



ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)

GRADO EN INGENIERÍA TELEMÁTICA

IDENTIFICACIÓN DE PATRONES DE COMPORTAMIENTO DE CONSUMO EN EL MERCADO ELÉCTRICO EN ENTORNOS BIG DATA

Autor: Beatriz Torreiro Mosquera

Directores: David Contreras Bárcena y Miguel Ángel Sanz Bobi

Madrid

Julio 2018

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título
Identificación de patrones de comportamiento de consumo en el mercado eléctrico en
entornos Big Data

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el
curso académico 2017/18 es de mi autoría, original e inédito y

no ha sido presentado con anterioridad a otros efectos.

El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido
tomada de otros documentos está debidamente referenciada.

Fdo.: Beatriz Torreiro Mosquera Fecha: 11 / 07 / 2018

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.: David Contreras Bárcena y Miguel Ángel Sanz Bobi Fecha: 11 / 07 / 2018

AUTORIZACIÓN PARA LA DIGITALIZACIÓN, DEPÓSITO Y DIVULGACIÓN EN RED DE PROYECTOS FIN DE GRADO, FIN DE MÁSTER, TESINAS O MEMORIAS DE BACHILLERATO

1º. Declaración de la autoría y acreditación de la misma.

El autor D^a.Beatriz Torreiro Mosquera DECLARA ser el titular de los derechos de propiedad intelectual de la obra: “Identificación de patrones de comportamiento de consumo en el mercado eléctrico en entornos Big Data”, que ésta es una obra original, y que ostenta la condición de autor en el sentido que otorga la Ley de Propiedad Intelectual.

2º. Objeto y fines de la cesión.

Con el fin de dar la máxima difusión a la obra citada a través del Repositorio institucional de la Universidad, el autor CEDE a la Universidad Pontificia Comillas, de forma gratuita y no exclusiva, por el máximo plazo legal y con ámbito universal, los derechos de digitalización, de archivo, de reproducción, de distribución y de comunicación pública, incluido el derecho de puesta a disposición electrónica, tal y como se describen en la Ley de Propiedad Intelectual. El derecho de transformación se cede a los únicos efectos de lo dispuesto en la letra a) del apartado siguiente.

3º. Condiciones de la cesión y acceso

Sin perjuicio de la titularidad de la obra, que sigue correspondiendo a su autor, la cesión de derechos contemplada en esta licencia habilita para:

- a) Transformarla con el fin de adaptarla a cualquier tecnología que permita incorporarla a internet y hacerla accesible; incorporar metadatos para realizar el registro de la obra e incorporar “marcas de agua” o cualquier otro sistema de seguridad o de protección.
- b) Reproducir la en un soporte digital para su incorporación a una base de datos electrónica, incluyendo el derecho de reproducir y almacenar la obra en servidores, a los efectos de garantizar su seguridad, conservación y preservar el formato.
- c) Comunicarla, por defecto, a través de un archivo institucional abierto, accesible de modo libre y gratuito a través de internet.
- d) Cualquier otra forma de acceso (restringido, embargado, cerrado) deberá solicitarse expresamente y obedecer a causas justificadas.
- e) Asignar por defecto a estos trabajos una licencia Creative Commons.
- f) Asignar por defecto a estos trabajos un HANDLE (URL *persistente*).

4º. Derechos del autor.

El autor, en tanto que titular de una obra tiene derecho a:

- a) Que la Universidad identifique claramente su nombre como autor de la misma
- b) Comunicar y dar publicidad a la obra en la versión que ceda y en otras posteriores a través de cualquier medio.
- c) Solicitar la retirada de la obra del repositorio por causa justificada.
- d) Recibir notificación fehaciente de cualquier reclamación que puedan formular terceras personas en relación con la obra y, en particular, de reclamaciones relativas a los derechos de propiedad intelectual sobre ella.

5º. Deberes del autor.

El autor se compromete a:

- a) Garantizar que el compromiso que adquiere mediante el presente escrito no infringe ningún derecho de terceros, ya sean de propiedad industrial, intelectual o cualquier otro.
- b) Garantizar que el contenido de las obras no atenta contra los derechos al honor, a la intimidad y a la imagen de terceros.
- c) Asumir toda reclamación o responsabilidad, incluyendo las indemnizaciones por daños, que pudieran ejercitarse contra la Universidad por terceros que vieran infringidos sus derechos e intereses a causa de la cesión.



ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)

GRADO EN INGENIERÍA TELEMÁTICA

IDENTIFICACIÓN DE PATRONES DE COMPORTAMIENTO DE CONSUMO EN EL MERCADO ELÉCTRICO EN ENTORNOS BIG DATA

Autor: Beatriz Torreiro Mosquera

Director: David Contreras Bárcena y Miguel Ángel Sanz Bobi

Madrid

Julio 2018

Agradecimientos

Este proyecto va dedicado a aquellas persona sin las cuales no hubiese sido posible.

En primer lugar, dar las gracias a los directores, David Contreras Bárcena y Miguel Ángel Sanz Bobi, que siempre han estado disponibles para ayudarme a lo largo de este trabajo.

A todas las personas que he conocido durante estos cuatro años, dentro y fuera de ICAI, que ahora puedo llamar amigos y han hecho de Madrid, mi segunda casa.

Por último, a mis dos pilares fundamentales, mi madre y mi hermana. A mi hermana pequeña, por hacer que me sienta orgullosa de ser su hermana mayor y darme las ganas necesarias para, cada día, intentar ser un mejor referente en su vida. A mi madre, por apoyarnos día a día, sacar fuerzas de donde no las hay para afrontar las adversidades y hacernos comprender, que las tres juntas, podemos conseguirlo todo.

IDENTIFICACIÓN DE PATRONES DE COMPORTAMIENTO DE CONSUMO EN EL MERCADO ELÉCTRICO EN ENTORNOS BIG DATA

Autor: Torreiro Mosquera, Beatriz

Directores: Contreras Bárcena, David y Sanz Bobi, Miguel Ángel

Entidad Colaboradora: ICAI - Universidad Pontificia Comillas

RESUMEN DEL PROYECTO

El objetivo principal de este proyecto reside en la utilización de técnicas de procesamiento Big Data con la finalidad de obtener información, sobre el consumo eléctrico en España, que sea de utilidad para las distribuidoras y generadoras eléctricas. Dicha información les permitirá conocer más a fondo las tendencias del mercado y las necesidades de sus clientes, lo que les facilitará la creación de ofertas orientadas a un tipo de cliente específico, que podrán ser usadas para mejorar el rendimiento de las empresas eléctricas. Para ello se ha llevado a cabo un análisis descriptivo sobre el consumo eléctrico que permite recaudar información acerca de como ciertos factores afectan al consumo de manera global. Teniendo en cuenta la información obtenida en dicho análisis, se procede a la identificación de patrones de comportamiento del mercado eléctrico español en función del tipo de actividad económica, el día de la semana, la estación del año y la provincia. Dichos patrones se han obtenido mediante la implementación, sobre el Cluster Big Data de ICAI, del algoritmo de agrupamiento K-Means, que crea grupos de datos en función del número de clúster, que es deducido a partir de una función de costes calculada.

Palabras clave: Big Data, Apache Spark, K-Means, Sector eléctrico, Patrones de comportamiento.

1. Introducción

La utilidad del Big Data recae en la capacidad de responder a preguntas que las empresas ni siquiera se habían formulado. La gran cantidad de datos permite, a las organizaciones, buscar tendencias, información que será empleada en la identificación de problemas y en la toma de decisiones. En la actualidad, numerosas organizaciones utilizan el Big Data con el fin de obtener información acerca de cómo reducir costes, lanzar nuevos productos y servicios o mejorar el rendimiento, [1].

El análisis Big Data se está adentrando, poco a poco, en todos los sectores del mercado, debido a la gran cantidad de información útil derivable de las grandes cantidades de datos. Siguiendo con la tendencia actual, no es de extrañar la introducción de técnicas de procesamiento Big Data en el sector eléctrico. Dichas técnicas hacen posible la detección de patrones de consumo, información que es de gran utilidad para las distribuidoras y generadoras eléctricas, ya que les permite prever el consumo futuro, así como amoldar las costumbres de sus clientes para beneficio propio.

2. Definición del proyecto

Con la realización de este proyecto se pretende alcanzar los siguientes objetivos:

- Obtener información acerca de como el tipo de actividad industrial, la estación, el día de la semana, el tipo de mercado, la localidad y la hora del día afectan al consumo eléctrico en España, a nivel global.
- Indagar acerca de las relaciones existente entre los factores anteriormente mencionados.
- Obtener patrones de comportamiento del consumo eléctrico de España, en función del día de la semana (día laboral, sábado o domingo), la estación (invierno, primavera, verano y otoño), el tipo de actividad económica (residencial o industrial) y la provincia (Madrid, Barcelona o Málaga) a lo largo del día.
- Proposición de herramientas comerciales que puedan emplear las distribuidoras y generadoras eléctricas para modificar las tendencias del consumo en su favor.

3. Descripción del modelo

Con la finalidad de extraer conclusiones de los grupos obtenidos, es necesario establecer un modelo de actuación que permita alcanzar dichas conclusiones y que sea desplegable sobre el Clúster Big Data de ICAI. En primer lugar, se deben aplicar los filtros pertinentes al conjunto de datos, para así excluir aquellos que no cumplen con los parámetros previamente definidos. A continuación, el conjunto de datos resultante se divide en dos grupos. Un grupo contendrá el 75 % del volumen de datos resultante y se empleará en el cálculo de los patrones de comportamiento. El 25 % restante será utilizado para determinar la validez de los grupos obtenidos.

Con el grupo de datos destinado a la predicción, se procederá al cálculo del número de clústers (k) a emplear en el algoritmo K-Means. Para calcular el número de clústers a emplear, es necesario construir la función de coste, que representa la suma total de las varianzas con respecto a los centroides calculados en función del número de clústers empleado. El valor óptimo de número de clústers coincide con el “codo” de la función coste, si se da el caso que el “codo” no es discernible a primera vista, se puede recurrir al cálculo de la disminución porcentual de la suma de varianzas en función del número de clústers, para escoger el mismo.

Una vez establecido el número de clústers a emplear, se procederá a la obtención de los centroides que identifican cada uno de los grupos, mediante la implementación del algoritmo K-Means sobre los datos de predicción. A continuación, se procede a la asignación de grupos de los datos de predicción y los de validación. Con los todos los datos asociados a un grupo de comportamiento se calculan los porcentajes de pertenencia de cada grupo, es decir, se obtiene, en forma de porcentaje, la cantidad de datos que pertenecen a cada grupo. Esta operación es realizada diferenciando los datos de predicción de los de validación, de forma que se hallarán por un lado los porcentajes de los datos de predicción y por otro los porcentajes de los datos de validación.

A continuación, es necesario determinar si el modelo obtenido es válido o no. Para ello se calcula la variación media por grupo mediante la comparación de los porcentajes de pertenencia provenientes de los datos de predicción con los procedentes de los datos de validación. Es decir, para cada grupo de comportamiento se halla la resta, en valor absoluto, del porcentaje de representación obtenido con los datos de predicción menos el porcentaje de representación calculado con los datos de validación.

Si la variación encontrada es lo suficientemente pequeña, se puede asumir dicha fluctuación y el modelo es válido. Por el contrario, si la variación media obtenida no es asumible, el modelo no es válido y es necesario volver a calcular el número de clústers a emplear.

El siguiente gráfico muestra de forma esquemática el modelo empleado.

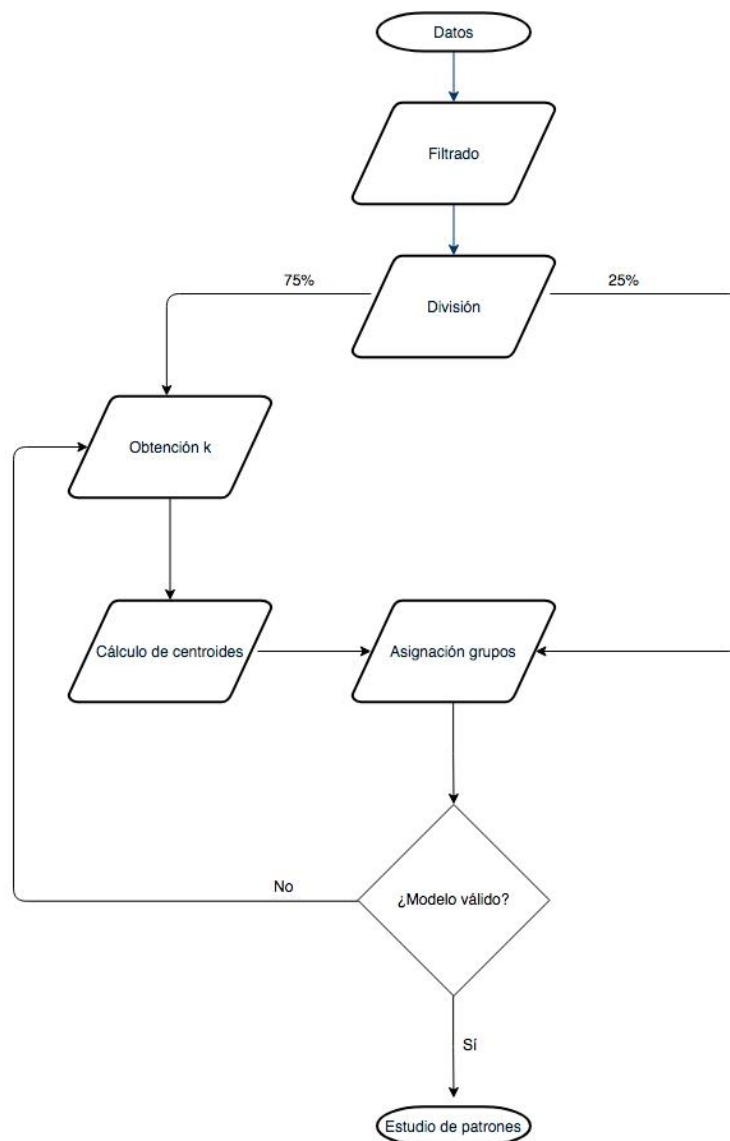


Figura 1.: Flujo del modelo empleado

4. Resultados

Los resultados obtenidos en este proyecto han sido exitosos ya que, para cada uno de los casos de estudio planteados a lo largo del mismo, se han obtenidos modelos válidos.

En la mayor parte de los casos, los grupos de comportamiento encontrados se pudieron reagrupar en función de la tendencia presentada a lo largo del día, residiendo la única diferencia en los niveles de consumo que tomaban. Dichos grupos permitieron una mayor comprensión acerca de los hábitos de los distintos tipos de consumidores, lo que posibilitó la proposición de posibles campañas de marketing que pueden llevar a cabo las distribuidoras eléctricas con el objetivo de optimizar sus plantas.

A continuación, se muestran dos gráficos que recogen los distintos patrones de comportamiento encontrados en el consumo residencial de un día laboral de invierno, el primero en la provincia de Madrid y el segundo en la de Barcelona.

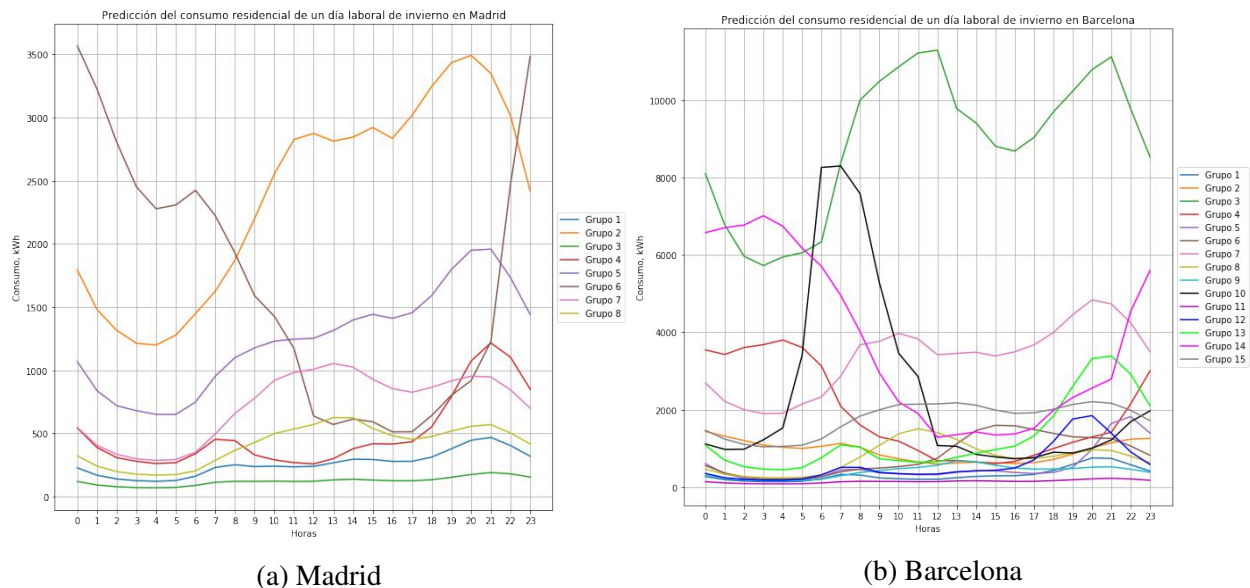


Figura 2.: Consumo residencial de un día laboral de invierno

5. Conclusiones

En el presente proyecto, se ha conseguido aplicar, de forma exitosa, el algoritmo de agrupamiento K-Means, con la finalidad de encontrar los distintos grupos de comportamiento que definen el consumo eléctrico en España. Dicha información ha sido empleada para la confección de herramientas comerciales que permitan, a la distribuidoras y generadoras eléctricas, alcanzar sus objetivos.

6. Referencias

[1] PowerData Especialistas en Gestión de Datos, “Big Data: ¿En qué consiste? Su importancia, desafíos y gobernabilidad.” <https://www.powerdata.es/big-data>, 2018.

IDENTIFICATION OF CONSUMER BEHAVIOR PATTERNS IN THE ELECTRIC MARKET IN BIG DATA ENVIRONMENTS

Author: Torreiro Mosquera, Beatriz

Supervisors: Contreras Bárcena, David and Sanz Bobi, Miguel Ángel

Collaborating Entity: ICAI - Universidad Pontificia Comillas

ABSTRACT

The main objective of this project lies in the use of Big Data processing techniques in order to gather information regarding the electricity consumption in Spain, which will be useful for electricity distributors and generators. This information will allow them to learn more about market trends and the needs of their clients, that will help them create offers aimed at a specific type of customer, which can be used to improve the performance of electricity companies. To do this, a descriptive analysis has been carried out on electricity consumption, which enables the gathering of information about how different factors affect the consumption as a whole. Taking into account the information obtained in this analysis, the next step is the identification of behavior patterns of the Spanish market according to the type of economic activity, the day of the week, the season of the year and the province. Those patterns have been obtained through the implementation, run on ICAI's Big Data Cluster, of the K-Means clustering algorithm, which creates data groups according to the cluster number, which is deduced from a calculated cost function.

Keywords: Big Data, Apache Spark, K-Means, Electric market, Behavior patterns.

1. Introduction

The usefulness of Big Data lies in its ability to answer questions that companies have not even thought of. The large amount of data enables companies to find tendencies, information that will be used in order to identify problems and make decisions. Currently, many companies use Big Data processing techniques to obtain insight in how to reduce costs, launch new products and services, or improve performance, [1].

Big Data analysis is entering all markets, one at a time, due to the huge amount of useful information that can be drawn from the large amount of data that is collected by companies everyday. Following this trend, it does not come as a surprise the introduction of Big Data processing techniques in the electric market. These techniques make possible the detection of consumption patterns, information that is very convenient for electric distributors and generators, as it allows them to predict future consumption levels, as well as, to shift the current consumption tendencies in their favor, by the launch of personalized offers.

2. Project definition

The following objectives are to be achieved:

- Obtain information regarding how the type of activity, the season, the day of the week, the type of market fare, the location and the time of the day affect electricity consumption in Spain, as a whole.
- Inquire about the existing relationships between the aforementioned factors.
- Obtain behavior patterns of the electricity consumption in Spain, depending on the type of day of the week (working day, Saturday or Sunday), the season (winter, spring, summer and fall), the type of economic activity (residential or industrial) and the province (Madrid, Barcelona or Málaga) throughout the day.
- Proposition of commercial tools that electric distributors and generators can use in order to modify current trends in their favor.

3. Description of the model

In order to draw conclusions from behavior patterns, it is necessary to establish a model that will help us reach those conclusions and that can be run in ICAI's Big Data Cluster. First of all, filters should be apply to the data, in order to exclude those sets of data that do not meet the requirements set by the parameters in each case of study. Next, the obtained data is divided in two groups, one will contain 75 % of the data and will be used to create the behavior patterns, while the other 25 % will be used to verify the reliability of those patterns.

Afterwards, with the data subgroup destined to the creation of the patterns, the number of clusters to be used in the K-Means algorithm will be calculated. In order to calculate the number of clusters (k) needed, it is necessary to construct the cost function, which represents the total sum of the variances with respect to the centroids calculated based on the number of clusters used. The optimal value for the number of clusters lies in the "elbow" of the cost function. If it is the case that the "elbow" is not discernible at first sight, the calculation of the percentage decrease of the sum of variances according to the number of clusters, can be used instead.

Once the number of clusters has been established, the centroids that identify each of the behavior patterns will be obtained by implementing the K-Means algorithm on the prediction data. Next, the assignment of patterns of prediction and validation data takes place. With all the data associated with a behavior group, the percentages of belonging to each group are calculated, that is, the amount of data belonging to each group is obtained as a percentage. This operation is carried out by differentiating the prediction data from the validation data, so that two sets of percentages are obtained.

Next, it is necessary to determine of the model obtained is valid or not. To do so, the average variation per group is calculated by comparing the percentages of belonging from the prediction data with those from the validation data. That is, for each behavior group, the difference in

absolute value between the percentage of belonging obtained from the prediction data and the one obtained from the validation data.

If the variation found is small enough, this fluctuation can be dealt with and the model is valid. On the other hand, if the average variation obtained is not acceptable, the model is not valid and it is necessary to recalculate the number of clusters to be used.

The following graphic shows schematically the model used.

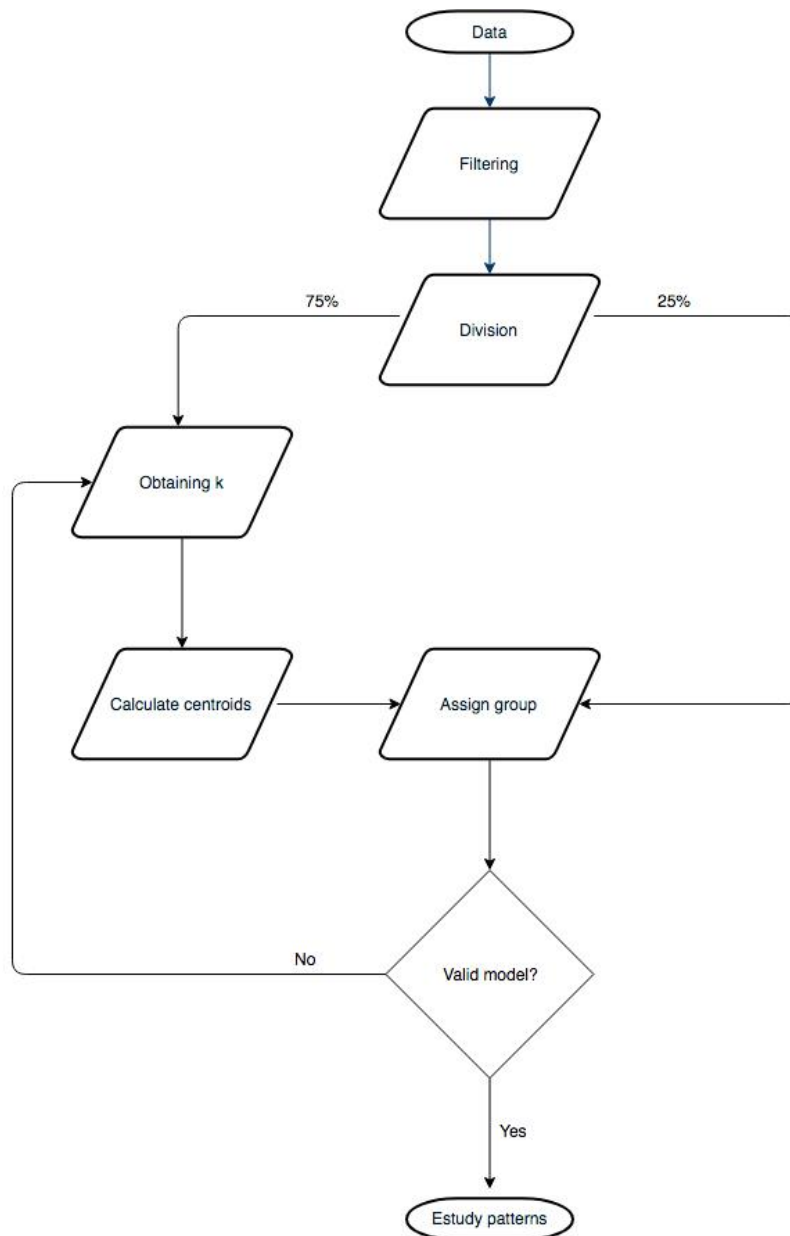


Figura 3.: Flow of the model used

4. Results

The results obtained in this project have been successful since valid models have been obtained for each one of the cases studied throughout the project.

In most cases, the obtained behavior patterns could be grouped according to their trend throughout the day, differentiating only in the levels of consumption. These groups allowed a greater understanding of the habits of different types of consumers, which made possible the proposal of marketing campaigns that can be carried out by the electricity distributors in order to optimize their plants.

Bellow are two graphs showing the different behavior patterns found in residential consumption of a winter working day, in the provinces of Madrid and Barcelona.

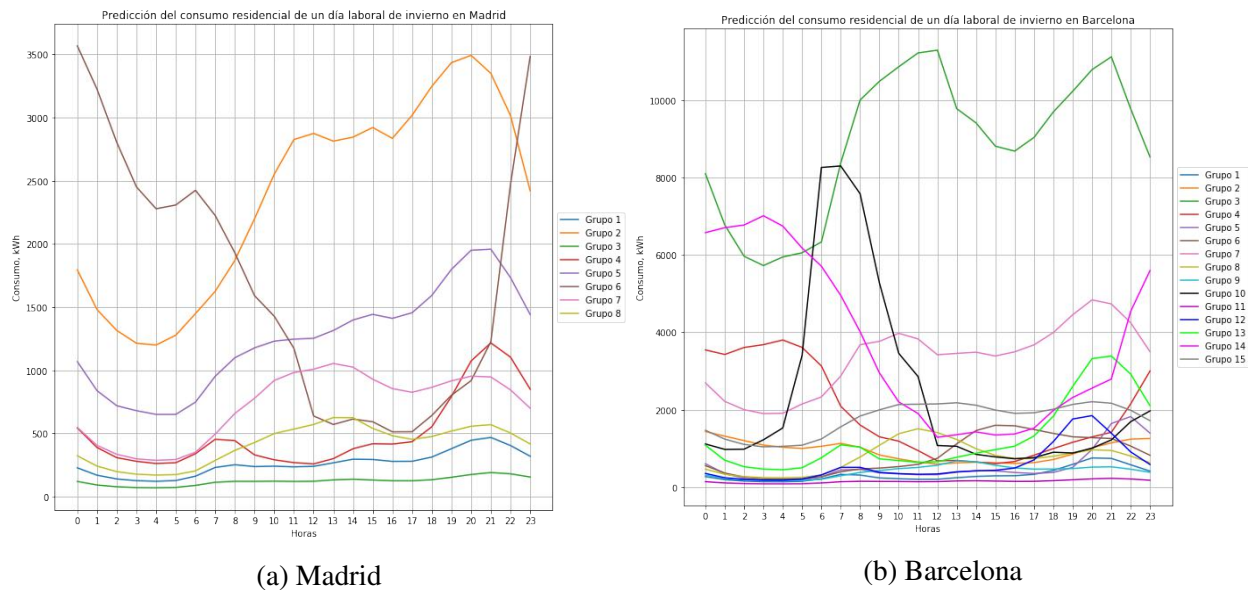


Figura 4.: Residential consumption of a winter working day

5. Conclusions

In this project, the K-Means clustering algorithm has been successfully applied, in order to find the different groups of behavior that define electricity consumption in Spain. This information has been used for the preparation of commercial tools that allow the electricity distributors and generators to achieve their objectives.

6. References

[1] PowerData Especialistas en Gestión de Datos, “Big Data: ¿En qué consiste? Su importancia, desafíos y gobernabilidad.” <https://www.powerdata.es/big-data>, 2018.

Índice de la memoria

Índice de figuras	III
Índice de tablas	VII
1. Introducción	1
2. Descripción De Las Tecnologías	3
2.1. Jupyter Notebook	3
2.2. Spark	3
2.3. Machine Learning	4
3. Estado De La Cuestión	5
3.1. El sector eléctrico en España	5
3.2. Técnicas de procesamiento Big Data	6
4. Descripción Del Modelo Desarrollado	9
4.1. Objetivos y Especificación	9
4.2. Datos	10
4.3. Algoritmos	11
4.3.1. Algoritmo K-Means	11
4.4. Plataforma de desarrollo	13
4.4.1. Arquitectura	13
4.4.2. Implementación	19
5. Análisis De Resultados	33
5.1. Análisis Descriptivo General	33
5.1.1. Consumo por tipo de actividad económica	34
5.1.2. Consumo por día de la semana	35
5.1.3. Consumo por hora del día	37
5.1.4. Consumo por municipios	38
5.1.5. Consumo por tipo de tarifa de mercado	40
5.2. Análisis Descriptivo De Las Provincias De Madrid, Barcelona Y Málaga	42
5.2.1. Comparativa basada en el tipo de actividad económica	44
5.2.2. Comparativa basada en el día de la semana	47
5.2.3. Comparativa basada en la hora del día	49
5.2.4. Comparativa basada en el tipo de tarifa de mercado	50
5.3. Análisis Descriptivo Del Tipo De Actividad Económica Industrial y Residencial	52
5.3.1. Comparativa basada en el día de la semana	54
5.3.2. Comparativa basada en la hora del día	55

5.3.3. Comparativa basada en la tarifa de mercado	58
5.4. Modelo de predicciones utilizando el algoritmo K-Means	60
5.4.1. Municipios de la provincia de Madrid	61
5.4.2. Municipios de la provincia de Barcelona	116
5.4.3. Municipios de la provincia de Málaga	154
6. Conclusiones Y Trabajos Futuros	195
Bibliografía	197
A. Nodo master	199
B. Nodo worker	201
C. Nodo edge	203
D. Clúster Big Data ICAI	205
E. Código del notebook K-Means, Madrid Residencial Invierno Diario	207
F. Consumo anual de los municipios	227

Índice de figuras

3.1. Herramientas Big Data	7
4.1. Número de clústers Vs Suma de las distancias euclídeas	12
4.2. Arquitectura Spark	14
4.3. Running Applications, plataforma web de Hadoop	15
4.4. Información de aplicación, plataforma web de Hadoop	15
4.5. Clúster Big Data ICAI	17
4.6. Planificación de los trabajos de Spark	18
4.7. Información ejecutores de Spark	18
4.8. Ejemplo de notebook de Jupyter	20
4.9. Ejemplo de pestaña Running de Jupyter	20
4.10. Lectura de los datos objeto de estudio	24
4.11. Variables empleadas durante la limpieza de datos	25
4.12. Filtrado en función del tipo de actividad económica	25
4.13. Filtrado en función del tipo de día de la semana	25
4.14. Filtrado en función de la provincia	26
4.15. Filtrado en función de la estación del año	26
4.16. Filtrado de aquellos datos que contiene niveles de consumo activo negativos	26
4.17. Generación del gráfico que muestra las tendencias de los patrones encontrados	27
4.18. Tiempo que tarda la aplicación en leer el fichero de datos	28
4.19. Tiempo que tarda la aplicación en filtrar por tipo de día de la semana	29
4.20. Tiempo que tarda la aplicación en realizar la operación show	29
4.21. Tiempo que tarda la aplicación en crear la función de coste del algoritmo K-Means	30
4.22. Tiempo que tarda la aplicación en calcular los porcentajes de pertenencia a los grupos	30
4.23. Tiempo que tarda la aplicación en calcular los márgenes de desviación	30
5.1. Distribución del consumo total diario	33
5.2. Distribución general del CNAE	34
5.3. Distribución del consumo por días de la semana	35
5.4. Consumo medio Residencial Vs Industrial	36
5.5. Consumo general por horas	38
5.6. Distribución del consumo total por municipios	39
5.7. Distribución de las tarifa de mercado	40
5.8. Distribución del consumo medio por horas para cada tarifa	41
5.9. Consumo total de la tarifa de mercado regulado vs libre	42
5.10. Comparativa de la distribución del consumo total de las provincias de Barcelona, Madrid y Málaga	44

5.11. Diagrama de sectores del CNAE de cada provincia	45
5.12. Comparativa de las distribuciones del consumo total en función del tipo de actividad económica	46
5.13. Comparativa del consumo medio de cada provincia en función del día de la semana	47
5.14. Comparativa del consumo medio de cada provincia en función del día de la semana y el tipo de actividad económica	48
5.15. Comparativa de la variación del consumo dependiendo de la hora del día	49
5.16. Diagrama de sectores de la distribución de tarifas de cada provincia	50
5.17. Porcentajes de CNAE en función del tipo de actividad económica y de la provincia	51
5.18. Porcentaje de cada actividad económica	52
5.19. Distribución del consumo residencial vs industrial	53
5.20. Variación del consumo industrial y residencial a lo largo de la semana	54
5.21. Variación del consumo medio a lo largo del día en función del CNAE	55
5.22. Variación del consumo medio en función del tipo de día de la semana	57
5.23. Diagrama de sectores de los tipos de tarifas de mercado en función del CNAE	58
5.24. Consumo medio de cada actividad económica por tarifa de mercado	59
5.25. Distribución de los tipos de tarifa de mercado en función del CNAE	59
5.26. Número de grupos en función de la suma de las distancias a los centroides	62
5.27. Consumo residencial de un día laboral de invierno en Madrid	62
5.28. Número de grupos en función de la suma de las distancias a los centroides	64
5.29. Consumo residencial de un sábado de invierno en Madrid	65
5.30. Consumo residencial de un domingo de invierno en Madrid	65
5.31. Número de grupos en función de la suma de las distancias a los centroides	67
5.32. Consumo industrial de un día laboral de invierno en Madrid	68
5.33. Número de grupos en función de la suma de las distancias a los centroides	69
5.34. Consumo industrial de un sábado de invierno en Madrid	70
5.35. Número de grupos en función de la suma de las distancias a los centroides	72
5.36. Consumo industrial de un domingo de invierno en Madrid	72
5.37. Número de grupos en función de la suma de las distancias a los centroides	74
5.38. Consumo residencial de un día laboral de primavera en Madrid	75
5.39. Número de grupos en función de la suma de las distancias a los centroides	76
5.40. Consumo residencial de un sábado de primavera en Madrid	77
5.41. Número de grupos en función de la suma de las distancias a los centroides	79
5.42. Consumo residencial de un domingo de primavera en Madrid	79
5.43. Número de grupos en función de la suma de las distancias a los centroides	81
5.44. Consumo industrial de un día laboral de primavera en Madrid	82
5.45. Número de grupos en función de la suma de las distancias a los centroides	84
5.46. Consumo industrial de un sábado de primavera en Madrid	85
5.47. Consumo industrial de un domingo de primavera en Madrid	85
5.48. Número de grupos en función de la suma de las distancias a los centroides	88
5.49. Consumo residencial de un día laboral de verano en Madrid	89
5.50. Consumo residencial de un sábado de verano en Madrid	89
5.51. Consumo residencial de un domingo de verano en Madrid	90
5.52. Número de grupos en función de la suma de las distancias a los centroides	92
5.53. Consumo industrial de un día laboral de verano en Madrid	93

5.54. Número de grupos en función de la suma de las distancias a los centroides	95
5.55. Consumo industrial de un sábado de verano en Madrid	96
5.56. Número de grupos en función de la suma de las distancias a los centroides	98
5.57. Consumo industrial de un domingo de verano en Madrid	98
5.58. Número de grupos en función de la suma de las distancias a los centroides	101
5.59. Consumo residencial de un día laboral de otoño en Madrid	101
5.60. Número de grupos en función de la suma de las distancias a los centroides	103
5.61. Consumo residencial de un sábado de otoño en Madrid	104
5.62. Número de grupos en función de la suma de las distancias a los centroides	106
5.63. Consumo residencial de un domingo de otoño en Madrid	107
5.64. Número de grupos en función de la suma de las distancias a los centroides	108
5.65. Consumo industrial de un día laboral de otoño en Madrid	109
5.66. Número de grupos en función de la suma de las distancias a los centroides	111
5.67. Consumo industrial de un sábado de otoño en Madrid	111
5.68. Número de grupos en función de la suma de las distancias a los centroides	114
5.69. Consumo industrial de un domingo de otoño en Madrid	114
5.70. Número de grupos en función de la suma de las distancias a los centroides	117
5.71. Consumo residencial de un día de invierno en Barcelona	118
5.72. Número de grupos en función de la suma de las distancias a los centroides	122
5.73. Consumo industrial de un día de invierno en Barcelona	123
5.74. Número de grupos en función de la suma de las distancias a los centroides	127
5.75. Consumo residencial de un día de primavera en Barcelona	128
5.76. Número de grupos en función de la suma de las distancias a los centroides	131
5.77. Consumo industrial de un día de primavera en Barcelona	132
5.78. Número de grupos en función de la suma de las distancias a los centroides	136
5.79. Consumo residencial de un día de verano en Barcelona	137
5.80. Número de grupos en función de la suma de las distancias a los centroides	140
5.81. Consumo industrial de un día de verano en Barcelona	141
5.82. Número de grupos en función de la suma de las distancias a los centroides	145
5.83. Consumo residencial de un día de otoño en Barcelona	146
5.84. Número de grupos en función de la suma de las distancias a los centroides	150
5.85. Consumo industrial de un día de otoño en Barcelona	151
5.86. Número de grupos en función de la suma de las distancias a los centroides	155
5.87. Consumo residencial de un día de invierno en Málaga	156
5.88. Número de grupos en función de la suma de las distancias a los centroides	160
5.89. Consumo industrial de un día de invierno en Málaga	161
5.90. Número de grupos en función de la suma de las distancias a los centroides	165
5.91. Consumo residencial de un día de primavera en Málaga	166
5.92. Número de grupos en función de la suma de las distancias a los centroides	170
5.93. Consumo industrial de un día de primavera en Málaga	171
5.94. Número de grupos en función de la suma de las distancias a los centroides	175
5.95. Consumo residencial de un día de verano en Málaga	176
5.96. Número de grupos en función de la suma de las distancias a los centroides	180
5.97. Consumo industrial de un día de verano en Málaga	181
5.98. Número de grupos en función de la suma de las distancias a los centroides	185

5.99. Consumo residencial de un día de otoño en Málaga	186
5.100 Número de grupos en función de la suma de las distancias a los centroides	190
5.101 Consumo industrial de un día de otoño en Málaga	191
A.1. Detalles nodo master	199
B.1. Detalles nodo worker	201
C.1. Detalles nodo edge	203
D.1. Clúster Big Data ICAI	206

Índice de tablas

5.1. Datos sobre la actividad industrial y residencial	34
5.2. Datos sobre el consumo en función del tipo de día	35
5.3. Municipios con el consumo más elevado	38
5.4. Consumo diario medio por tipo de tarifa	40
5.5. Consumo medio diario por provincias	43
5.6. Datos sobre la actividad industrial y residencial	53
5.7. Consumo industrial medio diario por día de la semana	57
5.8. Predicción del porcentaje de datos asociados a cada grupo	63
5.9. Validación del porcentaje de datos asociados a cada grupo	63
5.10. Predicción del porcentaje de datos asociados a cada grupo, sábado	66
5.11. Predicción del porcentaje de datos asociados a cada grupo, domingo	66
5.12. Validación del porcentaje de datos asociados a cada grupo, sábado	66
5.13. Validación del porcentaje de datos asociados a cada grupo, domingo	66
5.14. Predicción del porcentaje de datos asociados a cada grupo	69
5.15. Validación del porcentaje de datos asociados a cada grupo	69
5.16. Predicción del porcentaje de datos asociados a cada grupo	71
5.17. Validación del porcentaje de datos asociados a cada grupo	71
5.18. Predicción del porcentaje de datos asociados a cada grupo	73
5.19. Validación del porcentaje de datos asociados a cada grupo	73
5.20. Predicción del porcentaje de datos asociados a cada grupo	75
5.21. Validación del porcentaje de datos asociados a cada grupo	76
5.22. Predicción del porcentaje de datos asociados a cada grupo	78
5.23. Validación del porcentaje de datos asociados a cada grupo	78
5.24. Predicción del porcentaje de datos asociados a cada grupo	80
5.25. Validación del porcentaje de datos asociados a cada grupo	80
5.26. Predicción del porcentaje de datos asociados a cada grupo	83
5.27. Validación del porcentaje de datos asociados a cada grupo	83
5.28. Predicción del porcentaje de datos asociados a cada grupo, Sábado	86
5.29. Validación del porcentaje de datos asociados a cada grupo, Sábado	86
5.30. Predicción del porcentaje de datos asociados a cada grupo, Domingo	87
5.31. Validación del porcentaje de datos asociados a cada grupo, Domingo	87
5.32. Predicción del porcentaje de datos asociados a cada grupo, Diario	90
5.33. Predicción del porcentaje de datos asociados a cada grupo, Sábado	90
5.34. Predicción del porcentaje de datos asociados a cada grupo, Domingo	91
5.35. Validación del porcentaje de datos asociados a cada grupo, Diario	91
5.36. Validación del porcentaje de datos asociados a cada grupo, Sábado	91
5.37. Validación del porcentaje de datos asociados a cada grupo, Domingo	92

5.38. Predicción del porcentaje de datos asociados a cada grupo	94
5.39. Validación del porcentaje de datos asociados a cada grupo	94
5.40. Predicción del porcentaje de datos asociados a cada grupo	97
5.41. Validación del porcentaje de datos asociados a cada grupo	97
5.42. Predicción del porcentaje de datos asociados a cada grupo	99
5.43. Validación del porcentaje de datos asociados a cada grupo	100
5.44. Predicción del porcentaje de datos asociados a cada grupo	102
5.45. Validación del porcentaje de datos asociados a cada grupo	103
5.46. Predicción del porcentaje de datos asociados a cada grupo	105
5.47. Validación del porcentaje de datos asociados a cada grupo	105
5.48. Predicción del porcentaje de datos asociados a cada grupo	107
5.49. Validación del porcentaje de datos asociados a cada grupo	108
5.50. Predicción del porcentaje de datos asociados a cada grupo	110
5.51. Validación del porcentaje de datos asociados a cada grupo	110
5.52. Predicción del porcentaje de datos asociados a cada grupo	113
5.53. Predicción del porcentaje de datos asociados a cada grupo	113
5.54. Predicción del porcentaje de datos asociados a cada grupo	115
5.55. Validación del porcentaje de datos asociados a cada grupo	115
5.56. Predicción del porcentaje de datos asociados a cada grupo, Diario	119
5.57. Predicción del porcentaje de datos asociados a cada grupo, Sábado	120
5.58. Predicción del porcentaje de datos asociados a cada grupo, Domingo	120
5.59. Validación del porcentaje de datos asociados a cada grupo, Diario	120
5.60. Validación del porcentaje de datos asociados a cada grupo, Sábado	120
5.61. Validación del porcentaje de datos asociados a cada grupo, Domingo	121
5.62. Predicción del porcentaje de datos asociados a cada grupo, Diario	124
5.63. Predicción del porcentaje de datos asociados a cada grupo, Sábado	125
5.64. Predicción del porcentaje de datos asociados a cada grupo, Domingo	125
5.65. Validación del porcentaje de datos asociados a cada grupo, Diario	125
5.66. Validación del porcentaje de datos asociados a cada grupo, Sábado	125
5.67. Validación del porcentaje de datos asociados a cada grupo, Domingo	125
5.68. Predicción del porcentaje de datos asociados a cada grupo, Diario	129
5.69. Predicción del porcentaje de datos asociados a cada grupo, Sábado	129
5.70. Predicción del porcentaje de datos asociados a cada grupo, Domingo	129
5.71. Validación del porcentaje de datos asociados a cada grupo, Diario	130
5.72. Validación del porcentaje de datos asociados a cada grupo, Sábado	130
5.73. Validación del porcentaje de datos asociados a cada grupo, Domingo	130
5.74. Predicción del porcentaje de datos asociados a cada grupo, Diario	134
5.75. Predicción del porcentaje de datos asociados a cada grupo, Sábado	134
5.76. Predicción del porcentaje de datos asociados a cada grupo, Domingo	134
5.77. Validación del porcentaje de datos asociados a cada grupo, Diario	134
5.78. Validación del porcentaje de datos asociados a cada grupo, Sábado	135
5.79. Validación del porcentaje de datos asociados a cada grupo, Domingo	135
5.80. Predicción del porcentaje de datos asociados a cada grupo, Diario	138
5.81. Predicción del porcentaje de datos asociados a cada grupo, Sábado	138
5.82. Predicción del porcentaje de datos asociados a cada grupo, Domingo	139

5.83. Validación del porcentaje de datos asociados a cada grupo, Diario	139
5.84. Validación del porcentaje de datos asociados a cada grupo, Sábado	139
5.85. Validación del porcentaje de datos asociados a cada grupo, Domingo	139
5.86. Predicción del porcentaje de datos asociados a cada grupo, Diario	142
5.87. Predicción del porcentaje de datos asociados a cada grupo, Sábado	143
5.88. Predicción del porcentaje de datos asociados a cada grupo, Domingo	143
5.89. Validación del porcentaje de datos asociados a cada grupo, Diario	143
5.90. Validación del porcentaje de datos asociados a cada grupo, Sábado	143
5.91. Validación del porcentaje de datos asociados a cada grupo, Domingo	144
5.92. Predicción del porcentaje de datos asociados a cada grupo, Diario	147
5.93. Predicción del porcentaje de datos asociados a cada grupo, Sábado	148
5.94. Predicción del porcentaje de datos asociados a cada grupo, Domingo	148
5.95. Validación del porcentaje de datos asociados a cada grupo, Diario	148
5.96. Predicción del porcentaje de datos asociados a cada grupo, Sábado	148
5.97. Validación del porcentaje de datos asociados a cada grupo, Domingo	149
5.98. Predicción del porcentaje de datos asociados a cada grupo, Diario	152
5.99. Predicción del porcentaje de datos asociados a cada grupo, Sábado	153
5.100. Predicción del porcentaje de datos asociados a cada grupo, Domingo	153
5.101. Validación del porcentaje de datos asociados a cada grupo, Diario	153
5.102. Validación del porcentaje de datos asociados a cada grupo, Sábado	153
5.103. Validación del porcentaje de datos asociados a cada grupo, Domingo	154
5.104. Predicción del porcentaje de datos asociados a cada grupo, Diario	158
5.105. Predicción del porcentaje de datos asociados a cada grupo, Sábado	158
5.106. Predicción del porcentaje de datos asociados a cada grupo, Domingo	158
5.107. Validación del porcentaje de datos asociados a cada grupo, Diario	158
5.108. Validación del porcentaje de datos asociados a cada grupo, Sábado	159
5.109. Validación del porcentaje de datos asociados a cada grupo, Domingo	159
5.110. Predicción del porcentaje de datos asociados a cada grupo, Diario	163
5.111. Predicción del porcentaje de datos asociados a cada grupo, Sábado	163
5.112. Predicción del porcentaje de datos asociados a cada grupo, Domingo	163
5.113. Validación del porcentaje de datos asociados a cada grupo, Diario	163
5.114. Validación del porcentaje de datos asociados a cada grupo, Sábado	164
5.115. Validación del porcentaje de datos asociados a cada grupo, Domingo	164
5.116. Predicción del porcentaje de datos asociados a cada grupo, Diario	168
5.117. Predicción del porcentaje de datos asociados a cada grupo, Sábado	168
5.118. Predicción del porcentaje de datos asociados a cada grupo, Domingo	168
5.119. Validación del porcentaje de datos asociados a cada grupo, Diario	168
5.120. Validación del porcentaje de datos asociados a cada grupo, Sábado	169
5.121. Validación del porcentaje de datos asociados a cada grupo, Domingo	169
5.122. Predicción del porcentaje de datos asociados a cada grupo, Diario	173
5.123. Predicción del porcentaje de datos asociados a cada grupo, Sábado	173
5.124. Predicción del porcentaje de datos asociados a cada grupo, Domingo	173
5.125. Validación del porcentaje de datos asociados a cada grupo, Diario	173
5.126. Validación del porcentaje de datos asociados a cada grupo, Sábado	174
5.127. Validación del porcentaje de datos asociados a cada grupo, Domingo	174

5.128	Predicción del porcentaje de datos asociados a cada grupo, Diario	178
5.129	Predicción del porcentaje de datos asociados a cada grupo, Sábado	178
5.130	Predicción del porcentaje de datos asociados a cada grupo, Domingo	178
5.131	Validación del porcentaje de datos asociados a cada grupo, Diario	178
5.132	Validación del porcentaje de datos asociados a cada grupo, Sábado	179
5.133	Validación del porcentaje de datos asociados a cada grupo, Domingo	179
5.134	Predicción del porcentaje de datos asociados a cada grupo, Diario	183
5.135	Predicción del porcentaje de datos asociados a cada grupo, Sábado	183
5.136	Predicción del porcentaje de datos asociados a cada grupo, Domingo	183
5.137	Validación del porcentaje de datos asociados a cada grupo, Diario	183
5.138	Validación del porcentaje de datos asociados a cada grupo, Sábado	184
5.139	Validación del porcentaje de datos asociados a cada grupo, Domingo	184
5.140	Predicción del porcentaje de datos asociados a cada grupo, Diario	188
5.141	Predicción del porcentaje de datos asociados a cada grupo, Sábado	188
5.142	Predicción del porcentaje de datos asociados a cada grupo, Domingo	188
5.143	Validación del porcentaje de datos asociados a cada grupo, Diario	188
5.144	Predicción del porcentaje de datos asociados a cada grupo, Sábado	189
5.145	Validación del porcentaje de datos asociados a cada grupo, Domingo	189
5.146	Predicción del porcentaje de datos asociados a cada grupo, Diario	193
5.147	Predicción del porcentaje de datos asociados a cada grupo, Sábado	193
5.148	Predicción del porcentaje de datos asociados a cada grupo, Domingo	193
5.149	Validación del porcentaje de datos asociados a cada grupo, Diario	193
5.150	Validación del porcentaje de datos asociados a cada grupo, Sábado	193
5.151	Validación del porcentaje de datos asociados a cada grupo, Domingo	194
F.2.	Consumo total anual de todos los municipios	245

1. Introducción

El término Big Data hace referencia a un conjunto de datos cuyo tamaño, complejidad y tasa de crecimiento, implican grandes dificultades en su gestión y análisis, mediante tecnologías y herramientas tradicionales,[1]. La utilidad del Big Data recae en la capacidad de responder a preguntas que las empresas ni siquiera se habían formulado. La gran cantidad de datos permite a las organizaciones buscar tendencias, información que será empleada en la identificación de problemas y en la toma de decisiones. En la actualidad, numerosas organizaciones utilizan el Big Data con el fin de obtener información acerca de cómo reducir costes, lanzar nuevos productos y servicios o mejorar el rendimiento.

Ante esta nueva tendencia en el mercado actual, nos debemos plantear la siguiente pregunta, ¿En qué sectores se pueden aplicar técnicas de procesamiento Big Data? La respuesta es bien sencilla, en aquellos sectores que de una manera u otra recauden información sobre sus clientes. Esta información puede abarcar desde el mero hecho de la realización de una compra de un determinado producto en un determinado momento, a datos más detallados acerca del tiempo que un cliente ha pasado navegando en una página web y de ese tiempo que fracción corresponde a qué producto.

Una vez se ha recaudado una cantidad de datos considerable, las técnicas de procesamiento Big Data entran en juego. El objetivo será conseguir cualquier tipo de información que permita a las empresas crear perfiles más personalizados acerca de sus clientes. Una vez se encuentren en poder de este conocimiento, las empresas tienen la capacidad de lanzar ofertas o productos orientados a un sector concreto de la población.

El análisis Big Data se está adentrando, poco a poco, en todos los sectores del mercado, debido a la gran cantidad de información útil derivable de las grandes cantidades de datos. Siguiendo con la tendencia actual, no es de extrañar la introducción de técnicas de procesamiento Big Data en el sector eléctrico.

Gracias a la obtención de patrones de consumo en el sector eléctrico, es posible determinar cómo, el consumo varía a lo largo del tiempo. Una vez obtenida dicha información, se puede estudiar como distintos factores afectan al mismo. Agentes a tener en consideración incluyen, el clima, la estación del año, el nivel de ingresos de las familias o empresas, los hábitos de consumo,... Si los datos de los que se disponen comprenden un intervalo de tiempo del orden de lustros o décadas, se puede examinar como factores tecnológicos, como la introducción de automóviles eléctricos, afectan el consumo de la electricidad.

La cantidad de datos de la que se dispone solamente abarca un intervalo de tiempo de poco más de un año, por lo que este tipo de conclusiones se escapan del marco de este proyecto. El objetivo principal de este trabajo es la detección de perfiles de consumo en el mercado eléctrico. Una vez obtenida esta información, las empresas eléctricas tendrán la capacidad de predecir cuánta energía eléctrica es necesaria y saber si se pueden rebajar los picos o puntas de uso de la energía motivando al usuario para que consuma a otras horas. Además se podrán lanzar productos y campañas de marketing mucho más focalizadas en un tipo de cliente determinado.

2. Descripción De Las Tecnologías

De las tecnologías utilizadas en la realización de este proyecto, caben destacar las siguientes. En primer lugar, el uso de Jupyter Notebook, como entorno de desarrollo, donde el lenguaje de programación elegido ha sido Python, debido al gran número de librerías que posee. De entre las librerías utilizadas cabe destacar *pyspark*, que permite disfrutar del modelo Spark en Python. Además, se han utilizado algoritmos de Machine Learning, como el algoritmo K-Means.

2.1. Jupyter Notebook

El código generado en la realización de este proyecto, se ha desarrollado utilizando Jupyter Notebook. Jupyter Notebook es una aplicación web open-source que permite crear y compartir documentos que contienen código fuente, ecuaciones, visualizaciones y texto. Entre sus posibles usos caben destacar, la limpieza y la transformación de datos, simulaciones numéricas, modelos estadísticos, visualización de información y Machine Learning,[2].

Las libretas están estructuradas como una secuencia de celdas, que contienen secciones de código. Cada una de las celdas que forman una libreta se pueden ejecutar en el orden deseado, lo que permite relacionar cada una de ellas con una operación determinada. De esta forma, el desarrollo se convierte en un proceso mucho más interactivo, ya que después de la ejecución de cada celda el resultado aparecerá mostrado debajo de la misma.

2.2. Spark

Spark es un motor de procesamiento de datos que puede ser utilizado en un gran rango de escenarios. En muchos casos, es utilizado para realizar de forma rápida consultas, análisis y transformaciones sobre los datos, [3]. Fue creado en 2009 en la Universidad de Berkely, desde entonces ha ido ganando popularidad como motor de procesamiento Big Data y suplantando el puesto de tecnologías anteriores. Hoy en día forma parte de Apache.

Spark es capaz de lidiar con grandes cantidades de información, del orden de petabytes por segundo, [3]. Está estructurado como un sistema de computación distribuido a través de numerosos clústers. Además, soporta una gran cantidad de lenguajes entre los que se encuentran Java, Scala, R o Python, que ha sido el escogido para el desarrollo de este proyecto.

Los principales usos de esta tecnología son los siguientes:

- Spark Streaming and Processing: cuya función es lidiar con grandes flujos de datos y analizarlos a medida que llegan.
- Machine Learning: utilizado para la extracción de patrones y predicciones.
- Análisis Interactivo del Flujo: proporciona la capacidad de modificar las consultas sobre los datos a medida que se obtiene resultados de consultas anteriores.
- Integración de Datos: permite combinar datos procedentes de distintas fuentes de información.

2.3. Machine Learning

Machine Learning (aprendizaje automático), es una aplicación derivada de la Inteligencia Artificial, que permite a un sistema aprender automáticamente y mejorar a través de la experiencia, sin la necesidad de ser específicamente programado, [4]. Este proceso se basa en la aplicación de algoritmos sobre un conjunto de datos que permiten derivar patrones y obtener un modelo de predicción, de acuerdo con ciertas características especificadas. Estos algoritmos se pueden clasificar en cuatro grandes grupos, algoritmos supervisados, algoritmos no supervisados, algoritmos semi-supervisados y algoritmos de refuerzo, [4].

Los algoritmos supervisados, son aquellos que son capaces de extrapolar lo aprendido con datos de entrenamiento a nuevos conjuntos de datos. En la fase de entrenamiento se entregan, al algoritmo, parejas de datos y etiquetas, de esta forma aprende a distinguir las características que diferencian a los datos pertenecientes a cada uno de los grupos. Si el entrenamiento ha sido completado con éxito, el algoritmo sabrá asignar nuevos datos a sus correspondientes grupos. A este grupo pertenecen algoritmos como árboles de decisión o Naïve Bayes.

Los algoritmos no supervisados, son aquellos que utilizan información no clasificada durante la fase de entrenamiento. Estos algoritmos se utilizan para descubrir patrones de datos no clasificados. A este grupo pertenecen el algoritmo K-Means, que es explicado en detalle en el apartado 4.3.1.

Los algoritmos semi-supervisados, forman una categoría comprendida entre los algoritmos supervisados y los no supervisados. Se caracterizan por el hecho de que, normalmente, utilizan una pequeña porción de datos no clasificados y un gran porcentaje de datos no clasificados. La ventaja de estos algoritmos recae en el hecho de que el proceso de clasificación de los datos consume tanto tiempo como recursos. Además, el propio proceso de clasificación puede sesgar los datos. Por este motivo, los algoritmos semi-supervisados, mejora la detección de patrones, [5].

Los algoritmos de refuerzo, utilizan un proceso a base de prueba y error para descubrir las características de un sistema. El objetivo del aprendizaje por refuerzo es desenmascarar que acciones deben de ser realizadas en distintos momentos para maximizar la recompensa obtenida,[6].

3. Estado De La Cuestión

El objetivo principal de este proyecto es encontrar patrones en el consumo eléctrico en España, mediante la utilización de técnica de procesamiento Big Data. Por lo que esta sección se centrará en situar la evolución el sector eléctrico en España, así como el origen de las técnicas de procesamiento Big Data.

3.1. El sector eléctrico en España

Hasta el año 1998, la actividad del sector eléctrico en España estaba dividida entre unas pocas empresas que poseían monopolios en distintas regiones. A partir de la aprobación de la Ley 54/1997, del 27 de noviembre, el mercado eléctrico se abrió a terceros, lo que supuso la creación de un mercado basado en la negociación de la energía y la reducción de la intervención pública, [7]. A raíz de la liberación del sector eléctrico, la competencia dentro del mismo, se vio incrementada.

El suministro de energía eléctrica se define como la entrega de energía a través de las redes de transporte y distribución mediante contraprestación económica en las condiciones de regularidad y calidad que resulten exigibles,[7]. El suministro de energía consta de diferentes actividades. En primer lugar, la generación de la energía, que ha de ser transportada hasta las redes de distribución, para luego poder comercializarse.

El grupo Red Eléctrica, es el encargado de suministrar electricidad en España. Dentro del grupo se encuentra Red Eléctrica de España, que es el responsable de la actividad eléctrica en España, representado un 92 % de la actividad del grupo,[8].

Desde el año 2014, el grupo Red Eléctrica, está siguiendo las bases de un plan estratégico que terminará en el 2019. Este plan se basa en mejorar la eficiencia de la gestión del TSO (Gestor de Red de Transporte) y de las redes de fibra óptica. El crecimiento está siendo potenciado mejorando las herramientas de la operación del sistema, así como apostando por nuevos mercados rentables,[9]. En este último ámbito, se centra el proyecto, cuyo objetivo es encontrar patrones en el consumo eléctrico en España. De esta forma las empresas podrán diseñar estrategias que les ayuden a alcanzar sus objetivos, teniendo como referencia las tendencias del consumo actual.

3.2. Técnicas de procesamiento Big Data

La necesidad de las técnicas de procesamiento Big Data, recae en el hecho de que las herramientas y técnicas de procesamiento de datos tradicionales, no fueron diseñadas para lidiar con los volúmenes de información manejados actualmente. Ante esta situación, se han desarrollado nuevos procedimientos de tratamiento y almacenamiento de datos, adaptados a los grandes volúmenes de datos utilizados en la actualidad.

Big Data es un término que hace referencia a cualquier volumen de datos, ya sean estructurados, semiestructurados o no estructurados, que tienen el potencial de ser recopilados para obtener información, [10]. Además, los conjuntos de datos que denominamos Big Data se definen en función de cuatro características fundamentales: el volumen de datos, que puede llegar al orden de petabyte o exabytes; la variedad de fuentes y formatos de almacenamiento, la veracidad de la información y, finalmente, por la velocidad con la que se procesan.

El procesamiento Big Data consta de dos fases principales, la adquisición de datos y el procesamiento de los mismos. A su vez, cada una de estas fases puede subdividirse en tareas más concretas, a continuación, se describe este proceso de manera más detallada, [11].

1. Etapa de preparación:

- Fase inicial: el objetivo de esta fase es determinar la rentabilidad y factibilidad de la implementación de una solución Big Data. Para ello es necesario tener en cuenta el tipo de información que se quiere obtener, así como los recursos necesarios.
- Fase de planificación: en esta etapa se definen los recursos que se van a emplear.
- Fase de diseño: se desarrolla tanto la infraestructura como la arquitectura necesaria para soportar el procesamiento de los datos.
- Fase de implementación: despliegue de la infraestructura anteriormente diseñada.

2. Etapa de recopilación, análisis de datos y generación del valor

- Fase de recopilación: se evalúan los datos obtenidos y se estudia su viabilidad, así como su utilidad.
- Fase de análisis: se aplican distintos algoritmos sobre el conjunto de datos, con el objetivo de encontrar la información definida en la fase inicial.
- Fase de agregación de valor: en esta etapa los resultados obtenidos tras la aplicación de los algoritmos son analizados con el fin de extraer resultados relevantes al objeto de estudio.

Con el fin de llevar acabo este proceso, es necesario contar con los siguientes elementos de una arquitectura Big Data,[11].

- Fuentes: páginas web, transacciones bancarias, registros históricos de la compañía, ...
- Capa de almacenamiento: encargada de recoger los datos y guardarlos hasta su procesamiento.

- Capa de análisis: lleva a cabo el tratamiento de los datos empleando los algoritmos pertinentes.
- Capa de consumo: extracción de conclusiones.

Para desarrollar el procedimiento detallado anteriormente, es necesario disponer de las herramientas adecuadas. Las tecnologías Big Data se pueden clasificar del siguiente modo,[12].

- Motores de procesamiento: realizan cálculos en sistemas distribuidos.
- Ingestión: sistemas de recolección de datos.
- Mensajería: intercambio de información entre diferentes elementos software.
- Almacenamiento: guardan el gran volumen de datos recolectado.
- Bibliotecas de consulta: facilitan el acceso a los datos.
- Gestión de recursos: planificación y asignación de recursos dentro del clúster.
- Bibliotecas de Machine Learning: implementación de diversos algoritmos para llevar a cabo la clasificación y el análisis de los datos.

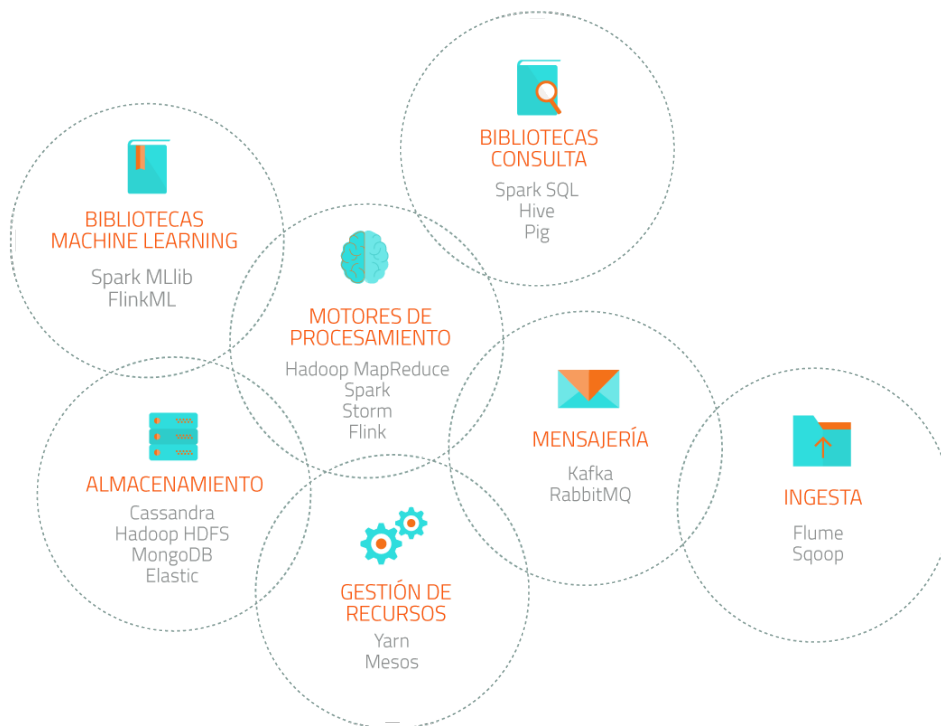


Figura 3.1.: Herramientas Big Data

4. Descripción Del Modelo Desarrollado

4.1. Objetivos y Especificación

El objetivo principal de este proyecto, es la identificación de patrones de comportamiento en el consumo eléctrico de España, en función de la provincia, la estación del año, el tipo de actividad industrial, el día de la semana y la hora del día. Las provincias elegidas para este estudio son Barcelona, Madrid y Málaga. Dichas provincias han sido elegidas de forma arbitraria, teniendo como meta el estudio de provincias localizadas en distintas zonas geográficas, así como con niveles población diferentes. La provincia de Barcelona constituye el nexo entre las tres provincias, por un lado presenta una gran población, característica que la relaciona con la provincia de Madrid, además, se encuentra localizada en la costa, lo que la relaciona con la provincia de Málaga. Con el fin de alcanzar dicho objetivo, es necesario realizar un estudio previo acerca de los datos objeto de estudio.

En primer lugar, se realizará un análisis descriptivo sobre todo el conjunto de datos. La finalidad del análisis descriptivo es conocer, a grandes rasgos, cómo distintos factores afectan al consumo eléctrico. Para ello se desarrollarán diversos ejecutables que permitan la extracción de información acerca del consumo dependiendo de varios factores, mediante diversos tipos de gráficos entre los cuales caben destacar, el diagrama de barras, el diagramas de sectores y el diagrama de cajas. Las variables independientes que se emplearán en esta sección del proyecto son: el tipo de actividad económica (residencial o industrial), el día de la semana (día laboral, sábado o domingo), la hora del día, el municipio y el tipo de tarifa de mercado.

Dado que el objetivo fundamental de este estudio es la obtención de patrones de consumo para distintas provincias, estaciones, día y tipos de actividades industriales a lo largo del día, se realizará un estudio de mayor envergadura acerca de algunos de estos factores. Primero, el estudio se centrará en obtener datos acerca de como cada factor se afecta a cada una de las tres provincias objeto de estudio. A continuación, se llevará a cabo el mismo análisis, pero esta vez en torno al tipo de actividad económica. Con el estudio, en profundidad, de estos dos factores, ya se cubren todas las facetas relevantes del resto de factores, por lo que no será necesario dedicarles una sección.

Por último, se elaborarán los patrones que rigen el consumo eléctrico en función de la provincia (Barcelona, Madrid y Málaga), el tipo de actividad económica (residencial e industrial), el tipo de día de la semana (día laboral, sábado y domingo) a lo largo del día. Para ello se empleará un algoritmo de agrupamiento denominado K-Means. Una vez obtenidos los distintos grupos de comportamiento, para cada uno de los casos planteados, se podrá trasladar dicha información a las

distribuidoras eléctricas. A las empresas energéticas les conviene aplanar el gráfico del consumo a lo largo del día, para poder optimizar las centrales de producción de energía. Con la información obtenida tras la implementación del algoritmo K-Means, pueden plantearse el lanzamiento de determinadas ofertas a grupos de clientes específicos, con la finalidad de redistribuir su consumo a lo largo del día y así alcanzar un consumo más estable a lo largo del día.

4.2. Datos

Se disponen de datos del consumo eléctrico en España desde el 12 de septiembre de 2013 hasta el 19 de octubre de 2015, es decir, poco más de dos años. A pesar de que los datos fueron recogidos hace tres años, siguen teniendo validez para este proyecto, ya que se trata de ilustrar la importancia de las técnicas de Machine Learning en el sector eléctrico, a través de la extracción de información útil a nivel empresarial.

Cada una de las entradas de la base de datos que contiene el consumo eléctrico en España, dispone de la siguiente información útil:

- DÍA: indica el día en el que el dato fue recolectado.
- HX: hora del día a la que corresponde, con $X \in [1, 25]$.
- ACTIVA_HX: consumo de energía activa, con $X \in [1, 25]$, en adelante toda referencia al consumo eléctrico se centrará en la energía activa.
- REACTIVA_HX: consumo de energía reactiva, con $X \in [1, 25]$.
- DE_MUNICIP: municipio donde se recogió el dato de consumo.
- CNAE: Clasificación Nacional de Actividades Económicas, indica el origen del consumo, ya sea industrial o residencial
- MERCADO: cliente con una tarifa regulada o con tarifa de mercado libre

Todas estas variables, serán las utilizadas a lo largo de este proyecto para filtrar el volumen de datos de acuerdo con los parámetros empleados en cada sección. La base de datos de la que se dispone, posee más columnas que no guardan relación con el estudio a realizar, por lo que se han excluido completamente. Además, a lo largo del proyecto se necesitarán más variables como la de estación, tipo del día o provincia, estas columnas serán creadas y utilizadas dependiendo de las necesidades encontradas en cada uno de los casos de estudio.

4.3. Algoritmos

A continuación, se explica en detalle el algoritmo utilizado en la realización de este proyecto.

4.3.1. Algoritmo K-Means

El algoritmo K-Means, es un algoritmo no supervisado de clústering, es decir, tiene como objetivo agrupar los datos en k categorías distintas en función de unos parámetros determinados. El agrupamiento de los datos se realiza a través de un proceso iterativo que tiene por objetivo minimizar la suma de las distancias cuadráticas de cada uno de los datos al centro del clúster al que pertenecen, [13].

$$\sum_{j=1}^k \sum_{i=1}^n ||x_i^{(j)} - c_j||^2 \quad (4.1)$$

Donde $x_i^{(j)}$ representa un punto del conjunto de datos, c_j representa el centro del clúster, k el número de clústers escogido y n el número total de puntos que forman el conjunto de datos.

Los pasos que comprenden el proceso iterativo son los siguientes:

1. k centros son colocados aleatoriamente sobre el plano que forma el conjunto de datos.
2. Se asigna cada elemento del grupo de datos al centro más cercano
3. Se optimiza la posición de cada uno de los centros, minimizando la distancia cuadrática total
4. Los pasos 2 y 3 se repiten hasta que los centros se vuelven inamovibles.

Escoger el número de clústers con los que el algoritmo va a trabajar, es una decisión de gran relevancia ya que las agrupaciones encontradas depende en gran medida de ello. Si el número de clústers elegido es muy pequeño, la varianza de cada uno de los datos con respecto al centro del clúster al que pertenecen aumentará. Por el contrario, si el número de clústers es muy pequeño la varianza disminuirá, siendo cero en el caso extremo en el que el número de clústers es igual al número de muestras. Un método que sirve para determinar el número de clústers apropiado, consiste en representar la suma de las distancias euclídeas de cada punto a su centroide correspondiente en función del número de clústers empleados. El número de clústers deseado será aquel que se corresponda con el codo del gráfico, como se muestra en la siguiente imagen, [14].

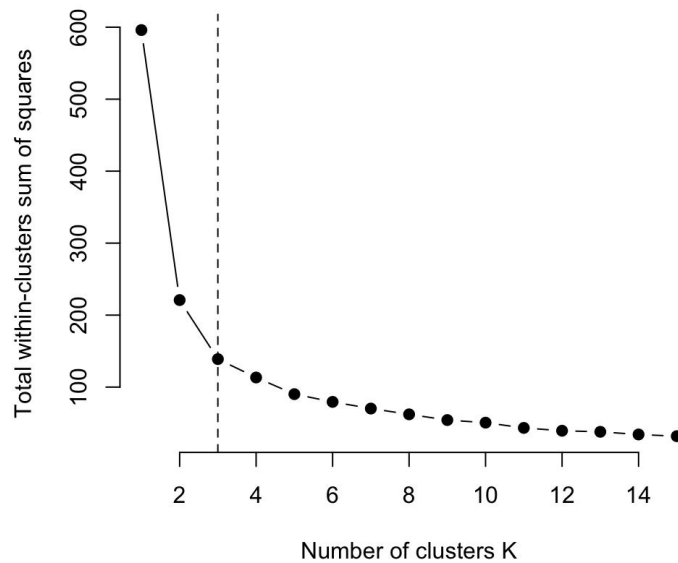


Figura 4.1.: Número de clústers Vs Suma de las distancias euclídeas

La mayor desventaja de este algoritmo recae en el hecho de que el resultado final depende, en gran medida, de la posición inicial de los centros y del número de clústers elegido. Es decir, si se ejecuta, varias veces, este algoritmo sobre un mismo conjunto de datos y con un número fijo de clústers, se obtendrán resultados distintos, ya que la posición inicial diferirá en cada ejecución. Por lo tanto, con la utilización de este algoritmo se puede obtener una posición de centros que proporciona una distancia cuadrática total mínima local, no global. Para mitigar este efecto, se debe ejecutar varias veces antes de extraer conclusiones de los datos obtenidos, [13].

4.4. Plataforma de desarrollo

El objetivo principal de esta sección es confeccionar una imagen acerca de todos los elementos que han hecho posible el desarrollo de este proyecto, tanto desde el punto de vista de la arquitectura como de la implementación.

En primer lugar, es necesario entender el framework sobre el que se basa todo el proyecto. Apache Spark es un framework computacional distribuido de código abierto y propósito general con un motor de procesamiento de datos en memoria capaz de extraer, transformar y cargar datos (ETL), análisis de datos, machine learning y procesamiento gráfico sobre grandes volúmenes de datos, ya estén en reposo o no, mediante el uso de APIs concisas de alto nivel para los lenguajes de programación: Scala, Python (el empleado en este proyecto), Java, R y SQL, [15]. Además, Spark se puede ejecutar en local, en clústers o en la nube, el clúster empleado en este proyecto se ejecuta sobre Hadoop YARN.

Desde un punto de vista de alto nivel, una aplicación Spark crea RDDs partiendo de ciertos datos de entrada, para después realizar las operaciones oportunas y devolver o almacenar los resultados. Resilient Distributed Dataset (RDD) es la abstracción fundamental de Apache Spark. Una RDD es una colección distribuida, es decir, los datos son guardados en distintas particiones.

De entre las características que poseen este tipo de colecciones, destacan las siguientes, [15]:

- Inmutable: una RDD no cambia una vez creada, para modificarla es necesario crear una nueva.
- Evaluación perezosa: las RDDs definen el flujo de información a seguir, pero no cuando se ejecuta.
- Paralelismo: los datos son procesados en paralelo.
- Cacheable: los datos se pueden guardar de forma persistente en memoria o en disco, que es más rápido.

4.4.1. Arquitectura

Spark emplea una arquitectura basada en la relación maestro/esclavo, [15]. Cada aplicación de Spark llama al módulo SparkContext, que representa la conexión al clúster de Spark y a través del cual se puede interactuar con las variables guardadas en el mismo, [16]. El contexto de Spark se comunica con el encargado de coordinar todos los procesos, el clúster Manager (master node). La función principal del clúster Manager es la distribución de los recursos entre todas las aplicaciones activas. Una vez conectado, Spark puede crear ejecutores, encargados de la computación y del almacenamiento de datos, en los nodos trabajadores (worker node) del clúster. A continuación, manda el código de la aplicación a los ejecutores. Finalmente, SparkContext manda tareas que serán procesadas por los ejecutores, [17]. La imagen de la Figura 4.2, refleja la arquitectura descrita.

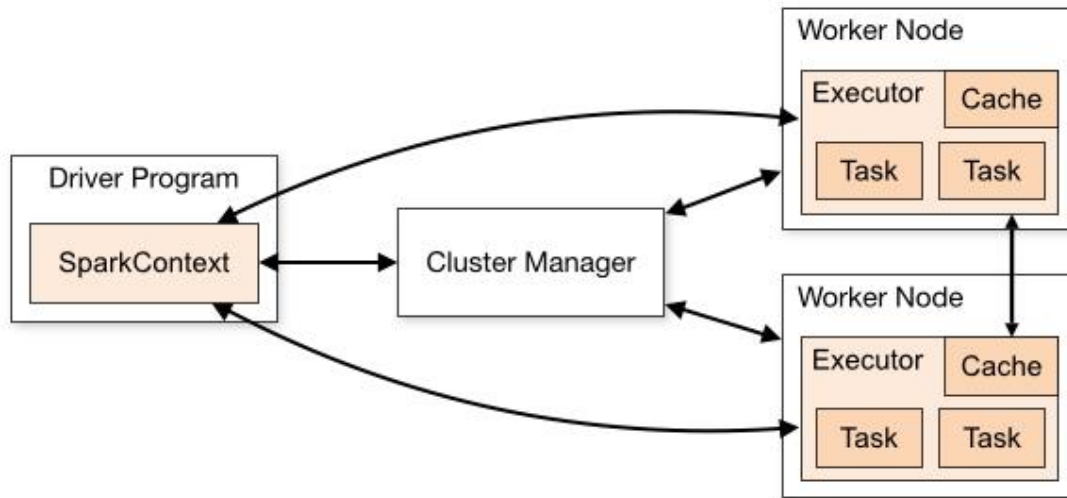


Figura 4.2.: Arquitectura Spark

Las principales ventajas de esta arquitectura son la siguientes, [17]:

- Cada aplicación tiene sus propios ejecutores, lo que permite aislar una aplicación de otra.
- Spark soporta una gran cantidad de administradores de clústers.
- SparkContext debe ser accesible desde todos los nodos que forman el clúster, siendo preferible que formen parte de la misma LAN (Local Area Network).

El clúster empleado en este proyecto es el Clúster Big Data ICAI. Dicho clúster posee 1 TB de memoria RAM, 96 TB de disco duro y 192 cores, todo ello repartido en ocho servidores Huawei. La arquitectura del clúster está compuesta por tres nodos master y cuatro nodos worker. Además, presenta un nodo edge. Este nodo funciona como una pasarela de salida y punto de conexión para los usuarios finales. En el esquema de la Figura 4.5, muestra la arquitectura seguida por el clúster empleado. La documentación referente a los nodos empleados, así como una imagen del rack empleado, se encuentra en los anexos A, B, C y D.

Con el fin de poder monitorizar las aplicaciones que se están ejecutando en el clúster, Hadoop ofrece una plataforma web dedicada a ello. A través de la misma, se puede acceder a información relacionada con la cantidad de recursos que se están empleado a nivel general y por cada aplicación.

La siguiente imagen, Figura 4.3, es una captura de pantalla de la ventana de la plataforma web de Hadoop que muestra una lista sobre las aplicaciones que se están ejecutando, que, en este caso, es una llamada “Consumos Electricos I”. En la parte superior de dicha ventana, se recoge la información relativa a los recursos empleados por el clúster, bajo la etiqueta “Cluster Metrics”. Este apartado de la ventana, indica que la aplicación que se está ejecutando está empleando 2.50 GB de los 400 GB disponibles y 2 vcores de los 256 vcores disponibles. A través del link que se encuentra en la columna “ID”, se puede acceder a información más específica acerca dicha aplicación.

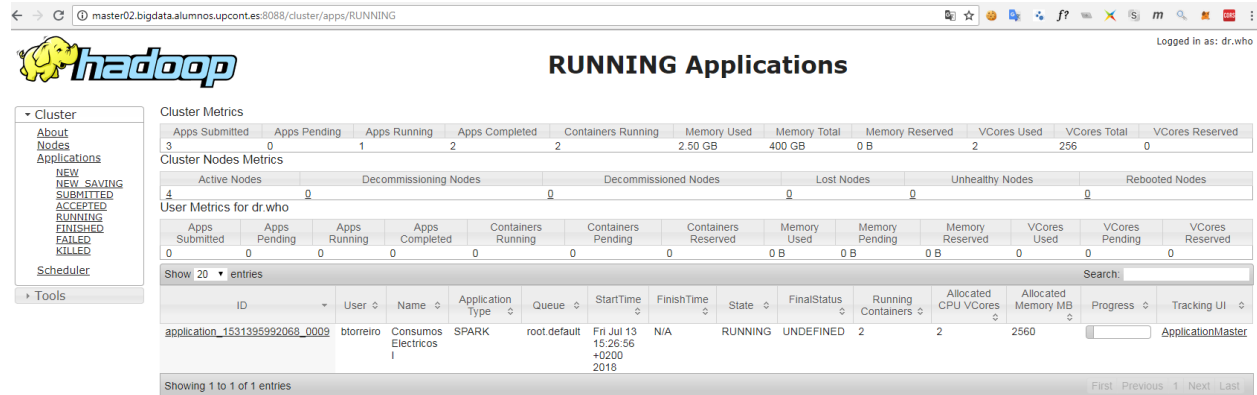


Figura 4.3.: Running Applications, plataforma web de Hadoop

Al hacer click sobre el enlace de la columna “ID”, se accede a la página mostrada en la Figura 4.4. Esta ventana muestra información acerca del usuario que lanzó la aplicación, la hora a la que empezó la ejecución, cuánto tiempo lleva y los recursos que ha empleado hasta el momento. Además, en la tabla de la parte inferior, se indica el nodo sobre el que se está ejecutando, que en este caso el nodo worker número uno.

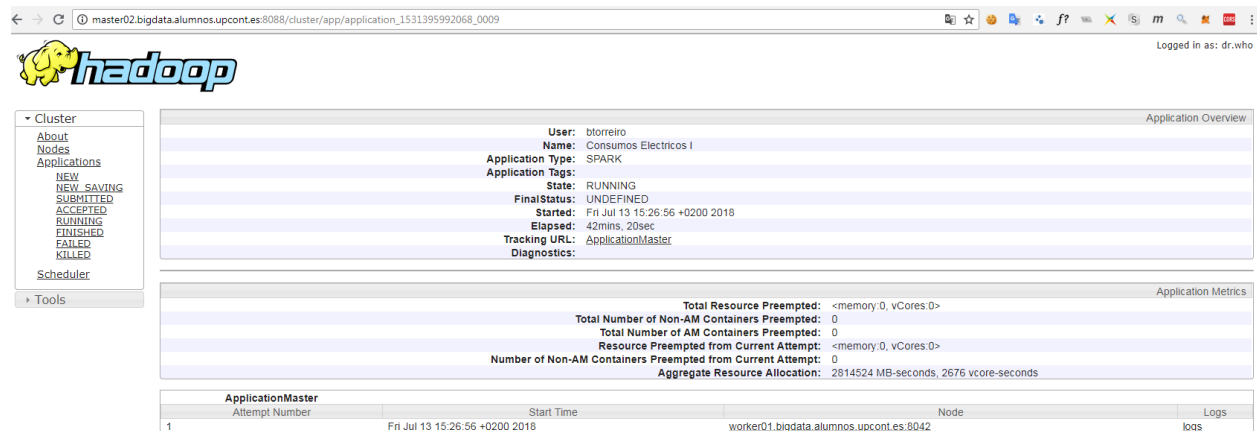


Figura 4.4.: Información de aplicación, plataforma web de Hadoop

La información mostrada en la plataforma web de Hadoop, también es accesible desde el terminal, como es de esperar. Hadoop YARN proporciona una librería con la que se puede acceder a información relativa tanto a los procesos que están en ejecución como aquellos que ya han terminado. Además, posibilita el lanzamiento de órdenes mediante comandos que permiten, entre otras cosas, llevar a cabo tareas de administración o matar aplicaciones. A continuación, se adjuntan dos ejemplos de comandos lanzados a YARN mientras la aplicación “Consumos Electricos I” está activa.

El siguiente comando devuelve una lista de las aplicaciones activas. Dado que solo se está ejecutando una aplicación, solamente aparece lista la aplicación “Consumos Electricos I”. Además, la información mostrada, relativa a la aplicación, es la misma que en la ventana principal de la plataforma web de Hadoop, Figura 4.3.

```
[root@master01 ~]# yarn application -list
```

Total number of applications (application-types: [] and states: [SUBMITTED, ACCEPTED, RUNNING]):1

Application-Id	Application-Type	User	Queue	Application-Name	State
application_1531395992068_0009	SPARK	btorreiro	root.default	Consumos Electricos I	RUNNING
Final-State	Progress	Tracking-URL			
UNDEFINED	10%	http://192.168.80.33:4040			

El ejemplo inferior, muestra el comando que se utiliza para matar las aplicaciones. Para ejecutar dicho comando se ha tomado el ID que representa la aplicación “Consumos Electricos I”.

```
[root@master01 ~]# yarn application -kill
application_1531395992068_0009
```

Killing application application_1531395992068_0009
 18/07/13 16:12:35 INFO impl.YarnClientImpl:
 Killed application application_1531395992068_0009

DAGScheduler

Un elemento primordial a la hora de ejecutar cualquier aplicación es el planificador, ya que debe establecer el orden en el que se van a ejecutar cada una de las acciones necesarias. DAGScheduler es la capa de planificación de Apache Spark que implementa la planificación basada en etapas. Transforma un plan de ejecución lógico en uno físico, [15]. El módulo DAGScheduler es creado durante el proceso de inicialización de SparkContext. Cuando se llama a una acción, SparkContext entrega la planificación lógica a DAGScheduler, que se encargará de crear y seguir las etapas necesarias. El DAGScheduler recibe dicho nombre ya que, para cada trabajo, computa un grafo acíclico dirigido, que de sus siglas en inglés se conoce como DAG.

