo, Japon., 1998.
- [22] J. C. R. Wei Lu, «Approach and Applications of Constrained ICA.,» *IEEE Transactions on Neural Networks.*, vol. 16, pp. 203-215, 2005.
- [23] C. R. Y. S. Udayan Ghose, «Socio-Economic Characterization of Student's Data using ICA and Cluster Analysis,» de *8th IEEE International Conference on Industrial Informatics*, Osaka, 2010.

Capítulo 8: Anexos

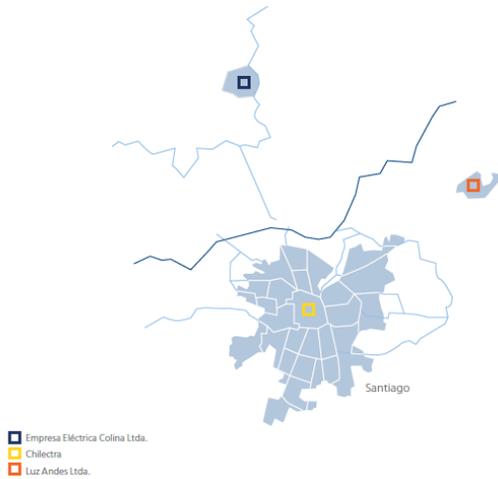


Figura 1: Área de concesión de la compañía.

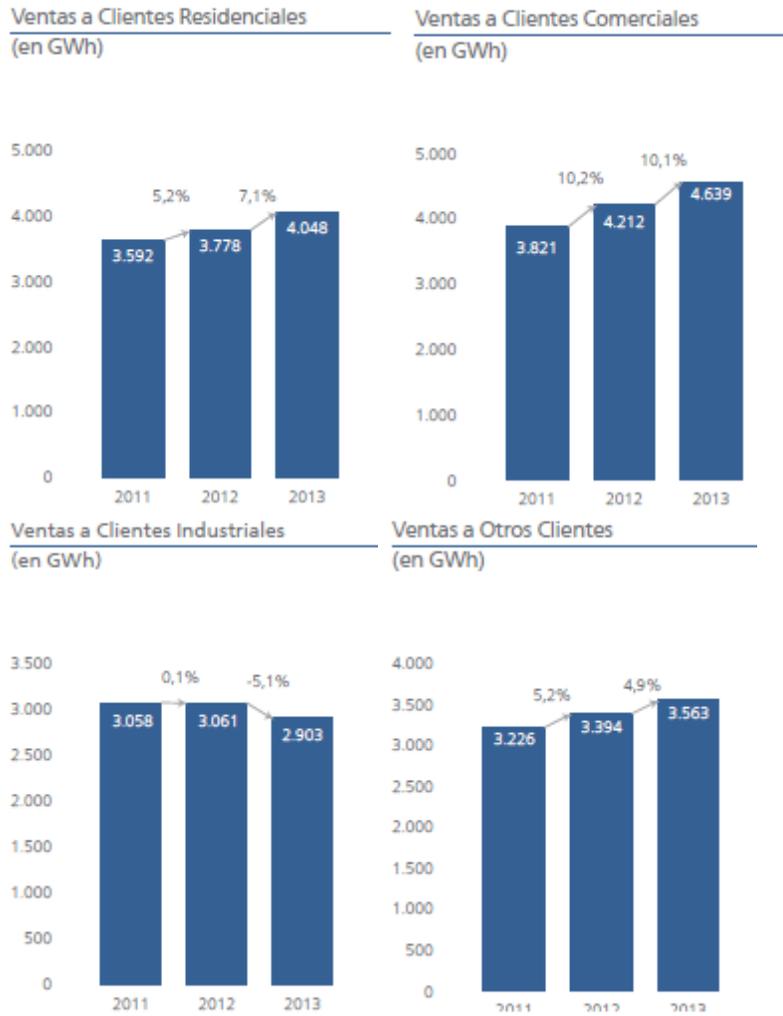


Figura 2: Crecimientos año 2013 por tipo de cliente según tipificación Gerencia Comercial.

Evolución Nº Clientes
(miles de clientes)

Ventas Totales de Energía
(en GWh)

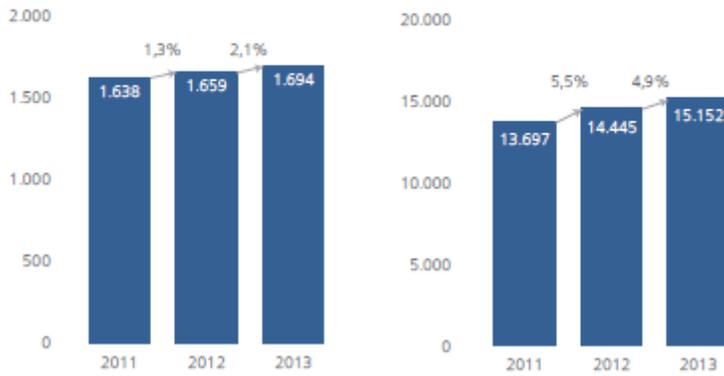


Figura 3: Crecimientos año 2013 cantidad total de clientes y ventas totales físicas de energía.

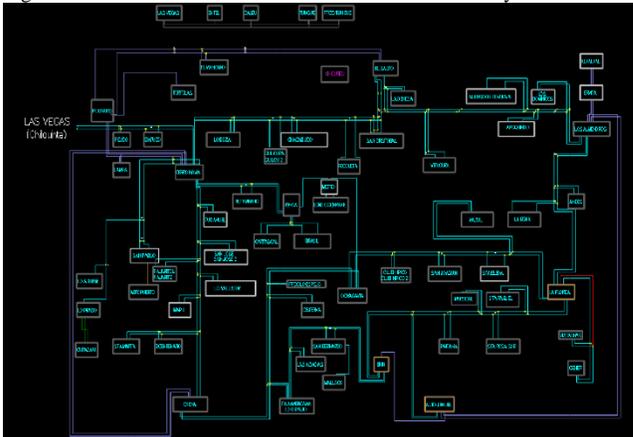


Figura 4: Representación gráfica del anillo de concesión.

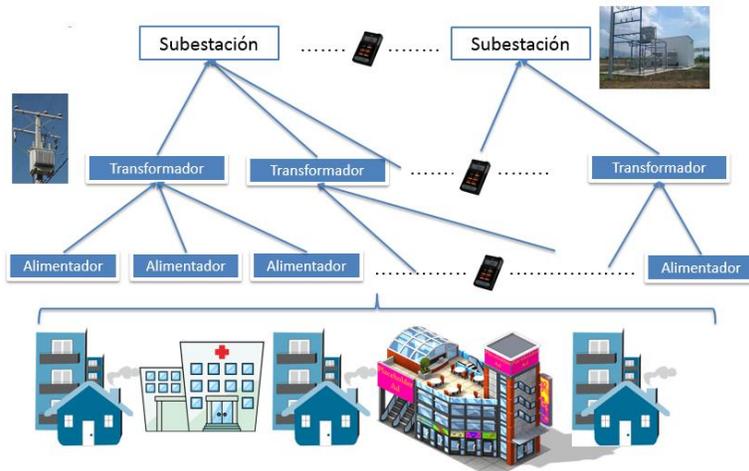
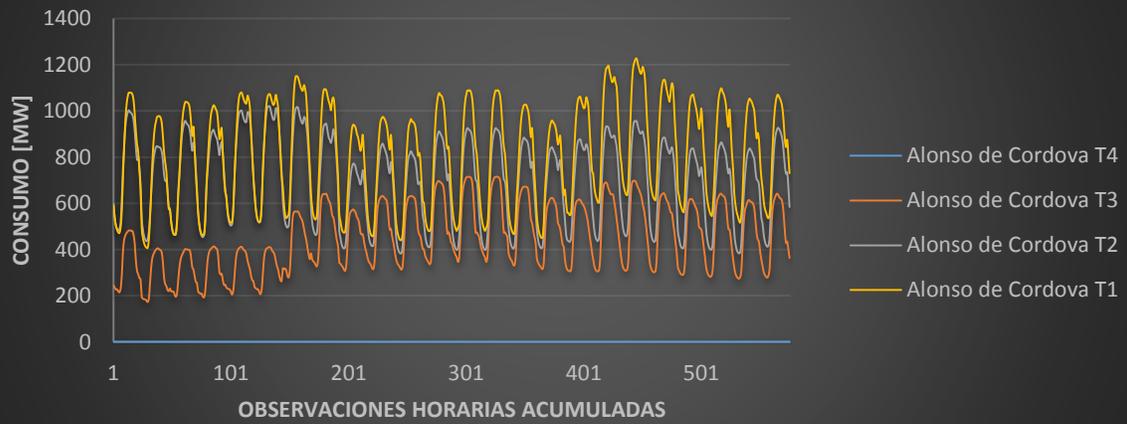


Figura 5: Esquema conceptual anillo de concesión.

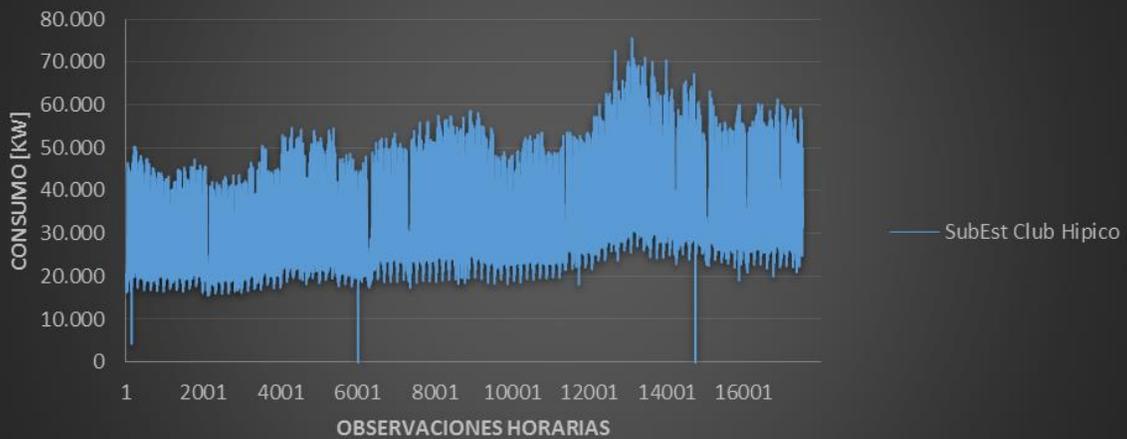
Visualización Consumo por Subestación



Visualización Consumo por Transformador



Visualización Consumo por Subestación



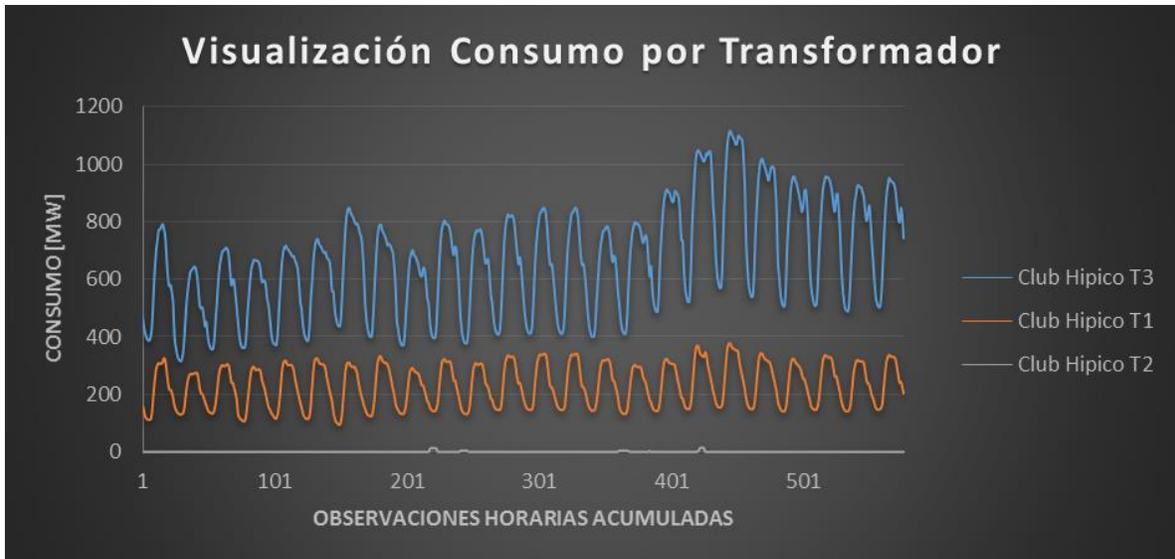


Figura 6: Distintas visualizaciones para la data proveniente de las subestaciones Club Hípico y Alonso Córdova, además de sus respectivos transformadores.

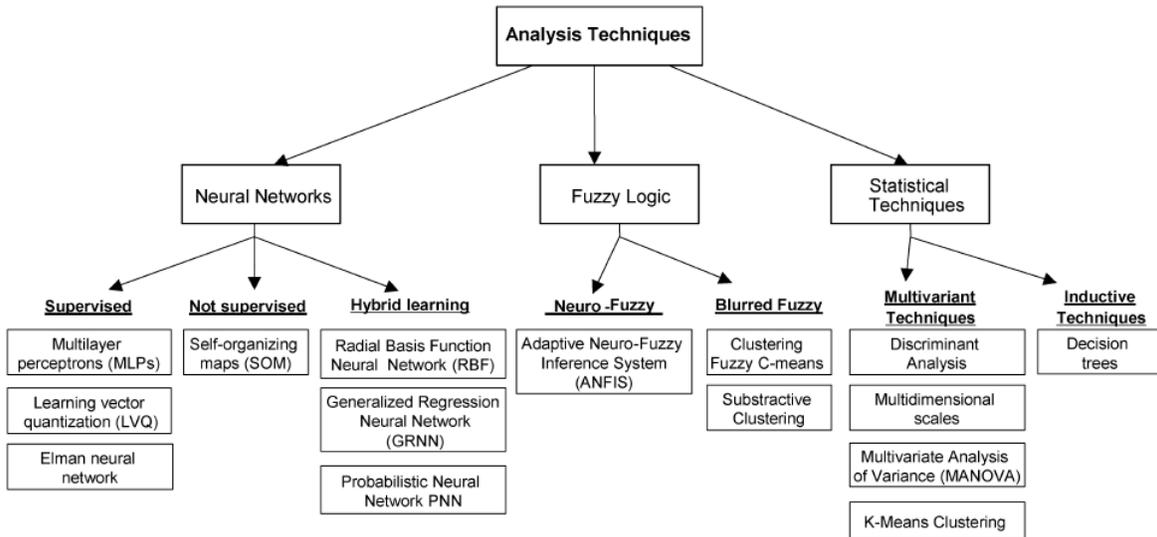


Figura 7: Distintas técnicas utilizadas para segmentar curvas de carga asociadas a sistemas de potencia eléctrica.