El DAGScheduler se basa en dos conceptos básicos, trabajos y etapas, a los que controla a través de registros y contadores internos. Además, se encarga de elegir el nodo en el que se van a ejecutar las próximas tareas en función del estado del caché. También lleva la cuenta de las RDDs que son cacheadas para evitar la repetición de tareas.

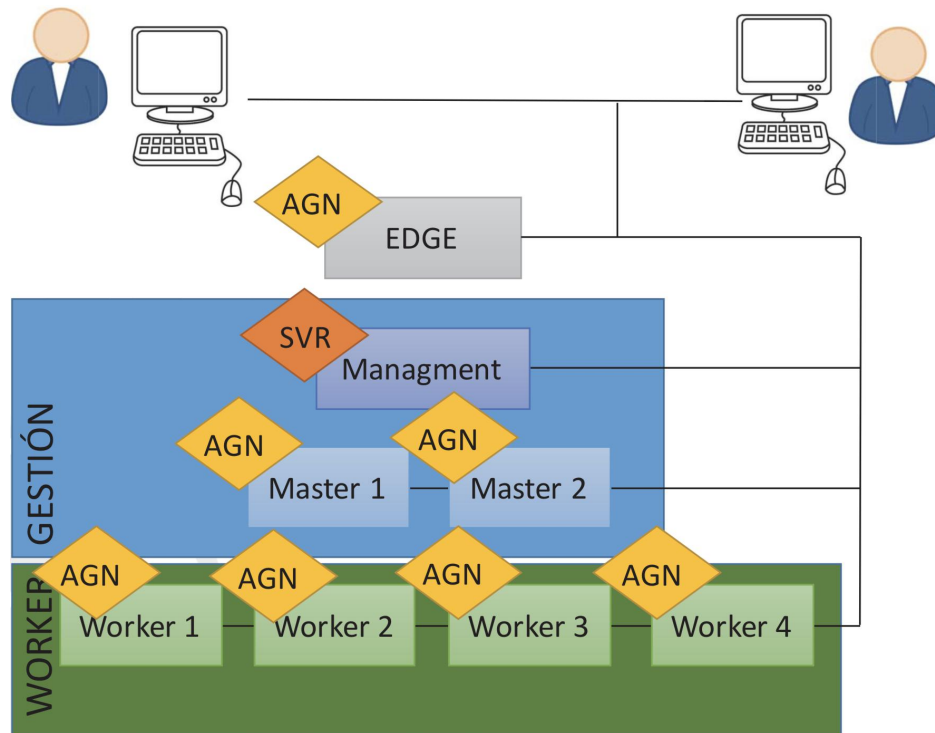


Figura 4.5.: Clúster Big Data ICAI

Todo este proceso puede ser seguido gracias a la plataforma web de Spark. La interfaz web permite monitorizar e interactuar con los distintos trabajos que se están o han sido ejecutados. Esta plataforma es lanzada cada vez que se crea una instancia de SparkContext, [15]. A través de la misma, se puede recoger información acerca de los trabajos, las etapas de los mismos, el almacenamiento, el entorno, los ejecutores, ... básicamente, cualquier tipo de información que sea recogida por los listeners de Spark.

A continuación, se incluyen dos imágenes, Figura 4.6 y Figura 4.7. La primera de ellas muestra una línea temporal de los trabajos ejecutados en la aplicación “Consumos Electricos I”. Se puede observar que cuatro trabajos han sido completados y que el algoritmo de planificación empleado es FIFO (First In First Out). En la segunda image, Figura 4.7, se recauda información relativa a los ejecutores. Para la aplicación “Consumos Electricos I” se han creado tres ejecutores, dos de los cuales ya han muerto; lo que se corresponde con el diagrama de tiempo de la Figura 4.6, en la que se puede ver la duración de la vida de dos de los tres ejecutores creados.



Figura 4.6.: Planificación de los trabajos de Spark

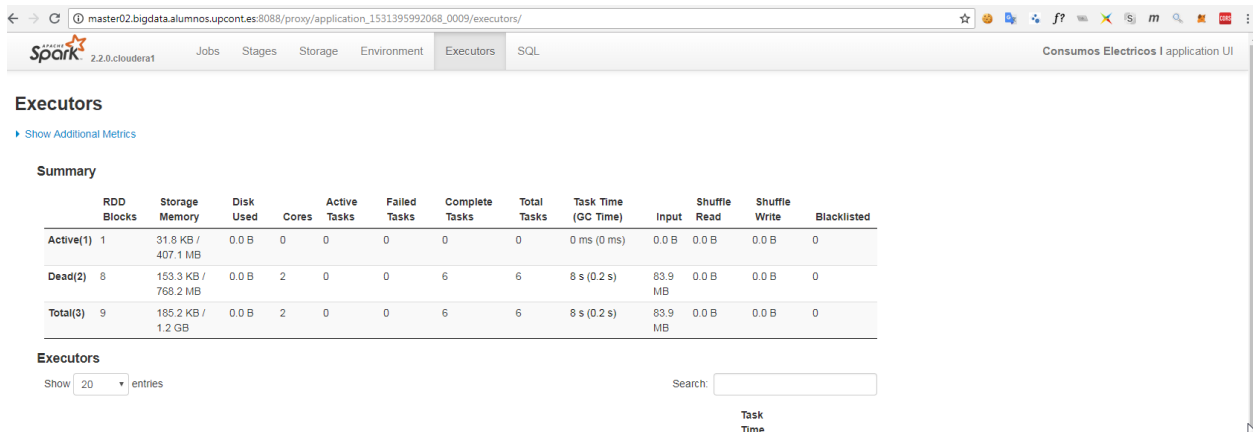


Figura 4.7.: Información ejecutores de Spark

4.4.2. Implementación

Una vez estudiada la arquitectura utilizada por el clúster sobre el que se ha desarrollado este proyecto, es necesario adentrarse en cómo y qué herramientas se han empleado en la creación de los distintos programas desarrollados con la finalidad de encontrar patrones del consumo eléctrico en España.

4.4.2.1. PySpark

Durante el desarrollo del presente proyecto se ha llevado a cabo la creación de distintos notebooks de Jupyter, en los que se ha creado el código empleado para obtener los resultados deseados. Dicho código ha sido escrito en Python 2.7, utilizando numerosas librerías: numéricas, gráficas, de machine learning, ... Sin embargo, de entre todas, las más importante es la librería PySpark, ya que es a través de la cual las líneas de código escritas se relacionan con la arquitectura Spark.

Apache Spark está escrito en Scala, con el fin de poder ser empleado por el lenguaje de programación Python, se creó la librería PySpark, [18]. La utilización de esta librería permite trabajar con RDDs desde Python. A partir de este momento, se pueden explotar todas las ventajas que ofrece Apache Spark, sin la necesidad de aprender un nuevo lenguaje de programación.

Jupyter Notebook es una aplicación popular que permite crear y ejecutar código en un entorno web. La gracia de esta herramienta recae en el uso de celdas, que permiten la ejecución del código en unidades definidas por el usuario; de esta forma, se incrementa el nivel de interacción con resultados intermedios, además de permitir una gran flexibilidad a la hora de ejecutar un mismo código. Al juntar la interactividad aportada por Jupyter con la escalabilidad de Spark, se obtiene una combinación muy poderosa que está siendo empleada en la actualidad por grandes empresas, como Google o IBM.

La imagen, Figura 4.8, muestra un notebook de Jupyter, en concreto el notebook “Consumos Electricos I”. A primera vista, se pueden observar como el código está dividido en secciones, cada una de ellas recibe el nombre de celda y comprenden la unidad mínima ejecutable. Como es necesario en cualquier programa de Python, en primer lugar, se definen las librerías empleadas durante el código. De entre las utilizadas, destaca la presencia de la librería PySpark, ya mencionada, que permite la integración de Apache Spark y Python. Además, cabe mencionar, que las dos celdas siguientes, están dedicadas a la creación de la configuración necesaria para ejecutar un programa basado en PySpark.

La plataforma web de Jupyter, también permite saber que aplicaciones están activas. Como, en el caso de este proyecto, los programas se estaban ejecutando sobre el clúster Big Data de la universidad, cada vez que se terminaba de trabajar sobre un programa era necesario hacer un “shutdown” del mismo, para que no siguiese consumiendo recursos que podían ser empleados por otras personas. La Figura 4.9, representa la ventana de Jupyter que muestra las aplicaciones que están activas y a través de la cual se puede matar las aplicaciones que ya no van a ser utilizadas.

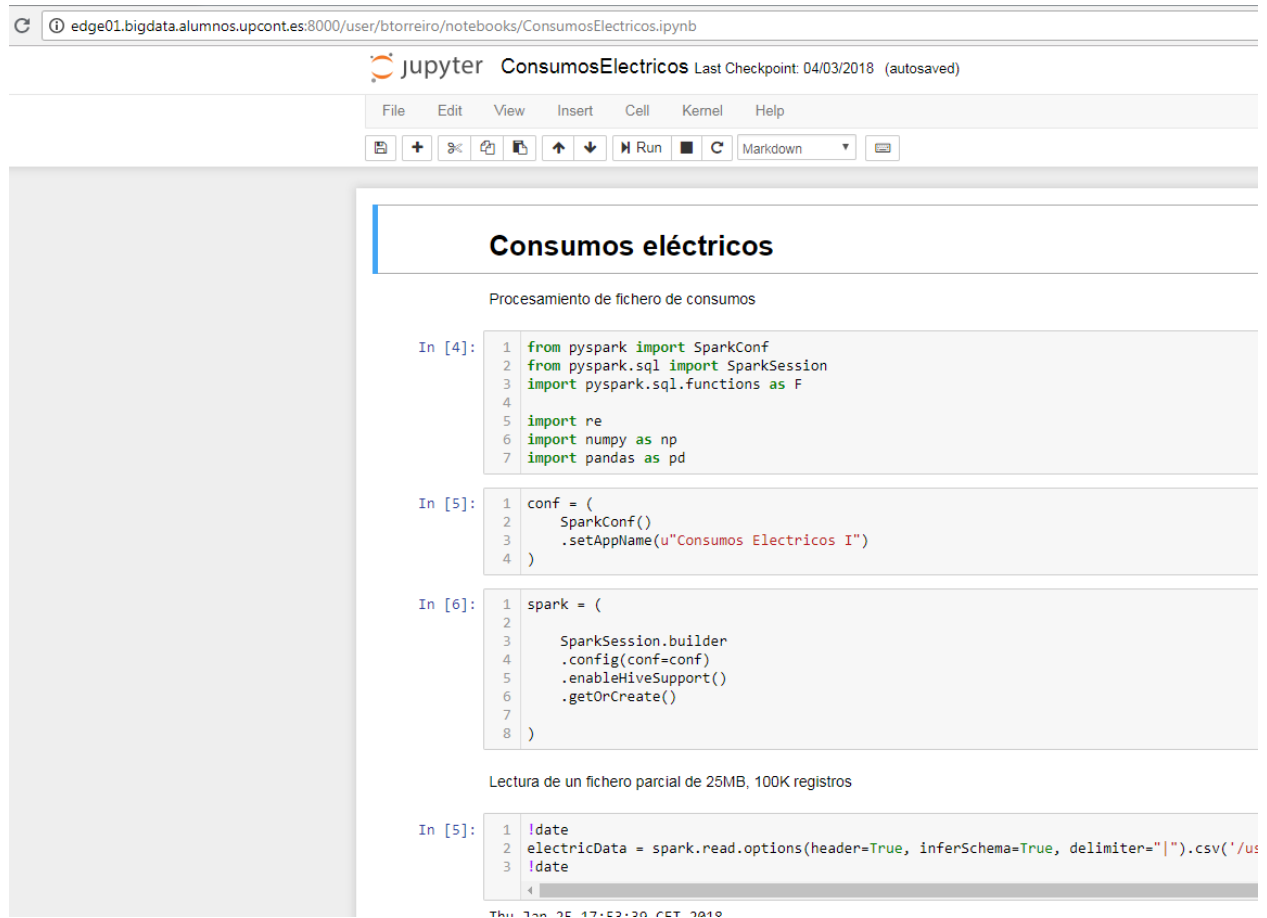


Figura 4.8.: Ejemplo de notebook de Jupyter

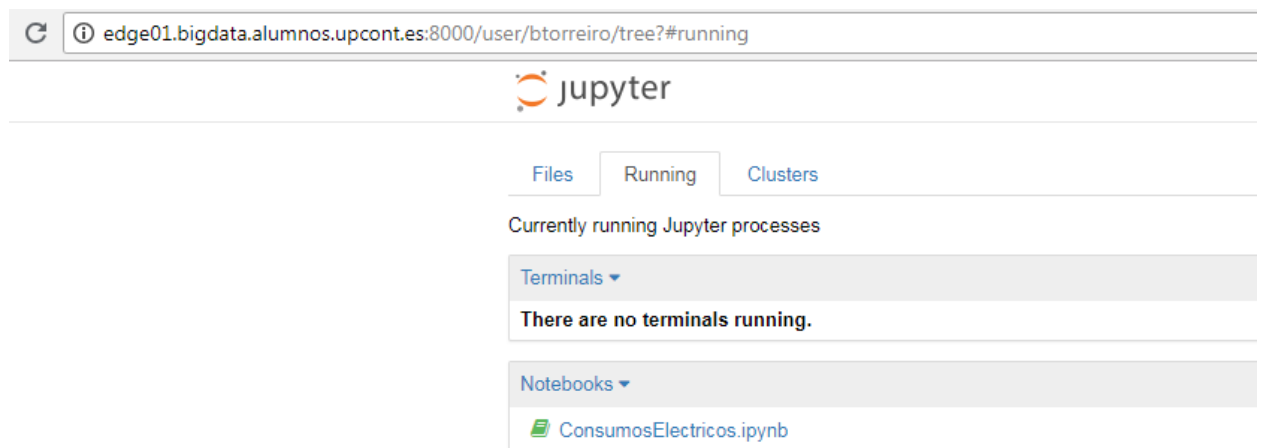


Figura 4.9.: Ejemplo de pestaña Running de Jupyter

Elementos críticos de PySpark

Los programas basados en Apache Spark, en este caso a través de la librería PySpark, pueden o deben hacer uso de las siguientes herramientas principales: DataFrames, transformaciones y acciones.

DataFrame

Un Dataset es una colección distribuida de datos, fue introducido en la versión 1.6 de Spark y proporciona todos los beneficios de las RDDs más los del motor de ejecución optimizado de Spark SQL, [19]. Los Datasets pueden ser contruidos a partir de objetos JVM (Java Virtual Machine) y ser manipulados utilizando transformaciones, como pueden ser map, filter, ...

Un DataFrame es un Dataset organizado en columnas. Es conceptualmente equivalente a una tabla de una base de datos relacional, pero optimizada. Un DataFrame puede ser construido a partir de una gran variedad de fuentes, archivos de datos, tablas Hive, bases de datos externas o RDDs ya existentes, [19].

Los DataFrames presentan las siguientes características, [20]:

- Inmutable: Una vez creado un DataFrame, no se pueden cambiar, se puede transformar después de aplicar las operaciones de transformación habilitadas.
- Evaluación perezosa: una tarea no es ejecutada hasta que ocurre una acción.
- Distribuido

Gracias a dichas características, los DataFrames son capaces de ofrecer las siguientes ventajas, [20]:

- Los DataFrames han sido diseñados para procesar grandes colecciones de datos estructurados o semiestructurados.
- La organización por columnas de los DataFrames, permite a Apache Spark, entender los esquemas de datos, lo que se traduce en una mejoría del rendimiento durante las ejecuciones.
- Los DataFrames tienen la capacidad de lidiar con petabytes de datos.
- Los DataFrames soportan una gran variedad de tipos de datos, como se ha mencionado anteriormente.
- Posee una API para diferentes lenguajes de programación como Python, R, Scala y Java.

Transformaciones

Las transformaciones son operaciones perezosas realizadas sobre un RDD que crean un nuevo RDD, por ejemplo: `map`, `filter`, `join`, `randomSplit`, ... Es decir, las transformaciones son funciones que toman un RDD como dato de entrada y generan uno o varios RDD nuevos, tras la aplicación de las operaciones pertinentes. Algunas transformaciones pueden ser canalizadas, con la finalidad de mejorar el rendimiento, [15].

El hecho de que las transformaciones son operaciones perezosas, quiere decir que aunque se vayan encadenando transformaciones sobre un mismo RDD, Spark no las ejecutará a menos que se ejecute una acción sobre la misma, [21]. Lo que permite a Spark guardar todas las transformaciones de un mismo RDD sin necesidad de guardar en memoria todos los RDDs que se van creando. Lamentablemente, si se ejecutan dos acciones sobre un RDD que es el resultado de una transformación, dicha transformación será calculada dos veces. Sin embargo, para evitar la redundancia de cálculo, Spark posee mecanismos de persistencia, [21].

Algunos ejemplos de transformaciones empleadas durante el desarrollo del proyecto son:

- *map(func)*: Devuelve un nuevo RDD tras pasar cada elemento del RDD original a través de la función introducida como parámetro.
- *filter(func)*: Realiza un filtrado de los elementos del RDD original para devolver un nuevo RDD con los datos filtrados.
- *distinct()*: Devuelve los elementos de los RDD que son distintos.

Acciones

Las acciones son operaciones RDD que devuelven valores no RDD. Materializan un valor en un programa de Spark, es decir, son operaciones RDD que devuelven cualquier tipo de valores, menos RDD, para ello desencadenan una transformación RDD, [15]. Las acciones permiten mandar datos desde los ejecutores al driver.

Algunos ejemplos de acciones empleadas durante el desarrollo del proyecto son:

- *reduce(func)*: Agrega los elementos usando la función introducida como parámetro, que debe ser conmutativa y asociativa.
- *collect()*: convierte un RDD en un array.
- *count()*: devuelve el número de elementos del RDD.

Rendimiento

A la hora de lanzar un programa, siempre se desea que el tiempo de ejecución del mismo sea el mínimo posible. Para alcanzar dicho objetivo es necesario familiarizarse con el lenguaje de programación empleado y que técnicas o filosofías se deben y no se deben emplear, con la finalidad de alcanzar el mejor rendimiento posible. A lo largo de este apartado se presentan ciertas ideas, [22], que se deben tener en mente a la hora de crear un programa basado en PySpark si se desea obtener un buen rendimiento.

Uno de los factores que más influyen en el tiempo de ejecución de cualquier programa, es el tamaño de los datos que se están manejando. Cuánto mayor sea el número de datos del que se disponga, mayor será el tiempo de ejecución. Está claro que en la mayor parte de las ocasiones, se tiene muy poco control sobre el tamaño del archivo de datos sobre el que se tiene que trabajar. Sin embargo, en función de los resultados que se quieran conseguir se puede filtrar el conjunto de datos inicial para conseguir un conjunto de datos con un tamaño más asequible que reduzca el tiempo de procesamiento. Es decir, si la operación se puede realizar sobre un subconjunto de los datos de los que se dispone, en primer lugar se debe crear un nuevo RDD que contenga esos datos, para luego realizar las operaciones pertinentes.

La acción *count()* tiene como finalidad devolver el número de elementos del RDD sobre el que se aplique. Debido a su naturaleza, no soporta el paralelismo, lo que ralentiza la ejecución del programa. Por ese motivo se debe evitar el uso de la operación *count()* a menos que sea necesario. También por motivos de paralelismo, se debe evitar el uso de bucles, ya que realizan operaciones sobre el conjunto de datos para cada uno de los elementos por orden, lo que anula totalmente las ventajas ofrecidas por el paralelismo e incrementa el tiempo de ejecución notablemente.

Uno de los grandes problemas cuando se trata con bases de datos distribuidas es la serialización. Formatos que tardan en serializar objetos o consumen una gran cantidad de recursos al hacerlo, ralentizan enormemente el cálculo. A menudo, esto es uno de los principales factores que se deben tener en cuenta para mejorar el rendimiento de las aplicaciones Spark.

Pandas es un paquete de Python que permite trabajar con conjuntos de datos estructurados o no. Este paquete, también implementa su propia versión de DataFrame. Mientras que los DataFrames de Spark son paralelizables, los DataFrames de Pandas no, por lo que es necesario tener cuidado a la hora de utilizarlos ya que empeoran el rendimiento de las aplicaciones considerablemente.

En ocasiones, la serialización del DataFrame de Pandas es una ventaja frente al DataFrame de SQL, como por ejemplo a la hora de pasar una lista de valores para la creación de un gráfico. Si el tamaño de datos manejado fuese del orden de los 100 MB no habría ningún problema de velocidad relacionado con el uso de Pandas, sin embargo, con el tamaño del conjunto de datos manejado en este proyecto, el uso de Pandas supone una bajada drástica en la velocidad de procesamiento. Por este motivo, es muy importante saber identificar en que situaciones se debe emplear y no abusar, para conseguir el mejor rendimiento posible.

4.4.2.2. Descripción del código

A lo largo del desarrollo del presente proyecto se han creado numerosos notebook de Jupyter con el fin de recaudar información acerca de los datos objeto de estudio. Con el objetivo de ilustrar como se han implementado los notebooks creados se ha tomado como ejemplo el notebook “K-Means, Madrid Residencial Invierno Diario”, ya que se trata de uno de los más completos.

Los pasos básicos seguidos en todos los programas son los siguientes:

1. Lectura
2. Limpieza
3. Transformaciones
4. Procesamiento
5. Visualización

En primer lugar, tras la creación de la configuración y de la instancia SparkContext, es necesario leer los datos sobre los que se va a trabajar. Para ello se emplea el comando:

```
electricData = spark.read
.options(header=True, inferSchema=True, delimiter="|")
.csv('/user/david/data/DATATHON_secuencial.csv').cache()
```

Además, como se puede observar en la Figura 4.10, se ha añadido una notación, que permite ver, de manera exacta, el tiempo empleado en la lectura de dicho fichero, que en este caso es poco más de un minuto. Este intervalo de tiempo está relacionado con el tamaño del conjunto de datos, que en este caso es de 8GB.

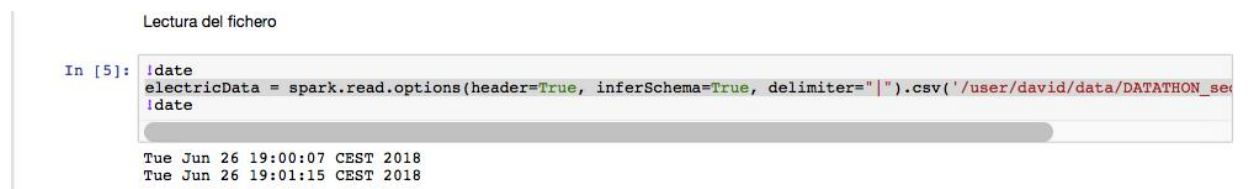


Figura 4.10.: Lectura de los datos objeto de estudio

A continuación, es necesario seleccionar solamente los datos que se van a emplear durante la etapa de procesamiento. Para ello, la herramienta básica utilizada es la transformación *filter(func)*, que permite crear un nuevo RDD a partir del original, seleccionando solamente los datos que cumplan las condiciones establecidas por la función que se pasada como parámetro.

Con la finalidad de hacer las notebooks creadas reutilizables para distintos casos de uso, se ha creado una celda dedicada a la declaración de las variables que determinarán los valores de los filtros empleados. Como se puede observar en la Figura 4.11, en este caso, las variables toman

os valores necesarios para solamente recoger los conjuntos de datos que pertenezca al consumo residencial, a la provincia de Madrid, a los días laborales y a la estación de invierno.

Variables

```
In [6]: cnae = 'T1' # Residencial
cnae_s = 'residencial'
provincia = 3.0 # Madrid
provincia_s = u'Madrid'
dia = 1.0 # Laboral
dia_s = u'día laboral'
estacion = 1.0 # Invierno
estacion_s = u'invierno'
```

Figura 4.11.: Variables empleadas durante la limpieza de datos

El RDD construido a partir de los datos objeto de estudio, no presenta columnas que permitan filtrar en función del día de la semana, la provincia o la estación del año. Ante esta situación se ha creado un nuevo RDD que contiene tres nuevas columnas que permiten el filtrado por los parámetros mencionados. La creación de este nuevo RDD ha sido posible gracias a la utilización de la transformación *withColumn(newColName, colExpr)* que permite añadir una columna al RDD sobre el que se aplica la transformación. La primera vez que esta transformación es empleada es con la finalidad de crear una nueva columna que recoja las suma de todo el consumo activo, antes de comenzar las filtraciones. Las figuras 4.12, 4.14, 4.13 y 4.15, muestran el proceso de filtrado de los datos objeto de estudio en función de los parámetros mencionados anteriormente.

Creación de las nuevas columnas

```
In [7]: electricData = electricData.withColumn("ACTIVA_TOTAL", sum(electricData[x] for x in electricData.columns if x.sta:

Sólo consumo residencial

In [8]: electricData = electricData.filter(F.col('CNAE') == cnae)
```

Figura 4.12.: Filtrado en función del tipo de actividad económica

Columna de tipo de día

```
In [12]: import datetime
def dayOfWeekElectric(fecha):
    fecha = str(fecha)
    diaSemana = datetime.datetime(int(fecha[:4]), int(fecha[4:6]), int(fecha[6:8])).weekday()+1
    if(diaSemana<6):
        return '1'
    elif(diaSemana<7):
        return '2'
    else:
        return '3'

In [13]: dayOfWeekElectricUDF = F.udf(lambda z: dayOfWeekElectric(z))
electricData = electricData.withColumn("TIPO_DIA", dayOfWeekElectricUDF(F.col('DIA')).cast('float'))

In [14]: electricData = electricData.filter(F.col('TIPO_DIA') == dia)
```

Figura 4.13.: Filtrado en función del tipo de día de la semana

Columna de provincias

```
In [9]: def buscarProvincia(municipio):
coruna = {"SADA", "A CORUÑA", "OLEIROS", "ARTEIXO", "FISTERRA", "CARBALLO", "MONFERO", "CULLEREDO", "CAMBRE", "MIÑO"}
barcelona = {"PIERA", "SANTA COLOMA DE GRAMENET", "SANTA MARGARIDA DE MONBUI", "SABADELL", "SITGES", "BADIA"}
madrid = {"MEJORADA DEL CAMPO", "ALCALA DE HENARES", "MADRID", "COSLADA", "POZUELO DE ALARCON", "ARGANDA DEL REY"}
malaga = {"MARBELLA", "FUENGIROLA", "VELEZ-MALAGA", "TORROX", "ESTEPONA", "VALLE DE ABDALAJIS", "MALAGA", "BENALMADE"}
if municipio in coruna:
    return 1.0
elif municipio in barcelona:
    return 2.0
elif municipio in madrid:
    return 3.0
elif municipio in malaga:
    return 4.0
else:
    return 0

In [10]: buscarProvinciaUDF = F.udf(lambda z: buscarProvincia(z))
electricData = electricData.withColumn("PROVINCIA", buscarProvinciaUDF(F.col("DE_MUNICIP")).cast('float'))

In [11]: electricData = electricData.filter(F.col("PROVINCIA") == provincia)
```

Figura 4.14.: Filtrado en función de la provincia

Columna de estación del año

```
In [15]: import datetime
def buscarEstacion(fecha):
    fecha = str(fecha)
    md = int(fecha[4:8])
    if ((md >= 320) and (md <= 619)):
        s = 2.0 # primavera
    elif ((md >= 620) and (md <= 921)):
        s = 3.0 # verano
    elif ((md >= 922) and (md <= 1220)):
        s = 4.0 # otoño
    elif ((md >= 1221) or (md <= 319)):
        s = 1.0 # invierno
    else:
        raise IndexError(u"Fecha inválida")
    return s

In [16]: buscarEstacionUDF = F.udf(lambda z: buscarEstacion(z))
electricData = electricData.withColumn("ESTACION", buscarEstacionUDF(F.col("DIA")).cast('float'))

In [17]: electricData = electricData.filter(F.col("ESTACION") == estacion)
```

Figura 4.15.: Filtrado en función de la estación del año

Además, como el análisis se ha centrado en el consumo activo, no en el reactivo, es necesario eliminar de los datos de trabajo aquellos que presenten niveles de consumo activo negativo, ya que esto se debe a errores en la recogida de los mismo. Para ello se aplica el filtro mostrado en la Figura 4.16, para todas las horas del día.

```
In [18]: electricData = electricData.filter((F.col("ACTIVA_H1") > 0) & (F.col("ACTIVA_H2") > 0) & (F.col("ACTIVA_H3") > 0))
```

Figura 4.16.: Filtrado de aquellos datos que contiene niveles de consumo activo negativos

A continuación, comienza la parte del código de dedicada al procesamiento de los datos transformados y seleccionados. En este caso, se han llevado a cabo las operaciones necesarias para implementar el algoritmo K-Means, lo que implica el uso de transformaciones y acciones, el código empleado se encuentra en el Anexo E. En otros casos, solamente se han realizado acciones con

la finalidad de obtener datos estadísticos acerca de los datos.

Por último, es necesario representar de forma visual los resultados obtenidos. Para ello se ha empleado la librería matplotlib de Python, que ofrece una sencilla API a través de la cual se puede construir todo tipo de gráficos, desde simple gráficos de barras hasta gráficos polares o logarítmicos. Durante el desarrollo del presente proyecto, todos los gráficos se han obtenido utilizando esta librería. Los tipos de gráficos creados son: diagramas de sectores, diagramas de cajas, gráficos de líneas (simples y múltiples) y gráficos de barras (separadas y superpuestas). La siguiente imagen, Figura 4.17, muestra el código empleado en la generación del gráfico que muestra las tendencias de los patrones encontrados tras la ejecución del algoritmo K-Means, a lo largo del día.

```
In [41]: import matplotlib.pyplot as plt

plt.figure(1,figsize = (10,10))

plt.plot(centers[0], '-', label=u'Grupo 1', color='#1f77b4')
plt.plot(centers[1], '-', label=u'Grupo 2', color='#ff7f0e')
plt.plot(centers[2], '-', label=u'Grupo 3', color='#2ca02c')
plt.plot(centers[3], '-', label=u'Grupo 4', color='#d62728')
plt.plot(centers[4], '-', label=u'Grupo 5', color='#9467bd')
plt.plot(centers[5], '-', label=u'Grupo 6', color='#8c564b')
plt.plot(centers[6], '-', label=u'Grupo 7', color='#e377c2')
plt.plot(centers[7], '-', label=u'Grupo 8', color='#bcbd22')

plt.title(u'Predicción del consumo ' + cnae_s + ' de un ' + dia_s + ' de ' + estacion_s + ' en ' + provincia_s)
plt.ylabel('Consumo, kWh')
plt.xlabel('Horas')
plt.xticks(np.arange(0, 24,1))
plt.grid()
plt.legend(loc='center left', bbox_to_anchor=(1, 0.5))
plt.show()
```

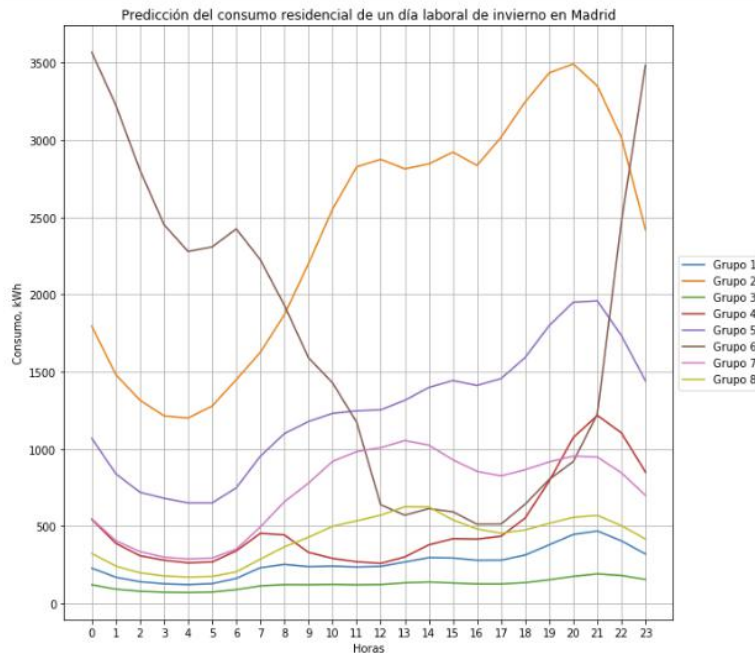


Figura 4.17.: Generación del gráfico que muestra las tendencias de los patrones encontrados

4.4.2.3. Rendimiento de procesamiento

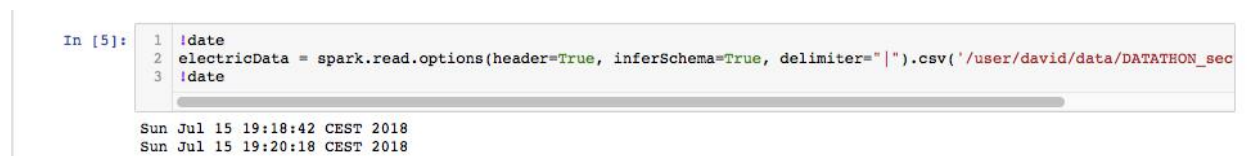
A la hora de desarrollar los numerosos notebooks se ha intentado aprovechar al máximo los recursos ofrecidos para el clúster, para ello se ha diseñado el código teniendo en mente el paralelismo, ya que mejora notablemente el rendimiento de las aplicaciones PySpark, aunque no siempre ha sido posible. A lo largo de este apartado se comentarán que celdas, de los notebooks creados forman cuellos de botellas, los tiempos obtenidos se han obtenidos cuando la aplicación “K-Meas, Madrid Residencial Invierno Diario” era la única que se estaba ejecutando en el clúster. Se ha elegido esta aplicación ya que recoge la mayor parte de las transformaciones y acciones aplicadas sobre los datos a lo largo del proyecto.

En la sección 4.4.2.1, se recoge una serie de operaciones y hábitos que se deben evitar para mejorar la eficiencia de las aplicaciones. Las ideas esenciales son las siguientes:

- Reducir el conjunto de datos sobre el que se realizan las operaciones en la medida de lo posible
- Evitar la acción `count()`
- Evitar el uso de bucles
- Uso moderado de la librería Pandas

Con estas ideas en mente, se procede al análisis del rendimiento de las distintas celdas que forman la aplicación “K-Means, Madrid Residencial Invierno Diario”.

La primera acción que demora considerablemente el tiempo de ejecución es la lectura del fichero, en este caso tarda algo más de minuto y medio. Lo cual no es de extrañar teniendo en cuenta que este es el tamaño del fichero es de 8GB.



```
In [5]: 1 |date
        2 electricData = spark.read.options(header=True, inferSchema=True, delimiter="|").csv('/user/david/data/DATATHON_sec
        3 |date
```

Sun Jul 15 19:18:42 CEST 2018
Sun Jul 15 19:20:18 CEST 2018

Figura 4.18.: Tiempo que tarda la aplicación en leer el fichero de datos

Durante todo el proceso de limpieza de datos, el tiempo de ejecución empleado es mínimo. Esto ha sido posible gracias al empleo de la transformación `withColumn(newColumnName, colExpr)` en la creación de nuevas columnas que permitan el uso de la transformación `filter(func)`. Es más, desde la lectura hasta el filtrado total de los datos, tan solo transcurren 3 segundos.

```

In [12]: 1 |date
          2 |import datetime
          3 |def dayOfWeekElectric(fecha):
          4 |    fecha = str(fecha)
          5 |    diaSemana = datetime.datetime(int(fecha[:4]), int(fecha[4:6]), int(fecha[6:8])).weekday()+1
          6 |    if(diaSemana<6):
          7 |        return '1'
          8 |    elif(diaSemana<7):
          9 |        return '2'
         10 |    else:
         11 |        return '3'
         12 |date

Sun Jul 15 19:20:19 CEST 2018
Sun Jul 15 19:20:19 CEST 2018

In [13]: 1 |date
          2 |dayOfWeekElectricUDF = F.udf(lambda z: dayOfWeekElectric(z))
          3 |electricData = electricData.withColumn("TIPO_DIA", dayOfWeekElectricUDF(F.col('DIA')).cast('float'))
          4 |date

Sun Jul 15 19:20:20 CEST 2018
Sun Jul 15 19:20:20 CEST 2018

In [14]: 1 |date
          2 |electricData = electricData.filter(F.col('TIPO_DIA') == dia)
          3 |date

Sun Jul 15 19:20:20 CEST 2018
Sun Jul 15 19:20:20 CEST 2018

```

Figura 4.19.: Tiempo que tarda la aplicación en filtrar por tipo de día de la semana

Todas las operaciones necesarias antes de poder construir la función de costes del algoritmo K-Means, también se realizan casi inmediatamente. Solamente se consume mucho tiempo al mostrar por pantalla parte de uno de los DataFrames creados. La función *show(n)* es el empleado para mostrar los primeros *n* elementos del DataFrame sobre el que se ejecuta, el único motivo por el que se ejecuta ese comando es para verificar la correcta creación de la columna ID que será necesaria en la implementación del algoritmo K-Means. Esta operación requiere más tiempo, 9 segundos, debido a que no es paralelizable y que la operación de escritura por pantalla suele ser lenta.

```

In [22]: 1 |date
          2 |trainingData.select('ID', 'ACTIVA_H1', 'ACTIVA_H2').show(5)
          3 |date

Sun Jul 15 19:22:21 CEST 2018
+-----+-----+
| ID | ACTIVA_H1 | ACTIVA_H2 |
+-----+-----+
| 1 | 101 | 89 |
| 2 | 149 | 126 |
| 3 | 153 | 57 |
| 4 | 318 | 270 |
| 5 | 323 | 194 |
+-----+-----+
only showing top 5 rows

Sun Jul 15 19:22:30 CEST 2018

```

Figura 4.20.: Tiempo que tarda la aplicación en realizar la operación show

La celda que más tiempo tarda en ejecutarse es la empleada para la obtención de la función de costes del algoritmo K-Means. Este se debe, principalmente, al uso de un bucle for, que imposibilita la paralelización, aumentando considerablemente el tiempo de ejecución. Además, las operaciones llevas a cabo para los distintos valores de *k*, realizan el proceso iterativo propio del algoritmo K-Means. Todo ello se acumula hasta alcanzar un tiempo de ejecución de 26 minutos y 13 segundos, que es algo menos del tiempo total de ejecución, que es igual a 1 hora 3 minutos y 48 segundos.

```

In [27]: 1 |date
          2 cost = np.zeros(30)
          3 for k in range(2,30):
          4     kmeans = KMeans().setK(k).setSeed(1).setFeaturesCol("features")
          5     model = kmeans.fit(df_kmeans.sample(False,0.1, seed=42))
          6     cost[k] = model.computeCost(df_kmeans)
          7 |date

Sun Jul 15 19:22:47 CEST 2018
Sun Jul 15 19:59:00 CEST 2018

```

Figura 4.21.: Tiempo que tarda la aplicación en crear la función de coste del algoritmo K-Means

La siguiente operación que no es relativamente rápida es el cálculo de los porcentajes de pertenencia, ya que implica la utilización de la acción `count()` que ya se ha establecido que es lenta. En un principio el cálculo de cada porcentaje suponía la realización de la acción `count()` dos veces, una para saber el número de entradas de la base de datos que pertenecían al grupo en cuestión y otra para determinar el total de los elementos. Ante la redundancia de este segundo `count()`, se decidió calcularlo tan solo una vez y guardar el valor en una variable.

```

In [41]: 1 |date
          2 total = df_pred.count()
          3 |date

Sun Jul 15 20:01:11 CEST 2018
Sun Jul 15 20:02:00 CEST 2018

In [42]: 1 |date
          2 total = float(total)
          3 |date

Sun Jul 15 20:02:00 CEST 2018
Sun Jul 15 20:02:00 CEST 2018

In [43]: 1 |date
          2 porcentaje_g1 = (df_pred.filter(F.col('prediction') == 0).count() / total) * 100
          3 porcentaje_g1
          4 |date

Sun Jul 15 20:02:00 CEST 2018
Sun Jul 15 20:02:48 CEST 2018

```

Figura 4.22.: Tiempo que tarda la aplicación en calcular los porcentajes de pertenencia a los grupos

Por último, otro conjunto de celdas que tardan más de un minuto en ejecutarse son las dedicadas a la obtención de los márgenes de desviación. Esto se debe al uso de un bucle `for` con la finalidad de sumar y restar a cada valor diario de un patrón su desviación típica.

```

In [52]: 1 |date
          2 grupol = df_pred.filter(F.col('prediction') == 0)
          3 grupol = grupol.select('features').toPandas()
          4 centroide = centers[0]
          5 horas_max_g1 = [centroide[0] + grupol.features.map(lambda x: x[0]).std()]
          6 horas_min_g1 = [centroide[0] - grupol.features.map(lambda x: x[0]).std()]
          7 for i in range(1,24):
          8     horas_max_g1.append(centroide[i] + grupol.features.map(lambda x: x[i]).std())
          9     aux = centroide[i] - grupol.features.map(lambda x: x[i]).std()
          10     if aux >= 0:
          11         horas_min_g1.append(aux)
          12     else:
          13         horas_min_g1.append(0)
          14 |date

Sun Jul 15 20:07:40 CEST 2018
Sun Jul 15 20:08:58 CEST 2018

```

Figura 4.23.: Tiempo que tarda la aplicación en calcular los márgenes de desviación

4.4.2.4. Problemas

Los principales problemas encontrados a lo largo del desarrollo de los distintos notebooks, están relacionados con la forma de pensar y el rendimiento del código.

Al comienzo del desarrollo las celdas de código escritas están diseñadas empleando la mentalidad de las bases de datos relacionales, abstracción que no está tan desencaminada gracias a los DataFrames, pero que no se puede mantener si se quiere mejorar el rendimiento de las mismas. Uno de los primeros problemas encontrados al iniciar el desarrollo de los notebooks dedicados a la búsqueda de patrones, residía en cómo realizar el filtrado de los datos por columnas que todavía no existían. La primera versión de la lógica diseñada para la creación de dichas columnas, funcionaba cuando se probaba con partes del conjunto total de datos, pero debido a la ineficiencia de la misma, no era trasladable al conjunto total ya que el servidor se quedaba sin memoria.

Los principales problemas con los que contaba la primera versión del código residían en el uso de bucles y en el abuso del concepto *union*. La idea básica detrás de la lógica inicial era recorrer cada uno de los elementos que formaban el conjunto de datos uno a uno y comprobar si pertenecían a una categoría determinada, por ejemplo, la provincia de Madrid. Si se cumplía dicha condición entonces el elemento era añadido al DataFrame que contenía todos los elementos que satisfacían la condición. Los dos grandes problemas de este código residían en el uso de un bucle *for* y de la operación *union*.

Spark no está diseñado para la utilización de bucles, si no para el uso del paralelismo. Es más, con cada iteración del bucle el tiempo de procesamiento del siguiente elemento aumenta. Además, la operación *union* no guarda todos los elementos en zonas consecutivas de memoria, lo que ralentiza la lectura del nuevo RDD creado. Por estos motivos, la primera versión de dicho código, no era capaz de procesar los 8GB de datos que constituye el fichero de datos empleado.

Ante este problema, un cambio drástico en el código era necesario, motivo por el cual se decidió emplear la transformación *withColumn(newColName,colExpr)*. Esta transformación hace uso del paralelismo, ya que la misma operación se puede estar aplicando a varios elementos a la vez, lo que reduce de manera drástica el tiempo de procesamiento, así como el espacio de memoria malgastado. Gracias a este cambio se consiguió crear un código capaz de crear un nuevo RDD a partir del original más una columna nueva en cuestión de segundos.

Sigue habiendo partes del código que emplean bucles y otras operaciones no paralelizables que ralentizan de manera significativa el tiempo de ejecución. Sin embargo, no se han encontrado maneras más eficientes de realizar la operación deseada, como es el caso del uso de la acción *count* y el bucle empleado en la construcción de la función de costes durante la implementación del algoritmo K-Means.

5. Análisis De Resultados

5.1. Análisis Descriptivo General

El objetivo del análisis descriptivo es conocer, de forma general, cómo distintos factores afectan a los datos objeto de estudio. En este caso, disponemos de datos sobre el consumo eléctrico en España durante aproximadamente dos años. Para ello, se han desarrollado diversos programas que permiten extraer información acerca del consumo dependiendo de los factores de estudio.

En primer lugar, para tener a modo de referencia, se ha calculado el consumo medio de energía activa total en un día en España en un año, que es igual a 8257.61 kWh. Este valor constituye la media diaria de nuestra población objeto de estudio, que será utilizada como referencia para observar la desviación de las medias de cada una de las muestras que se van a estudiar.

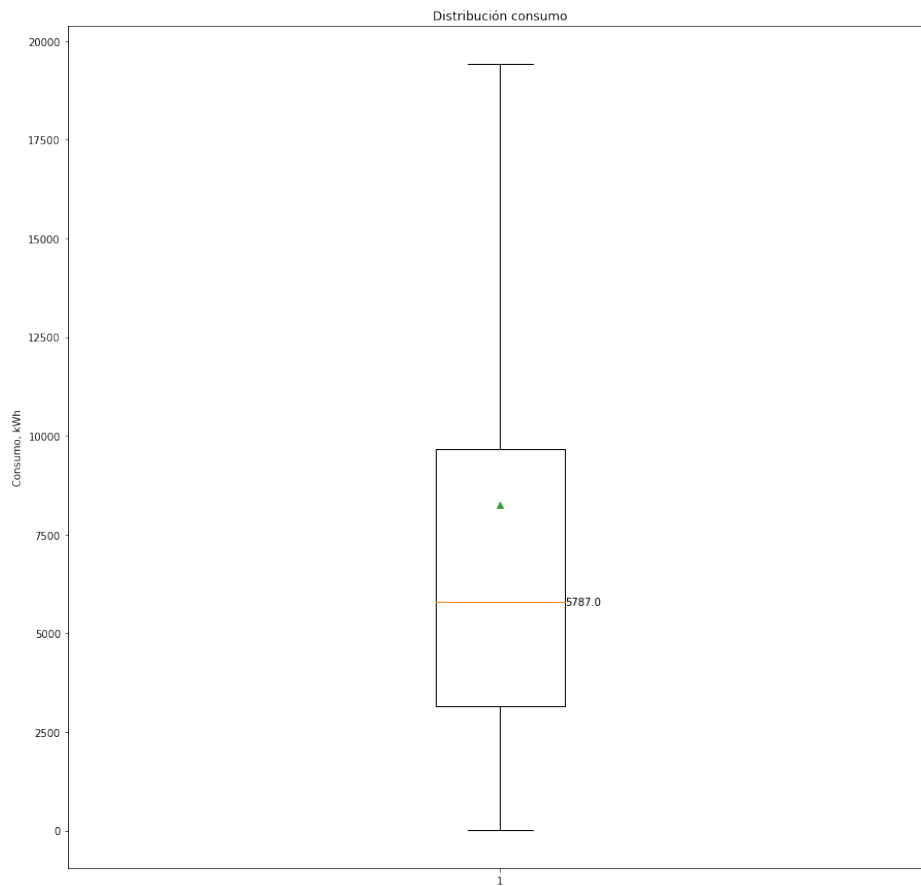


Figura 5.1.: Distribución del consumo total diario

Un primer acercamiento, antes de profundizar en como cada uno los distintos factores objeto de estudio afectan al consumo eléctrico en España, consiste en visualizar la distribución del consumo diario total, Figura 5.1. En primer lugar cabe destacar que el 50 % de los valores de la distribución se encuentran entre 3152.0 kWh y 9655.0 kWh, estando la mediana en 5787.0 kWh. Además, la diferencia existente entre la media (triángulo verde) y la mediana deja inferir una gran cantidad de outliers, éstos pueden ser debidos a la presencia de grandes industrias que generan un consumo mayor, en comparación con el consumo generado por viviendas, que presentan el mayor porcentaje de origen de consumo, como se muestra en la sección 5.1.1.

5.1.1. Consumo por tipo de actividad económica

El conjunto de datos del que se dispone, se puede dividir en dos grandes grupos en función del código de Clasificación Nacional de Actividades Económicas (CNAE). En nuestro caso, solamente hay dos actividades económicas presentes, actividad industrial y residencial. El siguiente diagrama de sectores, Figura 5.2, representa el porcentaje de la energía activa total que proviene de cada una de ellas. Como se puede observar, las actividades industriales tan solo comprenden el 24.5 % de la energía activa total consumida, frente al 75.5 % de la actividad residencial.

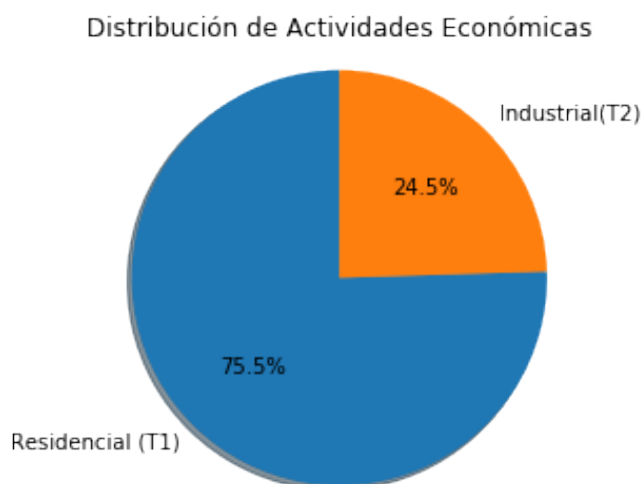


Figura 5.2.: Distribución general del CNAE

A pesar de que el consumo industrial representa un porcentaje menor del consumo de energía activa total, cada uno de los miembros de este grupo generan en promedio, un consumo mayor, como se muestra en la siguiente tabla, Tabla 5.1.

	RESIDENCIAL	INDUSTRIAL
Consumo Medio	7481.98	7399.44
Desviación Standard	12130.27	21570.61

Tabla 5.1.: Datos sobre la actividad industrial y residencial

5.1.2. Consumo por día de la semana

A fin de crear una imagen inicial acerca de las variaciones que se producen en el consumo eléctrico dependiendo del día de la semana, ya sean días laborables, sábados o domingos, se han calculado las medias diarias para cada una de las situaciones expuestas. Los valores obtenidos son los siguientes:

	CONSUMO MEDIO (kWh)	MEDIANA (kWh)
Días Laborables	8347.44	5834.00
Sábados	8064.24	5657.00
Domingos	8000.53	5672.00

Tabla 5.2.: Datos sobre el consumo en función del tipo de día

Se puede observar una pequeña caída en el consumo eléctrico durante el fin de semana, siendo más pronunciada los sábados. Esta tendencia se puede deber a la suspensión de las actividades industriales, ya que estas son las que generan un consumo medio mayor. Con el fin de obtener más información acerca de las variaciones en el consumo dependiendo del tipo de día se han creado los siguientes diagramas de cajas. Estos diagramas, muestran la distribución del consumo eléctrico para los días laborales, los sábados y los domingos, a excepción de los outliers.

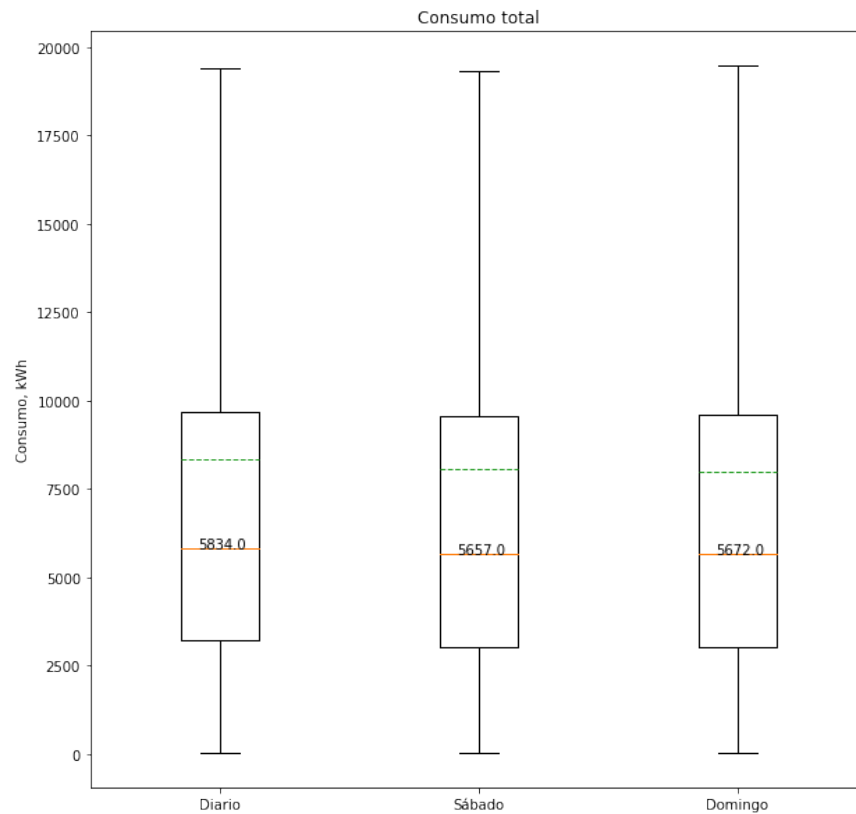


Figura 5.3.: Distribución del consumo por días de la semana

En el diagrama anterior, se ve representada la media, línea verde de puntos, y el valor de la mediana. La diferencia entre ambas, proporciona una idea acerca del sesgo que introducen los valores extraños, no mostrados en el diagrama de cajas.

Con el fin de indagar en la relación existente entre las variaciones del consumo en los distintos tipos de días con el tipo de consumo, se ha creado el siguiente gráfico. Este diagrama, muestra la variación del consumo medio residencial e industrial para cada tipo de día de la semana. Del mismo se pueden extraer las siguientes conclusiones. En primer lugar, cabe destacar la constancia del consumo residencial medio. Por el contrario, se puede observar la reducción existente en el consumo medio industrial, lo que respalda una reducción en el consumo de este sector.

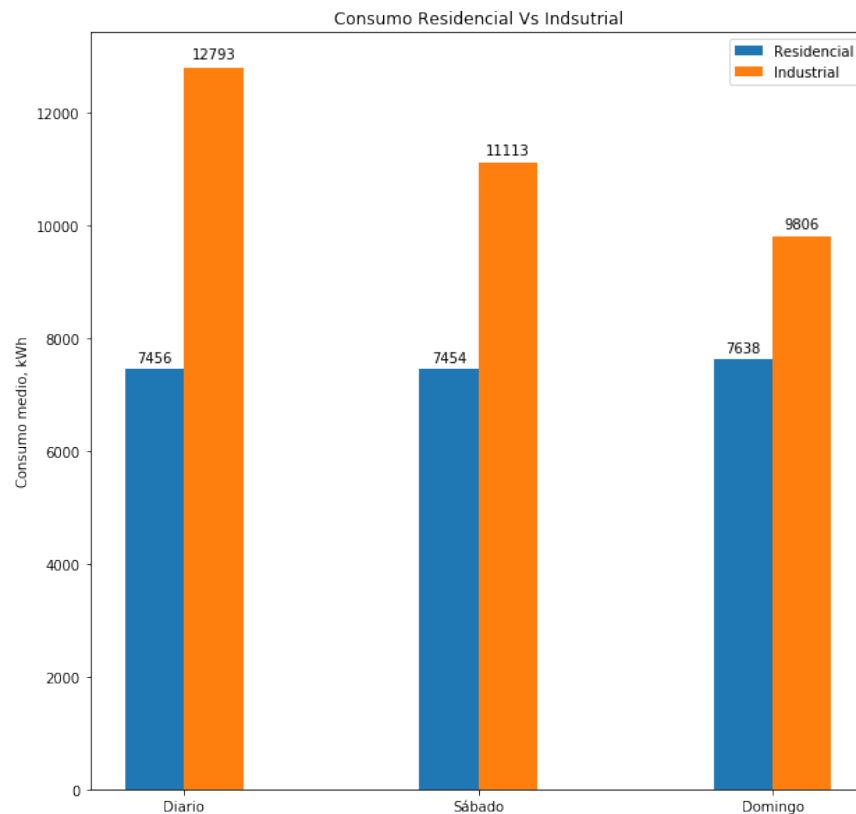


Figura 5.4.: Consumo medio Residencial Vs Industrial

5.1.3. Consumo por hora del día

Con respecto a la variación del consumo eléctrico en función de la hora del día, se puede observar, Figura 5.5, como las horas de menor consumo son las de madrugada mientras que a medida que avanza el día el consumo va en aumento. Concretamente, el mínimo se produce a las 5:00h, cuando tan solo se consumen 192.42 kWh; mientras que el máximo consumo se alcanza a las 22:00h, momento en el que es igual a 496.20 kWh.

En una vivienda, la tendencia encontrada no levanta sospecha alguna ya que encaja con el flujo de actividad. Durante las horas de sueño, el consumo se deberá principalmente al generado por la calefacción o el aire acondicionado. Una vez empieza el nuevo día los consumos van aumentando a medida que se hace uso de los distintos electrodomésticos y aparatos que necesitan conectarse a la red eléctrica. Además, al ponerse el sol el consumo también sufre una subida debida a la necesidad de encender la luz eléctrica al no haber luz natural, para después volver a sufrir una bajada al empezar las horas de sueño. Se puede considerar que la tendencia de subida se prolonga entre las 5:00h y las 22:00h. Teniendo esto en mente, cabe destacar la presencia de un máximo relativo en torno a las 15:00h, además del máximo absoluto a las 22:00h. Estos picos presentan relación con las horas de comidas, ya que la presencia en las viviendas es mayor y se hace uso de potentes electrodomésticos. En la sección 5.3, se amplía el estudio acerca de las tendencias en el consumo residencial.

Por el contrario, si nos centramos en actividades industriales, cuya tendencia no está suficientemente reflejada en el diagrama de barras de la Figura 5.5 debido al bajo porcentaje de datos industriales, cabe esperar un diagrama un tanto distinto. En primer lugar, las jornadas laborales no suelen alargarse hasta las 22:00h, por lo que cabe esperar una bajada en el consumo antes de esa hora. Además, muchas empresas realizan paradas al mediodía para que los trabajadores puedan comer, por lo que si la actividad cesa durante esas horas, también se esperará un consumo menor. Asimismo, se supone una mayor estabilidad en las horas laborables ya que la actividad es siempre la misma o similar, en cada una de las empresas. En la sección 5.3, se amplía el estudio acerca de las tendencias en el consumo generado por la industria.

El diagrama de cajas mostrado en la Figura 5.5, se puede observar como el 50% de los valores medios de consumo por horas se encuentran entre 253.14 kWh y 407.60 kWh, siendo la mediana igual a 380.40 kWh. Además, si se suman los consumos medios de cada una de las horas, se obtiene el siguiente valor 8257.57 kWh que se aproxima al consumo medio diario (8257.61 kWh), como era de esperar.

Cabe destacar que ambos diagramas hacen referencia al consumo medio total, por lo que la tendencia representada no refleja fielmente las tendencias de grupos más específicos de datos, como aquellos generados en actividades industriales o en residencias, en ciudades grandes o pequeñas, en distintas estaciones del año, ... Un análisis acerca de las tendencias en función de cada uno de los factores objeto de estudio, se lleva a cabo en secciones venideras.

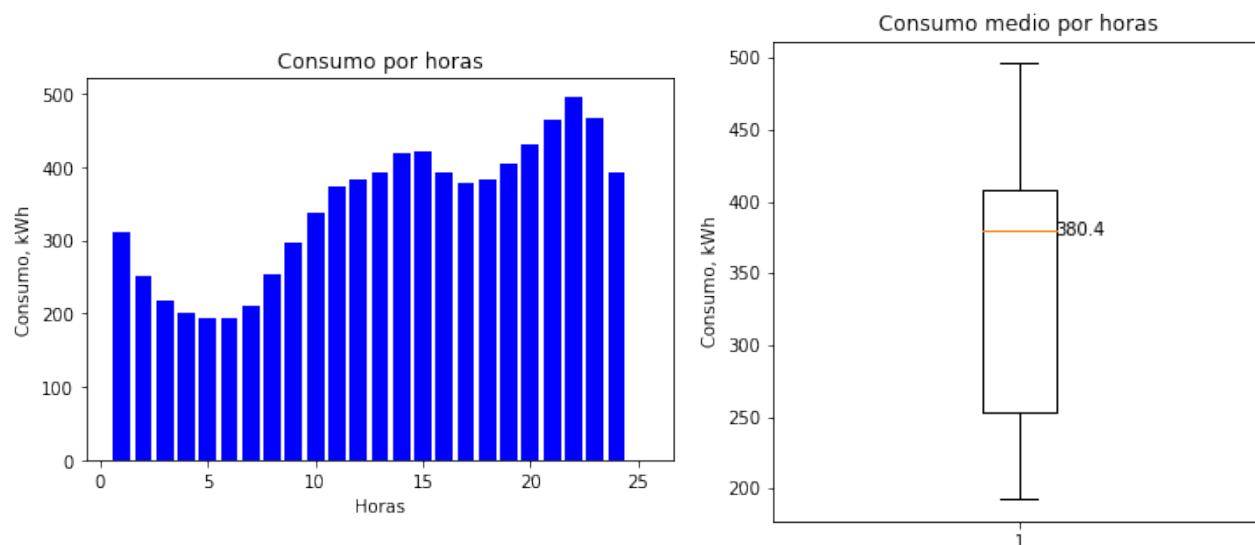


Figura 5.5.: Consumo general por horas

5.1.4. Consumo por municipios

Otro de los factores de estudio es el municipio del que proceden los datos. El objetivo de este factor es obtener conclusiones relacionadas con las diferencias en el consumo en distintas zonas geográficas y en áreas con diferente densidad de población. Con el fin de determinar que poblaciones generan un mayor y menor consumo se ha creado una lista que contiene el consumo medio de energía activa de cada uno de los municipios de los que se tienen datos, Tabla F.2.

Con el fin de obtener una imagen general a cerca de la distribución del consumo anual de los municipios españoles se ha creado el diagrama de cajas mostrado, Figura 5.6. Del mismo, que no muestra los valores extraños, cabe destacar la gran diferencia existente entre el valor del consumo anual total medio, representado por el triángulo verde, y la mediana del consumo, mostrada en la imagen. Esta desviación se debe a la presencia de outliers, cuyos valores elevan la media en gran medida.

Con el objetivo de entender qué municipios generan un consumo tan elevado en comparación con el resto y por qué motivo, se han extraído de la Tabla F.2, los cinco municipios con el consumo anual más elevado.

MUNICIPIO	CONSUMO ANUAL TOTAL(kWh)
Madrid	16980998236
Sant Boi de Llobregat	10044053219
Rubi	8431887781
Mairena del Alcor	8072565136
L'Hospitalet de Llobregat	7332389656

Tabla 5.3.: Municipios con el consumo más elevado

Las dos provincias del Reino de España que presentan la renta bruta media más alta son Madrid y Barcelona, [23]. Por lo que no es de extrañar que cuatro de los cinco municipios que presentan el mayor consumo pertenezcan a una de estas dos provincias, la excepción es Mairena del Alcor. De este hecho se puede inferir la existencia de una relación entre la riqueza de un municipio y el consumo eléctrico que genera, cuánto mayor sea la riqueza, mayor es el consumo. Además, grandes municipios tienden a tener una mayor presencia industrial, lo cual incrementa el consumo eléctrico, como fue determinado en la sección 5.1.1.

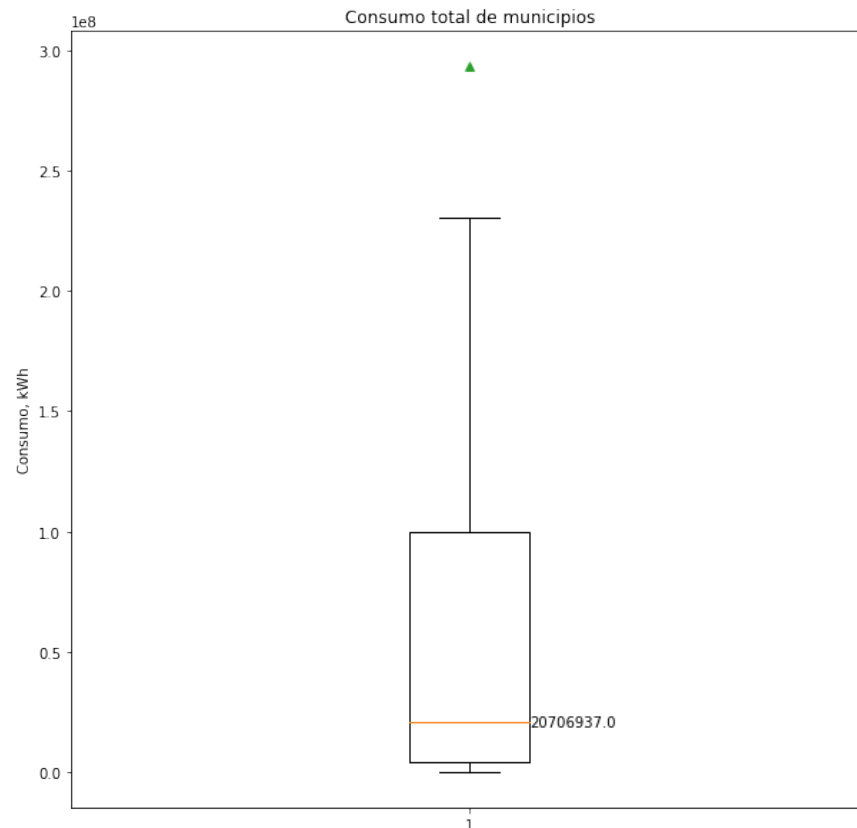


Figura 5.6.: Distribución del consumo total por municipios

Con el fin de extraer conclusiones más precisas para un determinado colectivo, se ha decidido desarrollar un estudio más profundo sobre las provincias de Madrid, Málaga y Barcelona, sección 5.2.

5.1.5. Consumo por tipo de tarifa de mercado

En el sector eléctrico hay dos tipos de mercados, el mercado regulado y el mercado libre. La principal diferencia entre ambos es el precio que se paga por el número de kWh consumidos. El mercado regulado es aquel en el que el precio por kW cambia cada hora de cada día dependiendo de la relación entre la oferta y la demanda en cada momento, [24]. De este modo, los consumidores pagarán más o menos en su factura dependiendo de las horas a la que consuman más energía. Por el contrario, el mercado libre es aquel en el que el precio es fijado por la distribuidora. En este caso el gasto generado depende únicamente del número de kW consumidos, independientemente de la hora.

Teniendo como finalidad obtener que porcentaje, del consumo eléctrico objeto de estudio, pertenece al mercado regulado y que porcentaje pertenece al mercado libre, se ha creado el siguiente diagrama de sectores, Figura 5.7. Como se puede observar, un 50.6% corresponde al mercado regulado, mientras que el mercado libre abarca un 49.4% de los datos.

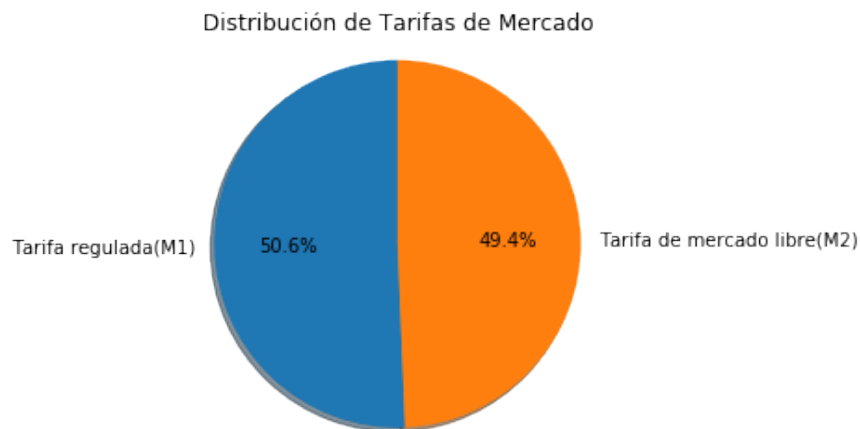


Figura 5.7.: Distribución de las tarifa de mercado

Para obtener más información acerca de la diferencia de coste entre los mercados, se ha calculado el consumo diario medio para cada una de las tarifas. Los valores obtenidos son los siguientes:

TARIFA	CONSUMO MEDIO(kWh)
Mercado Regulado	7511.34
Mercado Libre	9195.13

Tabla 5.4.: Consumo diario medio por tipo de tarifa

La tabla anterior muestra que la diferencia existente entre los consumos diarios medios de las dos tarifas, generando un mayor consumo medio la tarifa de mercado libre. Ante esta situación, surgen las siguientes preguntas. ¿Cuál es el motivo por cual la ambas tarifas abarcan prácticamente el mismo porcentaje del consumo total? ¿Existirá una relación con las horas en las que se consume? Para averiguar la respuesta a esta última pregunta se han creado los siguientes gráficos, que muestran la variación en el consumo para cada una de las dos tarifas.

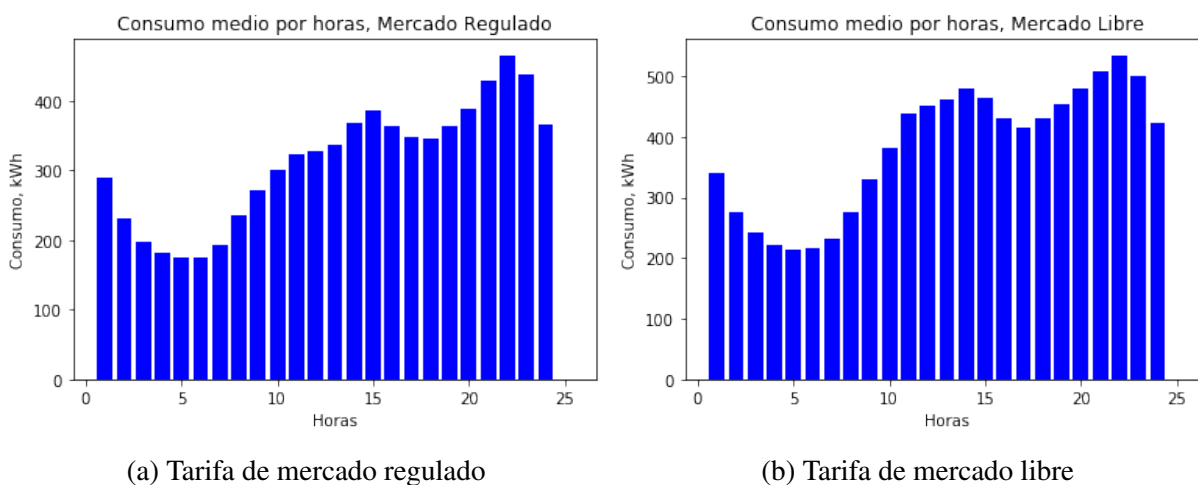


Figura 5.8.: Distribución del consumo medio por horas para cada tarifa

La distribución del consumo medio por horas presenta una forma similar para ambas tarifas de mercado. La única diferencia apreciable es el hecho de que los valores de la tarifa de mercado libre son más elevados, lo que es de esperar, teniendo en cuenta que el consumo medio de la tarifa de mercado libre es mayor que el de la tarifa de mercado regulado. Con estas tendencias, se puede deducir que la mayor parte de los clientes que tienen contratado la tarifa de mercado regulado consumen menos que aquellos que tienen contratada la tarifa de mercado libre.

Ante esta situación, se sospecha una relación entre el tipo de actividad económica y la tarifa de mercado contratada. Es decir, aquellos clientes que generan consumo del tipo residencial son más propensos a contratar la tarifa de mercado regulada. Por el contrario, aquellos clientes cuyo consumo proviene de actividades industriales, son más propensos a contratar la tarifa de mercado libre. Con el fin de comprobar la veracidad de esta hipótesis se ha construido un diagrama, 5.9, que presenta la relación entre el tipo de actividad económica y el tipo de tarifa contratada.

En el diagrama siguiente se puede observar como en ambas tarifas la actividad económica dominante es la residencial. Cabe destacar el hecho de que aunque en ambas predomine el consumo de origen residencial, en la tarifa de mercado libre las actividades industriales presentan un mayor peso sobre el consumo total. Por este motivo, y teniendo en cuenta que el consumo medio de las actividades industriales es mayor que el de las residenciales, Sección 5.1.1, no es de extrañar que el consumo medio de la tarifa libre sea también el mayor. Además, se puede observar como el mercado regulado recoge un mayor porcentaje del consumo total, como también se visualiza en la Figura 5.7.

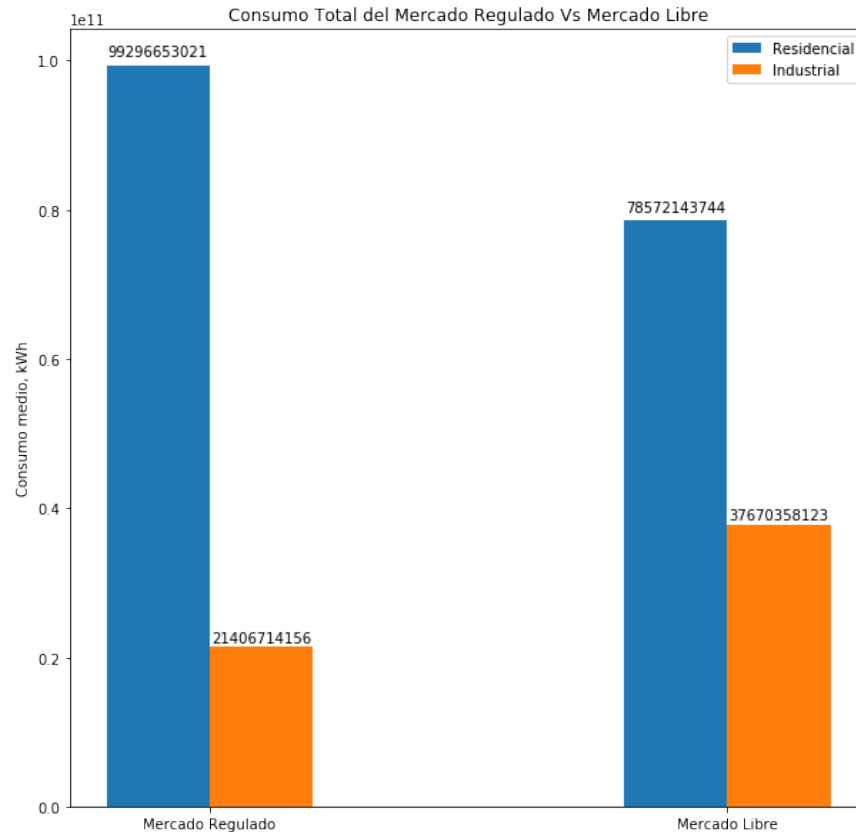


Figura 5.9.: Consumo total de la tarifa de mercado regulado vs libre

5.2. Análisis Descriptivo De Las Provincias De Madrid, Barcelona Y Málaga

El objetivo fundamental de este trabajo recae en encontrar patrones en el consumo eléctrico. Estos patrones se utilizarán para perfilar el consumo eléctrico en España a lo largo de un determinado día, en una determinada zona geográfica, para una estación del año especificada. Teniendo como finalidad alcanzar este objetivo, se ha decidido reducir los municipios objeto de estudio a aquellos pertenecientes a las provincias de Madrid, Barcelona y Málaga. En esta sección, se desarrolla una comparativa entre estas cuatro provincias, teniendo en cuenta los principales factores que afectan al consumo; el tipo de actividad económica, el día de la semana, la hora del día y el tipo de tarifa de mercado.

En primer lugar, para obtener una imagen general acerca de la situación de cada una de las provincias, se han calculado el consumo eléctrico medio diario y el consumo total de cada una de ellas. Los valores son los siguientes:

PROVINCIAS	CONSUMO MEDIO (kWh)	CONSUMO TOTAL (kWh)
Barcelona	7652.41	43077623598
Madrid	8011.50	19653413866
Málaga	7063.64	11352399127

Tabla 5.5.: Consumo medio diario por provincias

Gracias a los datos anteriores, se pueden ordenar las provincias en función del consumo medio diario que generan, así como en función del consumo total. Las ordenaciones quedan de la siguiente forma:

- En función del consumo medio diario: *Málaga < Barcelona < Madrid*
- En función del consumo total: *Málaga < Madrid < Barcelona*

Ambas ordenaciones tan solo se diferencian en el orden de Madrid y Barcelona. En primer lugar, cabe destacar el hecho de que la provincia de Barcelona presenta un consumo medio menor que el de la provincia de Madrid y, sin embargo, presenta el mayor consumo total con mucha diferencia. De esta observación se puede inferir que en esta provincia el consumo residencial abarca un porcentaje mayor en comparación con la provincia de Madrid, de ahí que presente un consumo medio menor a pesar de generar el mayor consumo total, la veracidad de esta afirmación se trata en la sección 5.2.1. Esta observación se puede extrapolar al resto de provincias, cuanto mayor sea el consumo medio mayor tiende a ser el porcentaje de consumo industrial frente al residencial, ya que el primer tipo de actividad económica genera un consumo medio mayor que el segundo. Por el contrario, las diferencias en el consumo total pueden estar asociadas con la población de los distintos municipios que conforman la provincia, el clima o el número de empresas industriales.

Con la finalidad de indagar en las distribuciones del consumo de cada una de las cuatro provincias objeto de estudio, se han construido los siguientes diagramas de cajas, Figura 5.10. En este gráfico, se han excluido los valores extraños.

Los diagramas siguientes reflejan la distribución del consumo de cada una de las provincias objeto de estudio. Se puede observar la gran diferencia existente entre el valor del consumo medio, representado por la línea verde de puntos, y la mediana del consumo, causada por la presencia de valores extraños. A pesar de la diferencia, se aprecia que, la provincia que alcanza los valores más bajos de consumo es Málaga. Los motivos por los que esta provincia presenta un consumo ligeramente menor se exponen a lo largo de esta sección. Además, cabe destacar la similitud entre las distribuciones de la provincia de Málaga y Barcelona.

Ante esta situación, surgen diversas preguntas. ¿Existe relación entre el consumo y el clima de cada provincia? ¿Será que en las provincias de Madrid y Barcelona existe tal diferencia entre el consumo de la capital y el consumo de los municipios de alrededor que se deberían de estudiar por separado? En esta sección se intentará resolver las preguntas planteadas.

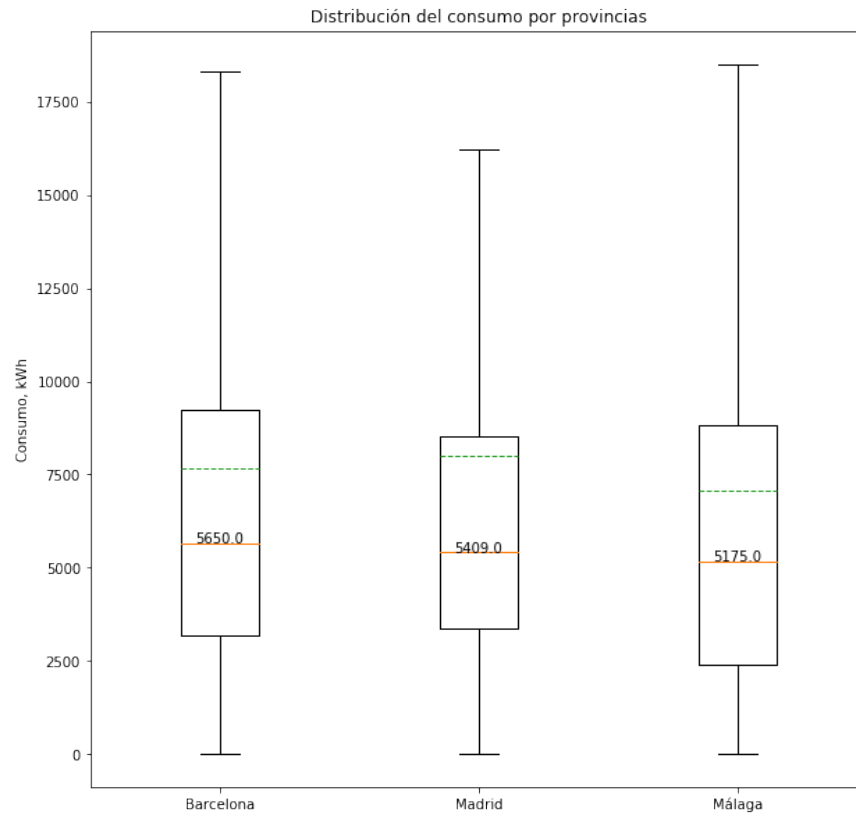


Figura 5.10.: Comparativa de la distribución del consumo total de las provincias de Barcelona, Madrid y Málaga

5.2.1. Comparativa basada en el tipo de actividad económica

A continuación, se realiza una comparación entre las cuatro provincias objeto de estudio en función del tipo de actividad económica, residencial o industrial. En la sección 5.1.1 se establece como el consumo del tipo residencial es más común y, por el contrario, las actividades industriales generan un consumo medio más elevado.

En primer lugar, para obtener una imagen general de la distribución de los distintos tipos de actividades económicas en las distintas provincias, se han generado los siguientes diagramas de sectores, que representan el porcentaje del consumo total abarcado por cada una de las actividades económicas.

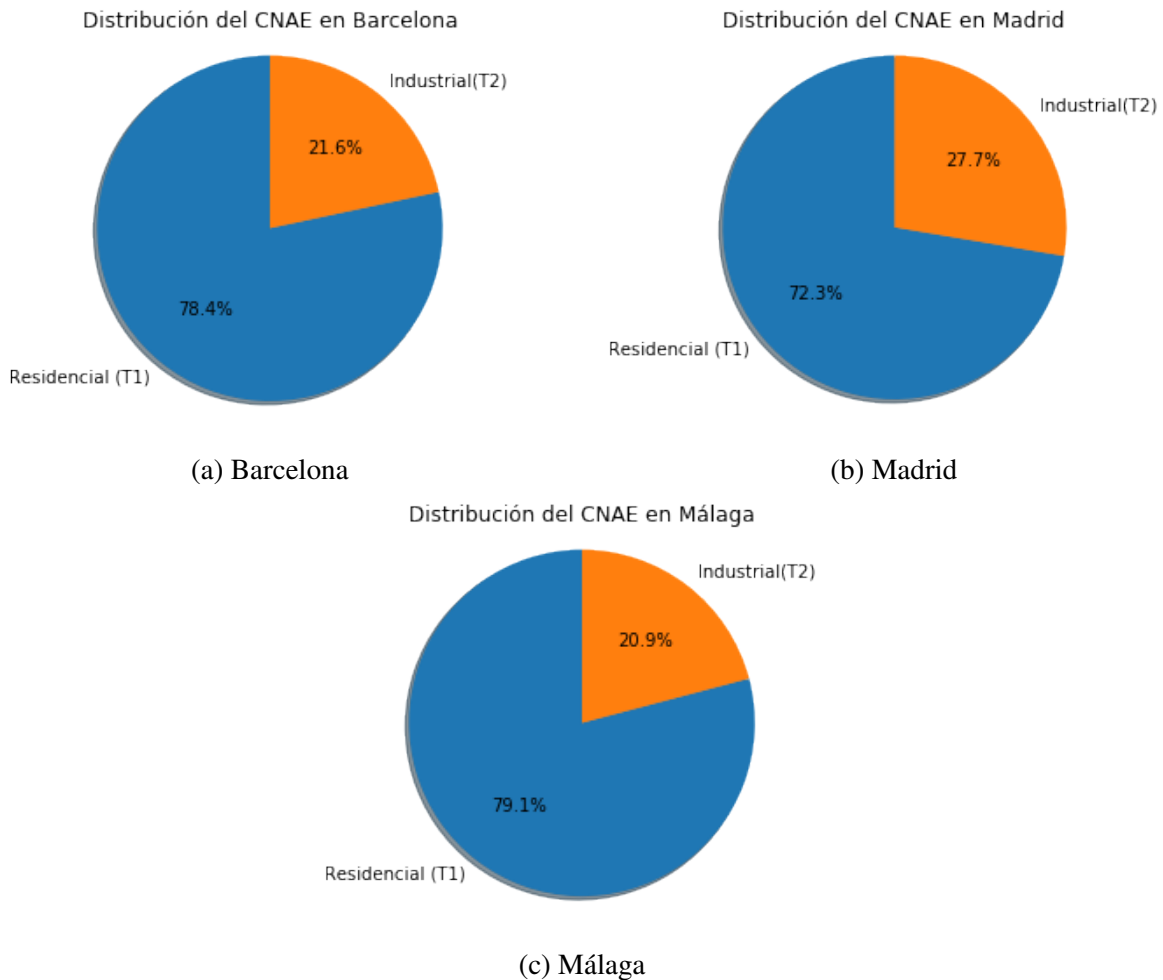


Figura 5.11.: Diagrama de sectores del CNAE de cada provincia

Los diagramas de sectores anteriores muestran que la provincia con mayor presencia de actividades industriales es Madrid. Debido a que las actividades industriales generan un mayor consumo, cabe esperar que aquellas provincias con un porcentaje mayor de actividad industrial presenten un consumo medio mayor. Por el contrario, una mayor proporción de consumo industrial sobre el consumo total, no garantiza tener una mediana mayor. Esta idea se ve reflejada en los siguientes diagramas de cajas, que representan, para cada provincia, la distribución de cada uno de los tipos de actividades económicas, excluyendo los valores extraños. Además, el objetivo de este gráfico es ver las diferencias existentes en las distribuciones de cada tipo de actividad en las distintas provincias, lo que proporciona información acerca de los hábitos de consumo de cada actividad en las distintas zonas geográficas.

De los diagramas de cajas siguientes, en primer lugar, cabe destacar el hecho de que la distribución del consumo residencial, en general, presenta un rango mucho menor que la distribución del consumo industrial, lo que indica que los datos están más concentrados en torno a la mediana. También se puede observar como la media, representada por la línea verde de puntos es mayor en casi todos los casos, lo que indica la presencia de outliers positivos, que no están representados en

el gráfico. Una vez más, aparece reflejada la tendencia del consumo industrial a tomar valores más elevados.

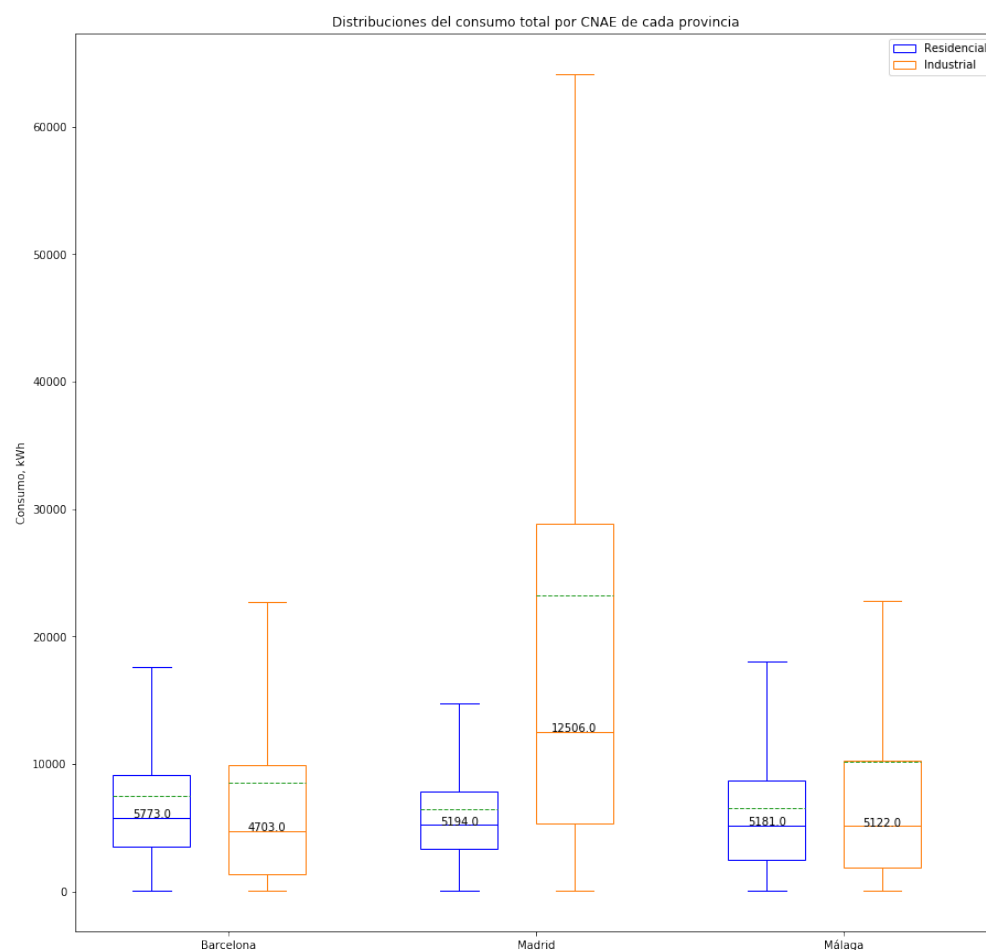


Figura 5.12.: Comparativa de las distribuciones del consumo total en función del tipo de actividad económica

Por otro lado, la provincia cuyas distribuciones presentan los valores más elevados del consumo industrial y, por tanto, mayor diferencia respecto al consumo residencial, es Madrid. Además, se puede observar, como Málaga es la provincia con un menor consumo tanto residencial, mientras que Barcelona presenta el menor consumo industrial. Este hecho puede tener relación con el clima o con el uso de otras fuentes de energía, como solamente se dispone de datos sobre el consumo eléctrico, es imposible saber la diferencia entre la cantidad de energía eléctrica consumida y la cantidad total de energía consumida.

5.2.2. Comparativa basada en el día de la semana

Anteriormente, en la sección 5.1.2, se establece como la diferencia existente en la distribución en los distintos días de la semana es prácticamente mínima. Esto se debe a que el consumo residencial, que representa la mayor parte del consumo, permanece casi inalterado. Por el contrario, si que se infería una disminución en el consumo de origen industrial en los fines de semana, aunque debido a su poco peso sobre el consumo total, no se apreciaba en la distribución del consumo total.

El objetivo de esta sección, es apreciar las diferencias, si es que existen, en las variaciones del consumo dependiendo del día de la semana, ya sea un día de diario, un sábado o un domingo, para cada una de las provincias objeto de estudio. Para alcanzar dicho objetivo, en primer lugar, se ha calculado el consumo medio de cada provincia, para cada uno de los tres grupos en los que se han dividido los días de la semana. Los valores calculados están representados en el siguiente gráfico.

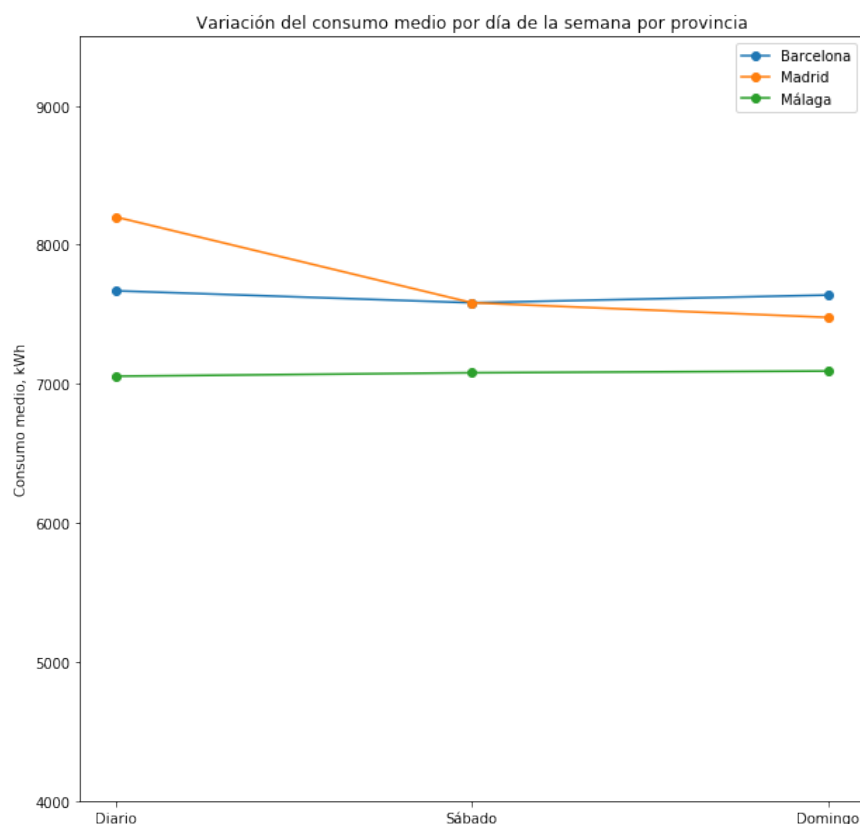


Figura 5.13.: Comparativa del consumo medio de cada provincia en función del día de la semana

En el diagrama anterior, se puede observar una tendencia descendente, es decir, el consumo medio durante el fin de semana es menor que durante los días laborables, tendencia que ya había sido descubierta en la sección 5.1.2. La única excepción es la provincia de Málaga, cuyo consumo máximo se da el domingo. Además, la provincia de Madrid presenta la pendiente más pronunciada, lo que puede ser debido a una mayor reducción del consumo industrial en comparación con el resto de provincias. Con el objetivo de determinar si la hipótesis anterior es correcta se ha creado

el siguiente gráfico, que muestra la variación del consumo por días de la semana desgranado en función del tipo de actividad económica para cada una de las provincias.

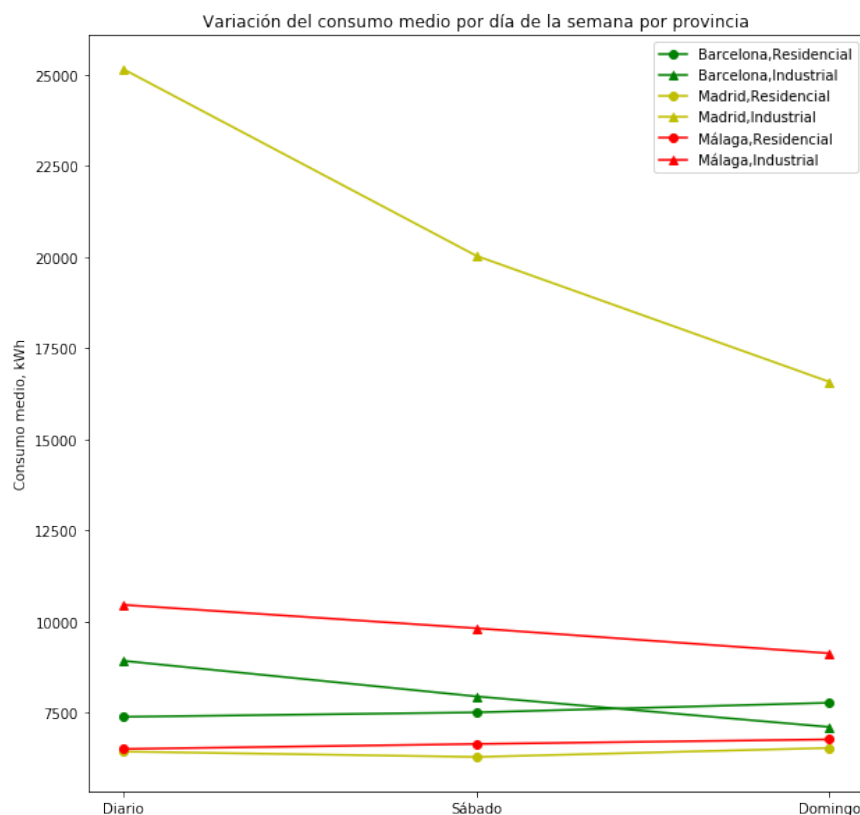


Figura 5.14.: Comparativa del consumo medio de cada provincia en función del día de la semana y el tipo de actividad económica

En primer lugar, resaltar que, para todas las provincias, el consumo industrial, para cada una de las categorías, es mayor que el residencial, lo que no es de extrañar, como se expuso en la sección 5.1.1. Cabe destacar que la provincia con el mayor consumo medio total a lo largo de la semana, Madrid, es la que presentan un mayor consumo industrial, lo que concuerda con el hecho de que sean las provincias con un mayor porcentaje de consumo industrial, Sección 5.2.1. Además, la provincia de Barcelona presenta la menor variación entre el consumo industrial y residencial medio.

5.2.3. Comparativa basada en la hora del día

En la sección 5.1.3, se realiza un análisis inicial acerca de como el consumo eléctrico medio varía en función de la hora. De dicho análisis se extraen las siguientes tendencias en el consumo. En primer lugar, la variación del consumo se puede dividir en dos fases, una de ascenso y otra de descenso. Desde, aproximadamente, las 5:00 hasta las 22:00, el consumo presenta una tendencia ascendente, mientras que desde las 23:00 hasta las 4:00 el consumo medio desciende a medida que pasan las horas. Este comportamiento no es de extrañar ya que coincide con el flujo de actividad de una vivienda a lo largo del día. Se profundiza en las relaciones de este patrón con los distintos tipos de actividades económicas en la sección 5.3.

Con el fin de observar las variaciones en el consumo a lo largo del día, para cada provincia, se ha creado el siguiente gráfico.

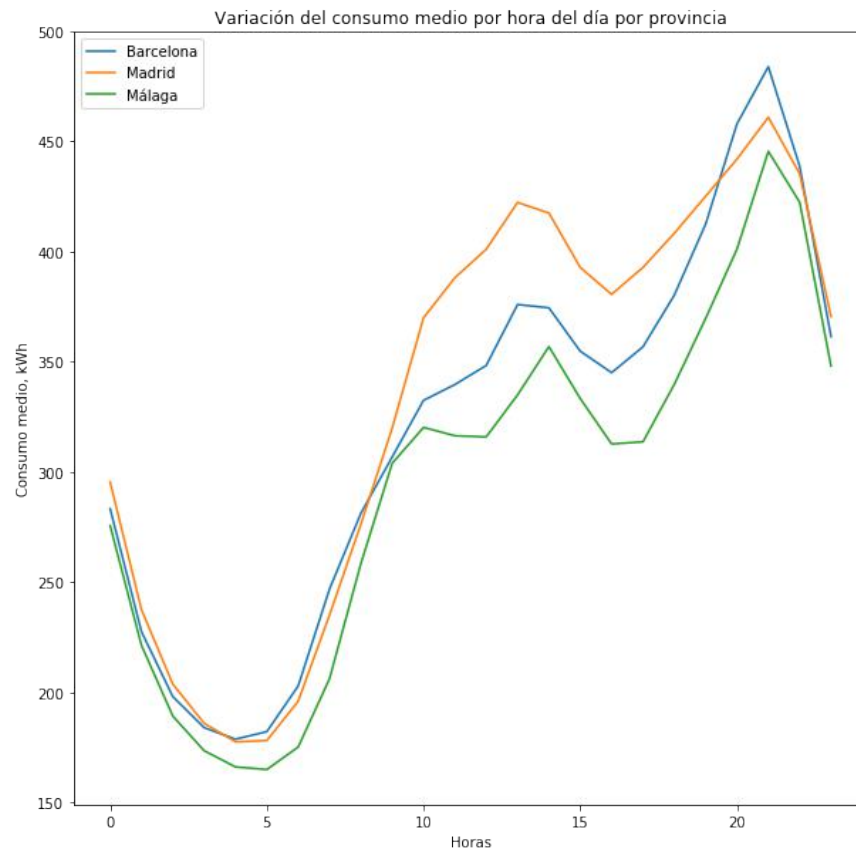


Figura 5.15.: Comparativa de la variación del consumo dependiendo de la hora del día

En el gráfico anterior, se puede observar como las tendencias descritas en la sección 5.1.3 se mantienen. Además, la provincia presenta, claramente, los niveles de consumo más bajos para cada una de las horas, lo que concuerda con lo observado en los apartados anteriores.

5.2.4. Comparativa basada en el tipo de tarifa de mercado

Las dos tarifas de mercado de las que se disponen son la regulada y la libre. La diferencia entre ellas reside en el hecho de que en la tarifa regulada el precio de la energía depende de la ley de la oferta y la demanda, mientras que en la tarifa libre, el precio por hora es fijo. En la sección 5.1.5, se establece que, aproximadamente, el 50 % del consumo eléctrico total corresponde a cada una de las tarifas. Sin embargo, el consumo medio de la tarifa libre es mayor que el de la tarifa regulada. Ante esta situación, se encuentra una relación existente entre el tipo de actividad económica y el tipo de tarifa de mercado. Es decir, la tarifa de mercado libre presenta un porcentaje de consumo industrial mayor, en comparación con la tarifa regulada, lo que concuerda con el hecho de que presente un consumo medio mayor.

El objetivo de este apartado es comprobar si las tendencias encontradas en el estudio general del consumo eléctrico, son aplicables a cada una de las provincias objeto de estudio. Para ello, en primer lugar se ha calculado el porcentaje correspondiente a cada tarifa de mercado, para cada una de las provincias.

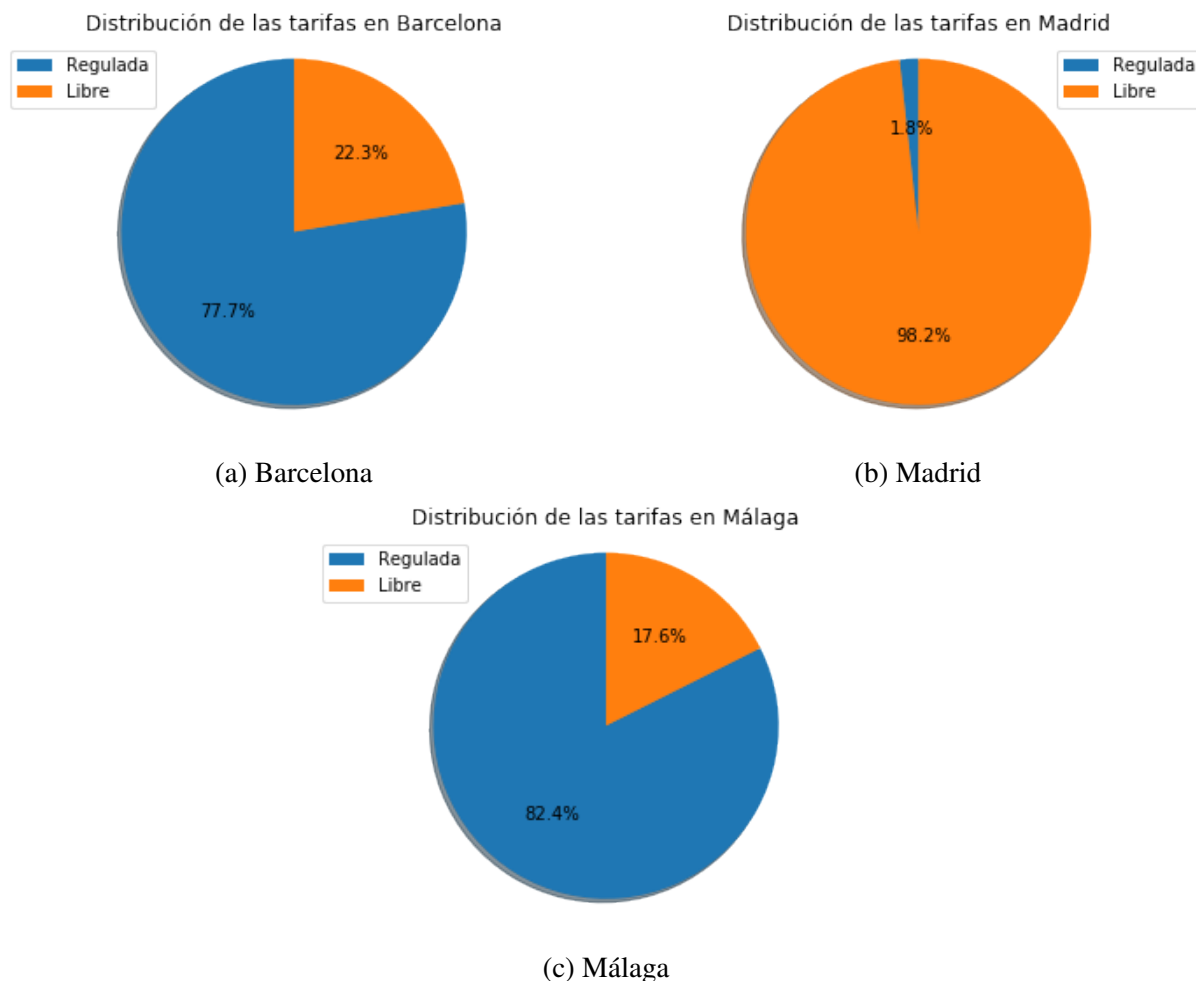


Figura 5.16.: Diagrama de sectores de la distribución de tarifas de cada provincia

Salta a la vista el hecho de que ninguna de las provincias presenta la proporción observada en la sección 5.1.5. Las cuatro provincias se pueden separar en dos grandes grupos; por un lado, Madrid, que presentan una clara mayoría de tarifa libre y por el otro lado, Barcelona y Málaga, que se decantan por la tarifa regulada. Ante esta situación, es necesario, evaluar el por qué de esta división tan clara, para ello se recurre a la influencia de otros factores.

A grandes rasgos, se puede considerar que la provincia de Madrid se caracteriza por presentar un consumo medio mayor, sección 5.2, y por poseer una mayor proporción de consumo industrial; lo contrario es aplicable al grupo formado por Barcelona y Málaga. En la sección 5.1.5, se plantea una relación existente entre el tipo de actividad económica y el tipo de tarifa de mercado, de tal manera que aquellos clientes con mayor proporción de consumo residencial, tienden a contratar la tarifa regulada, mientras que aquellos clientes en los que el consumo industrial presenta mayor peso, se decantan por la tarifa libre. La lógica detrás de esta relación es la siguiente. La tarifa libre ofrece un precio fijo sobre el consumo eléctrico, independientemente de la hora del día. Esta oferta es favorable para la industria, ya que necesita consumir energía a lo largo de todo el día, de ahí que haya un porcentaje mayor de consumo industrial en la tarifa libre.

Con el fin de indagar más profundamente acerca de esta relación, se ha creado el siguiente diagrama, que representa el porcentaje de consumo industrial y residencial que corresponde a cada provincia en función del tipo de tarifa de mercado.

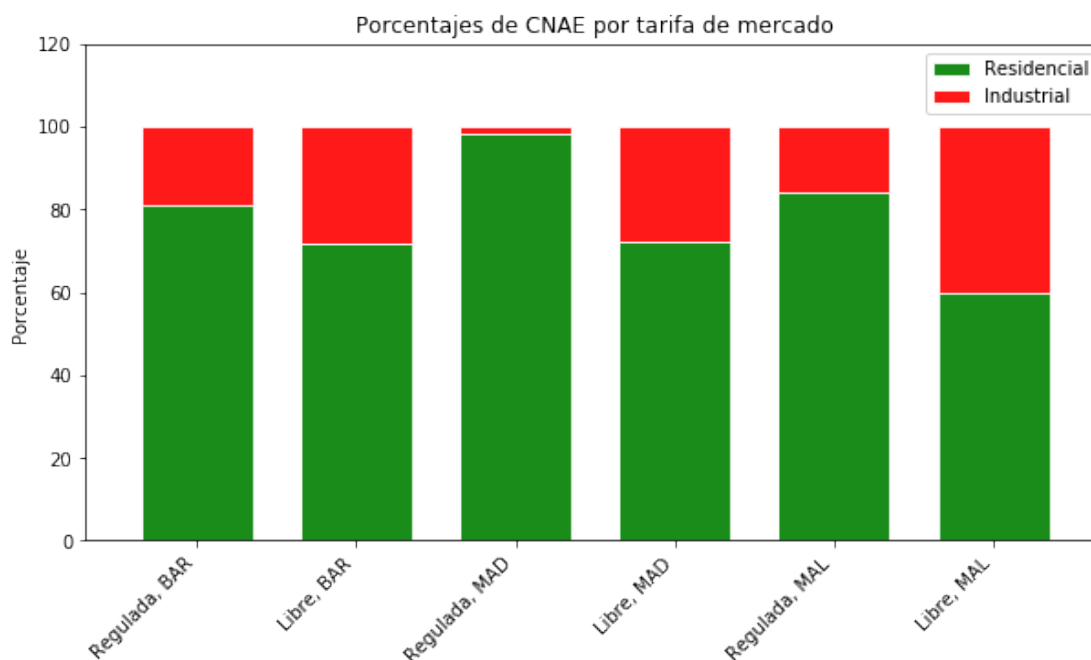


Figura 5.17.: Porcentajes de CNAE en función del tipo de actividad económica y de la provincia

A primera vista es apreciable que el tipo de actividad económica predominante es la residencial, lo que no es de extrañar, teniendo en cuenta los resultados obtenidos en las secciones 5.1.1 y 5.2.1. Además, se puede observar como el consumo de tipo industrial, abarca un mayor porcentaje en la tarifa de mercado libre, lo cual encaja con lo observado en la sección 5.1.5.

La principal diferencia entre los dos grupo previamente definidos, recae en que, para la provincia de Madrid, el porcentaje de consumo industrial presente en la tarifa regulada es ínfimo, en comparación con el grupo formado por Barcelona y Málaga.

5.3. Análisis Descriptivo Del Tipo De Actividad Económica Industrial y Residencial

A lo largo de las secciones se ha ido manifestando la importancia del tipo de actividad económica, ya que juega un papel crucial en las tendencias encontradas estudiando el resto de los factores. Debido a ello, el objetivo de esta sección es profundizar más acerca de la relación existente entre el tipo de actividad económica y el resto de factores, como son el día de la semana, la hora de la semana y el tipo de tarifa de mercado.

En primer lugar es necesario recapitular la información más general que ya se conoce acerca del consumo industrial y residencial. Como se muestra en la sección 5.1.1, los porcentajes correspondientes a cada una de las actividades son los siguientes:

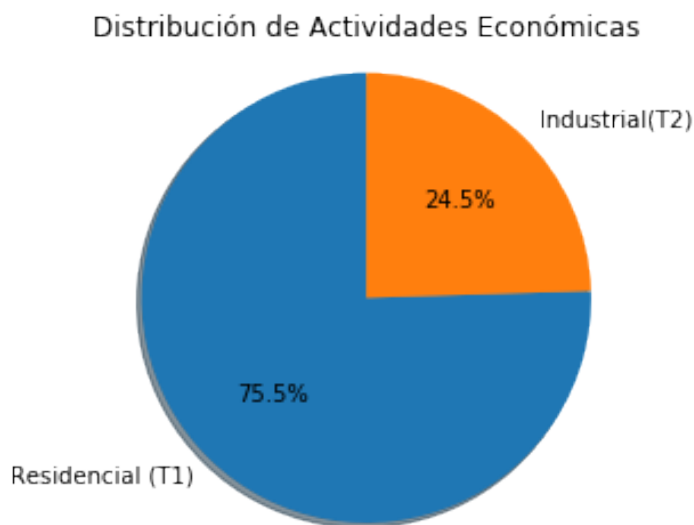


Figura 5.18.: Porcentaje de cada actividad económica

Además, también se han calculado los valores del consumo medio, tanto para el consumo industrial como para el consumo residencial, los resultados son los siguientes:

	RESIDENCIAL	INDUSTRIAL
Consumo Medio	7481.97	12130.27
Desviación Standard	7399.44	21570.61

Tabla 5.6.: Datos sobre la actividad industrial y residencial

Con el fin de recaudar más información acerca de las diferencias existentes entre el consumo residencial e industrial, se han creado los siguientes diagramas de cajas, excluyendo los valores extraños, que representan la distribución del consumo para cada una de las dos actividades económicas.

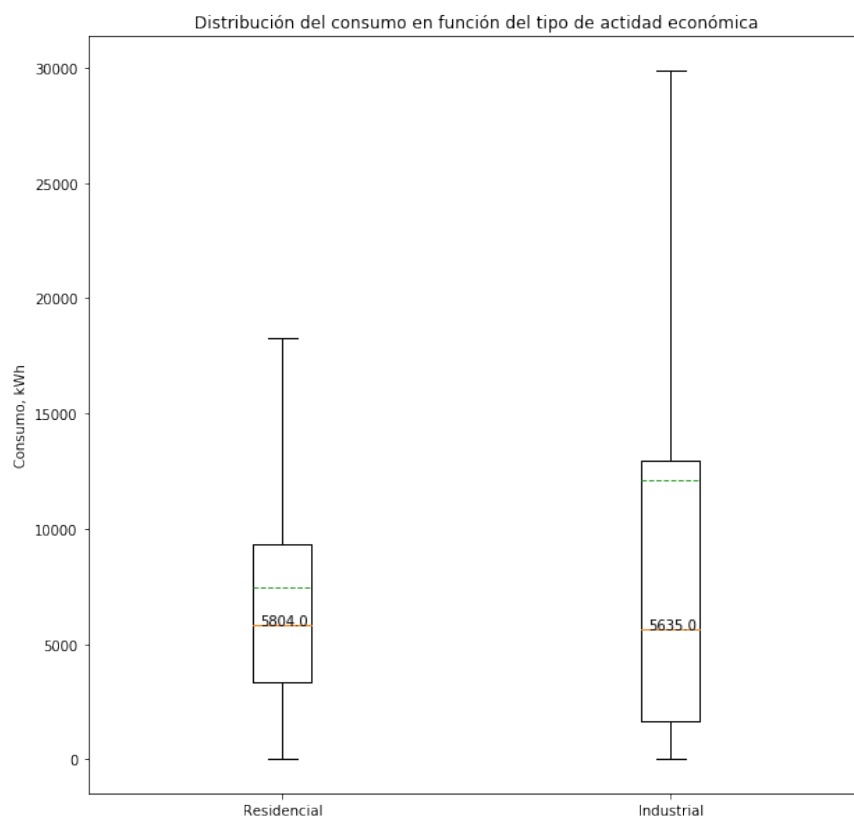


Figura 5.19.: Distribución del consumo residencial vs industrial

En los diagramas anteriores, se puede inferir la diferencia existente en el rango de las dos distribuciones, así como las variaciones entre las medias, representadas por las líneas de puntos, y las medianas. En primer lugar, la distribución del consumo residencial presenta un rango mucho menor que el consumo industrial, lo que contribuye a un menor sesgo con respecto a la media. Este sesgo es debido a la presencia de valores extraños, que no están representados en el diagrama.

5.3.1. Comparativa basada en el día de la semana

El objetivo principal de este apartado es determinar como el consumo de ambas actividades económicas varía en función del día de la semana. En la sección 5.1.2, se observa como el consumo eléctrico medio decrece durante el fin de semana. Una de las causas de este decrecimiento es la reducción del consumo industrial durante el fin de semana, lo cual concuerda con el cese de las actividades laborales. En esta sección, se recauda más información acerca de este fenómeno.

Con el fin de observar la tendencia del consumo eléctrico a lo largo de la semana se ha creado el siguiente gráfico, que muestra las variaciones del consumo medio para tipo de actividad económica para cada día de la semana.

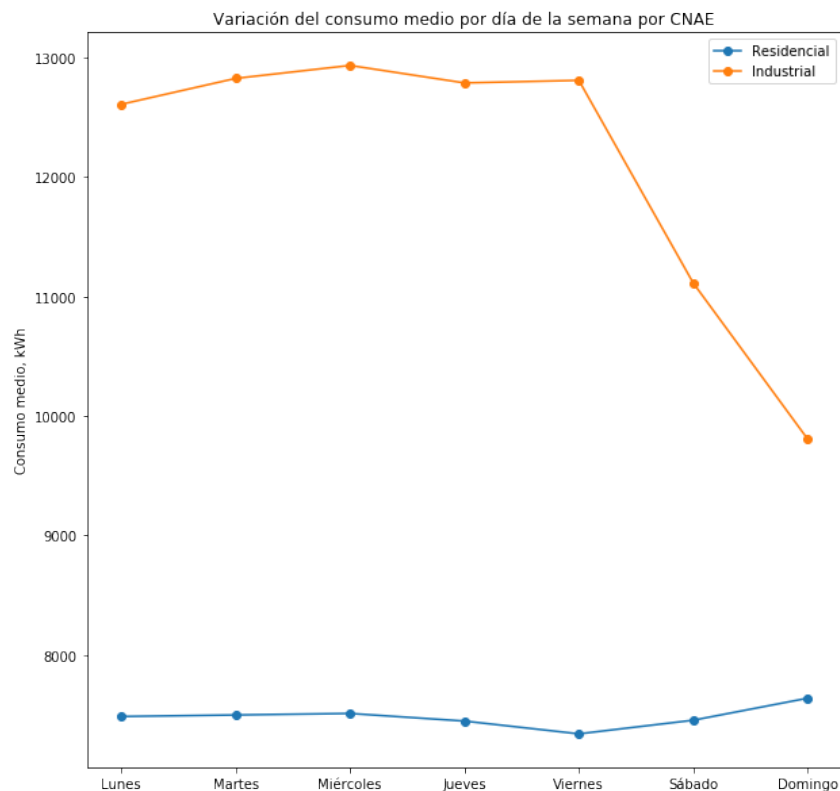


Figura 5.20.: Variación del consumo industrial y residencial a lo largo de la semana

Del gráfico anterior se extraen las siguientes conclusiones. En primer lugar, se aprecia como el consumo industrial presenta valores mucho más altos para cada día de la semana, lo cual no es de extrañar, teniendo en cuenta que el consumo medio industrial total es mayor que el residencial. Además, se confirma la hipótesis planteada, durante los días laborales, el consumo eléctrico se mantiene bastante estable, sin embargo, el consumo industrial presenta una gran caída durante el fin de semana. Por el contrario, el consumo residencial se mantiene bastante estable durante toda la semana.

5.3.2. Comparativa basada en la hora del día

A lo largo de la sección 5.1.3, se indaga acerca de las tendencias del consumo a lo largo de un día cualquiera. Se observa que, a lo largo del día, el consumo se puede dividir en dos etapas, una de crecimiento y otra de decrecimiento. La primera abarca desde, aproximadamente, las 5:00 hasta las 22:00, mientras que la decrecimiento el resto. Mirando más atentamente, se percibe una leve subida del consumo que genera un máximo relativo entorno a las 15:00.

El patrón observado concuerda perfectamente con la rutina de una vivienda. Durante las horas del sueño, el consumo es mínimo y a medida que avanza el día, y el sol cada vez se encuentra más bajo el consumo incrementa hasta llegar de nuevo al periodo de sueño. Por el contrario, si se mueve el foco a las actividades industriales, una tendencia un tanto distinta es de esperar. En primer lugar, la jornada laboral consta de ocho horas, por lo que el consumo debería de ser menor fuera de ese rango. Además, el factor luz se presupone que juega un papel menos importante teniendo en cuenta que en las empresas la mayor parte del consumo eléctrico es generado por la maquinaria y los aparatos electrónicos, por lo que el consumo debería ser más constante a lo largo de la jornada laboral.

Con el fin de verificar la veracidad de las hipótesis anteriormente planteadas, se han creado los siguientes gráficos. El objetivo de los mismo es mostrar la variación del consumo medio a lo largo del día, tanto para el consumo residencial como industrial.

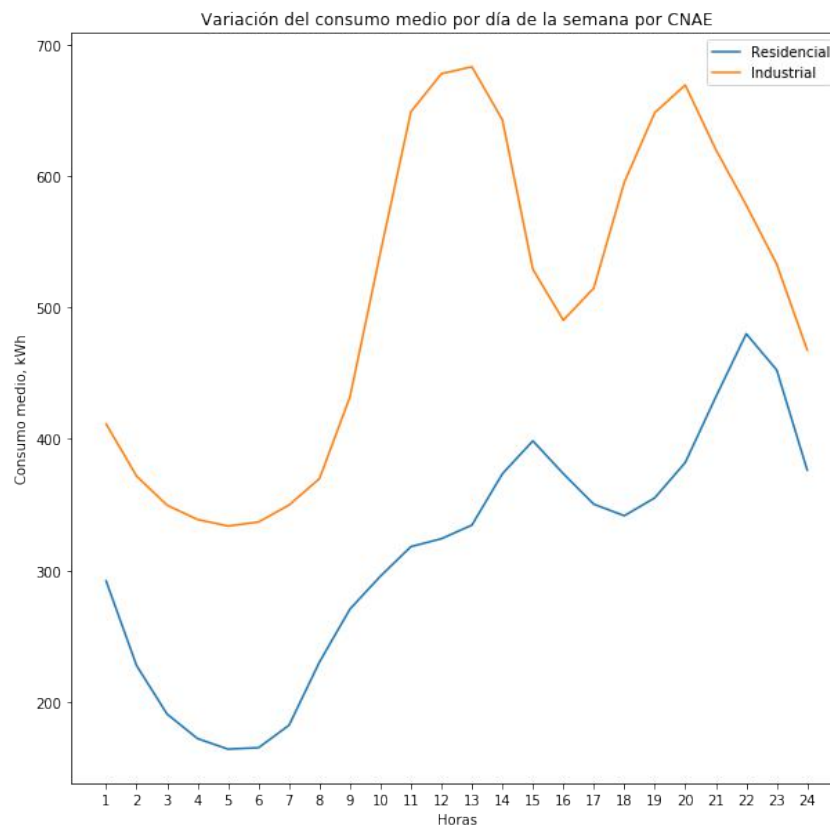


Figura 5.21.: Variación del consumo medio a lo largo del día en función del CNAE

A primera vista, se observa como en todo momento, el consumo industrial toma valores más elevados que el consumo residencial, lo cual no es de extrañar, teniendo en cuenta que el consumo medio de éste es más elevado, sección 5.1.1. Además, la mayor parte de las hipótesis planteadas, resultan que son correctas.

Con respecto al consumo residencial, se aprecian perfectamente las dos tendencias mencionadas, tanto la de subida como la de bajada. Incluso se mantiene el máximo relativo que aparece entorno a la hora de comer. Las similitudes entre el gráfico del consumo residencial y el diagrama de barras que representa las variaciones del consumo total, Figura 5.5, son de esperar, teniendo en cuenta que la mayor parte de los datos de consumo objeto de estudio son del tipo residencial.

Por el contrario, el consumo industrial presenta una forma muy distinta a la observada en el diagrama de barras de la Figura 5.5. En primer lugar, el consumo se comporta de la misma manera en dos intervalos de horas distintos. El primer intervalo comprende, aproximadamente, desde las 9:00 hasta las 16:00; mientras que el segundo abarca, aproximadamente, desde las 16:00 hasta 24:00. Cabe destacar, que entre las 15:00 y las 17:00 se produce una gran bajada en el consumo, la cual puede estar relacionada con el cese de la jornada laboral durante ese periodo de tiempo. En un principio, se asumía un gráfico más constante en las franjas de tiempo correspondientes a la jornada laboral, sin embargo, se perciben dos etapas, una de subida y otra de bajada, bien definidas.

En la sección 5.3.1, se observa la gran bajada que el consumo industrial experimenta durante los fines de semana y la constancia del consumo residencial a lo largo de toda la semana. Con estos datos en mente, se han creado los siguientes gráficos, que muestra la variación del consumo medio a lo largo del día para cada tipo de día la semana y para cada actividad económica.

En primer lugar, la figura 5.22a muestra la variación del consumo medio los días laborables, es decir, de lunes a viernes. Las diferencias existentes entre este gráfico y el descrito anteriormente, Figura 5.21, es despreciable, lo cual era de esperar teniendo en cuenta que este gráfico abarca la mayor parte de los días de la semana. El consumo residencial también sigue esta línea, ya que apenas varía durante el fin de semana.

Por el contrario, cierto cambio es apreciable en el consumo industrial. Cabe destacar, que la curva formada por los valores medios del consumo a lo largo del día, se ha desplazado, tomando ahora valores más bajos, como se predecía, teniendo en cuenta los resultados de la sección 5.3.1. Durante los días laborales, se pueden apreciar dos máximos, aproximadamente del mismo valor. El primero de ellos se da entorno a las 12:00, mientras que el segundo sucede entorno a las 20:00. Sin embargo, durante los fines de semana solo se puede apreciar un pico, que se da al mediodía los sábados y a última hora de la tarde durante los domingos.

La disminución sufrida por el consumo industrial durante los fines de semana, con respecto a los días laborales, es suficiente como para justificar la caída del valor del consumo medio observada en la figura 5.20. Para demostrarlo, se ha calculado el consumo medio para cada día de la semana, utilizando los datos con los que se crearon los diagramas siguientes, que son muy similares a los valores observados en el Figura 5.20.

Día de la semana	Consumo industrial medio diario (kWh)
Días Laboral	12793.92
Sábados	11113.04
Domingos	9806.82

Tabla 5.7.: Consumo industrial medio diario por día de la semana

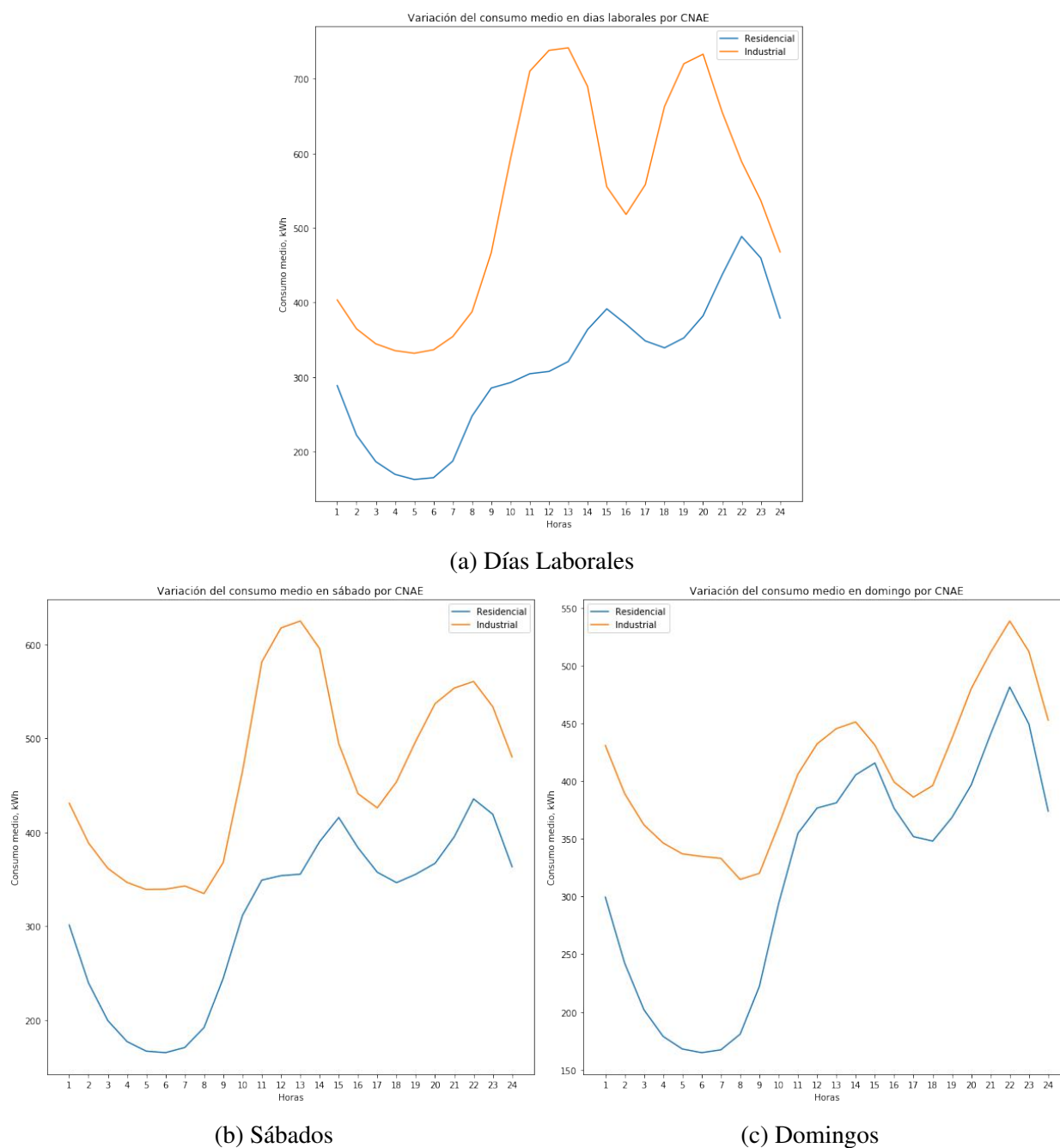


Figura 5.22.: Variación del consumo medio en función del tipo de día de la semana

5.3.3. Comparativa basada en la tarifa de mercado

En los datos objeto de estudio, hay dos tarifas de mercado, la tarifa regulada y la tarifa libre, que se diferencian en el modo de establecer el precio de la energía. Por un lado, la tarifa de mercado regulada, establece los precios en función de la ley de la oferta y la demanda, por lo que varía a lo largo del tiempo. Por el contrario, en la tarifa de mercado libre el precio de la energía se acuerda entre la distribuidora y el cliente y se mantiene fijo a lo largo del tiempo.

Con el fin de crear una imagen inicial acerca de la relación existente entre las tarifas de mercado y el tipo de actividad económica, se han creado los siguientes diagramas de sectores.

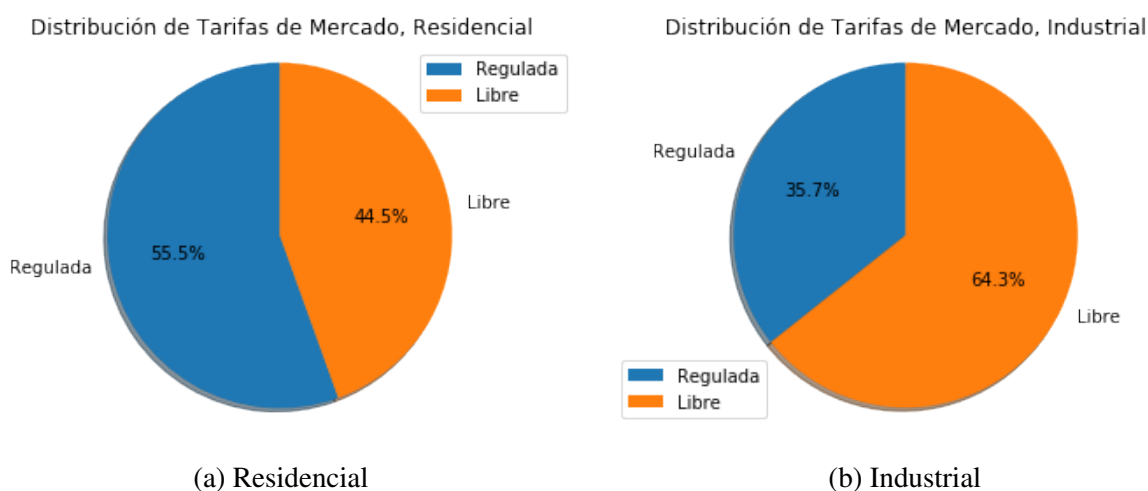


Figura 5.23.: Diagrama de sectores de los tipos de tarifas de mercado en función del CNAE

En los diagramas anteriores, se puede percibir cierta inclinación del consumo residencial hacia la tarifa de mercado regulada y del consumo industrial hacia la tarifa de mercado libre, lo cual ya se había inferido en la sección 5.1.5. Con el fin de recaudar más información acerca de la tendencia mencionada, se ha calculado el consumo medio para cada tarifa de mercado en cada tipo de actividad económica. Los resultados se han reflejado en el gráfico 5.24.

En la figura 5.24, se observa como el consumo medio mayor se da en el consumo industrial con la tarifa de mercado libre, lo cual no es de extrañar teniendo en cuenta que el consumo industrial medio es mayor que el residencial, sección 5.1.1 y que el consumo medio de la tarifa libre es mayor que el de la tarifa regulada, sección 5.1.5. Además, en la sección 5.3, se establece que los valores elevados del consumo medio para las actividades económicas industriales se deben a la presencia de valores extraños con niveles de consumo muy elevados que sesgan la media. Con esta idea en mente, se han creado los siguientes diagramas de cajas, Figura 5.25, con el fin de determinar si el elevado valor del consumo medio para la tarifa libre en el consumo proveniente de actividades industriales, se debe a la presencia de valores extraños o resulta que el rango de la distribución del consumo de los clientes que cumplen dichas condiciones toma valores mayores.

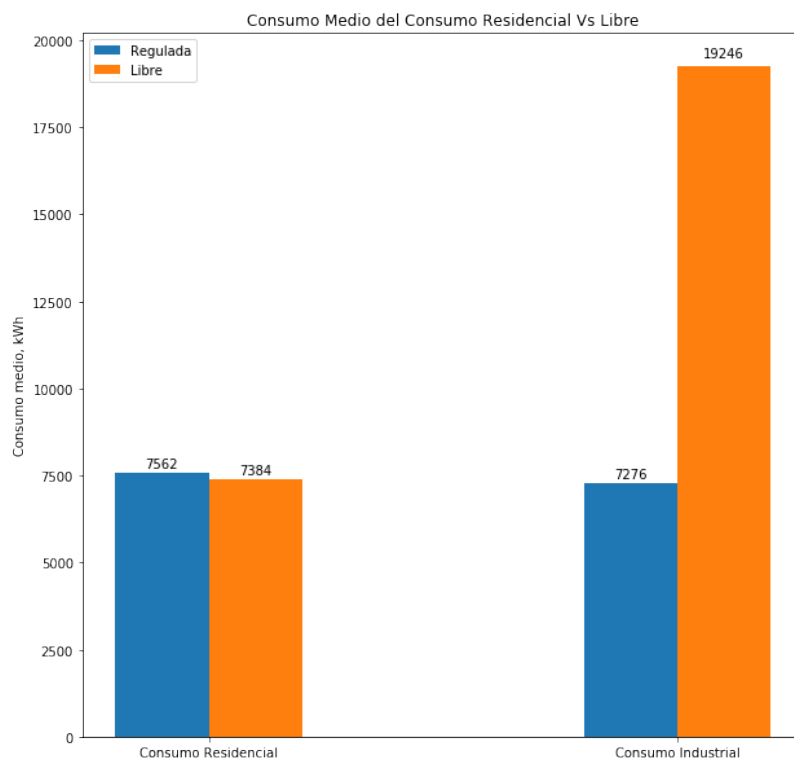


Figura 5.24.: Consumo medio de cada actividad económica por tarifa de mercado

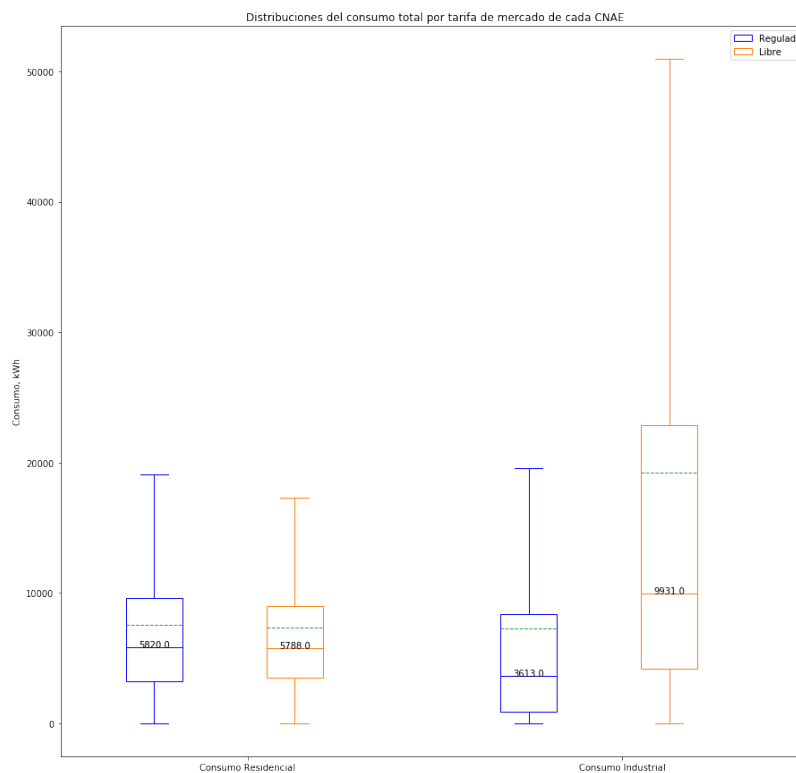


Figura 5.25.: Distribución de los tipos de tarifa de mercado en función del CNAE

La diferencia entre la media, representada por la línea verde de puntos, y la mediana indica el sesgo introducido por los valores extraños, no representados en los diagramas de cajas. Sin embargo, se mantiene el hecho de que los mayores valores de consumo se caracterizan por pertenecer a la tarifa de mercado libre y originarse en actividades económicas de tipo industrial. Por el contrario, los valores de consumo menores se dan en el consumo industrial para la tarifa regulada. Esta gran diferencia en los niveles de consumo dependiendo del tipo de tarifa de mercado, solamente se manifiesta en el consumo industrial. En el consumo residencial, los valores del consumo de la tarifa libre están ligeramente por encima de los de la tarifa regulada, pero la diferencia es mínima, comparando con la existente en el consumo industrial.

5.4. Modelo de predicciones utilizando el algoritmo K-Means

El principal objetivo de este apartado es mostrar los distintos perfiles de consumo observados a lo largo del día, en función de los siguientes factores; la provincia (Madrid, Barcelona y Málaga), el tipo de día de la semana, la estación del año y las actividades económicas, residencial o industrial. Gracias al cálculo de estos perfiles las distribuidoras eléctricas obtendrán información acerca de cuánta energía eléctrica se debe producir, así como información sobre distintos perfiles de usuarios, para los que podrá crear ofertas personalizadas.

Para obtener los distintos perfiles de consumo, se ha utilizado un algoritmo de agrupamiento denominado K-Means. A grandes rasgos, este algoritmo se encarga de la clasificación de los datos proporcionados en distintos grupos. El número de grupos creados depende de una variable que se le debe proporcionar, la cual normalmente se denota por k . Más información sobre el funcionamiento de este algoritmo se puede encontrar en la sección 4.3.1.

Los patrones de cada uno de los grupos, son obtenidos utilizando el 75 % de los datos objeto de estudio. El 25 % restante es empleado para determinar la fiabilidad de los perfiles obtenidos. Dicha fiabilidad se interpreta como la diferencia media existente entre los porcentajes de representación de cada grupo calculados en la fase de predicción, con los calculados en la fase de validación.

5.4.1. Municipios de la provincia de Madrid

En este apartado, el estudio se centra en los municipios pertenecientes a la provincia de Madrid. Es necesario recordar las conclusiones obtenidas en secciones anteriores acerca del consumo de la provincia de Madrid, ya que esta información será empleada a lo largo de este apartado.

La provincia de Madrid presenta un consumo medio igual a 8011.50 kWh, siendo la segunda provincia con el consumo medio más elevado. A su vez, este consumo, está formado por el generado por dos actividades económicas, la residencial y la industrial. En este caso, el consumo industrial representa un 27,7 % del consumo total, siendo la provincia con mayor presencia de este tipo de consumo. Además cabe destacar, que el consumo industrial sufre una gran caída durante los fines de semana, como se puede apreciar en la figura 5.14. También es necesario tener en cuenta que solo un 1,8 % del consumo generado pertenece a la tarifa regulada.

5.4.1.1. Perfiles del consumo de un día de invierno en la provincia de Madrid

El objetivo de este apartado es encontrar los perfiles que identifiquen a los distintos tipos de usuario según su consumo, tanto residencial como industrial, en cualquier día de invierno en la provincia de Madrid. La dinámica utilizada en este apartado es la misma que se va a emplear a lo largo de toda la sección.

En primer lugar, es necesario determinar el número de clústers en los que se deben agrupar los datos, valor que se obtiene a través de la creación de una gráfica que muestra la suma total de las varianzas para distintos números de clústers, k . Una vez determinado el número de clústers, se calculan los centroides de los mismos, que representan la evolución de cada patrón a lo largo de las horas del día. Además, es importante determinar la varianza de los patrones, ya que las empresas eléctricas necesitan saber el rango de valores entre los que se mueven los clientes que pertenecen a cada valor para poder estimar cuánta energía necesitan producir. A continuación, con el 25 % de los datos, que no han sido empleados en la creación del patrón, se comprueba el nivel de incertidumbre del modelo creado, es decir, el porcentaje de clientes que pertenecen a cada patrón debe ser aproximadamente igual al porcentaje calculado durante la creación de cada patrón.

Consumo residencial de un día laboral

El primer caso de esta sección predice los distintos patrones del consumo residencial un día laboral de invierno en Madrid. Para ello es necesario saber en cuántos clústers se pueden agrupar los datos pertenecientes a este grupo, por lo que se ha generado el gráfico de la figura 5.26.

En dicho gráfico, a primera vista, es complicado distinguir el “codo” de la función, por lo que se ha recurrido al cálculo del porcentaje de disminución de varianza introducido por cada aumento del número de clústers. Aplicando el método anterior, se ha observado que al pasar de siete a ocho clústers, la varianza disminuye en un 3.34 %. Por el contrario, si se aumenta el número de clústers a nueve, la varianza disminuye en un 1.43 %. Ante esta situación, se ha decidido emplear ocho clústers.

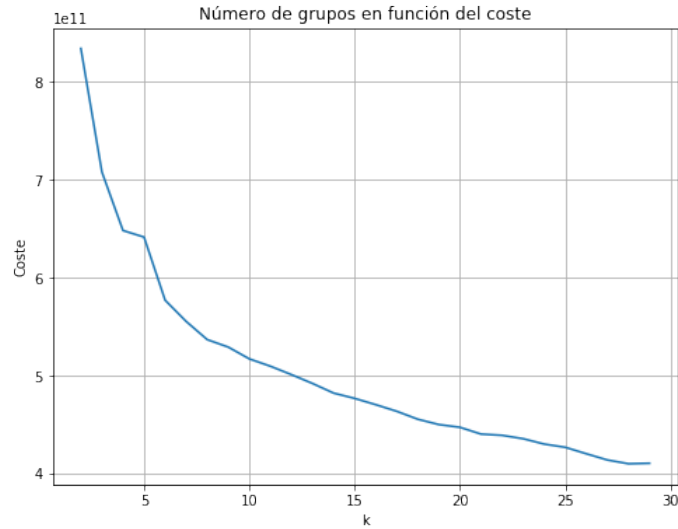


Figura 5.26.: Número de grupos en función de la suma de las distancias a los centroides

Con el número de clústers establecido, aplicando el algoritmo K-Means, se han obtenido los patrones representados en la figura 5.27. Además, en la figura de la derecha, se puede observar la desviación típica presente para cada patrón que es más notable en los patrones que poseen valores de consumo más elevados y representan un porcentaje ínfimo del conjunto de datos.

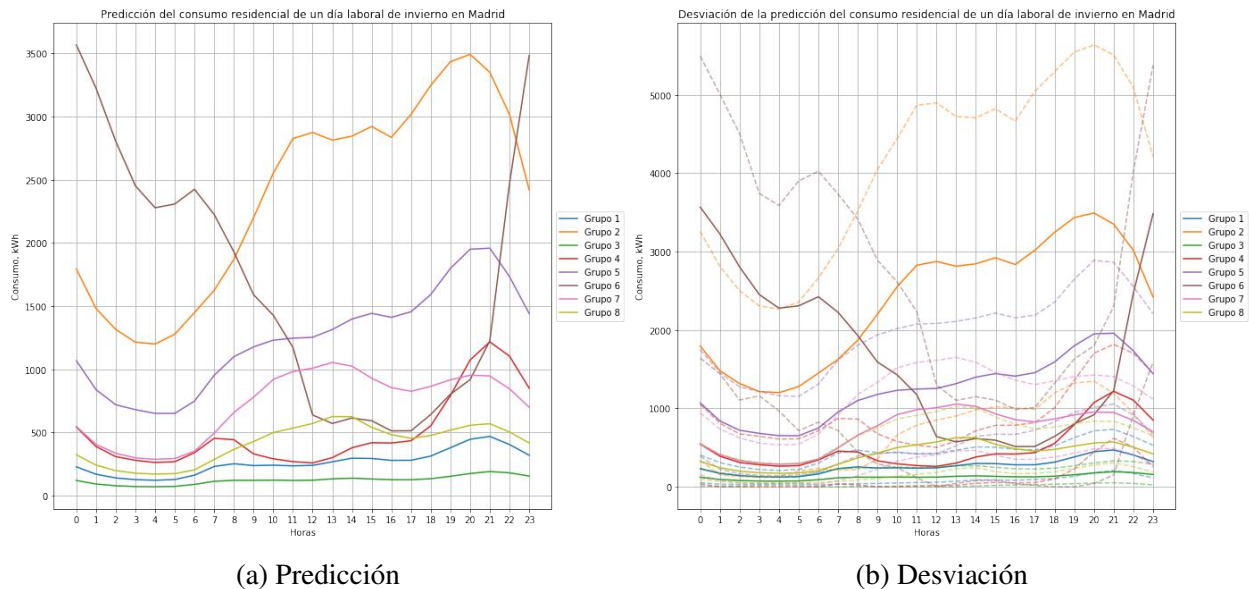


Figura 5.27.: Consumo residencial de un día laboral de invierno en Madrid

En la imagen, salta a la vista la extraña tendencia experimentada en el Grupo 6, los picos de consumo se encuentran durante las horas de sueño, mientras que por el día el consumo es mínimo, es decir, completamente al revés del resto de las tendencias. Este perfil coincide con el de profesionales que trabajan en el turno de noche, lo que encaja con el hecho de que sólo un 0.66 % de los

datos pertenezcan a este grupo, Tabla 5.8.

Siguiendo la línea esperada se encuentra el Grupo 2, que aunque la distribución del consumo no salta a la vista, sí que se caracteriza por poseer un consumo mayor y porque sólo un 0.41 % de los datos pertenecen al mismo. Lo que indica que este grupo está compuesto por individuos que tienen un consumo por encima de la media, lo que implica más gastos eléctricos dentro del hogar y se asocia con casas más grandes y una situación económica privilegiada.

Los grupos 5 y 4, se caracterizan por presentar, solamente, un pico de consumo, cuando habitualmente hay dos, uno a la hora de comer y otro a la hora de cenar. Por lo que a estos grupos pertenecerán aquellas personas que no acuden a comer a sus viviendas, que representan, en conjunto, un 10.88 % del total.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE(%)
Grupo 1	35.84 %	Grupo 5	1.56 %
Grupo 2	0.41 %	Grupo 6	0.66 %
Grupo 3	35.52 %	Grupo 7	4.28 %
Grupo 4	6.61 %	Grupo 8	15.13 %

Tabla 5.8.: Predicción del porcentaje de datos asociados a cada grupo

Teniendo como objetivo aplanar el gráfico del consumo, ya que de esta forma las productoras eléctricas optimizan las plantas de producción, los grupos cuya modificación del patrón tendrían más impacto son el Grupo 2 y el Grupo 6. Se podría plantear el lanzamiento de dos campañas publicitarias. Una de ellas se centraría en incentivar el consumo de los clientes pertenecientes al Grupo 6 durante las horas del día, mediante la propuesta de ofertas si consumen a dichas horas. Por otro lado a los clientes del Grupo 2 se les podría aumentar el precio de la energía pasado un determinado número de kW, para fomentar una reducción en el consumo.

Tras la obtención de los patrones, es necesario validarlos utilizando el 25 % restante de los datos. En este caso, se puede concluir que el modelo generado posee poca incertidumbre ya que la variación en la distribución de los datos en los distintos grupos es mínima, Tabla 5.9.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE(%)
Grupo 1	35.64 %	Grupo 5	1.64 %
Grupo 2	0.40 %	Grupo 6	0.63 %
Grupo 3	35.34 %	Grupo 7	4.33 %
Grupo 4	6.57 %	Grupo 8	15.44 %

Tabla 5.9.: Validación del porcentaje de datos asociados a cada grupo

Consumo residencial de los fines de semana

A pesar de que los patrones de consumo de los sábados y de los domingos se han generado de forma independiente, como el número de grupos obtenidos es el mismo y los patrones son muy similares, se estudian en conjunto. El número de clústers obtenidos en cada caso es igual a siete. En el caso del sábado, se ha elegido este valor teniendo en cuenta que al pasar de seis a siete clústers, la varianza disminuye en un 5.71 %, mientras que la pasar de siete a ocho clústers, tan solo disminuye en un 1.67 %, figura 5.28a. De la misma forma, en el caso de los domingos, el número de clústers escogido es igual a siete debido a que el incremento de seis a siete clústers disminuye la varianza en un 3.81 %, mientras que el aumento a ocho solamente genera una disminución del 1.96 %, figura 5.28b.

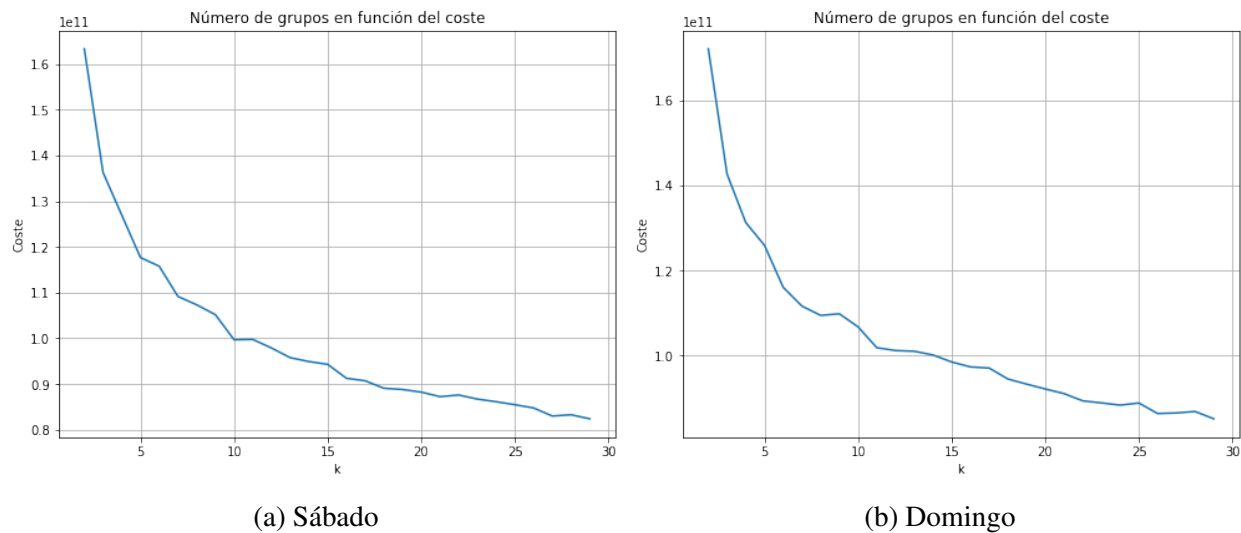


Figura 5.28.: Número de grupos en función de la suma de las distancias a los centroides

Una vez obtenidos los números de clústers a emplear, ya se pueden realizar las predicciones utilizando el 75 % de los datos objeto de estudio. Los resultados obtenidos se representan en las gráficas de la Figura 5.29 y la Figura 5.30.

Al igual que en el caso anterior, un grupo de clientes presenta una tendencia completamente distinta a la del resto, ya que sus picos de consumo se encuentran por la noche, mientras que por el día su consumo es mínimo, por lo que es probable que se jornada laboral se desarrolle por la noche. Es el caso del Grupo 5, en el caso del sábado y del Grupo 4 en el caso del domingo. En ambos caso el número de clientes que pertenecen a este patrón es mínimo, siendo un 0.75 % los sábados, Tabla 5.10, y un 0.46 % los domingos, Tabla 5.11.

También existe un conjunto de clientes cuyo consumo está muy por encima del consumo del resto de grupos. Siendo el Grupo 3 en las predicciones de los sábados y el Grupo 7 en la de los domingos. Dicho nivel de consumo se puede asociar con la posesión de viviendas más grandes o de coches eléctricos, lo que generaría un gran gasto eléctrico. El porcentaje de clientes que pertenece a este grupo es mínimo, siendo un 0.33 % los sábados y un 0.29 % los domingos.

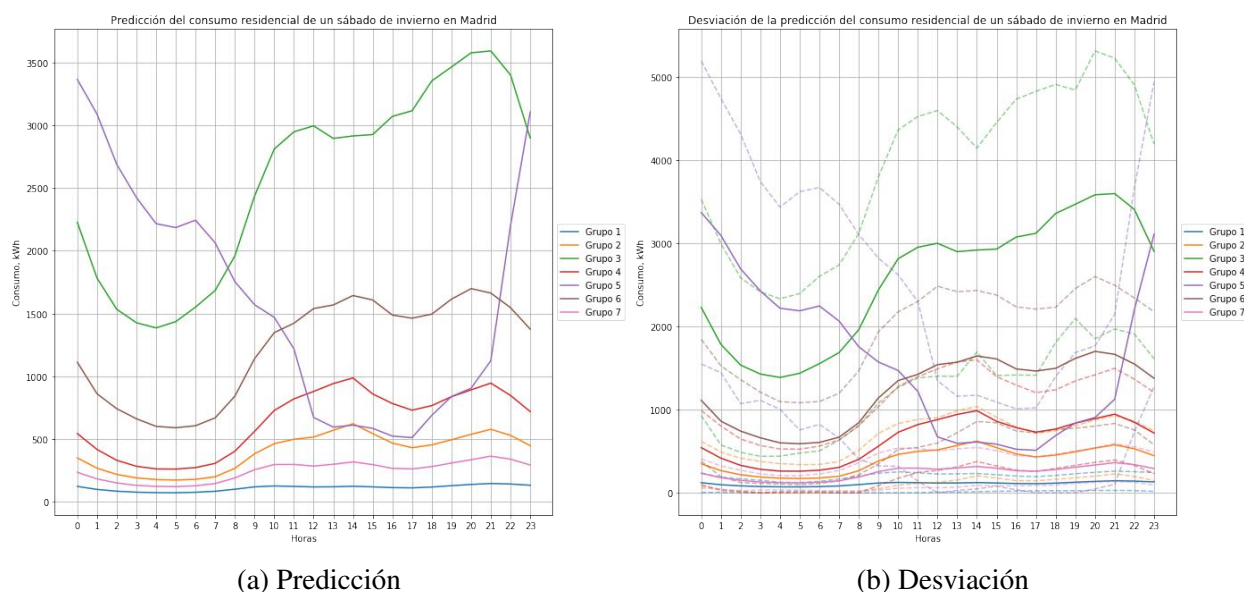


Figura 5.29.: Consumo residencial de un sábado de invierno en Madrid

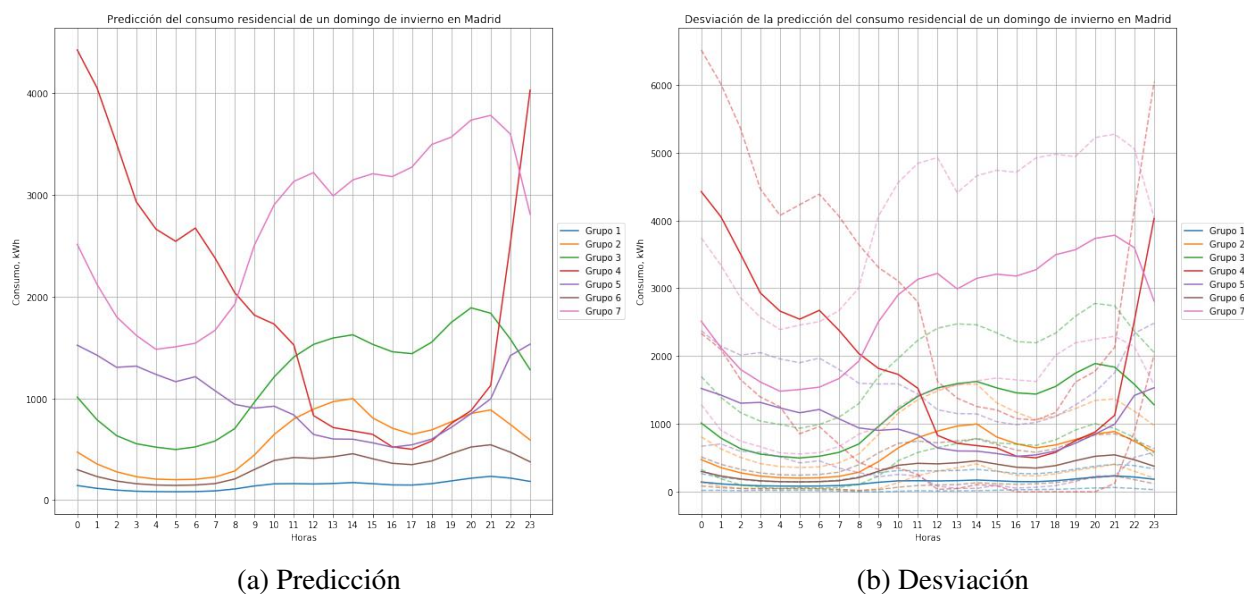


Figura 5.30.: Consumo residencial de un domingo de invierno en Madrid

Por último cabe destacar la presencia de un conjunto de clientes cuyos niveles de consumo son mínimos y se mantienen prácticamente constantes a lo largo del día. Es la situación del Grupo 1, tanto los sábados como los domingos. Además, este grupo es el que refleja el mayor porcentaje del consumo residencial en ambos casos, Tabla 5.10 y Tabla 5.11.

Con el objetivo de aplanar el consumo eléctrico a lo largo del día, se puede proponer, principalmente, dos campañas publicitarias. Una de ellas se centraría en la redistribución de los clientes de los grupos 4 y 5 que, para que desplacen parte de su consumo nocturno a las horas de luz. Esto

se puede conseguir mediante el lanzamiento de ofertas en las franjas horarias a las que se desea que se traslade el consumo. Además, se puede incentivar la reducción del consumo de los clientes del Grupo 7, mediante el aumento de precio del consumo tras sobrepasar cierto límite de consumo.

GRUPO	PORCENTAJE(%)	GRUPO	PORCENTAJE (%)
Grupo 1	30.89 %	Grupo 5	0.75 %
Grupo 2	22.12 %	Grupo 6	1.92 %
Grupo 3	0.33 %	Grupo 7	35.37 %
Grupo 4	8.61 %		

Tabla 5.10.: Predicción del porcentaje de datos asociados a cada grupo, sábado

GRUPO	PORCENTAJE(%)	GRUPO	PORCENTAJE (%)
Grupo 1	46.48 %	Grupo 5	1.35 %
Grupo 2	13.20 %	Grupo 6	36.18 %
Grupo 3	2.04 %	Grupo 7	0.29 %
Grupo 4	0.46 %		

Tabla 5.11.: Predicción del porcentaje de datos asociados a cada grupo, domingo

Con las predicciones calculadas, es necesario evaluar la fiabilidad de las mismas, para ello se comparan los porcentajes de clientes pertenecientes a cada grupos. En este caso, el modelo calculado es preciso, ya que las variaciones experimentadas en los porcentajes entre la predicción y la validación son del 0.22 % en el caso de los sábados y del 0.20 % en el caso de los domingos, Tabla 5.12 y Tabla 5.13.

GRUPO	PORCENTAJE(%)	GRUPO	PORCENTAJE (%)
Grupo 1	30.36 %	Grupo 5	0.85 %
Grupo 2	22.41 %	Grupo 6	1.90 %
Grupo 3	0.28 %	Grupo 7	35.20 %
Grupo 4	8.98 %		

Tabla 5.12.: Validación del porcentaje de datos asociados a cada grupo, sábado

GRUPO	PORCENTAJE(%)	GRUPO	PORCENTAJE (%)
Grupo 1	46.73 %	Grupo 5	1.18 %
Grupo 2	12.76 %	Grupo 6	36.47 %
Grupo 3	2.10 %	Grupo 7	0.37 %
Grupo 4	0.39 %		

Tabla 5.13.: Validación del porcentaje de datos asociados a cada grupo, domingo

Consumo industrial de un día laboral

Una vez estudiado el consumo residencial en invierno para todos los días de la semana, es el turno del consumo industrial. En primer lugar, antes de encontrar los patrones del consumo industrial de un día laboral de invierno en la provincia de Madrid, es necesario obtener el número de clústers a emplear. Para lograr dicho objetivo, se ha creado el siguiente gráfico, Figura 5.31. Como se puede observar, es “codo” de la función no se puede observar a primer vista, por lo que se ha recurrido al cálculo del porcentaje de disminución de la varianza. Al aumentar el número de clústers de once a doce, la varianza disminuye en un 5.33 %, mientras que al aumentar a trece, sólo se aprecia una disminución del 0.02 %, por lo que el número de clústers elegido es doce.

En la sección 5.3, se tratan las diferencias existentes entre el consumo residencial y el industrial. El consumo industrial se caracteriza por presentar un consumo medio mayor y por la simetría que presenta con respecto al centro de la jornada laboral. Además, el consumo industrial sufre una gran disminución durante los fines de semana. Con estas ideas presente, se pueden estudiar las diferencias existentes entre los patrones encontrados, Figure 5.32.

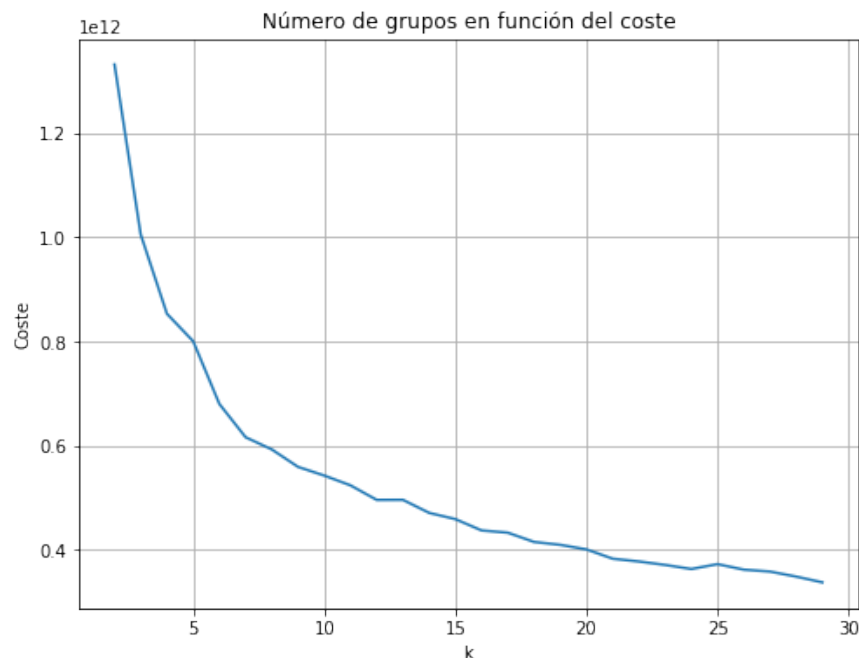


Figura 5.31.: Número de grupos en función de la suma de las distancias a los centroides

En primer lugar, es necesario comentar que a pesar de que se ha decidido emplear doce clústers, solamente once aparecen representados en la Figura 5.32. El Grupo 7 se ha eliminado de la gráfica ya que sólo recoge un conjunto de valores que son claramente superiores al resto, es decir, este patrón no recoge información útil para el análisis.

A primera vista, destaca el patrón seguido por los clientes pertenecientes a los Grupo 12, ya que presenta un consumo muy elevado que se concentra entre las 7:00 y las 15:00, principalmente,

lo cual no encaja con el patrón observado en la Figura 5.21 de la sección 5.3. Por este motivo, solamente el 0.17 %, Tabla 5.14, de los datos pertenecen a este grupo. De forma más sutil, el consumo del Grupo 5 y 3 sigue la misma tendencia.

También caben destacar las tendencias del Grupo 9 y del Grupo 2 cuyo consumo presenta el pico de último hora de la tarde, principio de la noche, más elevado que el pico del mediodía. Esto implica que son empresas cuya actividad industrial aumenta a la hora de la cena, por lo que puede ser el consumo de restaurantes, que durante los días de la semana tienen más clientela a la hora de cenar.

El resto de los patrones se pueden agrupar en dos categorías, aquellos grupos cuyo consumo se mantiene relativamente constante a lo largo de la jornada laboral (Grupo 11, 1 y 4); y los que presentan dos picos de consumo, uno por la mañana y otro por la tarde (Grupo 8, 6 y 10). Además, el consumo del Grupo 1 acapara el 45.87 %, Tabla 5.14, del consumo total, y es el grupo con el menor consumo. Lo que indica la gran diferencia existente entre las empresas, hay algunas pocas que generan un gran consumo, mientras que otras presentan niveles de consumo similares a los residenciales. Este hecho ya se infería en el diagrama de cajas de la Figura 5.19 de la sección 5.3.

Con las predicciones calculadas, es necesario evaluar la fiabilidad de las mismas, para ello se comparan los porcentajes de los clientes pertenecientes a cada grupo. En este caso, el modelo calculado presenta una mayor variación, en comparación con casos previos, Tabla 5.14 y Tabla 5.15. Este aumento de variación se debe a que las varianzas con las que se estaba lidiando eran mayores. Sin embargo, de media, un porcentaje presenta una variación igual a 0.15 %, que sigue siendo asumible.

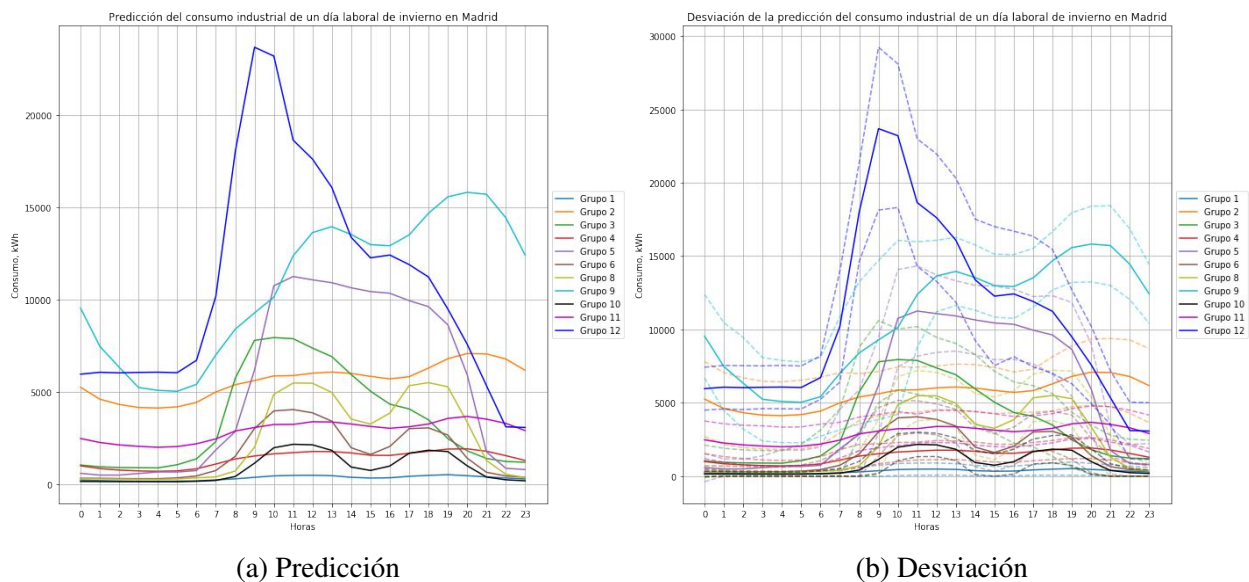


Figura 5.32.: Consumo industrial de un día laboral de invierno en Madrid

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	45.87 %	Grupo 5	0.77 %	Grupo 9	0.29 %
Grupo 2	1.77 %	Grupo 6	8.33 %	Grupo 10	16.53 %
Grupo 3	1.88 %	Grupo 7	0.003 %	Grupo 11	6.75 %
Grupo 4	14.53 %	Grupo 8	3.09 %	Grupo 12	0.17 %

Tabla 5.14.: Predicción del porcentaje de datos asociados a cada grupo

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	46.05 %	Grupo 5	0.70 %	Grupo 9	0.40 %
Grupo 2	1.83 %	Grupo 6	8.20 %	Grupo 10	16.06 %
Grupo 3	1.86 %	Grupo 7	0.00 %	Grupo 11	6.58 %
Grupo 4	14.84 %	Grupo 8	3.24 %	Grupo 12	0.24 %

Tabla 5.15.: Validación del porcentaje de datos asociados a cada grupo

Consumo industrial de un sábado

Como se ha ido haciendo hasta el momento, en primer lugar, es necesario calcular el número de clústers a emplear. En este caso, el “codo” de la función, Figura 5.33, está algo mejor definido que en los casos anteriores, pero aún así se ha recurrido al cálculo de la disminución relativa de suma de varianzas para seleccionar el número de clústers adecuado. Al aumentar el número de clústers de diez a once la suma de varianzas se ve reducida en un 9.16 %. Sin embargo, a partir de doce clústers, cada vez que se aumenta el número de clústers la suma de varianzas solamente varía en un 2 % aproximadamente, por lo que se ha elegido once como el número de clústers.

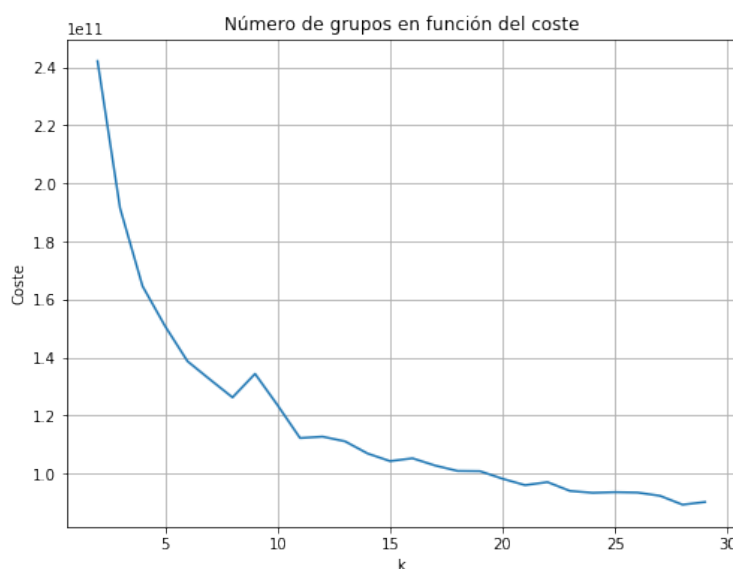


Figura 5.33.: Número de grupos en función de la suma de las distancias a los centroides

En este caso, aunque el número de clústers con el que se ha implementado el algoritmo K-Means es once, en la Figura 5.34, solamente se ven representados diez grupos. Esto se debe a que el Grupo 2 solamente reflejaba los valores de un conjunto de datos, por lo que no proporcionaba información que fuese útil en otros escenarios similares.

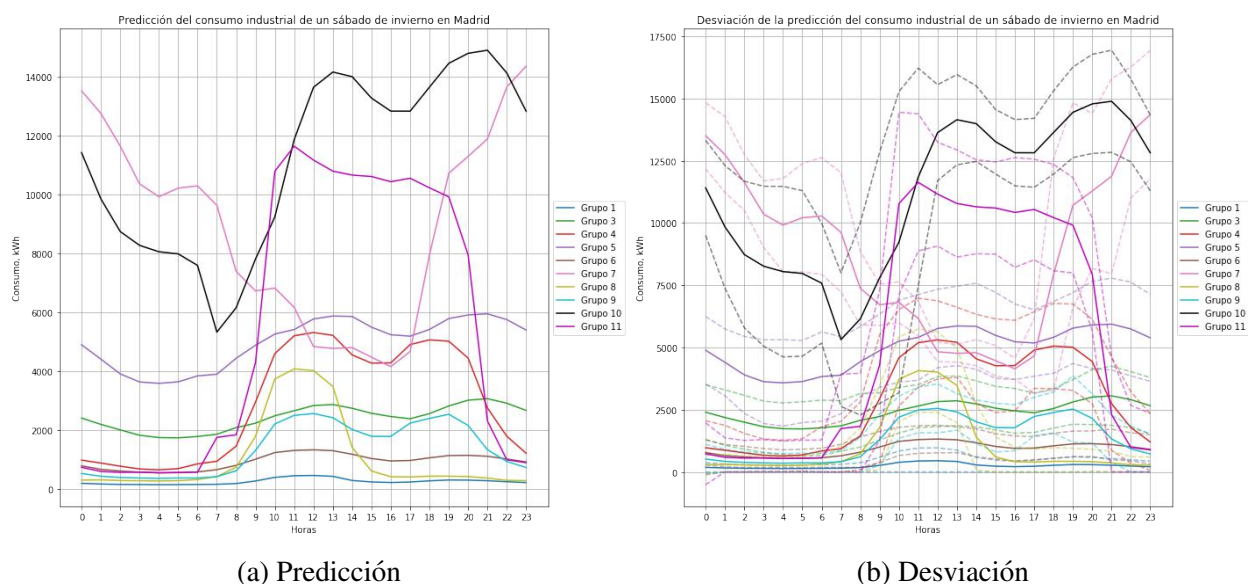


Figura 5.34.: Consumo industrial de un sábado de invierno en Madrid

A primera vista, destaca la presencia del Grupo 7 cuyos picos de consumo se dan por la noche, mientras que el consumo es mínimo durante las horas de luz. Este tipo de consumo se asocia con establecimientos cuya actividad principal se desarrolla durante la noche, como puede ser el caso de bares o discotecas. Lo cual también está respaldado por el hecho de que este patrón no se detectó durante los días laborales, cuando este tipo de locales suele permanecer cerrado. Este patrón refleja el 0.12 %, Tabla 5.16, del consumo industrial total de los sábados.