Measures	Forms	Comments	Examples and Applications
Minkowski distance	$D_p = \left(\sum_{i=1}^d x_{ji} - x_{ji} ^p \right)^{1/p}$	Metric. Invariant to any translation and rotation only for $n=2$ (Euclidean distance). Features with large values and variances tend to dominate over other features.	Fuzzy c -means with measures based on Minkowski family [130].
Euclidean distance	$D_2 = \left(\sum_{i=1}^d x_{ji} - x_{ji} ^2 \right)^{1/2}$	The most commonly used metric. Special case of Minkowski metric at $n=2$. Tend to form hyperspherical clusters.	K -means algorithm [191]
City-block distance	$D_1 = \sum_{i=1}^d x_{ji} - x_{ji} $	Special case of Minkowski metric at $n=1$. Tend to form hyperrectangular clusters.	Fuzzy ART [57]
Sup distance	$D_\infty = \max_{1 \leq i \leq d} x_{ji} - x_{ji} $	Special case of Minkowski metric at $n \rightarrow \infty$.	Fuzzy c -means with sup norm [39].
Mahalanobis distance	$D_M = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j)$, where \mathbf{S} is the within-group covariance matrix.	Invariant to any nonsingular linear transformation. \mathbf{S} is calculated based on all objects. Tend to form hyperellipsoidal clusters. When features are not correlated, squared Mahalanobis distance is equivalent to squared Euclidean distance. May cause some computational burden.	Ellipsoidal ART [13], Hyperellipsoidal clustering algorithm [194].
Pearson correlation	$D_p = (1 - r_p) / 2$, where $r_p = \frac{\sum_{i=1}^d (x_{ji} - \bar{x}_i)(x_{ji} - \bar{x}_j)}{\sqrt{\sum_{i=1}^d (x_{ji} - \bar{x}_i)^2 \sum_{i=1}^d (x_{ji} - \bar{x}_j)^2}}$	Not a metric. Derived from correlation coefficient. Unable to detect the magnitude of differences of two variables.	Widely used as the measure for analyzing gene expression data [80].
Point symmetry distance	$D_{ps} = \min_{\substack{j=1, \dots, N \\ j \neq i}} \frac{\ (\mathbf{x}_i - \mathbf{x}_j) + (\mathbf{x}_j - \mathbf{x}_i)\ }{\ (\mathbf{x}_i - \mathbf{x}_j)\ + \ (\mathbf{x}_j - \mathbf{x}_i)\ }$	Not a metric. Compute the distance between an object \mathbf{x}_i and a reference point \mathbf{x}_j . D_{ps} is minimized when a symmetric pattern exists.	SBKM (Symmetry-based K -means) [264].
Cosine similarity	$S_y = \cos \alpha = \frac{\mathbf{x}_i^T \mathbf{x}_j}{\ \mathbf{x}_i\ \ \mathbf{x}_j\ }$	Independent of vector length. Invariant to rotation, but not to linear transformations.	The most commonly used measure in document clustering [261].

Figura 8: Distintas nociones de distancia y/o similitud.

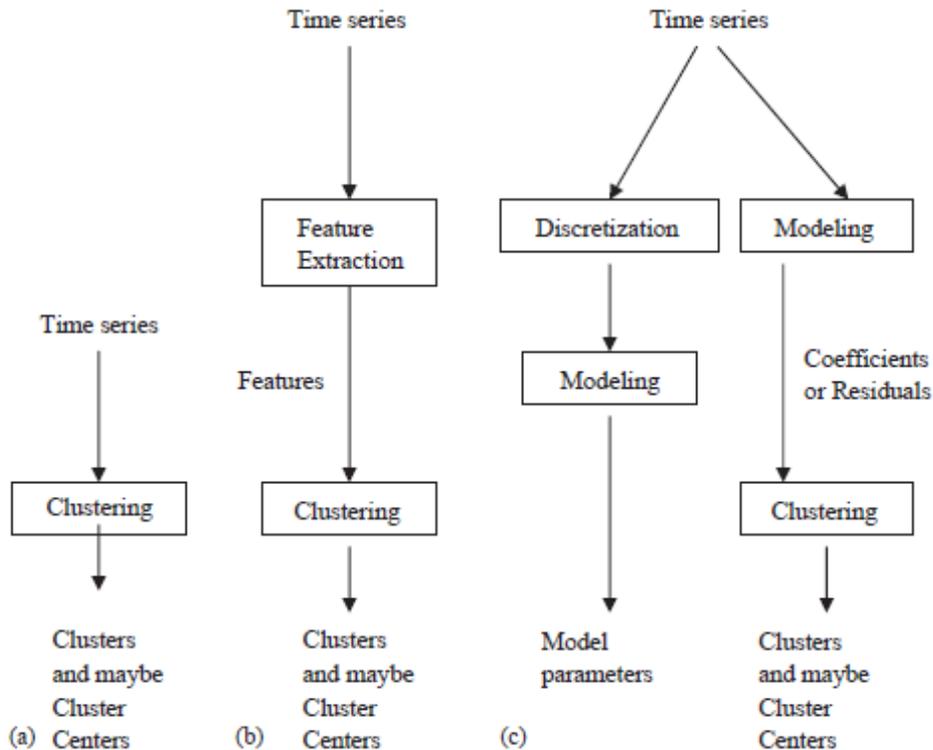


Figura 9: Raw-Data, Model-Based approach y Feature-Based approach para segmentar series de tiempo.

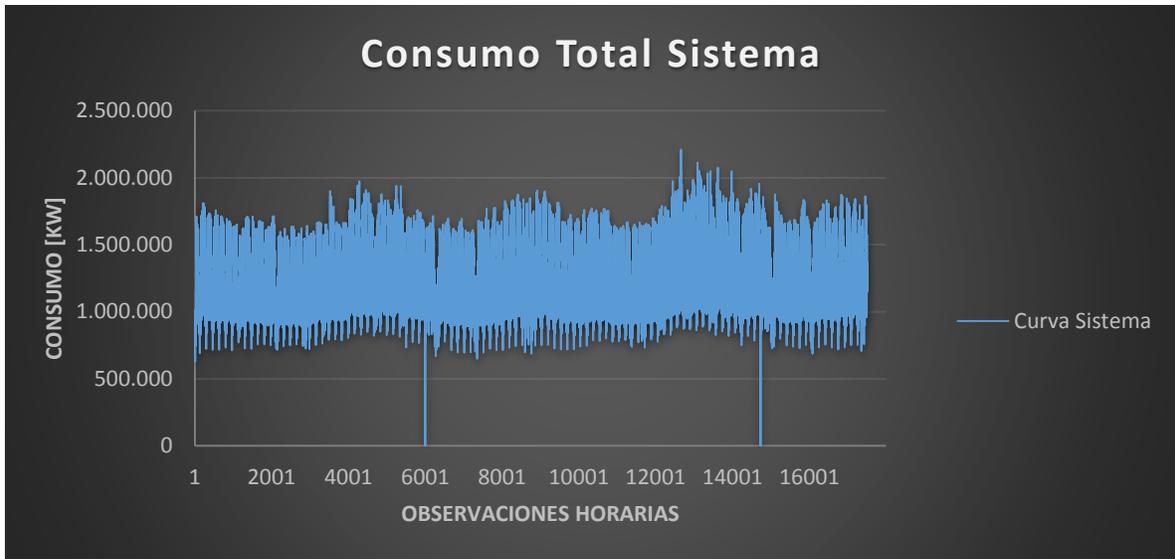


Figura 11: Consumo Total del Sistema mediante una visualización horaria para todo el período de estudio.

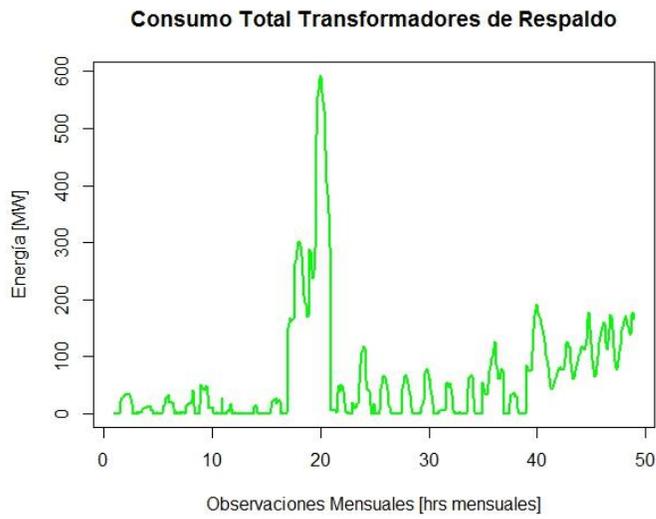


Figura 14: Serie de consumo total asociado a transformadores de respaldo.

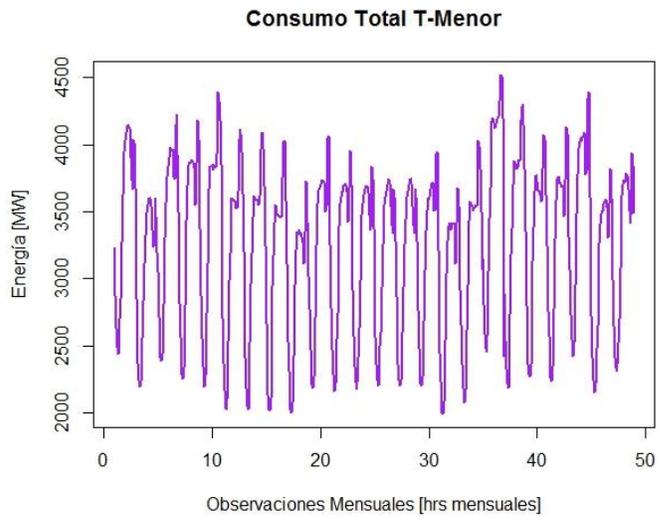


Figura 17: Serie de consumo total asociado a transformadores de menor consumo observado.

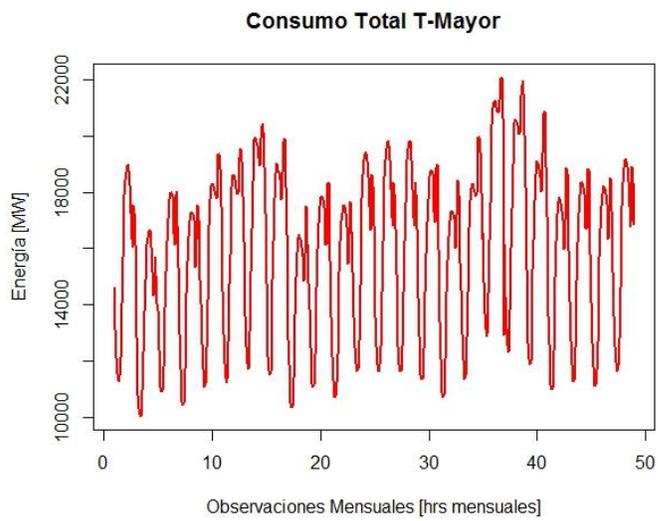


Figura 20: Serie de consumo total asociado a transformadores de mayor consumo observado.

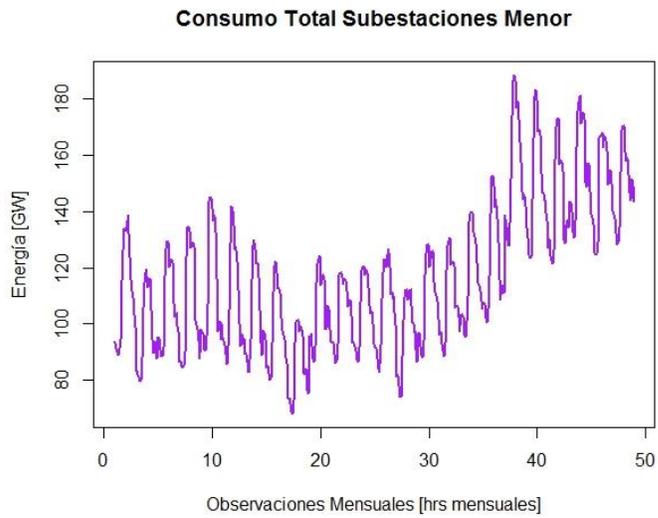


Figura 24: Serie de consumo total asociado a transformadores de mayor consumo observado.

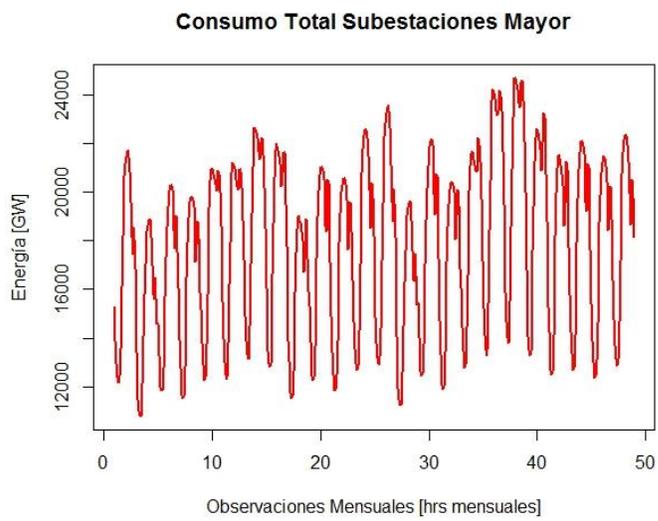


Figura 27: Serie de consumo total asociado a transformadores de mayor consumo observado.

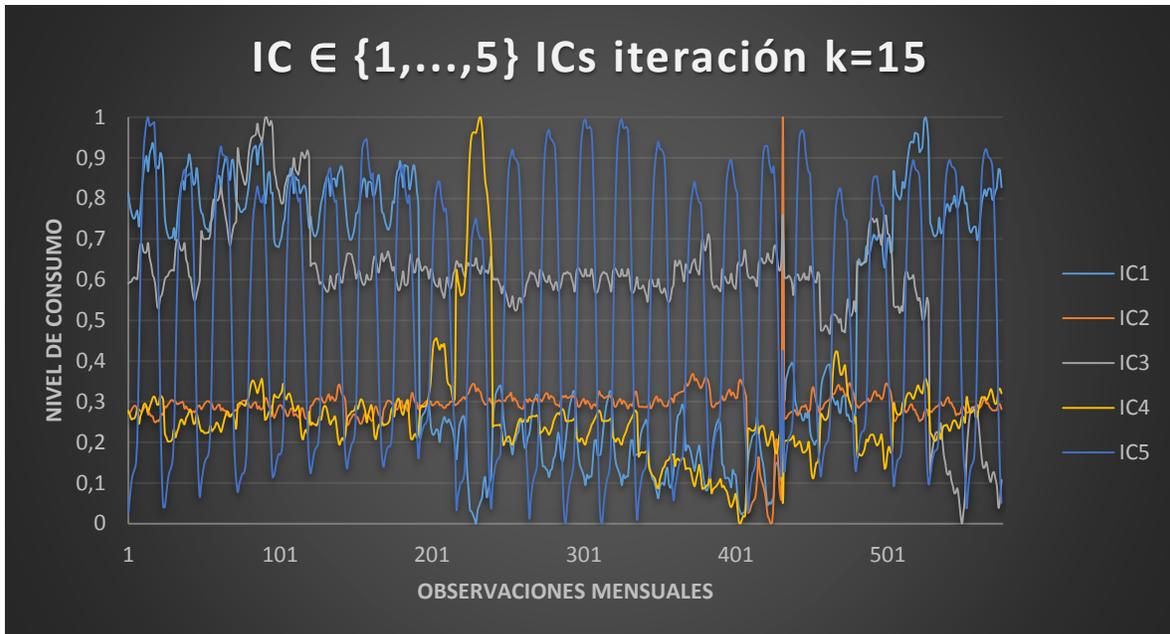


Figura 36: Visualización primeros 5 componentes independientes Experimento k=15.

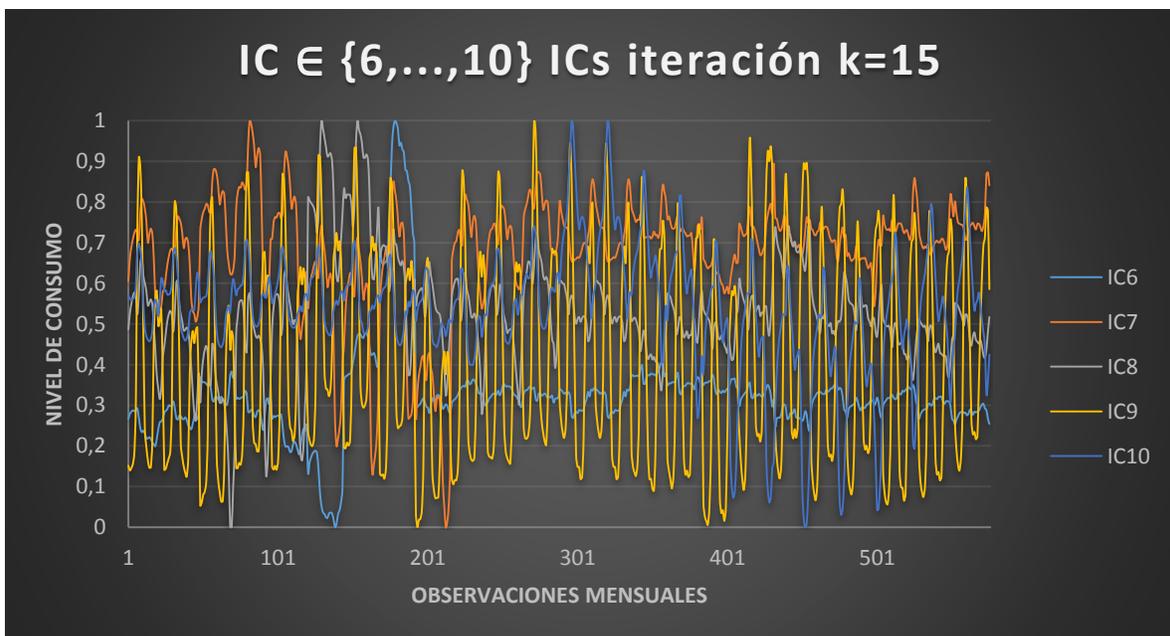


Figura 37: Visualización siguientes 5 componentes independientes Experimento k=15.