Además, llama la atención el hecho de que el Grupo 8 presenta un consumo prácticamente simétrico desde las 8:00h hasta las 15:00h. Este tipo de comportamiento es el propio de locales que solamente abren por la mañana los sábados, como puede ser el caso de panadería o quioscos. Este grupo representa el 8.56 %, Tabla 5.16, del consumo industrial de los sábados.

En la misma línea se encuentra en Grupo 11, que presenta un consumo constante a lo largo de la jornada laboral, cuando lo habitual en el consumo industrial es la presencia de una bajada del consumo durante las horas del mediodía. Motivo por el que solo un 0.59 %, Tabla 5.16, del consumo se ve reflejado en este patrón.

El resto de los grupos presentan la forma habitual del consumo industrial a lo largo del día, presentada en la sección 5.3.2. Residiendo la única diferencia en los niveles de consumo. Una vez más el grupo con el menor consumo, Grupo 1, representa el mayor porcentaje del consumo, en este caso el 53.54 %, Tabla 5.16.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	53.54 %	Grupo 5	2.25 %	Grupo 9	7.07 %
Grupo 2	0.02 %	Grupo 6	18.46 %	Grupo 10	0.38 %
Grupo 3	7.16 %	Grupo 7	0.12 %	Grupo 11	0.59 %
Grupo 4	1.86 %	Grupo 8	8.56 %		

Tabla 5.16.: Predicción del porcentaje de datos asociados a cada grupo

Con los grupos establecidos, es necesario determinar la estabilidad de los mismos. Para ello se comparan los porcentajes obtenidos para cada grupo en la fase de predicción, Tabla 5.16, y en la fase de validación, Tabla 5.17. De dicha comparación se obtiene una variación media de 0.32 % que es asumible, por lo que el modelo encontrado es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	53.18 %	Grupo 5	2.19 %	Grupo 9	6.92 %
Grupo 2	0.00 %	Grupo 6	17.46 %	Grupo 10	0.20 %
Grupo 3	8.04 %	Grupo 7	0.25 %	Grupo 11	0.87 %
Grupo 4	2.24 %	Grupo 8	8.65 %		

Tabla 5.17.: Validación del porcentaje de datos asociados a cada grupo

Consumo industrial de un domingo

En primer lugar, hay que averiguar el número de clústers que es necesario emplear, por lo que se ha generado una gráfica que representa la relación existente entre suma de las varianzas y el número de clústers, Figura 5.35. Para este escenario, el número de clústers elegido es quince. El principal motivo por el que se ha escogido este número es porque a partir de quince clústers el comportamiento errático de la función cesa y la disminución de la suma de varianzas avanza lentamente.

A pesar de que el algoritmo K-Means ha sido implementado utilizando quince clústers, la Figura 5.36, solamente muestra doce grupos. Esto se debe a que los grupos 2, 6 y 15 solo reflejaban información acerca de una o dos entradas de la base de datos, por lo que han sido eliminados del gráfico al no aportar información extrapolable a otros datos. Por ese mismo motivo, los porcentajes de los mencionados grupos en la tabla de validación toman el valor cero, Tabla 5.19.

Uno de los patrones que más destacan, es el seguido por el Grupo 3. Este patrón se caracteriza por presentar la forma del consumo industrial observada en la sección 5.3.2, pero en un intervalo de tiempo menor, desde las 11:00h hasta las 20:00h, aproximadamente. Este comportamiento puede coincidir con el de tiendas u otros locales que ofrecen servicio a terceros, que deciden reducir la jornada laboral durante los domingos. En la misma línea, destaca el Grupo 14, que solamente presenta consumo por la mañana; que se podría relacionar con comercios que sólo estén abiertos a

esas horas.

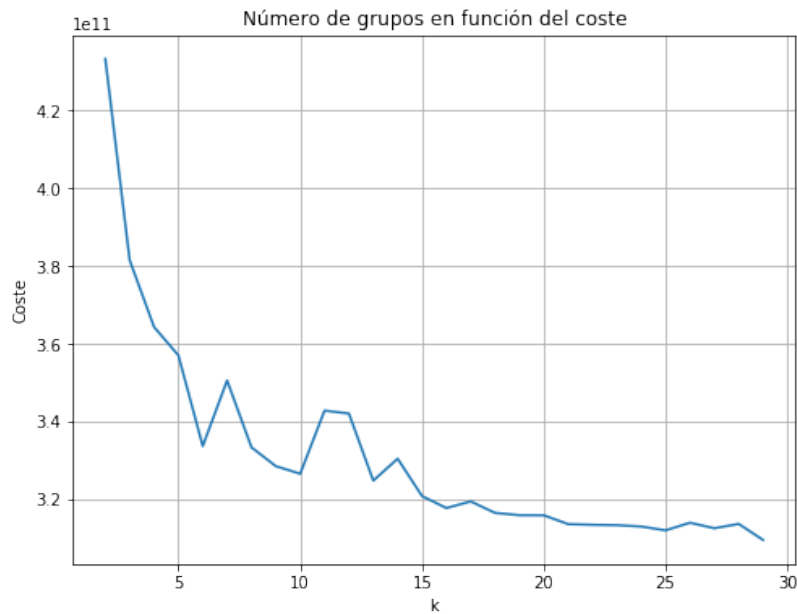


Figura 5.35.: Número de grupos en función de la suma de las distancias a los centroides

El resto de patrones siguen la forma esperada del consumo industrial de un domingo, diferenciándose en el nivel de los valores de consumo y en lo pronunciados que están los dos picos principales que caracterizan este tipo de consumo. Otra vez más, el grupo con el menor nivel de consumo es el que refleja un mayor porcentaje del consumo, en este caso el Grupo 1, con un 52.48% , Tabla 5.18.

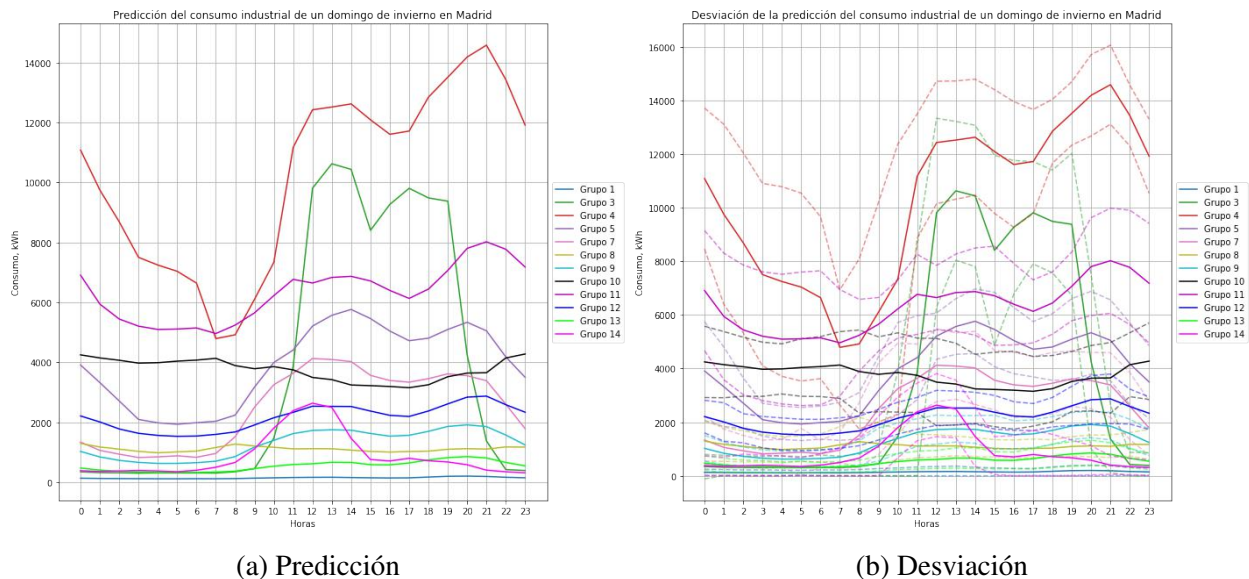


Figura 5.36.: Consumo industrial de un domingo de invierno en Madrid

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	52.48 %	Grupo 6	0.02 %	Grupo 11	1.01 %
Grupo 2	0.02 %	Grupo 7	1.90 %	Grupo 12	4.37 %
Grupo 3	0.35 %	Grupo 8	5.33 %	Grupo 13	21.68 %
Grupo 4	0.29 %	Grupo 9	7.70 %	Grupo 14	1.84 %
Grupo 5	1.10 %	Grupo 10	1.89 %	Grupo 15	0.03 %

Tabla 5.18.: Predicción del porcentaje de datos asociados a cada grupo

El hecho de haber generado grupos que solamente representaban uno o dos conjuntos de valores ha perjudicado a la fiabilidad del modelo. Sin embargo, las predicciones, Tabla 5.18, y las validaciones, Tabla 5.19, tan solo difieren en un 0.30 % de media, por lo que el modelo es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	50.78 %	Grupo 6	0.00 %	Grupo 11	0.57 %
Grupo 2	0.00 %	Grupo 7	2.12 %	Grupo 12	4.34 %
Grupo 3	0.47 %	Grupo 8	5.57 %	Grupo 13	22.04 %
Grupo 4	0.47 %	Grupo 9	8.16 %	Grupo 14	2.27 %
Grupo 5	1.09 %	Grupo 10	2.12 %	Grupo 15	0.00 %

Tabla 5.19.: Validación del porcentaje de datos asociados a cada grupo

5.4.1.2. Perfiles del consumo de un día de primavera en la provincia de Madrid

El objetivo de esta sección es encontrar los distintos patrones que representan el consumo residencial e industrial de un día de primavera en la provincia de Madrid. Para ello se va a seguir la misma dinámica empleada en la sección anterior, en primer lugar se obtiene el número de clústers a emplear, para después obtener los distintos patrones de consumo.

Consumo residencial un día laboral

Con el objetivo de encontrar el número de clústers a emplear, se ha generado el gráfico de la Figura 5.37. En dicho gráfico no se puede distinguir el “codo” de la función, por lo que es necesario recurrir al cálculo del porcentaje de disminución de la suma de varianzas para averiguar el número de clústers a emplear. En este caso el número de clústers elegido es doce. Utilizando doce clústers, se han obtenido los patrones representados en la Figura 5.38.

A primera vista, destaca la presencia del Grupo 2 cuyos valores de consumo son claramente superiores a los del resto. Sin embargo, sí que presenta la forma esperada del consumo residencial a lo largo de una jornada laboral, es decir, trayectoria ascendente durante las horas de luz y descendente durante la noche. Lo que llama la atención es el hecho de que el pico relativo de consumo experimentado, normalmente, a la hora de comer, se encuentra entre las 11:00 y las 13:00, un par

de horas antes de lo previsto. Además, presenta una pequeña depresión a las 18:00. Cabe destacar el hecho de que la desviación presente en este grupo es la mayor de todos los grupos, aunque solamente representa un 0.05 %, Tabla 5.20, del consumo residencial de un día laboral de primavera en Madrid.

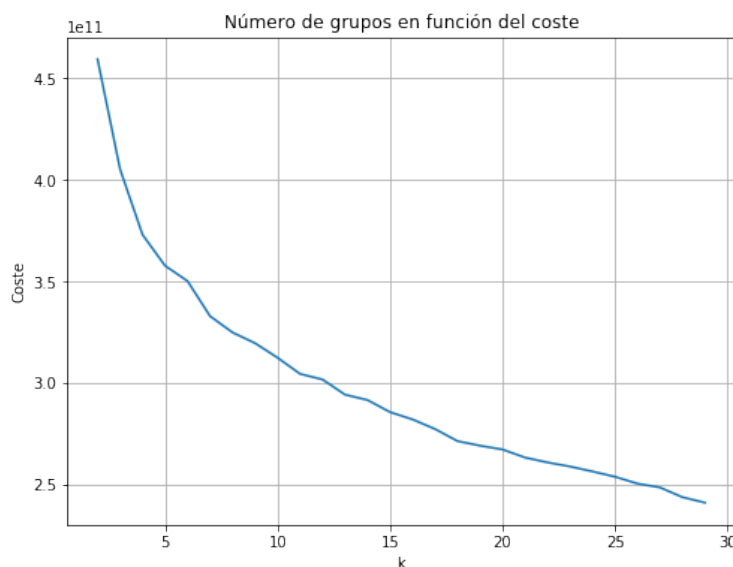


Figura 5.37.: Número de grupos en función de la suma de las distancias a los centroides

También destaca la tendencia seguida por el Grupo 11, ya que sus valores de consumo se mantienen relativamente constantes, en comparación con las variaciones entre el día y la noche esperadas en un consumo residencial, como se muestra en la sección 5.3.2. Sí que se experimenta cierta bajada durante las horas de sueño, a pesar de no ser tan evidente como se esperaba; esto se puede deber a la utilización de radiadores eléctricos o de aire acondicionado, dependiendo de como se presente la primavera, por la noche, lo que aumenta el consumo durante esas horas. Este comportamiento no es habitual, ya que solamente el 0.25 %, Tabla 5.20, del consumo residencial de un día laboral de primavera en Madrid, se ve representado por dicho patrón.

Además, el conjunto de grupos formado por el Grupo 5, 6, 7, 9 y 10; solamente presentan un pico diferenciado a lo largo del día, cada grupo a diferentes horas. Este tipo de tendencias se pueden asociar a hogares en los que el consumo se ve incrementado en franjas horarias muy determinadas debido a la presencia de personal de limpieza. Esta situación no cuadra especialmente bien con el Grupo 9, ya que este grupo experimenta el pico de consumo entre las 20:00 y las 23:00, por lo que es más probable que el pico de este grupo se deba a que es el único momento del día en el que hay personas en la vivienda, después de una larga jornada laboral.

Cabe resaltar el cambio en la magnitud de los picos experimentado por el Grupo 8. Normalmente, el pico de consumo del mediodía es menor que el experimentado a la hora de cenar. Sin embargo, en este caso el pico de la hora de la comida es mayor, lo que indica que por algún motivo estas viviendas generan un menor consumo durante la hora de la cena. Lo cual se puede

deber principalmente por dos motivos, o utilizan fuentes de luz no eléctricas o no necesitan la tanta energía eléctrica para cocinar ya que emplean cocinas de gas.

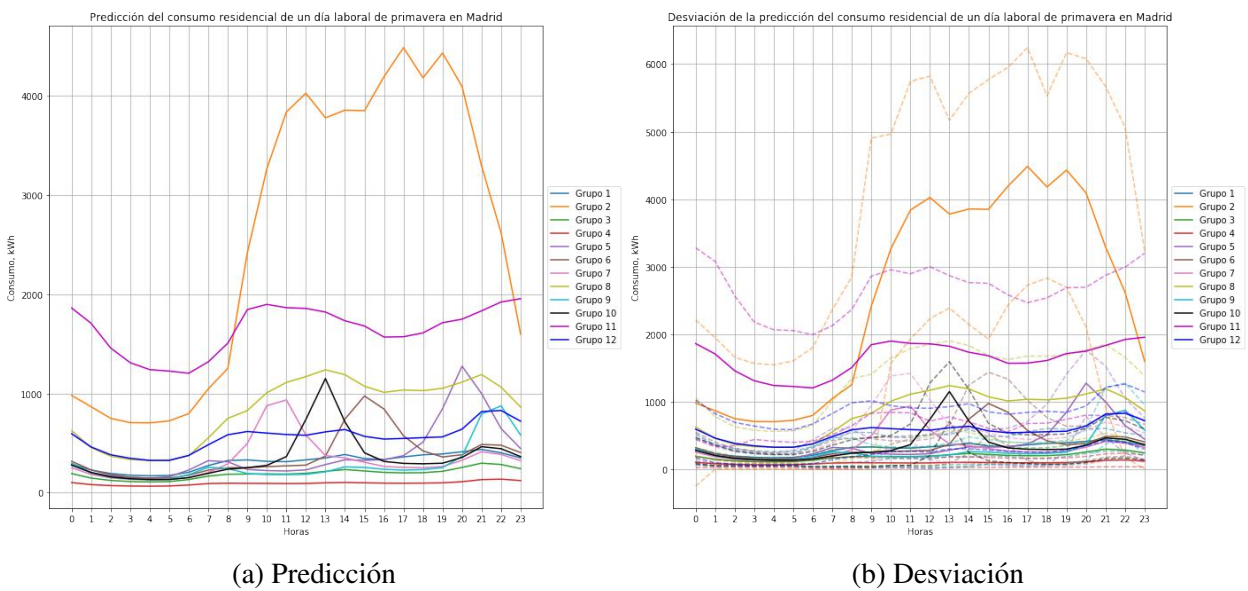


Figura 5.38.: Consumo residencial de un día laboral de primavera en Madrid

El resto de grupos representan la tendencia esperada a lo largo del día para un consumo residencial. El Grupo 3 y el Grupo 4 son los que representan un porcentaje mayor de este tipo de consumo, 61.76 % en conjunto, y además, son los que presentan los valores de consumo menores.

Uno de los principales objetivos de las distribuidoras eléctricas es conseguir aplanar y disminuir, lo máximo posible el gráfico, del consumo a lo largo del día, lo que les permitiría aumentar la rentabilidad de las plantas eléctricas. En esta situación, las distribuidoras eléctricas podrán lanzar, principalmente, dos campañas. Una con el objetivo de disminuir el consumo generado por los clientes pertenecientes al Grupo 2, lo cual se puede conseguir mediante el aumento del precio al sobrepasar cierto límite de consumo. Por otro lado, se puede lanzar una campaña con el objetivo repartir el consumo de aquellos grupos que presentan un pico pronunciado, a lo largo del día.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	13.75 %	Grupo 5	2.94 %	Grupo 9	6.67 %
Grupo 2	0.05 %	Grupo 6	3.10 %	Grupo 10	3.49 %
Grupo 3	32.66 %	Grupo 7	3.32 %	Grupo 11	0.25 %
Grupo 4	29.01 %	Grupo 8	1.14 %	Grupo 12	3.63 %

Tabla 5.20.: Predicción del porcentaje de datos asociados a cada grupo

Con los grupos formados, es necesario evaluar la fiabilidad de los mismo, por ese motivo se han comparado los porcentajes de representación del consumo de cada la tabla de predicción, Tabla 5.20, y de la tabla de validación, Tabla 5.21. Sorprendentemente, de media, el porcentaje de pertenencia a cada grupo, tan solo varía en un 0.03 %, lo cual confirma la fiabilidad del modelo.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	13.73 %	Grupo 5	2.94 %	Grupo 9	6.69 %
Grupo 2	0.05 %	Grupo 6	3.14 %	Grupo 10	3.43 %
Grupo 3	32.78 %	Grupo 7	3.32 %	Grupo 11	0.26 %
Grupo 4	29.00 %	Grupo 8	1.11 %	Grupo 12	3.56 %

Tabla 5.21.: Validación del porcentaje de datos asociados a cada grupo

Consumo residencial un sábado

Con el fin de obtener un conjunto de grupos que definan los patrones presentes en el consumo residencial de un sábado de primavera en la provincia de Madrid, es necesario obtener el número de clústers a emplear en la implementación del algoritmo K-Means. Una vez más, el “codo” de la función, Figura 5.39, no se puede apreciar de forma clara, por lo que se recurre al cálculo del porcentaje de disminución de la suma de varianzas para hallar el número de clústers que se va a utilizar. Al aumentar el número de clústers de once a doce, se consigue reducir la suma de las varianzas en un 3.51 %, sin embargo al seguir aumentando el número de clústers, esta cifra baja al 2.72 %, por lo que finalmente se ha decidido emplear doce clústers. Una vez escogido el número de clústers a emplear, ya se puede implementar el algoritmo K-Means, que ha generado los grupos representados en la Figura 5.40.

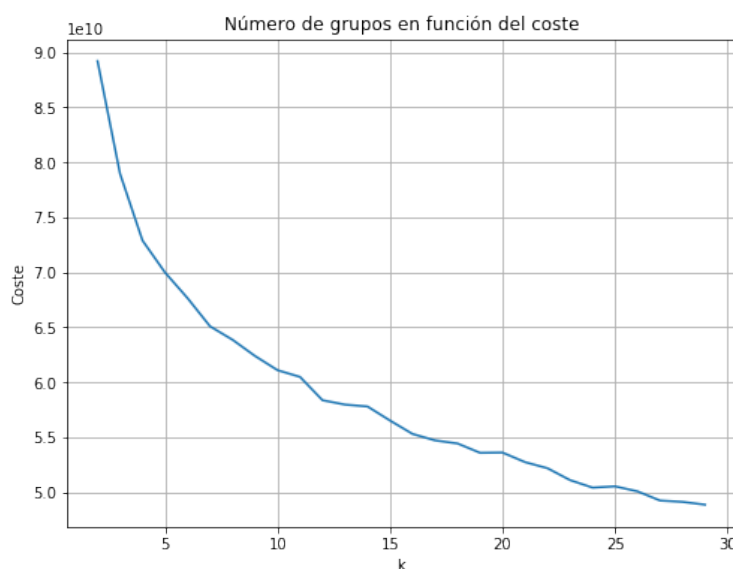


Figura 5.39.: Número de grupos en función de la suma de las distancias a los centroides

De entre los grupos generados, el primero que salta a la vista es el Grupo 12, ya presenta los valores de consumo más elevados. Además, llama la atención el hecho de que el pico de consumo correspondiente a la hora de cenar es menor que el de la hora de la comida, observación que se puede extrapolar al Grupo 10, que presenta una tendencia de consumo parecida pero tomado valores más bajos. Sin embargo, esta tendencia no es muy común, ya que solamente el 0.72 % del consumo residencial de un sábado de primavera en la provincia de Madrid se ve representado por el Grupo 10 o el 12.

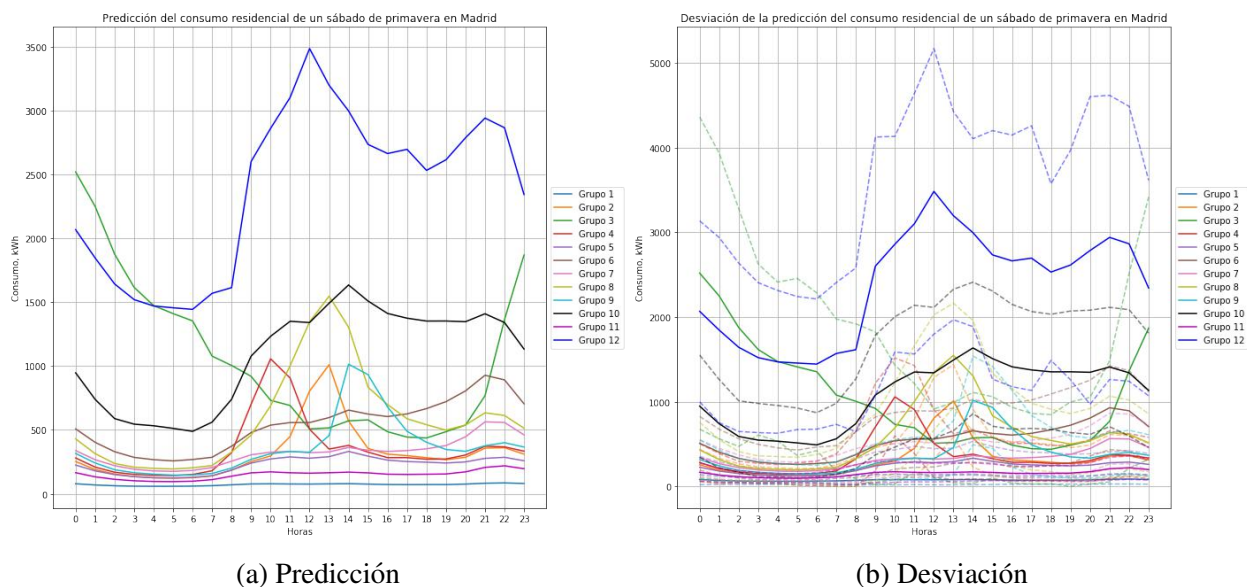


Figura 5.40.: Consumo residencial de un sábado de primavera en Madrid

También llama la atención el patrón seguido por el Grupo 3, cuyos picos de consumo se dan durante las horas de sueño, en vez de durante el día. Este patrón ya había sido observado durante los meses de invierno. Este comportamiento está relacionado con los hábitos de consumo de aquellas personas que trabajan durante por la noche. Tendencia que solamente es seguida por el 0.18 %, Tabla 5.22, del consumo residencial que se está estudiando.

Como en el caso del consumo residencial de un día laboral de primavera en Madrid, ciertos grupos solamente presentan un pico de consumo a lo largo del día, en este caso dichos grupos son el Grupo 2, 4, 8 y 9. Todos los picos se concentran en torno a la hora de comer, siendo el Grupo 8 el que presenta los valores de consumo más elevados. Este tipo de comportamiento se asocia son personas que solamente comen en sus viviendas, lo cual es muy habitual durante los fines de semana.

El resto de los grupos siguen con el esquema esperado del consumo residencial a lo largo de un día, resolviendo las diferencias en los niveles del consumo. Una vez más, los dos grupos con el nivel de consumo menor (Grupo 1 y 11), representan el mayor porcentaje del consumo, Tabla 5.22.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	21.01 %	Grupo 5	20.41 %	Grupo 9	4.63 %
Grupo 2	4.67 %	Grupo 6	4.26 %	Grupo 10	0.64 %
Grupo 3	0.18 %	Grupo 7	12.05 %	Grupo 11	26.97 %
Grupo 4	3.42 %	Grupo 8	1.68 %	Grupo 12	0.08 %

Tabla 5.22.: Predicción del porcentaje de datos asociados a cada grupo

En este caso, se pueden lanzar varias campañas publicitarias con el objetivo de aplanar los gráficos de consumo. En primer lugar, se puede lanzar una campaña centrada en el Grupo 12, con el objetivo de reducir sus niveles de consumo. Además, se puede lanzar otra campaña que tenga como fin desplazar los picos de consumo de las horas de sueño a la horas de luz del Grupo 3. Por último, se puede lanzar una campaña centrada en la eliminación de los picos de consumo puntuales de los grupos 2,4,8 y 9 y así evitar sobrecargas de demanda puntuales.

Teniendo el modelo establecido es necesario evaluarlo el 25 % de los datos destinados a la validación. Para ello, se han comparado los porcentajes de representación predichos de cada grupo con los obtenidos en la fase de validación. En este caso, la variación media por grupo es de 0.07 %, por lo que el modelo construido es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	20.86 %	Grupo 5	20.42 %	Grupo 9	4.92 %
Grupo 2	4.59 %	Grupo 6	4.27 %	Grupo 10	0.65 %
Grupo 3	0.16 %	Grupo 7	11.84 %	Grupo 11	26.99 %
Grupo 4	3.47 %	Grupo 8	1.73 %	Grupo 12	0.08 %

Tabla 5.23.: Validación del porcentaje de datos asociados a cada grupo

Consumo residencial un domingo

Teniendo como objetivo generar patrones que representen las distintas tendencias del consumo residencial un domingo de primavera en la provincia de Madrid, se debe obtener en número de clústers a emplear. Una vez más, el “codo” de la función, Figura 5.41, no se puede discernir a simple vista, por lo que es necesario recurrir al cálculo del porcentaje de disminución de la suma de varianzas a medida que se aumenta el número de clústers. En esta casuística, al aumentar el número de clústers de catorce a quince, se consigue reducir la suma de varianzas en un 1.32 %; por el contrario a medida que se aumenta el número de clústers, este valor se reduce por lo que se ha decidido emplear quince clústers.

Una vez decidido el número de clústers a emplear, se ha podido implementar el algoritmo K-Means, que ha generado los grupos mostrados en la Figura 5.42, junto con la desviación presente en cada grupo. Cabe mencionar que a pesar de haber decidido emplear quince clústers, en el gráfico solamente aparecen catorce grupos. Esto se debe a que el Grupo 12 solamente estaba formado por

un conjunto de datos, por lo que no aportaba información relevante al estudio y se ha eliminado del gráfico.

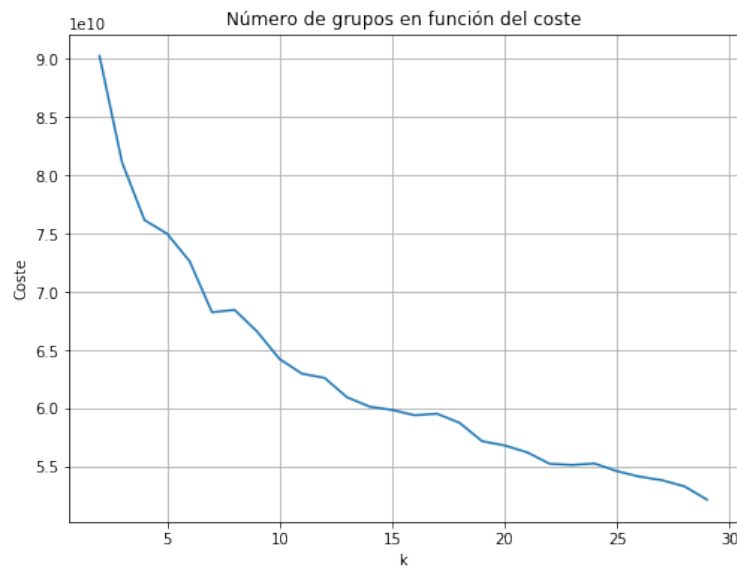


Figura 5.41.: Número de grupos en función de la suma de las distancias a los centroides

De los grupos obtenidos, los dos grupos que más se asemejan al patrón de consumo residencial obtenido en la sección 5.3.2, son el Grupo 7 y el Grupo 5. La principal diferencia entre ellos reside en el hecho de que el consumo del Grupo 7 es mayor, es más, es el grupo con mayores niveles de consumo, así como la mayor desviación. Además, llama la atención el hecho de que estos grupos, conjuntamente, tan solo representan el 0.70% del consumo residencial total de un domingo de primavera en Madrid, Tabla 5.24.

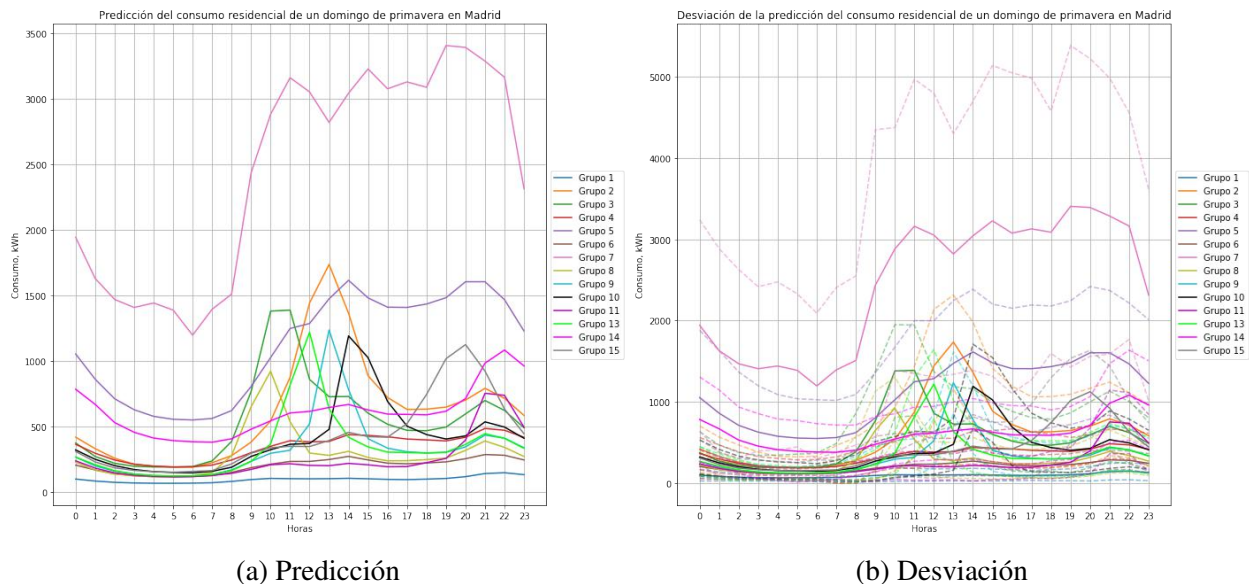


Figura 5.42.: Consumo residencial de un domingo de primavera en Madrid

Como en el caso anterior, hay un conjunto de grupos cuyo consumo a lo largo del día solamente presenta un pico de consumo, normalmente a la hora de la comida. Este comportamiento indica que los clientes pertenecientes a dicho grupos o no cenan o no comen en sus viviendas, lo cual es habitual durante los fines de semana.

El resto de los grupos presentan un patrón similar al encontrado en la sección 5.3.2 para el consumo residencial a lo largo de un día, pero con niveles de consumo menores. A este conjunto de grupos pertenecen los grupos 1 y 6, los dos con el consumo menor, que una vez más, son los grupos que abarcan un mayor porcentaje del consumo residencial de un domingo de primavera en la provincia de Madrid, Tabla 5.24.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	31.60 %	Grupo 6	29.72 %	Grupo 11	6.08 %
Grupo 2	1.53 %	Grupo 7	0.07 %	Grupo 12	0.001 %
Grupo 3	1.46 %	Grupo 8	3.27 %	Grupo 13	3.33 %
Grupo 4	10.56 %	Grupo 9	3.46 %	Grupo 14	2.28 %
Grupo 5	0.63 %	Grupo 10	3.48 %	Grupo 15	2.50 %

Tabla 5.24.: Predicción del porcentaje de datos asociados a cada grupo

Como en el caso anterior, con el objetivo de aplanar los niveles de consumo a lo largo del día se podría lanzar dos campañas. La finalidad de una de ellas estaría enfocada en reducir el nivel de los picos de aquellos grupos que presentan picos de consumo muy pronunciados, una forma de conseguirlo es aumentar el precio de la energía en esas franjas horarias. Por otro lado, una campaña tendría como objetivo reducir el nivel de consumo de los clientes pertenecientes al Grupo 12, para poder acercar sus valores de consumo a los del resto. Una forma de conseguir dicho objetivo puede estar basada en el aumento del precio de la energía tras sobrepasar ciertos niveles de consumo.

Con el modelo predictivo establecido, es necesario evaluar su fiabilidad. Para ello se comparan los porcentajes de pertenencia de cada grupo obtenidos en la fase de predicción, Tabla 5.24, con los obtenidos en la fase de validación, Tabla 5.25, utilizando el 25 % restante de los datos. El resultado de esta comparación es una variación media de cada grupo es de 0.11 %, lo que es asumible, y por ello el modelo es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	31.92 %	Grupo 6	29.68 %	Grupo 11	5.80 %
Grupo 2	1.46 %	Grupo 7	0.08 %	Grupo 12	0.00 %
Grupo 3	1.47 %	Grupo 8	3.41 %	Grupo 13	3.28 %
Grupo 4	10.79 %	Grupo 9	3.35 %	Grupo 14	2.36 %
Grupo 5	0.55 %	Grupo 10	3.47 %	Grupo 15	2.37 %

Tabla 5.25.: Validación del porcentaje de datos asociados a cada grupo

Consumo industrial un día laboral

En este caso, el “codo” de la función de coste, Figura 5.43, es más apreciable que en casos anteriores, pero aún así se ha recurrido al cálculo del porcentaje de disminución de la varianza en función del número de clústers para averiguar el número de clústers a emplear. Además, en este caso es especialmente importante debido al elevado orden de la suma de varianzas. Al aumentar el número de clústers de catorce a quince, la suma de varianzas disminuye en un 3.20 %, sin embargo, a medida que se aumenta el número de clústers, este valor disminuye, por lo que se ha decidido emplear quince clústers en este apartado.

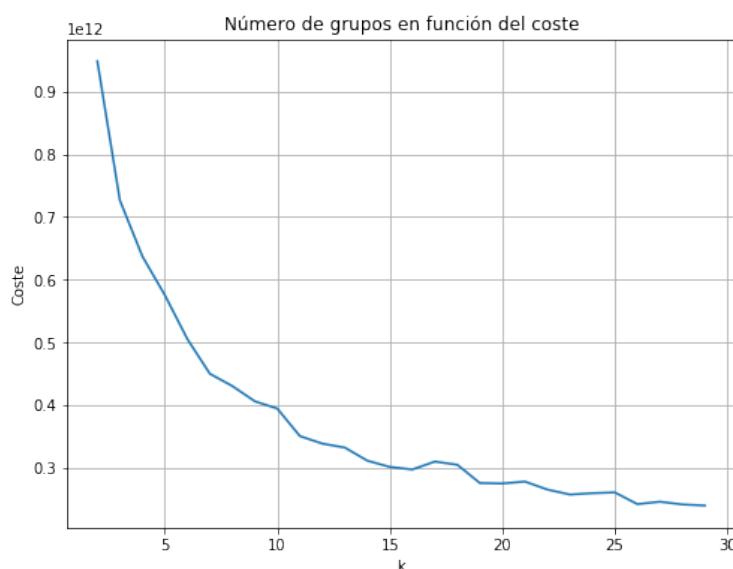


Figura 5.43.: Número de grupos en función de la suma de las distancias a los centroides

Con el número de clústers elegido, se ha podido implementar el algoritmo K-Means. Los resultados del algoritmo se muestra en la Figura 5.44, junto con la desviación de los grupos.

A primera vista, destaca el Grupo 15 debido a la peculiar tendencia que sigue a lo largo del día. El gran pico de consumo que se da a las 22:00, en conjunto con la depresión en el consumo desde las 7:00 hasta las 12:00 hace pensar que este tipo de patrón puede ser seguido por un restaurante. Inicia su actividad a las 12:00 ya que necesita empezar a preparar las comidas. A lo largo de la tarde sufre una pequeña depresión, espacio de tiempo entre que los últimos clientes del mediodía terminan sus platos y los primeros clientes del turno de noche empiezan a aparecer. A las 22:00 se encuentra el pico de consumo de la hora punta del servicio de cenas, las cuales se pueden alargar hasta entrada la madrugada, motivo por el cual el consumo no empieza a descender hasta las 5:00. Este patrón solamente representa el 0.12 %, Tabla 5.26, del consumo industrial de un día laboral de primavera en Madrid, lo cual no es de extrañar, teniendo en cuenta que durante la semana el índice de personas que cenar fueran de sus viviendas es bajo.

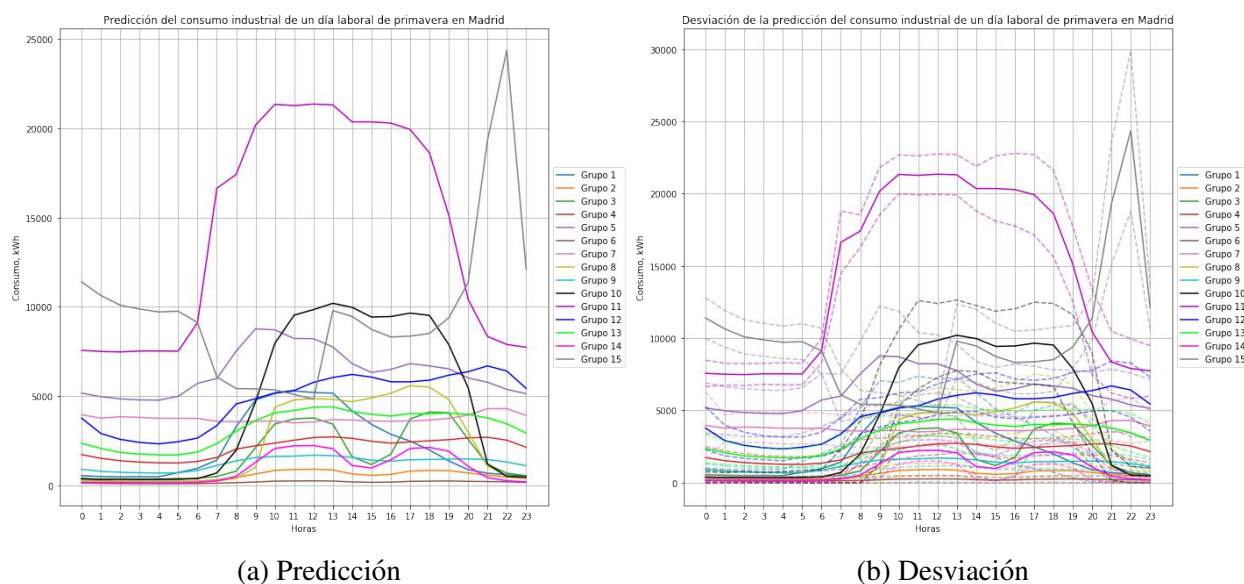


Figura 5.44.: Consumo industrial de un día laboral de primavera en Madrid

Además, cabe resaltar el hecho de que el Grupo 11, 10 y 8, presentan una forma poco común en el consumo industrial, siendo la forma habitual la encontrada en la sección 5.3.2. Normalmente, el consumo industrial a lo largo del día se caracteriza por la presencia de dos funciones cóncavas prácticamente idénticas, una en horario laboral antes de comer y otra en en horario laboral de tarde. Sin embargo estos dos grupos presentan un consumo relativamente constante a lo largo de la jornada laboral, lo que coincide con los hábitos de consumos de empresas que necesitan proveer servicio a lo largo de toda la jornada laboral. El conjunto de empresas que siguen este patrón de consumo, independientemente de los niveles de consumo, representa un 1.52% del consumo industrial de un día laboral de primavera en la provincia de Madrid.

También llama la atención la tendencia seguida por el Grupo 5 y el Grupo 1, que presenta una elevación, no muy pronunciada, desde las 7:00 hasta las 15:00, aproximadamente. Lo que se corresponden con actividades industriales en las que las horas de jornada laboral no impactan de forma relevante el consumo general de las instalaciones. Este tipo de actividades solamente representan el 2.32% del consumo empleado en este apartado.

El conjunto de grupo formado por el Grupo 3, 2 y 14 presentan la tendencia de consumo más similar a la encontrada en la sección 5.3.2. Motivo por el cual no extraña que en conjunto representen el 39.93% del consumo industrial de un día laboral de primavera en la provincia de Madrid.

Los grupos 12, 13 y 4 presenta un patrón a lo largo del día que recuerda al consumo residencial encontrado en la sección 5.3. Sufre una pequeña depresión durante las horas de sueño y a medida que avanza el día, el consumo asciende de forma prácticamente constante, a excepción del pico presenciado durante las horas del mediodía. Este comportamiento puede corresponder con el de locales que ofrecen servicios durante todo el día y tienen una mayor afluencia durante las horas de las comidas, como puede ser el caso de cafeterías o bares.

Con el objetivo en mente de aplanar el consumo eléctrico a lo largo del día para que así las centrales eléctricas optimicen sus instalaciones, se pueden lanzar, principalmente, dos campañas publicitarias. Una de ellas debería tener como objetivo reducir los niveles de consumo generados por los clientes pertenecientes al Grupo 11, lo cual se podría conseguir mediante el incremento del precio de la energía pasado cierto límite. Por otro lado, una segunda campaña podría tener como objetivo reducir el pico de consumo encontrado en torno a las 22:00 en el Grupo 15, para conseguirlo se podría emplear una técnica similar a la anteriormente descrita.

El resto de los grupos, es decir los grupos 6, 7 y 9, presentan un consumo prácticamente constante a lo largo del día. En especial destaca el Grupo 6 debido a que sus niveles de consumo son los más bajos. Lo que una vez se corresponde con el grupo que, individualmente, abarca un porcentaje mayor.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	1.64 %	Grupo 6	35.56 %	Grupo 11	0.11 %
Grupo 2	27.16 %	Grupo 7	1.77 %	Grupo 12	1.03 %
Grupo 3	3.13 %	Grupo 8	0.98 %	Grupo 13	2.81 %
Grupo 4	5.60 %	Grupo 9	9.40 %	Grupo 14	9.64 %
Grupo 5	0.63 %	Grupo 10	0.44 %	Grupo 15	0.12 %

Tabla 5.26.: Predicción del porcentaje de datos asociados a cada grupo

Con los distintos patrones encontrados, es hora de evaluar la fiabilidad del modelo comparando los porcentajes de permanencia a cada grupo con los porcentaje obtenidos utilizando el 25 % de los datos restantes, Tabla 5.26 y Tabla 5.27. Para este caso la variación media es igual a 0.19 % por lo que el modelo encontrado es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	1.59 %	Grupo 6	35.85 %	Grupo 11	0.08 %
Grupo 2	27.95 %	Grupo 7	1.67 %	Grupo 12	0.98 %
Grupo 3	3.29 %	Grupo 8	1.02 %	Grupo 13	2.71 %
Grupo 4	5.01 %	Grupo 9	8.88 %	Grupo 14	9.66 %
Grupo 5	0.65 %	Grupo 10	0.53 %	Grupo 15	0.13 %

Tabla 5.27.: Validación del porcentaje de datos asociados a cada grupo

Consumo industrial del fin de semana

En este caso se estudian los patrones encontrados en el consumo industrial de un sábado y domingo de primavera en la provincia de Madrid, ya que los patrones encontrados son muy similares. Para ello, en primer lugar, es necesario encontrar el número de clústers a emplear, motivo

por el cual se ha creado la función de costes encontrada en la Figura 5.45. Una vez más, el “codo” de la función no es apreciable a primera vista, en ninguna de las dos funciones, por lo que se ha recurrido al cálculo de la disminución porcentual de la suma de varianzas en función del número de clústers. Al aumentar el número de clústers de catorce a quince, la suma de varianzas sufre una disminución mayor que si se sigue aumentando el número de clústers, motivo por el cual se ha decidido emplear quince clústers en ambos casos.

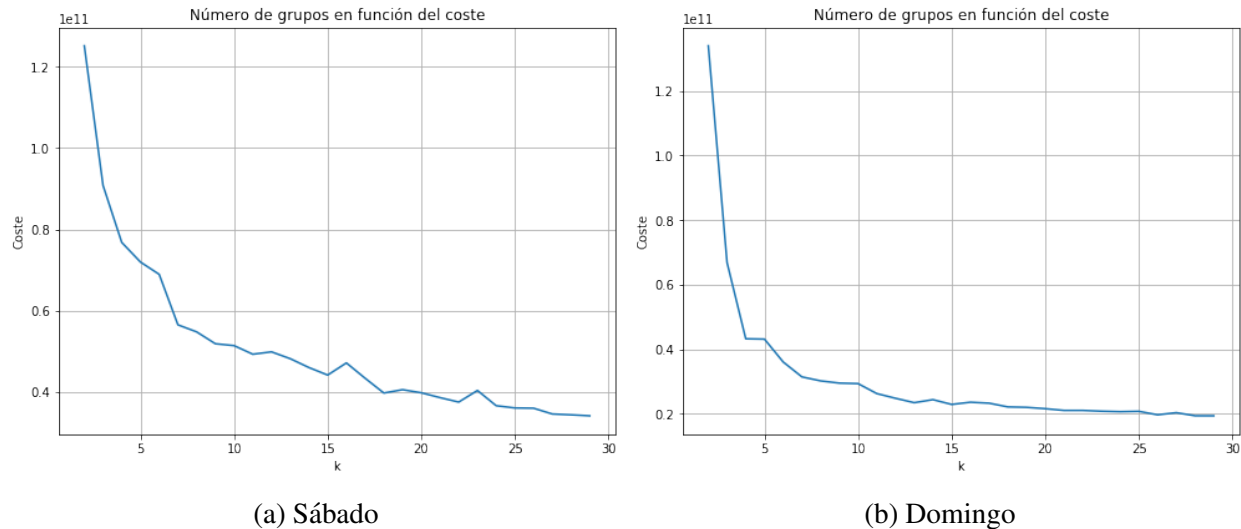


Figura 5.45.: Número de grupos en función de la suma de las distancias a los centroides

Con el número de clústers a emplear determinado, se puede implementar el algoritmo K-Means con el objetivo de encontrar los distintos patrones que reflejan el comportamiento de varios grupos de consumidores a lo largo de un fin de semana de primavera en la provincia de Madrid, Figura 5.46 y Figura 5.47. A pesar de que el número de clústers elegido es quince, en la imagen que representa los grupos del domingo, Figura 5.47, solamente aparecen catorce grupos, esto se debe a la presencia de un grupo, Grupo 5, que solamente representaba un conjunto de datos. Debido a la poca fiabilidad de dicho grupos, se ha decidido excluirlo del análisis.

Como en la caso del consumo industrial de un día laboral de primavera en la provincia de Madrid, el primer grupo que llama la atención, es el Grupo 2 en el caso del gráfico de los sábados y el Grupo 12, los domingos. Este grupo destaca por el hecho de que presenta un gran pico de consumo en torno a las 22:00 y una pequeña depresión entre las 7:00 y las 12:00. Como se comentaba en la casuística anterior, este comportamiento del consumo se puede corresponder con el de un restaurante, que sufre la mayor afluencia de clientes en el turno de noche. Además, cabe destacar que en para ambos días, el porcentaje del consumo representado por este grupo es ligeramente superior al 10 %, Tabla 5.28 y Tabla 5.30.

También cabe destacar la presencia de ciertos grupos cuya distribución del consumo no es la esperada, siendo este de tipo industrial. Estos grupos son el Grupo 8, 9 y 14 en el gráfico del sábado y los grupos 8 y 10 en el gráfico del domingo. La característica a resaltar de estos grupos es el hecho de que presentan un consumo relativamente constante durante un periodo de tiempo

determinado, que varía dependiendo del grupo. Por el contrario, en la sección 5.3.2, se encontraba una tendencia del consumo a lo largo del día, que presentaba dos picos, uno en horario de mañana y otro en horario de tarde.

Además, destacan los grupos 11 y 5, en el gráfico de los sábados y el Grupo 7 en el caso de los domingos. El motivo por el que destacan es el hecho de que presentan dos picos, como es de esperar en el consumo industrial, pero el segundo pico del día, que se presenta a distintas horas, es mayor que el primero. Lo que indica una mayor actividad en la segunda parte de la jornada laboral.

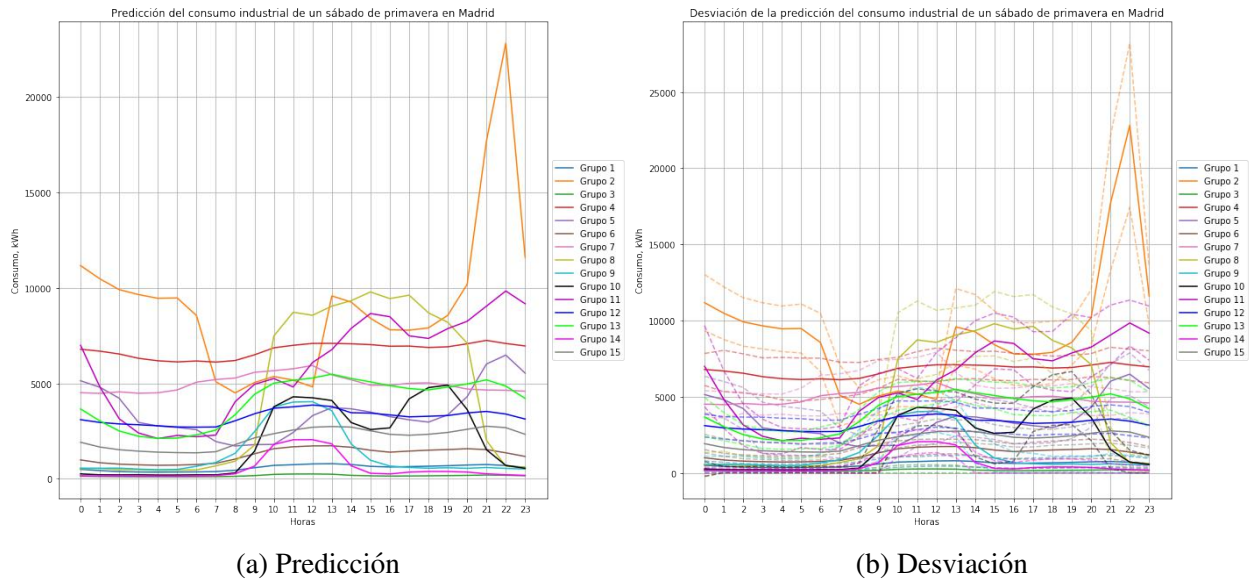


Figura 5.46.: Consumo industrial de un sábado de primavera en Madrid

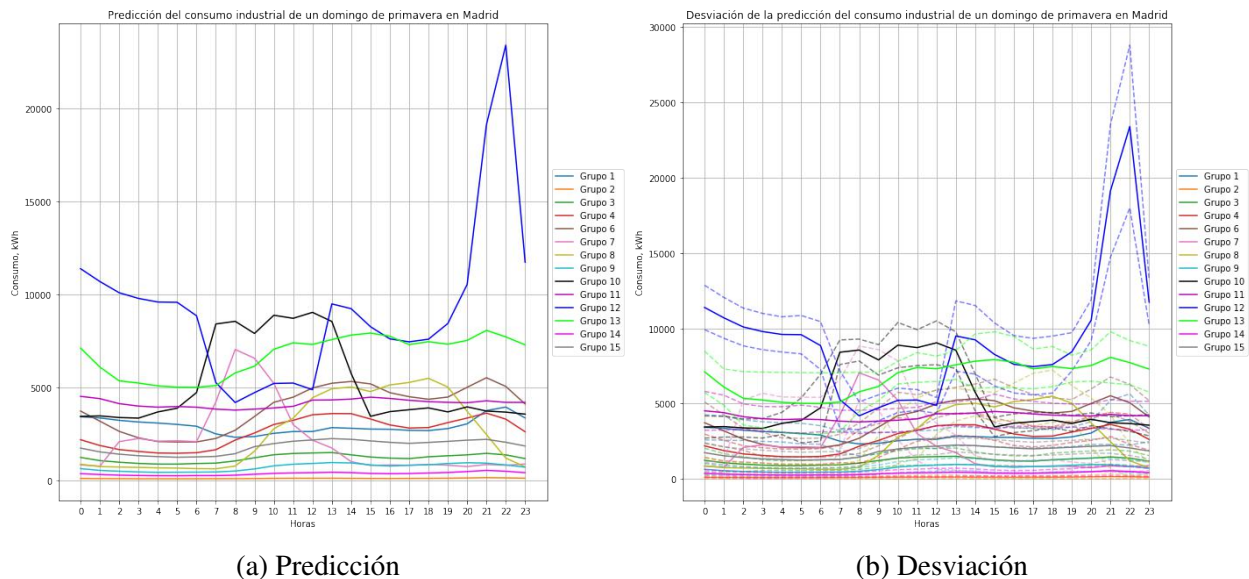


Figura 5.47.: Consumo industrial de un domingo de primavera en Madrid

Con diferencia, el grupo que presenta una tendencia de consumo a lo largo del día más parecida a la encontrada en la sección 5.3.2, es la seguida por el Grupo 10, en el gráfico de los sábados, que sorprendentemente, tan solo representa el 1.23 % del consumo industrial de un sábado de primavera en la provincia de Madrid.

El resto de los grupos presenta una tendencia similar a la esperada en un consumo industrial, pero cuyos picos no sobresalen tanto como en otros casos. Una vez más, los grupos con los niveles de consumo más bajos, representan el mayor porcentaje del consumo. En el caso de los sábados es el Grupo 3, que representa el 48.74 %, Tabla 5.28, del consumo y en el caso de los domingos es el Grupo 2, que representa el 45.74 %.

Con la finalidad de estabilizar los niveles de consumo a lo largo del día, para que así las centrales eléctricas optimicen sus instalaciones, se podría lanzar, principalmente, una oferta. El objetivo principal de esta oferta sería reducir el gran pico de consumo que se da en torno de las 22:00, en ambos días. Un método para reducir este consumo sería aumentar el precio de la energía en esa franja horaria o aumentarlo una vez sobrepasado cierto nivel de consumo.

Estando el modelo construido, es necesario evaluar la fiabilidad del mismo. Para medir la misma, se comparan los porcentajes de pertenencia obtenidos en la fase de predicción, en la que se usa el 75 % de los datos, con los porcentajes obtenidos en la fase de validación, en la que se usan el 25 % restante. En este caso ambos modelos son válidos, ya que, de media, la asignación de porcentajes varía un 0.38 % los sábados y un 0.18 % los domingos, lo que es perfectamente asumible. Esta pequeña diferencia entre el sábado y el domingo se debe a que el gráfico de coste del domingo decae más rápidamente que el del sábado.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	20.70 %	Grupo 6	9.17 %	Grupo 11	0.28 %
Grupo 2	0.11 %	Grupo 7	0.60 %	Grupo 12	2.16 %
Grupo 3	48.74 %	Grupo 8	0.35 %	Grupo 13	1.28 %
Grupo 4	0.42 %	Grupo 9	2.65 %	Grupo 14	6.66 %
Grupo 5	0.52 %	Grupo 10	1.23 %	Grupo 15	5.12 %

Tabla 5.28.: Predicción del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	21.78 %	Grupo 6	8.46 %	Grupo 11	0.41 %
Grupo 2	0.17 %	Grupo 7	0.46 %	Grupo 12	3.19 %
Grupo 3	47.53 %	Grupo 8	0.29 %	Grupo 13	1.04 %
Grupo 4	0.41 %	Grupo 9	3.03 %	Grupo 14	6.39 %
Grupo 5	0.54 %	Grupo 10	1.37 %	Grupo 15	4.94 %

Tabla 5.29.: Validación del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	1.39 %	Grupo 6	1.38 %	Grupo 11	1.05 %
Grupo 2	45.74 %	Grupo 7	0.10 %	Grupo 12	0.12 %
Grupo 3	6.75 %	Grupo 8	0.27 %	Grupo 13	0.48 %
Grupo 4	2.07 %	Grupo 9	11.56 %	Grupo 14	24.70 %
Grupo 5	0.01 %	Grupo 10	0.12 %	Grupo 15	4.25 %

Tabla 5.30.: Predicción del porcentaje de datos asociados a cada grupo, Domingo

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	1.35 %	Grupo 6	1.12 %	Grupo 11	1.35 %
Grupo 2	45.18 %	Grupo 7	0.23 %	Grupo 12	0.14 %
Grupo 3	6.80 %	Grupo 8	0.37 %	Grupo 13	0.98 %
Grupo 4	2.28 %	Grupo 9	11.41 %	Grupo 14	24.59 %
Grupo 5	0.00 %	Grupo 10	0.14 %	Grupo 15	4.05 %

Tabla 5.31.: Validación del porcentaje de datos asociados a cada grupo, Domingo

5.4.1.3. Perfiles del consumo de un día de verano en la provincia de Madrid

El objetivo de esta sección es encontrar los distintos patrones que representan el consumo residencial e industrial de un día de verano en la provincia de Madrid. Para ello se va a seguir la misma dinámica empleada en las secciones anteriores, en primer lugar se obtiene el número de clústers a emplear, para después obtener los distintos patrones de consumo.

Consumo residencial

Sorprendentemente, en este caso en particular, el número de clústers elegido para todos los días de la semana es el mismo. Este fenómeno ha sido posible gracias a la poca variación que presenta el consumo residencial en cuanto a valores de consumo y tendencia a lo largo de la semana, lo que ha dado lugar a una gran similitud en las pendientes de las funciones de coste, Figura 5.48, ya que los valores que toma la función de coste de los días laborales son mayores que en el caso de los días del fin de semana. Con el objetivo de encontrar el número de clústers a emplear se ha recurrido al cálculo del porcentaje de disminución de la suma de varianzas con cada aumento del número de clústers. Tras realizar esa operación para cada uno de los tres casos, se ha llegado a la conclusión de que el número de clústers a emplear es doce.

Con el número de clústers escogido, se puede implementar el algoritmo K-Means con la finalidad de obtener los distintos grupos de clientes que representan el consumo residencial en verano en la provincia de Madrid. A pesar de que el número de clústers elegido es doce, en los gráficos del sábado y del domingo, Figura 5.50 y Figura 5.51, tan solo se muestran once grupos. El motivo detrás diferencia reside en la presencia de dos grupos, uno en cada día, que solamente reflejaban

un conjunto de datos, por lo que no aportaban información fiable de cara a una predicción, razón por la que han sido eliminados del estudio.

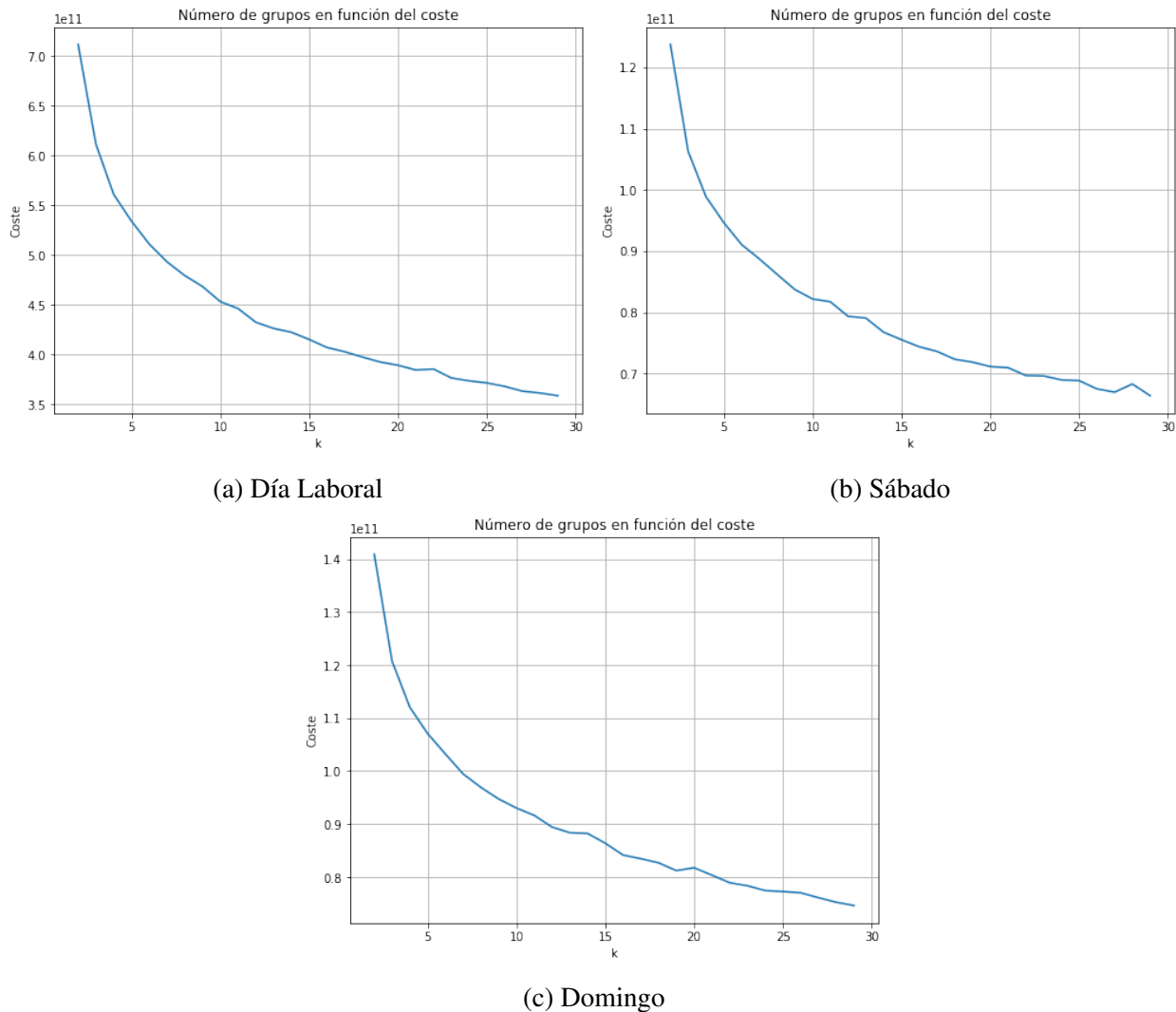


Figura 5.48.: Número de grupos en función de la suma de las distancias a los centroides

A primera vista, el grupo que más llama la atención es el Grupo 3, que presenta la misma tendencia en los tres casos. Destaca, principalmente, por poseer los valores del consumo más elevados, así como una mayor desviación. Mirando más fijamente los valores de las gráficas, se puede observar como los valores del consumo residencial durante la semana son superiores a los valores que toma el consumo durante el fin de semana. La forma de la tendencia es la esperada, ya que coincide con la forma encontrada para el consumo residencial en la sección 5.3.2, siendo la tendencia del gráfico del sábado un poco más dispar. Además, cabe destacar el hecho de que para todos los días de la semana, el Grupo 3, representa el menor porcentaje, Tabla ??, Tabla ?? y Tabla 5.34; de consumo residencial de un día de verano en la provincia de Madrid.

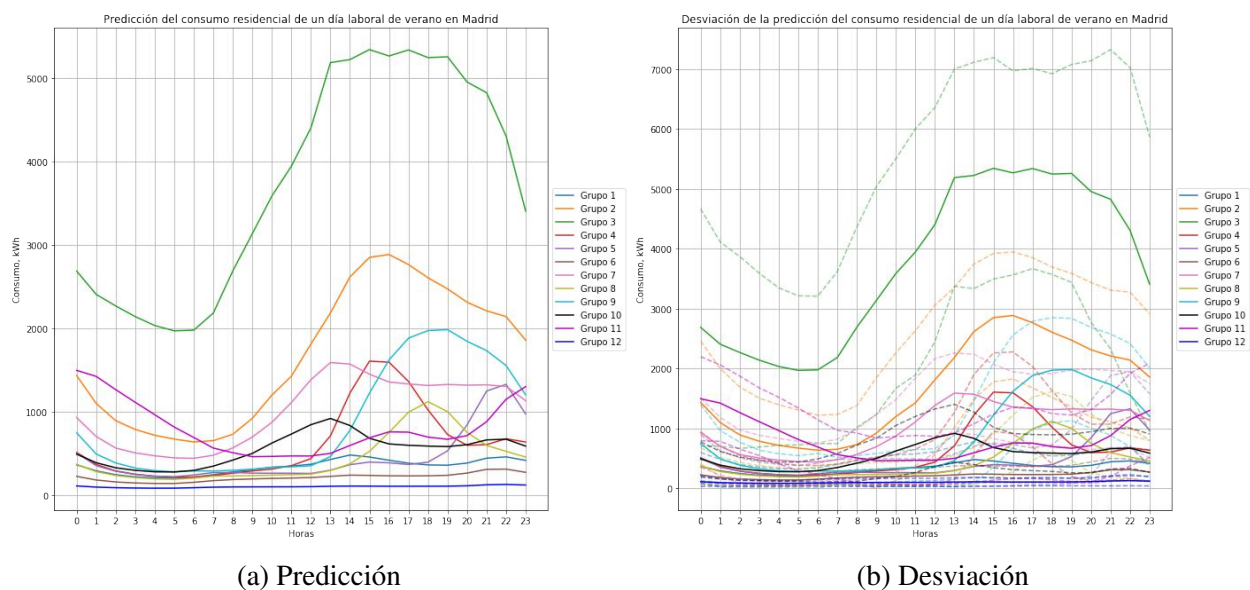


Figura 5.49.: Consumo residencial de un día laboral de verano en Madrid

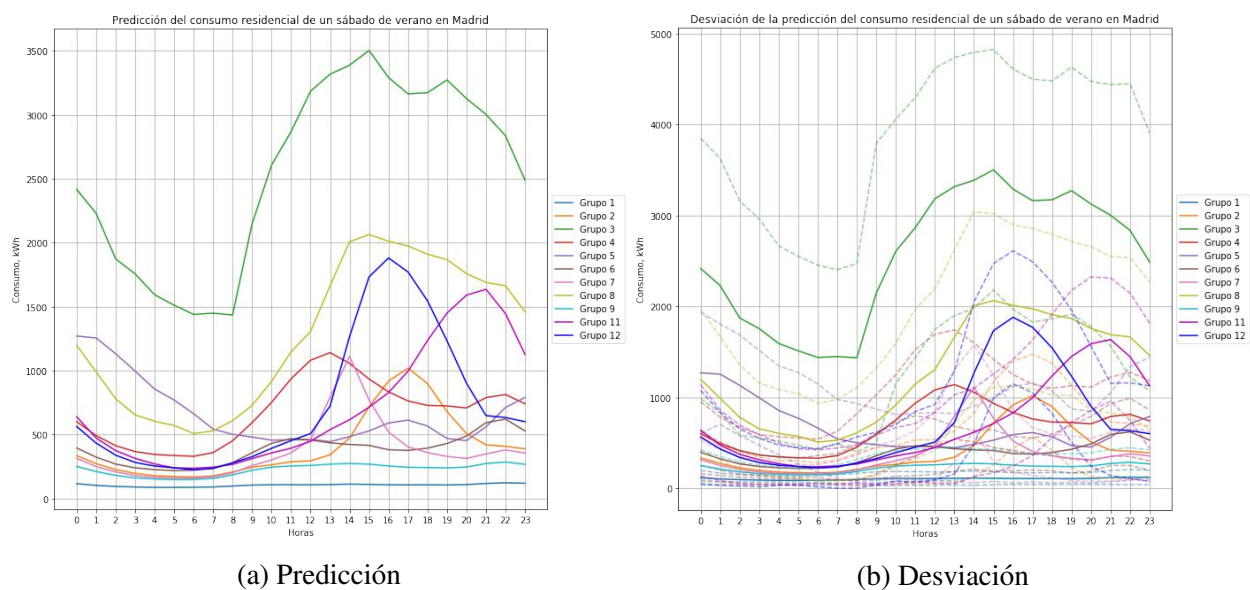


Figura 5.50.: Consumo residencial de un sábado de verano en Madrid

Siguiendo en esta línea, se encuentran los grupos 2, 9, 4, 5 y 8 en el gráfico de días laborales; los grupos 8, 12, 11, 4 y 2 en el gráfico de los sábados y los grupos 9, 1, 11, 8, 10 y 7 en el gráfico de los domingos. Estos grupos se caracterizan por presentar un máximo claro al día, aunque en cada grupo está presente a distintas horas de la tarde. Lo que más llama la atención acerca de estos grupos es el porcentaje del consumo tan pequeño que abarca cada uno de ellos, Tabla 5.32, Tabla 5.33 y Tabla 5.34, teniendo en cuenta que la tendencia que presentan era la esperada y hallada en la sección 5.3.2.

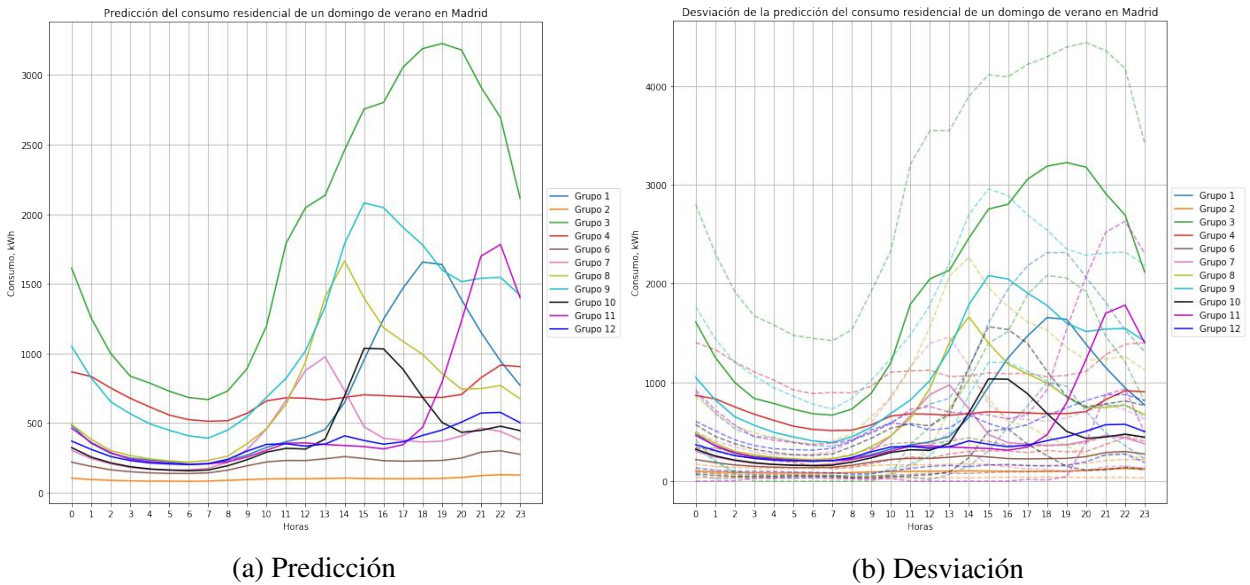


Figura 5.51.: Consumo residencial de un domingo de verano en Madrid

Con diferencias los grupos que presentan el comportamiento más extraño son el Grupo 11 en el caso de los días laborales, el Grupo 5 en el caso de los sábados y el Grupo 4 en el caso de los domingos. Estos patrones se caracterizan por sufrir una ligera caída durante las horas de luz, cuando se espera un aumento del consumo. Este tipo de comportamiento se puede corresponder con el de personas que trabajen en el turno de noche y por tanto no generan consumo en las franjas horarias habituales. Lo cual se respalda por el hecho de que cada uno de estos grupos representan, para su correspondiente tipo de día, en torno al 2 %, Tabla 5.32, Tabla 5.33 y Tabla 5.34, del consumo residencial en verano en la provincia de Madrid.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	16.70 %	Grupo 5	3.42 %	Grupo 9	1.30 %
Grupo 2	0.50 %	Grupo 6	31.70 %	Grupo 10	4.51 %
Grupo 3	0.05 %	Grupo 7	1.47 %	Grupo 11	1.30 %
Grupo 4	2.35 %	Grupo 8	3.49 %	Grupo 12	33.19 %

Tabla 5.32.: Predicción del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	40.82 %	Grupo 5	1.19 %	Grupo 9	31.92 %
Grupo 2	3.84 %	Grupo 6	10.75 %	Grupo 10	0.001 %
Grupo 3	0.09 %	Grupo 7	4.91 %	Grupo 11	1.40 %
Grupo 4	2.58 %	Grupo 8	0.82 %	Grupo 12	1.69 %

Tabla 5.33.: Predicción del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	1.98 %	Grupo 5	0.001 %	Grupo 9	1.37 %
Grupo 2	35.42 %	Grupo 6	31.78 %	Grupo 10	4.31 %
Grupo 3	0.32 %	Grupo 7	5.35 %	Grupo 11	1.49 %
Grupo 4	2.51 %	Grupo 8	2.13 %	Grupo 12	13.35 %

Tabla 5.34.: Predicción del porcentaje de datos asociados a cada grupo, Domingo

El resto de los grupos, presentan unos niveles de consumo demasiado bajos como para poder observar una tendencia clara en los gráficos mostrados. Aunque si que se puede intuir una subida durante las horas de luz, lo que nos inclina pensar que estos grupos también presentarán una tendencia similar a la encontrada en la sección 5.3.2.

Con el objetivo de aplanar el consumo a lo largo del día, para que así las centrales eléctricas hagan un uso más eficiente de sus instalaciones, se puede lanzar un tipo de campaña publicitaria. El objetivo de la campaña sería reducir los picos de consumo que se dan normalmente por las tardes. En función de los niveles de consumo se podría focalizar en una franjas horarios o en otras, mediante el incremento del precio de la energía en dichas franjas.

Con los modelos ya obtenidos, es necesario evaluar la fiabilidad de los mismos. Para ello se comparan los porcentajes de pertenencia a cada grupo obtenidos en la fase de predicción, Tabla 5.32, Tabla 5.33 y Tabla 5.34, con los obtenidos en la fase de validación, Tabla 5.35, Tabla 5.36 y Tabla 5.37. En este caso, la variación media de los grupos es de 0.08 % en los días laborales, de 0.10 % en los sábados y de 0.14 % en los domingos, lo que es perfectamente asumible, por lo que los modelos son válidos.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	16.62 %	Grupo 5	3.40 %	Grupo 9	1.34 %
Grupo 2	0.52 %	Grupo 6	31.89 %	Grupo 10	4.69 %
Grupo 3	0.05 %	Grupo 7	1.47 %	Grupo 11	1.34 %
Grupo 4	2.37 %	Grupo 8	3.44 %	Grupo 12	33.87 %

Tabla 5.35.: Validación del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	40.66 %	Grupo 5	1.23 %	Grupo 9	31.71 %
Grupo 2	3.66 %	Grupo 6	11.23 %	Grupo 10	0.00 %
Grupo 3	0.09 %	Grupo 7	4.86 %	Grupo 11	1.40 %
Grupo 4	2.64 %	Grupo 8	0.84 %	Grupo 12	1.69 %

Tabla 5.36.: Validación del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	1.90 %	Grupo 5	0.00 %	Grupo 9	1.37 %
Grupo 2	36.03 %	Grupo 6	31.39 %	Grupo 10	4.36 %
Grupo 3	0.28 %	Grupo 7	5.49 %	Grupo 11	1.50 %
Grupo 4	2.52 %	Grupo 8	1.99 %	Grupo 12	13.17 %

Tabla 5.37.: Validación del porcentaje de datos asociados a cada grupo, Domingo

Consumo industrial de un día laboral

Como en los casos anteriores, en primer lugar, es necesario determinar el número de clústers a emplear en la implementación del algoritmo K-Means. Para ello, se ha creado la función de coste de la Figura 5.52, que representa la suma total de las varianzas con respecto a los centroides, en función del número de clústers que se emplean. Como el “codo” de la función no es discernible a primera vista, se ha recurrido al cálculo de la disminución porcentual de la suma de varianzas como medida de referencia para determinar el número de clústers. Al aumentar el número de clústers de trece a catorce, la suma de varianzas sufre una disminución del 4.98 %, mientras que si se sigue aumentando el número de clústers esta cifra disminuye a un 0.5 %, motivo por el cual el número de clústers escogido es catorce.

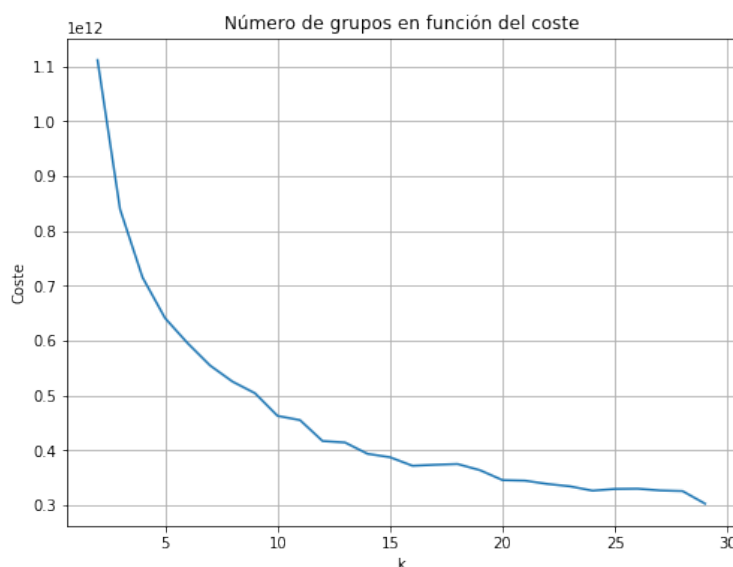


Figura 5.52.: Número de grupos en función de la suma de las distancias a los centroides

Con el número de clústers elegido, se ha podido implementar el algoritmo K-Means, cuyo objetivo es encontrar los patrones que reflejan el consumo industrial de un día laboral de verano en la provincia de Madrid. Los grupos obtenidos, junto con sus respectivas desviaciones típicas se encuentran representados en la Figura 5.53.

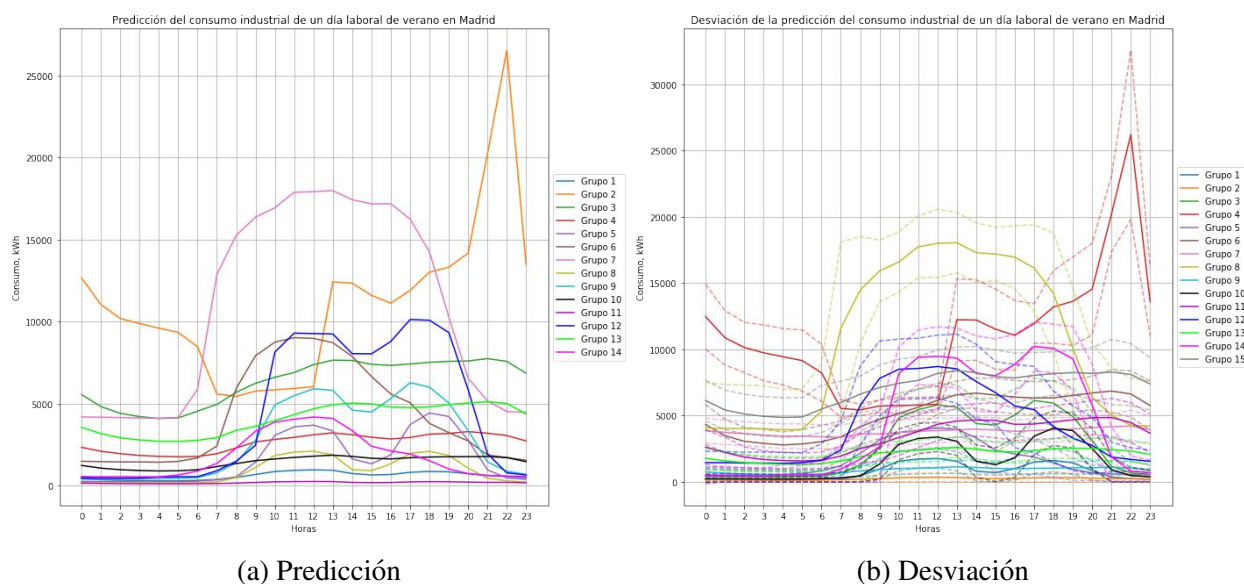


Figura 5.53.: Consumo industrial de un día laboral de verano en Madrid

A primera vista, el grupo que más destaca es el Grupo 7, debido a su forma totalmente cóncava, cuando el patrón de consumo industrial esperado tiende a ser muy similar, tanto en horario de mañana como de tarde, presentando dos picos simétricos de consumo, como se puede apreciar en la sección 5.3.2. Este tipo de comportamiento se asocia a empresas que mantienen una actividad estable a lo largo de toda su jornada laboral. Esta peculiar tendencia encontrada en el consumo industrial de un día laboral de verano en la provincia de Madrid, tan solo representa el 0.14 %, Tabla 5.38, del consumo total del mismo. Además, el Grupo 6 y el Grupo 14, solamente presentan un pico de consumo, aunque en estos casos el intervalo de elevación del consumo es menor.

También llama la atención la tendencia seguida por el Grupo 2. Curiosamente, en el consumo industrial de primavera, independientemente del tipo de día, de la provincia de Madrid, se encuentra un grupo con una tendencia prácticamente idéntica. Por el contrario, esta tendencia no se encuentra en el consumo industrial de los meses de invierno, por lo que se puede asociar a épocas del año más calurosas. Anteriormente, ya se había establecido la posible relación existente entre el consumo de un restaurante y la tendencia observada, debido a la depresión en el consumo desde las 7:00 hasta las 12:00, aproximadamente, el gran pico de consumo a las 22:00 y niveles de consumo similares en las franjas horarias restantes. Una vez más, este tipo de consumo representa un porcentaje ínfimo del consumo industrial, siendo en este caso igual a 0.14 %, Tabla 5.38.

Sin embargo, sí que existen un conjunto de grupos que presentan la tendencia esperada, tratando con consumo industrial. Es el caso de los grupos 12, 9, 5 y 8. Estos grupos, presentan los dos picos típicos del consumo industrial, uno por la mañana y otro por la tarde, además de la depresión del mediodía.

El resto de los grupos representados en la Figura 5.53, posee una tendencia mucho más plana a lo largo del día. Es muy probable que estos grupos presenten la forma típica del consumo industrial, ya que se aprecian unas pequeñas elevaciones del consumo en las franjas horarias que

se corresponden con los picos de consumo. Sin embargo, es difícil de discernir debido a la escala de la gráfica. Una vez más, el grupo con el nivel de consumo menor es el que representa una mayor proporción del consumo total, en este caso es el Grupo 11 que abarca el 37.86 %, Tabla 5.38, del consumo total.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	24.85 %	Grupo 6	0.53 %	Grupo 11	37.86 %
Grupo 2	0.14 %	Grupo 7	0.14 %	Grupo 12	0.60 %
Grupo 3	1.74 %	Grupo 8	9.68 %	Grupo 13	3.67 %
Grupo 4	4.75 %	Grupo 9	2.10 %	Grupo 14	2.57 %
Grupo 5	4.46 %	Grupo 10	6.93 %		

Tabla 5.38.: Predicción del porcentaje de datos asociados a cada grupo

Uno de los objetivos de este proyecto es proporcionar información útil para las distribuidoras de energía eléctrica para que puedan lanzar campañas publicitarias a determinados grupos de clientes. A las distribuidoras energéticas les conviene aplanar los niveles de consumo a lo largo del día para poder optimizar las instalaciones. En este caso en particular, con la finalidad de alcanzar este objetivo, se pueden lanzar dos campañas publicitarias. Una de ellas se centrará en los clientes pertenecientes al Grupo 2, para conseguir que se reduzca el gran pico de consumo que se sufre en torno a las 22:00. Por otro lado, una oferta se debería centrar en los clientes del Grupo 7, ya que posee niveles de consumo muy elevados a lo largo de toda la jornada laboral. En ambos casos, un método para conseguir reducir el consumo se puede basar en el incremento del precio de la energía en determinadas franjas horarias o al sobrepasar determinados niveles de consumo.