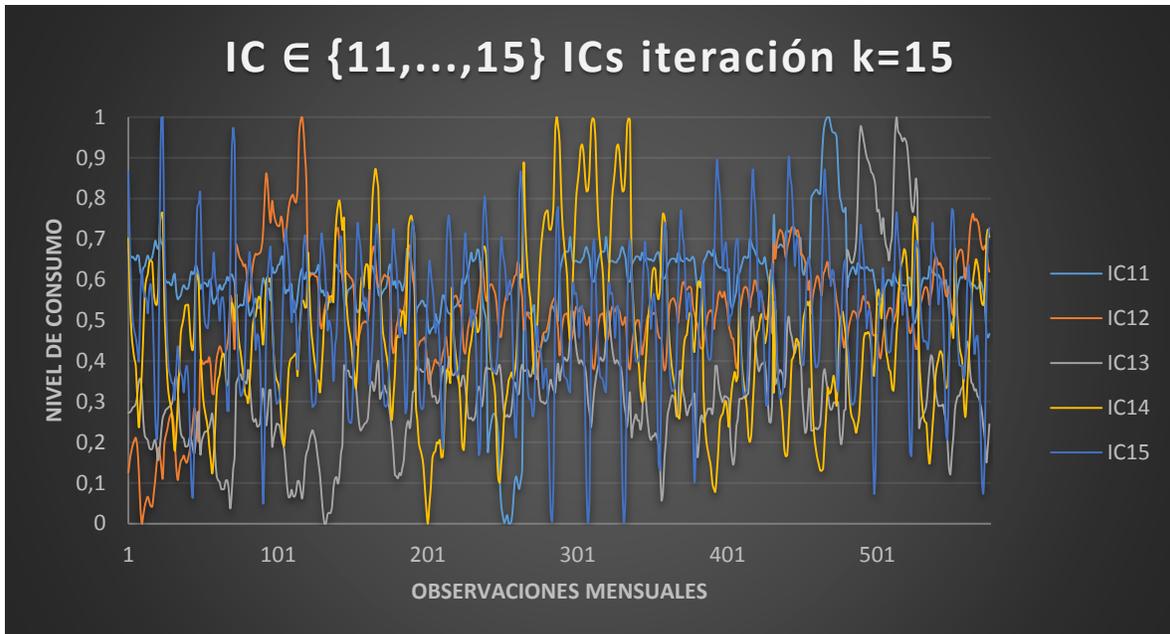


Figura 38: Visualización siguientes 5 componentes independientes Experimento k=15.

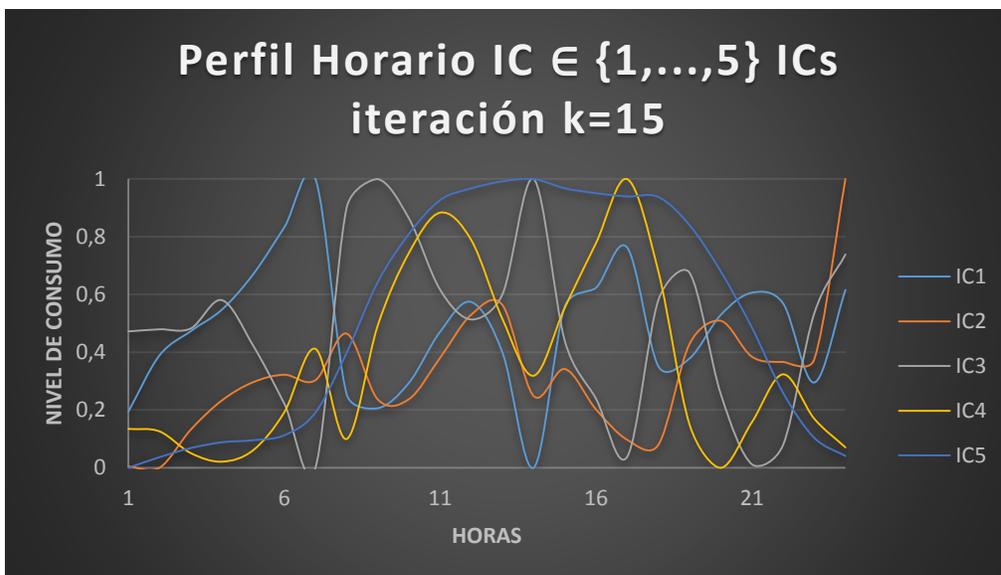


Figura 39: Visualización primeros 5 perfiles de carga asociados a los componentes independientes Experimento k=15.

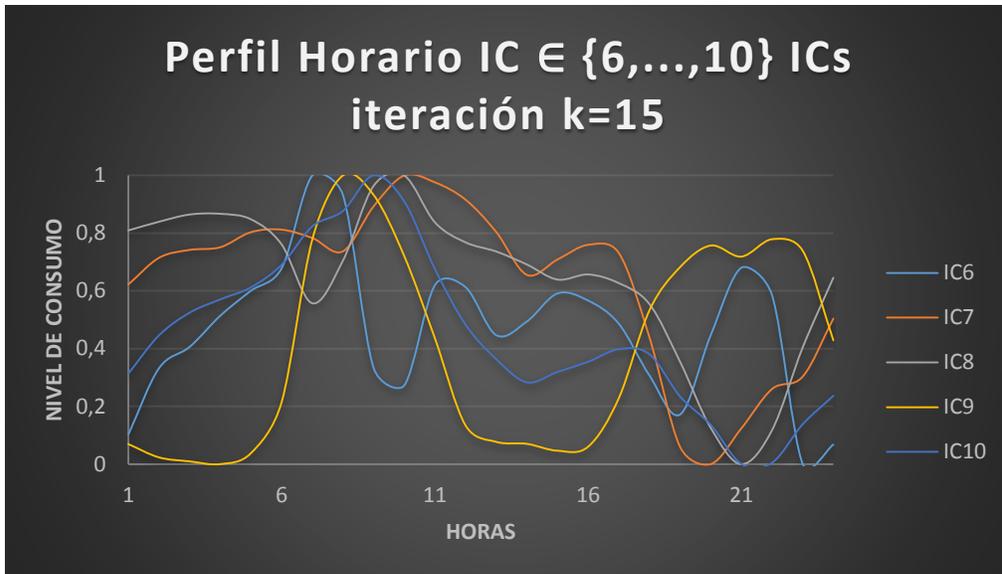


Figura 40: Visualización siguientes 5 perfiles de carga asociados a los componentes independientes Experimento k=15.

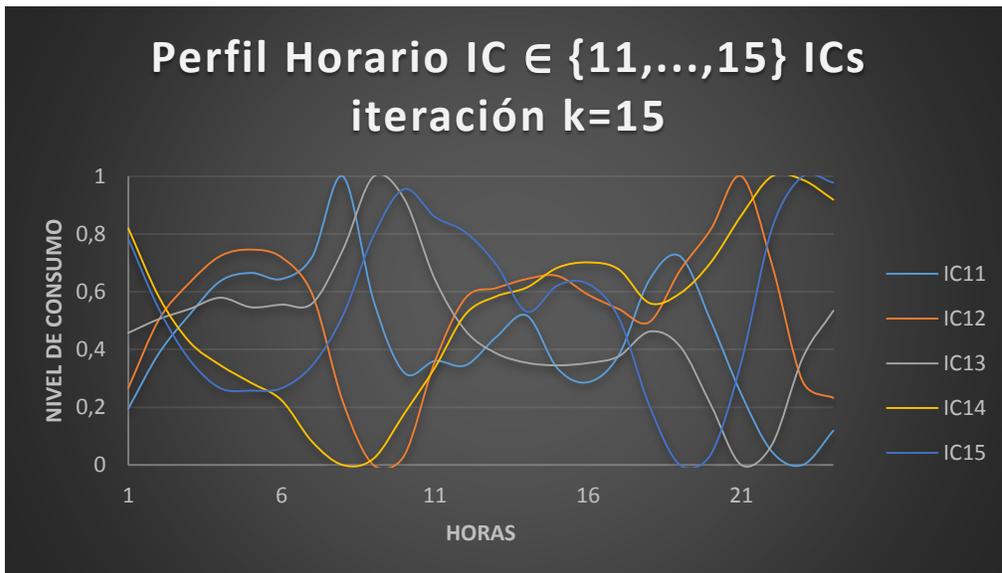


Figura 41: Visualización siguientes 5 perfiles de carga asociados a los componentes independientes Experimento k=15.

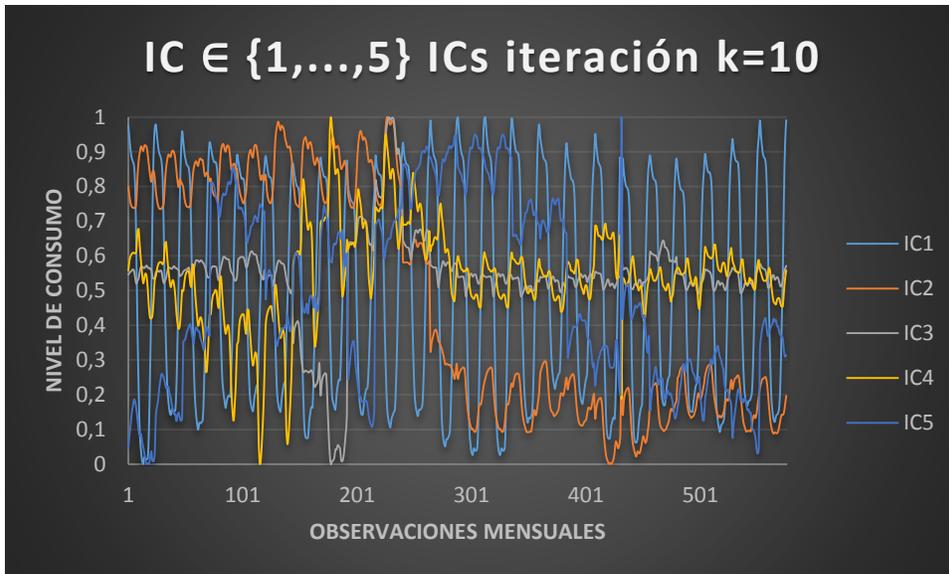


Figura 42: Visualización primeros 5 componentes independientes Experimento k=10.

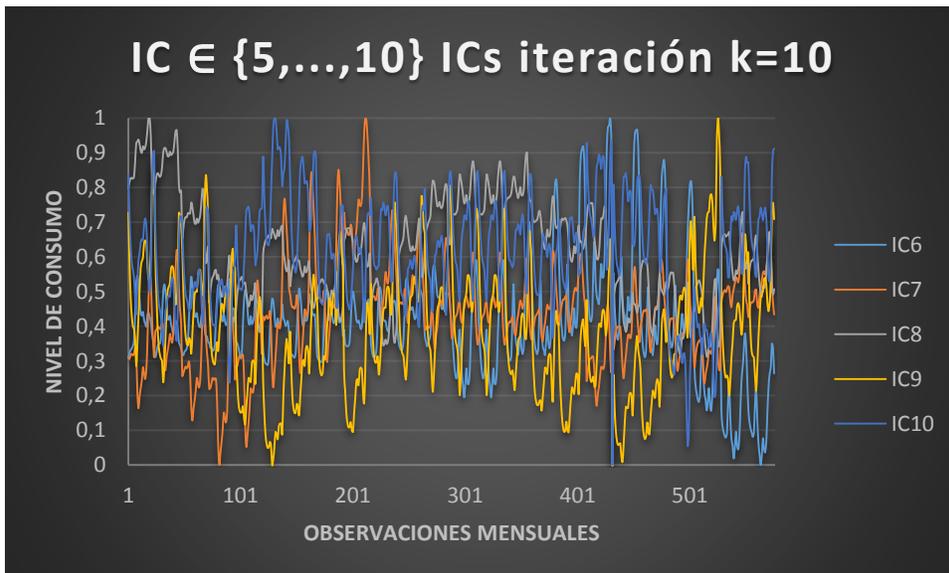


Figura 43: Visualización siguientes 5 componentes independientes Experimento k=10.

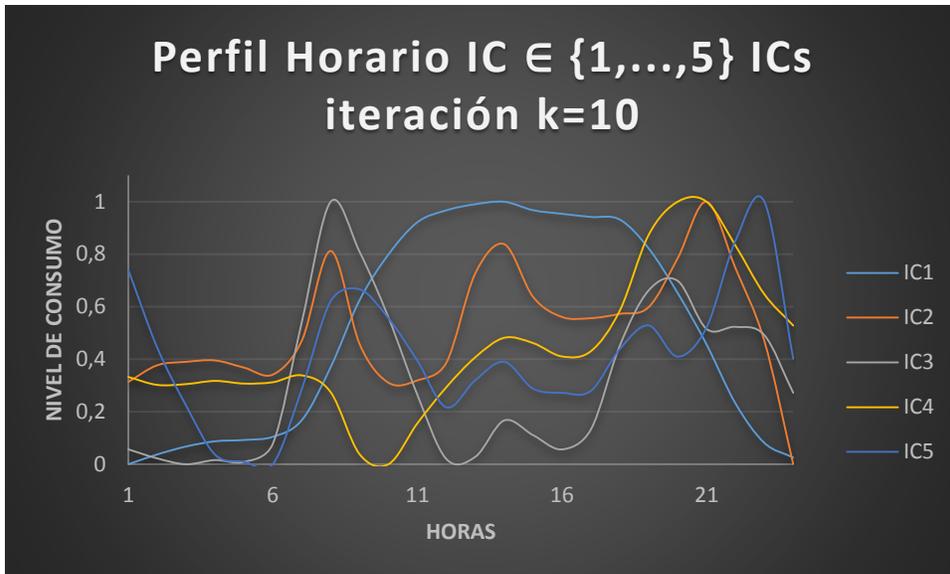


Figura 44: Visualización primeros 5 perfiles de carga asociados a los componentes independientes Experimento k=10.

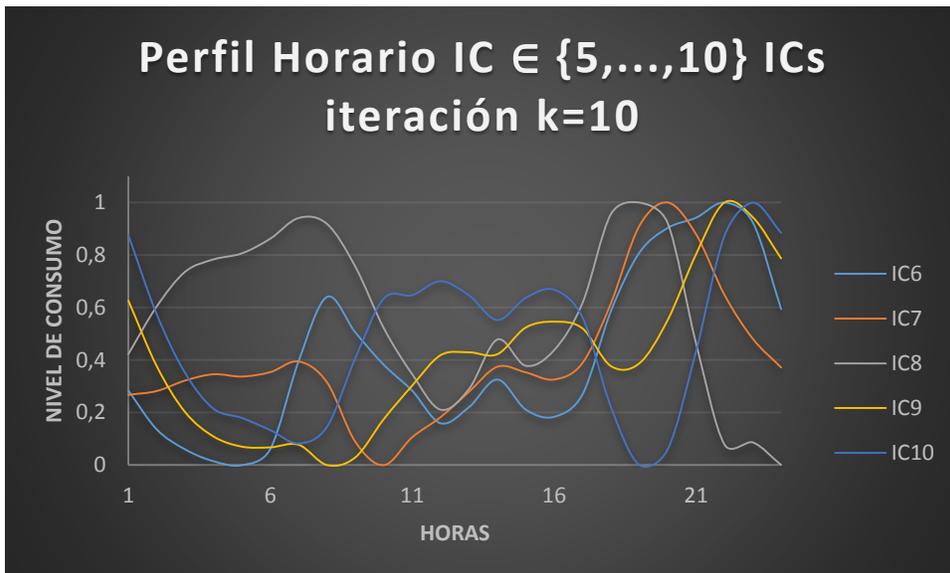


Figura 45: Visualización siguientes 5 perfiles de carga asociados a los componentes independientes Experimento k=10.