Con el modelo establecido, es necesario evaluar la fiabilidad del mismo. Para ello se comparan los porcentajes de representación de cada grupo obtenidos en la fase de predicción, Tabla 5.38, con los obtenidos en la fase de validación, Tabla 5.39. En este caso, la variación media entre los grupos es de 0.16 %, valor que es perfectamente asumible, por lo que el modelo calculado es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	24.55 %	Grupo 6	0.49 %	Grupo 11	38.48 %
Grupo 2	0.15 %	Grupo 7	0.13 %	Grupo 12	0.59 %
Grupo 3	1.57 %	Grupo 8	9.56 %	Grupo 13	3.85 %
Grupo 4	4.76 %	Grupo 9	2.12 %	Grupo 14	2.57 %
Grupo 5	3.97 %	Grupo 10	7.03 %		

Tabla 5.39.: Validación del porcentaje de datos asociados a cada grupo

Consumo industrial de un sábado

En primer lugar, es necesario determinar el número de clústers a emplear en la ejecución del algoritmo K-Means. Para ello, se ha calculado la función de coste de la Figura 5.54, que muestra la suma total de las varianzas con respecto a los centroides, en función del número de grupos que se utilizan. Dado que el “codo” de la función no se distingue a simple vista, se ha recurrido al cálculo de la disminución porcentual de la suma de varianzas como medida de referencia para determinar el número de grupos a emplear. Al aumentar el número de clústers de quince a dieciséis, se obtiene una disminución del 8.27 % de la suma de varianzas, sin embargo, al continuar aumentando el número de clústers, este valor se reduce, por lo que se ha decidido emplear dieciséis clústers.

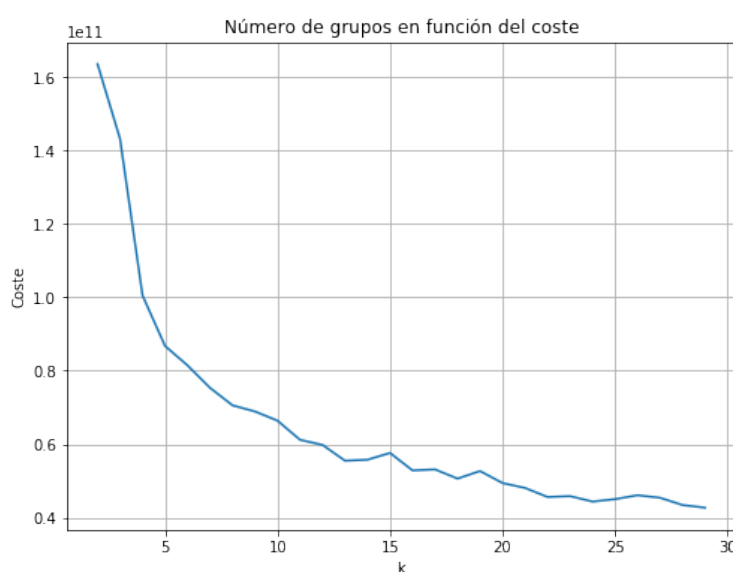


Figura 5.54.: Número de grupos en función de la suma de las distancias a los centroides

Una vez elegido el número de clústers a emplear, se ha ejecutado el algoritmo K-Means con la finalidad de encontrar los dieciséis grupos que definen el consumo industrial de un sábado de verano en la provincia de Madrid. Los grupos, junto con sus desviaciones típicas, se muestran en la Figura 5.55.

Como en el caso del consumo industrial de primavera y el consumo industrial de un día laboral de verano, nos encontramos con un grupo que presenta la tendencia reflejada por el Grupo 6, en este caso. Al igual que se ha comentado en casos anteriores, este tipo de comportamiento se asocia con el consumo generado por un restaurante, debido a la depresión en el consumo entre las 7:00 y las 12:00, aproximadamente, así como el gran pico de consumo que se da en torno a las 22:00 y los niveles de consumo prácticamente estables en la franjas horarias restantes. Una vez más, este tipo de consumo representa un porcentaje ínfimo del consumo industrial, siendo en este caso igual a 0.18 %, Tabla 5.40.

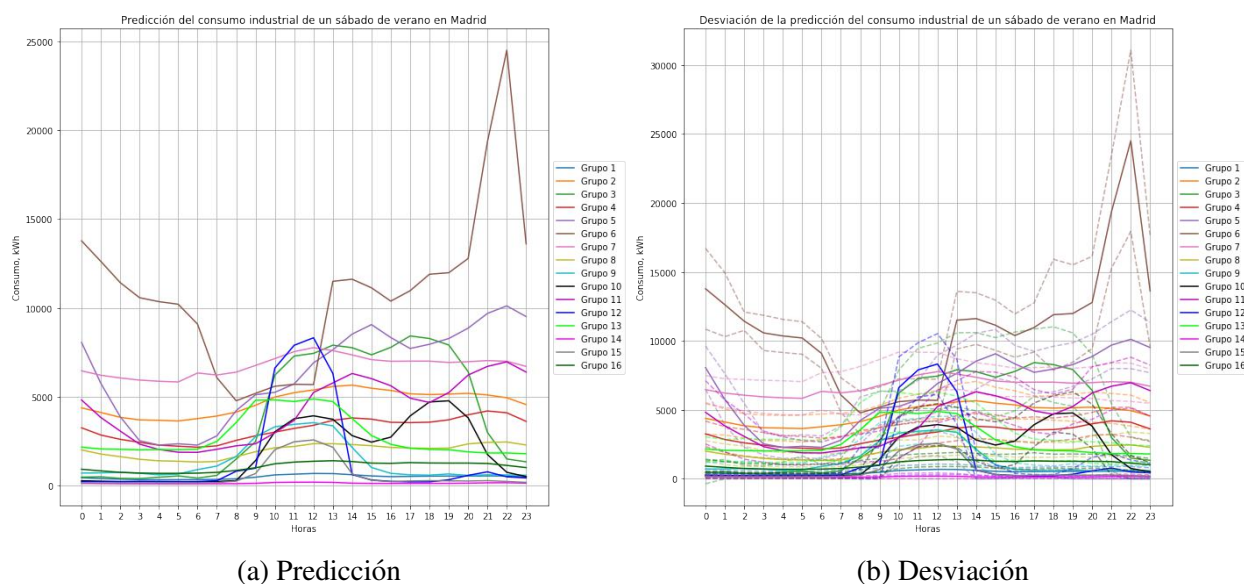


Figura 5.55.: Consumo industrial de un sábado de verano en Madrid

Además, llama la atención la presencia de un conjunto de grupos cuyo consumo se concentra entre las 9:00 y las 14:00, aproximadamente. Lo que refleja el patrón de consumo de establecimientos que tan solo ofrecen sus servicios en horario de mañana. Este conjunto está formado por el Grupo 9, 12, 13 y 14; que, en total, representan el 9.03 %, Tabla 5.40, del consumo industrial de un sábado de verano en la provincia de Madrid. En la misma línea, se encuentra el Grupo 3, que también presenta un consumo constante en una franja horaria, que en este caso abarca desde las 9:00 hasta las 21:00, lo que se traduce en una jornada de doce horas en la que el consumo eléctrico permanece prácticamente invariable.

Existe un conjunto de grupos que presentan la tendencia esperada de un consumo industrial. Es el caso de los grupos 5, 10 y 11. Estos grupos presentan los dos picos típicos del consumo industrial, uno por la mañana y otro por la tarde, así como la depresión durante la hora de comer.

El resto de los grupos representados en la Figura 5.55, poseen una tendencia prácticamente plana a lo largo del día. Es muy probable que estos grupos presenten la forma típica del consumo industrial, expuesta en la sección 5.3.2, ya que se aprecian unas pequeñas elevaciones del consumo en las franjas horarias que se corresponden con los picos de consumo. Sin embargo, es difícil de discernir debido a las magnitudes representadas en el eje Y. Una vez más, el grupo con el nivel de consumo menor es el que representa una mayor proporción del consumo total, en este caso es el Grupo 14 que abarca el 46.34 %, Tabla 5.40, del consumo total.

Las distribuidoras eléctricas buscan aplanar los niveles de consumo a lo largo del día para optimizar la producción de las centrales eléctricas. En este caso, con la finalidad de alcanzar este objetivo, se puede lanzar, principalmente, una campaña publicitaria. Esta campaña estará centrada en reducir el gran pico de consumo que sufre el Grupo 6 en torno a las 22:00. Una forma para conseguirlo se puede basar en el aumentar el precio de la energía a partir de las 22:00, lanzar ofertas para otras franjas horarias o aumentar el precio de la energía al sobrepasar cierto límite de consumo.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	22.18 %	Grupo 7	0.88 %	Grupo 13	1.02 %
Grupo 2	1.79 %	Grupo 8	4.32 %	Grupo 14	46.34 %
Grupo 3	0.54 %	Grupo 9	1.91 %	Grupo 15	5.57 %
Grupo 4	3.31 %	Grupo 10	1.24 %	Grupo 16	8.89 %
Grupo 5	0.46 %	Grupo 11	0.85		
Grupo 6	0.18 %	Grupo 12	0.53 %		

Tabla 5.40.: Predicción del porcentaje de datos asociados a cada grupo

Con el modelo establecido, es necesario evaluar la fiabilidad del mismo. Para ello, se comparan los porcentajes de representación de cada grupo obtenidos en la fase de predicción, Tabla 5.40, con los obtenidos en la fase de validación, Tabla 5.41. En este caso, la variación media entre los grupos es del 0.40 %, lo que es asumible y por tanto el modelo es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	23.78 %	Grupo 7	0.69 %	Grupo 13	1.01 %
Grupo 2	1.65 %	Grupo 8	5.24 %	Grupo 14	44.37 %
Grupo 3	0.28 %	Grupo 9	1.86 %	Grupo 15	5.93 %
Grupo 4	3.39 %	Grupo 10	1.57 %	Grupo 16	8.75 %
Grupo 5	0.44 %	Grupo 11	0.52		
Grupo 6	0.08 %	Grupo 12	0.48 %		

Tabla 5.41.: Validación del porcentaje de datos asociados a cada grupo

Consumo industrial de un domingo

En primer lugar, es necesario determinar el número de clústers a emplear en la ejecución del algoritmo K-Means. Para ello, se ha calculado la función de coste de la Figura 5.56, que muestra la suma total de las varianzas con respecto a los centroides, en función del número de grupos que se utilizan. Dado que el “codo” de la función no se distingue a simple vista, se ha recurrido al cálculo de la disminución porcentual de la suma de varianzas como medida de referencia para determinar el número de grupos a emplear. Al aumentar el número de clústers de nueve a diez, se obtiene una disminución del 6.20 % de la suma de varianzas, sin embargo, al continuar aumentando el número de clústers, este valor se reduce, por lo que se ha decidido emplear diez clústers.

Una vez escogido el número de clústers a emplear, se ha implementado el algoritmo K-Means, con la finalidad de encontrar los diez grupos que, en principio, definen el consumo industrial de un domingo de verano en la provincia de Madrid. Los grupos, junto con las desviaciones típicas, se encuentran en la Figura 5.57.

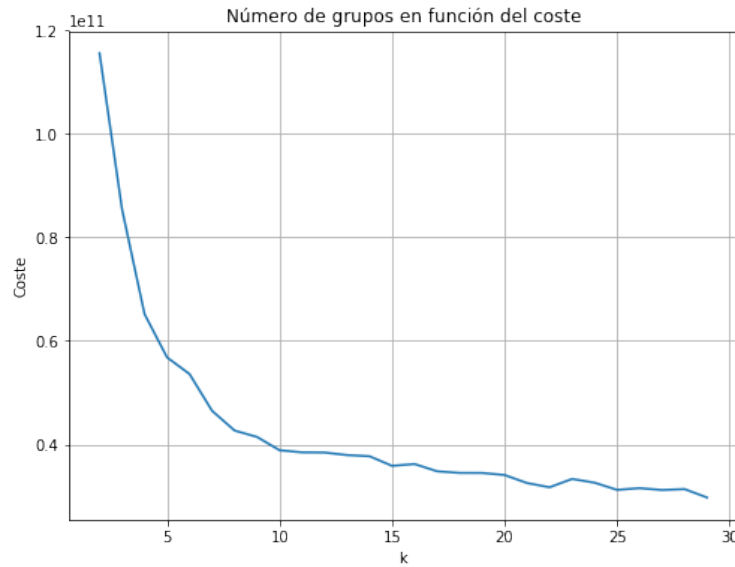


Figura 5.56.: Número de grupos en función de la suma de las distancias a los centroides

Como en el caso del consumo industrial de primavera y el consumo industrial de un día laboral y de un sábado de verano, nos encontramos con un grupo que presenta la tendencia reflejada por el Grupo 4, en este caso. Al igual que se ha comentado en casos anteriores, este tipo de comportamiento se asocia con el consumo generado por un restaurante, debido a la depresión en el consumo entre las 7:00 y las 12:00, aproximadamente, así como el gran pico de consumo que se da en torno a las 22:00 y los niveles de consumo prácticamente estables en la franjas horarias restantes. Una vez más, este tipo de consumo representa un porcentaje ínfimo del consumo industrial, siendo en este caso igual a 0.15 %, Tabla 5.42.

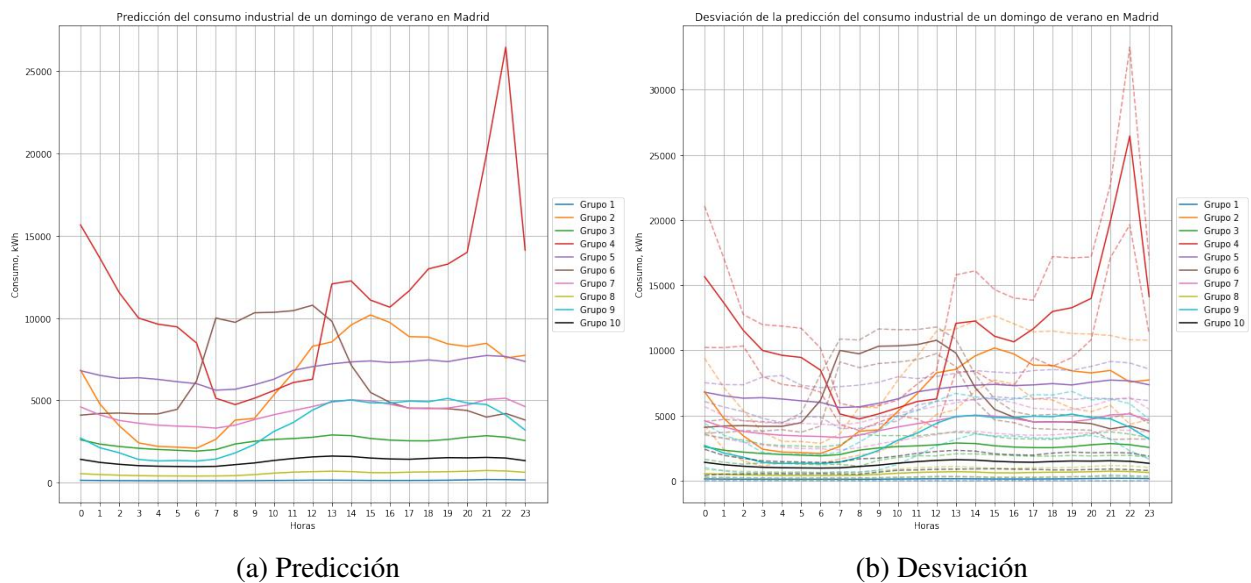


Figura 5.57.: Consumo industrial de un domingo de verano en Madrid

Además, llama la atención la presencia del Grupo 6 cuyo consumo se concentra entre las 9:00 y las 14:00, aproximadamente. Lo que refleja el patrón de consumo de establecimientos que tan solo ofrecen sus servicios en horario de mañana. En este caso, este grupo representa, solamente el 0.16 %, Tabla 5.42, del consumo industrial de un domingo de verano en la provincia de Madrid.

También destaca el Grupo 2, que posee una tendencia que recuerda a la del consumo residencial encontrada en la sección 5.3.2. Este tipo de comportamiento se corresponde con el de establecimientos que abren en torno a las 11:00 y mantienen un consumo estable hasta la hora de cierre, en torno a las 23:00. Este grupo tan solo representa el 0.32 %, Tabla 5.42, del consumo total de esta casuística.

El resto de los grupos representados en la Figura 5.57, poseen una tendencia prácticamente plana a lo largo del día. Es muy probable que estos grupos presenten la forma típica del consumo industrial, expuesta en la sección 5.3.2, ya que se aprecian unas pequeñas elevaciones del consumo en las franjas horarias que se corresponden con los picos de consumo. Sin embargo, es difícil de discernir debido a las magnitudes representadas en el eje Y. Una vez más, el grupo con el nivel de consumo menor es el que representa una mayor proporción del consumo total, en este caso es el Grupo 1 que abarca el 60.16 %, Tabla 5.42, del consumo total.

Las distribuidoras eléctricas buscan aplanar los niveles de consumo a lo largo del día para optimizar la producción de las centrales eléctricas. En este caso, con la finalidad de alcanzar este objetivo, se puede lanzar, principalmente, una campaña publicitaria. Esta campaña estará centrada en reducir el gran pico de consumo que sufre el Grupo 4 en torno a las 22:00. Una forma para conseguirlo se puede basar en aumentar el precio de la energía a partir de las 22:00, lanzar ofertas para otras franjas horarias o aumentar el precio de la energía al sobrepasar cierto límite de consumo.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	60.16 %	Grupo 6	0.16 %
Grupo 2	0.32 %	Grupo 7	2.61 %
Grupo 3	4.20 %	Grupo 8	21.33 %
Grupo 4	0.15 %	Grupo 9	1.68 %
Grupo 5	0.65 %	Grupo 10	8.69 %

Tabla 5.42.: Predicción del porcentaje de datos asociados a cada grupo

Con el modelo establecido, es necesario evaluar la fiabilidad del mismo. Para ello, se comparan los porcentajes de representación de cada grupo obtenidos en la fase de predicción, Tabla 5.42, con los obtenidos en la fase de validación, Tabla 5.43. En este caso, la variación media dentro los grupos es del 0.34 %, lo que es asumible y por tanto el modelo es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	59.39 %	Grupo 6	0.04 %
Grupo 2	0.48 %	Grupo 7	2.46 %
Grupo 3	3.97 %	Grupo 8	22.43 %
Grupo 4	0.08 %	Grupo 9	1.35 %
Grupo 5	0.99 %	Grupo 10	8.81 %

Tabla 5.43.: Validación del porcentaje de datos asociados a cada grupo

5.4.1.4. Perfiles del consumo de un día de otoño en la provincia de Madrid

El objetivo de esta sección es encontrar los distintos patrones que representan el consumo residencial e industrial de un día de otoño en la provincia de Madrid. Para alcanzar dicha finalidad, se va a seguir la dinámica empleada en las secciones anteriores, en primer lugar, se obtiene el número de clústers a emplear en cada caso, para después calcular los distintos patrones de consumo mediante la implementación del algoritmo K-Means.

Consumo residencial de un día laboral

En primer lugar, es necesario determinar el número de clústers a emplear. Para ello, se ha creado la función de coste de la Figura 5.58, que representa la suma total de las varianzas con respecto a los centroides, en función del número de clústers empleados. Como el “codo” de la función no es identificable a primera vista, se ha recurrido al cálculo de la disminución porcentual de la suma de varianzas como medida de referencia para determinar el número de clústers necesario. Al aumentar el número de clústers de catorce a quince, se consigue una disminución de la suma de varianzas del 2.09 %, sin embargo, si se sigue aumentando esta cifra, los valores se disminuyen, es decir, si se sigue aumentando el número de clústers, los nuevos clústers, cada vez aportan menor información. Motivo por el cual se ha decidido emplear quince clústers.

Con el número de clústers elegido, se ha podido implementar el algoritmo K-Means, cuyo objetivo es encontrar los patrones que reflejan el consumo residencial de un día laboral de otoño en la provincia de Madrid. Los grupos obtenidos, junto con sus respectivas desviaciones típicas se encuentran representados en la Figura 5.59.

De entre los grupos calculados, a primera vista, destaca la tendencia seguida por los clientes pertenecientes al Grupo 11. En conjunto, la tendencia sigue la directrices básicas de un consumo residencial a lo largo del día, es decir, a partir de las 7:00 de la mañana el consumo comienza a ascender, hasta la hora de cenar, a partir de la cual empieza el descenso propio de la horas de sueño. Sin embargo, en este caso, destaca el gran pico de consumo que tiene lugar en torno a las 20:00. Estos valores de consumo tan elevados, se relacionan con grandes viviendas o ineficientes, cuyos valores de consumo se encuentran muy por encima de la media. El patrón seguido por este grupo no es muy habitual ya que solamente el 0.04 %, Tabla 5.44 del consumo residencial de un día laboral de otoño en Madrid pertenece a este grupo.

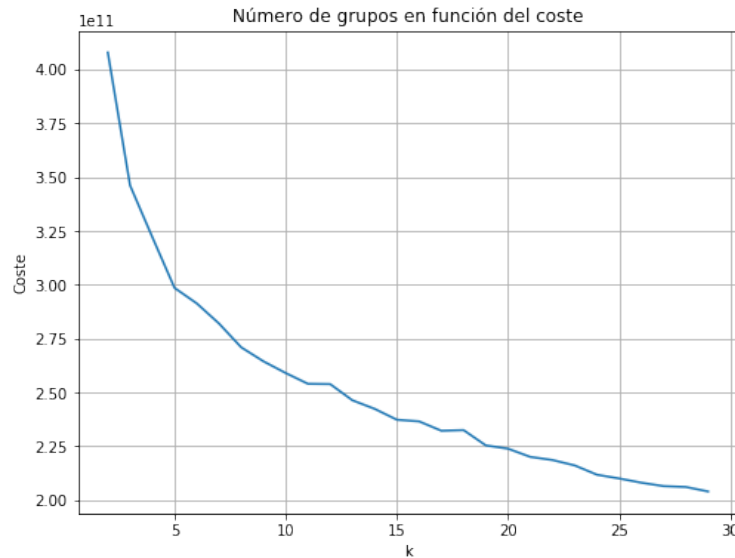


Figura 5.58.: Número de grupos en función de la suma de las distancias a los centroides

Un conjunto de grupos siguen la tendencia más tradicional del consumo residencial, encontrada en la sección 5.3.2. Estos grupos son el 1, 3, 4, 7, 9, 12, 13, 14 y 15. Cada uno presenta distintos niveles de consumo, pero la tendencia seguida es la misma. Crecimiento con el consumo a lo largo del día hasta alcanzar el pico correspondiente con la hora de la cena, para después descender durante las horas de sueño. En conjunto, estos grupos representan el 85.98 %, Tabla 5.44 del consumo residencial de esta casuística.

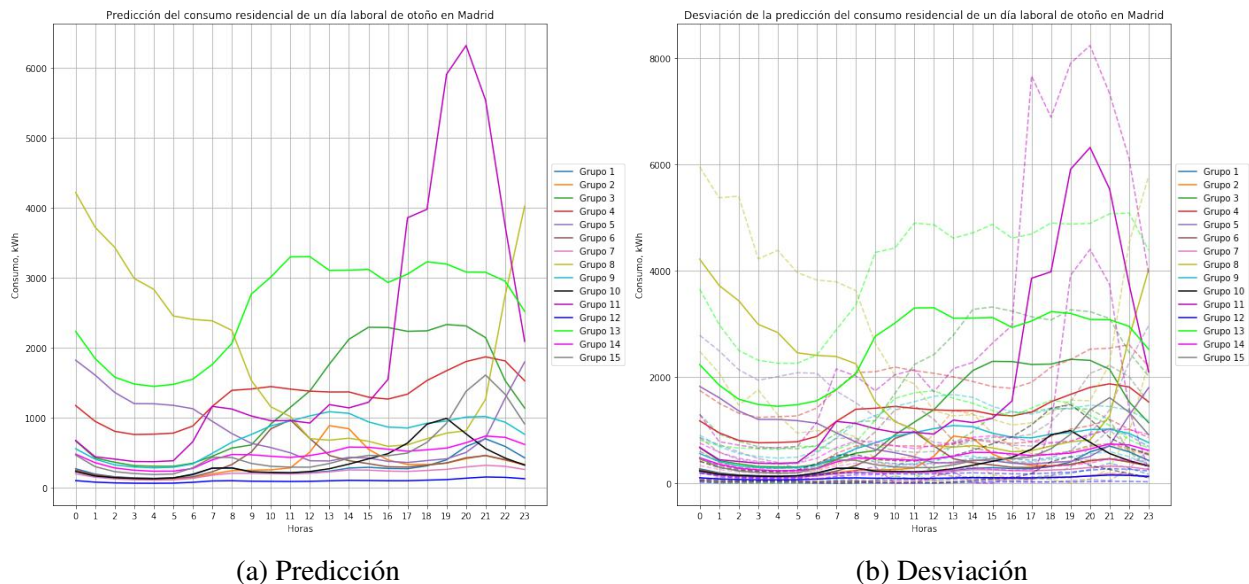


Figura 5.59.: Consumo residencial de un día laboral de otoño en Madrid

Además, cabe destacar la presencia de dos grupos cuyo patrón de consumo a lo largo del día es completamente contrario a los expuestos. Es el caso de los grupos 8 y 5. Estos grupos, presentan el pico de consumo en torno a las 00:00, para después descender hasta, aproximadamente, las 18:00 hora a la que empieza la etapa ascendiente. Es decir, los clientes pertenecientes a este grupo no generan consumo hasta media tarde, cuando lo habitual es que el consumo empiece la fase de ascenso a las 7:00. Tendencias similares ya se observaron en el caso del consumo residencial de un sábado de primavera, así como en todos los casos del consumo residencial de invierno. Este comportamiento se puede asociar con la agenda de aquellas personas que trabajen por la noche. Motivo por el cual, estos dos grupos, en conjunto, tan solo representan el 0.79 %, Tabla 5.44, del consumo.

También llama la atención la presencia de los grupos 2, 6 y 10 que se caracterizan por presentar un niveles de consumo elevados en determinadas franjas horarias, cada uno en una distinta. Lo que indica que esas son las franjas horarias en las que se concentra la mayor parte de la actividad en la vivienda, ya sea cocina, hacer la colada o encender la calefacción.

Como se ha comentado en apartados anteriores, uno de los principales objetivos de las distribuidoras eléctricas es conseguir aplanar los niveles de consumo a lo largo del día para optimizar la producción de las centrales eléctricas. En este caso, con la finalidad de alcanzar este objetivo se pueden lanzar, principalmente, dos campañas publicitarias. Una de ellas se centraría en incentivar el consumo de los clientes pertenecientes al Grupo 8 y 5 durante las horas del día, mediante la propuesta de oferta si consumen a dichas horas. Por otro lado, a los clientes del Grupo 11 se les podría aumentar el precio de la energía desde las 18:00 hasta las 22:00 o a partir de cierto nivel de consumo, para fomentar una reducción en el consumo.

Con el modelo predictivo establecido, es necesario evaluar su fiabilidad. Para ello se comparan los porcentajes de pertenencia de cada grupo obtenidos en la fase de predicción, Tabla 5.44, con los obtenidos en la fase de validación, Tabla 5.45, utilizando el 25 % restante de los datos. El resultado de esta comparación es una variación media por grupo de 0.06 %, lo que es asumible, y por ello el modelo es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	14.25 %	Grupo 6	3.99 %	Grupo 11	0.04 %
Grupo 2	5.39 %	Grupo 7	30.97 %	Grupo 12	28.57 %
Grupo 3	0.21 %	Grupo 8	0.20 %	Grupo 13	0.14 %
Grupo 4	0.64 %	Grupo 9	2.43 %	Grupo 14	6.83 %
Grupo 5	0.59 %	Grupo 10	3.77 %	Grupo 15	1.94 %

Tabla 5.44.: Predicción del porcentaje de datos asociados a cada grupo

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	14.36 %	Grupo 6	3.94 %	Grupo 11	0.06 %
Grupo 2	5.40 %	Grupo 7	30.86 %	Grupo 12	28.75 %
Grupo 3	0.19 %	Grupo 8	0.21 %	Grupo 13	0.15 %
Grupo 4	0.60 %	Grupo 9	2.31 %	Grupo 14	6.91 %
Grupo 5	0.55 %	Grupo 10	3.85 %	Grupo 15	1.85 %

Tabla 5.45.: Validación del porcentaje de datos asociados a cada grupo

Consumo residencial de un sábado

Antes de implementar el algoritmo K-Means, es necesario determinar el número de clústers a emplear. Para ello, se ha creado la función de coste de la Figura 5.60, que representa la suma total de las varianzas con respecto a los centroides, en función del número de clústers empleados. Como el “codo” de la función no se puede distinguir a primera vista, se ha calculado la disminución porcentual de la suma de varianzas como medida de referencia para determinar el número de clústers necesario. Al aumentar el número de clústers de diez a once, la suma de varianzas sufre una disminución del 2.85 %, sin embargo, al seguir aumentando el número de clústers este número se reduce a un 1.05 %, por lo que se han decidido emplear once clústers.

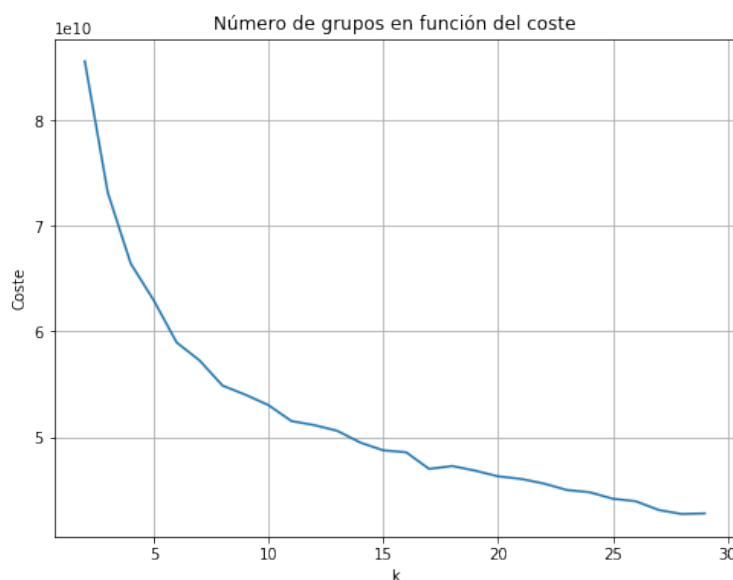


Figura 5.60.: Número de grupos en función de la suma de las distancias a los centroides

Con el número de clústers elegido, se ha podido implementar el algoritmo K-Means, cuyo objetivo es extraer los patrones que representan el consumo residencial de un sábado de otoño en la provincia de Madrid. Los grupos obtenidos, junto con sus respectivas desviaciones típicas se encuentran representados en la Figura 5.61.

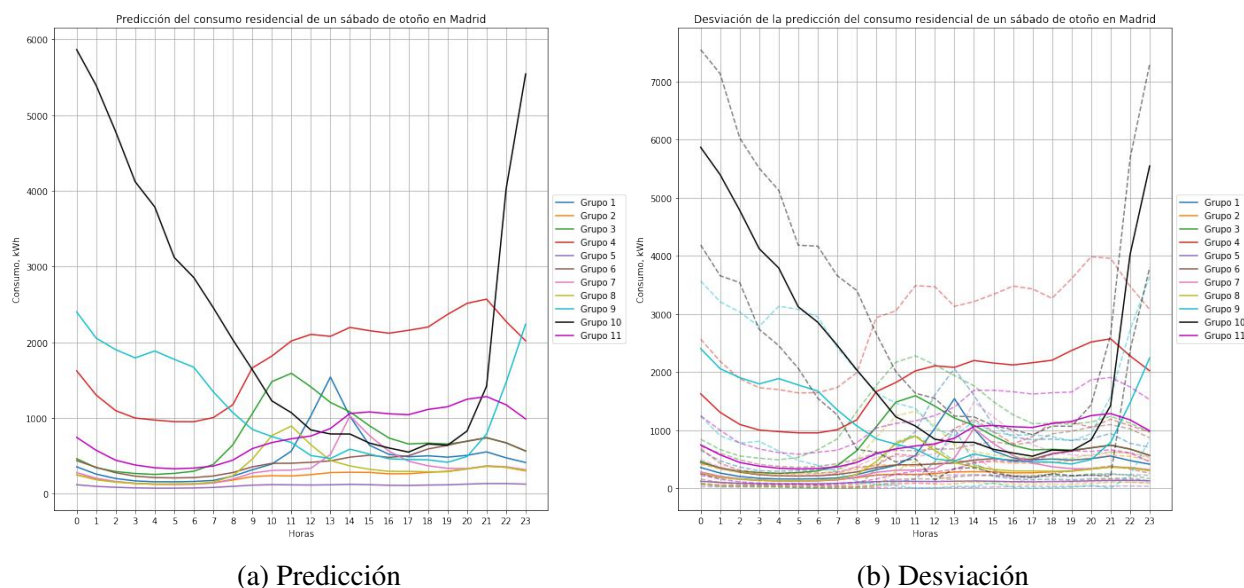


Figura 5.61.: Consumo residencial de un sábado de otoño en Madrid

A primera vista, los grupos que más destacan son el Grupo 10 y el Grupo 9. Estos grupos se caracterizan por presentar sus niveles de consumo más elevados durante las horas de sueño. Cuando lo esperado es que el consumo disminuya durante la noche y aumente a medida que avanza el día. Este tipo de tendencia se corresponden con el consumo generado por clientes que trabajan en el turno de noche. Este conjunto de grupos tan solo representa el 0.69 %, Tabla 5.46, del consumo residencial de un sábado de otoño en la provincia de Madrid.

Además, también llama la atención el hecho de que algunos grupos tan solo presenta un pico de consumo en torno a la hora de la comida, lo que contrasta con la tendencia de consumo residencial obtenida en la sección 5.3.2. En dicha sección, se encontraba una tendencia general para el consumo residencial en la que se podían distinguir dos picos en el consumo, uno a la hora de la comida y otro a la hora de la cena, siendo este último el más elevado. La tendencia seguida por los grupos 1, 3, 7 y 8 se corresponde con la de personas que solamente almuerzan en sus viviendas o generan muy poco consumo durante la hora de la cena, lo que se podría justificar si no emplean electrodomésticos para cocinar, por ejemplo.

El resto de los grupos, siguen el patrón de consumo residencial esperado con mayores o menores niveles de consumo. Este conjunto de grupos formado abarca el 81.3 %, Tabla 5.46, del consumo residencial de un sábado de otoño en la provincia de Madrid.

Las distribuidoras eléctricas desean aplanar los niveles de consumo a lo largo del día para optimizar la producción de las centrales eléctricas. En este caso, con la finalidad de alcanzar dicho objetivo, se puede lanzar, principalmente, una campaña publicitaria. El objetivo de dicha campaña será el de distribuir el consumo nocturno de los grupos 10 y 9 durante le resto de las horas del día. Para ello, se pueden lanzar ofertas durante las horas a las que se desea mover dicho consumo.

Con el modelo establecido, es necesario evaluar la veracidad del mismo. Para ello, se comparan los porcentajes de representación de cada grupo obtenidos en la fase de predicción, Tabla 5.46, con los obtenidos en la fase de validación, Tabla 5.47. En este caso, la variación media dentro de los grupos es del 0.22 %, lo que es asumible, por lo que el modelo es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	3.34 %	Grupo 5	35.18 %	Grupo 9	0.59 %
Grupo 2	31.13 %	Grupo 6	11.58 %	Grupo 10	0.10 %
Grupo 3	1.31 %	Grupo 7	6.97 %	Grupo 11	2.88 %
Grupo 4	0.53 %	Grupo 8	6.40 %		

Tabla 5.46.: Predicción del porcentaje de datos asociados a cada grupo

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	3.24 %	Grupo 5	35.07 %	Grupo 9	0.56 %
Grupo 2	32.00 %	Grupo 6	11.69 %	Grupo 10	0.09 %
Grupo 3	1.50 %	Grupo 7	6.60 %	Grupo 11	2.73 %
Grupo 4	0.54 %	Grupo 8	5.99 %		

Tabla 5.47.: Validación del porcentaje de datos asociados a cada grupo

Consumo residencial de un domingo

En primer lugar, es necesario determinar el número de clústers a emplear. Para ello, se ha creado la función de coste de la Figura 5.62, que representa la suma total de las varianzas con respecto a los centroides, en función del número de clústers empleados. Como el “codo” de la función no es identificable a primera vista, se ha recurrido al cálculo de la disminución porcentual de la suma de varianzas como medida de referencia para determinar el número de clústers necesario. Al aumentar el número de clústers de catorce a quince, se consigue una disminución de la suma de varianzas del 2.78 %, sin embargo, si se sigue aumentando esta cifra, los valores se disminuyen, es decir, si se sigue aumentando el número de clústers, los nuevos clústers, cada vez aportan menor información. Motivo por el cual se ha decidido emplear quince clústers.

Con el número de clústers elegido, se ha podido implementar el algoritmo K-Means, cuyo objetivo es encontrar los patrones que reflejan el consumo residencial de un día laboral de otoño en la provincia de Madrid. Los grupos obtenidos, junto con sus respectivas desviaciones típicas se encuentran representados en la Figura 5.63.

Cabe destacar la presencia de dos grupos cuyo patrón de consumo a lo largo del día es completamente contrario a los expuestos, patrón que también ha sido observado en los dos casos anteriores. Es el caso de los grupos 9 y 10. Estos grupos, presentan el pico de consumo en torno a las 00:00, para después descender hasta, aproximadamente, las 18:00 hora a la que empieza la etapa ascendente. Es decir, los clientes pertenecientes a este grupo no generan consumo hasta media

tarde, cuando lo habitual es que el consumo empiece la fase de ascenso a las 7:00. Este comportamiento se puede asociar con la agenda de aquellas personas que trabajen por la noche. Motivo por el cual, estos dos grupos, en conjunto, tan solo representan el 0.63 %, Tabla 5.48, del consumo.

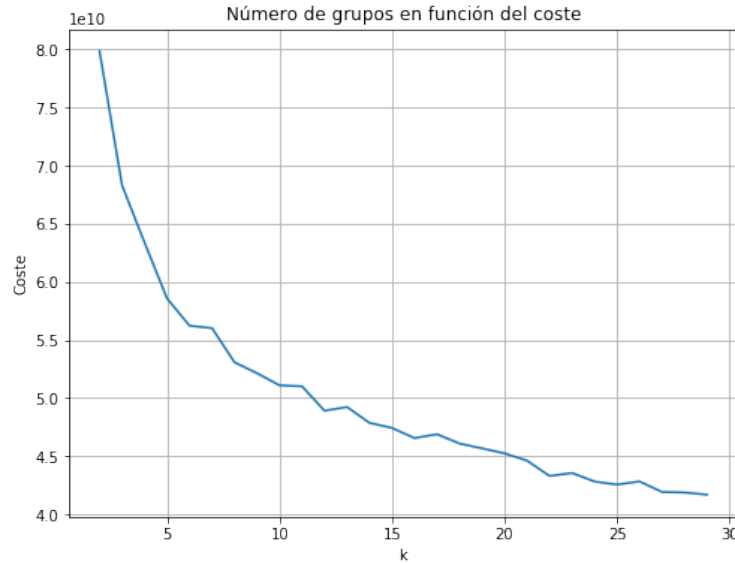


Figura 5.62.: Número de grupos en función de la suma de las distancias a los centroides

También destaca la tendencia seguida por el Grupo 5. Aunque a grandes rasgos, sigue el esquema esperado de un consumo residencial a lo largo del día, posee varios máximos relativos a lo largo del día, que no se encontraban en la tendencia general del consumo residencial encontrada en la sección 5.3.2. La rareza de esta tendencia tan solo es mantenida por el 0.04 %, Tabla 5.48, del consumo de este caso de estudio.

Siguiendo de forma más estricta la tendencia del consumo residencial encontrada en la sección 5.3.2, se encuentra los grupos 2, 3, 4, 6, 8, 11 y 13; con distintos niveles de consumo. El conjunto formado por estos grupos representa el 78.87 % del consumo residencial de un domingo de otoño en Madrid.

El resto de los grupos se caracterizan por presentar un pico claro a lo largo del día. Los grupo 15, 7, 14 y 12 lo sufren a la hora del mediodía, mientras que en el Grupo 1 se manifiesta durante la hora de la cena. Este comportamiento se corresponden con el de clientes que solamente realizan una de las comidas en sus viviendas o que apenas emplean electrodoméstico o la calefacción/aire acondicionado, cuando la mayoría lo hacen.

Uno de los principales objetivos de las distribuidoras eléctricas es conseguir aplanar los niveles de consumo a lo largo del día para optimizar la producción de las centrales eléctricas. En este caso, con la finalidad de alcanzar este objetivo se pueden lanzar, principalmente, dos campañas publicitarias. El objetivo de una de ellas debe ser redistribuir el consumo de los grupos 9 y 10, a lo largo de las horas de las horas de luz. Lo que se pueden conseguir mediante el lanzamiento de

ofertas en las franjas horarias a las que desea trasladar el consumo. Por otro lado, la otra campaña se puede centrar en la reducción del consumo de los clientes pertenecientes al Grupo 5, que se puede conseguir mediante el aumento del precio de la energía una vez se traspase cierto límite de consumo.

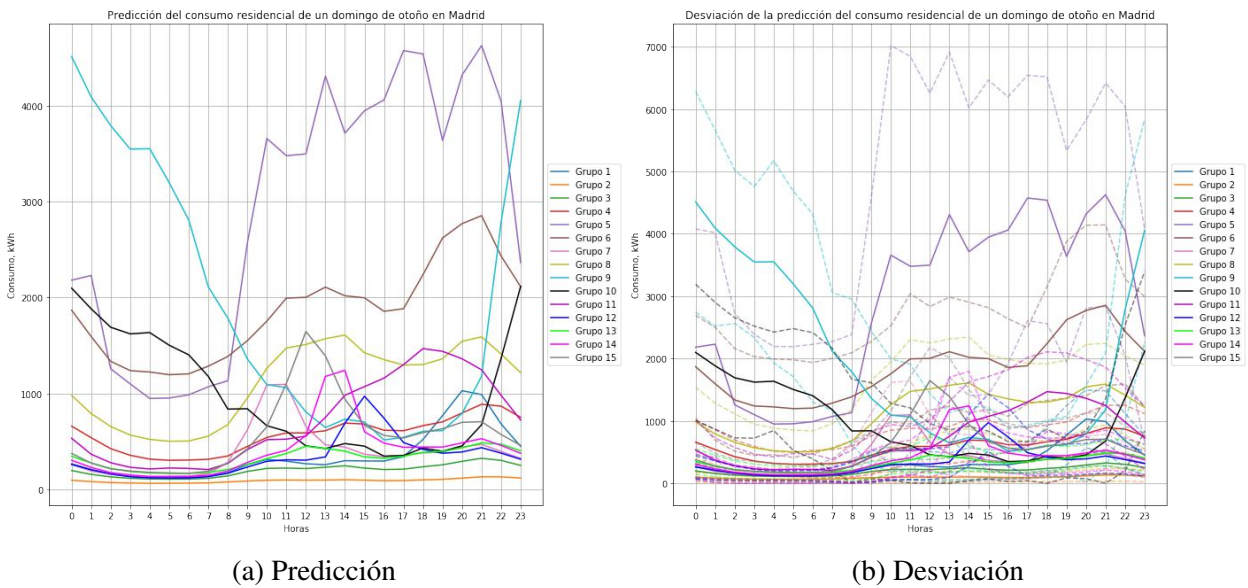


Figura 5.63.: Consumo residencial de un domingo de otoño en Madrid

Con el modelo predictivo establecido, es necesario evaluar la veracidad del mismo. Motivo por el que se comparan los porcentajes de representación de cada grupo obtenidos en la fase de predicción, Tabla 5.48, con los obtenidos en la fase de validación, Tabla 5.49. En este caso, el resultado de la comparación es una variación media por cada grupo de 0.12 %, lo que es asumible y por ello el modelo es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	4.53 %	Grupo 6	0.24 %	Grupo 11	1.36 %
Grupo 2	26.18 %	Grupo 7	4.08 %	Grupo 12	4.92 %
Grupo 3	30.00 %	Grupo 8	1.04 %	Grupo 13	15.26 %
Grupo 4	4.79 %	Grupo 9	0.18 %	Grupo 14	4.37 %
Grupo 5	0.04 %	Grupo 10	0.45 %	Grupo 15	2.56 %

Tabla 5.48.: Predicción del porcentaje de datos asociados a cada grupo

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	4.19 %	Grupo 6	0.24 %	Grupo 11	1.31 %
Grupo 2	26.41 %	Grupo 7	4.50 %	Grupo 12	4.97 %
Grupo 3	29.68 %	Grupo 8	1.03 %	Grupo 13	15.26 %
Grupo 4	4.82 %	Grupo 9	0.12 %	Grupo 14	4.54 %
Grupo 5	0.05 %	Grupo 10	0.43 %	Grupo 15	2.45 %

Tabla 5.49.: Validación del porcentaje de datos asociados a cada grupo

Consumo industrial de un día laboral

Una vez estudiado el consumo residencial en otoño para todos los días de la semana, es el turno del consumo industrial. En primer lugar, antes de encontrar los patrones del consumo industrial de un día laboral de otoño en la provincia de Madrid, es necesario obtener el número de clústers a emplear. Para lograr dicho objetivo, se ha creado el siguiente gráfico, Figura 5.64. Como se puede observar, es “codo” de la función no se puede observar a primer vista, por lo que se ha recurrido al cálculo del porcentaje de disminución de la varianza. Al aumentar el número de clústers de once a doce, la varianza disminuye en un 6.37 %, mientras que al aumentar a trece, sólo se aprecia una disminución del 1.70 %, por lo que el número de clústers elegido es doce.

En la sección 5.3, se tratan las diferencias existentes entre el consumo residencial y el industrial. El consumo industrial se caracteriza por presentar un consumo medio mayor y por la simetría que presenta con respecto al centro de la jornada laboral. Además, el consumo industrial sufre una gran disminución durante los fines de semana. Con estas ideas presente, se pueden estudiar las diferencias existentes entre los patrones encontrados, Figure 5.65.

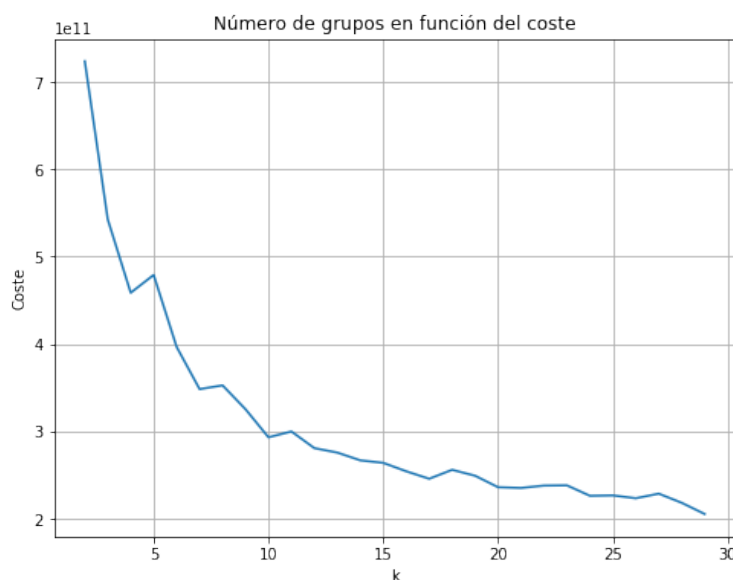


Figura 5.64.: Número de grupos en función de la suma de las distancias a los centroides

A primera vista, destaca el patrón seguido por los clientes pertenecientes a los Grupo 11, ya que presenta un consumo muy elevado que se concentra entre las 7:00 y las 15:00, principalmente, lo cual no encaja con el patrón observado en la Figura 5.21 de la sección 5.3. Por este motivo, solamente el 0.11 %, Tabla 5.50, de los datos pertenecen a este grupo. De forma más sutil, el consumo del Grupo 10 presenta una tendencia similar. Esta tendencia ya se observaba en el apartado acerca del consumo industrial de un día laboral de invierno, sin embargo no se encuentra en el consumo industrial de primavera o de verano, lo que lleva a pensar que existe una relación entre la presencia de este grupo y la temperatura de la época del año.

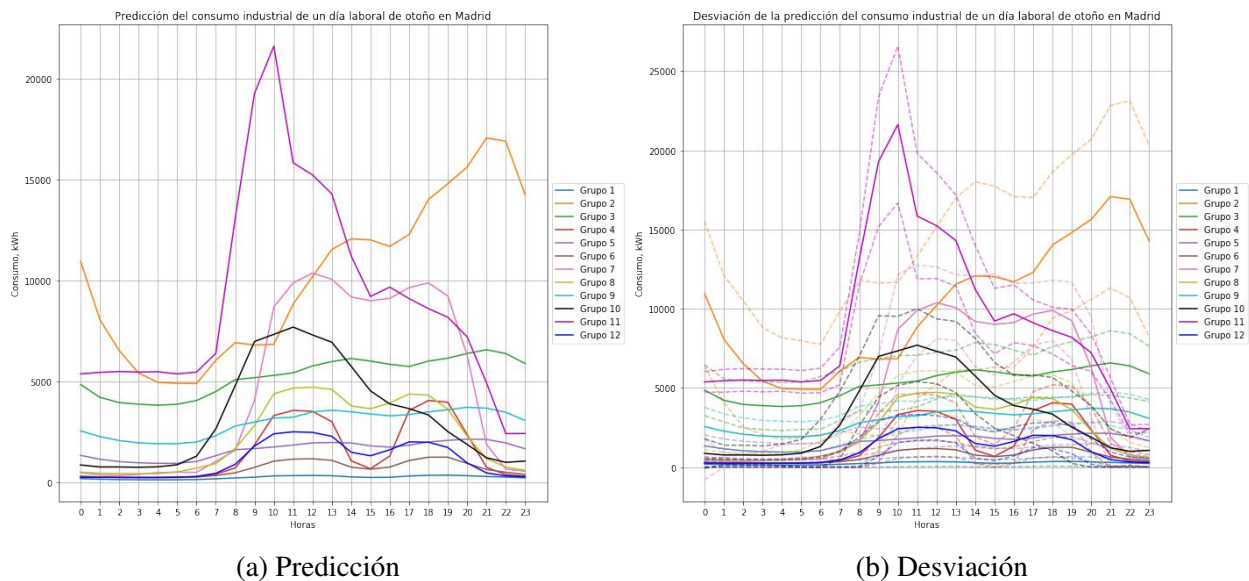


Figura 5.65.: Consumo industrial de un día laboral de otoño en Madrid

También caben destacar la tendencia del Grupo 2 cuyo consumo presenta una progresión ascendente a lo largo de la jornada laboral, asemejándose a la tendencia esperada de un consumo residencial. Esto implica que son empresas cuya actividad industrial aumenta a la hora de la cena, por lo que puede ser el consumo de restaurantes, que durante los días de la semana tienen más clientela a la hora de cenar.

El resto de los patrones se pueden agrupar en dos categorías, aquellos grupos cuyo consumo se mantiene relativamente constante a lo largo de la jornada laboral (Grupo 1, 3, 5 y 9); y los que presentan dos picos de consumo, uno por la mañana y otro por la tarde (Grupo 4, 6, 7, 8 y 12). Además, el consumo del Grupo 1 acapara el 39.27 %, Tabla 5.50, del consumo total, y es el grupo con el menor consumo. Lo que indica la gran diferencia existente entre las empresas, hay algunas pocas que generan un gran consumo, mientras que otras presentan niveles de consumo similares a los residenciales. Este hecho ya se infería en el diagrama de cajas de la Figura 5.19 de la sección 5.3.

Uno de los objetivos principales de este proyecto es proporcionar información útil para las distribuidoras de energía eléctrica para que puedan lanzar campañas publicitarias a determinados

grupos de clientes. A las distribuidoras energéticas les conviene aplanar los niveles de consumo a lo largo del día para poder optimizar las instalaciones. En este caso en particular, con la finalidad de alcanzar dicho objetivo, se pueden lanzar, principalmente, dos campañas publicitarias. El objetivo de una de ellas sería bajar los niveles de consumo de los clientes del Grupo 11, mientras que otra se podría en bajar los niveles de consumo del Grupo 2. Un método para conseguirlo se puede basar en el disminuir del precio de la energía eléctrica de ciertas franjas horarias para fomentar la redistribución del consumo o aumentar el precio tras sobrepasar ciertos niveles de energía.

Con las predicciones calculadas, es necesario evaluar la fiabilidad de las mismas, para ello se comparan los porcentajes de los clientes pertenecientes a cada grupos, Tabla 5.50 y Tabla 5.51. De media, un porcentaje presenta una variación igual a 0.26 %, que es asumible, por lo que el modelo es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	39.27 %	Grupo 5	9.01 %	Grupo 9	5.44 %
Grupo 2	0.44 %	Grupo 6	24.41 %	Grupo 10	1.08 %
Grupo 3	1.54 %	Grupo 7	0.64 %	Grupo 11	0.11 %
Grupo 4	4.24 %	Grupo 8	3.01 %	Grupo 12	10.81 %

Tabla 5.50.: Predicción del porcentaje de datos asociados a cada grupo

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	39.86 %	Grupo 5	8.42 %	Grupo 9	5.61 %
Grupo 2	0.41 %	Grupo 6	24.72 %	Grupo 10	1.00 %
Grupo 3	1.72 %	Grupo 7	0.55 %	Grupo 11	0.07 %
Grupo 4	4.00 %	Grupo 8	3.34 %	Grupo 12	10.30 %

Tabla 5.51.: Validación del porcentaje de datos asociados a cada grupo

Consumo industrial de un sábado

En primer lugar, es necesario determinar el número de clústers a emplear en la ejecución del algoritmo K-Means. Para ello, se ha calculado la función de coste de la Figura 5.66, que muestra la suma total de las varianzas con respecto a los centroides, en función del número de grupos que se utilizan. Dado que el “codo” de la función no se distingue a simple vista, se ha recurrido al cálculo de la disminución porcentual de la suma de varianzas como medida de referencia para determinar el número de grupos a emplear. Al aumentar el número de clústers de trece a catorce, se obtiene una disminución del 7.24 % de la suma de varianzas, sin embargo, al continuar aumentando el número de clústers, este valor se reduce, por lo que se ha decidido emplear catorce clústers.

Una vez elegido el número de clústers a emplear, se ha ejecutado el algoritmo K-Means con la finalidad de encontrar los dieciséis grupos que definen el consumo industrial de un

sábado de otoño en la provincia de Madrid. Los grupos, junto con sus desviaciones típicas, se muestran en la Figura 5.67.

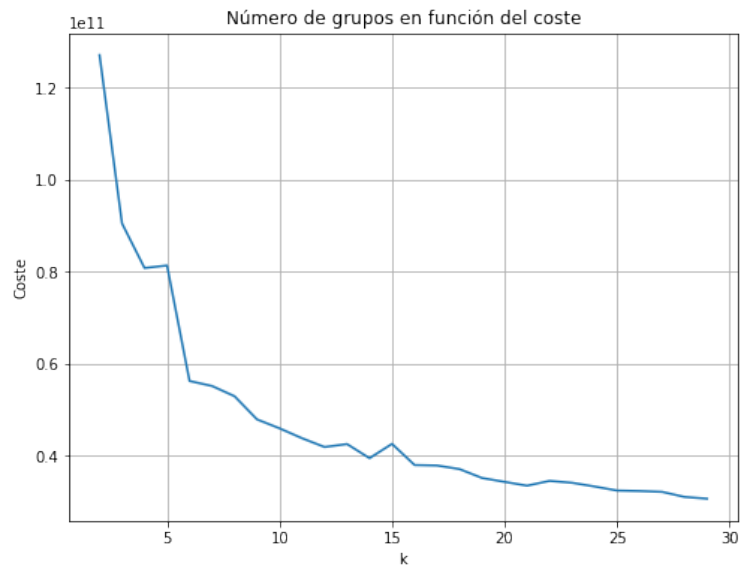


Figura 5.66.: Número de grupos en función de la suma de las distancias a los centroides

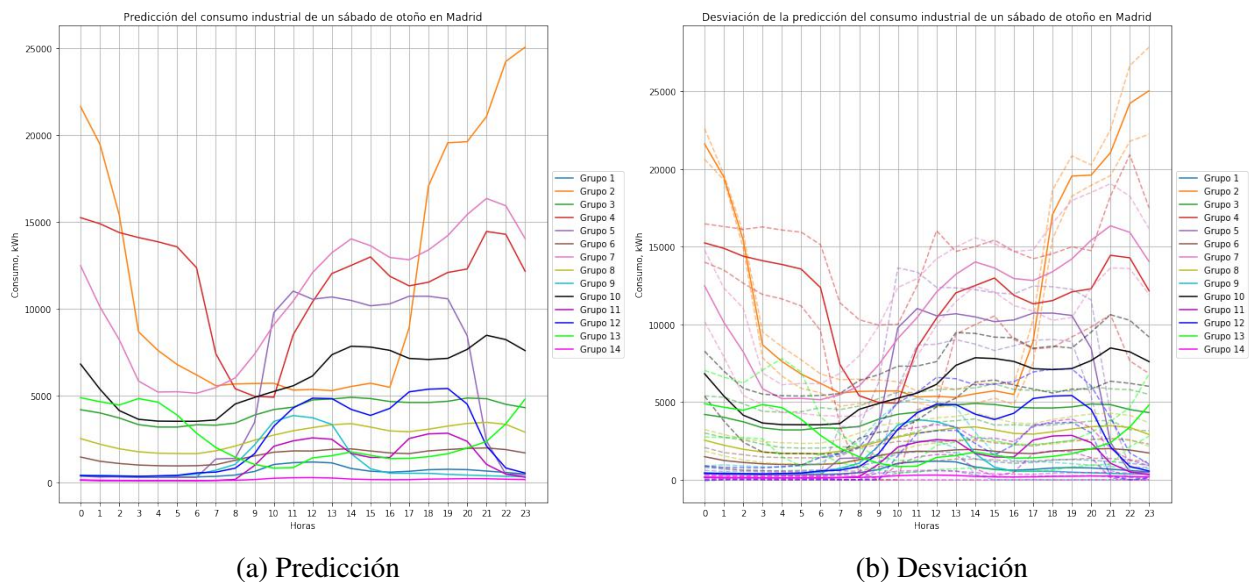


Figura 5.67.: Consumo industrial de un sábado de otoño en Madrid

A primera vista, destaca la tendencia seguida por los clientes pertenecientes al Grupo 2 y al Grupo 13. Tratando con consumo industrial, se espera encontrar dos picos de niveles de consumo similares, uno a media mañana y otro a media tarde, correspondiéndose con el punto medio de la jornada de mañana y de la de tarde, como se determinó en la sección 5.3.2. Sin embargo, en este caso, el consumo sufre una gran depresión desde las 3:00 hasta las 16:00 aproximadamente. Este

tipo de comportamiento se puede asociar a establecimientos que cierran por la mañanas, alargando su jornada laboral hasta la madrugada. Este puede ser el caso de bares que abran las puertas después de comer y cierren sobre las 3:00. El Grupo 2 tan solo representa el 0.07 %, Tabla 5.52, del consumo industrial de un sábado de otoño en la provincia de Madrid, mientras que el Grupo 13 abarca el 0.48 %. Una tendencia similar es seguida por el Grupo 4, aunque la franja horaria de depresión solo comprende de 7:00 a 11:00 aproximadamente, mostrando un consumo industrial más típico el resto del día.

También cabe destacar la tendencia del Grupo 7 y del Grupo 10, cuyo consumo presenta una progresión ascendente a lo largo de la jornada laboral, asemejándose a la tendencia esperada de un consumo residencial. Esto implica que son empresas cuya actividad industrial aumenta a la hora de la cena, por lo que puede ser el consumo de restaurantes, que presentan mayor afluencia de clientes a la hora de la cena.

Además, llama la atención el patrón del Grupo 5, ya que mantiene un nivel de consumo relativamente constante a lo largo de la jornada laboral. Lo que indica que se corresponde con el consumo de empresas que mantienen los mismo niveles de consumo durante la jornada, como puede ser el caso de oficinas.

Un par de grupos, concretamente el Grupo 1 y el Grupo 9, solamente presentan un pico a lo largo del día, en torno al mediodía. Este comportamiento se asocia con empresas que tienen mayor actividad durante las horas del mediodía, motivo por el cual generan un mayor consumo durante dicha franja horaria, que se puede asociar con cafeterías o pequeños locales que solamente abren por las mañanas los fines de semana.

El resto de los patrones se pueden agrupar en dos categorías, aquellos grupos cuyo consumo se mantiene relativamente constante a lo largo de la jornada laboral (Grupo 3, 6, 8 y 14); y los que presentan dos picos de consumo, uno por la mañana y otro por la tarde (Grupo 11 y 12). Además, el consumo del Grupo 14 acapara el 51.53 %, Tabla 5.52, del consumo total, y es el grupo con el menor consumo. Lo que indica la gran diferencia existente entre las empresas, hay algunas pocas que generan un gran consumo, mientras que otras presentan niveles de consumo similares a los residenciales. Este hecho ya se infería en el diagrama de cajas de la Figura 5.19 de la sección 5.3.

Las distribuidoras eléctricas buscan aplanar los niveles de consumo a lo largo del día para optimizar la producción de las centrales eléctricas. En este caso, con la finalidad de alcanzar este objetivo, se puede lanzar, principalmente, una campaña publicitaria. Esta campaña estará centrada en redistribuir el consumo del Grupo 2, principalmente, para disminuir el parón de consumo que sufre a primeras horas de la mañana. Una forma de conseguir dicho objetivo se puede basar en lanzar ofertas de reducción del coste en determinadas franjas horarias y de esa forma promover el consumo en dichas franjas horarias.

Con el modelo establecido, es necesario evaluar la fiabilidad del mismo. Para ello, se comparan los porcentajes de representación de cada grupo obtenidos en la fase de predicción, Tabla 5.52, con los obtenidos en la fase de validación, Tabla 5.53. En este caso, la variación media entre los grupos es del 0.34 %, lo que es asumible y por tanto el modelo es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	22.77 %	Grupo 6	8.23 %	Grupo 11	2.55 %
Grupo 2	0.07 %	Grupo 7	0.23 %	Grupo 12	0.89 %
Grupo 3	2.19 %	Grupo 8	4.38 %	Grupo 13	0.48 %
Grupo 4	0.11 %	Grupo 9	5.13 %	Grupo 14	51.53 %
Grupo 5	0.43 %	Grupo 10	1.03 %		

Tabla 5.52.: Predicción del porcentaje de datos asociados a cada grupo

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	22.66 %	Grupo 6	8.23 %	Grupo 11	3.90 %
Grupo 2	0.13 %	Grupo 7	0.26 %	Grupo 12	1.52 %
Grupo 3	1.98 %	Grupo 8	4.10 %	Grupo 13	0.53 %
Grupo 4	0.07 %	Grupo 9	6.14 %	Grupo 14	49.93 %
Grupo 5	0.40 %	Grupo 10	0.92 %		

Tabla 5.53.: Predicción del porcentaje de datos asociados a cada grupo

Consumo industrial de un domingo

En primer lugar, es necesario determinar el número de clústers a emplear en la ejecución del algoritmo K-Means. Para ello, se ha calculado la función de coste de la Figura 5.68, que muestra la suma total de las varianzas con respecto a los centroides, en función del número de grupos que se utilizan. Dado que el “codo” de la función no se distingue a simple vista, se ha recurrido al cálculo de la disminución porcentual de la suma de varianzas como medida de referencia para determinar el número de grupos a emplear. Al aumentar el número de clústers de nueve a diez, se obtiene una disminución del 10.36 % de la suma de varianzas, sin embargo, al continuar aumentando el número de clústers, este valor se reduce, por lo que se ha decidido emplear diez clústers.

Una vez escogido el número de clústers a emplear, se ha implementado el algoritmo K-Means, con la finalidad de encontrar los diez grupos que, en principio, definen el consumo industrial de un domingo de otoño en la provincia de Madrid. Los grupos, junto con las desviaciones típicas, se encuentran en la Figura 5.69.

A primera vista, destaca la tendencia seguida por los clientes pertenecientes al Grupo 8. Tratando con consumo industrial, se espera encontrar dos picos de niveles de consumo similares, uno a media mañana y otro a media tarde, correspondiéndose con el punto medio de la jornada de mañana y de la de tarde, como se determinó en la sección 5.3.2. Sin embargo, en este caso, el consumo sufre una gran depresión desde las 6:00 hasta las 12:00 aproximadamente, siendo el consumo prácticamente constante el resto del día. Este tipo de comportamiento se puede asociar a establecimientos que cierran por la mañanas, alargando su jornada laboral hasta la madrugada.

Este puede ser el caso de bares. El Grupo 8 tan solo representa el 0.07 %, Tabla 5.54, del consumo industrial de un domingo de otoño en la provincia de Madrid.

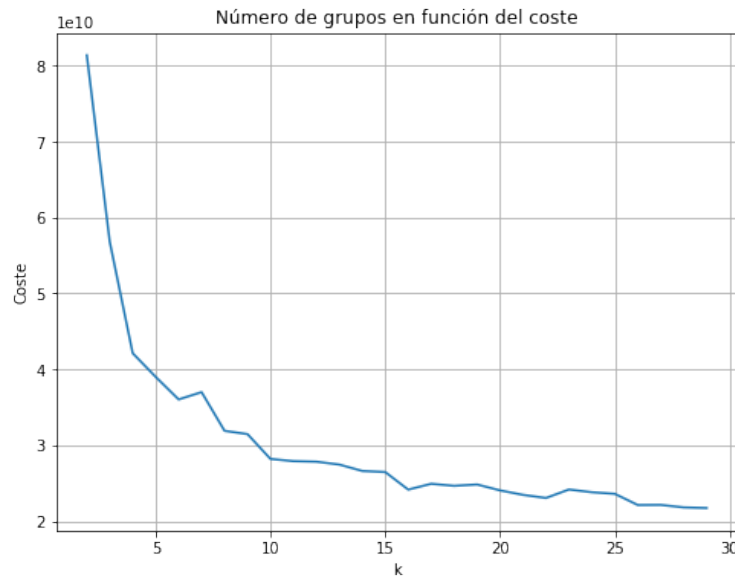


Figura 5.68.: Número de grupos en función de la suma de las distancias a los centroides

También cabe destacar la tendencia del Grupo 2 y del Grupo 6, cuyo consumo presenta una progresión ascendente a lo largo de la jornada laboral, asemejándose a la tendencia esperada de un consumo residencial. Esto implica que son empresas cuya actividad industrial aumenta a la hora de la cena, por lo que puede ser el consumo de restaurantes, que presentan mayor afluencia de clientes a la hora de la cena.

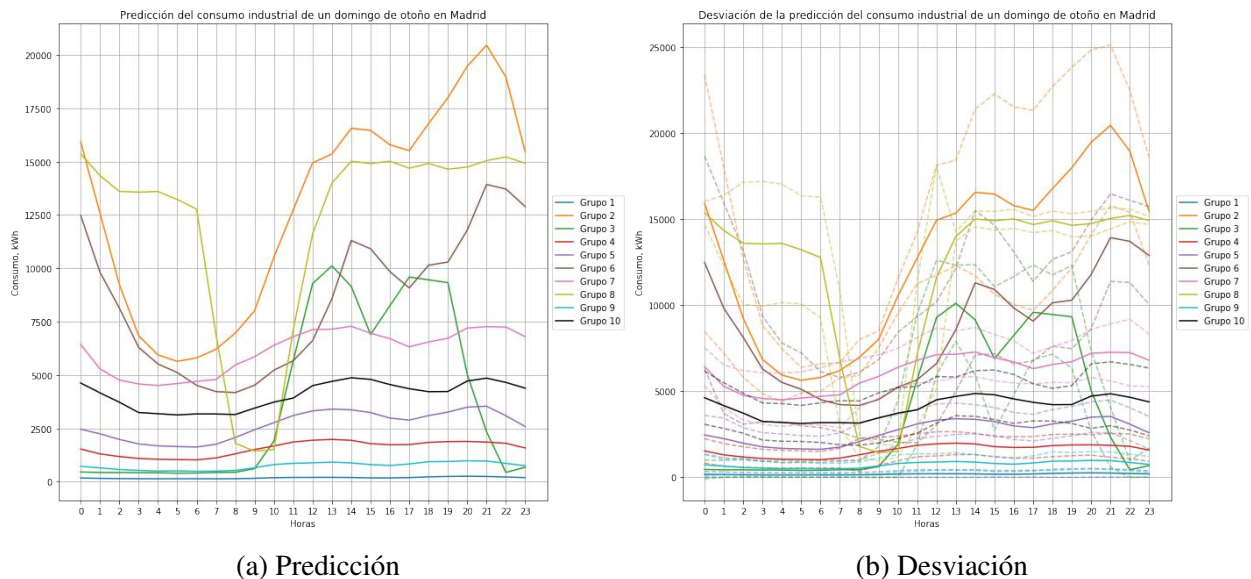


Figura 5.69.: Consumo industrial de un domingo de otoño en Madrid

El resto de los patrones se pueden agrupar en dos categorías, aquellos grupos cuyo consumo se mantiene relativamente constante a lo largo de la jornada laboral (Grupo 1, 4, 5, 7, 9 y 10); y los que presentan dos picos de consumo, uno por la mañana y otro por la tarde (Grupo 3). Además, el consumo del Grupo 1 acapara el 66.47 %, Tabla 5.54, del consumo total, y es el grupo con el menor consumo. Lo que indica la gran diferencia existente entre las empresas, hay algunas pocas que generan un gran consumo, mientras que otras presentan niveles de consumo similares a los residenciales.

Uno de los objetivos principales de este proyecto es proporcionar información útil para las distribuidoras de energía eléctrica para que puedan lanzar campañas publicitarias a determinados grupos de clientes. A las distribuidoras energéticas les conviene aplanar los niveles de consumo a lo largo del día para poder optimizar las instalaciones. En este caso en particular, con la finalidad de alcanzar dicho objetivo, se pueden lanzar, principalmente, una campaña publicitaria. El objetivo de dicha campaña será conseguir la redistribución del consumo de los clientes pertenecientes a los grupos 2 y 8, principalmente. Para alcanzar dicho objetivo, se pueden lanzar ofertas de reducción del precio de la energía en aquellas franjas horarias a las que deseamos que se traslade dicho consumo. Por ejemplo, en la caso del Grupo 8, se le puede oferta una reducción del precio entre las 8:00 y las 10:00, franja horaria de consumo mínimo, con la esperanza de que traslade parte de su consumo a dicho periodo de tiempo.

Con las predicciones calculadas, es necesario evaluar la fiabilidad de las mismas, para ello se comparan los porcentajes de los clientes pertenecientes a cada grupos, Tabla 5.54 y Tabla 5.55. De media, un porcentaje presenta una variación igual a 0.26 %, que es asumible, por lo que el modelo es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	66.47 %	Grupo 6	0.29 %
Grupo 2	0.12 %	Grupo 7	0.81 %
Grupo 3	0.29 %	Grupo 8	0.07 %
Grupo 4	8.70 %	Grupo 9	17.11 %
Grupo 5	4.03 %	Grupo 10	2.09 %

Tabla 5.54.: Predicción del porcentaje de datos asociados a cada grupo

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	66.76 %	Grupo 6	0.29 %
Grupo 2	0.22 %	Grupo 7	0.79 %
Grupo 3	0.14 %	Grupo 8	0.07 %
Grupo 4	7.97 %	Grupo 9	16.87 %
Grupo 5	4.98 %	Grupo 10	1.94 %

Tabla 5.55.: Validación del porcentaje de datos asociados a cada grupo

5.4.2. Municipios de la provincia de Barcelona

En este apartado, el estudio se centra en los municipios pertenecientes a la provincia de Barcelona. Es necesario recordar las conclusiones obtenidas en secciones anteriores acerca del consumo de la provincia de Barcelona, ya que esta información será empleada a lo largo de este apartado.

La provincia de Barcelona presenta un consumo medio igual a 7652.41 kWh. A su vez, este consumo, está formado por el generado por dos actividades económicas, la residencial y la industrial. En este caso, el consumo industrial representa un 21.6% del consumo total. Además cabe destacar, que el consumo industrial sufre una gran caída durante los fines de semana, como se puede apreciar en la figura 5.14. También es necesario tener en cuenta que un 77.7% de su consumo se rige por la tarifa regulada, a diferencia de la provincia de Madrid, en la que dominaba la tarifa libre.

5.4.2.1. Perfiles del consumo de un día de invierno en la provincia de Barcelona

El objetivo de esta sección es encontrar los distintos patrones que representan el consumo residencial e industrial de un día de invierno en la provincia de Barcelona. Para alcanzar dicha finalidad, se va a seguir la dinámica empleada en las secciones anteriores, en primer lugar, se obtiene el número de clústers a emplear en cada caso, para después calcular los distintos patrones de consumo mediante la implementación del algoritmo K-Means.

Consumo residencial

A pesar de que el número de clústers empleado en cada tipo de día de la semana es diferente, se ha decidido juntar el estudio de todo el consumo residencial de un día de invierno en la provincia de Barcelona en un solo análisis debido a las similitudes presentes entre los grupos encontrados. Con la finalidad de obtener el número de clústers a emplear en cada tipo de día de la semana, se ha recurrido al cálculo de la disminución porcentual de la suma de varianzas en función del número de grupos utilizados, ya que las funciones de coste, Figura 5.70, calculadas carecen de un “codo” distinguible a primera vista. Tras finalizar el proceso, se ha decidido emplear quince clústers en el consumo de los días laborales, trece en el de los sábados y catorce en el de los domingos.

Una vez escogidos el número de clústers a emplear en cada caso, se ha implementado el algoritmo K-Means, con el objetivo de encontrar los grupos que definen el consumo residencial de un día de invierno en la provincia de Barcelona. Los grupos, junto con las desviaciones típicas, se encuentran en la Figura 5.71. Cabe mencionar, que a pesar de que el número de clústers elegido para los domingos es catorce, tan solo se representan trece grupos. El motivo detrás de la ausencia del Grupo 5, recae en el hecho de que solamente representaba un conjunto de datos, por lo que no aportaba información relevante a la predicción.

A primera vista, en todos los tipos de día, destaca la presencia de un grupo que sobresale por encima del resto, debido a los valores tan elevados que toma a lo largo del día. Es el caso del Grupo 3 en los días laborales y del Grupo 11 en los sábados y los domingos. Además, no sigue la tendencia ascendente a lo largo del día que se espera del consumo residencial, como se estableció en la

sección 5.3.2. Sorprendentemente, encajaría menor con el gráfico esperado del consumo industrial, motivo por el cual representa un porcentaje ínfimo del consumo residencial de un día de invierno en la provincia de Barcelona, Tabla 5.56, Tabla 5.57 y Tabla 5.58. Este tipo de comportamiento se puede dar en viviendas en las que durante las mañanas la limpieza sea realizada por una persona externa, que termina su jornada antes de la hora de comer, y a la hora de cenar el residente vuelve al hogar.

También llama la atención la presencia de un grupo cuyo consumo decae durante las horas de luz, es el caso de los grupos 4 y 14 en los días laborales, los grupos 4 y 6 en el caso de los sábados y los grupos 2 y 10 en el de los domingos. Este patrón es totalmente opuesto al esperado cuando se trata de consumo residencial. Es muy probable que los clientes que pertenecen a estos grupos trabajen en el turno de noche, motivo por el cual la depresión de su consumo se da durante el día.

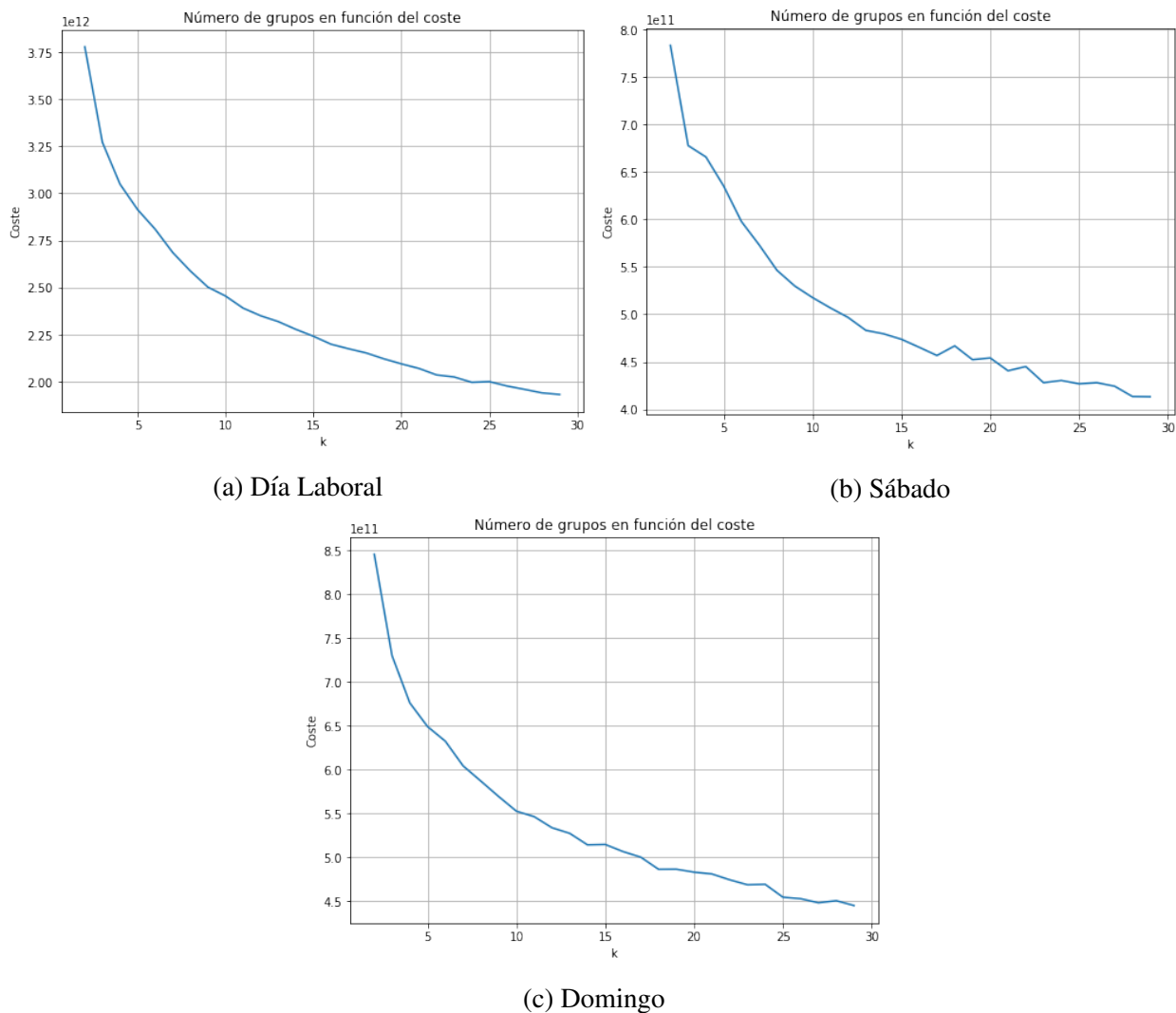


Figura 5.70.: Número de grupos en función de la suma de las distancias a los centroides

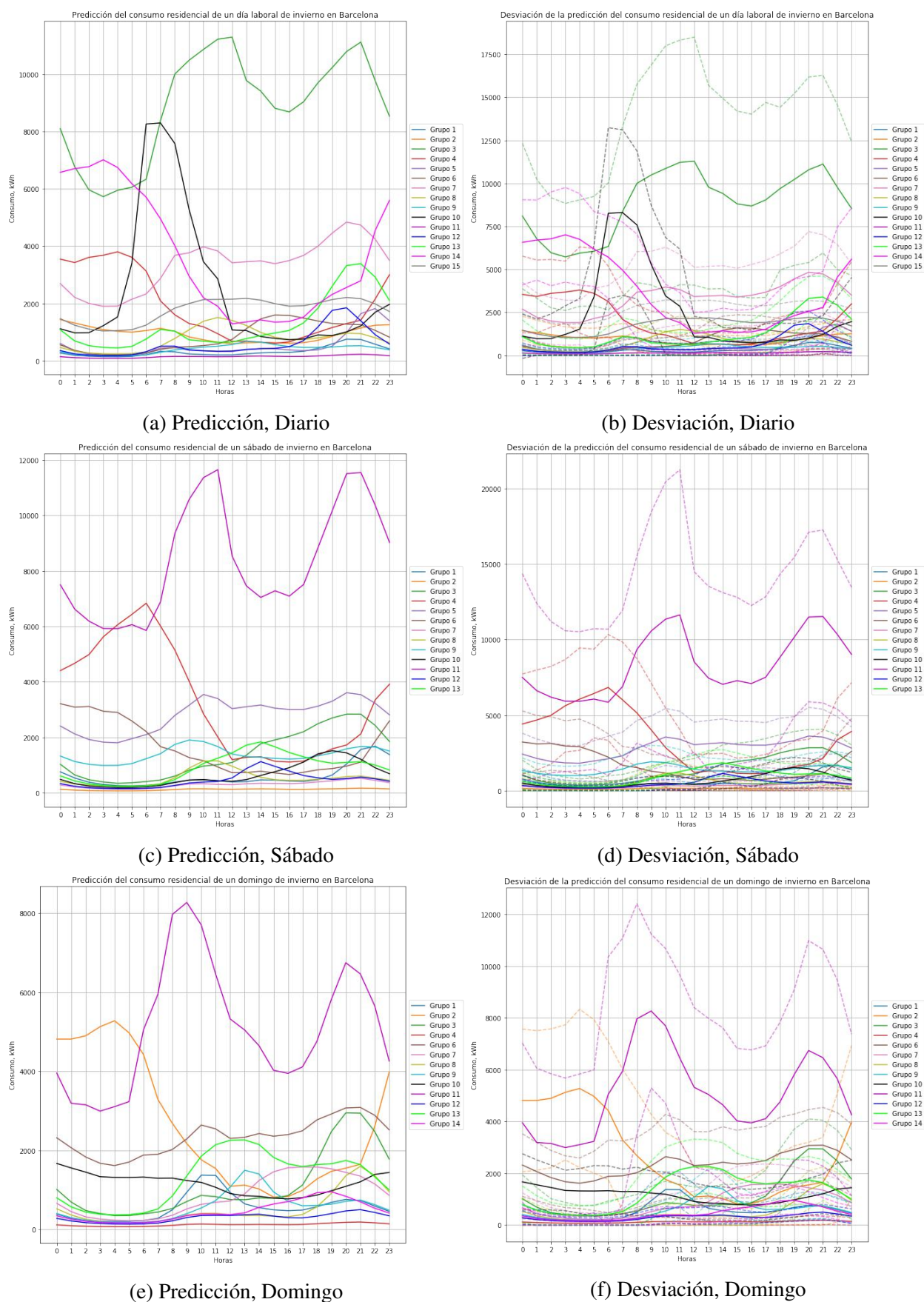


Figura 5.71.: Consumo residencial de un día de invierno en Barcelona

Sorprendentemente, hay un conjunto de grupos, que solamente presentan un pico de consumo a lo largo del día a la hora del mediodía, es el caso de los grupos 8 y 9 en el gráfico de los días laborales, los grupos 8, 12 y 13 en el de los sábados y los grupos 1 y 9 en el de los domingos. Este tipo de comportamiento se asocia con el de aquellas personas que tan solo emplean recursos eléctricos durante la hora de la comida, lo que implica que o bien reducen drásticamente el uso de aparatos eléctricos o emplean otras fuentes de energía en ese espacio de tiempo.

El resto de los grupos representados, reflejan el patrón de consumo residencial esperado, tomando valores mayores o menores y experimentando mayor o menor diferencia entre los dos picos principales del día. Una vez más se da el que grupo con los niveles de consumo menores a lo largo de todo el día, es el que representa un mayor porcentaje del consumo. Es el caso del Grupo 11 en el gráfico de los días laborales, el Grupo 2 en el de los sábados y el Grupo 4 en el de los domingos.

Con la finalidad de aplanar los niveles de consumo a lo largo del día, ya que de esta forma las distribuidoras eléctricas pueden optimizar sus centrales, se pueden lanzar, principalmente, dos campañas publicitarias. Una de ellas se debe centrar en la reducción de los niveles de consumo de aquellos que sobresalen por encima del resto. Por otro lado, se podría buscar la redistribución del consumo de aquellos grupos que siguen una tendencia contraria a la esperada. Esto se puede conseguir mediante el lanzamiento de ofertas de energía en determinadas franjas horarias, medida que motiva la redistribución del consumo. Además, se puede establecer un límite de consumo a partir del cual el precio de la electricidad aumenta para fomentar la reducción del consumo.