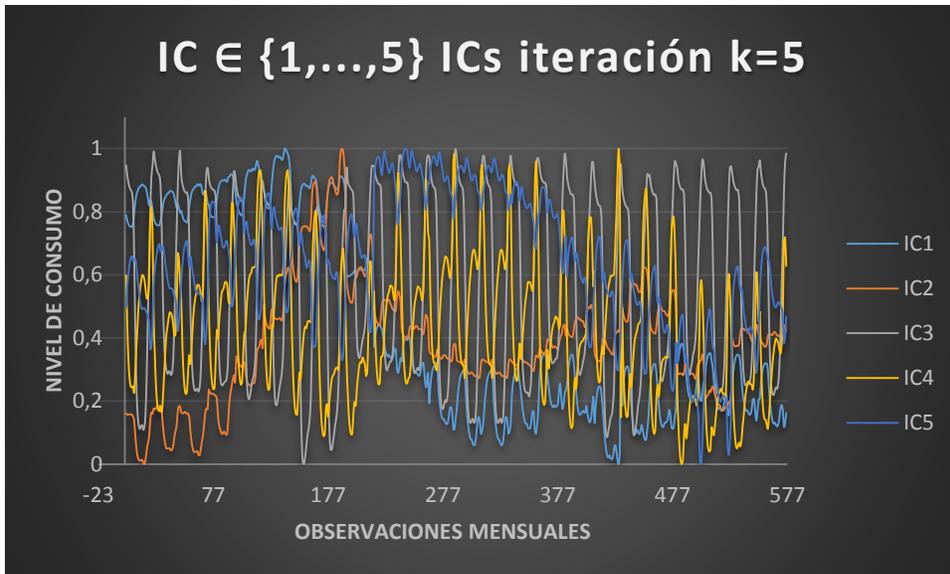


Figura 46: Visualización primeros 5 componentes independientes Experimento k=5.

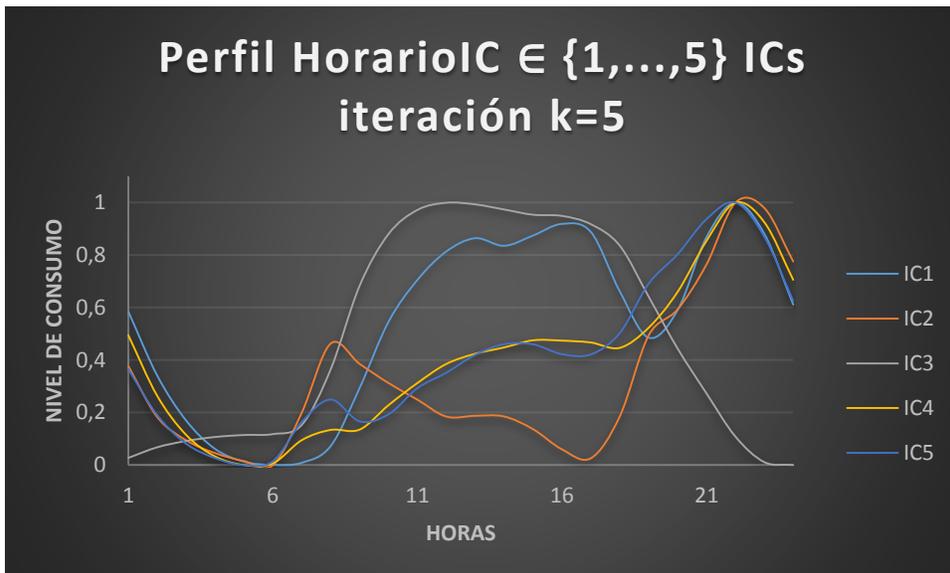


Figura 47: Visualización primeros 5 perfiles de carga asociados a los componentes independientes Experimento k=5.

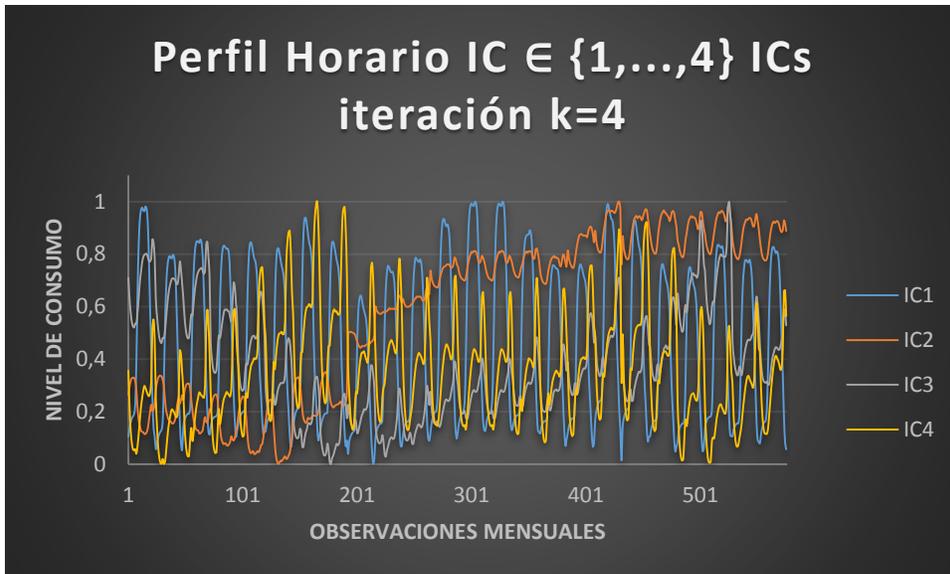


Figura 48: Visualización primeros 5 componentes independientes Experimento k=4.

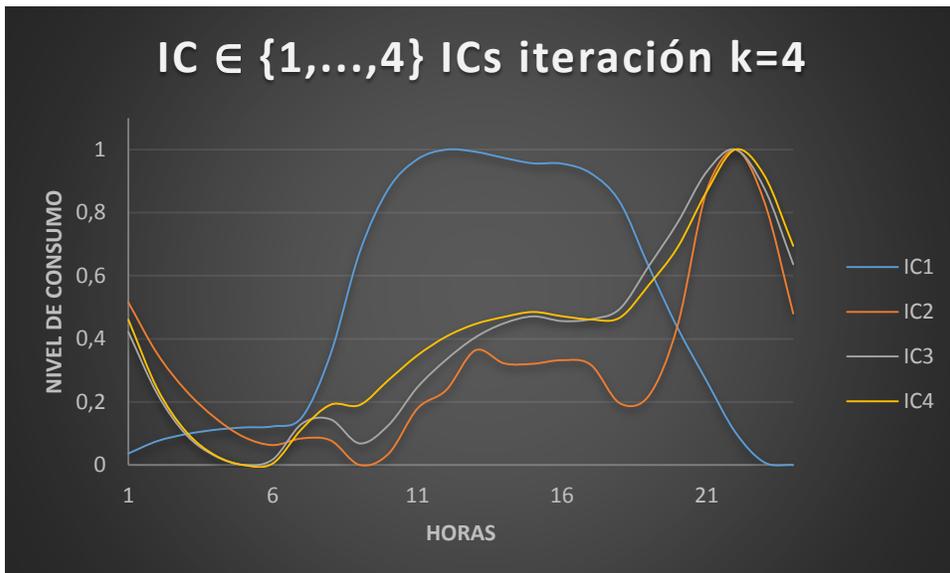


Figura 49: Visualización primeros 4 perfiles de carga asociados a los componentes independientes Experimento k=4.

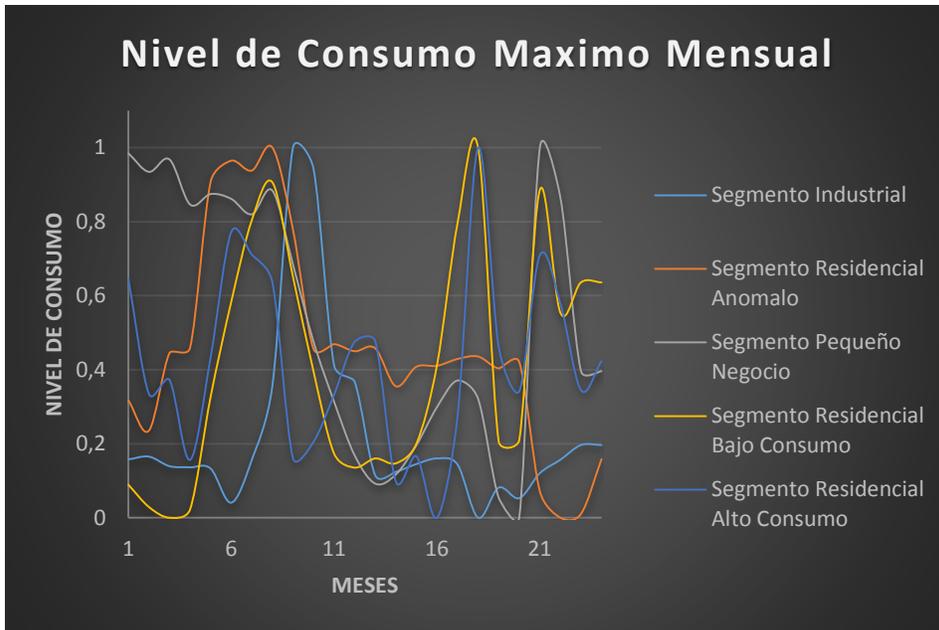


Figura 51: Mximo Nivel de Consumo Mensual para los distintos segmento encontrados.

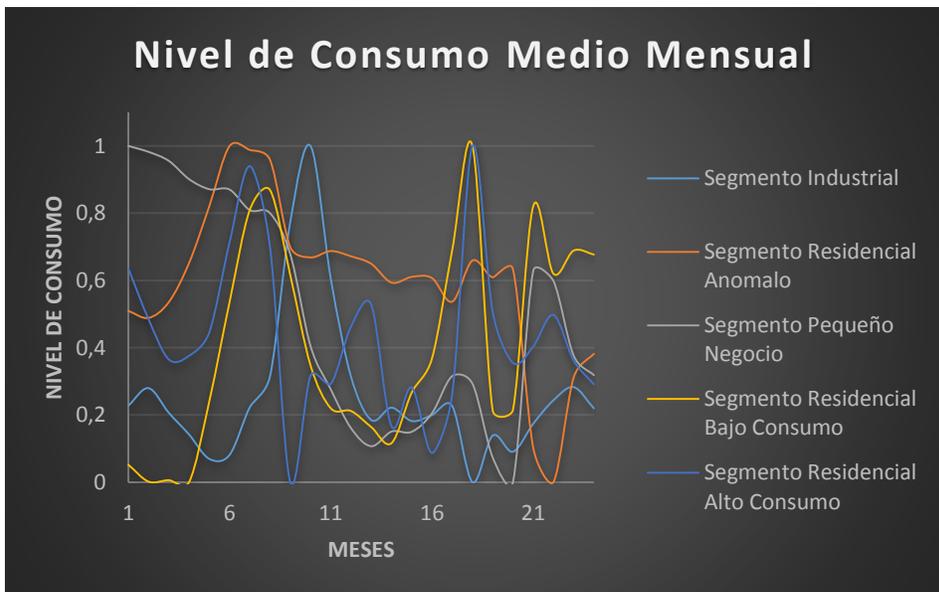


Figura 52: Nivel de Consumo Medio Mensual para los distintos segmento encontrados.

Factor de carga mensual

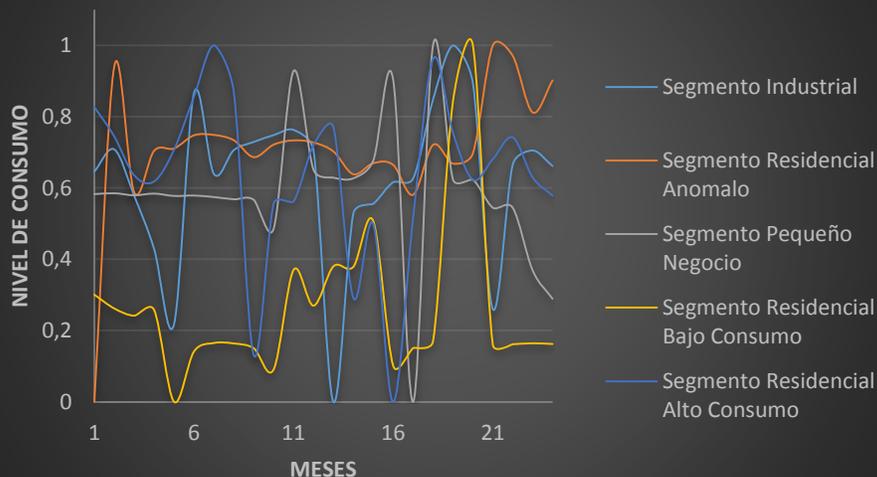


Figura 53: Factor de Carga Mensual para los distintos segmento encontrados.

	<u>Ind</u>	<u>Res_A</u>	<u>Peq_Neg</u>	<u>Res_BC</u>	<u>Res_AC</u>
ALONSO DE CORDOVA T1	27,9%	28,7%	3,1%	30,8%	9,5%
ALONSO DE CORDOVA T2	28,3%	20,2%	8,8%	25,4%	17,3%
ALONSO DE CORDOVA T3	16,7%	24,3%	18,4%	11,5%	29,2%
ALONSO DE CORDOVA T4	20,7%	14,3%	35,2%	5,4%	24,3%
ALTAMIRANO T1	19,7%	18,7%	25,5%	16,4%	19,6%
ALTAMIRANO T2	21,2%	17,3%	24,6%	19,4%	17,6%
ANDES T1	20,1%	18,1%	29,9%	15,1%	16,9%
ANDES T2	21,4%	17,1%	30,0%	14,3%	17,2%
APOQUINDO T1	20,4%	25,2%	18,1%	27,3%	9,0%
APOQUINDO T2	20,9%	23,2%	22,2%	24,0%	9,7%
APOQUINDO T3	20,4%	21,8%	26,3%	17,4%	14,1%

Tabla 5: Market-Share por segmento para los primeros 11 transformadores.

	<u>Ind</u>	<u>Res_A</u>	<u>Peq_Neg</u>	<u>Res_BC</u>	<u>Res_AC</u>
BATUCO-1	20,7%	15,1%	32,3%	8,7%	23,2%
BATUCO-2	17,3%	18,7%	27,0%	16,4%	20,6%
BATUCO-3	21,5%	14,5%	23,0%	17,7%	23,2%
BRASIL T1	24,2%	22,8%	11,0%	21,0%	20,9%
BRASIL T2	25,5%	19,0%	10,1%	21,2%	24,2%
BRASIL T3	20,7%	14,3%	35,2%	5,4%	24,3%
CARRASCAL T1	21,0%	17,9%	28,7%	11,7%	20,6%
CARRASCAL T2	21,8%	16,0%	29,2%	12,6%	20,4%
CARRASCAL T3	20,7%	14,2%	35,1%	5,4%	24,5%
CHACABUCO T1	24,2%	17,4%	26,3%	8,2%	23,9%
CHACABUCO T2	27,3%	16,4%	26,4%	7,6%	22,4%

Tabla 6: Market-Share por segmento para los siguientes 11 transformadores.

	<u>Ind</u>	<u>Res A</u>	<u>Peq Neg</u>	<u>Res BC</u>	<u>Res AC</u>
CHACABUCO T3	27,3%	12,6%	28,4%	7,6%	24,1%
CHACABUCO T4	20,8%	14,2%	35,2%	5,4%	24,3%
CHACABUCO T6	20,4%	14,4%	34,4%	5,4%	25,4%
CHACABUCO T8	6,7%	20,4%	38,7%	7,3%	27,0%
CISTERNA T1	19,6%	18,8%	30,5%	17,6%	13,6%
CISTERNA T2	20,7%	19,0%	30,0%	17,0%	13,3%
CISTERNA T3	17,5%	20,8%	28,7%	23,8%	9,2%
CLUB HIPICO T1	21,9%	16,9%	26,8%	10,8%	23,7%
CLUB HIPICO T2	20,7%	14,3%	35,2%	5,4%	24,3%
CLUB HIPICO T3	19,6%	31,2%	20,8%	22,4%	5,9%
CLUB HIPICO T4	21,7%	17,2%	27,3%	9,4%	24,3%

Tabla 7: Market-Share por segmento para los siguientes 11 transformadores.

	<u>Ind</u>	<u>Res A</u>	<u>Peq Neg</u>	<u>Res BC</u>	<u>Res AC</u>
EL MANZANO T1	18,5%	12,7%	36,3%	6,1%	26,4%
EL MANZANO T3	20,7%	14,3%	35,1%	5,5%	24,5%
LA DEHESA T1	20,3%	18,6%	31,6%	13,3%	16,1%
LA DEHESA T2	21,3%	16,6%	28,9%	11,9%	21,3%
LA DEHESA T3	20,3%	17,0%	33,0%	10,3%	19,4%
LA DEHESA T4	20,7%	14,3%	35,2%	5,4%	24,3%
LA REINA T1	20,7%	14,4%	35,2%	5,4%	24,3%
LA REINA T2	23,4%	19,4%	25,7%	18,6%	12,8%
LA REINA T3	9,1%	27,1%	34,4%	29,3%	0,0%
LA REINA T4	20,5%	22,7%	30,3%	24,1%	2,3%
LO AGUIRRE T1	20,4%	14,3%	31,9%	8,3%	25,0%

Tabla 8: Market-Share por segmento para los siguientes 11 transformadores.

	<u>Ind</u>	<u>Res A</u>	<u>Peq Neg</u>	<u>Res BC</u>	<u>Res AC</u>
LO BOZA T1	23,2%	16,0%	25,2%	7,3%	28,4%
LO BOZA T2	22,0%	13,6%	27,4%	7,4%	29,5%
LO BOZA T3	16,4%	18,0%	39,9%	8,1%	17,6%
LO BOZA T4	16,5%	6,4%	29,1%	16,1%	31,9%
LORD COCHRANE T1	27,8%	9,7%	10,1%	23,8%	28,6%
LORD COCHRANE T2	28,3%	24,0%	3,1%	15,8%	28,8%
LORD COCHRANE T3	19,1%	16,7%	32,9%	5,1%	26,1%
LOS DOMINICOS T1	20,8%	14,3%	35,1%	5,5%	24,4%
LOS DOMINICOS T2	24,9%	17,9%	28,6%	13,4%	15,2%
LOS DOMINICOS T3	14,7%	25,8%	21,4%	21,5%	16,6%
MACUL T1	20,8%	14,3%	35,2%	5,4%	24,3%

Tabla 9: Market-Share por segmento para los siguientes 11 transformadores.

	<u>Ind</u>	<u>Res A</u>	<u>Peq Neg</u>	<u>Res BC</u>	<u>Res AC</u>
MACUL T3	20,7%	18,3%	19,8%	26,1%	15,2%
MACUL T4	17,3%	22,2%	26,7%	31,6%	2,3%
MACUL T6	21,5%	16,1%	31,1%	10,2%	21,2%
MAIPU T1	21,6%	14,8%	31,4%	6,7%	25,5%
MAIPU T2	20,4%	15,1%	27,5%	12,5%	24,5%
MAIPU T4	21,4%	17,2%	25,9%	9,0%	26,5%
OCHAGAVIA T1	20,6%	14,4%	34,9%	5,4%	24,6%
OCHAGAVIA T2	22,9%	21,9%	23,0%	24,5%	7,7%
PAJARITOS T1	17,0%	20,1%	26,9%	35,3%	0,8%
PAJARITOS T2	17,6%	18,4%	29,1%	26,3%	8,6%
PAJARITOS T3	19,2%	16,5%	33,2%	14,1%	17,0%

Tabla 10: Market-Share por segmento para los siguientes 11 transformadores.

	<u>Ind</u>	<u>Res A</u>	<u>Peq Neg</u>	<u>Res BC</u>	<u>Res AC</u>
PAJARITOS T4	20,8%	14,3%	35,2%	5,4%	24,3%
POLPAICO T1	20,1%	11,7%	33,8%	7,3%	27,1%
PUDAHUEL T1	18,1%	15,4%	31,5%	18,4%	16,5%
PUDAHUEL T2	18,7%	16,3%	32,9%	18,2%	13,9%
PUDAHUEL T3	21,2%	15,0%	38,6%	4,2%	21,0%
QUILICURA T1	20,1%	14,2%	34,9%	8,8%	22,0%
QUILICURA T2	20,7%	17,8%	24,5%	18,9%	18,1%
QUILICURA T3	21,6%	16,7%	28,4%	9,4%	23,9%
QUILICURA T4	20,1%	17,7%	30,0%	11,1%	21,1%
RECOLETA T1	27,5%	12,3%	23,1%	11,3%	25,9%
RECOLETA T2	25,3%	13,8%	32,4%	7,7%	20,9%

Tabla 11: Market-Share por segmento para los siguientes 11 transformadores.

	<u>Ind</u>	<u>Res A</u>	<u>Peq Neg</u>	<u>Res BC</u>	<u>Res AC</u>
RECOLETA T3	0,7%	19,2%	41,1%	13,5%	25,5%
RECOLETA T4	23,7%	23,4%	23,6%	26,5%	2,9%
SAN CRISTOBAL T1	30,2%	13,4%	11,5%	19,5%	25,5%
SAN CRISTOBAL T2	26,2%	14,8%	5,7%	18,6%	34,7%
SAN CRISTOBAL T3	21,6%	17,0%	21,7%	25,9%	13,7%
SAN CRISTOBAL T4	30,6%	10,9%	24,9%	10,7%	23,0%
SAN CRISTOBAL T5	0,0%	43,4%	31,3%	4,6%	20,7%
SAN JOAQUIN T1	24,9%	20,6%	13,7%	19,0%	21,8%
SAN JOAQUIN T2	20,6%	14,9%	32,7%	10,8%	21,0%
SAN JOAQUIN T3	20,3%	19,7%	19,9%	21,2%	18,8%
SAN JOAQUIN T4	22,8%	15,7%	28,9%	7,7%	25,0%

Tabla 12: Market-Share por segmento para los siguientes 11 transformadores.

	<u>Ind</u>	<u>Res A</u>	<u>Peq Neg</u>	<u>Res BC</u>	<u>Res AC</u>
SAN JOSE T1	20,3%	16,7%	32,3%	11,9%	18,8%
SAN JOSE T2	17,9%	20,1%	31,5%	13,9%	16,7%
SAN JOSE T3	19,8%	15,9%	33,9%	11,0%	19,4%
SAN JOSE T4	13,9%	24,0%	31,2%	22,4%	8,6%
SAN PABLO T1	26,6%	11,7%	57,7%	0,0%	4,0%
SAN PABLO T2	16,4%	17,8%	21,3%	13,9%	30,6%
SANTA ELENA T1	21,0%	14,3%	31,5%	6,6%	26,5%
SANTA ELENA T2	12,5%	31,6%	13,4%	22,2%	20,2%
SANTA ELENA T3	36,0%	0,0%	48,2%	4,6%	11,2%
SANTA ELENA T4	20,4%	26,3%	12,3%	26,2%	14,8%
SANTA MARTA T1	18,4%	17,8%	32,1%	21,0%	10,7%

Tabla 13: Market-Share por segmento para los siguientes 11 transformadores.

	<u>Ind</u>	<u>Res A</u>	<u>Peq Neg</u>	<u>Res BC</u>	<u>Res AC</u>
SANTA MARTA T2	18,2%	19,0%	31,6%	14,1%	17,1%
SANTA MARTA T3	20,7%	14,3%	35,2%	5,5%	24,2%
SANTA MARTA T4	21,0%	12,6%	39,0%	8,3%	19,1%
SANTA RAQUEL T1	17,6%	22,0%	31,8%	26,6%	2,0%
SANTA RAQUEL T2	17,8%	16,6%	32,0%	18,7%	15,0%
SANTA RAQUEL T4	19,5%	16,1%	35,8%	11,9%	16,7%
SANTA RAQUEL T5	21,3%	16,1%	31,7%	9,1%	21,9%
VITACURA T1	23,8%	22,1%	4,1%	24,5%	25,4%
VITACURA T2	27,9%	25,5%	10,4%	14,0%	22,2%
VITACURA T3	29,2%	22,7%	0,0%	22,3%	25,8%
VITACURA T4	27,2%	9,2%	7,1%	28,4%	28,1%

Tabla 14: Market-Share por segmento para los siguientes 11 transformadores.

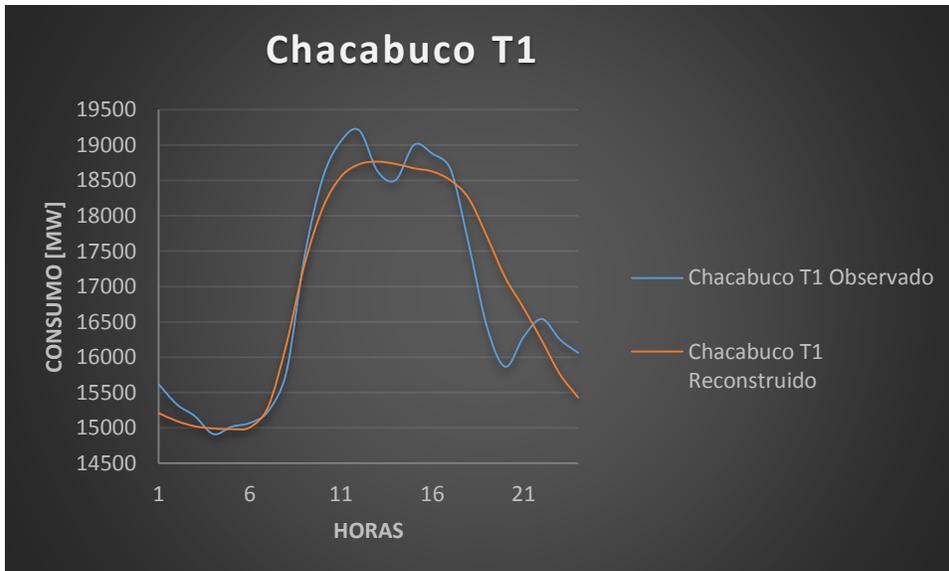


Figura 59: Perfil observado vs. Perfil reconstruido Chacabuco T1.

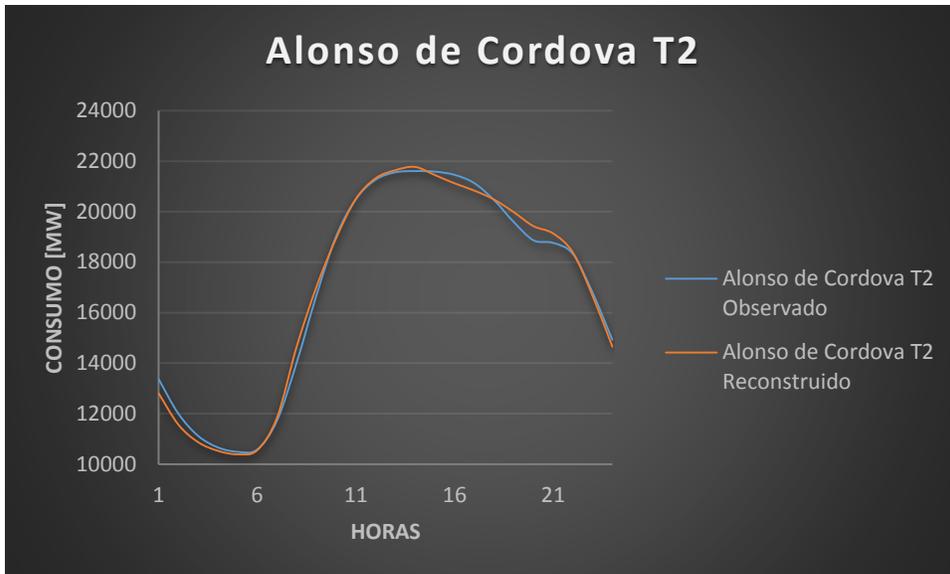


Figura 60: Perfil observado vs. Perfil reconstruido Alonso de Cordova T2.

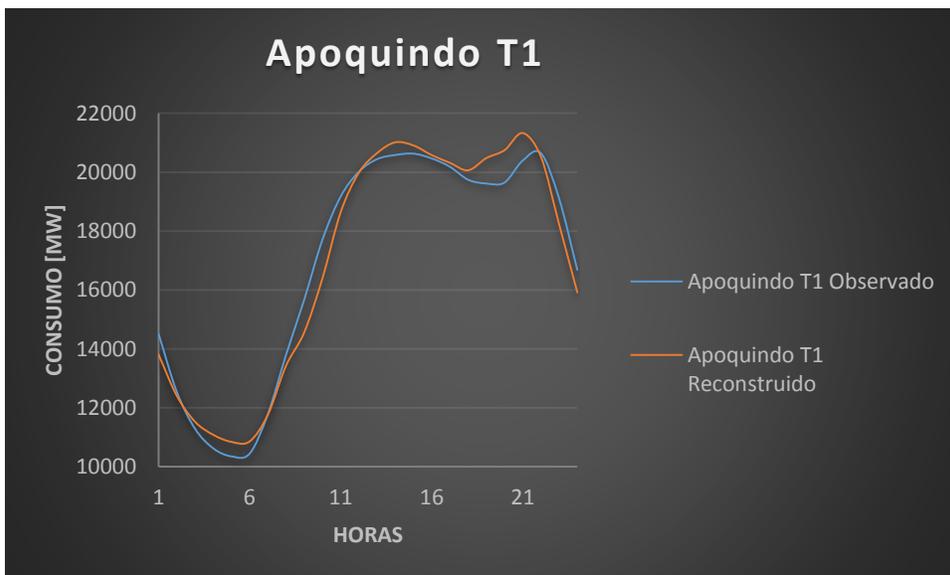


Figura 61: Perfil observado vs. Perfil reconstruido Apoquindo T1.

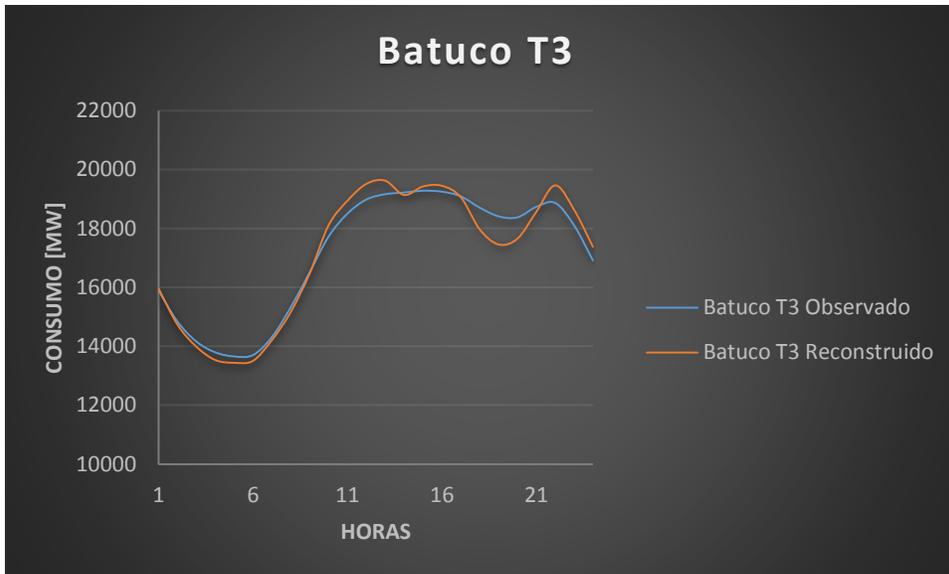


Figura 62: Perfil observado vs. Perfil reconstruido Batuco T3.