Con el modelo establecido, es necesario evaluar la fiabilidad del mismo. Para ello, se comparan los porcentajes de representación de cada grupos obtenidos en la fase de predicción, Tabla 5.56, Tabla 5.57 y Tabla 5.58, con los obtenidos en la fase de validación, Tabla 5.59, Tabla 5.60 y Tabla 5.61. En este caso, la variación es del 0.02 % en los días laborales, del 0.07 % en los sábados y de 0.12 % los domingos, valores que son asumibles, por lo que el modelo calculado es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	19.92 %	Grupo 6	3.29 %	Grupo 11	42.09 %
Grupo 2	2.25 %	Grupo 7	0.22 %	Grupo 12	4.57 %
Grupo 3	0.02 %	Grupo 8	3.68 %	Grupo 13	1.25 %
Grupo 4	0.36 %	Grupo 9	16.48 %	Grupo 14	0.09 %
Grupo 5	4.47 %	Grupo 10	0.06 %	Grupo 15	1.26 %

Tabla 5.56.: Predicción del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	4.97 %	Grupo 6	0.54 %	Grupo 11	0.02 %
Grupo 2	35.37 %	Grupo 7	29.29 %	Grupo 12	9.88 %
Grupo 3	1.37 %	Grupo 8	6.94 %	Grupo 13	3.49 %
Grupo 4	0.16 %	Grupo 9	1.87 %		
Grupo 5	0.51 %	Grupo 10	5.82 %		

Tabla 5.57.: Predicción del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	5.54 %	Grupo 6	0.82 %	Grupo 11	0.07 %
Grupo 2	0.31 %	Grupo 7	3.98 %	Grupo 12	29.28 %
Grupo 3	1.58 %	Grupo 8	5.92 %	Grupo 13	1.84 %
Grupo 4	32.11 %	Grupo 9	6.53 %	Grupo 14	10.08
Grupo 5	0.00 %	Grupo 10	1.96 %		

Tabla 5.58.: Predicción del porcentaje de datos asociados a cada grupo, Domingo

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	19.91 %	Grupo 6	3.31 %	Grupo 11	42.15 %
Grupo 2	2.24 %	Grupo 7	0.22 %	Grupo 12	4.53 %
Grupo 3	0.01 %	Grupo 8	3.69 %	Grupo 13	1.24 %
Grupo 4	0.37 %	Grupo 9	16.45 %	Grupo 14	0.10 %
Grupo 5	4.46 %	Grupo 10	0.06 %	Grupo 15	1.26 %

Tabla 5.59.: Validación del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	4.80 %	Grupo 6	0.52 %	Grupo 11	0.01 %
Grupo 2	35.67 %	Grupo 7	29.19 %	Grupo 12	9.89 %
Grupo 3	1.33 %	Grupo 8	6.98 %	Grupo 13	3.56 %
Grupo 4	0.13 %	Grupo 9	1.81 %		
Grupo 5	0.49 %	Grupo 10	5.63 %		

Tabla 5.60.: Validación del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	5.53 %	Grupo 6	0.77 %	Grupo 11	0.05 %
Grupo 2	0.33 %	Grupo 7	3.98 %	Grupo 12	29.57 %
Grupo 3	1.47 %	Grupo 8	6.09 %	Grupo 13	1.83 %
Grupo 4	31.51 %	Grupo 9	6.51 %	Grupo 14	10.34
Grupo 5	0.00 %	Grupo 10	2.02 %		

Tabla 5.61.: Validación del porcentaje de datos asociados a cada grupo, Domingo

Consumo industrial

Una vez más, a pesar de que el número de clústers empleado en cada tipo de día de la semana es diferente, se ha decidido juntar el estudio de todo el consumo industrial de un día de invierno en la provincia de Barcelona, en un sólo análisis debido a las similitudes presentes entre los grupos encontrados. Con el objetivo de concretar el número de clústers a emplear en cada tipo de día de la semana, se ha recurrido al cálculo de la disminución porcentual de la suma de varianzas en función del número de grupos utilizados, ya que las funciones de coste calculadas, Figura 5.72, carecen de “codo” discernible a primera vista. Tras completar el proceso, se ha decidido emplear diez clúster en el caso de los días laborales, doce en los sábados y diez en los domingos.

Con el número de clústers elegido para cada tipo de día de la semana, se ha podido implementar el algoritmo K-Means, con la finalidad de obtener los distintos grupos de clientes que representan el consumo industrial a lo largo de un día de invierno en la provincia de Barcelona. Los grupos, junto con sus respectivas desviaciones típicas, se encuentran en la Figura 5.73. Cabe mencionar, que a pesar de que el número de clústers elegido para los sábados es doce, tan solo se representan once grupos, falta el Grupo 7. Esto se debe a que dicho grupo solamente representa un conjunto de datos, por lo que no aporta información relevante a la predicción.

Uno de los primeros grupos que llaman la atención al observar los gráficos de la Figura 5.73, es el Grupo 4 de los días laborales, por poseer los niveles de consumo más elevados. Tratando con consumo de tipo industrial, se esperan niveles de consumo elevados, como ya se había observado en la sección 5.3 y como se demuestra si se comparan los valores de consumo de los grupos de consumo industrial, Figura 5.73, con los niveles del consumo residencial, Figura 5.71. Sin embargo, estos los niveles presentados por el Grupo 4 no son los habituales ya que solamente se da en el 0.002 %, Tabla 5.62 del consumo industrial de un día de invierno en la provincia de Barcelona. Además, este grupo resalta por la tendencia descendente que presenta desde la madrugada hasta las 17:00, exceptuando el pico de consumo de las 7:00. Este tipo de comportamiento se podría corresponder con el de locales que abren a media tarde y continúan trabajando hasta las primeras horas de la mañana, como puede ser el caso de establecimiento de comida que se centren en dar servicio de cena.

Un conjunto de grupos que llama la atención es el formado por aquellos que presentan, aproximadamente dos niveles de consumo. Los máximos de dicho consumo se dan desde aproximadamente las 17:00 hasta las 7:00. Este conjunto de grupos está formado por los grupos 7 y 9 en

el gráfico de días laborales, por los grupos 3 y 5 en los sábados y por los grupos 2, 5 y 7 en los domingos. Este tipo de comportamiento se puede asociar al seguido por locales que comienzan su actividad a media tarde y continúan su actividad durante toda la noche. Sin embargo, este patrón es seguido por un porcentaje mínimo de clientes, Tabla 5.62, Tabla 5.63, Tabla 5.64. Hay otros dos grupos que siguen, de forma más sutil, un patrón similar, es el caso del Grupo 6 de los sábados y del Grupo 10 de los domingos.

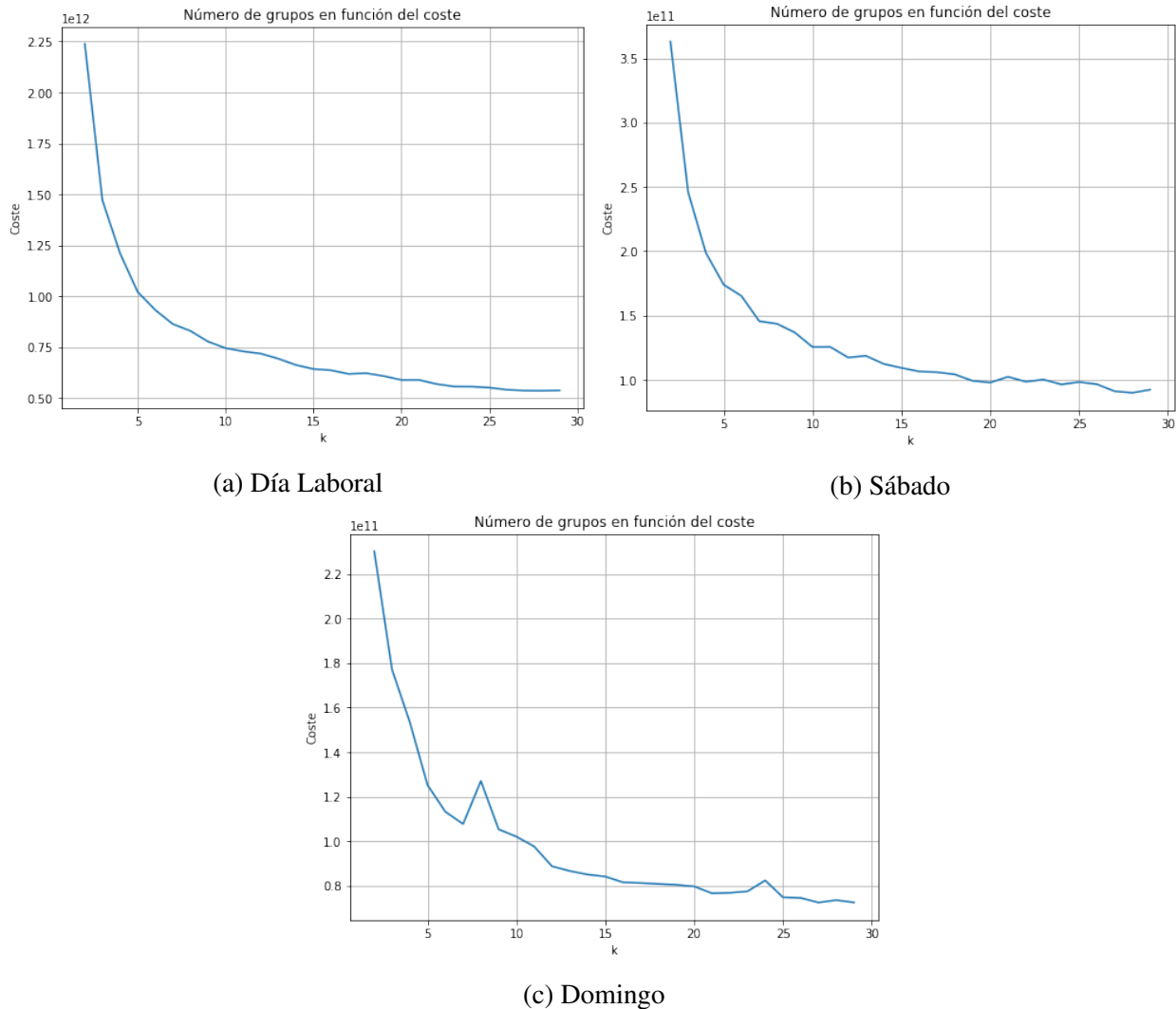
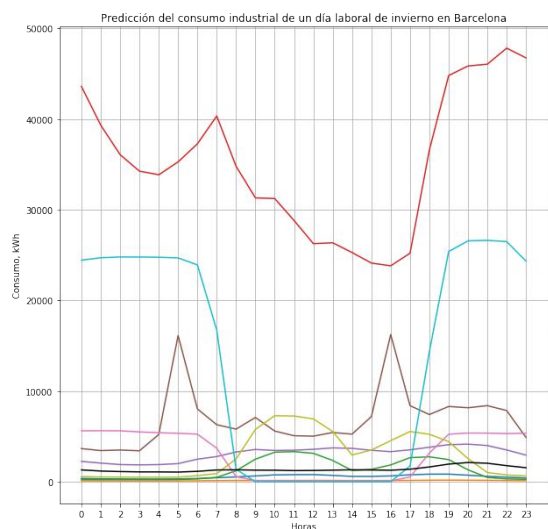
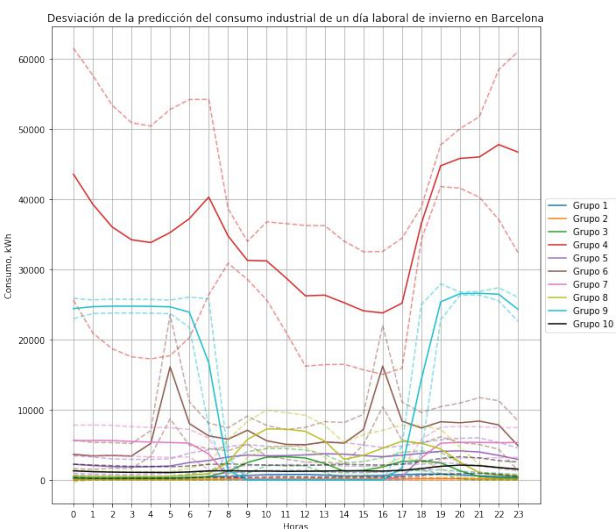


Figura 5.72.: Número de grupos en función de la suma de las distancias a los centroides

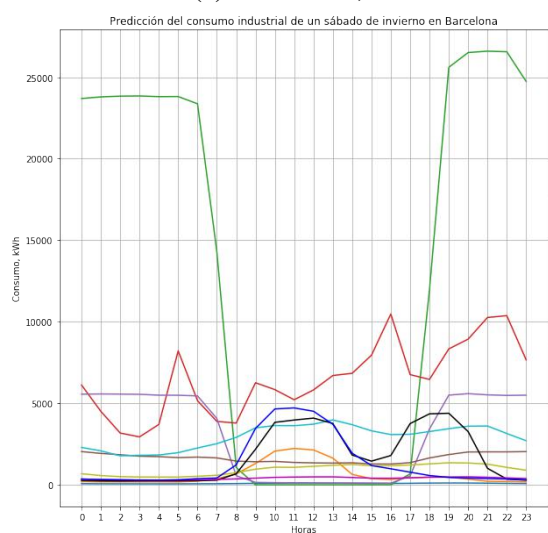
También llama la atención la presencia de unos grupos que poseen dos picos muy claros, uno por la mañana y otro por la tarde, que solamente se prolongan durante un par de horas, mientras que durante el resto del día el consumo es prácticamente constante. Es el caso del Grupo 6, en el caso del consumo de un día laboral, del Grupo 4, aunque con una tendencia más ascendente, en el caso de los sábados y del Grupo 11 en el caso de los domingos. Lo más extraño es que estos picos se dan sobre las 5:00 y las 16:00. Sin embargo el porcentaje de consumo que se ve representado por dichos patrones es mínimo, Tabla 5.62, Tabla 5.63, Tabla 5.64.



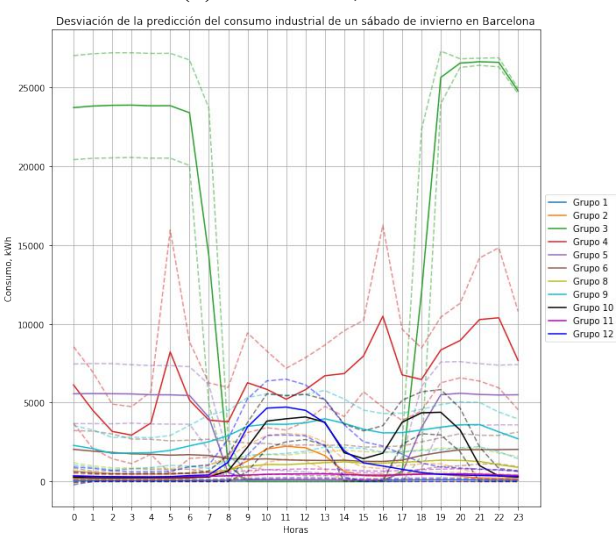
(a) Predicción, Diario



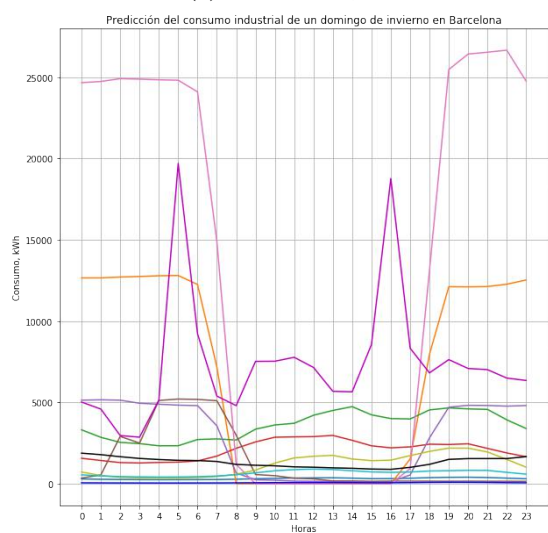
(b) Desviación, Diario



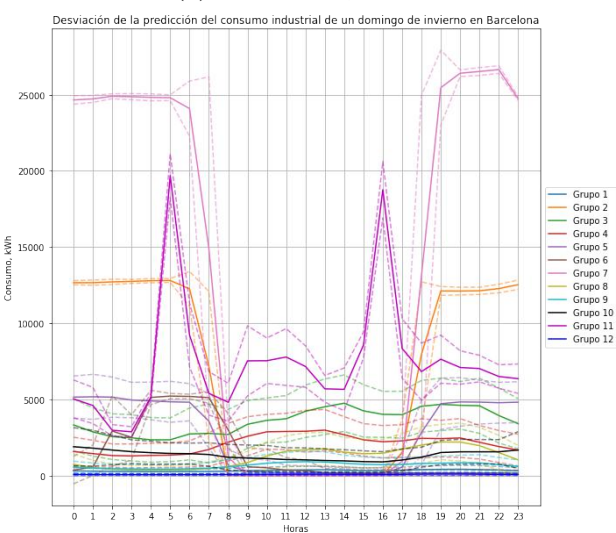
(c) Predicción, Sábado



(d) Desviación, Sábado



(e) Predicción, Domingo



(f) Desviación, Domingo

Figura 5.73.: Consumo industrial de un día de invierno en Barcelona

En el gráfico de los sábados, Figura 5.73, se pueden observar algunos grupos que tan solo presentan un pico a lo largo del día, durante el horario de mañana. Se trata de los grupos 2 y 12. Este tipo de comportamiento se puede asociar al seguido por locales que tan solo ofrecen sus servicios por las mañanas los fines de semana, como puede ser el caso de quioscos.

Algunos de los grupos calculados, presentan una tendencia muy similar a la encontrada para el consumo industrial en la sección 5.3.2. El patrón consumo que caracteriza al consumo industrial presenta dos picos claros de consumo, uno por la mañana y otro por la tarde. Es el caso de los grupos 3 y 8 de los días laborales, del Grupo 10 de los sábados y el Grupo 7 de los domingos. Además, cabe mencionar el resto de grupos que presentan una tendencia más constante a lo largo del día. En estos grupos se puede inferir una pequeña subida del consumo a las horas donde deberían encontrar los picos principales, es muy probable que no se puedan distinguir por la escala de la imagen, pero que si que sigan dicha tendencia. Una vez más, los grupos con los niveles de consumo más bajos, son los que representan un porcentaje mayor del consumo industrial de un día de invierno en la provincia de Barcelona; es el caso del Grupo 2 en el gráfico de los días laborales, del Grupo 1 en el de los sábados y del Grupo 12 en el de los domingos.

Uno de los objetivos principales de este proyecto es proporcionar información útil para las distribuidoras de energía eléctrica para que puedan lanzar campañas de publicidad orientadas a determinados grupos de clientes. A las distribuidoras energéticas les conviene aplanar los niveles de consumo a lo largo del día para poder optimizar la producción de las centrales. En este caso, principalmente, se pueden lanzar dos campañas publicitarias. Una de ellas tendría como objetivo reducir los niveles de consumo de los clientes pertenecientes al Grupo 4 de los días laborales, ya que presentan unos niveles de consumo muy superiores al resto. Este objetivo se podría conseguir mediante la subida del precio de la energía una vez sobrepasado cierto nivel de consumo. Por otro lado, una segunda campaña se podría centrar en la redistribución del consumo de aquellos grupos que presentan una gran depresión entre las 8:00 y las 17:00. Para conseguir dicha redistribución, se les podrían lanzar ofertas en las que se indicasen determinadas franjas horarias en las que el precio fuese más reducido, con la esperanza de que desplazasen su consumo a dichas franjas.

Con el modelo predictivo establecido, es necesario evaluar la veracidad del mismo. Motivo por el cual se comparan los porcentajes de representación de cada grupo obtenidos en la fase de predicción, Tabla 5.62, Tabla 5.63, Tabla 5.64, con los obtenidos en la fase de validación, Tabla 5.65, Tabla 5.66, Tabla 5.67. En este caso, el resultado de la comparación es una variación media por cada grupo del 0.08 % en los días laborales, del 0.09 % en los sábados y del 0.12 % en los domingos. Estos valores son perfectamente asumibles, por lo que el modelo calculado es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	23.78 %	Grupo 6	0.04 %
Grupo 2	65.12 %	Grupo 7	0.61 %
Grupo 3	4.11 %	Grupo 8	0.71 %
Grupo 4	0.002 %	Grupo 9	0.03 %
Grupo 5	0.79 %	Grupo 10	3.81 %

Tabla 5.62.: Predicción del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	56.48 %	Grupo 5	0.58 %	Grupo 9	0.77 %
Grupo 2	3.26 %	Grupo 6	1.72 %	Grupo 10	0.80 %
Grupo 3	0.04 %	Grupo 7	0.003 %	Grupo 11	29.09 %
Grupo 4	0.11 %	Grupo 8	6.18 %	Grupo 12	0.96 %

Tabla 5.63.: Predicción del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	28.78 %	Grupo 5	0.64 %	Grupo 9	9.98 %
Grupo 2	0.04 %	Grupo 6	0.04 %	Grupo 10	1.69 %
Grupo 3	0.36 %	Grupo 7	0.02 %	Grupo 11	0.03 %
Grupo 4	1.15 %	Grupo 8	2.21 %	Grupo 12	55.06 %

Tabla 5.64.: Predicción del porcentaje de datos asociados a cada grupo, Domingo

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	24.01 %	Grupo 6	0.05 %
Grupo 2	65.80 %	Grupo 7	0.62 %
Grupo 3	4.06 %	Grupo 8	0.74 %
Grupo 4	0.002 %	Grupo 9	0.04 %
Grupo 5	0.77 %	Grupo 10	3.89 %

Tabla 5.65.: Validación del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	56.61 %	Grupo 5	0.56 %	Grupo 9	0.79 %
Grupo 2	3.13 %	Grupo 6	1.75 %	Grupo 10	0.66 %
Grupo 3	0.02 %	Grupo 7	0.00 %	Grupo 11	29.31 %
Grupo 4	0.09 %	Grupo 8	5.95 %	Grupo 12	1.13 %

Tabla 5.66.: Validación del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	28.99 %	Grupo 5	0.58 %	Grupo 9	10.08 %
Grupo 2	0.01 %	Grupo 6	0.03 %	Grupo 10	1.48 %
Grupo 3	0.52 %	Grupo 7	0.04 %	Grupo 11	0.04 %
Grupo 4	1.24 %	Grupo 8	2.36 %	Grupo 12	54.61 %

Tabla 5.67.: Validación del porcentaje de datos asociados a cada grupo, Domingo

5.4.2.2. Perfiles del consumo de un día de primavera en la provincia de Barcelona

El objetivo de esta sección es encontrar los distintos patrones que representan el consumo residencial e industrial de un día de primavera en la provincia de Barcelona. Para alcanzar dicha finalidad, se va a seguir la dinámica empleada en las secciones anteriores, en primer lugar, se obtiene el número de clústers a emplear en cada caso, para después calcular los distintos patrones de consumo mediante la implementación del algoritmo K-Means.

Consumo residencial

Sorprendentemente, en este caso, el número de clústers elegido para todos los días de la semana es el mismo. Con el objetivo de encontrar el número de clústers a emplear en cada tipo de día de la semana, ya que el “codo” de las funciones de coste, Figura 5.74 es difícil de discernir a simple vista, se ha recurrido al cálculo del porcentaje de la disminución de la suma de varianzas con cada aumento del número de clústers. Tras realizar esa operación para cada uno de los tres casos, se ha llegado a la conclusión de que el número de clústers a emplear es quince.

Una vez elegido el número de clústers a emplear, se ha implementado el algoritmo K-Means con la finalidad de encontrar quince grupos que definen el consumo residencial de un día de primavera en la provincia de Barcelona. Los grupos, junto con sus desviaciones típicas, se encuentran en la Figura 5.75. A pesar de que el número de clústers elegido es quince, los gráficos de los días laborales y los domingos tan solo reflejan información relativa a catorce grupos. Esto se debe a la eliminación de un par de grupos que tan solo recogían información acerca de un conjunto de datos, por lo que no aportaban información útil para este estudio.

De entre los grupos calculados para cada tipo de día de la semana, salta a la vista la tendencia seguida por el Grupo 3 de los domingos, Figura 5.75, debido a sus elevados niveles de consumo. Además, llama la atención el hecho de que presenta una tendencia que no es la esperada para un consumo de tipo residencial, como se estableció en la sección 5.3.2. El patrón esperado en un consumo residencial presenta una tendencia ascendente a lo largo de todas las horas de luz, poseyendo un pico relativo en torno a la hora de comer, mientras que el consumo desciende a lo largo de la noche. Sin embargo, en este caso, la subida solamente se produce a partir de las 12:00, aunque sí que mantiene el pico relativo en torno a la hora de la comida. Este tendencia tan sol representa el 0.003 %, Tabla 5.68, del consumo residencial de un domingo de primavera en la provincia de Barcelona.

Siguiendo la tendencia esperada del consumo residencial de forma muy clara se encuentra el Grupo 15 en los días laborales, el Grupo 11 en los sábados y el Grupo 11 en los domingos. A pesar de que reflejan la tendencia esperada para un consumo de tipo residencial, cada uno de ellos abarca un porcentaje mínimo del consumo residencial de un día de primavera en la provincia de Barcelona, Tabla 5.68, Tabla 5.69 y Tabla 5.70, debido a los elevados valores que presenta.

También llama la atención la presencia de un grupo cuyo consumo decae durante las horas de luz, es el caso del Grupo 2 en los días laborales, el Grupo 10 en el caso de los sábados y el Grupo 4 en el de los domingos. Este patrón es totalmente opuesto al esperado cuando se trata de consumo

residencial. Es muy probable que los clientes que pertenecen a estos grupos trabajen en el turno de noche, motivo por el cual la depresión de su consumo se da durante el día.

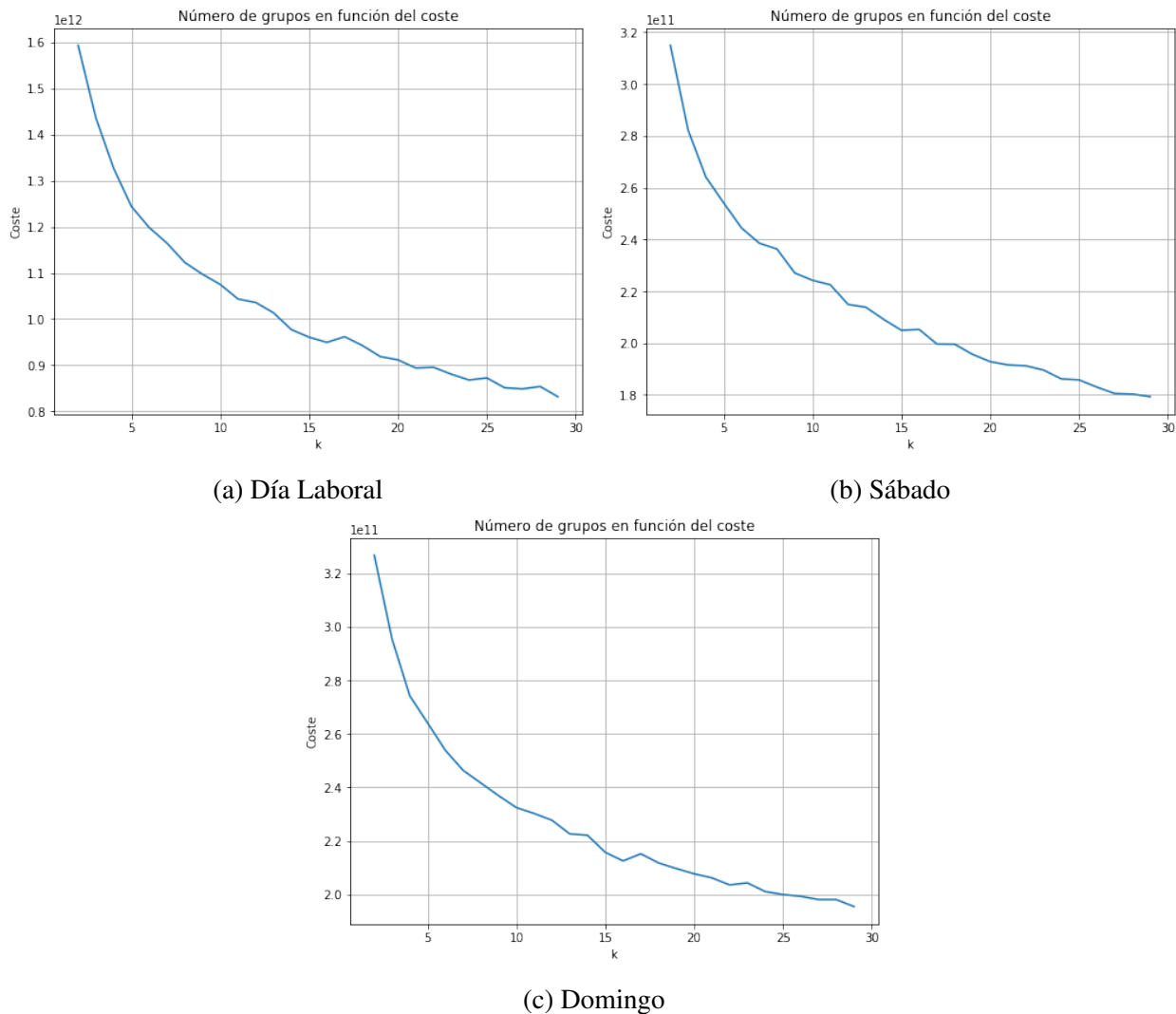


Figura 5.74.: Número de grupos en función de la suma de las distancias a los centroides

Además, hay un conjunto de grupos, que solamente presentan un pico de consumo a lo largo del día a la hora del mediodía, es el caso de los grupos 4, 5, 10 y 11 en el gráfico de los días laborales, los grupos 2, 3, 9, 12 y 13 en el de los sábados y los grupos 2, 9 y 14 en el de los domingos. Este tipo de comportamiento se asocia con el de aquellas personas que tan solo emplean recursos eléctricos durante la hora de la comida, lo que implica que o bien reducen drásticamente el uso de aparatos eléctricos o emplean otras fuentes de energía en ese espacio de tiempo.

El resto de los grupos representados, reflejan el patrón de consumo residencial esperado, tomando valores mayores o menores y experimentando mayor o menor diferencia entre los dos picos principales del día. Una vez más se da el que grupo con los niveles de consumo menores a lo largo de todo el día, es el que representa un mayor porcentaje del consumo.

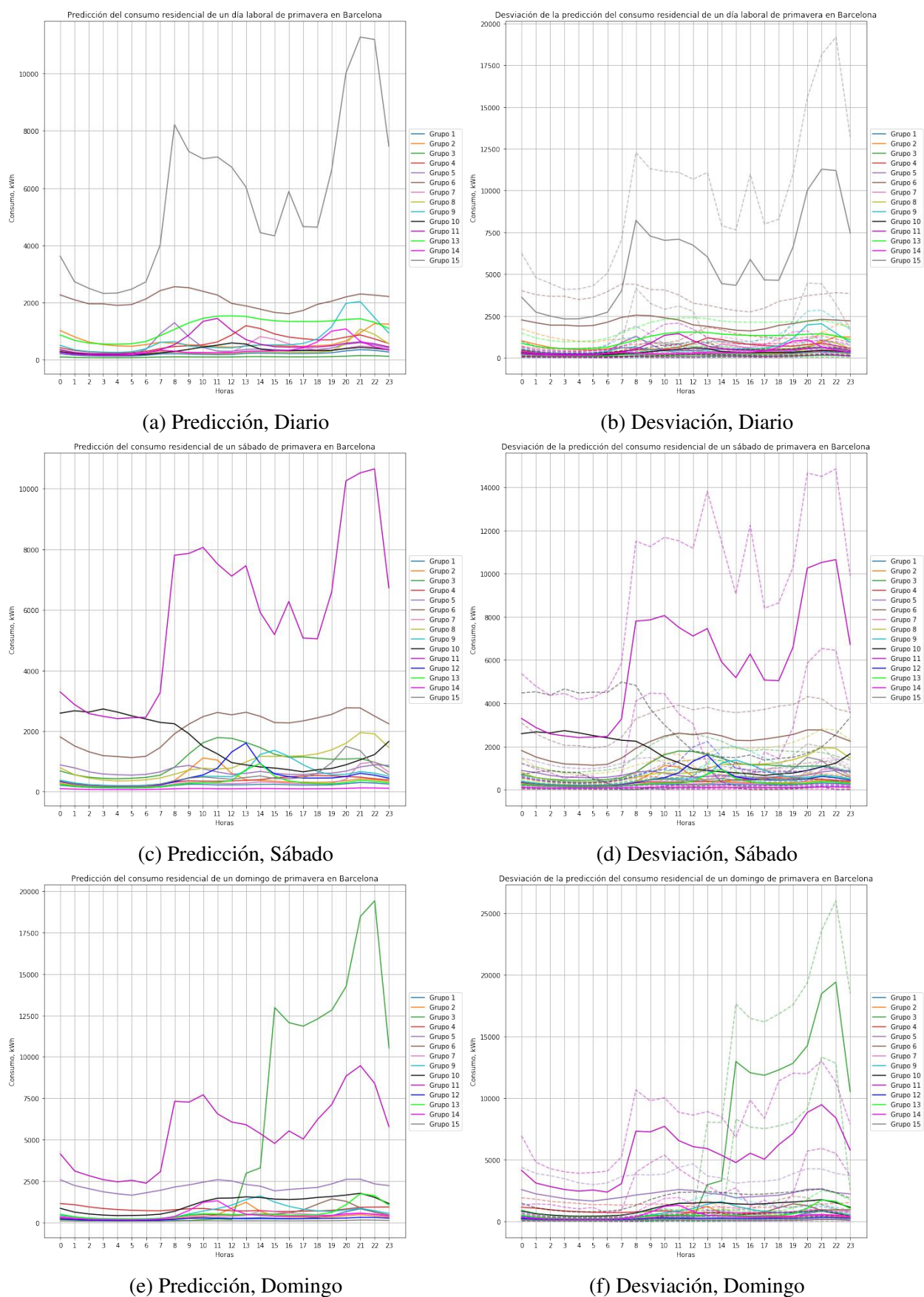


Figura 5.75.: Consumo residencial de un día de primavera en Barcelona

Con la finalidad de aplanar los niveles de consumo a lo largo del día, ya que de esta forma las distribuidoras eléctricas pueden optimizar sus centrales, se pueden lanzar, principalmente, dos campañas publicitarias. Una de ellas se debe centrar en la reducción de los niveles de consumo de aquellos que sobresalen por encima del resto. Por otro lado, se podría buscar la redistribución del consumo de aquellos grupos que siguen una tendencia contraria a la esperada. Esto se puede conseguir mediante el lanzamiento de ofertas de energía en determinadas franjas horarias, medida que motiva la redistribución del consumo. Además, se puede establecer un límite de consumo a partir del cual el precio de la electricidad aumenta para fomentar la reducción del consumo.

Con el modelo establecido, es necesario evaluar la fiabilidad del mismo. Para ello, se comparan los porcentajes de representación de cada grupos obtenidos en la fase de predicción, Tabla 5.68, Tabla 5.69 y Tabla 5.70, con los obtenidos en la fase de validación, Tabla 5.71, Tabla 5.72 y Tabla 5.73. En este caso, la variación es del 0.32 % en los días laborales, del 0.04 % en los sábados y de 0.08 % los domingos, valores que son asumibles, por lo que el modelo calculado es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	31.10 %	Grupo 6	0.19 %	Grupo 11	1.92 %
Grupo 2	1.72 %	Grupo 7	5.94 %	Grupo 12	0.001 %
Grupo 3	29.94 %	Grupo 8	7.04 %	Grupo 13	0.91 %
Grupo 4	2.97 %	Grupo 9	1.39 %	Grupo 14	4.77 %
Grupo 5	2.30 %	Grupo 10	9.79 %	Grupo 15	0.02 %

Tabla 5.68.: Predicción del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	29.86 %	Grupo 6	0.17 %	Grupo 11	0.02 %
Grupo 2	4.20 %	Grupo 7	5.62 %	Grupo 12	2.83 %
Grupo 3	1.00 %	Grupo 8	0.93 %	Grupo 13	6.90 %
Grupo 4	9.90 %	Grupo 9	2.81 %	Grupo 14	31.12 %
Grupo 5	1.98 %	Grupo 10	0.14 %	Grupo 15	2.52 %

Tabla 5.69.: Predicción del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	8.68 %	Grupo 6	2.63 %	Grupo 11	0.02 %
Grupo 2	5.32 %	Grupo 7	10.32 %	Grupo 12	29.34 %
Grupo 3	0.003 %	Grupo 8	0.002 %	Grupo 13	2.18 %
Grupo 4	1.24 %	Grupo 9	2.66 %	Grupo 14	3.84 %
Grupo 5	0.17 %	Grupo 10	1.07 %	Grupo 15	32.53 %

Tabla 5.70.: Predicción del porcentaje de datos asociados a cada grupo, Domingo

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	31.06 %	Grupo 6	0.20 %	Grupo 11	1.94 %
Grupo 2	1.72 %	Grupo 7	5.93 %	Grupo 12	0.00 %
Grupo 3	29.96 %	Grupo 8	6.93 %	Grupo 13	0.89 %
Grupo 4	3.00 %	Grupo 9	1.42 %	Grupo 14	4.75 %
Grupo 5	2.26 %	Grupo 10	9.94 %	Grupo 15	0.01 %

Tabla 5.71.: Validación del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	29.92 %	Grupo 6	0.21 %	Grupo 11	0.01 %
Grupo 2	4.21 %	Grupo 7	5.58 %	Grupo 12	2.81 %
Grupo 3	0.98 %	Grupo 8	0.93 %	Grupo 13	6.97 %
Grupo 4	9.98 %	Grupo 9	2.71 %	Grupo 14	31.14 %
Grupo 5	1.89 %	Grupo 10	0.11 %	Grupo 15	2.55 %

Tabla 5.72.: Validación del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	8.71 %	Grupo 6	2.56 %	Grupo 11	0.02 %
Grupo 2	5.33 %	Grupo 7	10.35 %	Grupo 12	29.73 %
Grupo 3	0.002 %	Grupo 8	0.00 %	Grupo 13	21.7 %
Grupo 4	1.25 %	Grupo 9	2.78 %	Grupo 14	3.70 %
Grupo 5	0.16 %	Grupo 10	0.99 %	Grupo 15	32.25 %

Tabla 5.73.: Validación del porcentaje de datos asociados a cada grupo, Domingo

Consumo industrial

En este caso, el número de clústers elegido para todos los días de la semana es el mismo. Con el objetivo de encontrar el número de clústers a emplear en cada tipo de día de la semana, ya que el “codo” de las funciones de coste, Figura 5.76, es difícil de discernir a simple vista, se ha recurrido al cálculo del porcentaje de la disminución de la suma de varianzas con cada aumento del número de clústers. Tras realizar esa operación para cada uno de los tres casos, se ha llegado a la conclusión de que el número de clústers a emplear es quince.

Una vez elegido el número de clústers a emplear, se ha implementado el algoritmo K-Means con la finalidad de encontrar quince grupos que definen el consumo industrial de un día de primavera en la provincia de Barcelona. Los grupos, junto con sus desviaciones típicas, se encuentran en la Figura 5.77. A pesar de que el número de clústers elegido es quince, los gráficos de los días laborales tan solo reflejan información realativa a catorce grupos. Lo cual se debe a que cierto grupo

solamente representaba información acerca de un conjunto de datos, por lo que no proporcionaba información útil para este estudio.

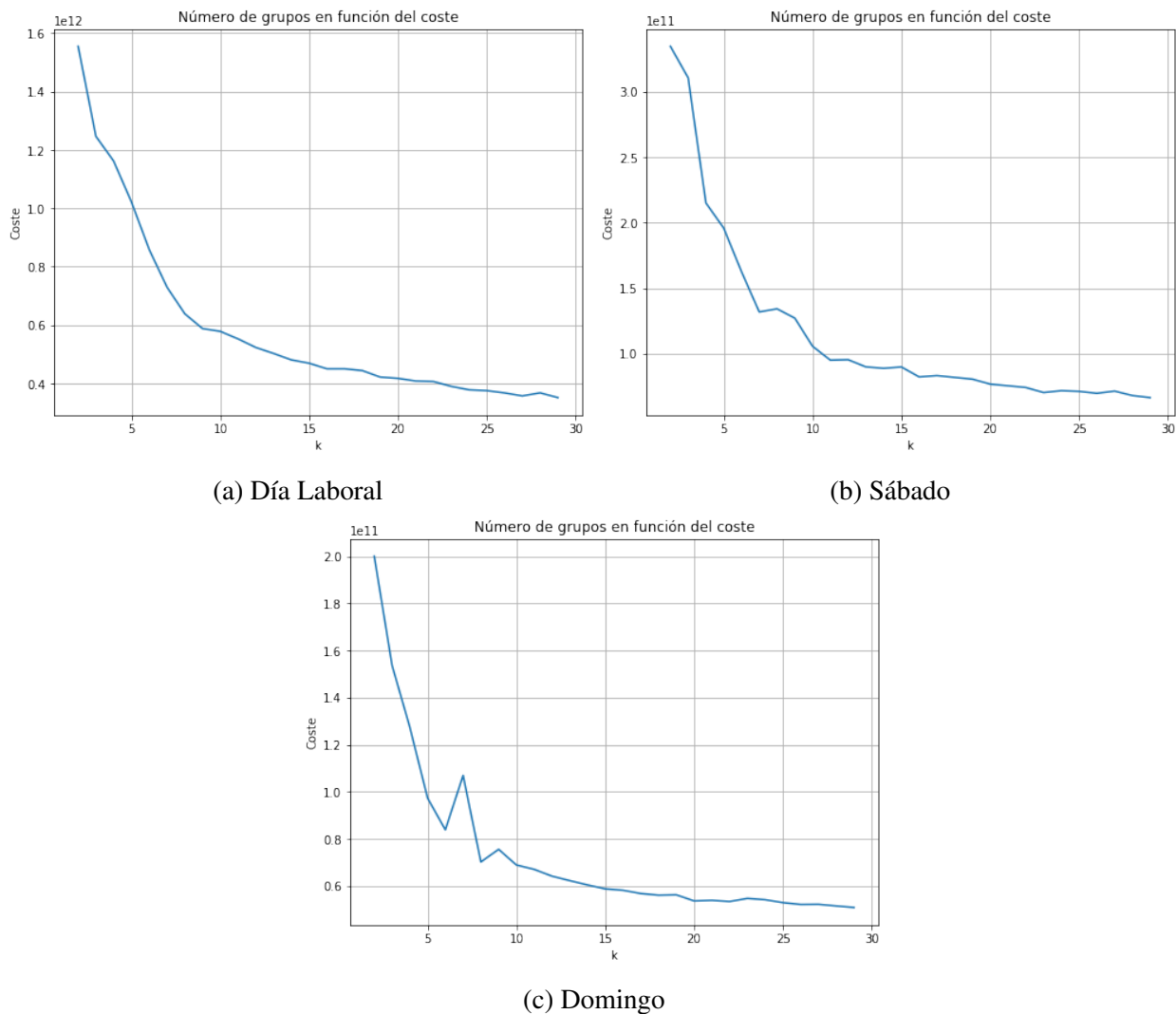


Figura 5.76.: Número de grupos en función de la suma de las distancias a los centroides

Un conjunto de grupos que llama la atención es el formado por aquellos que presentan, aproximadamente dos niveles de consumo. Los máximos de dicho consumo se dan desde aproximadamente las 21:00 hasta las 7:00. Este conjunto de grupos está formado por los grupos 7, 9 y 11 en el gráfico de días laborales, por los grupos 2, 5, 9, 13 y 14 en los sábados y por los grupos 2, 4, 5 y 9 en los domingos. Este tipo de comportamiento se puede asociar al seguido por locales que comienzan su actividad sobre las 21:00 y continúan su actividad durante toda la noche. Sin embargo, este patrón es seguido por un porcentaje mínimo de clientes, Tabla 5.74, Tabla 5.75, Tabla 5.76. Patrones similares ya se había detectado en el caso del consumo industrial de un día de invierno en la provincia de Barcelona, resolviendo la principal diferencia en el hecho de que en el consumo de invierno la etapa de consumo máximo empezaba a las 17:00 en vez de a las 21:00.

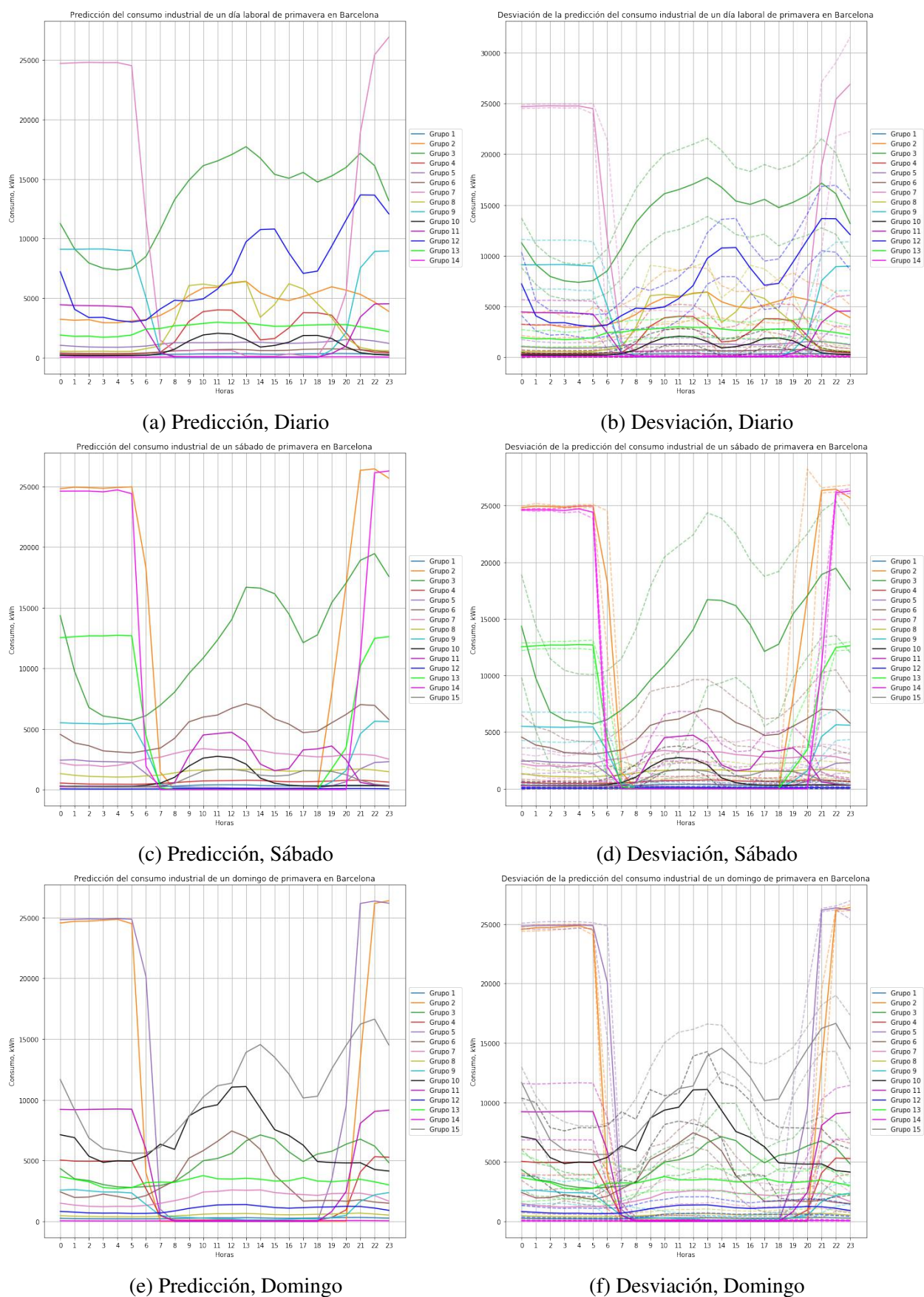


Figura 5.77.: Consumo industrial de un día de primavera en Barcelona

En cada uno de los tipos de día de la semana destaca la presencia de grupos cuyo patrón asemeja al esperado si se tratase de consumo residencial, sección 5.3.2. Es el caso del Grupo 12 en los días laborales, del Grupo 3 en los sábados y del Grupo 15 en los domingos. Dicho consumo presenta una progresión ascendente a lo largo de la jornada laboral, asemejándose a la tendencia esperada de un consumo residencial. Esto implica que son empresas cuya actividad industrial aumenta a la hora de la cena, por lo que puede ser el consumo de restaurantes, que presentan mayor afluencia de clientes a la hora de la cena. Patrones similares fueron encontrados en el consumo industrial de un día de otoño en la provincia de Madrid.

Durante el fin de semana, se ha detectado la presencia de un conjunto de grupos que tan solo presentan un pico de consumo al día, en vez de los dos picos que se esperan cuando se trata de consumo industrial. Es el caso del Grupo 10 de los sábados y de los grupos 6 y 10 de los domingos. Este tipo de comportamiento se asocia con el de aquellas empresas que tan solo ofrecen sus servicios en horario de mañana; siendo una minoría, como se puede observar en la Tabla 5.75 y Tabla 5.76.

El resto de los patrones se pueden agrupar en dos categorías, aquellos grupos cuyo consumo se mantiene relativamente constante a lo largo de la jornada laboral y los que presentan dos picos de consumo, uno por la mañana y otro por la tarde. Una vez más, los grupos con los niveles de consumo menores son los que abarcan un mayor porcentaje del consumo objeto de estudio, en este caso es el Grupo 14 en los días laborales, el Grupo 12 en los sábados y el Grupo 14 los domingos. Lo que refleja la gran diferencia existente entre las empresas, hay algunas pocas que generan un gran consumo, mientras que otras presentan niveles de consumo similares a los residenciales.

Uno de los objetivos principales de este proyecto es proporcionar información útil para las distribuidoras de energía eléctrica para que puedan lanzar campañas de publicidad orientadas a determinados grupos de clientes. A las distribuidoras energéticas les conviene aplanar los niveles de consumo a lo largo del día para poder optimizar la producción de las centrales. En este caso, principalmente, se pueden lanzar una campañas publicitarias. Tendría como objetivo conseguir la redistribución del consumo de aquellos grupos que presentan una gran depresión entre las 7:00 y las 21:00. Para conseguir dicha redistribución, se les podrían lanzar ofertas en las que se indicasen determinadas franjas horarias en las que el precio fuese más reducido, con la esperanza de que desplazasen su consumo a dichas franjas.

Con las predicciones calculadas, es necesario evaluar la fiabilidad de las mismas, para ello se comparan los porcentajes de los clientes pertenecientes a cada grupos. Para ello se calcula la variación entre los porcentajes de pertenencia a cada grupo encontrados en la fase de predicción, Tabla 5.74, Tabla 5.75 y Tabla 5.76, con los obtenidos en la fase de validación, Tabla 5.77, Tabla 5.78 y Tabla 5.79. De media, un porcentaje presenta una variación igual a 0.04 % los días laborales, 0.14 % los sábados y 0.15 % los domingos, que es asumible, por lo que el modelo es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	29.93 %	Grupo 6	13.85 %	Grupo 11	0.56 %
Grupo 2	0.27 %	Grupo 7	0.02 %	Grupo 12	0.05 %
Grupo 3	0.01 %	Grupo 8	0.15 %	Grupo 13	0.78 %
Grupo 4	1.22 %	Grupo 9	0.12 %	Grupo 14	46.76 %
Grupo 5	2.88 %	Grupo 10	3.42 %	Grupo 15	0.001 %

Tabla 5.74.: Predicción del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	29.54 %	Grupo 6	0.26 %	Grupo 11	0.54 %
Grupo 2	0.01 %	Grupo 7	0.67 %	Grupo 12	52.25 %
Grupo 3	0.03 %	Grupo 8	1.84 %	Grupo 13	0.03 %
Grupo 4	10.38 %	Grupo 9	0.47 %	Grupo 14	0.02 %
Grupo 5	0.45 %	Grupo 10	1.73 %	Grupo 15	1.76 %

Tabla 5.75.: Predicción del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	28.42 %	Grupo 6	0.05 %	Grupo 11	0.12 %
Grupo 2	0.02 %	Grupo 7	1.13 %	Grupo 12	3.62 %
Grupo 3	0.15 %	Grupo 8	13.38 %	Grupo 13	0.33 %
Grupo 4	0.43 %	Grupo 9	0.44 %	Grupo 14	51.86 %
Grupo 5	0.02 %	Grupo 10	0.01 %	Grupo 15	0.03 %

Tabla 5.76.: Predicción del porcentaje de datos asociados a cada grupo, Domingo

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	30.04 %	Grupo 6	13.76 %	Grupo 11	0.55 %
Grupo 2	0.22 %	Grupo 7	0.04 %	Grupo 12	0.06 %
Grupo 3	0.03 %	Grupo 8	0.17 %	Grupo 13	0.82 %
Grupo 4	1.20 %	Grupo 9	0.11 %	Grupo 14	46.75 %
Grupo 5	2.96 %	Grupo 10	3.29 %	Grupo 15	0.00 %

Tabla 5.77.: Validación del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	30.33 %	Grupo 6	0.25 %	Grupo 11	0.46 %
Grupo 2	0.02 %	Grupo 7	0.73 %	Grupo 12	51.80 %
Grupo 3	0.06 %	Grupo 8	1.96 %	Grupo 13	0.02 %
Grupo 4	10.27 %	Grupo 9	0.45 %	Grupo 14	0.02 %
Grupo 5	0.38 %	Grupo 10	1.79 %	Grupo 15	1.46 %

Tabla 5.78.: Validación del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	28.91 %	Grupo 6	0.05 %	Grupo 11	0.03 %
Grupo 2	0.00 %	Grupo 7	0.97 %	Grupo 12	3.80 %
Grupo 3	0.15 %	Grupo 8	12.71 %	Grupo 13	0.32 %
Grupo 4	0.33 %	Grupo 9	0.40 %	Grupo 14	52.26 %
Grupo 5	0.00 %	Grupo 10	0.01 %	Grupo 15	0.07 %

Tabla 5.79.: Validación del porcentaje de datos asociados a cada grupo, Domingo

5.4.2.3. Perfiles del consumo de un día de verano en la provincia de Barcelona

El objetivo de esta sección es encontrar los distintos patrones que representan el consumo residencial e industrial de un día de verano en la provincia de Barcelona. Para alcanzar dicha finalidad, se va a seguir la dinámica empleada en las secciones anteriores, en primer lugar, se obtiene el número de clústers a emplear en cada caso, para después calcular los distintos patrones de consumo mediante la implementación del algoritmo K-Means.

Consumo residencial

El número de clústers elegido para todos los días de la semana es el mismo. Con el objetivo de encontrar el número de clústers a emplear en cada tipo de día de la semana, ya que el “codo” de las funciones de coste, Figura 5.78, es difícil de discernir a simple vista, se ha recurrido al cálculo del porcentaje de la disminución de la suma de varianzas con cada aumento del número de clústers. Tras realizar esa operación para cada uno de los tres casos, se ha llegado a la conclusión de que el número de clústers a emplear es doce.

Una vez elegido el número de clústers a emplear, se ha implementado el algoritmo K-Means con la finalidad de encontrar doce grupos que definen el consumo residencial de un día de verano en la provincia de Barcelona. Los grupos, junto con sus desviaciones típicas, se encuentran en la Figura 5.79.

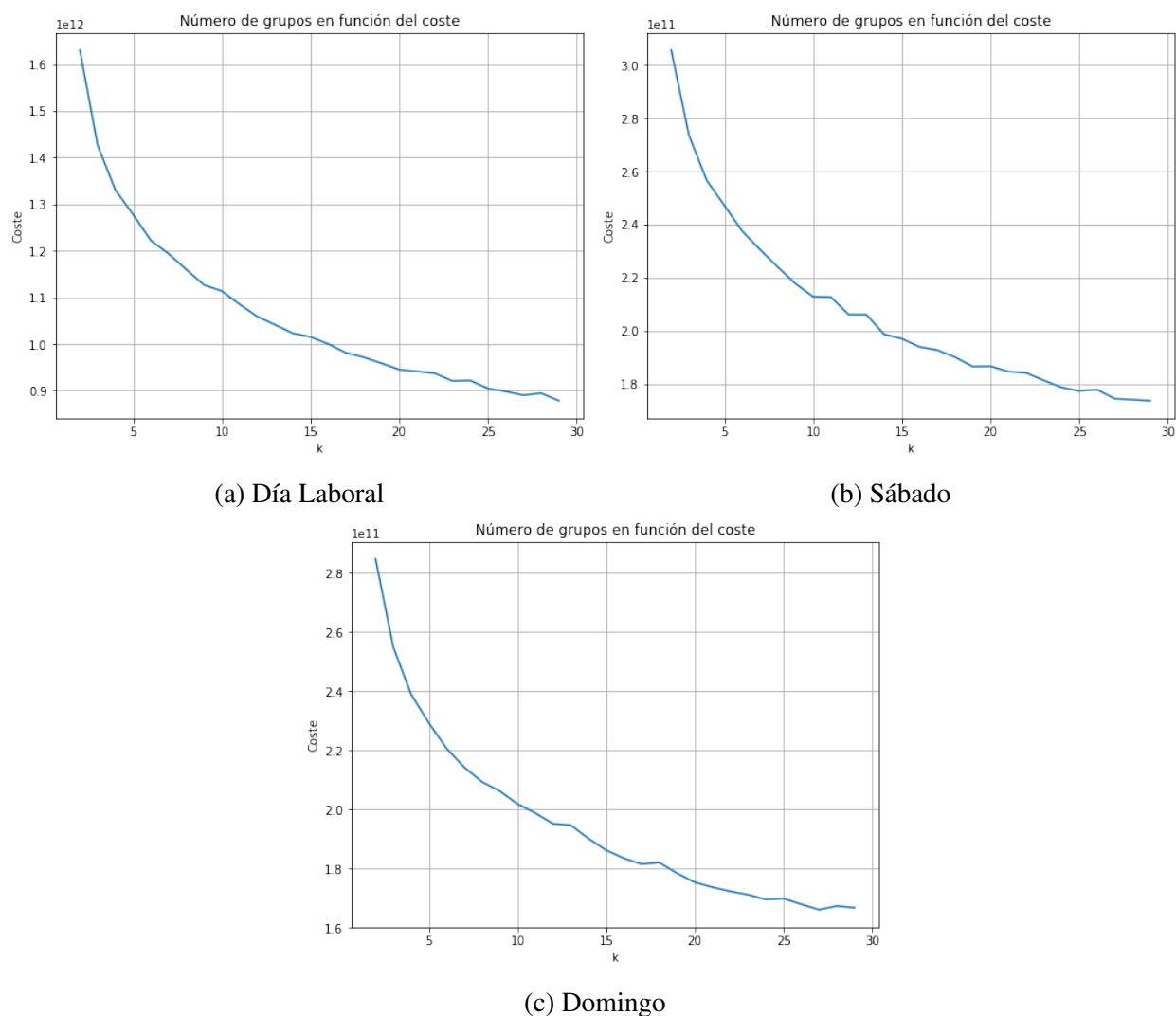
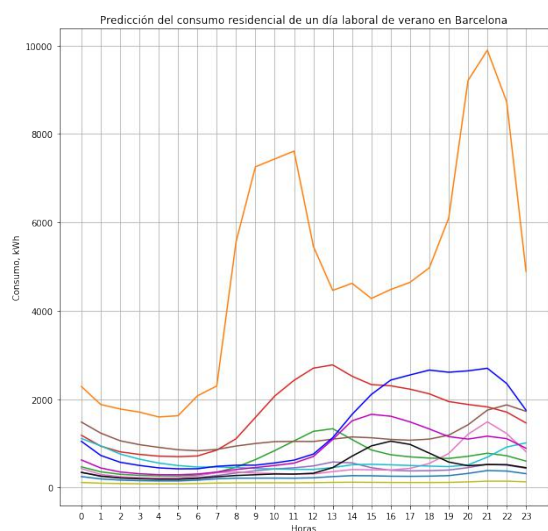


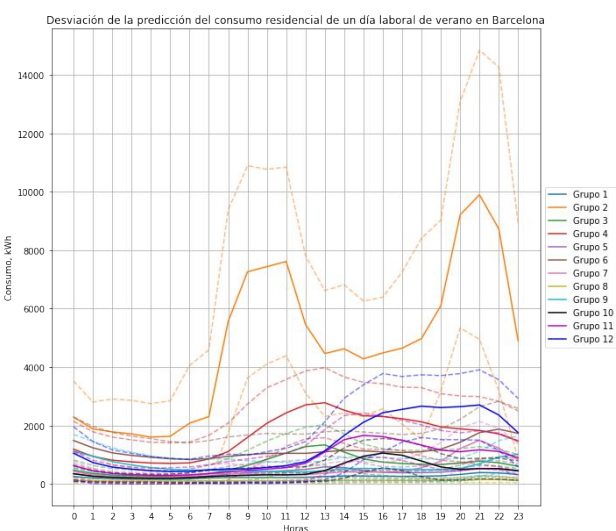
Figura 5.78.: Número de grupos en función de la suma de las distancias a los centroides

A primera vista, en todos los tipos de día, destaca la presencia de un grupo que sobresale por encima del resto, debido a los valores tan elevados que toma a lo largo del día. Es el caso del Grupo 2 en los días laborales y en los sábados y del Grupo 8 los domingos. Además, no sigue la tendencia ascendente a lo largo del día que se espera del consumo residencial, como se estableció en la sección 5.3.2. Encajaría mejor con el gráfico esperado del consumo industrial, motivo por el cual representa un porcentaje ínfimo del consumo residencial de un día de invierno en la provincia de Barcelona, Tabla 5.80, Tabla 5.81 y Tabla 5.82. Este tipo de comportamiento se puede dar en viviendas en las que durante las mañanas la limpieza sea realizada por una persona externa, que termina su jornada antes de la hora de comer, y a la hora de cenar el residente vuelve al hogar.

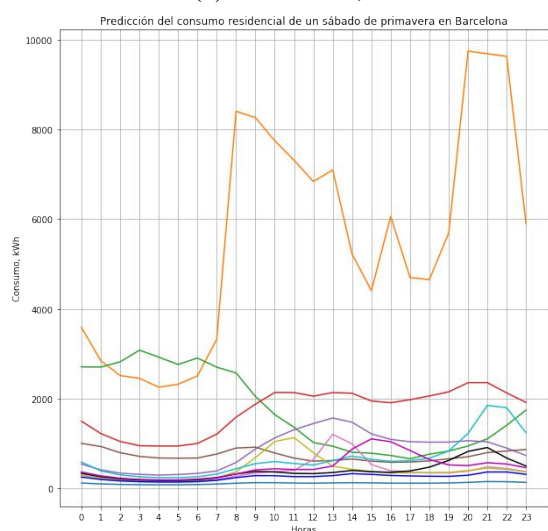
También llama la atención la presencia de un grupo cuyo consumo decae durante las horas de luz, es el caso del Grupo 3 en los sábados. Este patrón es totalmente opuesto al esperado cuando se trata de consumo residencial. Es muy probable que los clientes que pertenecen a estos grupos trabajen en el turno de noche, motivo por el cual la depresión de su consumo se da durante el día.



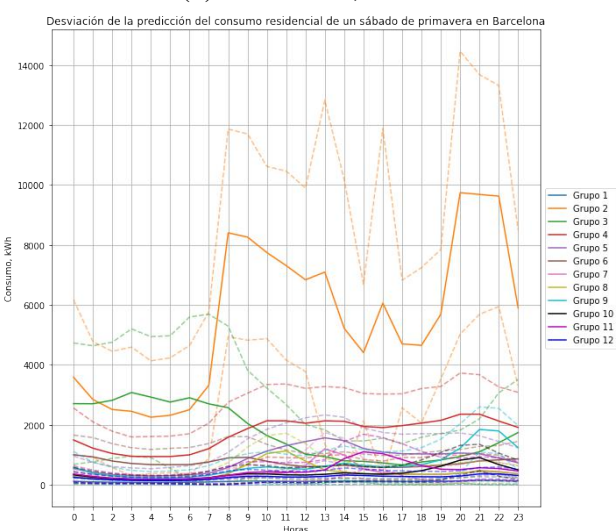
(a) Predicción, Diario



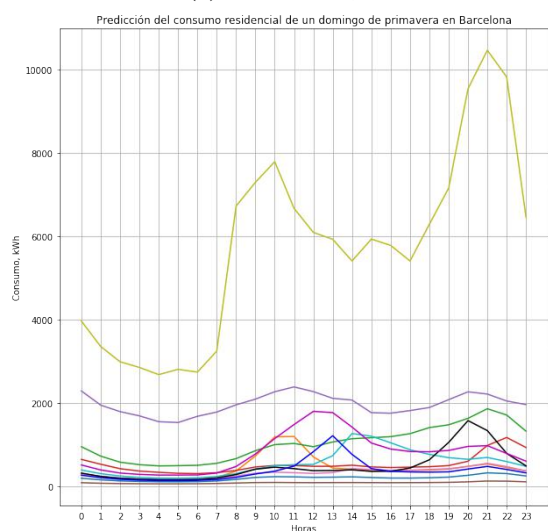
(b) Desviación, Diario



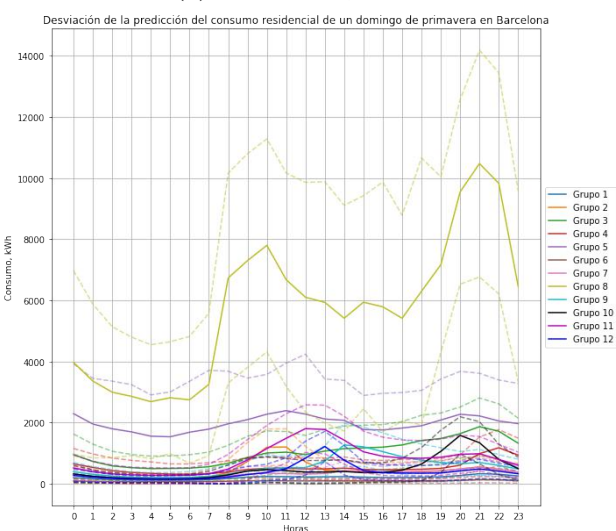
(c) Predicción, Sábado



(d) Desviación, Sábado



(e) Predicción, Domingo



(f) Desviación, Domingo

Figura 5.79.: Consumo residencial de un día de verano en Barcelona

Sorprendentemente, hay un conjunto de grupos, cuyo pico del mediodía toma valores más elevados que los del pico de la hora de cenar. Es el caso de los grupos 3, 4, 10 y 11 en el gráfico de los días laborales, los grupos 5, 7, 8 y 11 en el de los sábados y los grupos 2, 9, 11 y 12 en el de los domingos. Este tipo de comportamiento se asocia con el de aquellas personas que tan solo emplean recursos eléctricos durante la hora de la comida, lo que implica que o bien reducen drásticamente el uso de aparatos eléctricos o emplean otras fuentes de energía en ese espacio de tiempo.

El resto de los grupos representados, reflejan el patrón de consumo residencial esperado, tomando valores mayores o menores y experimentando mayor o menor diferencia entre los dos picos principales del día. Una vez más se da el que grupo con los niveles de consumo menores a lo largo de todo el día, es el que representa un mayor porcentaje del consumo. Es el caso del Grupo 8 en el gráfico de los días laborales, el Grupo 1 en el de los sábados y el Grupo 1 en el de los domingos.

Con la finalidad de aplanar los niveles de consumo a lo largo del día, ya que de esta forma las distribuidoras eléctricas pueden optimizar sus centrales, se pueden lanzar, principalmente, una campañas publicitaria. Dicha campaña se debe centrar en la reducción de los niveles de consumo de aquellos que sobresalen por encima del resto. Esto se puede conseguir mediante el establecimiento de un límite de consumo a partir del cual el precio de la electricidad aumenta para fomentar la reducción del consumo.

Con el modelo establecido, es necesario evaluar la fiabilidad del mismo. Para ello, se comparan los porcentajes de representación de cada grupos obtenidos en la fase de predicción, Tabla 5.80, Tabla 5.81 y Tabla 5.82, con los obtenidos en la fase de validación, Tabla 5.83, Tabla 5.84 y Tabla 5.85. En este caso, la variación es del 0.06 % en los días laborales, del 0.87 % en los sábados y de 0.07 % los domingos, valores que son asumibles, por lo que el modelo calculado es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	31.54 %	Grupo 5	13.77 %	Grupo 9	2.68 %
Grupo 2	0.02 %	Grupo 6	1.11 %	Grupo 10	4.58 %
Grupo 3	3.37 %	Grupo 7	4.35 %	Grupo 11	2.15 %
Grupo 4	0.45 %	Grupo 8	35.38 %	Grupo 12	0.41 %

Tabla 5.80.: Predicción del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	39.62 %	Grupo 5	1.93 %	Grupo 9	1.73 %
Grupo 2	0.02 %	Grupo 6	1.53 %	Grupo 10	8.36 %
Grupo 3	0.11 %	Grupo 7	5.97 %	Grupo 11	4.34 %
Grupo 4	0.35 %	Grupo 8	4.68 %	Grupo 12	31.36 %

Tabla 5.81.: Predicción del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	31.13 %	Grupo 5	0.25 %	Grupo 9	3.44 %
Grupo 2	4.07 %	Grupo 6	27.98 %	Grupo 10	3.12 %
Grupo 3	1.25 %	Grupo 7	17.20 %	Grupo 11	1.69 %
Grupo 4	4.07 %	Grupo 8	0.02 %	Grupo 12	5.79 %

Tabla 5.82.: Predicción del porcentaje de datos asociados a cada grupo, Domingo

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	31.59 %	Grupo 5	13.91 %	Grupo 9	2.67 %
Grupo 2	0.01 %	Grupo 6	1.10 %	Grupo 10	4.62 %
Grupo 3	3.30 %	Grupo 7	4.19 %	Grupo 11	2.25 %
Grupo 4	0.44 %	Grupo 8	35.51 %	Grupo 12	0.40 %

Tabla 5.83.: Validación del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	39.83 %	Grupo 5	1.95 %	Grupo 9	1.73 %
Grupo 2	0.02 %	Grupo 6	1.57 %	Grupo 10	8.33 %
Grupo 3	0.11 %	Grupo 7	5.81 %	Grupo 11	4.51 %
Grupo 4	0.39 %	Grupo 8	4.61 %	Grupo 12	31.15 %

Tabla 5.84.: Validación del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	31.16 %	Grupo 5	0.22 %	Grupo 9	3.43 %
Grupo 2	4.13 %	Grupo 6	27.69 %	Grupo 10	3.11 %
Grupo 3	1.21 %	Grupo 7	17.33 %	Grupo 11	1.63 %
Grupo 4	4.20 %	Grupo 8	0.05 %	Grupo 12	5.85 %

Tabla 5.85.: Validación del porcentaje de datos asociados a cada grupo, Domingo

Consumo industrial

Una vez estudiado el consumo residencial en verano para todos los días de la semana, es el turno del consumo industrial. En primer lugar, antes de encontrar los patrones del consumo industrial de un día de verano en la provincia de Barcelona, es necesario obtener el número de clústers a emplear. Para lograr dicho objetivo, se ha creado el siguiente gráfico, Figura 5.80. Como se puede observar, es “codo” de la función no se puede observar a primer vista, por lo que se ha recurrido al cálculo del porcentaje de disminución de la varianza. Al finalizar el proceso se concluye que el

número de clústers a emplear es doce en los días laborales y quince en los fines de semana. A pesar de haber elegido distintas cantidades de clústers, todos los días de la semana se van a estudiar a la vez debido a las similitudes presentes en los patrones encontrados.

Con el número de clústers elegido para cada tipo de día de la semana, se ha podido implementar el algoritmo K-Means, con la finalidad de obtener los distintos grupos de clientes que representan el consumo industrial a lo largo de un día de verano en la provincia de Barcelona. Los grupos, junto con sus respectivas desviaciones típicas, se encuentran en la Figura 5.81.

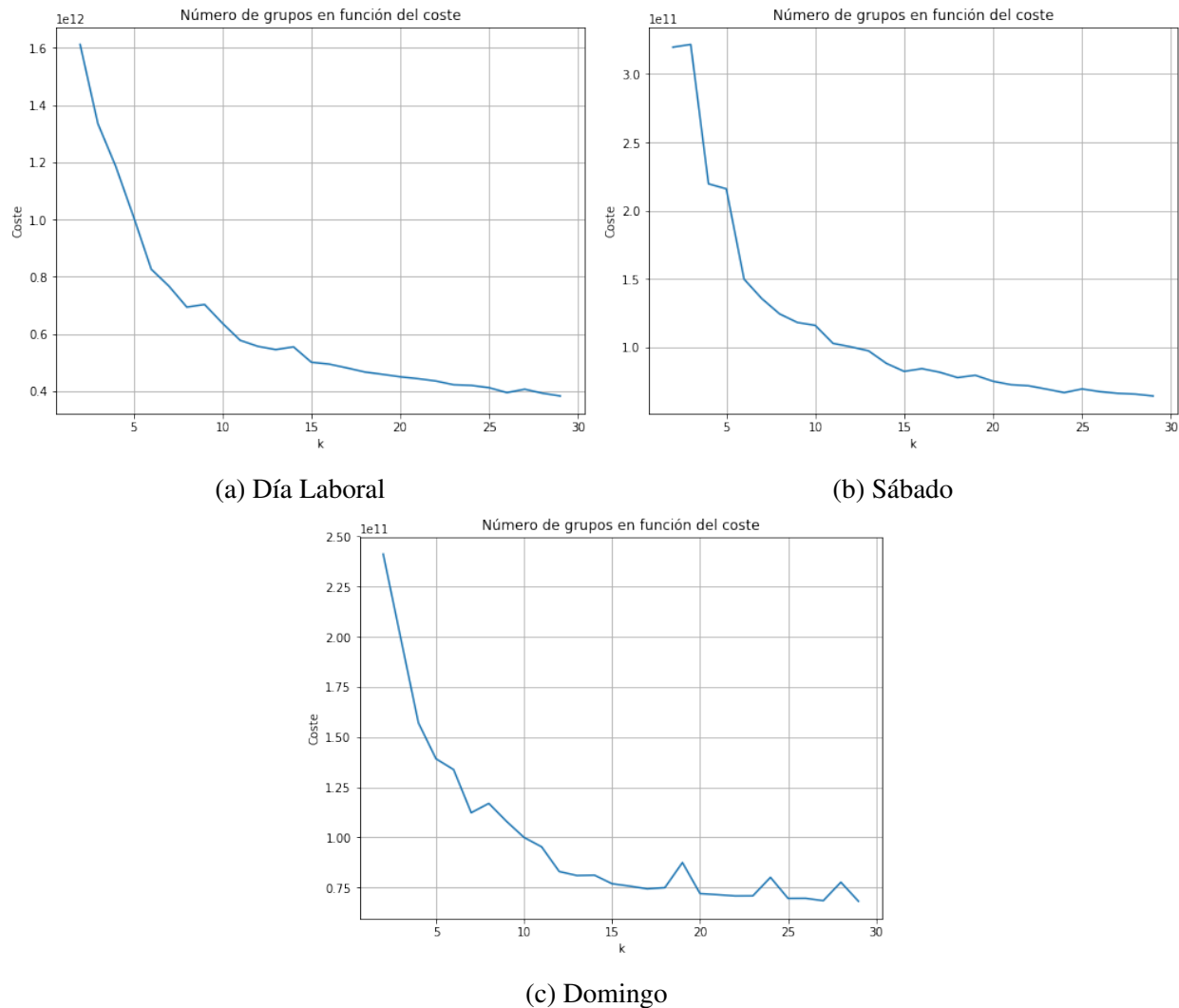


Figura 5.80.: Número de grupos en función de la suma de las distancias a los centroides

En el gráfico de los fines de semana, Figura 5.81, se pueden observar algunos grupos que tan solo presentan un pico a lo largo del día, durante el horario de mañana. Se trata de los grupos 4, 10 y 11 en el caso de los sábados y del Grupo 10 en el caso de los domingos. Este tipo de comportamiento se puede asociar al seguido por locales que tan solo ofrecen sus servicios por las mañanas los fines de semana, como puede ser el caso de quioscos.

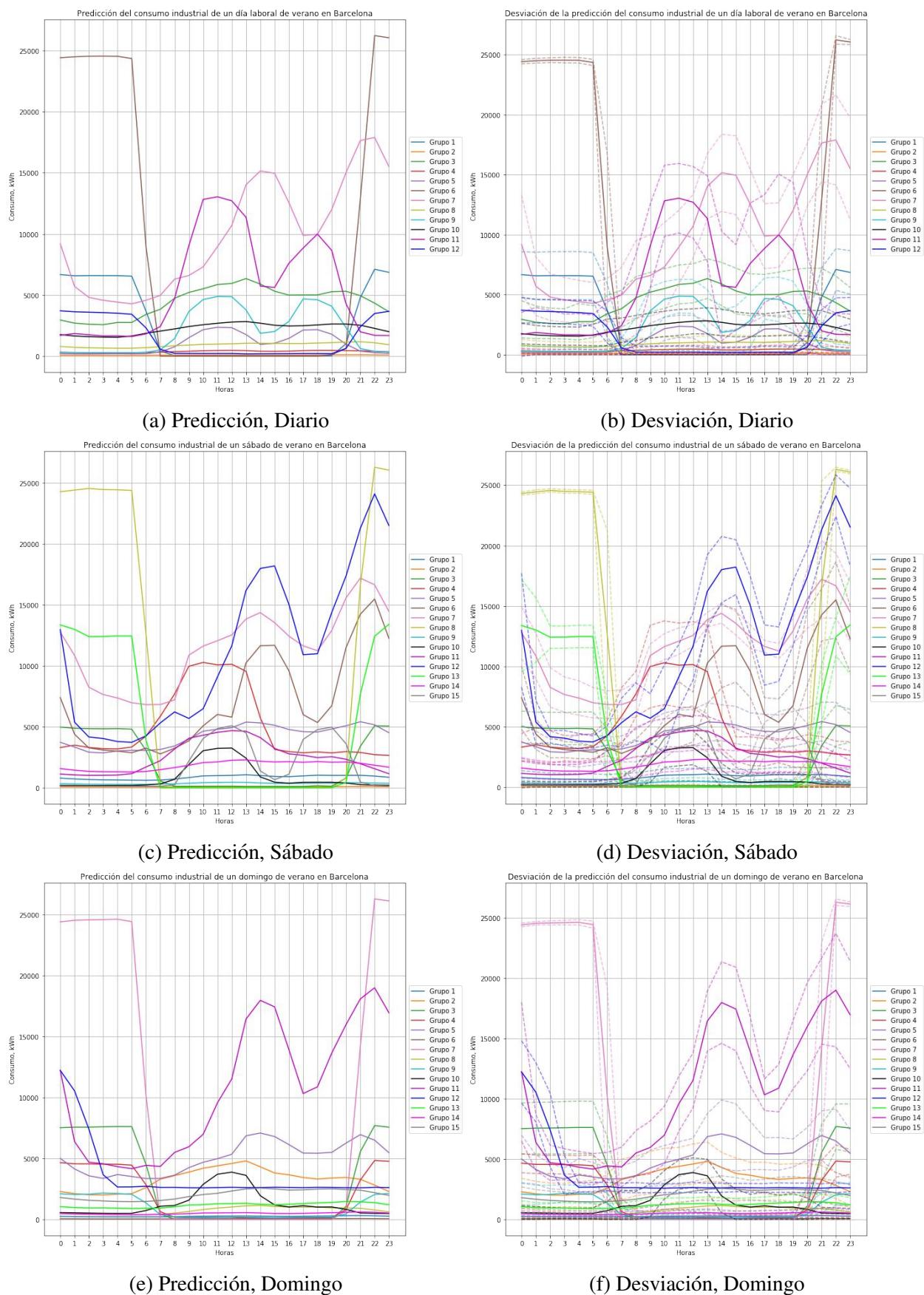


Figura 5.81.: Consumo industrial de un día de verano en Barcelona

Un conjunto de grupos que llama la atención es el formado por aquellos que presentan, aproximadamente dos niveles de consumo. Los máximos de dicho consumo se dan desde aproximadamente las 20:00 hasta las 7:00. Este conjunto de grupos está formado por los grupos 1, 6, y 12 en el gráfico de días laborales, por los grupos 3, 8, y 12 en los sábados y por los grupos 3, 4, 7, 9 y 12 en los domingos. Este tipo de comportamiento se puede asociar al seguido por locales que comienzan su actividad a media tarde y continúan su actividad durante toda la noche. Sin embargo, este patrón es seguido por un porcentaje mínimo de clientes, Tabla 5.86, Tabla 5.87, Tabla 5.88.

Algunos de los grupos calculados, presentan una tendencia muy similar a la encontrada para el consumo industrial en la sección 5.3.2. El patrón consumo que caracteriza al consumo industrial presenta dos picos claros de consumo, uno por la mañana y otro por la tarde. Es el caso de los grupos 3, 5, 7, 9, 10 y 11 de los días laborales, de los grupos 5, 6, 7, 12 y 15 de los sábados y el los grupos 2, 5, 8, 11 y 12 de los domingos, Figura 5.81. Además, cabe mencionar que el resto de grupos presentan una tendencia más constante a lo largo del día. En estos grupos se puede inferir una pequeña subida del consumo a las horas donde deberían encontrar los picos principales, es muy probable que no se puedan distinguir por la escala de la imagen, pero que si que sigan dicha tendencia. Una vez más, los grupos con los niveles de consumo más bajos, son los que representan un porcentaje mayor del consumo industrial de un día de verano en la provincia de Barcelona; es el caso de los grupos 2, 4 y 8 en el gráfico de los días laborales, de los grupos 1, 2, 9 y 14 en el de los sábados y del Grupo 12 en el de los domingos.

Uno de los objetivos principales de este proyecto es proporcionar información útil para las distribuidoras de energía eléctrica para que puedan lanzar campañas de publicidad orientadas a determinados grupos de clientes. A las distribuidoras energéticas les conviene aplanar los niveles de consumo a lo largo del día para poder optimizar la producción de las centrales. En este caso, principalmente, se pueden lanzar una campañas publicitaria. Dicha campaña tendría como objetivo conseguir la redistribución del consumo de aquellos grupos que presentan una gran depresión entre las 7:00 y las 20:00. Para conseguir dicha redistribución, se les podrían lanzar ofertas en las que se indicasen determinadas franjas horarias en las que el precio fuese más reducido, con la esperanza de que desplazasen su consumo a dichas franjas.

Con el modelo predictivo establecido, es necesario evaluar la veracidad del mismo. Motivo por el cual se comparan los porcentajes de representación de cada grupo obtenidos en la fase de predicción, Tabla 5.86, Tabla 5.87, Tabla 5.88, con los obtenidos en la fase de validación, Tabla 5.89, Tabla 5.90, Tabla 5.91. En este caso, el resultado de la comparación es una variación media por cada grupo del 0.06 % en los días laborales, del 0.13 % en los sábados y del 0.13 % en los domingos. Estos valores son perfectamente asumibles, por lo que el modelo calculado es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	0.27 %	Grupo 5	3.53 %	Grupo 9	1.30 %
Grupo 2	55.04 %	Grupo 6	0.02 %	Grupo 10	1.50 %
Grupo 3	0.44 %	Grupo 7	0.06 %	Grupo 11	0.10 %
Grupo 4	31.38 %	Grupo 8	5.94 %	Grupo 12	0.51 %

Tabla 5.86.: Predicción del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	6.63 %	Grupo 6	0.04 %	Grupo 11	0.56 %
Grupo 2	56.92 %	Grupo 7	0.03 %	Grupo 12	0.04 %
Grupo 3	0.59 %	Grupo 8	0.02 %	Grupo 13	0.03 %
Grupo 4	0.08 %	Grupo 9	30.88 %	Grupo 14	1.63 %
Grupo 5	0.41 %	Grupo 10	1.76 %	Grupo 15	0.39 %

Tabla 5.87.: Predicción del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	27.89 %	Grupo 6	50.78 %	Grupo 11	0.05 %
Grupo 2	0.45 %	Grupo 7	0.03 %	Grupo 12	0.03 %
Grupo 3	0.16 %	Grupo 8	3.30 %	Grupo 13	2.37 %
Grupo 4	0.44 %	Grupo 9	0.41 %	Grupo 14	12.57 %
Grupo 5	0.23 %	Grupo 10	0.29 %	Grupo 15	0.98 %

Tabla 5.88.: Predicción del porcentaje de datos asociados a cada grupo, Domingo

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	0.22 %	Grupo 5	3.64 %	Grupo 9	1.31 %
Grupo 2	54.80 %	Grupo 6	0.03 %	Grupo 10	1.58 %
Grupo 3	0.41 %	Grupo 7	0.08 %	Grupo 11	0.08 %
Grupo 4	31.38 %	Grupo 8	6.04 %	Grupo 12	0.54 %

Tabla 5.89.: Validación del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	6.36 %	Grupo 6	0.04 %	Grupo 11	0.45 %
Grupo 2	57.77 %	Grupo 7	0.02 %	Grupo 12	0.04 %
Grupo 3	0.66 %	Grupo 8	0.03 %	Grupo 13	0.01 %
Grupo 4	0.09 %	Grupo 9	30.68 %	Grupo 14	1.50 %
Grupo 5	0.38 %	Grupo 10	1.57 %	Grupo 15	0.41 %

Tabla 5.90.: Validación del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	27.43 %	Grupo 6	51.56 %	Grupo 11	0.04 %
Grupo 2	0.45 %	Grupo 7	0.01 %	Grupo 12	0.03 %
Grupo 3	0.13 %	Grupo 8	3.48 %	Grupo 13	2.33 %
Grupo 4	0.35 %	Grupo 9	0.29 %	Grupo 14	12.51 %
Grupo 5	0.23 %	Grupo 10	0.19 %	Grupo 15	0.99 %

Tabla 5.91.: Validación del porcentaje de datos asociados a cada grupo, Domingo

5.4.2.4. Perfiles del consumo de un día de otoño en la provincia de Barcelona

El objetivo de esta sección es encontrar los distintos patrones que representan el consumo residencial e industrial de un día de otoño en la provincia de Barcelona. Para alcanzar dicha finalidad, se va a seguir la dinámica empleada en las secciones anteriores, en primer lugar, se obtiene el número de clústers a emplear en cada caso, para después calcular los distintos patrones de consumo mediante la implementación del algoritmo K-Means.

Consumo residencial

En este caso, el número de clústers elegido para todos los días de la semana es el mismo. Con el objetivo de encontrar el número de clústers a emplear en cada tipo de día de la semana, ya que el “codo” de las funciones de coste, Figura 5.82, es difícil de discernir a simple vista, se ha recurrido al cálculo del porcentaje de la disminución de la suma de varianzas con cada aumento del número de clústers. Tras realizar esa operación para cada uno de los tres casos, se ha llegado a la conclusión de que el número de clústers a emplear es quince.

Una vez elegido el número de clústers a emplear, se ha implementado el algoritmo K-Means con la finalidad de encontrar quince grupos que definen el consumo residencial de un día de otoño en la provincia de Barcelona. Los grupos, junto con sus desviaciones típicas, se encuentran en la Figura 5.83. A pesar de que el número de clústers elegido es quince, los gráficos de los domingos tan solo reflejan información realativa a catorce grupos. Lo cual se debe a que un grupo solamente representaba información acerca de un conjunto de datos, por lo que no proporcionaba información útil para este estudio.