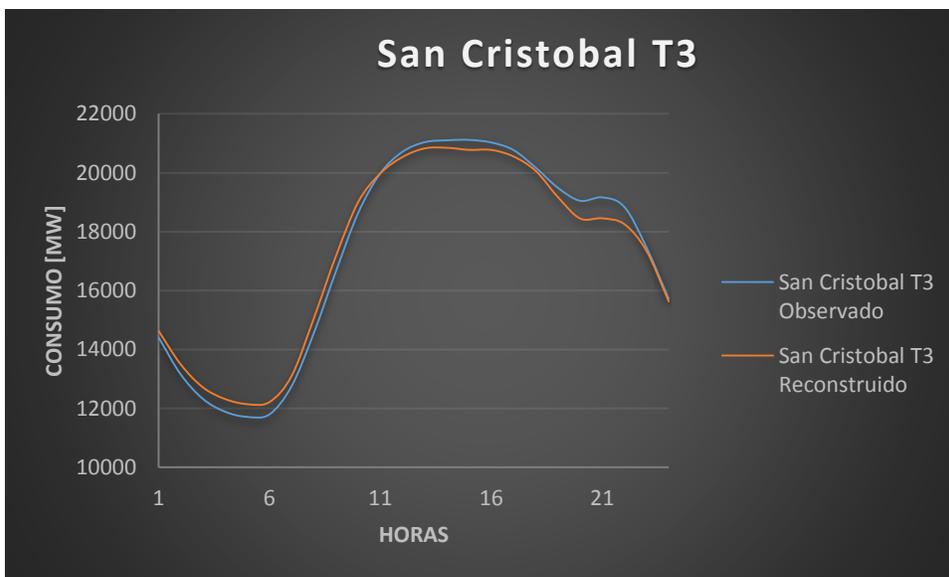


Figura 63: Perfil observado vs. Perfil reconstruido San Cristóbal T3.

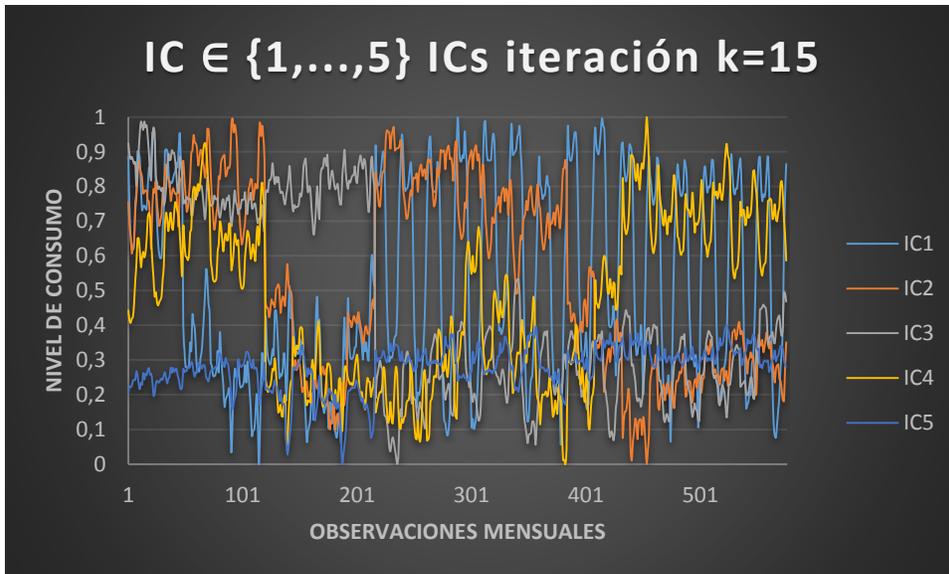


Figura 65: Visualización primeros 5 componentes independientes Experimento k=15.

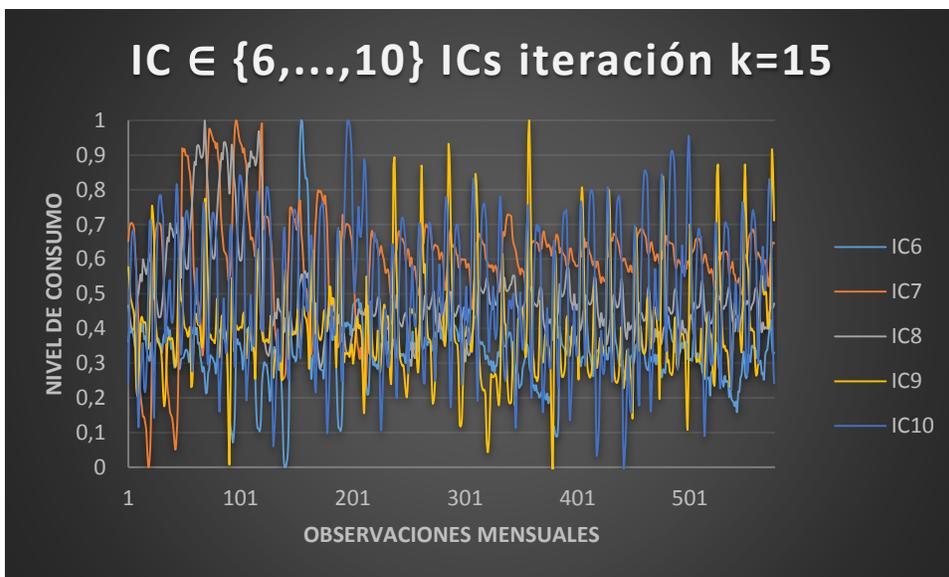


Figura 66: Visualización siguientes 5 componentes independientes Experimento k=15.

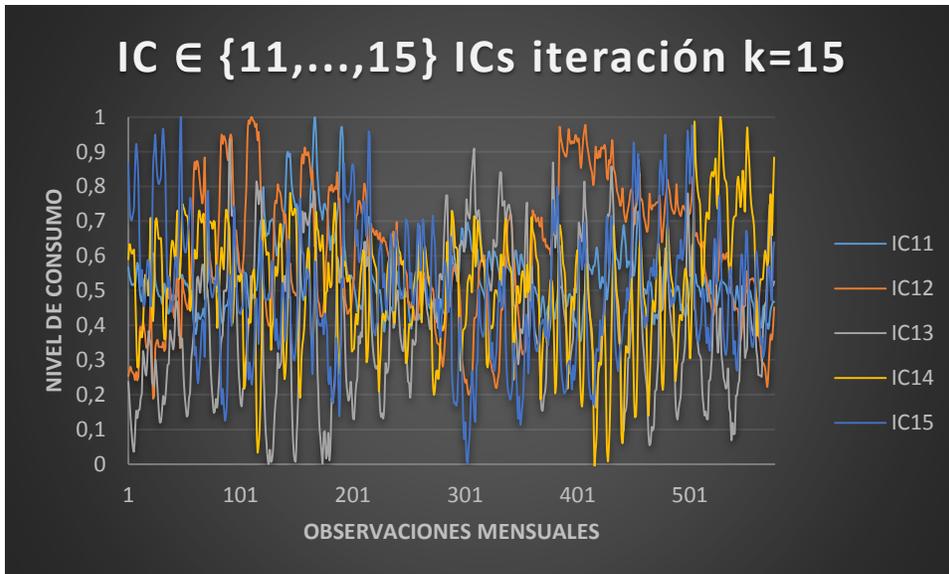


Figura 67: Visualización siguientes 5 componentes independientes Experimento k=15.



Figura 68: Visualización primeros 5 perfiles de carga asociados a los componentes independientes Experimento k=15.



Figura 69: Visualización siguientes 5 perfiles de carga asociados a los componentes independientes Experimento k=15.

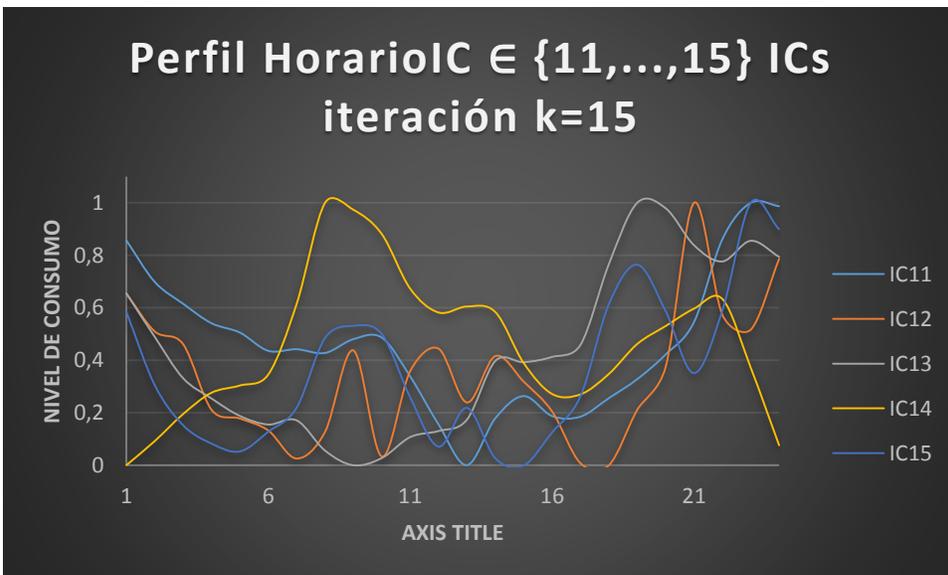


Figura 70: Visualización siguientes 5 perfiles de carga asociados a los componentes independientes Experimento k=15.

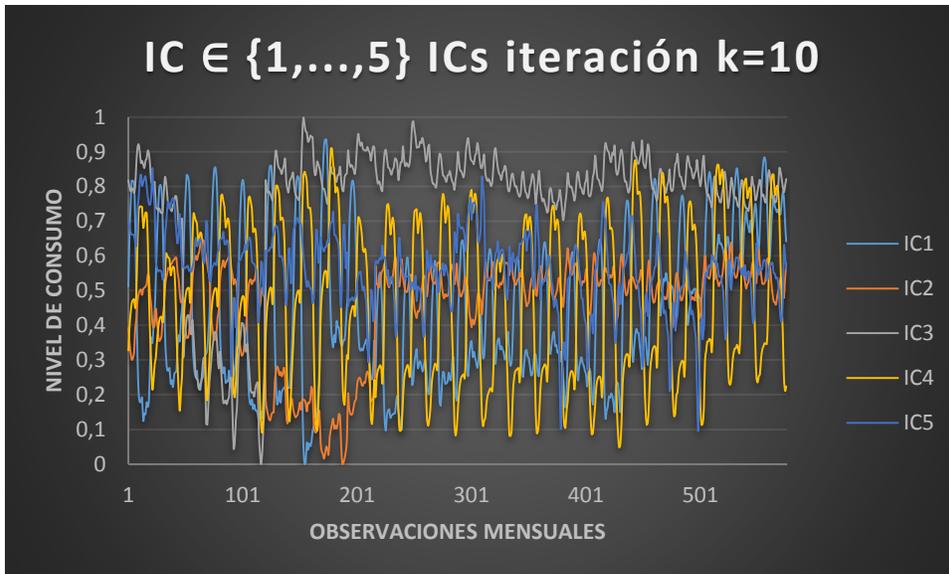


Figura 71: Visualización primeros 5 componentes independientes Experimento k=10.

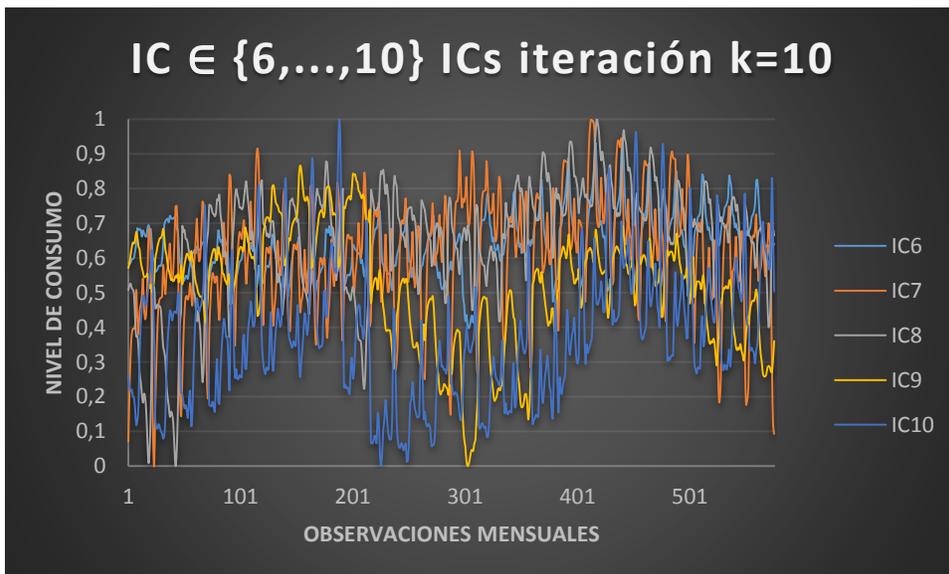


Figura 72: Visualización siguientes 5 componentes independientes Experimento k=10.

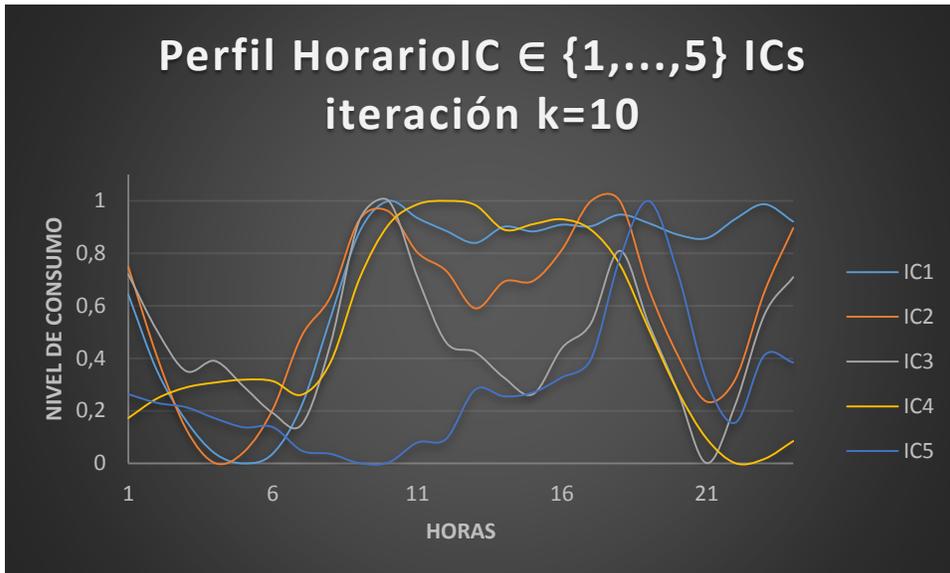


Figura 73: Visualización primeros 5 perfiles de carga asociados a los componentes independientes Experimento k=10.

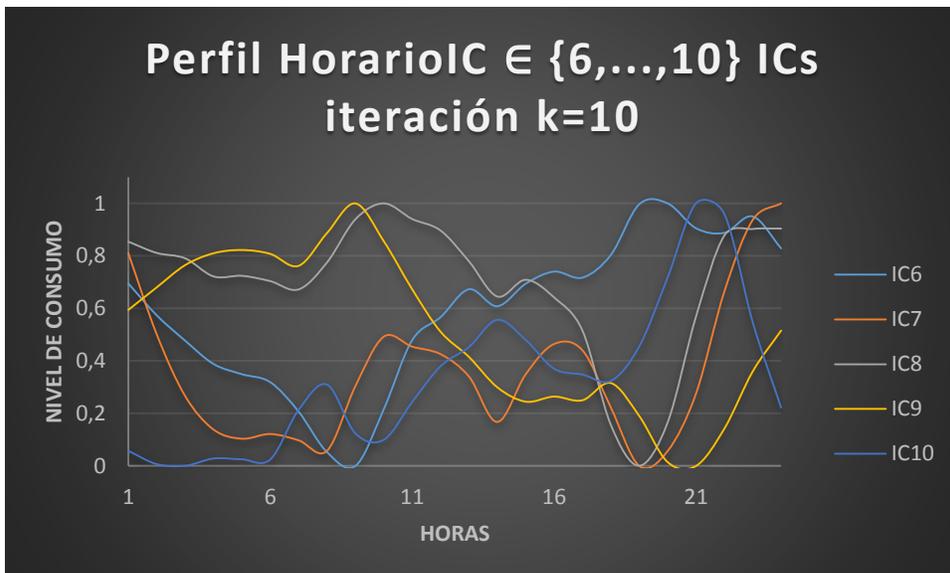


Figura 74: Visualización siguientes 5 perfiles de carga asociados a los componentes independientes Experimento k=10.

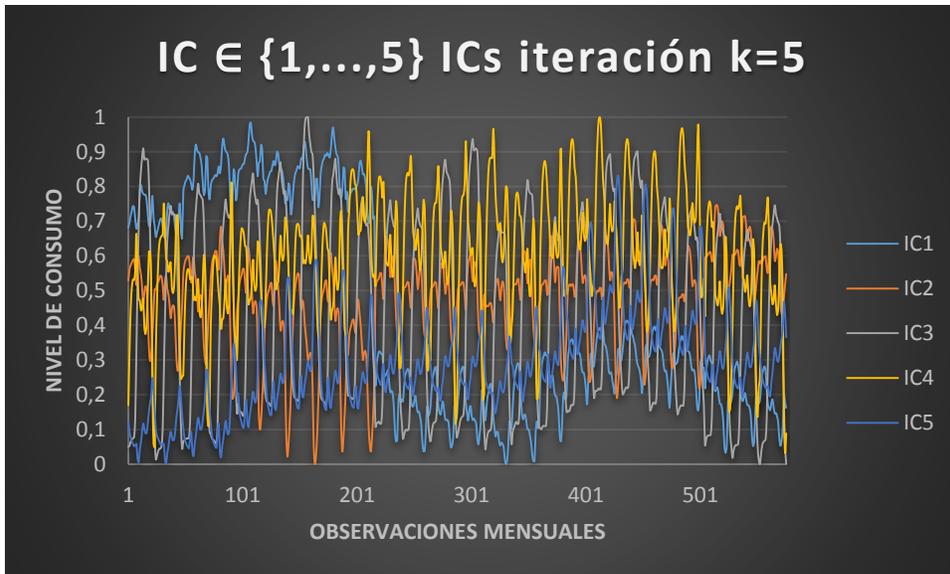


Figura 75: Visualización primeros 5 componentes independientes Experimento k=5.

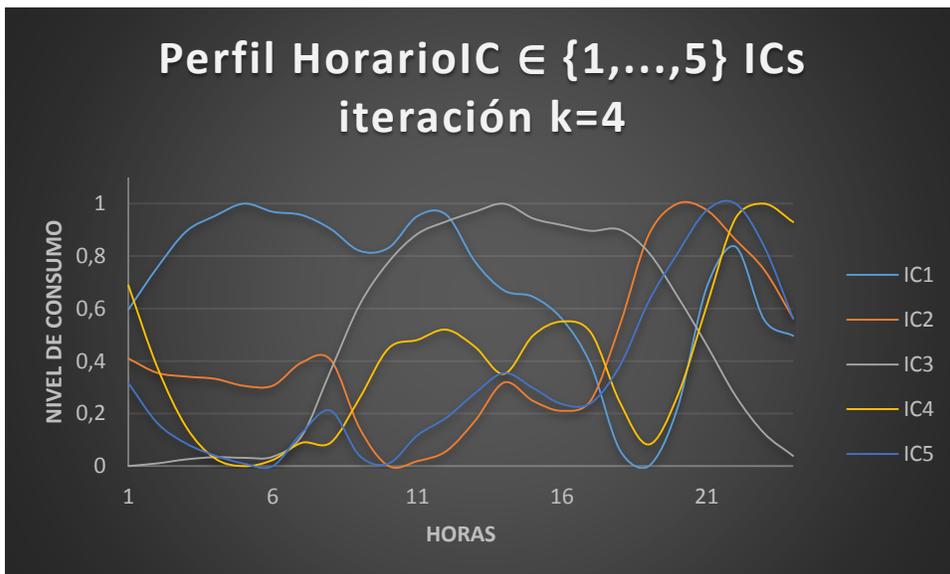


Figura 76: Visualización 5 perfiles de carga asociados a los componentes independientes Experimento k=10.

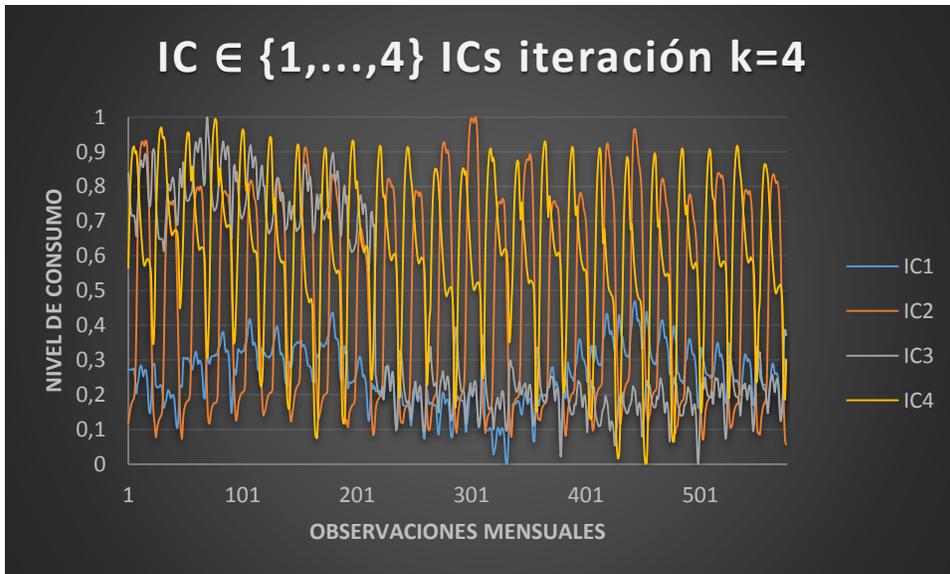


Figura 77: Visualización 4 componentes independientes Experimento k=4.

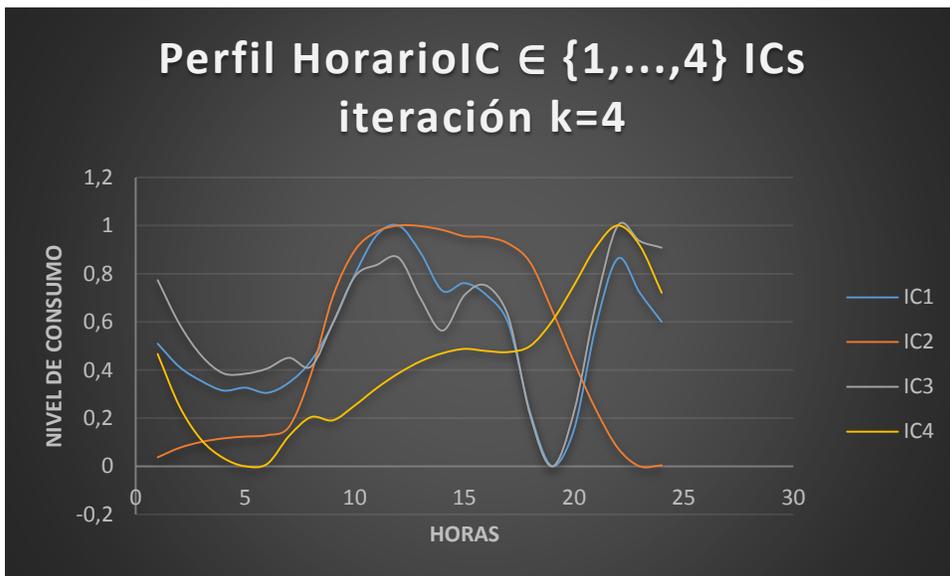


Figura 78: Visualización 4 perfiles de carga asociados a los componentes independientes Experimento k=4.

	Ind	Res_AC	Peq_Neg	Res_A	Res_BC
ALONSO DE CORDOVA	30,6%	7,3%	15,8%	26,7%	19,5%
ALTAMIRANO	20,0%	16,0%	29,2%	19,9%	14,8%
ALTO JAHUEL	17,5%	12,9%	32,3%	31,9%	5,5%
ANDES	19,5%	13,5%	29,3%	24,4%	13,2%
APOQUINDO	11,8%	20,4%	18,9%	17,8%	31,0%
BATUCO	23,8%	33,7%	25,1%	11,1%	6,3%
BICENTENARIO	10,6%	10,9%	35,1%	35,4%	8,0%
BRASIL	17,3%	9,6%	24,9%	31,8%	16,5%
BUIN CGE	17,5%	12,9%	32,3%	31,9%	5,5%
CALEU	17,4%	12,8%	32,4%	31,9%	5,5%
CARRASCAL	17,5%	12,3%	30,6%	28,3%	11,2%

Tabla 19: Market-share por segmento para las primeras 11 subestaciones.

	Ind	Res_AC	Peq_Neg	Res_A	Res_BC
CHACABUCO	0,0%	16,0%	35,2%	33,5%	15,3%
CHICUREO	14,3%	14,3%	32,1%	32,1%	7,3%
CISTERNA	28,7%	9,6%	28,2%	9,6%	23,9%
CLUB HIPICO	0,2%	9,7%	29,1%	32,3%	28,8%
EL MANZANO	18,0%	11,9%	33,4%	31,6%	5,1%
ENTEL	17,5%	12,9%	32,3%	31,9%	5,5%
FLORIDA 12	19,3%	13,2%	31,1%	28,7%	7,7%
LA DEHESA	17,8%	13,8%	28,7%	25,3%	14,3%
LA REINA	25,0%	7,4%	25,8%	6,4%	35,5%
LO AGUIRRE	18,2%	12,9%	32,0%	31,1%	5,8%
LO BOZA	9,8%	23,8%	34,5%	24,7%	7,2%

Tabla 20: Market-share por segmento para las primeras 11 subestaciones.

	Ind	Res_AC	Peq_Neg	Res_A	Res_BC
LO PRADO 012	17,4%	12,8%	32,3%	31,9%	5,5%
LO PRADO 044	17,3%	12,9%	32,3%	31,9%	5,5%
LO VALLEDOR	19,2%	25,9%	26,0%	19,7%	9,2%
LORD COCHRANE	25,4%	3,8%	24,8%	38,6%	7,4%
LOS DOMINICOS	17,9%	12,8%	27,2%	25,6%	16,6%
MACUL	28,3%	22,3%	23,2%	11,3%	14,9%
MAIPU	23,2%	29,9%	24,4%	22,4%	0,0%
METRO 110-02	17,4%	12,9%	32,3%	31,9%	5,5%
OCHAGAVIA	19,0%	16,9%	28,6%	23,0%	12,6%
PAJARITOS	36,2%	7,2%	28,0%	0,0%	28,6%
PANAMERICANA	19,1%	12,4%	31,7%	31,9%	4,9%

Tabla 21: Market-share por segmento para las primeras 11 subestaciones.

	Ind	Res_AC	Peq_Neg	Res_A	Res_BC
POLPAICO 23	18,5%	9,7%	34,6%	32,3%	4,9%
PUDAHUEL	26,2%	10,2%	31,0%	17,7%	14,9%
QUILICURA	22,8%	15,6%	27,6%	21,3%	12,7%
RECOLETA	24,1%	6,5%	29,9%	17,7%	21,9%
RUNGUE	16,6%	12,8%	32,6%	32,1%	6,0%
SAN CRISTOBAL	17,1%	0,0%	12,8%	36,7%	33,4%
SAN JOAQUIN	17,3%	27,2%	22,6%	16,0%	16,9%
SAN JOSE	25,8%	5,0%	29,9%	13,5%	25,8%
SAN PABLO	18,6%	27,1%	28,2%	23,9%	2,3%
SANTA ELENA	18,6%	14,3%	22,6%	19,0%	25,4%
SANTA MARTA	21,7%	14,0%	30,5%	17,3%	16,6%
SANTA RAQUEL	29,6%	14,9%	27,2%	8,9%	19,5%
VITACURA	10,0%	20,7%	0,0%	38,5%	30,8%

Tabla 22: Market-share por segmento para las primeras 11 subestaciones.

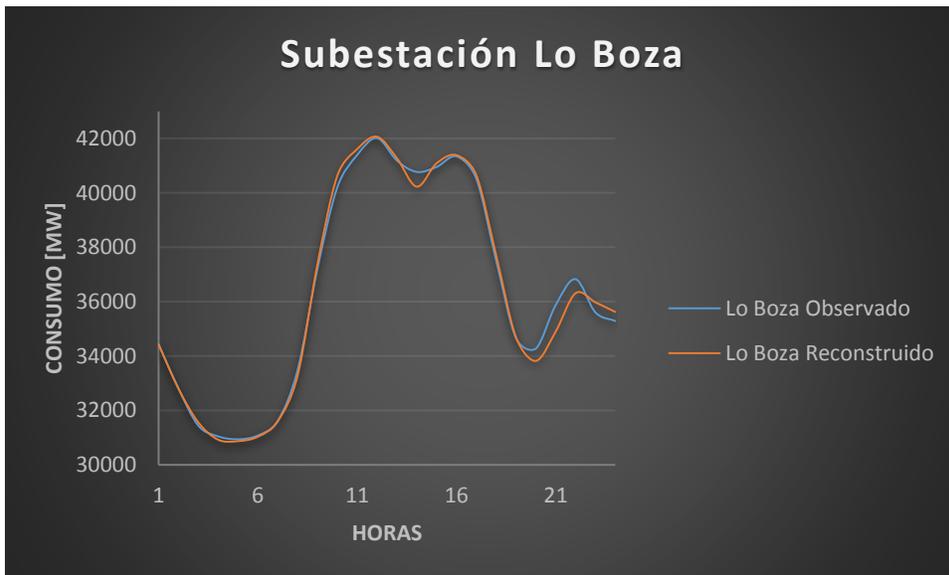


Figura 82: Perfil observado vs. Perfil reconstruido subestación Lo Boza.

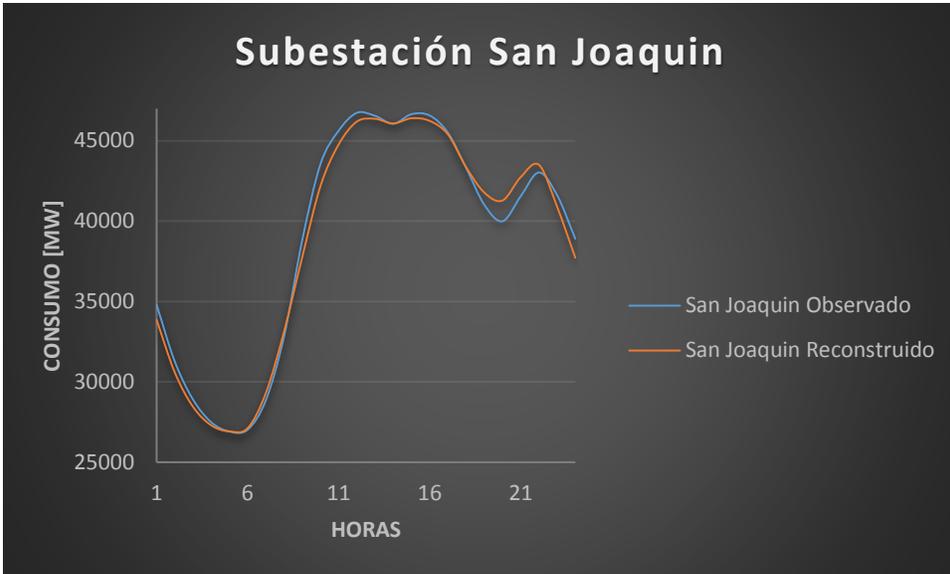


Figura 83: Perfil observado vs. Perfil reconstruido subestación San Joaquín.

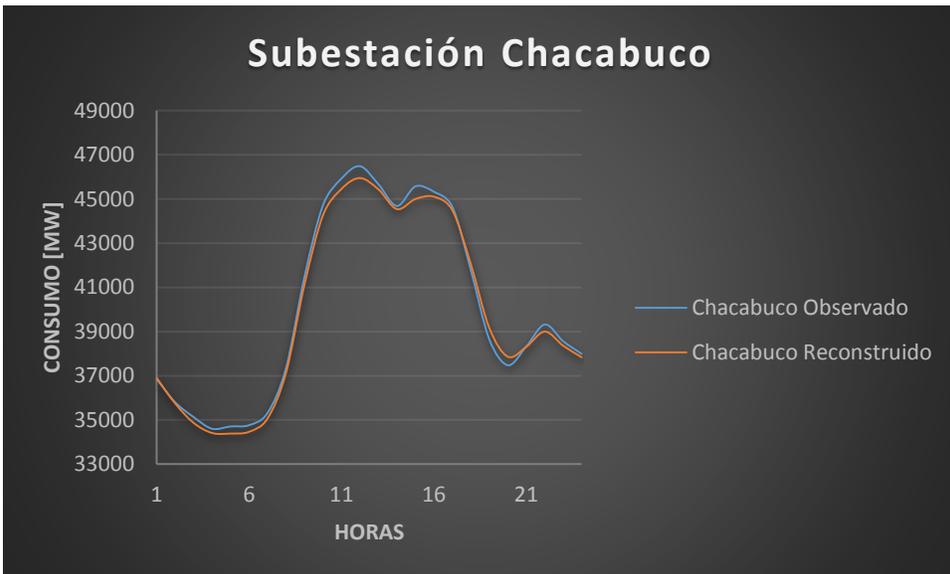


Figura 84: Perfil observado vs. Perfil reconstruido subestación Chacabuco.