A primera vista, en todos los tipos de día, destaca la presencia de un grupo que sobresale por encima del resto, debido a los valores tan elevados que toma a lo largo del día. Es el caso del Grupo 3 en los días laborales, del Grupo 4 en los sábados y del Grupo 8 en los domingos. Además, no siguen la tendencia ascendente a lo largo del día que se espera del consumo residencial, como se estableció en la sección 5.3.2. Sorprendentemente, encajaría mejor con el gráfico esperado del consumo industrial, motivo por el cual representa un porcentaje ínfimo del consumo residencial de un día de invierno en la provincia de Barcelona, Tabla 5.92, Tabla 5.93 y Tabla 5.94. Este tipo de comportamiento se puede dar en viviendas en las que durante las mañanas la limpieza sea realizada por una persona externa, que termina su jornada antes de la hora de comer, y a la hora de cenar el residente vuelve al hogar.

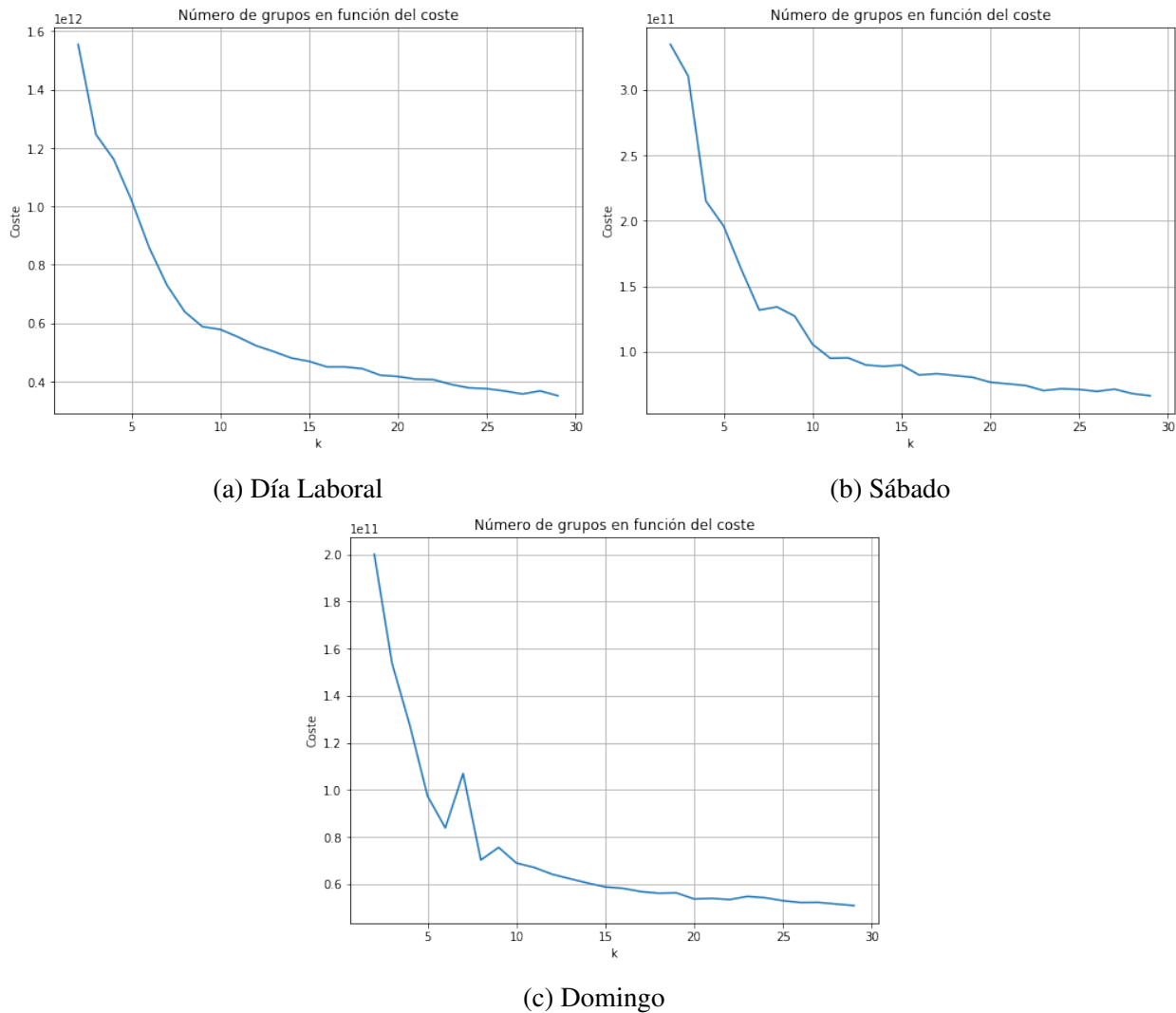


Figura 5.82.: Número de grupos en función de la suma de las distancias a los centroides

También llama la atención la presencia de un conjunto de grupos cuyo consumo decae durante las horas de luz, es el caso de los grupos 2 y 15 en los días laborales, los grupos 5 y 6 en el caso de los sábados y los grupos 4 y 13 en el de los domingos. Este patrón es totalmente opuesto al esperado cuando se trata de consumo residencial. Es muy probable que los clientes que pertenecen a estos grupos trabajen en el turno de noche, motivo por el cual la depresión de su consumo se da durante el día. Además, representan un pequeño porcentaje del consumo residencial de un día de otoño en la provincia de Barcelona, Tabla 5.92, Tabla 5.93 y Tabla 5.94.

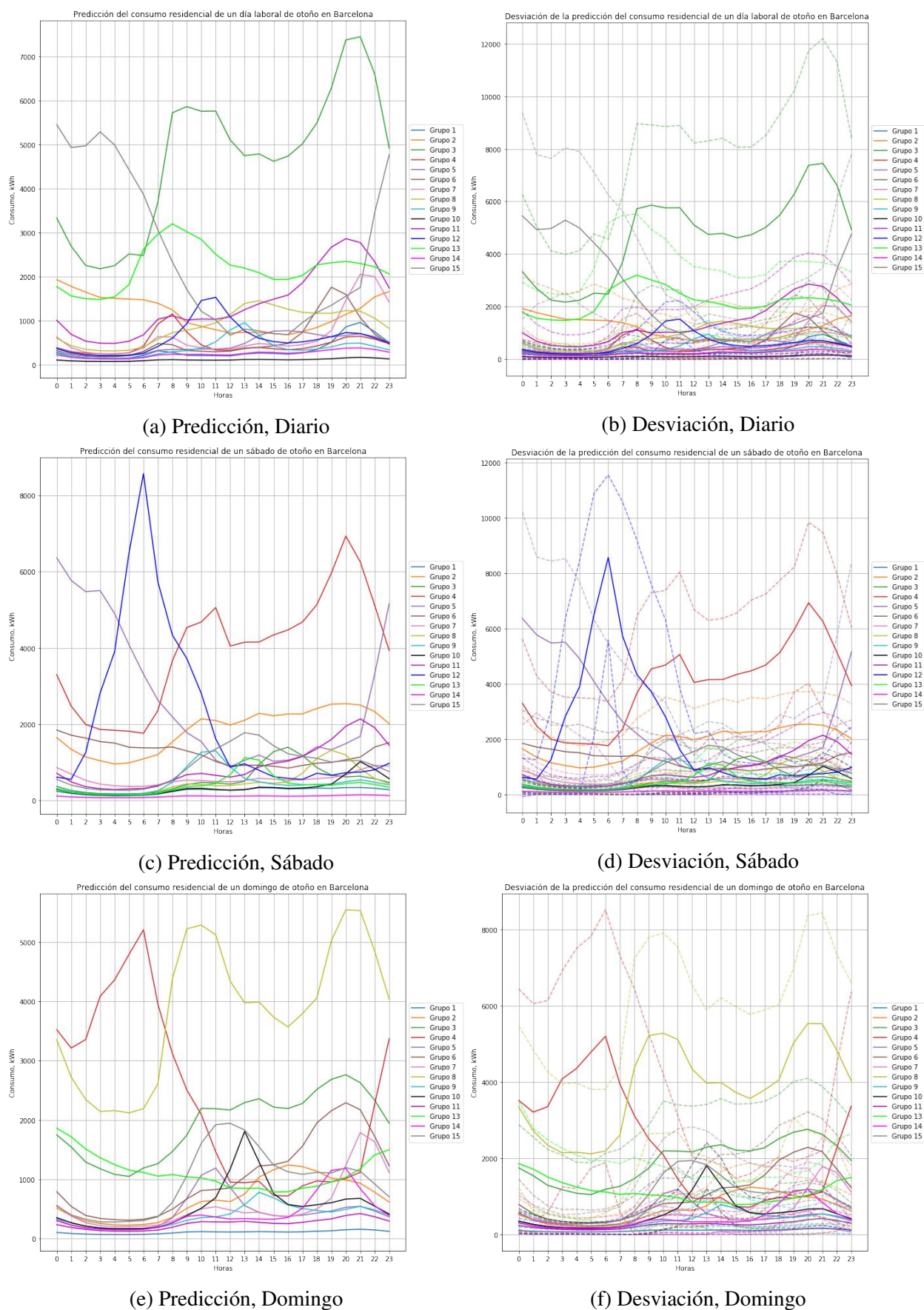


Figura 5.83.: Consumo residencial de un día de otoño en Barcelona

Sorprendentemente, hay un conjunto de grupos, que solamente presentan un pico de consumo cuya posición a lo largo del tiempo varía entre las 5:00 y las 16:00, aproximadamente. Es el caso de los grupos 4, 8, 9, 12 y 13 en el gráfico de los días laborales, los grupos 3, 9, 12, 13 y 15 en el de los sábados y los grupos 5, 9 10 y 15 en el de los domingos. Este tipo de comportamiento se asocia con el de aquellas personas que tan solo emplean recursos eléctricos durante la hora de la comida, lo que implica que o bien reducen drásticamente el uso de aparatos eléctricos o emplean otras fuentes de energía en ese espacio de tiempo.

El resto de los grupos representados, reflejan el patrón de consumo residencial esperado, tomando valores mayores o menores y experimentando mayor o menor diferencia entre los dos picos principales del día. Una vez más se da el que grupo con los niveles de consumo menores a lo largo de todo el día, es el que representa un mayor porcentaje del consumo.

Con la finalidad de aplanar los niveles de consumo a lo largo del día, ya que de esta forma las distribuidoras eléctricas pueden optimizar sus centrales, se pueden lanzar, principalmente, dos campañas publicitarias. Una de ellas se debe centrar en la reducción de los niveles de consumo de aquellos que sobresalen por encima del resto. Por otro lado, se podría buscar la redistribución del consumo de aquellos grupos que siguen una tendencia contraria a la esperada. Esto se puede conseguir mediante el lanzamiento de ofertas de energía en determinadas franjas horarias, medida que motiva la redistribución del consumo. Además, se puede establecer un límite de consumo a partir del cual el precio de la electricidad aumenta para fomentar la reducción del consumo.

Con el modelo establecido, es necesario evaluar la fiabilidad del mismo. Para ello, se comparan los porcentajes de representación de cada grupos obtenidos en la fase de predicción, Tabla 5.92, Tabla 5.93 y Tabla 5.94, con los obtenidos en la fase de validación, Tabla 5.95, Tabla 5.96 y Tabla 5.97. En este caso, la variación es del 0.06 % en los días laborales, del 0.12 % en los sábados y de 0.08 % los domingos, valores que son asumibles, por lo que el modelo calculado es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	9.60 %	Grupo 6	3.09 %	Grupo 11	0.79 %
Grupo 2	0.70 %	Grupo 7	2.24 %	Grupo 12	2.31 %
Grupo 3	0.05 %	Grupo 8	2.44 %	Grupo 13	0.31 %
Grupo 4	3.29 %	Grupo 9	6.40 %	Grupo 14	30.31 %
Grupo 5	7.68 %	Grupo 10	30.69 %	Grupo 15	0.12 %

Tabla 5.92.: Predicción del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	29.39 %	Grupo 6	0.85 %	Grupo 11	2.11 %
Grupo 2	0.62 %	Grupo 7	3.81 %	Grupo 12	0.06 %
Grupo 3	3.60 %	Grupo 8	4.12 %	Grupo 13	7.43 %
Grupo 4	0.07 %	Grupo 9	4.35 %	Grupo 14	33.18 %
Grupo 5	0.11 %	Grupo 10	7.94 %	Grupo 15	2.35 %

Tabla 5.93.: Predicción del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	31.31 %	Grupo 6	1.44 %	Grupo 11	28.95 %
Grupo 2	4.29 %	Grupo 7	2.60 %	Grupo 12	0.001 %
Grupo 3	0.52 %	Grupo 8	0.08 %	Grupo 13	0.97 %
Grupo 4	0.16 %	Grupo 9	12.02 %	Grupo 14	6.04 %
Grupo 5	5.92 %	Grupo 10	3.91 %	Grupo 15	1.79 %

Tabla 5.94.: Predicción del porcentaje de datos asociados a cada grupo, Domingo

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	9.55 %	Grupo 6	3.08 %	Grupo 11	0.80 %
Grupo 2	0.76 %	Grupo 7	2.21 %	Grupo 12	2.30 %
Grupo 3	0.04 %	Grupo 8	2.45 %	Grupo 13	0.30 %
Grupo 4	3.26 %	Grupo 9	6.57 %	Grupo 14	30.00 %
Grupo 5	7.75 %	Grupo 10	30.82 %	Grupo 15	0.12 %

Tabla 5.95.: Validación del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	29.30 %	Grupo 6	0.74 %	Grupo 11	2.00 %
Grupo 2	0.58 %	Grupo 7	4.06 %	Grupo 12	0.05 %
Grupo 3	3.78 %	Grupo 8	3.96 %	Grupo 13	7.22 %
Grupo 4	0.06 %	Grupo 9	4.18 %	Grupo 14	33.29 %
Grupo 5	0.12 %	Grupo 10	8.30 %	Grupo 15	2.38 %

Tabla 5.96.: Predicción del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	31.58 %	Grupo 6	1.48 %	Grupo 11	29.01 %
Grupo 2	4.13 %	Grupo 7	2.68 %	Grupo 12	0.00 %
Grupo 3	0.53 %	Grupo 8	0.06 %	Grupo 13	0.98 %
Grupo 4	0.10 %	Grupo 9	11.87 %	Grupo 14	5.91 %
Grupo 5	5.88 %	Grupo 10	3.97 %	Grupo 15	1.82 %

Tabla 5.97.: Validación del porcentaje de datos asociados a cada grupo, Domingo

Consumo industrial

Una vez más, a pesar de que el número de clústers empleado en cada tipo de día de la semana es diferente, se ha decidido juntar el estudio de todo el consumo industrial de un día de otoño en la provincia de Barcelona, en un sólo análisis debido a las similitudes presentes entre los grupos encontrados. Con el objetivo de concretar el número de clústers a emplear en cada tipo de día de la semana, se ha recurrido al cálculo de la disminución porcentual de la suma de varianzas en función del número de grupos utilizados, ya que las funciones de coste calculadas, Figura 5.84, carecen de “codo” discernible a primera vista. Tras completar el proceso, se ha decidido emplear catorce clúster en el caso de los días laborales, catorce en los sábados y doce en los domingos.

Una vez elegido el número de clústers a emplear, se ha implementado el algoritmo K-Means con la finalidad de encontrar los grupos que definen el consumo industrial de un día de otoño en la provincia de Barcelona. Los grupos, junto con sus desviaciones típicas, se encuentran en la Figura 5.85.

Un conjunto de grupos que llama la atención es el formado por aquellos que presentan, aproximadamente dos niveles de consumo. Los máximos de dicho consumo se dan desde aproximadamente las 16:00 hasta las 8:00. Este conjunto de grupos está formado por los grupos 2, 3 y 14 en el gráfico de días laborales, por los grupos 4, 7, 10, 13 y 14 en los sábados y por los grupos 3, 4, 7 y 12 en los domingos. Este tipo de comportamiento se puede asociar al seguido por locales que comienzan su actividad después de comer y continúan su actividad durante toda la noche. Sin embargo, este patrón es seguido por un porcentaje mínimo de clientes, Tabla 5.98, Tabla 5.99, Tabla 5.100. Patrones similares ya se había detectado en el caso del consumo industrial del resto de las estaciones, encontrándose la principal diferencia en la duración exacta de la depresión del consumo.

También destaca la presencia de unos grupo que poseen dos picos muy claros, uno por la mañana y otro por la tarde, que solamente se prolongan durante un par de horas, mientras que durante el resto del día el consumo es prácticamente constante. Es el caso del Grupo 7 del gráfico del consumo en los días laborales, este patrón no se detecta durante los fines de semana. Lo más extraño es que estos picos se dan sobre las 5:00 y las 16:00. Sin embargo el porcentaje de consumo que se ve representado por dichos patrones es mínimo, Tabla 5.98.

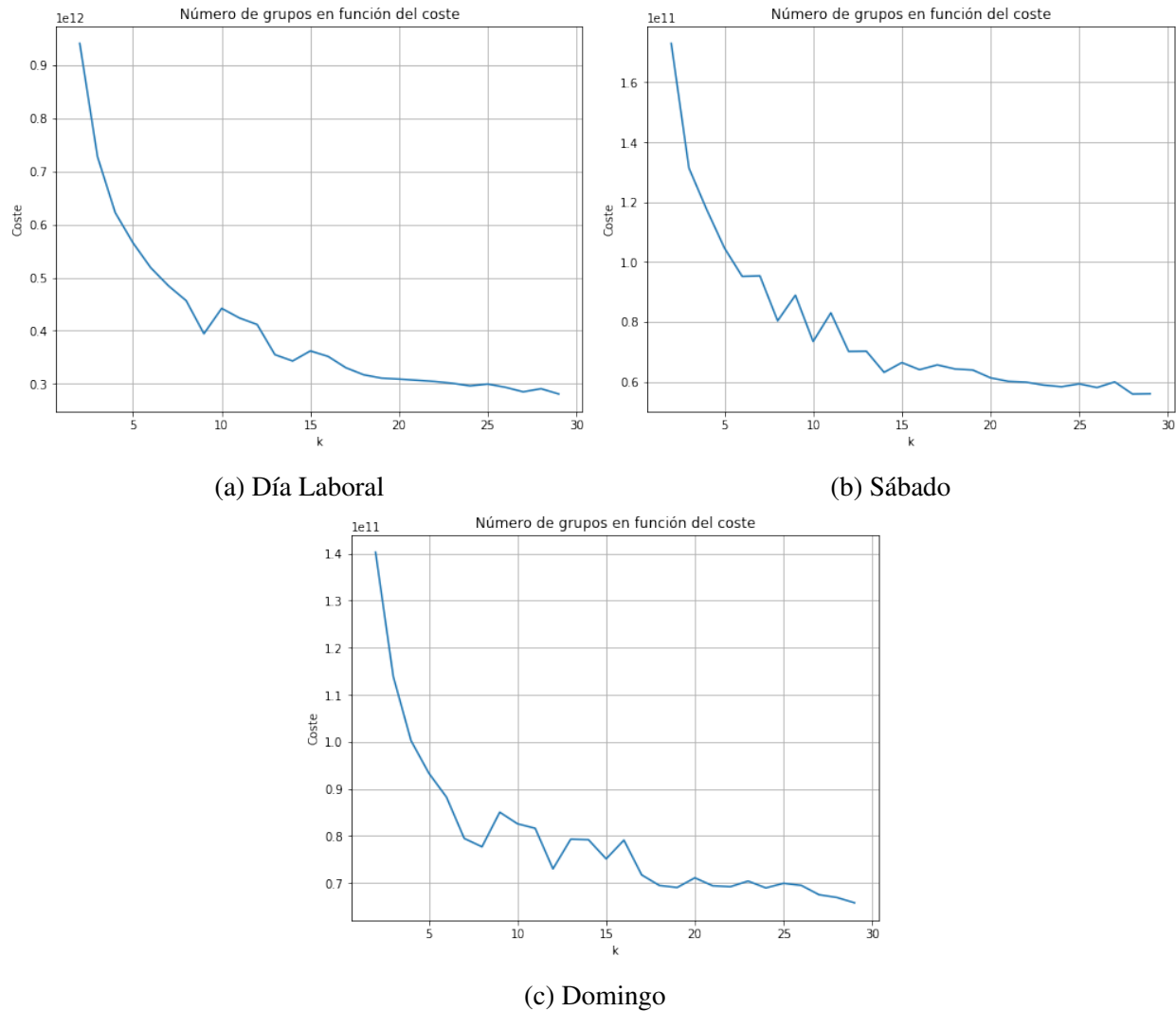
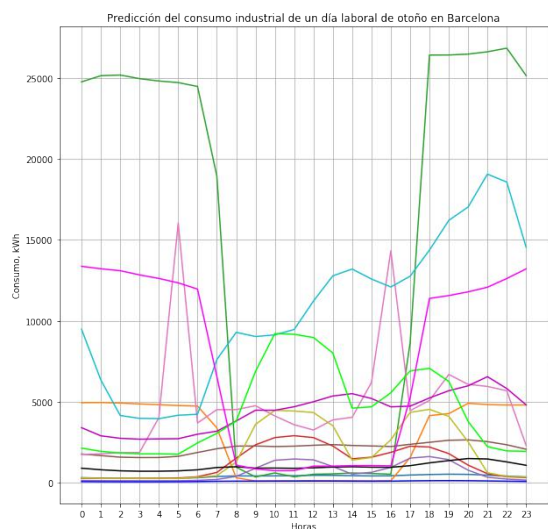


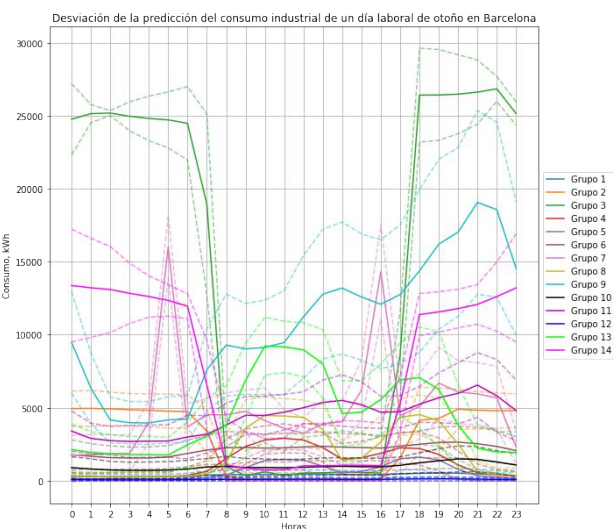
Figura 5.84.: Número de grupos en función de la suma de las distancias a los centroides

En cada uno de los tipos de día de la semana destaca la presencia de grupos cuyo patrón asemeja al esperado si se tratase de consumo residencial, sección 5.3.2. Es el caso de los grupos 9 y 11 en los días laborales, del Grupo 2 en los sábados y del Grupo 6 en los domingos. Dicho consumo presenta una progresión ascendente a lo largo de la jornada laboral, asemejándose a la tendencia esperada de un consumo residencial. Esto implica que son empresas cuya actividad industrial aumenta a la hora de la cena, por lo que puede ser el consumo de restaurantes, que presentan mayor afluencia de clientes a la hora de la cena. Patrones similares fueron encontrados en el consumo industrial de un día de primavera en la provincia de Barcelona y en el consumo industrial de otoño en la provincia de Madrid.

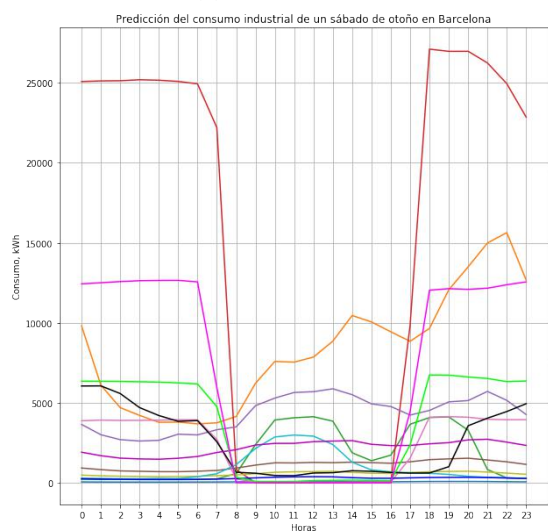
En el gráfico de los sábados, Figura 5.85, se pueden observar algunos grupos que tan solo presentan un pico a lo largo del día, durante el horario de mañana, se trata del Grupo 9. Este tipo de comportamiento se puede asociar al seguido por locales que tan solo ofrecen sus servicios por las mañanas los fines de semana, como puede ser el caso de quioscos.



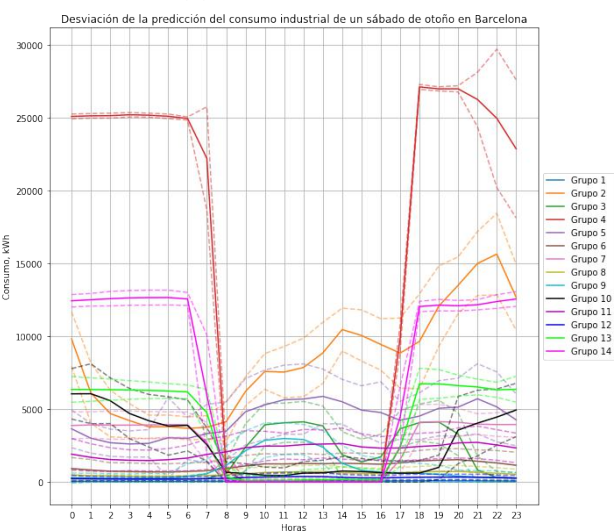
(a) Predicción, Diario



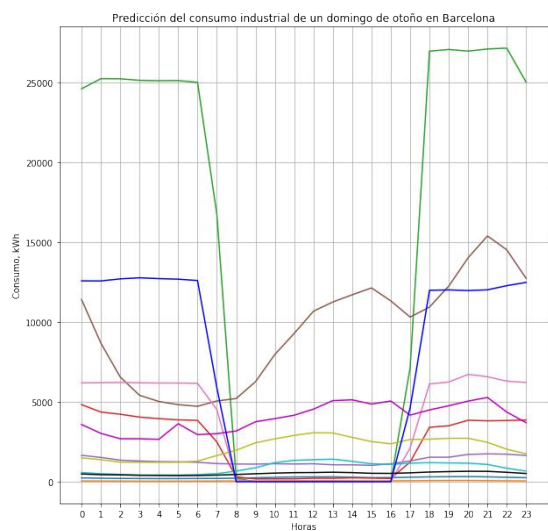
(b) Desviación, Diario



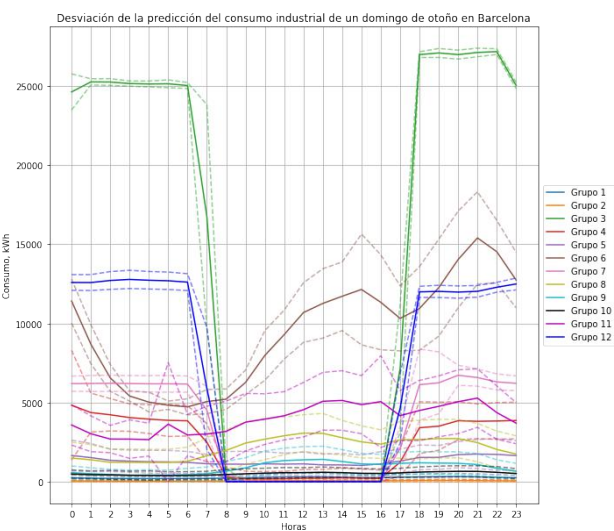
(c) Predicción, Sábado



(d) Desviación, Sábado



(e) Predicción, Domingo



(f) Desviación, Domingo

Figura 5.85.: Consumo industrial de un día de otoño en Barcelona

Algunos de los grupos calculados, presentan una tendencia muy similar a la encontrada para el consumo industrial en la sección 5.3.2. El patrón consumo que caracteriza al consumo industrial presenta dos picos claros de consumo, uno por la mañana y otro por la tarde. Es el caso de los grupos 4, 5, 8 y 13 de los días laborales, del los grupos 3 y 5 de los sábados y el los grupos 8 y 9 de los domingos, Figura 5.85. Además, cabe mencionar que el resto de grupos presentan una tendencia más constante a lo largo del día. En estos grupos se puede inferir una pequeña subida del consumo a las horas donde e deberían encontrar los picos principales, es muy probable que no se puedan distinguir por la escala de la imagen, pero que si que sigan dicha tendencia. Una vez más, los grupos con los niveles de consumo más bajos, son los que representan un porcentaje mayor del consumo industrial de un día de otoño en la provincia de Barcelona; es el caso del Grupo 12 en el gráfico de los días laborales, del Grupo 1 en el de los sábados y del Grupo 2 en el de los domingos.

Uno de lo objetivos principales de este proyecto es proporcionar información útil para las distribuidoras de energía eléctrica para que puedan lanzar campañas de publicidad orientadas a determinados grupos de clientes. A las distribuidoras energéticas les conviene aplanar los niveles de consumo a lo largo del día para poder optimizar la producción de las centrales. En este caso, principalmente, se pueden lanzar una campañas publicitaria. Dicha campaña tendría como objetivo conseguir la redistribución del consumo de aquellos grupos que presentan una gran depresión entre las 8:00 y las 16:00. Para conseguir dicha redistribución, se les podrían lanzar ofertas en las que se indicasen determinadas franjas horarias en las que el precio fuese más reducido, con la esperanza de que desplazasen su consumo a dichas franjas.

Con el modelo predictivo establecido, es necesario evaluar la veracidad del mismo. Motivo por el cual se comparan los porcentajes de representación de cada grupo obtenidos en la fase de predicción, Tabla 5.98, Tabla 5.99, Tabla 5.100, con los obtenidos en la fase de validación, Tabla 5.101, Tabla 5.102, Tabla 5.103. En este caso, el resultado de la comparación es una variación media por cada grupo del 0.10% en los días laborales, del 0.10% en los sábados y del 0.15% en los domingos. Estos valores son perfectamente asumibles, por lo que el modelo calculado es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	31.17 %	Grupo 6	1.39 %	Grupo 11	0.25 %
Grupo 2	0.65 %	Grupo 7	0.02 %	Grupo 12	52.50 %
Grupo 3	0.03 %	Grupo 8	1.02 %	Grupo 13	0.11 %
Grupo 4	2.25 %	Grupo 9	0.03 %	Grupo 14	0.04 %
Grupo 5	4.99 %	Grupo 10	5.55 %		

Tabla 5.98.: Predicción del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	47.96 %	Grupo 6	4.33 %	Grupo 11	1.18 %
Grupo 2	0.07 %	Grupo 7	0.50 %	Grupo 12	29.14 %
Grupo 3	0.59 %	Grupo 8	13.69 %	Grupo 13	0.17 %
Grupo 4	0.03 %	Grupo 9	1.88 %	Grupo 14	0.04 %
Grupo 5	0.30 %	Grupo 10	0.12 %		

Tabla 5.99.: Predicción del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	28.10 %	Grupo 5	1.86 %	Grupo 9	3.88 %
Grupo 2	49.70 %	Grupo 6	0.03 %	Grupo 10	14.03 %
Grupo 3	0.04 %	Grupo 7	0.18 %	Grupo 11	0.30 %
Grupo 4	0.59 %	Grupo 8	1.25 %	Grupo 12	0.04 %

Tabla 5.100.: Predicción del porcentaje de datos asociados a cada grupo, Domingo

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	31.59 %	Grupo 6	1.24 %	Grupo 11	0.24 %
Grupo 2	0.55 %	Grupo 7	0.02 %	Grupo 12	52.37 %
Grupo 3	0.03 %	Grupo 8	1.12 %	Grupo 13	0.14 %
Grupo 4	2.41 %	Grupo 9	0.02 %	Grupo 14	0.02 %
Grupo 5	4.82 %	Grupo 10	5.43 %		

Tabla 5.101.: Validación del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	48.21 %	Grupo 6	3.93 %	Grupo 11	1.11 %
Grupo 2	0.06 %	Grupo 7	0.56 %	Grupo 12	29.00 %
Grupo 3	0.74 %	Grupo 8	13.70 %	Grupo 13	0.13 %
Grupo 4	0.03 %	Grupo 9	2.07 %	Grupo 14	0.02 %
Grupo 5	0.30 %	Grupo 10	0.14 %		

Tabla 5.102.: Validación del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	28.65 %	Grupo 5	1.92 %	Grupo 9	4.19 %
Grupo 2	49.30 %	Grupo 6	0.02 %	Grupo 10	13.77 %
Grupo 3	0.00 %	Grupo 7	0.17 %	Grupo 11	0.27 %
Grupo 4	0.55 %	Grupo 8	1.16 %	Grupo 12	0.02 %

Tabla 5.103.: Validación del porcentaje de datos asociados a cada grupo, Domingo

5.4.3. Municipios de la provincia de Málaga

En este apartado, el estudio se centra en los municipios pertenecientes a la provincia de Málaga. Es necesario recordar las conclusiones obtenidas en secciones anteriores acerca del consumo de la provincia de Málaga, ya que esta información será empleada a lo largo de este apartado.

La provincia de Málaga presenta un consumo medio igual a 7063.64 kWh, siendo la provincia con el consumo medio menor. A su vez, este consumo, esta formado por el generado por dos actividades económicas, la residencial y la industrial. En este caso, el consumo industrial representa un 20.9 % del consumo total. Además cabe destacar, que el consumo industrial sufre una gran caída durante los fines de semana, como se puede apreciar en la figura 5.14. También es necesario tener en cuenta que un 82.4 % de su consumo se rige por la tarifa regulada, a diferencia de la provincia de Madrid, en la que dominaba la tarifa libre.

5.4.3.1. Perfiles del consumo de un día de invierno en la provincia de Málaga

El objetivo de esta sección es encontrar los distintos patrones que representan el consumo residencial e industrial de un día de invierno en la provincia de Málaga. Para alcanzar dicha finalidad, se va a seguir la dinámica empleada en las secciones anteriores, en primer lugar, se obtiene el número de clústers a emplear en cada caso, para después calcular los distintos patrones de consumo mediante la implementación del algoritmo K-Means.

Consumo residencial

Sorprendentemente, en este caso, el número de clústers elegido para todos los días de la semana es el mismo. Con el objetivo de encontrar el número de clústers a emplear en cada tipo de día de la semana, ya que el “codo” de las funciones de coste, Figura 5.86 es difícil de discernir a simple vista, se ha recurrido al cálculo del porcentaje de la disminución de la suma de varianzas con cada aumento del número de clústers. Tras realizar esa operación para cada uno de los tres casos, se ha llegado a la conclusión de que el número de clústers a emplear es doce.

Una vez elegido el número de clústers a emplear, se ha implementado el algoritmo K-Means con la finalidad de encontrar doce grupos que definen el consumo residencial de un día de invierno en la provincia de Málaga. Los grupos, junto con sus desviaciones típicas, se encuentran en la Figura 5.87.

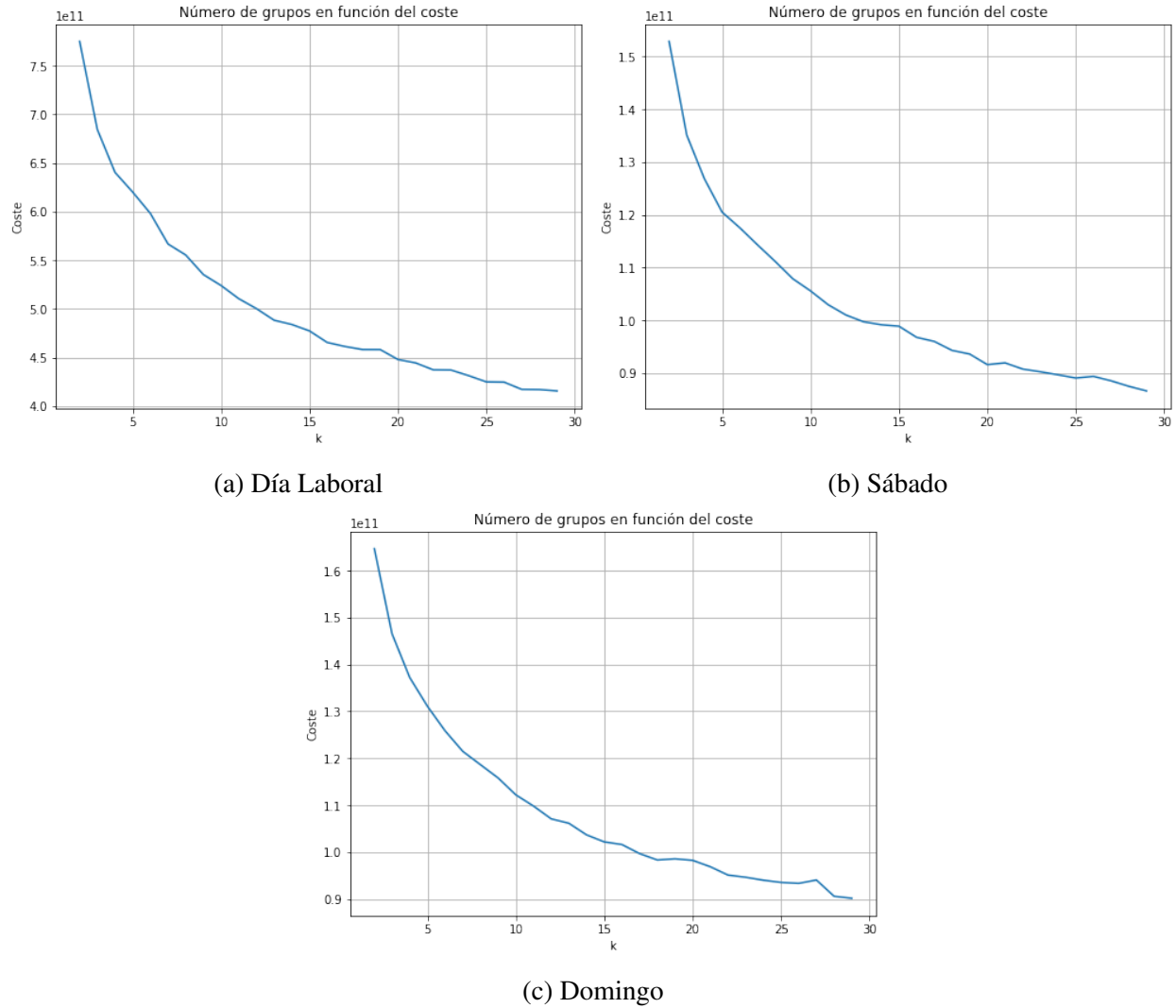


Figura 5.86.: Número de grupos en función de la suma de las distancias a los centroides

A primera vista, en todos los tipos de día, destaca la presencia de un grupo que sobresale por encima del resto, debido a los valores tan elevados que toma a lo largo del día. Es el caso del Grupo 12 en los días laborales, del Grupo 7 en los sábados y del Grupo 2 en los domingos. Además, presentan un patrón mucho más errático a lo largo del día en comparación con el encontrado en la sección 5.3.2. Motivo por el que tal vez representan un porcentaje ínfimo del consumo residencial de un día de invierno en la provincia de Málaga, Tabla 5.104, Tabla 5.105 y Tabla 5.106.

También llama la atención la presencia de un conjunto de grupos cuyo consumo decae durante las horas de luz, es el caso del Grupo 9 en los días laborales, el Grupo 8 en el caso de los sábados y el Grupo 5 en el de los domingos. Este patrón es totalmente opuesto al esperado cuando se trata de consumo residencial. Es muy probable que los clientes que pertenecen a estos grupos trabajen en el turno de noche, motivo por el cual la depresión de su consumo se da durante el día. Además, representan un pequeño porcentaje del consumo residencial de un día de invierno en la provincia de Málaga, Tabla 5.104, Tabla 5.105 y Tabla 5.106.

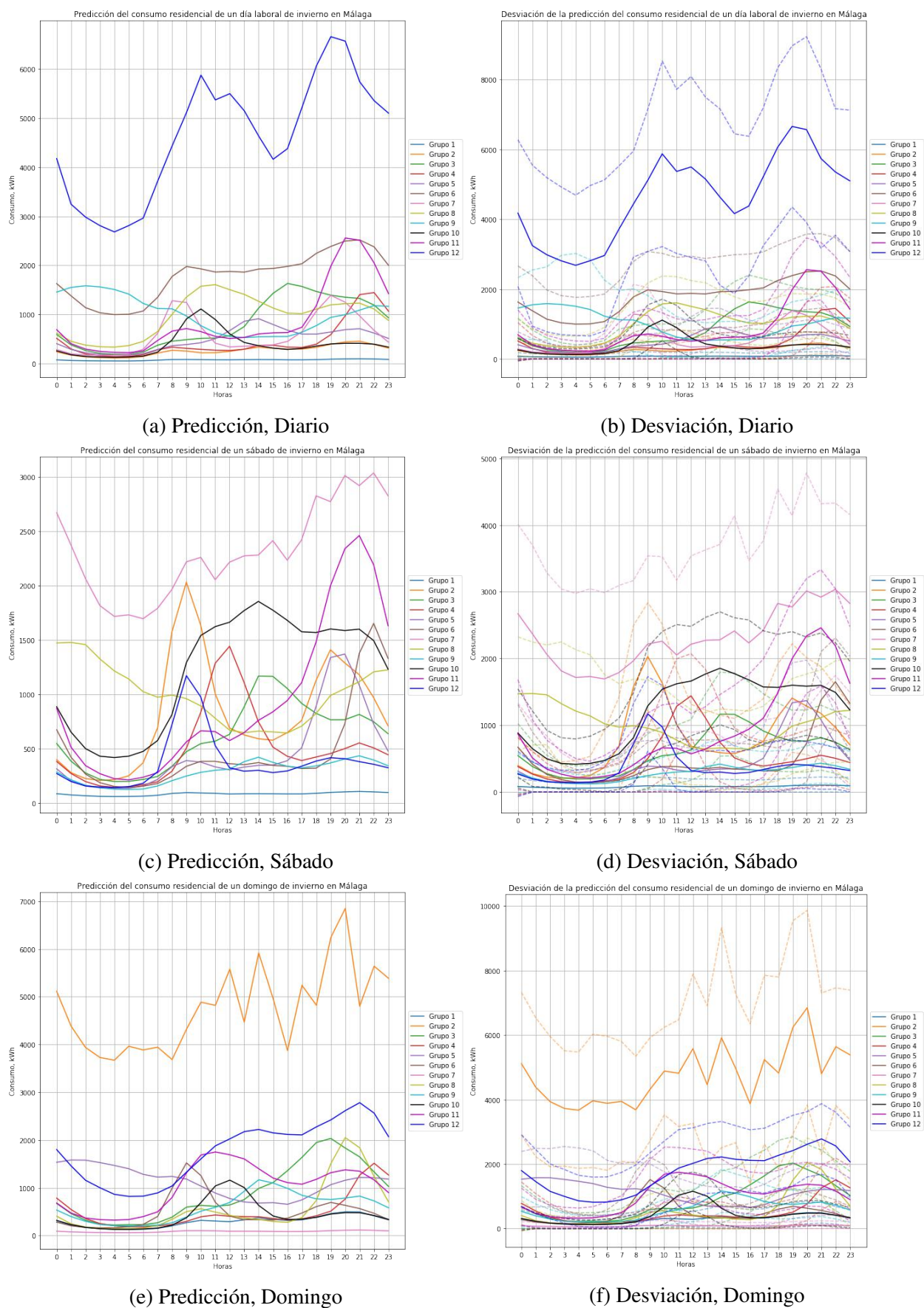


Figura 5.87.: Consumo residencial de un día de invierno en Málaga

Se puede observar un conjunto de grupos que no siguen la tendencia ascendente a lo largo del día que se espera del consumo residencial, como se estableció en la sección 5.3.2. Es el caso de los grupos 7 y 8 en los gráficos de los días laborales, el Grupo 2 en los de los sábados y el Grupo 11 en los de los domingos. Sorprendentemente, encajaría menor con el gráfico esperado del consumo industrial, razón por la que tal vez representan un porcentaje ínfimo del consumo residencial de un día de invierno en la provincia de Málaga, Tabla 5.104, Tabla 5.105 y Tabla 5.106. Este tipo de comportamiento se puede dar en viviendas en las que durante las mañanas la limpieza sea realizada por una persona externa, que termina su jornada antes de la hora de comer, y a la hora de cenar el residente vuelve al hogar.

Además, cabe destacar el conjunto de grupos, que solamente presentan el pico de mayor consumo por las mañanas. La posición de dicho pico varía a lo largo del tiempo entre las 8:00 y las 14:00, aproximadamente. Es el caso de los grupos 5 y 10 en el gráfico de los días laborales, los grupos 3, 4 y 12 en el de los sábados y los grupos 6, 9 y 10 en el de los domingos. Este tipo de comportamiento se asocia con el de aquellas personas que tan solo emplean recursos eléctricos durante la hora de la comida, lo que implica que o bien reducen drásticamente el uso de aparatos eléctricos o emplean otras fuentes de energía en ese espacio de tiempo.

Entre los grupos encontrados resalta la presencia de un tipo nuevo de patrón, que no se había manifestado anteriormente. Este patrón se caracteriza por seguir la tendencia esperada de un consumo residencial, hasta el mediodía. A partir de dicho momento, el consumo se vuelve prácticamente constante, es decir, pierde la tendencia de subida propia de un consumo residencial. Es el caso del Grupo 3 de los días laborales y de los grupos 9 y 10 de los sábados. Este tipo de comportamiento puede ser el seguido por aquellas personas que pasan toda la tarde en sus viviendas y no abusan en un periodo concreto de tiempo del uso de electrodomésticos u otros aparatos que generen un gran consumo eléctrico.

El resto de los grupos representados, reflejan el patrón de consumo residencial esperado, tomando valores mayores o menores y experimentando mayor o menor diferencia entre los dos picos principales del día. Una vez más se da el que grupo con los niveles de consumo menores a lo largo de todo el día, es el que representa un mayor porcentaje del consumo.

Con la finalidad de aplanar los niveles de consumo a lo largo del día, ya que de esta forma las distribuidoras eléctricas pueden optimizar sus centrales, se pueden lanzar, principalmente, tres campañas publicitarias. Una de ellas se debe centrar en la reducción de los niveles de consumo de aquellos que sobresalen por encima del resto. Por otro lado, se podría buscar la redistribución del consumo de aquellos grupos que siguen una tendencia de consumo descendente durante las horas de luz. Por último, una campaña se podría centrar en reducir la diferencia de ciertos picos de consumo con respecto al consumo generado el resto del día. Esto se puede conseguir mediante el lanzamiento de ofertas de energía en determinadas franjas horarias, medida que motiva la redistribución del consumo. Además, se puede establecer un límite de consumo a partir del cual el precio de la electricidad aumenta para fomentar la reducción del consumo.

Con el modelo establecido, es necesario evaluar la fiabilidad del mismo. Para ello, se comparan los porcentajes de representación de cada grupo obtenidos en la fase de predicción, Tabla

5.104, Tabla 5.105 y Tabla 5.106, con los obtenidos en la fase de validación, Tabla 5.107, Tabla 5.108 y Tabla 5.109. En este caso, la variación es del 0.10 % en los días laborales, del 0.13 % en los sábados y de 0.23 % los domingos, valores que son asumibles, por lo que el modelo calculado es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	38.18 %	Grupo 5	9.94 %	Grupo 9	1.28 %
Grupo 2	26.37 %	Grupo 6	0.89 %	Grupo 10	5.54 %
Grupo 3	2.53 %	Grupo 7	3.64 %	Grupo 11	2.12 %
Grupo 4	6.65 %	Grupo 8	2.78 %	Grupo 12	0.06 %

Tabla 5.104.: Predicción del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	39.66 %	Grupo 5	5.63 %	Grupo 9	24.28 %
Grupo 2	2.17 %	Grupo 6	4.41 %	Grupo 10	2.04 %
Grupo 3	7.21 %	Grupo 7	0.49 %	Grupo 11	2.20 %
Grupo 4	4.49 %	Grupo 8	1.83 %	Grupo 12	5.60 %

Tabla 5.105.: Predicción del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	24.91 %	Grupo 5	1.43 %	Grupo 9	7.52 %
Grupo 2	0.04 %	Grupo 6	4.73 %	Grupo 10	6.14 %
Grupo 3	2.45 %	Grupo 7	40.82 %	Grupo 11	2.95 %
Grupo 4	4.43 %	Grupo 8	3.68 %	Grupo 12	0.89 %

Tabla 5.106.: Predicción del porcentaje de datos asociados a cada grupo, Domingo

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	38.53 %	Grupo 5	9.87 %	Grupo 9	1.27 %
Grupo 2	26.11 %	Grupo 6	0.84 %	Grupo 10	5.58 %
Grupo 3	2.48 %	Grupo 7	3.68 %	Grupo 11	1.99 %
Grupo 4	6.82 %	Grupo 8	2.77 %	Grupo 12	0.07 %

Tabla 5.107.: Validación del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	39.51 %	Grupo 5	5.45 %	Grupo 9	24.54 %
Grupo 2	2.33 %	Grupo 6	4.55 %	Grupo 10	2.10 %
Grupo 3	7.27 %	Grupo 7	0.50 %	Grupo 11	1.89 %
Grupo 4	4.35 %	Grupo 8	1.93 %	Grupo 12	5.58 %

Tabla 5.108.: Validación del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	25.78 %	Grupo 5	1.61 %	Grupo 9	7.26 %
Grupo 2	0.05 %	Grupo 6	4.78 %	Grupo 10	6.27 %
Grupo 3	2.34 %	Grupo 7	40.34 %	Grupo 11	2.62 %
Grupo 4	4.23 %	Grupo 8	3.69 %	Grupo 12	1.05 %

Tabla 5.109.: Validación del porcentaje de datos asociados a cada grupo, Domingo

Consumo industrial

A pesar de que el número de clústers empleado en cada tipo de día de la semana es diferente, se ha decidido juntar el estudio de todo el consumo industrial de un día de invierno en la provincia de Málaga, en un sólo análisis debido a las similitudes presentes entre los grupos encontrados. Con el objetivo de concretar el número de clústers a emplear en cada tipo de día de la semana, se ha recurrido al cálculo de la disminución porcentual de la suma de varianzas en función del número de grupos utilizados, ya que las funciones de coste calculadas, Figura 5.88, carecen de “codo” discernible a primera vista. Tras completar el proceso, se ha decidido emplear doce clúster en el caso de los días laborales y trece en los sábados y domingos.

Una vez elegido el número de clústers a emplear, se ha implementado el algoritmo K-Means con la finalidad de encontrar los grupos que definen el consumo industrial de un día de invierno en la provincia de Málaga. Los grupos, junto con sus desviaciones típicas, se encuentran en la Figura 5.89. A pesar de que el número de clústers elegido para los fines de semana es trece, el gráfico de los sábados tan solo representa doce grupos, mientras que el del domingo, representa once. Esto se debe a que el Grupo 12 de los sábados y los grupos 6 y 12 de los domingos, han sido eliminados ya que solo reflejaban información acerca de un conjunto de datos, por lo que no aportaba información relevante al estudio predictivo.

A primera vista, destaca un conjunto de grupos que presentan los valores de consumo más elevados, así como un patrón errático, que no se asemeja a la forma esperada de un consumo del tipo industrial, como la encontrada en la sección 5.3.2. Es el caso de los grupos 2, 4 y 12 de los días laborales y de los grupos 7 y 13 de los sábados. Dichos patrones presentan grandes fluctuaciones a lo largo del día y alcanzan valores de consumo muy elevados, llegando a tener picos de más de 30000 kWh. Cabe destacar que de entre estos patrones, los del sábado presentan valores de consumo más elevados por la mañana, lo que se puede asociar a empresas que solamente trabajen media

jornada los sábados. Sin embargo, los patrones erráticos de los días laborales, solamente presentan una clara depresión en las horas de sueño, además, los grupos 2 y 4 presentan una forma más similar a la esperada para un consumo industrial. La tendencia errática y los valores de consumo tan elevados no representan la mayor parte del consumo industrial de un día de invierno en la provincia de Málaga, es más, representan porcentajes mínimos de dicho consumo, Tabla 5.110 y Tabla 5.111.

Siguiendo con esta línea, se encuentran el Grupo 2 de los sábados y el Grupo 11 de los domingos. Estos patrones también presentan una tendencia errática, pero con niveles de consumo menores, en torno a los 9000 kWh. No muestran en ninguna franja horaria una depresión clara que indique el cese de la actividad industrial. Estos grupos tampoco representan un gran porcentaje del consumo objeto de estudio de esta sección, Tabla 5.111 y Tabla 5.112.

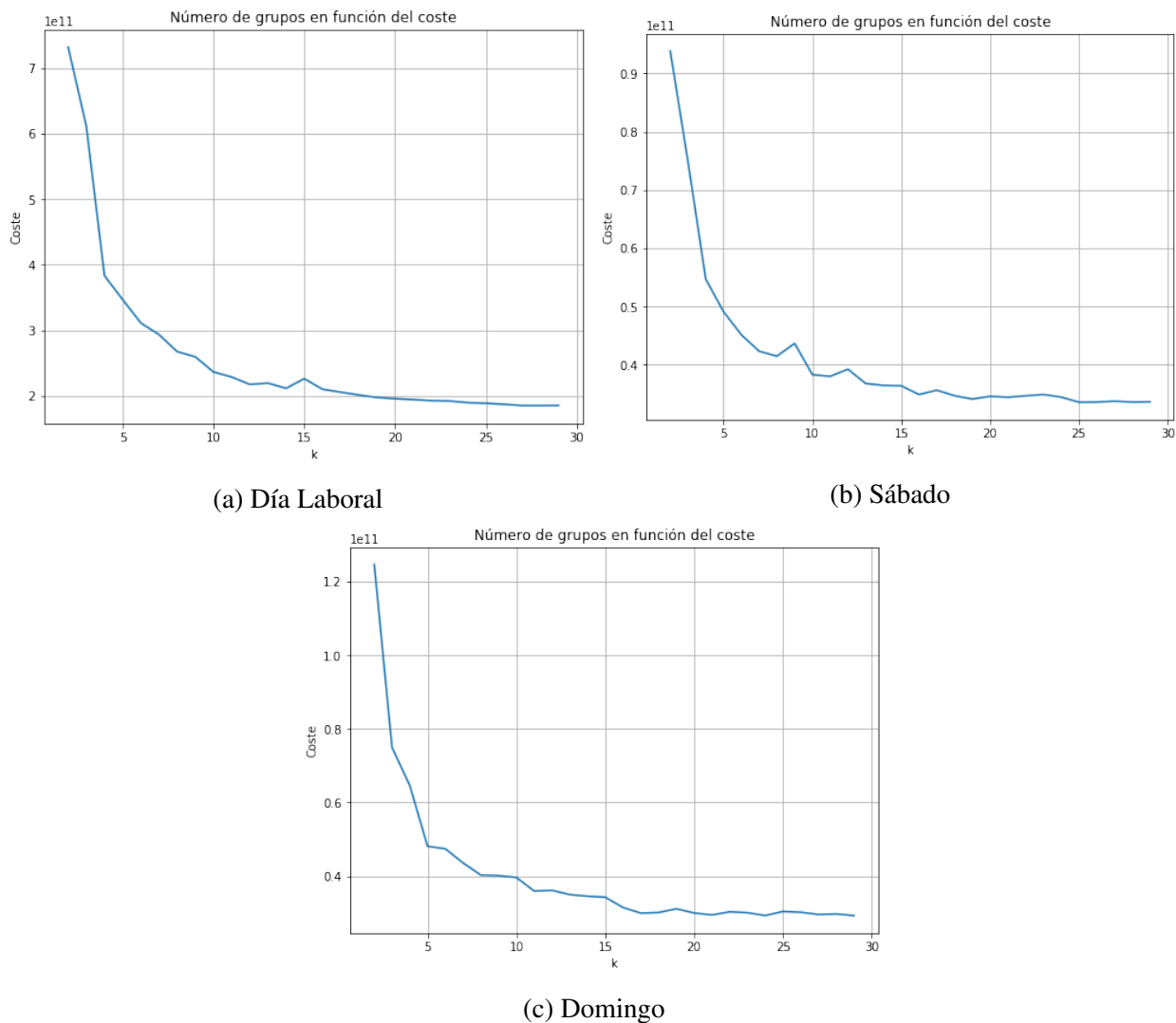


Figura 5.88.: Número de grupos en función de la suma de las distancias a los centroides



Figura 5.89.: Consumo industrial de un día de invierno en Málaga

Un conjunto de grupos que llama la atención es el formado por aquellos que presentan, aproximadamente dos niveles de consumo. Los máximos de dicho consumo se dan desde aproximadamente las 17:00 hasta las 9:00. Este conjunto de grupos está formado por los grupos 3, 8 y 9 en el gráfico de días laborales, por los grupos 3, 4 y 5 en los sábados y por los grupos 2, 4, 5, 8 y 13 en los domingos. Este tipo de comportamiento se puede asociar al seguido por locales que comienzan su actividad después de comer y continúan su actividad durante toda la noche. Sin embargo, este patrón es seguido por un porcentaje mínimo de clientes, Tabla 5.110, Tabla 5.111, Tabla 5.112. Patrones similares ya se había detectado en el caso del consumo industrial de la provincia de Barcelona, encontrándose la principal diferencia en la duración exacta de la depresión del consumo, dependiendo de la estación.

La tendencia de consumo industrial encontrada en la sección 5.3.2, se caracteriza por la presencia de dos picos claros de consumo, uno por la mañana y otro por la tarde. Variantes de esta tendencia son seguidas por los siguientes grupos. Por un lado, cabe mencionar que ciertos grupos siguen un patrón que no presentan el pico de consumo que se espera por la tarde. Alcanzan un pico por la mañana y el resto de la jornada laboral presentan una tendencia descendente. Es el caso del Grupo 10 en los días laborales, del Grupo 9 en el caso de los sábados y del Grupo 10 de los domingos. Por otro lado, se encuentra el Grupo 7 de los domingos, que presenta una tendencia creciente a lo largo de la jornada laboral hasta alcanzar un único pico durante la tarde, asemejándose al consumo residencial. Este conjunto de grupos, ya presentan un porcentaje más relevante del consumo industrial de un día de invierno en la provincia de Málaga, Tabla 5.110, Tabla 5.111, Tabla 5.112.

El Grupo 11 de los días laborales, Figura 5.89, es el que presenta una tendencia más similar a la encontrada en la sección 5.3.2 para el consumo industrial. Además, cabe mencionar que el resto de grupos presentan una tendencia más constante a lo largo del día. En estos grupos se puede inferir una pequeña subida del consumo a las horas donde deberían encontrar los picos principales, es muy probable que no se puedan distinguir por la escala de la imagen, pero que si que sigan dicha tendencia. Una vez más, los grupos con los niveles de consumo más bajos, son los que representan un porcentaje mayor del consumo industrial de un día de invierno en la provincia de Málaga; es el caso del Grupo 1 para todos los días de la semana, Tabla 5.110, Tabla 5.111, Tabla 5.112.

Uno de los objetivos principales de este proyecto es proporcionar información útil para las distribuidoras de energía eléctrica para que puedan lanzar campañas de publicidad orientadas a determinados grupos de clientes. A las distribuidoras energéticas les conviene aplanar los niveles de consumo a lo largo del día para poder optimizar la producción de las centrales. En este caso, principalmente, se pueden lanzar dos campañas publicitaria. Una campaña tendría como objetivo conseguir la redistribución del consumo de aquellos grupos que presentan una gran depresión entre las 9:00 y las 17:00. Para conseguir dicha redistribución, se les podrían lanzar ofertas en las que se indicasen determinadas franjas horarias en las que el precio fuese más reducido, con la esperanza de que desplazasen su consumo a dichas franjas. Otra tendría como objetivo reducir los niveles de consumo de aquellos grupos que sobresalen por encima del resto y presenta una tendencia errática, lo cual se puede conseguir mediante el establecimiento de un límite de consumo a partir del cual el precio de la electricidad aumenta para fomentar la reducción del consumo.

Con el modelo predictivo establecido, es necesario evaluar la veracidad del mismo. Motivo por el cual se comparan los porcentajes de representación de cada grupo obtenidos en la fase de predicción, Tabla 5.110, Tabla 5.111, Tabla 5.112, con los obtenidos en la fase de validación, Tabla 5.113, Tabla 5.114, Tabla 5.115. En este caso, el resultado de la comparación es una variación media por cada grupo del 0.16% en los días laborales, del 0.42% en los sábados y del 0.21% en los domingos. Estos valores son perfectamente asumibles, por lo que el modelo calculado es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	62.13 %	Grupo 5	1.47 %	Grupo 9	2.00 %
Grupo 2	0.06 %	Grupo 6	23.07 %	Grupo 10	0.44 %
Grupo 3	0.93 %	Grupo 7	6.01 %	Grupo 11	3.59 %
Grupo 4	0.04 %	Grupo 8	0.22 %	Grupo 12	0.04 %

Tabla 5.110.: Predicción del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	61.87 %	Grupo 6	0.45 %	Grupo 11	6.43 %
Grupo 2	0.15 %	Grupo 7	0.05 %	Grupo 12	0.02 %
Grupo 3	0.22 %	Grupo 8	1.16 %	Grupo 13	0.03 %
Grupo 4	0.74 %	Grupo 9	4.07 %		
Grupo 5	1.96 %	Grupo 10	22.86 %		

Tabla 5.111.: Predicción del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	63.95 %	Grupo 6	0.02 %	Grupo 11	0.17 %
Grupo 2	0.21 %	Grupo 7	6.31 %	Grupo 12	0.03 %
Grupo 3	5.87 %	Grupo 8	0.69 %	Grupo 13	0.54 %
Grupo 4	0.93 %	Grupo 9	1.07 %		
Grupo 5	0.61 %	Grupo 10	18.60 %		

Tabla 5.112.: Predicción del porcentaje de datos asociados a cada grupo, Domingo

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	61.41 %	Grupo 5	1.65 %	Grupo 9	1.970 %
Grupo 2	0.09 %	Grupo 6	23.50 %	Grupo 10	0.43 %
Grupo 3	0.77 %	Grupo 7	6.26 %	Grupo 11	3.55 %
Grupo 4	0.02 %	Grupo 8	0.31 %	Grupo 12	0.04 %

Tabla 5.113.: Validación del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	59.61 %	Grupo 6	0.25 %	Grupo 11	6.69 %
Grupo 2	0.05 %	Grupo 7	0.00 %	Grupo 12	0.00 %
Grupo 3	0.25 %	Grupo 8	1.08 %	Grupo 13	0.10 %
Grupo 4	1.18 %	Grupo 9	4.013 %		
Grupo 5	1.97 %	Grupo 10	24.72 %		

Tabla 5.114.: Validación del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	64.08 %	Grupo 6	0.00 %	Grupo 11	0.19 %
Grupo 2	0.19 %	Grupo 7	6.73 %	Grupo 12	0.00 %
Grupo 3	6.68 %	Grupo 8	0.66 %	Grupo 13	1.42 %
Grupo 4	0.62 %	Grupo 9	1.00 %		
Grupo 5	0.33 %	Grupo 10	18.10 %		

Tabla 5.115.: Validación del porcentaje de datos asociados a cada grupo, Domingo

5.4.3.2. Perfiles del consumo de un día de primavera en la provincia de Málaga

El objetivo de esta sección es encontrar los distintos patrones que representan el consumo residencial e industrial de un día de primavera en la provincia de Málaga. Para alcanzar dicha finalidad, se va a seguir la dinámica empleada en las secciones anteriores, en primer lugar, se obtiene el número de clústers a emplear en cada caso, para después calcular los distintos patrones de consumo mediante la implementación del algoritmo K-Means.

Consumo residencial

Una vez más, a pesar de que el número de clústers empleado en cada tipo de día de la semana es diferente, se ha decidido juntar el estudio de todo el consumo residencial de un día de primavera en la provincia de Málaga, en un sólo análisis debido a las similitudes presentes entre los grupos encontrados. Con el objetivo de concretar el número de clústers a emplear en cada tipo de día de la semana, se ha recurrido al cálculo de la disminución porcentual de la suma de varianzas en función del número de grupos utilizados, ya que las funciones de coste calculadas, Figura 5.90, carecen de “codo” discernible a primera vista. Tras completar el proceso, se ha decidido emplear quince clúster en el caso de los días laborales, quince en los sábados y once en los domingos.

Una vez elegido el número de clústers a emplear, se ha implementado el algoritmo K-Means con la finalidad de encontrar los grupos que definen el consumo residencial de un día de primavera en la provincia de Málaga. Los grupos, junto con sus desviaciones típicas, se encuentran en la Figura 5.91. A pesar del número de clústers elegido en cada caso, los gráficos de los sábados presentan un grupo menos y los gráficos de los domingos presentan dos grupos menos. Lo cual se

debe a que ciertos grupos solamente representaba información acerca de un conjunto de datos, por lo que no proporcionaban información útil para este estudio y se han eliminado.

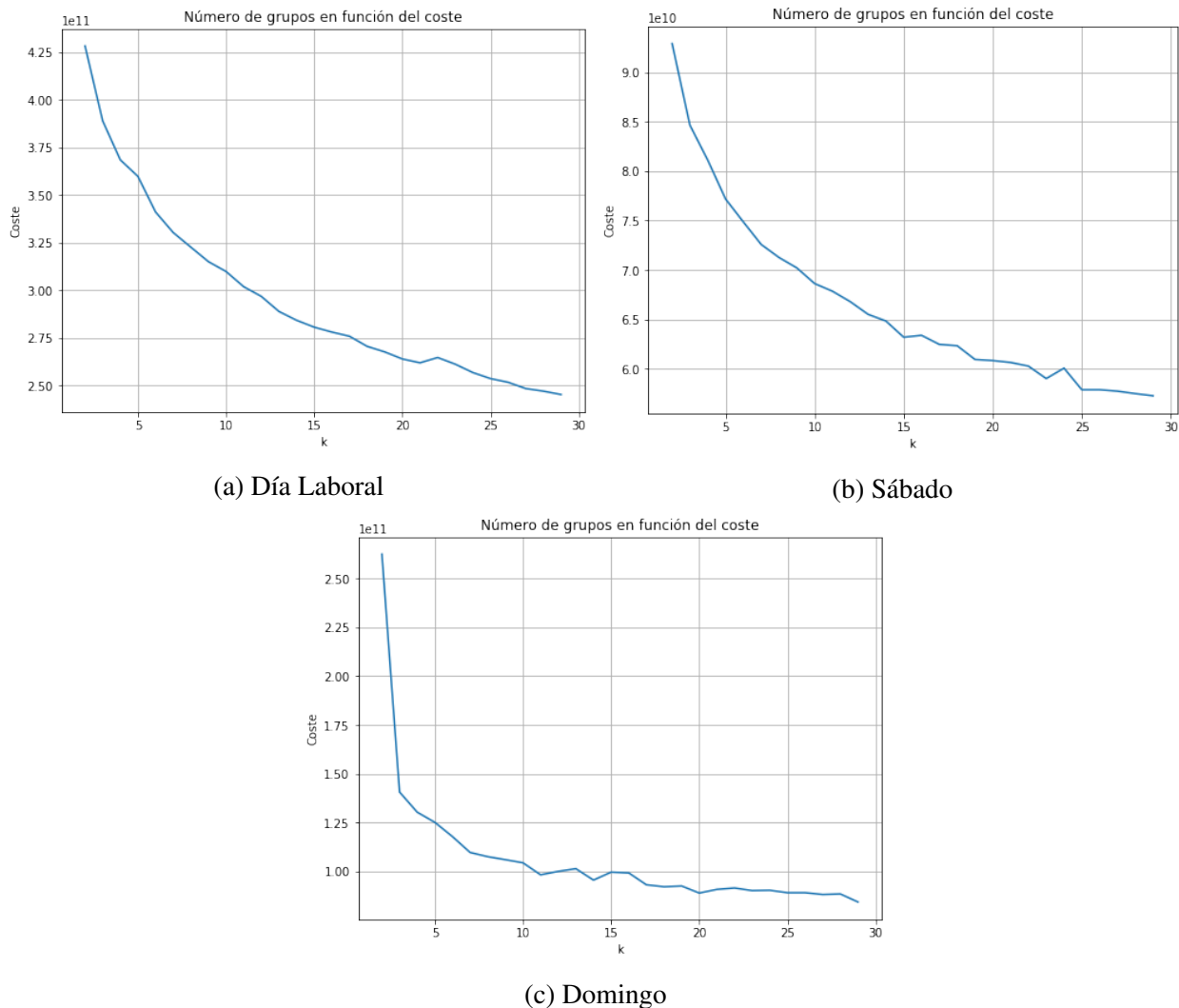


Figura 5.90.: Número de grupos en función de la suma de las distancias a los centroides

A primera vista, en todos los tipos de día, destaca la presencia de un grupo que sobresale por encima del resto, debido a los valores tan elevados que toma a lo largo del día. Es el caso del Grupo 13 en los días laborales, del Grupo 7 en los sábados y de los grupo 2 y 4 en los domingos. Además, no siguen la tendencia ascendente a lo largo del día que se espera del consumo residencial, como se estableció en la sección 5.3.2. Sorprendentemente, encajaría mejor con el gráfico esperado del consumo industrial, motivo por el cual representa un porcentaje ínfimo del consumo residencial de un día de primavera en la provincia de Málaga, Tabla 5.116, Tabla 5.117 y Tabla 5.118. Este tipo de comportamiento se puede dar en viviendas en las que durante las mañanas la limpieza sea realizada por una persona externa, que termina su jornada antes de la hora de comer, y a la hora de cenar el residente vuelve al hogar.

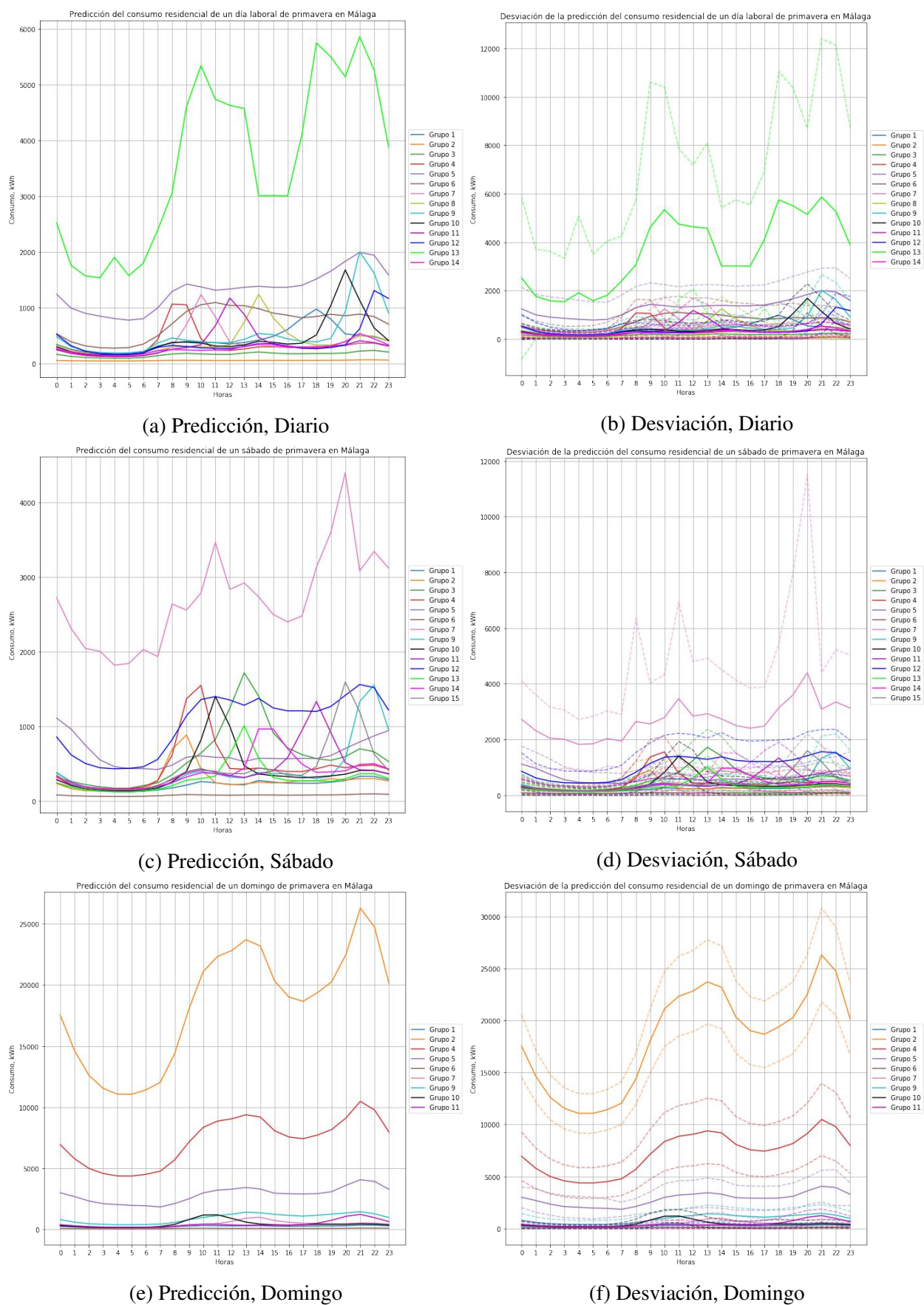


Figura 5.91.: Consumo residencial de un día de primavera en Málaga

También llama la atención la presencia de un grupo cuyo consumo decae durante las horas de luz, es el caso del Grupo 5 en los sábados. Este patrón es totalmente opuesto al esperado cuando se trata de consumo residencial. Es muy probable que los clientes que pertenecen a estos grupos trabajen en el turno de noche, motivo por el cual la depresión de su consumo se da durante el día. Además, representan un pequeño porcentaje del consumo residencial de un día de primavera en la provincia de Málaga, Tabla 5.116, Tabla 5.117 y Tabla 5.118.

Además, cabe destacar el conjunto de grupos, que solamente presentan el pico de mayor consumo por las mañanas. La posición de dicho pico varía a lo largo del tiempo entre las 8:00 y las 18:00, aproximadamente. Es el caso de los grupos 1, 4, 7, 8 y 11 en el gráfico de los días laborales, los grupos 2, 3, 4, 10, 11, 13 y 14 en el de los sábados y los grupos 7 y 10 en el de los domingos. Este tipo de comportamiento se asocia con el de aquellas personas que tan solo emplean recursos eléctricos durante la hora de la comida, lo que implica que o bien reducen drásticamente el uso de aparatos eléctricos o emplean otras fuentes de energía en ese espacio de tiempo.

El patrón detectado por primera vez en el consumo residencial de un día de invierno en la provincia de Málaga, se vuelve a manifestar en este caso, por lo que se puede asumir que guarda relación con los hábitos de consumo de los habitantes de la provincia de Málaga en particular. Este patrón se caracteriza por seguir la tendencia esperada de un consumo residencial, hasta el mediodía. A partir de dicho momento, el consumo se vuelve prácticamente constante, es decir, pierde la tendencia de subida propia de un consumo residencial. Es el caso del Grupo 6 de los días laborales y del Grupo 12 de los sábados. Este tipo de comportamiento puede ser el seguido por aquellas personas que pasan toda la tarde en sus viviendas y no abusan en un periodo concreto de tiempo del uso de electrodomésticos u otros aparatos que generen un gran consumo eléctrico.

El resto de los grupos representados, reflejan el patrón de consumo residencial esperado, tomando valores mayores o menores y experimentando mayor o menor diferencia entre los dos picos principales del día. Una vez más se da el que grupo con los niveles de consumo menores a lo largo de todo el día, es el que representa un mayor porcentaje del consumo.

Con la finalidad de aplanar los niveles de consumo a lo largo del día, ya que de esta forma las distribuidoras eléctricas pueden optimizar sus centrales, se pueden lanzar, principalmente, dos campañas publicitarias. Una de ellas se debe centrar en la reducción de los niveles de consumo de aquellos que sobresalen por encima del resto. Por otro lado, una campaña se podría centrar en reducir la diferencia de ciertos picos de consumo con respecto al consumo generado el resto del día. Esto se puede conseguir mediante el lanzamiento de ofertas de energía en determinadas franjas horarias, medida que motiva la redistribución del consumo. Además, se puede establecer un límite de consumo a partir del cual el precio de la electricidad aumenta para fomentar la reducción del consumo.

Con el modelo establecido, es necesario evaluar la fiabilidad del mismo. Para ello, se comparan los porcentajes de representación de cada grupos obtenidos en la fase de predicción, Tabla 5.116, Tabla 5.117 y Tabla 5.118, con los obtenidos en la fase de validación, Tabla 5.119, Tabla 5.120 y Tabla 5.121. En este caso, la variación es del 0.07 % en los días laborales, del 0.14 % en los sábados y de 0.13 % los domingos, valores que son asumibles, por lo que el modelo es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	3.66 %	Grupo 6	2.20 %	Grupo 11	3.97 %
Grupo 2	31.25 %	Grupo 7	3.50 %	Grupo 12	2.86 %
Grupo 3	26.20 %	Grupo 8	4.10 %	Grupo 13	0.03 %
Grupo 4	3.14 %	Grupo 9	1.90 %	Grupo 14	14.38 %
Grupo 5	0.52 %	Grupo 10	2.29 %	Grupo 15	0.001 %

Tabla 5.116.: Predicción del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	26.06 %	Grupo 6	38.76 %	Grupo 11	2.80 %
Grupo 2	4.60 %	Grupo 7	0.07 %	Grupo 12	1.21 %
Grupo 3	2.03 %	Grupo 8	0.002 %	Grupo 13	6.00 %
Grupo 4	2.14 %	Grupo 9	2.99 %	Grupo 14	5.67 %
Grupo 5	1.77 %	Grupo 10	3.15 %	Grupo 15	2.78 %

Tabla 5.117.: Predicción del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	32.64 %	Grupo 5	0.15 %	Grupo 9	2.23 %
Grupo 2	0.02 %	Grupo 6	44.28 %	Grupo 10	5.37 %
Grupo 3	0.008 %	Grupo 7	9.49 %	Grupo 11	5.76 %
Grupo 4	0.06 %	Grupo 8	0.002 %		

Tabla 5.118.: Predicción del porcentaje de datos asociados a cada grupo, Domingo

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	3.65 %	Grupo 6	2.23 %	Grupo 11	4.04 %
Grupo 2	31.40 %	Grupo 7	3.41 %	Grupo 12	2.88 %
Grupo 3	25.83 %	Grupo 8	4.17 %	Grupo 13	0.04 %
Grupo 4	3.25 %	Grupo 9	1.98 %	Grupo 14	14.40 %
Grupo 5	0.46 %	Grupo 10	2.27 %	Grupo 15	0.00 %

Tabla 5.119.: Validación del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	26.14 %	Grupo 6	38.58 %	Grupo 11	2.7 %
Grupo 2	4.85 %	Grupo 7	0.06 %	Grupo 12	1.24 %
Grupo 3	2.09 %	Grupo 8	0.00 %	Grupo 13	5.48 %
Grupo 4	2.11 %	Grupo 9	3.32 %	Grupo 14	5.91 %
Grupo 5	1.77 %	Grupo 10	3.08 %	Grupo 15	2.66 %

Tabla 5.120.: Validación del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	32.68 %	Grupo 5	0.20 %	Grupo 9	2.50 %
Grupo 2	0.01 %	Grupo 6	44.47 %	Grupo 10	5.20 %
Grupo 3	0.00 %	Grupo 7	9.63 %	Grupo 11	5.27 %
Grupo 4	0.05 %	Grupo 8	0.00 %		

Tabla 5.121.: Validación del porcentaje de datos asociados a cada grupo, Domingo

Consumo industrial

Una vez estudiado el consumo residencial en primavera para todos los días de la semana, es el turno del consumo industrial. En primer lugar, antes de encontrar los patrones del consumo industrial de un día de primavera en la provincia de Málaga, es necesario obtener el número de clústers a emplear. Para lograr dicho objetivo, se ha creado el siguiente gráfico, Figura 5.92. Como se puede observar, es “codo” de la función no se puede observar a primer vista, por lo que se ha recurrido al cálculo del porcentaje de disminución de la varianza. Al finalizar el proceso se concluye que el número de clústers a emplear es diez en los días laborales y once en los fines de semana. A pesar de haber elegido distintas cantidades de clústers, todos los días de la semana se van a estudiar a la vez debido a las similitudes presentes en los patrones encontrados.

Con el número de clústers elegido para cada tipo de día de la semana, se ha podido implementar el algoritmo K-Means, con la finalidad de obtener los distintos grupos de clientes que representan el consumo industrial a lo largo de un día de verano en la provincia de Málaga. Los grupos, junto con sus respectivas desviaciones típicas, se encuentran en la Figura 5.93. Aunque el número de clústers elegido es once en los fines de semana, los gráficos tan solo reflejan información relativa a diez grupos, el Grupo 11 de los sábados y el Grupo 3 de los domingos ha sido eliminados. Lo cual se debe a que cierto grupo solamente representaba información acerca de un conjunto de datos, por lo que no proporcionaba información útil para este estudio.

A primera vista, destaca un conjunto de grupos que presentan los valores de consumo más elevados. Es el caso del Grupo 2 de los días laborales y del Grupo 6 de los domingos. Dichos patrones presentan una tendencia similar a la esperada para el consumo industrial a lo largo del día y alcanzan valores de consumo muy elevados, llegando a tener picos de más de 20000 kWh. Este tipo de comportamiento de consumo se asocia con el de empresas con grandes instalaciones que

generan un gran consumo, manteniendo el patrón habitual del consumo industrial. Estas empresas son una minoría en el consumo industrial de invierno en la provincia de Málaga, Tabla 5.122 y Tabla 5.124.

También cabe destacar la presencia de un grupo que presenta una tendencia errática a los largo del día, patrones similares se observaron en el consumo industrial de un día de primavera en la provincia de Málaga. Se trata del Grupo 2 de los sábados. Dicho grupo presentan valores de consumo más elevados por la mañana, lo que se puede asociar a empresas que solamente trabajen media jornada los sábados.

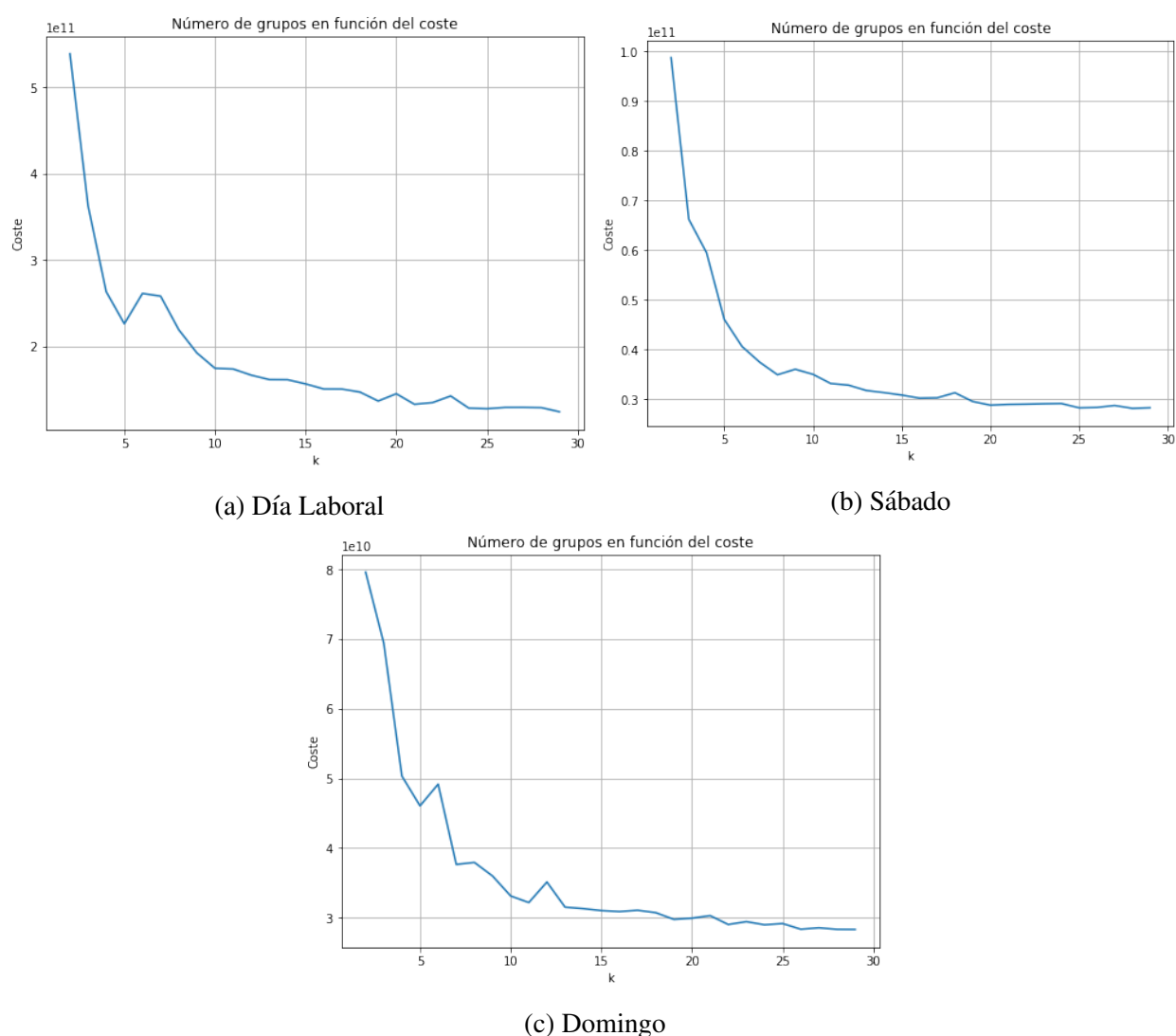


Figura 5.92.: Número de grupos en función de la suma de las distancias a los centroides



Figura 5.93.: Consumo industrial de un día de primavera en Málaga

Un conjunto de grupos que llama la atención es el formado por aquellos que presentan, aproximadamente dos niveles de consumo. Los máximos de dicho consumo se dan desde aproximadamente las 20:00 hasta las 8:00. Este conjunto de grupos está formado por los grupos 3, 4 y 5 en el gráfico de días laborales, por los grupos 4 y 10 en los sábados y por los grupos 2, 5 y 11 en los domingos. Este tipo de comportamiento se puede asociar al seguido por locales que comienzan su actividad después de comer y continúan su actividad durante toda la noche. Sin embargo, este patrón es seguido por un porcentaje mínimo de clientes, Tabla 5.122, Tabla 5.123, Tabla 5.124. Patrones similares ya se había detectado en el caso del consumo industrial de la provincia de Barcelona, así como en el consumo industrial de un día de invierno en la provincia de Málaga, encontrándose la principal diferencia en la duración exacta de la depresión del consumo, dependiendo de la estación.