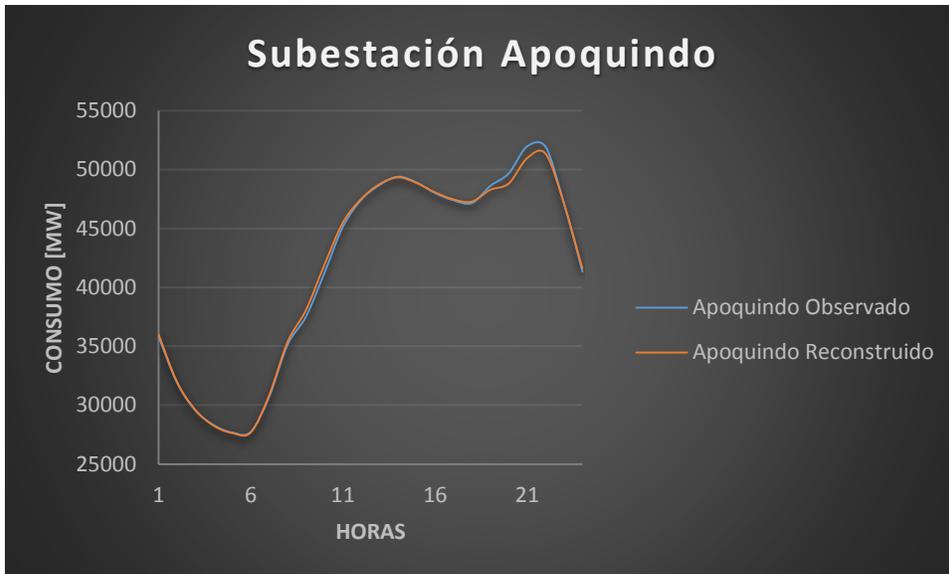


Figura 85: Perfil observado vs. Perfil reconstruido subestación Apoquindo.

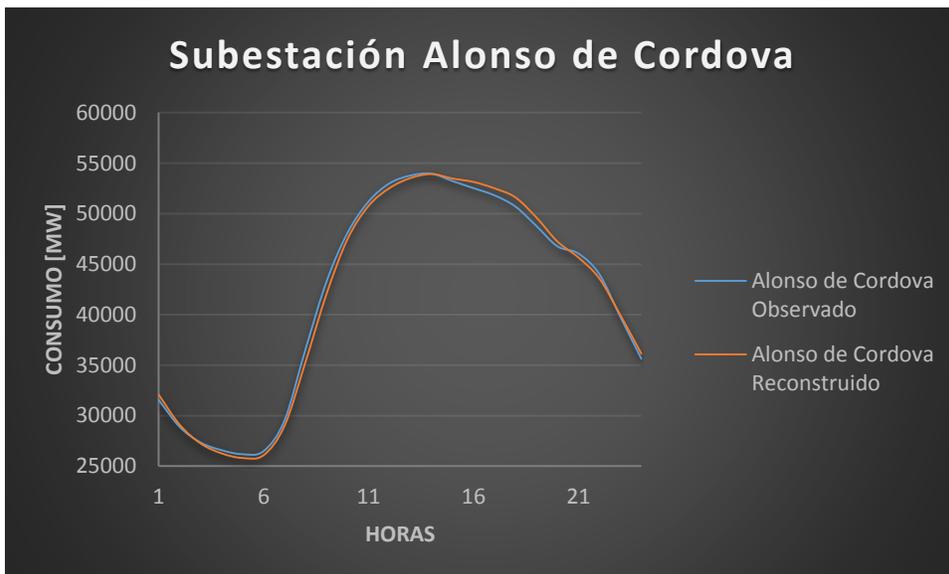


Figura 86: Perfil observado vs. Perfil reconstruido subestación Alonso de Córdoba.

Capítulo 9: Apéndice

9.1 Caracterización de K-Means

Supongamos que se tiene una colección de objetos $\mathbf{x}_j \in \mathbf{R}^d$ con $\mathbf{j} \in \{1, \dots, N\}$ y que se desean organizar en $\mathbf{C} = \{\mathbf{C}_1, \dots, \mathbf{C}_K\}$ subconjuntos. Se define la siguiente función de error cuadrático

$$J(\mathbf{Y}, \mathbf{M}) = \sum_{i=1}^K \sum_{j=1}^N \gamma_{ij} \|\mathbf{x}_j - \mathbf{m}_i\|^2 \quad \text{con } \mathbf{Y}_{\{y_{ij}\}} \text{ una matriz de partición donde } y_{ij} = \begin{cases} 1 & \text{si } \mathbf{x}_j \in \text{cluster } i \\ 0 & \text{en caso contrario} \end{cases}$$

Además $\sum_{i=1}^K y_{ij} = 1 \forall j$.

Dado lo anterior el algoritmo básico de K-means funciona de la siguiente forma:

- 1) Se inicializa una K-partición en forma aleatoria o en base a algún conocimiento a priori de los objetos calculándose la matriz

$$\mathbf{M} = [\mathbf{m}_1, \dots, \mathbf{m}_K] \text{ con } \mathbf{m}_i \text{ media de la cantidad objetos en el cluster } i$$
- 2) Se asigna cada objeto de la colección al clúster más cercano denominado \mathbf{C}_w bajo la siguiente regla:

$$\text{Si } \|\mathbf{x}_j - \mathbf{m}_w\| < \|\mathbf{x}_j - \mathbf{m}_i\| \text{ entonces } \mathbf{x}_j \in \mathbf{C}_w$$
- 3) Se recalcula la matriz de partición \mathbf{M} en base a la partición actual.
- 4) Se repiten los pasos 2) y 3) hasta que no hallan cambios en cada cluster o hasta que la función anterior converja a un ϵ dado.

9.2 Caracterización del algoritmo Fast Fourier Transformation

La transformada de Fourier descompone la señal temporal en una colección de sinusoides dadas por: $f(t) = \frac{a_0}{2} + \sum_{r=1}^K a_r \cos(rt) + \sum_{r=1}^K b_r \sin(rt)$ tal que

$$\begin{aligned} a_0 &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) dt \\ b_r &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \sin(rt) dt \\ a_r &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \cos(rt) dt \\ b_r &= \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \sin(rt) dt \end{aligned}$$

Con $r=1,2,3,\dots$

Esto permite caracterizar cada señal temporal a través de una constante $\mathbf{a}_0, \mathbf{a}_r, \mathbf{b}_r$ las cuales representan la magnitud de cada curva sinusoidal a una frecuencia dada por r .

Luego la suma de todas las magnitudes a todas las frecuencias permite reconstruir la señal original. El algoritmo de Fast Fourier Transformation permite realizar este procedimiento en forma eficiente y extraer los distintos coeficientes.

9.3 Código SQL y extracción de la data

```
SELECT Month([FECHA]) AS M, Hour([HORA])+1 AS H, Avg([MWh]) AS PotMedia, [Año
2013_DS14].[SUBESTACION
DS14] AS Expr1
FROM [Año
2013_DS14]
GROUP BY Month([FECHA]), Hour([HORA])+1, [Año
2013_DS14].[SUBESTACION
DS14]
ORDER BY Hour([HORA])+1;
SELECT Month([FECHA]) AS M, Day([FECHA]) AS DIA, Hour([HORA])+1 AS H, Avg([MWh]) AS
PotMedia, [Año
2013_DS14].[SUBESTACION
DS14] AS Expr1
FROM [Año
2013_DS14]
WHERE [Año
2013_DS14].[SUBESTACION
DS14]="Subest"
GROUP BY Month([FECHA]), Day([FECHA]), Hour([HORA])+1, [Año
2013_DS14].[SUBESTACION
DS14]
ORDER BY Day([FECHA]);
```

9.4 Script en R para generar cada Experimento del algoritmo fastICA

```
# Cargar Paquetes #
library("fastICA", lib.loc=~R/win-library/3.2")
library("zoo", lib.loc=~R/win-library/3.2")
library("signal", lib.loc=~R/win-library/3.2")
library("psych", lib.loc=~R/win-library/3.2")
library("PerformanceAnalytics", lib.loc=~R/win-library/3.2")
library("ggplot2", lib.loc=~R/win-library/3.2")
#Funciones de computo de errores#
# Funcion que retorna RMSE
rmse <- function(error,x){
  sqrt(mean(error^2))/(max(x)-min(x))}
# Funcion que retorna MAE
mae <- function(error){
  mean(abs(error))}
#Funcion que retorna MAPE
mape <- function(y, yhat){
  mean(abs((y - yhat)/y))}
#Funcion para invertir curvas en caso de ser necesario
invY<-function(x){
  1/x}
#####Cargar data_set#####
setwd("C:\\Users\\Sebastian\\Desktop\\Practica 3 y Memoria\\Inputs Modelos\\Consumo Regulado por
SubEst")
data_set<-read.table("Consumo Horario Mensual por Subest.txt", header=TRUE, sep="\t", dec=",")
data_set<-na.omit(data_set)
data_set<-as.data.frame.matrix(data_set)
attach(data_set)
data_set<-data_set[,2:length(data_set)]
consumo_total<-rowSums(data_set)
setwd("C:\\Users\\Sebastian\\Desktop\\Practica 3 y Memoria\\Resultados Modelo\\ICs_Subest")
write.csv(consumo_total, file="Consumo_total.csv")
###ICA k=POR FIJAR###
for (k in 1:114){
  iter<-k
  X<-data_set
  Model<-fastICA(X, k, alg.typ="deflation", fun="logcosh", alpha=2, method="R", row.norm=FALSE,
maxit=100, tol=0.001, verbose=TRUE)
  MM<-Model$A
  ICs<-Model$S
  X.1<-Model$X
```

```

write.csv(ICs, file="ICs_iter.csv")
write.csv(MM, file="MM_iter.csv")
#Reconstruir X#
J<-ICs%*%MM
for (i in 1:length(X)) {J[,i]<-J[,i]+mean(X[,i])}
error<-c(0)
errorMAE<-c(0)
errorRMSE<-c(0)
errorMAPE<-c(0)
for (i in 1:length(X)) {error<-c(error,X[,i]-J[,i])}
for (i in 1:length(X)) {errorRMSE<-c(errorRMSE,rmse(error[i],X[,i]))}
for (i in 1:length(X)){errorMAE<-c(errorMAE,mae(error[i+1]))}
for (i in 1:length(X)){errorMAPE<-c(errorMAPE,mape(X[,i],J[,i]))}
consumo_total_iter<-rowSums(J)
error_sist<-consumo_total_iter-consumo_total_iter
rmse_sist<-rmse(error_sist, consumo_total)
mae_sist<-mae(error_sist)
mape_sist<-mape(consumo_total,consumo_total_iter)
plot(consumo_total,type="l", lwd=2, xlab="hours", ylab="Observed Energy Consumption [MW]",
main="Observed Energy Consumption and ICA reconstruction")
points(consumo_total_iter, type="l", col="red", lwd=0.5)
write.csv(rmse_sist, file="rmse_sist_15.csv")
write.csv(mae_sist, file="mae_sist_15.csv")
write.csv(mape_sist, file="mape_sist_15.csv")
write.csv(errorMAE, file="MAE15.csv")
write.csv(errorMAPE, file="MAPE15.csv")
write.csv(errorRMSE, file="RMSE15.csv")
}

```

9.5 Script en R para el cálculo de las correlaciones

```

#####Cargar data_sets#####
setwd("C:\\Users\\Sebastian\\Desktop\\Practica 3 y Memoria\\Resultados Modelo\\Correlaciones")
data_set_IMACEC<-read.table("IMACEC.txt", header=FALSE, sep="\t", dec=",")
data_set_ics<-read.table("Consumo_Segmentos_subest.txt", header=TRUE, sep="\t", dec=",")
data_set_ics<-data_set_ics[,2:6]
data_set_IMACEC<-data_set_IMACEC[,2]
#Correlaciones#
cor.test(data_set_IMACEC, data_set_ics[,1],method=c("pearson"))
cor.test(data_set_IMACEC, data_set_ics[,1],method=c("kendall"), exact=NULL)
cor.test(data_set_IMACEC, data_set_ics[,1],method=c("spearman"))
par(mfrow=c(2,3), font=1, font.lab=1, font.axis=2)
plot(data_set_ics[,1],data_set_IMACEC, ylab="IMACEC", xlab="Segmento Industrial")
plot(data_set_ics[,2],data_set_IMACEC, ylab="IMACEC", xlab="Segmento Residencial Anomalo")
plot(data_set_ics[,3],data_set_IMACEC, ylab="IMACEC", xlab="Pequeño Negocio")
plot(data_set_ics[,4],data_set_IMACEC, ylab="IMACEC", xlab="Residencial Bajo Consumo")
plot(data_set_ics[,5],data_set_IMACEC, ylab="IMACEC", xlab="Residencial Alto Consumo")
#####Cargar data_sets#####
setwd("C:\\Users\\Sebastian\\Desktop\\Practica 3 y Memoria\\Resultados Modelo\\Correlaciones")
data_set_IMACEC<-read.table("IMACEC_trans.txt", header=FALSE, sep="\t", dec=",")
data_set_ics<-read.table("Consumo_Segmentos_transf.txt", header=TRUE, sep="\t", dec=",")

```

```

data_set_ics<-data_set_ics[,2:6]
data_set_IMACEC<-data_set_IMACEC[1:12,2]
#Correlaciones#
cor.test(data_set_IMACEC, data_set_ics[,1],method=c("pearson"))
cor.test(data_set_IMACEC, data_set_ics[,1],method=c("kendall"), exact=NULL)
cor.test(data_set_IMACEC, data_set_ics[,1],method=c("spearman"))
cor.test(data_set_IMACEC, data_set_ics[,2],method=c("pearson"))
cor.test(data_set_IMACEC, data_set_ics[,2],method=c("kendall"), exact=NULL)
cor.test(data_set_IMACEC, data_set_ics[,2],method=c("spearman"))
cor.test(data_set_IMACEC, data_set_ics[,3],method=c("pearson"))
cor.test(data_set_IMACEC, data_set_ics[,3],method=c("kendall"), exact=NULL)
cor.test(data_set_IMACEC, data_set_ics[,3],method=c("spearman"))
cor.test(data_set_IMACEC, data_set_ics[,4],method=c("pearson"))
cor.test(data_set_IMACEC, data_set_ics[,4],method=c("kendall"), exact=NULL)
cor.test(data_set_IMACEC, data_set_ics[,4],method=c("spearman"))
cor.test(data_set_IMACEC, data_set_ics[,5],method=c("pearson"))
cor.test(data_set_IMACEC, data_set_ics[,5],method=c("kendall"), exact=NULL)
cor.test(data_set_IMACEC, data_set_ics[,5],method=c("spearman"))
par(mfrow=c(2,3), font=1, font.lab=1, font.axis=2)
plot(data_set_ics[,1],data_set_IMACEC, ylab="IMACEC", xlab="Segmento Industrial")
abline(lm(data_set_IMACEC~data_set_ics[,1]), col="red")
plot(data_set_ics[,2],data_set_IMACEC, ylab="IMACEC", xlab="Segmento Residencial
Anomalo")
abline(lm(data_set_IMACEC~data_set_ics[,2]), col="red")
plot(data_set_ics[,3],data_set_IMACEC, ylab="IMACEC", xlab="Pequeño Negocio")
abline(lm(data_set_IMACEC~data_set_ics[,3]), col="red")
plot(data_set_ics[,4],data_set_IMACEC, ylab="IMACEC", xlab="Residencial Bajo
Consumo")
abline(lm(data_set_IMACEC~data_set_ics[,4]), col="red")
plot(data_set_ics[,5],data_set_IMACEC, ylab="IMACEC", xlab="Residencial Alto
Consumo")
abline(lm(data_set_IMACEC~data_set_ics[,5]), col="red")

```