La tendencia de consumo industrial encontrada en la sección 5.3.2, se caracteriza por la presencia de dos picos claros de consumo, uno por la mañana y otro por la tarde. Variantes de esta tendencia son seguidas por los siguientes grupos. Por un lado, cabe mencionar que ciertos grupos siguen un patrón que no presentan el pico de consumo que se espera por la tarde. Alcanzan un pico por la mañana y el resto de la jornada laboral presentan una tendencia descendente. Es el caso de los grupos 9 y 10 en los días laborales y del Grupo 9 en el caso de los sábados.

Un conjunto de grupos sí que presenta una tendencia más similar a la encontrada en la sección 5.3.2 para el consumo industrial. Es el caso del Grupo 7 de los días laborales, de los grupos 5 y 7 de los sábados y del Grupo 7 de los domingos. Además, cabe mencionar que el resto de grupos presentan una tendencia más constante a lo largo del día. En estos grupos se puede inferir una pequeña subida del consumo a las horas donde deberían encontrar los picos principales, es muy probable que no se puedan distinguir por la escala de la imagen, pero que si que sigan dicha tendencia. Una vez más, los grupos con los niveles de consumo más bajos, son los que representan un porcentaje mayor del consumo industrial de un día de primavera en la provincia de Málaga; es el caso del Grupo 1 para todos los días de la semana, Tabla 5.122, Tabla 5.123, Tabla 5.124.

Uno de los objetivos principales de este proyecto es proporcionar información útil para las distribuidoras de energía eléctrica para que puedan lanzar campañas de publicidad orientadas a determinados grupos de clientes. A las distribuidoras energéticas les conviene aplanar los niveles de consumo a lo largo del día para poder optimizar la producción de las centrales. En este caso, principalmente, se pueden lanzar dos campañas publicitaria. Una campaña tendría como objetivo conseguir la redistribución del consumo de aquellos grupos que presentan una gran depresión entre las 8:00 y las 20:00. Para conseguir dicha redistribución, se les podrían lanzar ofertas en las que se indicasen determinadas franjas horarias en las que el precio fuese más reducido, con la esperanza de que desplazasen su consumo a dichas franjas. Otra tendría como objetivo reducir los niveles de consumo de aquellos grupos que sobresalen por encima del resto, lo cual se puede conseguir mediante el establecimiento de un límite de consumo a partir del cual el precio de la electricidad aumenta para fomentar la reducción del consumo.

Con el modelo predictivo establecido, es necesario evaluar la veracidad del mismo. Motivo por el cual se comparan los porcentajes de representación de cada grupo obtenidos en la fase de predicción, Tabla 5.122, Tabla 5.123, Tabla 5.124, con los obtenidos en la fase de validación, Tabla

5.125, Tabla 5.126, Tabla 5.127. En este caso, el resultado de la comparación es una variación media por cada grupo del 0.16% en los días laborales, del 0.26% en los sábados y del 0.51% en los domingos. Estos valores son perfectamente asumibles, por lo que el modelo calculado es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	58.45 %	Grupo 6	28.43 %
Grupo 2	0.12 %	Grupo 7	6.58 %
Grupo 3	0.20 %	Grupo 8	2.46 %
Grupo 4	1.15 %	Grupo 9	0.49 %
Grupo 5	1.90 %	Grupo 10	0.22 %

Tabla 5.122.: Predicción del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	58.61 %	Grupo 5	0.38 %	Grupo 9	3.25 %
Grupo 2	0.12 %	Grupo 6	27.71 %	Grupo 10	1.79 %
Grupo 3	1.04 %	Grupo 7	0.26 %	Grupo 11	0.01 %
Grupo 4	1.04 %	Grupo 8	5.79 %		

Tabla 5.123.: Predicción del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	50.50 %	Grupo 5	1.18 %	Grupo 9	8.98 %
Grupo 2	1.07 %	Grupo 6	0.16 %	Grupo 10	31.29 %
Grupo 3	0.02 %	Grupo 7	0.88 %	Grupo 11	1.07 %
Grupo 4	0.30 %	Grupo 8	4.54 %		

Tabla 5.124.: Predicción del porcentaje de datos asociados a cada grupo, Domingo

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	58.29 %	Grupo 6	27.97 %
Grupo 2	0.17 %	Grupo 7	7.06 %
Grupo 3	0.20 %	Grupo 8	2.63 %
Grupo 4	1.22 %	Grupo 9	0.46 %
Grupo 5	1.86 %	Grupo 10	0.15 %

Tabla 5.125.: Validación del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	59.35 %	Grupo 5	0.71 %	Grupo 9	2.49 %
Grupo 2	0.13 %	Grupo 6	27.54 %	Grupo 10	1.51 %
Grupo 3	1.16 %	Grupo 7	0.13 %	Grupo 11	0.00 %
Grupo 4	1.29 %	Grupo 8	5.69 %		

Tabla 5.126.: Validación del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	53.04 %	Grupo 5	1.28 %	Grupo 9	8.36 %
Grupo 2	1.19 %	Grupo 6	0.19 %	Grupo 10	30.77 %
Grupo 3	0.00 %	Grupo 7	0.47 %	Grupo 11	0.85 %
Grupo 4	0.09 %	Grupo 8	3.75 %		

Tabla 5.127.: Validación del porcentaje de datos asociados a cada grupo, Domingo

5.4.3.3. Perfiles del consumo de un día de verano en la provincia de Málaga

El objetivo de esta sección es encontrar los distintos patrones que representan el consumo residencial e industrial de un día de verano en la provincia de Málaga. Para alcanzar dicha finalidad, se va a seguir la dinámica empleada en las secciones anteriores, en primer lugar, se obtiene el número de clústers a emplear en cada caso, para después calcular los distintos patrones de consumo mediante la implementación del algoritmo K-Means.

Consumo residencial

En este caso, el número de clústers elegido para todos los días de la semana es el mismo. Con el objetivo de encontrar el número de clústers a emplear en cada tipo de día de la semana, ya que el “codo” de las funciones de coste, Figura 5.94 es difícil de discernir a simple vista, se ha recurrido al cálculo del porcentaje de la disminución de la suma de varianzas con cada aumento del número de clústers. Tras realizar esa operación para cada uno de los tres casos, se ha llegado a la conclusión de que el número de clústers a emplear es quince.

Una vez elegido el número de clústers a emplear, se ha implementado el algoritmo K-Means con la finalidad de encontrar doce grupos que definen el consumo residencial de un día de verano en la provincia de Málaga. Los grupos, junto con sus desviaciones típicas, se encuentran en la Figura 5.95.

A primera vista, en todos los tipos de día, destaca la presencia de un grupo que sobresale por encima del resto, debido a los valores tan elevados que toma a lo largo del día. Es el caso del Grupo 10 en los días laborales, del Grupo 12 en los sábados y del Grupo 4 en los domingos. Además, no siguen la tendencia ascendente a lo largo del día que se espera del consumo residencial, como se estableció en la sección 5.3.2. Sorprendentemente, el Grupo 10 de los sábados encajaría menor

con el gráfico esperado del consumo industrial, motivo por el cual representa un porcentaje ínfimo del consumo residencial de un día de verano en la provincia de Málaga, Tabla 5.129. Este tipo de comportamiento se puede dar en viviendas en las que durante las mañanas la limpieza sea realizada por una persona externa, que termina su jornada antes de la hora de comer, y a la hora de cenar el residente vuelve al hogar. Los otros dos grupos, presentan la tendencia esperada al tratar con consumo residencial pero con valores de consumo mayores, lo que se asocia a casas más grandes, que generan un mayor consumo. Estos grupos también representan un porcentaje mínimo del consumo, Tabla 5.128 y Tabla 5.82.

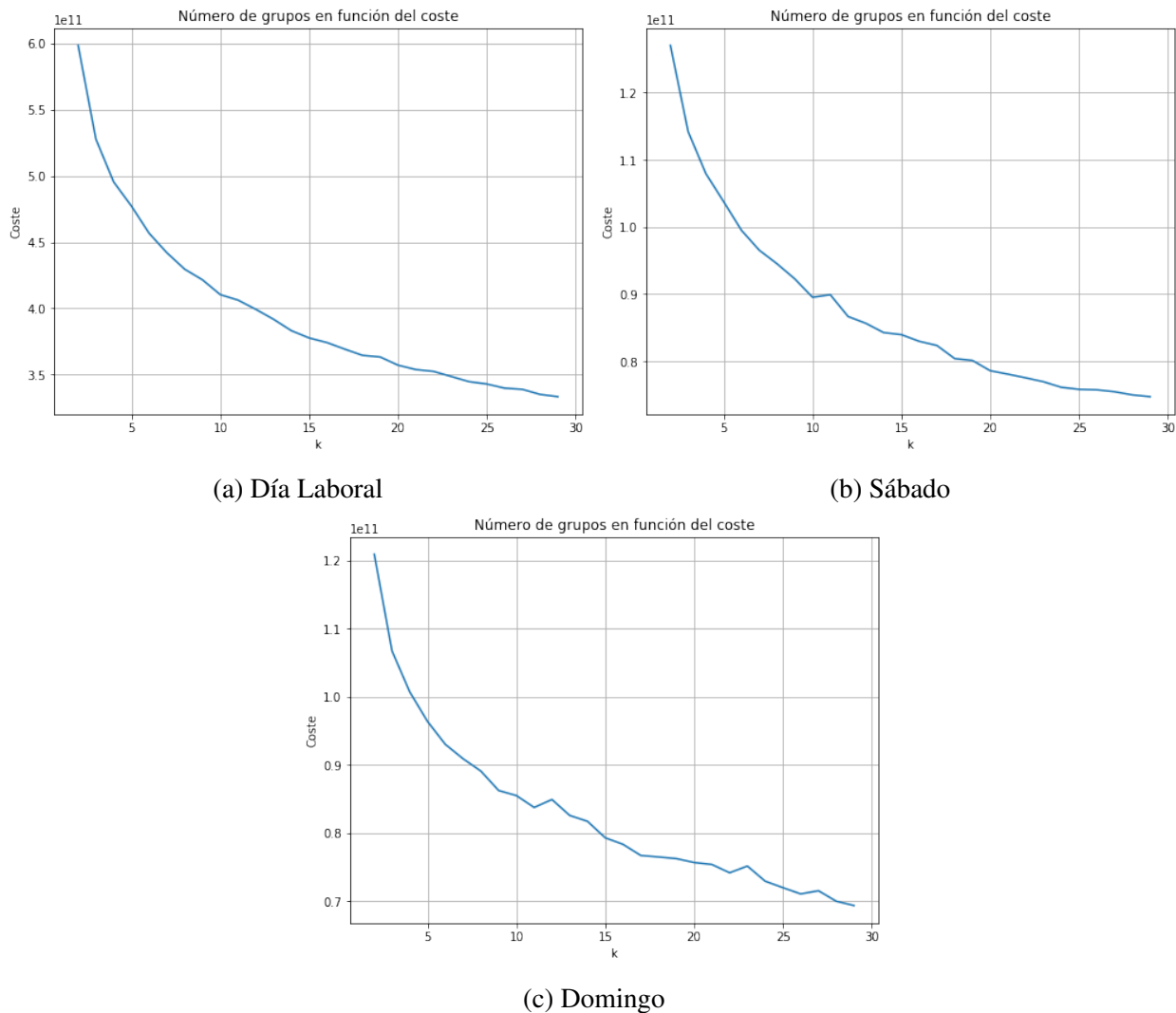


Figura 5.94.: Número de grupos en función de la suma de las distancias a los centroides

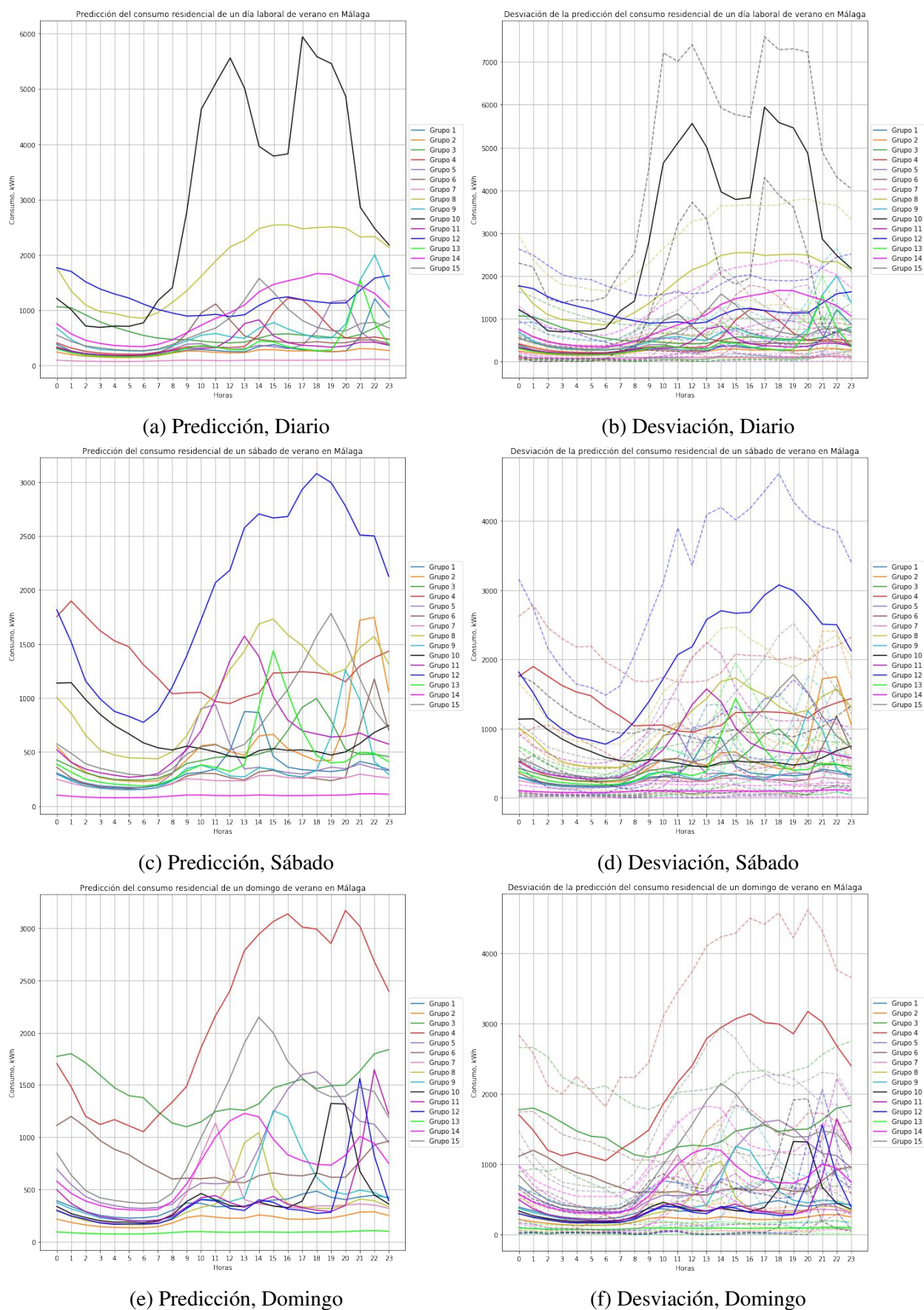


Figura 5.95.: Consumo residencial de un día de verano en Málaga

También llama la atención la presencia de un conjunto de grupos cuyo consumo decae durante las horas de luz, es el caso de los grupos 3 y 12 de los días laborales, de los grupos 4 y 10 de los sábados y de los grupos 3 y 6 de los domingos. Este patrón es totalmente opuesto al esperado cuando se trata de consumo residencial. Es muy probable que los clientes que pertenecen a estos grupos trabajen en el turno de noche, motivo por el cual la depresión de su consumo se da durante el día. Además, representan un pequeño porcentaje del consumo residencial de un día de verano en la provincia de Málaga, Tabla 5.128, Tabla 5.129 y Tabla 5.130.

Además, cabe destacar el conjunto de grupos, que solamente presentan el pico de mayor consumo entre las primeras horas de la mañana y la mitad de la tarde. La posición de dicho pico varía a lo largo del tiempo entre las 8:00 y las 18:00, aproximadamente. Es el caso de los grupos 4, 6, 11 y 15 en el gráfico de los días laborales, los grupos 1, 3, 5, 8, 11 y 13 en el de los sábados y los grupos 7, 8, 9, 14 y 15 en el de los domingos. Este tipo de comportamiento se asocia con el de aquellas personas que tan solo emplean recursos eléctricos durante la hora de la comida, lo que implica que o bien reducen drásticamente el uso de aparatos eléctricos o emplean otras fuentes de energía en ese espacio de tiempo.

El resto de los grupos representados, reflejan el patrón de consumo residencial esperado, tomando valores mayores o menores y experimentando mayor o menor diferencia entre los dos picos principales del día. Una vez más se da el que grupo con los niveles de consumo menores a lo largo de todo el día, es el que representa un mayor porcentaje del consumo.

Con la finalidad de aplanar los niveles de consumo a lo largo del día, ya que de esta forma las distribuidoras eléctricas pueden optimizar sus centrales, se pueden lanzar, principalmente, dos campañas publicitarias. Una de ellas se debe centrar en la reducción de los niveles de consumo de aquellos que sobresalen por encima del resto. Por otro lado, una campaña se podría centrar en reducir la diferencia de ciertos picos de consumo con respecto al consumo generado el resto del día. Esto se puede conseguir mediante el lanzamiento de ofertas de energía en determinadas franjas horarias, medida que motiva la redistribución del consumo. Además, se puede establecer un límite de consumo a partir del cual el precio de la electricidad aumenta para fomentar la reducción del consumo.

Con el modelo establecido, es necesario evaluar la fiabilidad del mismo. Para ello, se comparan los porcentajes de representación de cada grupo obtenidos en la fase de predicción, Tabla 5.128, Tabla 5.129 y Tabla 5.130, con los obtenidos en la fase de validación, Tabla 5.131, Tabla 5.132 y Tabla 5.133. En este caso, la variación es del 0.07 % en los días laborales, del 0.16 % en los sábados y de 0.15 % los domingos, valores que son asumibles, por lo que el modelo calculado es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	4.85 %	Grupo 6	3.62 %	Grupo 11	8.45 %
Grupo 2	27.43 %	Grupo 7	34.36 %	Grupo 12	0.81 %
Grupo 3	2.09 %	Grupo 8	0.34 %	Grupo 13	4.48 %
Grupo 4	3.44 %	Grupo 9	2.10 %	Grupo 14	1.54 %
Grupo 5	3.83 %	Grupo 10	0.04 %	Grupo 15	2.63 %

Tabla 5.128.: Predicción del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	6.94 %	Grupo 6	5.67 %	Grupo 11	2.11 %
Grupo 2	3.03 %	Grupo 7	26.77 %	Grupo 12	0.21 %
Grupo 3	4.30 %	Grupo 8	1.16 %	Grupo 13	3.50 %
Grupo 4	0.60 %	Grupo 9	4.78 %	Grupo 14	32.86 %
Grupo 5	4.54 %	Grupo 10	1.97 %	Grupo 15	1.53 %

Tabla 5.129.: Predicción del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	11.04 %	Grupo 6	1.51 %	Grupo 11	3.27 %
Grupo 2	27.38 %	Grupo 7	4.18 %	Grupo 12	4.48 %
Grupo 3	0.55 %	Grupo 8	5.43 %	Grupo 13	29.90 %
Grupo 4	0.15 %	Grupo 9	3.52 %	Grupo 14	2.88 %
Grupo 5	1.50 %	Grupo 10	3.44 %	Grupo 15	0.81 %

Tabla 5.130.: Predicción del porcentaje de datos asociados a cada grupo, Domingo

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	4.74 %	Grupo 6	3.54 %	Grupo 11	8.58 %
Grupo 2	27.55 %	Grupo 7	34.06 %	Grupo 12	0.80 %
Grupo 3	2.11 %	Grupo 8	0.36 %	Grupo 13	4.43 %
Grupo 4	3.43 %	Grupo 9	2.13 %	Grupo 14	1.55 %
Grupo 5	3.97 %	Grupo 10	0.05 %	Grupo 15	2.72 %

Tabla 5.131.: Validación del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	6.90 %	Grupo 6	5.60 %	Grupo 11	2.08 %
Grupo 2	2.79 %	Grupo 7	27.48 %	Grupo 12	0.24 %
Grupo 3	4.18 %	Grupo 8	1.24 %	Grupo 13	3.19 %
Grupo 4	0.63 %	Grupo 9	4.79 %	Grupo 14	33.23 %
Grupo 5	4.36 %	Grupo 10	1.93 %	Grupo 15	1.37 %

Tabla 5.132.: Validación del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	11.03 %	Grupo 6	1.57 %	Grupo 11	3.21 %
Grupo 2	28.08 %	Grupo 7	4.12 %	Grupo 12	4.52 %
Grupo 3	0.50 %	Grupo 8	5.33 %	Grupo 13	29.42 %
Grupo 4	0.14 %	Grupo 9	3.63 %	Grupo 14	2.91 %
Grupo 5	1.61 %	Grupo 10	3.13 %	Grupo 15	0.80 %

Tabla 5.133.: Validación del porcentaje de datos asociados a cada grupo, Domingo

Consumo industrial

A pesar de que el número de clústers empleado en cada tipo de día de la semana es diferente, se ha decidido juntar el estudio de todo el consumo industrial de un día de invierno en la provincia de Málaga, en un sólo análisis debido a las similitudes presentes entre los grupos encontrados. Con el objetivo de concretar el número de clústers a emplear en cada tipo de día de la semana, se ha recurrido al cálculo de la disminución porcentual de la suma de varianzas en función del número de grupos utilizados, ya que las funciones de coste calculadas, Figura 5.96, carecen de “codo” discernible a primera vista. Tras completar el proceso, se ha decidido emplear diez clúster en el caso de los días laborales y doce en los sábados y domingos.

Una vez elegido el número de clústers a emplear, se ha implementado el algoritmo K-Means con la finalidad de encontrar los grupos que definen el consumo industrial de un día de verano en la provincia de Málaga. Los grupos, junto con sus desviaciones típicas, se encuentran en la Figura 5.97.

A primera vista, destaca un conjunto de grupos que presentan los valores de consumo más elevados, así como un patrón errático, que no se asemeja a la forma esperada de un consumo del tipo industrial, como la encontrada en la sección 5.3.2. Es el caso de los grupos 2 y 6 de los días laborales, del Grupo 12 de los sábados y del Grupo 6 de los domingos. Dichos patrones presentan grandes fluctuaciones a lo largo del día y alcanzan valores de consumo muy elevados, llegando a tener picos de más de 20000 kWh. Sin embargo, los patrones erráticos del Grupo 2 de los días laborales y de los fines de semana, solamente presentan una clara depresión en las horas de sueño, además, el Grupo 6 de los días laborales presenta una forma más similar a la esperada para un consumo industrial. La tendencia errática y los valores de consumo tan elevados no representan la

mayor parte del consumo industrial de un día de verano en la provincia de Málaga, es más, representan porcentajes mínimos de dicho consumo, Tabla 5.134, Tabla 5.135 y Tabla 5.136.

Siguiendo con esta línea, se encuentra el Grupo 8 de los sábados. Este patrón también presentan una tendencia errática, pero con niveles de consumo menores, en torno a los 8000 kWh. No muestran en ninguna franja horaria una depresión clara que indique el cese de la actividad industrial. Estos grupos tampoco representan un gran porcentaje del consumo objeto de estudio de esta sección, Tabla 5.135.

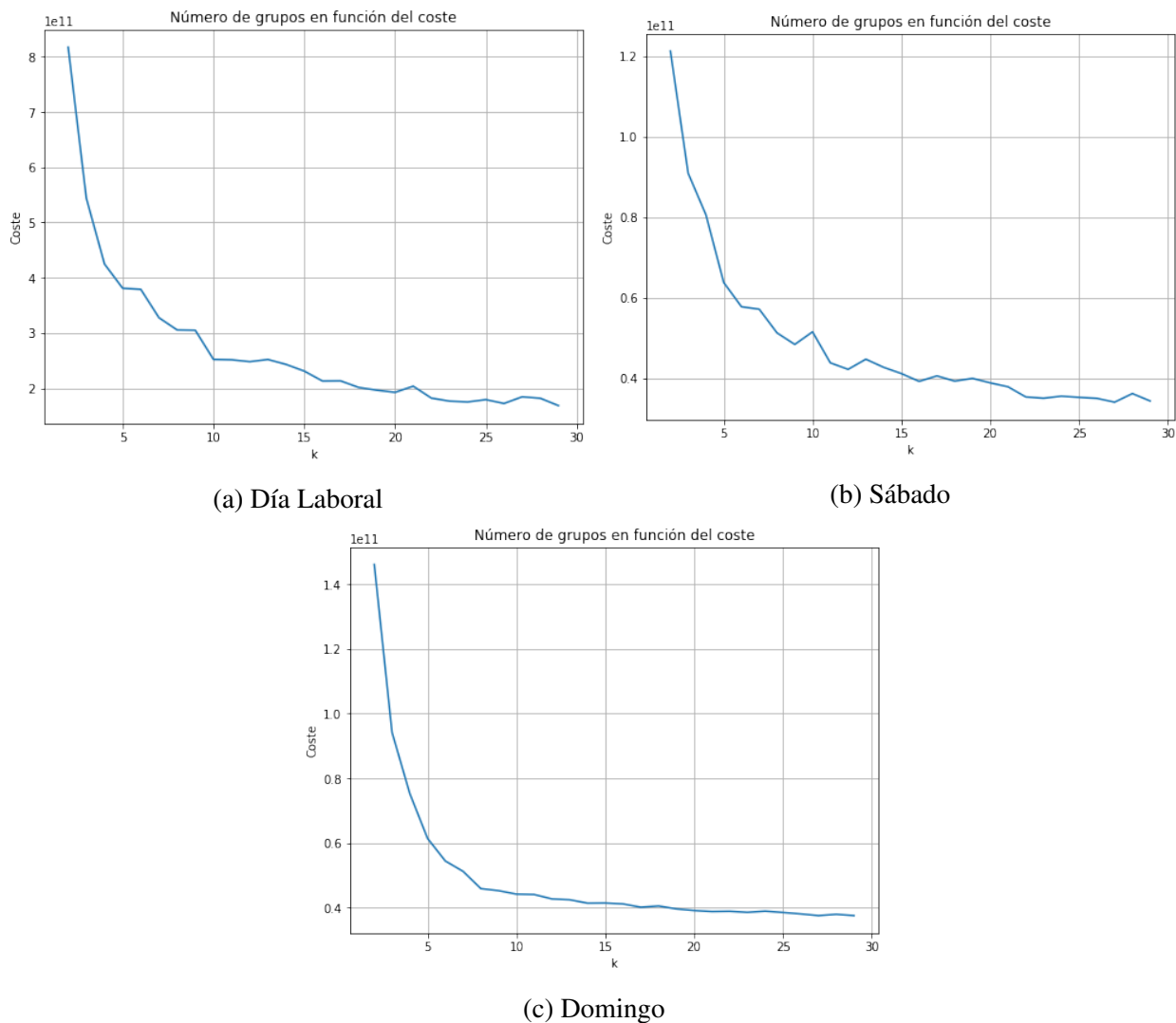
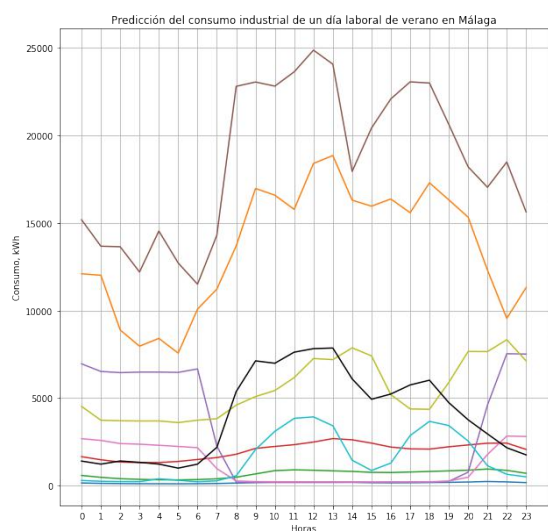
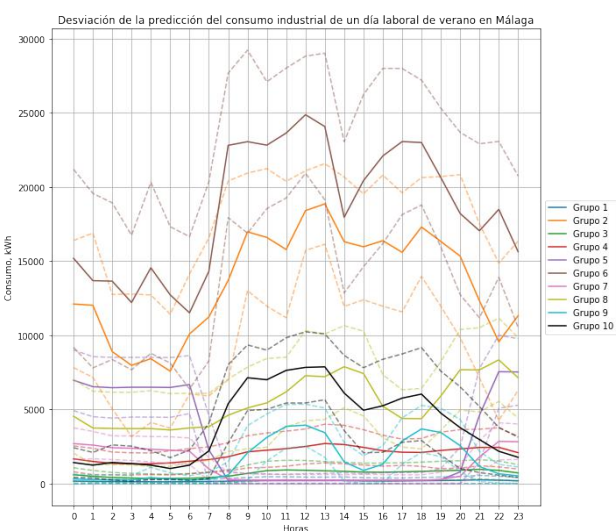


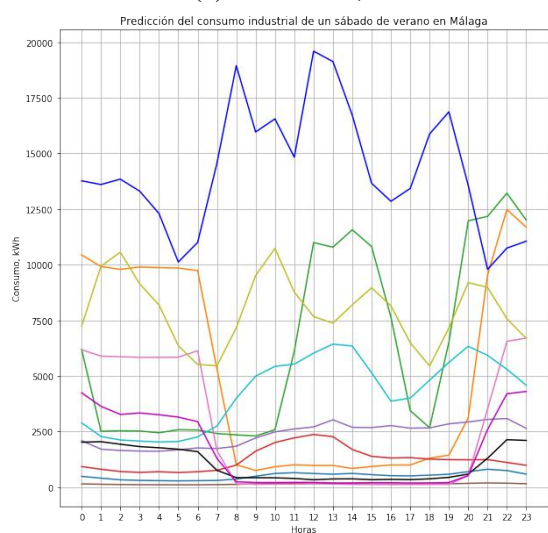
Figura 5.96.: Número de grupos en función de la suma de las distancias a los centroides



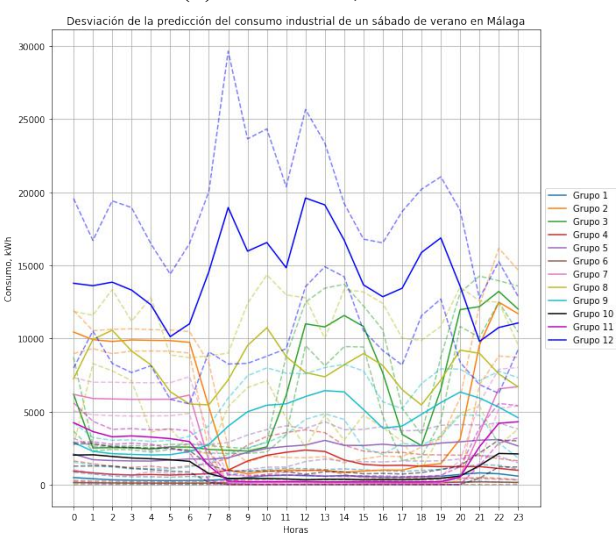
(a) Predicción, Diario



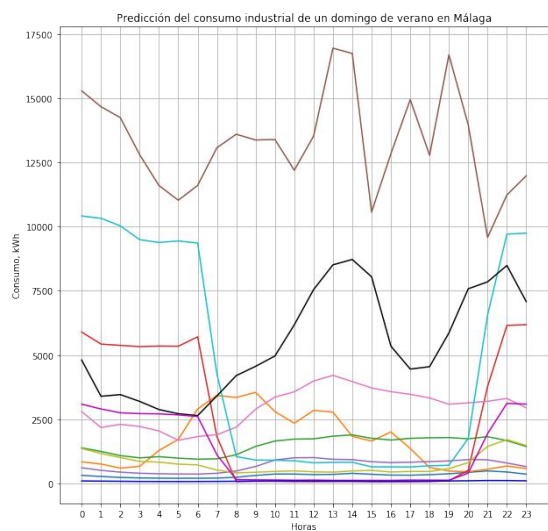
(b) Desviación, Diario



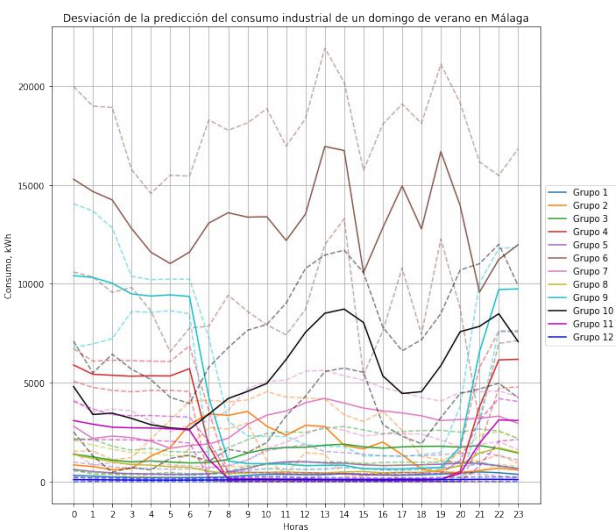
(c) Predicción, Sábado



(d) Desviación, Sábado



(e) Predicción, Domingo



(f) Desviación, Domingo

Figura 5.97.: Consumo industrial de un día de verano en Málaga

Un conjunto de grupos que llama la atención es el formado por aquellos que presentan, aproximadamente dos niveles de consumo. Los máximos de dicho consumo se dan desde aproximadamente las 20:00 hasta las 8:00. Este conjunto de grupos está formado por los grupos 5 y 7 en el gráfico de días laborales, por los grupos 2, 7, 10 y 11 en los sábados y por los grupos 4, 9 y 11 en los domingos. Este tipo de comportamiento se puede asociar al seguido por locales que comienzan su actividad después de comer y continúan su actividad durante toda la noche. Sin embargo, este patrón es seguido por un porcentaje mínimo de clientes, Tabla 5.134, Tabla 5.135, Tabla 5.136. Patrones similares ya se había detectado en el caso del consumo industrial de la provincia de Barcelona, así como en los casos de consumo industrial de la provincia de Málaga estudiados hasta ahora, lo que indica cierta relación con la zona geográfica de las provincias y con su clima. La principal diferencia entre los distintos casos se encuentra en la duración exacta de la depresión del consumo, dependiendo de la estación.

La tendencia de consumo industrial encontrada en la sección 5.3.2, se caracteriza por la presencia de dos picos claros de consumo, uno por la mañana y otro por la tarde. Variantes de esta tendencia son seguidas por los siguientes grupos. Por un lado, cabe mencionar que ciertos grupos siguen un patrón que no presentan el pico de consumo que se espera por la tarde. Alcanzan un pico por la mañana y el resto de la jornada laboral presentan una tendencia descendente. Es el caso del Grupo 10 en los días laborales, del Grupo 4 en el caso de los sábados y del los grupos 2 y 7 de los domingos. Este conjunto de grupos, presentan un porcentaje no muy relevante del consumo industrial de un día de verano en la provincia de Málaga, Tabla 5.134, Tabla 5.135, Tabla 5.136.

Los grupos 8 y 9 de los días laborales, los grupo 3 y 9 de los sábados y el Grupo 10 de los domingos, Figura 5.97, son los que presentan una tendencia más similar a la encontrada en la sección 5.3.2 para el consumo industrial. Además, cabe mencionar que el resto de grupos presentan una tendencia más constante a lo largo del día. En estos grupos se puede inferir una pequeña subida del consumo a las horas donde deberían encontrar los picos principales, es muy probable que no se puedan distinguir por la escala de la imagen, pero que si que sigan dicha tendencia. Una vez más, los grupos con los niveles de consumo más bajos, son los que representan un porcentaje mayor del consumo industrial de un día de invierno en la provincia de Málaga; es el caso del Grupo 1 los días laborales, el Grupo 6 los sábados y el Grupo 12 los domingos, Tabla 5.134, Tabla 5.135, Tabla 5.136.

Uno de los objetivos principales de este proyecto es proporcionar información útil para las distribuidoras de energía eléctrica para que puedan lanzar campañas de publicidad orientadas a determinados grupos de clientes. A las distribuidoras energéticas les conviene aplanar los niveles de consumo a lo largo del día para poder optimizar la producción de las centrales. En este caso, principalmente, se pueden lanzar dos campañas publicitaria. Una campaña tendría como objetivo conseguir la redistribución del consumo de aquellos grupos que presentan una gran depresión entre las 8:00 y las 20:00. Para conseguir dicha redistribución, se les podrían lanzar ofertas en las que se indicasen determinadas franjas horarias en las que el precio fuese más reducido, con la esperanza de que desplazasen su consumo a dichas franjas. Otra tendría como objetivo reducir los niveles de consumo de aquellos grupos que sobresalen por encima del resto y presenta una tendencia errática, lo cual se puede conseguir mediante el establecimiento de un límite de consumo a partir del cual el precio de la electricidad aumenta para fomentar la reducción del consumo.

Con el modelo predictivo establecido, es necesario evaluar la veracidad del mismo. Motivo por el cual se comparan los porcentajes de representación de cada grupo obtenidos en la fase de predicción, Tabla 5.134, Tabla 5.135, Tabla 5.136, con los obtenidos en la fase de validación, Tabla 5.137, Tabla 5.138, Tabla 5.139. En este caso, el resultado de la comparación es una variación media por cada grupo del 0.29 % en los días laborales, del 0.30 % en los sábados y del 0.37 % en los domingos. Estos valores son perfectamente asumibles, por lo que el modelo calculado es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	71.30 %	Grupo 6	0.07 %
Grupo 2	0.07 %	Grupo 7	2.30 %
Grupo 3	18.55 %	Grupo 8	0.34 %
Grupo 4	3.74 %	Grupo 9	1.95 %
Grupo 5	0.97 %	Grupo 10	0.71 %

Tabla 5.134.: Predicción del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	24.74 %	Grupo 5	1.83 %	Grupo 9	0.55 %
Grupo 2	0.19 %	Grupo 6	63.38 %	Grupo 10	1.88 %
Grupo 3	0.10 %	Grupo 7	0.76 %	Grupo 11	0.86 %
Grupo 4	5.55 %	Grupo 8	0.06 %	Grupo 12	0.08 %

Tabla 5.135.: Predicción del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	30.48 %	Grupo 5	9.08 %	Grupo 9	0.30 %
Grupo 2	0.35 %	Grupo 6	0.12 %	Grupo 10	0.50 %
Grupo 3	4.23 %	Grupo 7	1.05 %	Grupo 11	1.57 %
Grupo 4	0.80 %	Grupo 8	2.50 %	Grupo 12	49.02 %

Tabla 5.136.: Predicción del porcentaje de datos asociados a cada grupo, Domingo

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	70.15 %	Grupo 6	0.05 %
Grupo 2	0.06 %	Grupo 7	2.58 %
Grupo 3	19.38 %	Grupo 8	0.35 %
Grupo 4	3.97 %	Grupo 9	1.70 %
Grupo 5	1.04 %	Grupo 10	0.72 %

Tabla 5.137.: Validación del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	24.23 %	Grupo 5	1.97 %	Grupo 9	1.00 %
Grupo 2	0.23 %	Grupo 6	62.17 %	Grupo 10	2.13 %
Grupo 3	0.08 %	Grupo 7	0.89 %	Grupo 11	0.93 %
Grupo 4	6.14 %	Grupo 8	0.04 %	Grupo 12	0.19 %

Tabla 5.138.: Validación del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	32.12 %	Grupo 5	8.09 %	Grupo 9	0.23 %
Grupo 2	0.20 %	Grupo 6	0.08 %	Grupo 10	0.39 %
Grupo 3	3.60 %	Grupo 7	1.17 %	Grupo 11	1.52 %
Grupo 4	0.78 %	Grupo 8	2.93 %	Grupo 12	48.89 %

Tabla 5.139.: Validación del porcentaje de datos asociados a cada grupo, Domingo

5.4.3.4. Perfiles del consumo de un día de otoño en la provincia de Málaga

El objetivo de esta sección es encontrar los distintos patrones que representan el consumo residencial e industrial de un día de otoño en la provincia de Málaga. Para alcanzar dicha finalidad, se va a seguir la dinámica empleada en las secciones anteriores, en primer lugar, se obtiene el número de clústers a emplear en cada caso, para después calcular los distintos patrones de consumo mediante la implementación del algoritmo K-Means.

Consumo residencial

En este caso, el número de clústers elegido para todos los días de la semana es el mismo. Con el objetivo de encontrar el número de clústers a emplear en cada tipo de día de la semana, ya que el “codo” de las funciones de coste, Figura 5.98 es difícil de discernir a simple vista, se ha recurrido al cálculo del porcentaje de la disminución de la suma de varianzas con cada aumento del número de clústers. Tras realizar esa operación para cada uno de los tres casos, se ha llegado a la conclusión de que el número de clústers a emplear es quince.

Una vez elegido el número de clústers a emplear, se ha implementado el algoritmo K-Means con la finalidad de encontrar quince grupos que definen el consumo residencial de un día de invierno en la provincia de Málaga. Los grupos, junto con sus desviaciones típicas, se encuentran en la Figura 5.99.

A primera vista, llama la atención un conjunto de grupos, que solamente presentan un pico de consumo cuya posición a lo largo del tiempo varía entre las 3:00 y las 4:00, aproximadamente. Dicho pico se eleva por encima de los 6000 kWh. Es el caso del Grupo 5 en los días laborales,

del Grupo 9 los sábados y del Grupo 7 los domingos. Otros grupos también presentan un único pico de consumo cuya posición se encuentra entre las 9:00 y las 18:00; aunque estos picos presentan valores de consumo mucho menores, Figura 5.99. Es el caso de los grupos 8, 9, 13, 14 y 15 en el gráfico de los días laborales, los grupos 1, 2, 10, 13, 14 y 15 en el de los sábados y los grupos 5, 8 y 13 en el de los domingos. Este tipo de comportamiento se asocia con el de aquellas personas que tan solo emplean recursos eléctricos durante la hora de la comida, lo que implica que o bien reducen drásticamente el uso de aparatos eléctricos o emplean otras fuentes de energía en ese espacio de tiempo. Además, los picos más elevados de consumo se asocian a casas más grandes, donde el consumo energético es mayor.

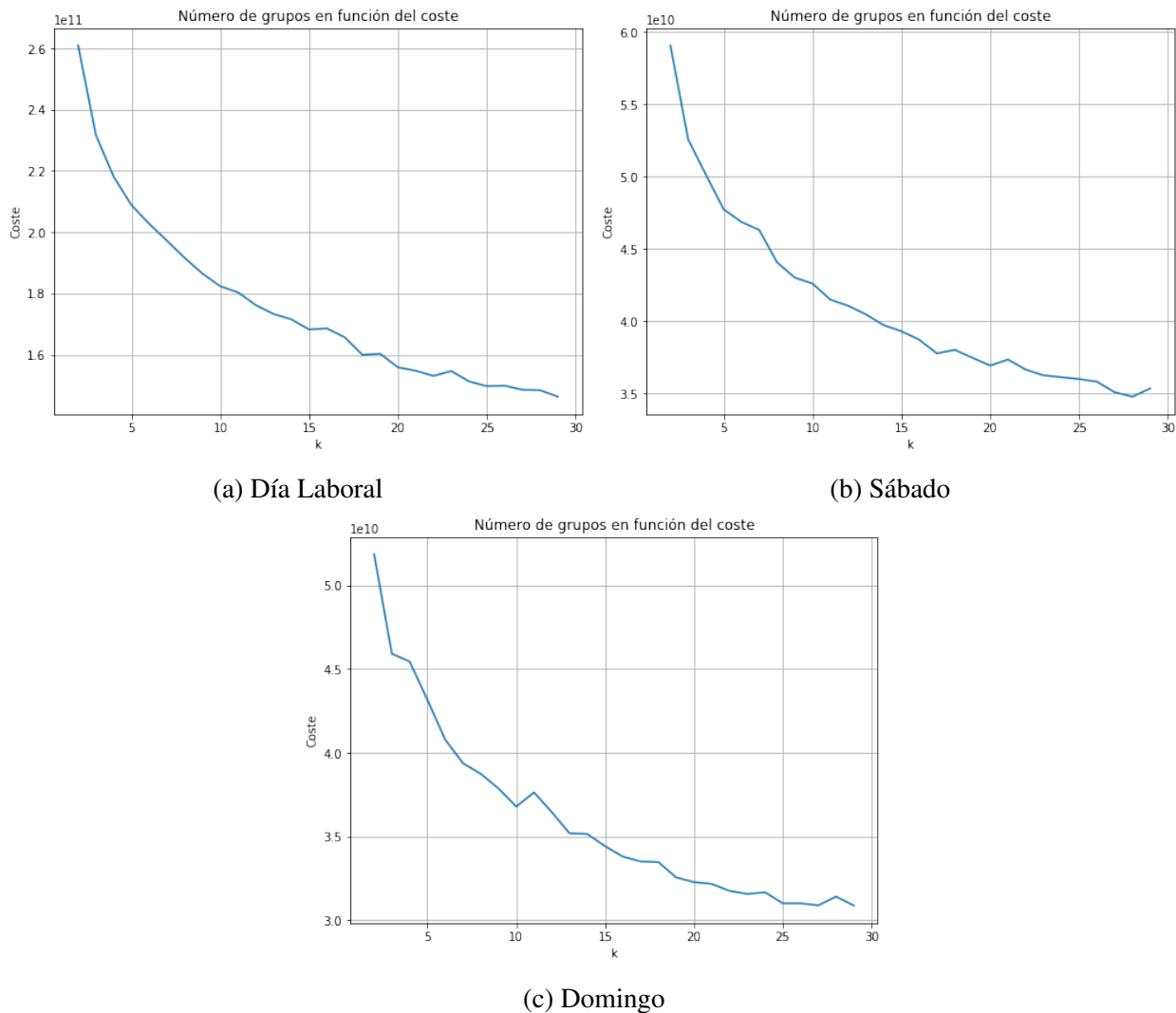


Figura 5.98.: Número de grupos en función de la suma de las distancias a los centroides

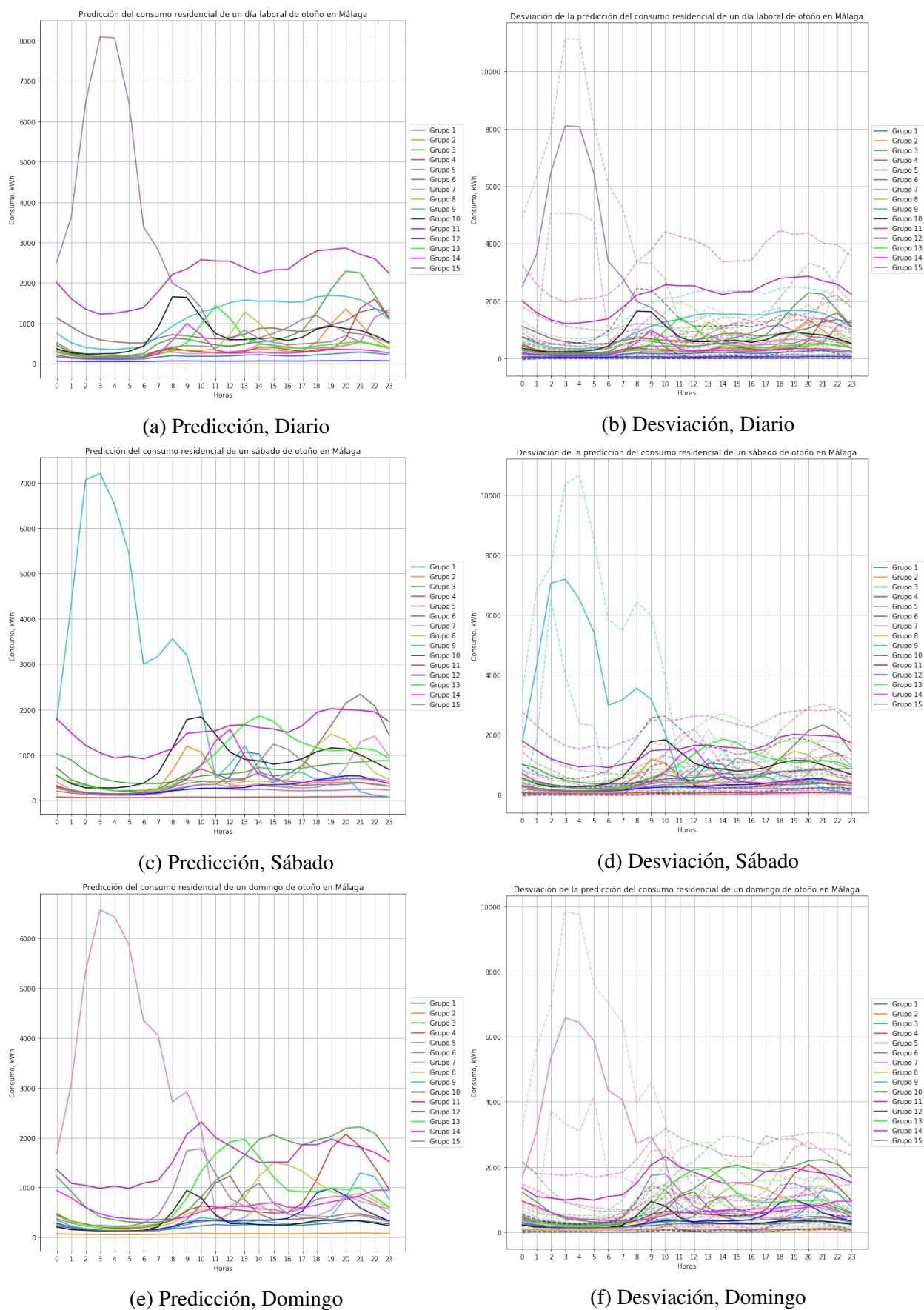


Figura 5.99.: Consumo residencial de un día de otoño en Málaga

También llama la atención la presencia de un grupo cuyo consumo decae durante las horas de luz, es el caso del Grupo 15 de los domingos. Este patrón es totalmente opuesto al esperado cuando se trata de consumo residencial. Es muy probable que los clientes que pertenecen a estos grupos trabajen en el turno de noche, motivo por el cual la depresión de su consumo se da durante el día. Además, representan un pequeño porcentaje del consumo residencial de un día de invierno en la provincia de Málaga, Tabla 5.142.

El patrón detectado por primera vez en el consumo residencial de un día de otoño en la provincia de Málaga, se vuelve a manifestar en este caso. Este patrón se caracteriza por seguir la tendencia esperada de un consumo residencial, hasta el mediodía. A partir de dicho momento, el consumo se vuelve prácticamente constante, es decir, pierde la tendencia de subida propia de un consumo residencial. Es el caso del Grupo 9 de los días laborales, del Grupo 11 de los sábados y del Grupo 3 de los domingos. Este tipo de comportamiento puede ser el seguido por aquellas personas que pasan toda la tarde en sus viviendas y no abusan en un periodo concreto de tiempo del uso de electrodomésticos u otros aparatos que generen un gran consumo eléctrico.

Se puede observar un grupo que no siguen la tendencia ascendente a lo largo del día que se espera del consumo residencial, como se estableció en la sección 5.3.2 Es el caso del Grupo 11 de los días laborales y de los domingos. Sorprendentemente, encajaría mejor con el gráfico esperado del consumo industrial, razón por la que tal vez representan un porcentaje muy pequeño del consumo residencial de un día de otoño en la provincia de Málaga, Tabla 5.140, Tabla 5.141 y Tabla 5.142. Este tipo de comportamiento se puede dar en viviendas en las que durante las mañanas la limpieza sea realizada por una persona externa, que termina su jornada antes de la hora de comer, y a la hora de cenar el residente vuelve al hogar.

El resto de los grupos representados, reflejan el patrón de consumo residencial esperado, tomando valores mayores o menores y experimentando mayor o menor diferencia entre los dos picos principales del día. Una vez más se da el que grupo con los niveles de consumo menores a lo largo de todo el día, es el que representa un mayor porcentaje del consumo.

Con la finalidad de aplanar los niveles de consumo a lo largo del día, ya que de esta forma las distribuidoras eléctricas pueden optimizar sus centrales, se pueden lanzar, principalmente, una campañas publicitarias. La campaña se debería centrar en reducir la diferencia de ciertos picos de consumo con respecto al consumo generado el resto del día. Esto se puede conseguir mediante el lanzamiento de ofertas de energía en determinadas franjas horarias, medida que motiva la redistribución del consumo. Además, se puede establecer un límite de consumo a partir del cual el precio de la electricidad aumenta para fomentar la reducción del consumo.

Con el modelo establecido, es necesario evaluar la fiabilidad del mismo. Para ello, se comparan los porcentajes de representación de cada grupo obtenidos en la fase de predicción, Tabla 5.140, Tabla 5.141 y Tabla 5.142, con los obtenidos en la fase de validación, Tabla 5.143, Tabla 5.144 y Tabla 5.145. En este caso, la variación es del 0.10 % en los días laborales, del 0.23 % en los sábados y de 0.20 % los domingos, valores que son asumibles, por lo que el modelo es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	24.92 %	Grupo 6	1.72 %	Grupo 11	0.17 %
Grupo 2	4.37 %	Grupo 7	13.34 %	Grupo 12	33.71 %
Grupo 3	1.26 %	Grupo 8	4.03 %	Grupo 13	2.75 %
Grupo 4	2.91 %	Grupo 9	0.98 %	Grupo 14	4.78 %
Grupo 5	0.03 %	Grupo 10	1.35 %	Grupo 15	3.78 %

Tabla 5.140.: Predicción del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	5.42 %	Grupo 6	1.03 %	Grupo 11	0.61 %
Grupo 2	4.28 %	Grupo 7	2.96 %	Grupo 12	12.55 %
Grupo 3	3.09 %	Grupo 8	3.31 %	Grupo 13	1.39 %
Grupo 4	35.80 %	Grupo 9	0.02 %	Grupo 14	3.02 %
Grupo 5	21.79 %	Grupo 10	1.29 %	Grupo 15	3.46 %

Tabla 5.141.: Predicción del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	24.81 %	Grupo 6	4.33 %	Grupo 11	0.38 %
Grupo 2	38.77 %	Grupo 7	0.02 %	Grupo 12	5.95 %
Grupo 3	0.70 %	Grupo 8	2.17 %	Grupo 13	1.32 %
Grupo 4	1.87 %	Grupo 9	3.59 %	Grupo 14	3.07 %
Grupo 5	1.39 %	Grupo 10	5.37 %	Grupo 15	6.24 %

Tabla 5.142.: Predicción del porcentaje de datos asociados a cada grupo, Domingo

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	24.98 %	Grupo 6	1.57 %	Grupo 11	0.13 %
Grupo 2	4.19 %	Grupo 7	13.54 %	Grupo 12	33.36 %
Grupo 3	1.28 %	Grupo 8	4.13 %	Grupo 13	2.90 %
Grupo 4	3.10 %	Grupo 9	1.00 %	Grupo 14	4.72 %
Grupo 5	0.02 %	Grupo 10	1.33 %	Grupo 15	3.74 %

Tabla 5.143.: Validación del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	5.74 %	Grupo 6	0.99 %	Grupo 11	0.49 %
Grupo 2	4.19 %	Grupo 7	3.02 %	Grupo 12	11.72 %
Grupo 3	3.14 %	Grupo 8	2.97 %	Grupo 13	1.52 %
Grupo 4	36.35 %	Grupo 9	0.01 %	Grupo 14	2.94 %
Grupo 5	22.41 %	Grupo 10	1.31 %	Grupo 15	3.19 %

Tabla 5.144.: Predicción del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	25.13 %	Grupo 6	4.16 %	Grupo 11	0.44 %
Grupo 2	38.85 %	Grupo 7	0.00 %	Grupo 12	6.56 %
Grupo 3	0.53 %	Grupo 8	2.10 %	Grupo 13	1.25 %
Grupo 4	1.77 %	Grupo 9	3.62 %	Grupo 14	3.21 %
Grupo 5	1.66 %	Grupo 10	4.80 %	Grupo 15	5.93 %

Tabla 5.145.: Validación del porcentaje de datos asociados a cada grupo, Domingo

Consumo industrial

En este caso, el número de clústers elegido para todos los días de la semana es el mismo. Con el objetivo de encontrar el número de clústers a emplear en cada tipo de día de la semana, ya que el “codo” de las funciones de coste, Figura 5.100 es difícil de discernir a simple vista, se ha recurrido al cálculo del porcentaje de la disminución de la suma de varianzas con cada aumento del número de clústers. Tras realizar esa operación para cada uno de los tres casos, se ha llegado a la conclusión de que el número de clústers a emplear es diez.

Una vez elegido el número de clústers a emplear, se ha implementado el algoritmo K-Means con la finalidad de encontrar diez grupos que definen el consumo residencial de un día de otoño en la provincia de Málaga. Los grupos, junto con sus desviaciones típicas, se encuentran en la Figura 5.101.

A primera vista, destaca un conjunto de grupos que presentan los valores de consumo más elevados, así como un patrón errático, que no se asemeja a la forma esperada de un consumo del tipo industrial, como la encontrada en la sección 5.3.2. Es el caso de los grupos 2, 3 y 6 de los días laborales, de los grupos 2, 5 y 7 de los sábados y los grupos 2 y 8 de los domingos. Dichos patrones presentan grandes fluctuaciones a lo largo del día y alcanzan valores de consumo muy elevados, llegando a tener picos de más de 30000 kWh, en el caso de los sábados. Cabe destacar que de entre estos patrones, el Grupo 7 de los sábados presenta valores de consumo más elevados por la mañana, lo que se puede asociar a empresas que solamente trabajen media jornada los sábados. Sin embargo, el resto de los patrones erráticos del sábados y los de los días laborales, solamente presentan una clara depresión en las horas de sueño, además, el Grupo 6 de los días laborales y los grupos 2 y 5 de los sábados, presentan una forma más similar a la esperada para un consumo

industrial. Los patrones erráticos de los domingos, presentan niveles de consumo algo menores. No muestran en ninguna franja horaria una depresión clara que indique el cese de la actividad industrial. La tendencia errática y los valores de consumo tan elevados no representan la mayor parte del consumo industrial de un día de invierno en la provincia de Málaga, es más, representan porcentajes mínimos de dicho consumo, Tabla 5.146, Tabla 5.147 y Tabla 5.148.

Un conjunto de grupos que llama la atención es el formado por aquellos que presentan, aproximadamente dos niveles de consumo. Los máximos de dicho consumo se dan desde aproximadamente las 17:00 hasta las 9:00. Este conjunto de grupos está formado por los grupos 4 y 10 en el gráfico de días laborales, por los grupos 4, 6 y 9 en los sábados y por los grupos 3, 5, 7 y 10 en los domingos. Este tipo de comportamiento se puede asociar al seguido por locales que comienzan su actividad a media tarde y continúan su actividad durante toda la noche. Sin embargo, este patrón es seguido por un porcentaje mínimo de clientes, Tabla 5.146, Tabla 5.147, Tabla 5.148.

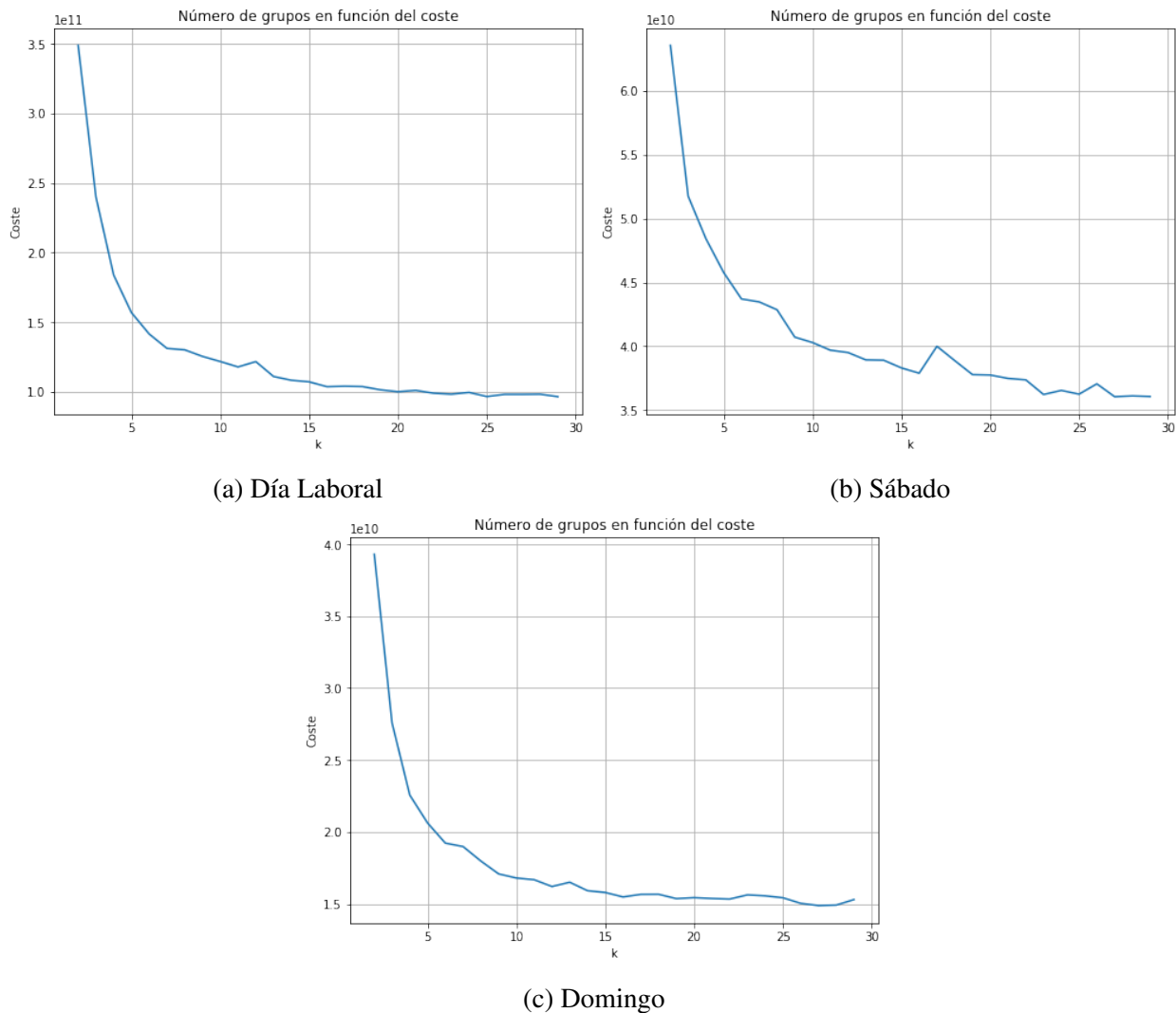


Figura 5.100.: Número de grupos en función de la suma de las distancias a los centroides

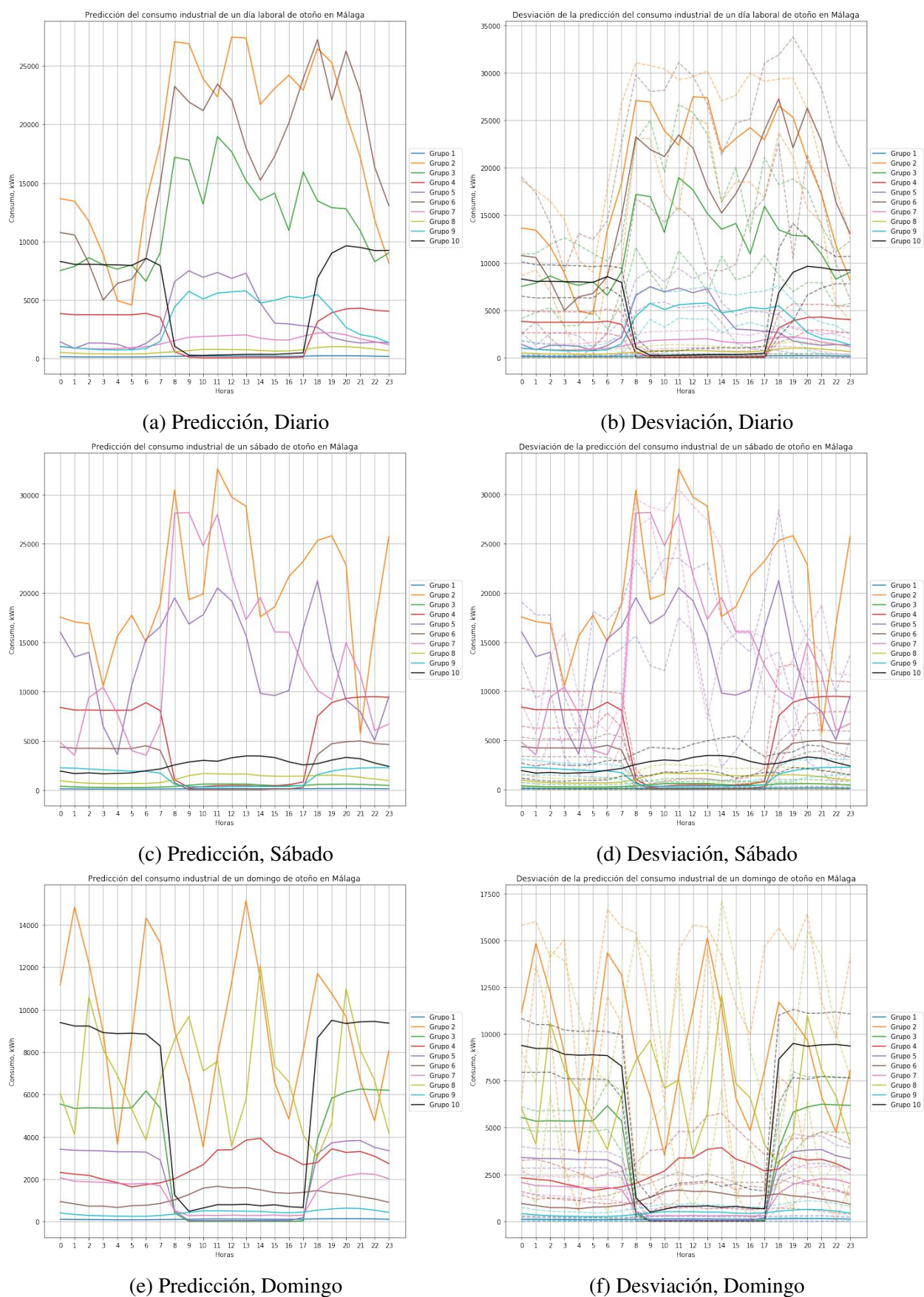


Figura 5.101.: Consumo industrial de un día de otoño en Málaga

La tendencia de consumo industrial encontrada en la sección 5.3.2, se caracteriza por la presencia de dos picos claros de consumo, uno por la mañana y otro por la tarde. Sin embargo esta tendencia no es la seguida por los grupos 5 y 9 de los días laborales. Dichos grupos sufren una subida del consumo en torno a las 8:00 y después mantienen el mismo nivel de consumo a lo largo de toda la jornada laboral. Cabe mencionar, que el caso del Grupo 5 esta tan solo se extiende hasta las 14:00, aproximadamente, por lo que dicho grupo representa a las empresas que tan solo trabajan en horario de mañana.

Un conjunto de grupos sí que presenta una tendencia más similar a la encontrada en la sección 5.3.2 para el consumo industrial. Es el caso de los grupos 7 y 8 de los días laborales y de los grupos 4, 6 y 9 de los domingos. Además, cabe mencionar que el resto de grupos presentan una tendencia más constante a lo largo del día. En estos grupos se puede inferir una pequeña subida del consumo a las horas donde deberían encontrar los picos principales, es muy probable que no se puedan distinguir por la escala de la imagen, pero que si que sigan dicha tendencia. Una vez más, los grupos con los niveles de consumo más bajos, son los que representan un porcentaje mayor del consumo industrial de un día de otoño en la provincia de Málaga; es el caso del Grupo 1 para todos los días de la semana, Tabla 5.146, Tabla 5.147, Tabla 5.148.

Uno de los objetivos principales de este proyecto es proporcionar información útil para las distribuidoras de energía eléctrica para que puedan lanzar campañas de publicidad orientadas a determinados grupos de clientes. A las distribuidoras energéticas les conviene aplanar los niveles de consumo a lo largo del día para poder optimizar la producción de las centrales. En este caso, principalmente, se pueden lanzar dos campañas publicitarias. Una campaña tendría como objetivo conseguir la redistribución del consumo de aquellos grupos que presentan una gran depresión entre las 9:00 y las 17:00. Para conseguir dicha redistribución, se les podrían lanzar ofertas en las que se indicasen determinadas franjas horarias en las que el precio fuese más reducido, con la esperanza de que desplazasen su consumo a dichas franjas. Otra tendría como objetivo reducir los niveles de consumo de aquellos grupos que sobresalen por encima del resto y presenta una tendencia errática, lo cual se puede conseguir mediante el establecimiento de un límite de consumo a partir del cual el precio de la electricidad aumenta para fomentar la reducción del consumo.

Con el modelo predictivo establecido, es necesario evaluar la veracidad del mismo. Motivo por el cual se comparan los porcentajes de representación de cada grupo obtenidos en la fase de predicción, Tabla 5.146, Tabla 5.147, Tabla 5.148, con los obtenidos en la fase de validación, Tabla 5.149, Tabla 5.150, Tabla 5.151. En este caso, el resultado de la comparación es una variación media por cada grupo del 0.24% en los días laborales, del 0.36% en los sábados y del 0.57% en los domingos. Estos valores son perfectamente asumibles, por lo que el modelo calculado es válido.

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	73.36 %	Grupo 6	0.05 %
Grupo 2	0.05 %	Grupo 7	5.93 %
Grupo 3	0.04 %	Grupo 8	17.51 %
Grupo 4	2.00 %	Grupo 9	0.50 %
Grupo 5	0.15 %	Grupo 10	0.40 %

Tabla 5.146.: Predicción del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	63.35 %	Grupo 6	1.11 %
Grupo 2	0.02 %	Grupo 7	0.05 %
Grupo 3	25.73 %	Grupo 8	6.03 %
Grupo 4	0.46 %	Grupo 9	2.08 %
Grupo 5	0.07 %	Grupo 10	1.09 %

Tabla 5.147.: Predicción del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	64.27 %	Grupo 6	5.49 %
Grupo 2	0.08 %	Grupo 7	1.39 %
Grupo 3	0.79 %	Grupo 8	0.08 %
Grupo 4	0.60 %	Grupo 9	25.95 %
Grupo 5	1.09 %	Grupo 10	0.25 %

Tabla 5.148.: Predicción del porcentaje de datos asociados a cada grupo, Domingo

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	74.43 %	Grupo 6	0.11 %
Grupo 2	0.03 %	Grupo 7	5.32 %
Grupo 3	0.06 %	Grupo 8	17.11 %
Grupo 4	1.86 %	Grupo 9	0.45 %
Grupo 5	0.14 %	Grupo 10	0.45 %

Tabla 5.149.: Validación del porcentaje de datos asociados a cada grupo, Diario

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	62.28 %	Grupo 6	1.89s %
Grupo 2	0.07 %	Grupo 7	0.07 %
Grupo 3	26.53 %	Grupo 8	5.52 %
Grupo 4	0.29 %	Grupo 9	2.03 %
Grupo 5	0.07 %	Grupo 10	1.24 %

Tabla 5.150.: Validacion del porcentaje de datos asociados a cada grupo, Sábado

GRUPO	PORCENTAJE (%)	GRUPO	PORCENTAJE (%)
Grupo 1	65.91 %	Grupo 6	5.70 %
Grupo 2	0.00 %	Grupo 7	1.59 %
Grupo 3	0.84 %	Grupo 8	0.25 %
Grupo 4	1.17 %	Grupo 9	23.53 %
Grupo 5	0.92 %	Grupo 10	0.08 %

Tabla 5.151.: Validación del porcentaje de datos asociados a cada grupo, Domingo

6. Conclusiones Y Trabajos Futuros

Cabe recordar que el objetivo principal del presente proyecto es la obtención de patrones de comportamiento del consumo eléctrico en España. Una vez obtenidos dichos patrones, se extrae información que permite elaborar herramientas empresariales que serán entregada a las distribuidoras y generadoras eléctricas para que puedan mejorar su rentabilidad.

Para ello, en primer lugar, se lleva a cabo un análisis descriptivo en el que se explora como distintos factores afecta al consumo eléctrico a nivel general. Dicho análisis ha asentando las bases sobre las que se sustenta este proyecto, ya que se han perfilado los patrones que, a grandes rasgos, son seguidos por los grupos de comportamiento obtenidos posteriormente.

Tras la creación de una imagen inicial acerca de como diversos parámetros afectan al consumo eléctrico en España, se ha proseguido a la obtención de los patrones de comportamiento que rigen el consumo eléctrico, mediante la implementación del algoritmo K-Means. Antes de llevar a cabo la implementación del algoritmo se han filtrado los datos, con la finalidad de excluir aquellos que no cumplieran con los valores deseados de las variables independientes para cada caso de estudio. Dichas variables independientes están compuestas por: el tipo de actividad económica (residencial o industrial), la estación del año, el tipo del día (día laboral, sábado o domingo) y la provincia (Madrid, Barcelona o Málaga). De esta forma se han encontrado grupos de comportamiento que permiten dar respuesta a preguntas del estilo ¿Que tendencias se encuentran en el consumo residencial de un domingo de otoño en la provincia de Málaga?

Durante este proyecto, se han conseguido crear modelos exitosos que reflejan los patrones de comportamiento presenten para cada uno de los setenta y dos casos estudiados. Para cada uno de los cuales, se ha calculado el número de grupos a emplear, se ha implementado el algoritmo de agrupamiento K-Means y se ha verificado su fiabilidad. Además, se han propuesto técnicas que podrán ser empleadas por las distribuidoras y generadoras eléctricas, para conseguir uno de sus objetivos principales, un gráfico de consumo más plano a lo largo del tiempo, ya que de esta forma se mejora la rentabilidad de las centrales eléctricas.

Este proyecto ha explorado todos los casos que se planteaban, pero no abarca todas las posibilidades que ofrece el conjunto de datos sobre el que se ha realizado. En primer lugar, tan solo se han estudiado las provincias de Barcelona, Madrid y Málaga cuando se dispone de información acerca de toda la geografía española. Además, se pueden incluir nuevos factores a los ya empleados, como por ejemplo, el tipo de mercado, ya sea libre o regulado. Por estas dos vías se le puede dar continuidad al proyecto a la vez que se amplía el foco del mismo.

Hoy en día, las técnicas de procesamiento Big Data se están adentrando, uno a uno, en todos los mercados. Este proyecto tenía como fuente de datos el consumo del sector eléctrico, pero las técnicas y la filosofía empleadas, se pueden extrapolar a otros sectores eléctricos, así como a otros mercados completamente distintos. La idea fundamental es común, fijar un objetivo y emplear técnicas de procesamiento de datos que permitan alcanzar dicho objetivo.

Bibliografía

- [1] PowerData Especialistas en Gestión de Datos, “Big Data: ¿En qué consiste? Su importancia, desafíos y gobernabilidad.” <https://www.powerdata.es/big-data>, 2018.
- [2] Jupyter. <http://jupyter.org/index.html>, 2018.
- [3] MapR, “Spark 101: What is it, what it does, and why it matters.” <https://mapr.com/blog/spark-101-what-it-what-it-does-and-why-it-matters/>, 2018.
- [4] Expert System, “What is machine learning? a definition.” <http://www.expertsystem.com/machine-learning-definition/>, 2018.
- [5] Data Science, “What is semi-supervised learning?.” <https://www.datascience.com/blog/what-is-semi-supervised-learning>, 2018.
- [6] Fernando Sancho Caparrini, “Aprendizaje por refuerzo: algoritmo q learning.” <http://www.cs.us.es/~fsancho/?e=109>, 2017.
- [7] Ministerio de Energía, Turismo y Agenda Digital, “Energía Eléctrica, Estructura del sector.” <http://www.minetad.gob.es/energia/electricidad/Paginas/sectorElectrico.aspxs>.
- [8] Red Eléctrica de España, “El grupo Red Eléctrica.” <http://www.ree.es/es/conocenos/el-grupo-red-electrica>, 2018.
- [9] Red Eléctrica de España, “Estrategia.” <http://www.ree.es/es/conocenos/estrategia>, 2018.
- [10] TechTarget, “Big Data.” <https://searchdatacenter.techtarget.com/es/definicion/Big-data>, 2018.
- [11] Ve, “Qué es Big Data: fases y elementos.” <https://www.ve.com/es/blog/que-es-big-data-fases-elementos>, 2018.
- [12] Paradigma, “Ecosistema Big Data.” <https://www.paradigmadigital.com/ecosistema-big-data/>, 2018.
- [13] “A tutorial on clustering algorithms.” https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html, 2018.
- [14] “Determining the optimal number of clusters: 3 must known methods - unsupervised machine learning.” <http://www.sthda.com/english/wiki/print.php?id=239>, 2018.

- [15] Jacek Laskowski, “Mastering Apache Spark.” <https://legacy.gitbook.com/book/jaceklaskowski/mastering-apache-spark/details>, 2018.
- [16] <https://spark.apache.org/docs/2.0.2/api/java/org/apache/spark/SparkContext.html>.
- [17] Apache Spark, “Cluster Mode Overview.” <https://spark.apache.org/docs/latest/cluster-overview.html>, 2018.
- [18] Tutorials Point, “PySpark - Introduction.” https://www.tutorialspoint.com/pyspark/pyspark_introduction.html.
- [19] <https://spark.apache.org/docs/latest/sql-programming-guide.html#datasets-and-dataframes>.
- [20] Analytics Vidhya, “Complete Guide on DataFrame Operations in PySpark.” <https://www.analyticsvidhya.com/blog/2016/10/spark-dataframe-and-operations/>.
- [21] “Capítulo III Conceptos básicos de Spark.” http://reader.digitalbooks.pro/content/preview/books/41061/book/OEBPS/Text/capitulo_3.html.
- [22] “Apache Spark - Best Practices and Tuning.” <https://legacy.gitbook.com/book/umbertogriffo/apache-spark-best-practices-and-tuning/details>.
- [23] Agencia Tributaria, “Renta bruta y Renta disponible por CCAA, provincia y tamaño de población.” http://www.agenciatributaria.es/AEAT/Contenidos_Comunes/La_Agencia_Tributaria/Estadisticas/Publicaciones/sites/irpfmunicipios/2015/jrubikf48d05cb152743679af21148bb70ea9d96b4addc5.html, 2015.
- [24] Endesa, “Los 2 mercados eléctricos: el libre y el regulado.” <https://www.endesaclientes.com/mercado-libre-mercado-regulado-pvpc>, 2018.

A. Nodo master

Server RH2288 V3 - Masters			
RH2288 V3			
Base Configuration			
02311GGX	BC4M20HGSB	RH2288 V3 (12*3.5inch HDD Chassis)(Only for oversea,except Japan)H22M-03	1
02310YKM	BC1M01FGED	SM233 Onboard NIC,2x10GE Electrical Interface(X540),RJ45	1
02311AFU	BC1M06FAN	8038 Fan module	4
02131042	WEPW80015	460W GOLD AC Power Module	2
Broadwell EP CPU			
02311NEQ	BC1M56CPU	Intel Xeon E5-2640 v4(2.4GHz/10-core/25MB/90W) Processor (with heatsink)	2
DDR4 Memory			
06200213	N24DDR402	DDR4 RDIMM Memory,16GB,2400MT/s,2Rank(1G*8bit),1.2V, ECC	8
Hard Disk(include Front Panel)			
02311AYT	N2000ST7W3	HDD,2000GB,SATA 6Gb/s,7.2K rpm,64MB,3.5inch(3.5inch Drive Bay)	4
RAID Card and Other Accessories			
02311JEA	BC1M51ESMQ	SR320BC(LSI2208) SAS/SATA RAID Card,RAID0,1,5,6,10,50,60,6Gb/s,1GB Cache,used for RH2288 V3/RH2288H V3's 12HDD chassis	1
Cables and Optional Equipments			
21240434	EGUIDER01	2U Static Rail Kit	1
Soporte			
Technical Support Service			
88134UGQ	02311GGX-88134UGQ-48	RH2288 V3(12HDD EXP Chassis)H22M-03 For oversea(except Japan)-Co-Care Standard 9x5xNBD Service-48Month(s)	3
			Subtotal Serv

Figura A.1.: Detalles nodo master

B. Nodo worker

Equipamiento Huawei			UDS
Server RH2288 V3 - Worker			
RH2288 V3			
Base Configuration			
02311GGX	BC4M20HG5B	RH2288 V3 (12*3.5inch HDD Chassis)(Only for oversea,except Japan)H22M-03	1
02310YKM	BC1M01FGED	SM233 Onboard NIC,2x10GE Electrical Interface(X540),RJ45	1
02311AFU	BC1M06FAN	8038 Fan module	4
02310QWX	EN3MCACC	750W Platinum AC Power Module	2
Broadwell EP CPU			
02311NEW	BC1M62CPU	Intel Xeon E5-2680 v4(2.4GHz/14-core/35MB/120W) Processor (with heatsink)	2
DDR4 Memory			
06200213	N24DDR402	DDR4 RDIMM Memory,16GB,2400MT/s,2Rank(1G*8bit),1.2V, ECC	8
Hard Disk(include Front Panel)			
02311AYT	N2000ST7W3	HDD,2000GB,SATA 6Gb/s,7.2K rpm,64MB,3.5inch(3.5inch Drive Bay)	8
RAID Card and Other Accessories			
02311JEA	BC1M51ESMQ	SR320BC(LSI2208) SAS/SATA RAID Card,RAID0,1,5,6,10,50,60,6Gb/s,1GB Cache,used for RH2288 V3/RH2288H V3's 12HDD chassis	1
Cables and Optional Equipments			
21240434	EGUIDER01	2U Static Rail Kit	1
Soporte			
Technical Support Service			
88134UGQ	02311GGX-88134UGQ-48	RH2288 V3(12HDD EXP Chassis)H22M-03 For oversea(except Japan)-Co-Care Standard 9x5xNBD Service-48Month(s)	4
Subtotal Ser			

Figura B.1.: Detalles nodo worker

C. Nodo edge

Server RH2288 V3 - Edge			
<u>RH2288 V3</u>			
Base Configuration			
02311GGX	BC4M20HGSB	RH2288 V3 (12*3.5inch HDD Chassis)(Only for oversea,except Japan)H22M-03	1
02310YKM	BC1M01FGED	SM233 Onboard NIC,2x10GE Electrical Interface(X540),RJ45	1
02311AFU	BC1M06FAN	8038 Fan module	4
02131042	WEPW80015	460W GOLD AC Power Module	2
Broadwell EP CPU			
02311NEQ	BC1M56CPU	Intel Xeon E5-2640 v4(2.4GHz/10-core/25MB/90W) Processor (with heatsink)	2
DDR4 Memory			
06200213	N24DDR402	DDR4 RDIMM Memory,16GB,2400MT/s,2Rank(1G*8bit),1.2V, ECC	4
Hard Disk(include Front Panel)			
02311AYT	N2000ST7W3	HDD,2000GB,SATA 6Gb/s,7.2K rpm,64MB,3.5inch(3.5inch Drive Bay)	4
RAID Card and Other Accessories			
02311JEA	BC1M51ESMQ	SR320BC(LSI2208) SAS/SATA RAID Card,RAID0,1,5,6,10,50,60,6Gb/s,1GB Cache,used for RH2288 V3/RH2288H V3's 12HDD chassis	1
Cables and Optional Equipments			
21240434	EGUIDER01	2U Static Rail Kit	1
<u>Soporte</u>			
Technical Support Service			
88134UGQ	02311GGX-88134UGQ-48	RH2288 V3(12HDD EXP Chassis)H22M-03 For oversea(except Japan)-Co-Care Standard 9x5xNBD Service-48Month(s)	

Figura C.1.: Detalles nodo edge

D. Clúster Big Data ICAI



Figura D.1.: Clúster Big Data ICAI

E. Código del notebook K-Means, Madrid Residencial Invierno Diario

```
In [1]: from pyspark import SparkConf
        from pyspark.sql import SparkSession
        import pyspark.sql.functions as F

        import re
        import numpy as np
        import pandas as pd
```

```
In [2]: from pyspark.ml.clustering import KMeans
```

```
In [3]: conf = (
        SparkConf()
        .setAppName(u"K-Means, Madrid Residencial Invierno Diario")
        )
```

```
In [4]: spark = (
        SparkSession.builder
        .config(conf=conf)
        .enableHiveSupport()
        .getOrCreate()
        )
```

Lectura del fichero

```
In [5]: !date
        electricData = spark.read.options(header=True, inferSchema=True, de
        limiter="|").csv('/user/david/data/DATATHON_secuencial.csv').cache(
        )
        !date
```

```
Tue Jun 26 19:00:07 CEST 2018
Tue Jun 26 19:01:15 CEST 2018
```

Variables

```
In [6]: cnae = 'T1' # Residencial
        cnae_s = 'residencial'
        provincia = 3.0 # Madrid
        provincia_s = u'Madrid'
        dia = 1.0 # Laboral
        dia_s = u'día laboral'
        estacion = 1.0 # Invierno
        estacion_s = u'invierno'
```

Creación de las nuevas columnas

```
In [7]: electricData = electricData.withColumn("ACTIVA_TOTAL", sum(electricData[x] for x in electricData.columns if x.startswith("ACTIVA_H")).cast('float'))
```

Sólo consumo residencial

```
In [8]: electricData = electricData.filter(F.col('CNAE') == cnae)
```

Columna de provincias

```
In [9]: def buscarProvincia(municipio):
    coruna = ["SADA",u"A CORUÑA","OLEIROS","ARTEIXO","FISTERRA","CARBALLO","MONFERO","CULLEREDO","CAMBRE",u'MIÑO','BERGONDO','SANTIAGO DE COMPOSTELA']
    barcelona = [u"PIERA",u"SANTA COLOMA DE GRAMENET",u"SANTA MARGA RIDA DE MONTBUI",u"SABADELL",u"SITGES",u"BADIA DEL VALLES",u"SANT ANDREU DE LA BARCA",u"TARADELL",u"ARENYS DE MAR",u"ODENA",u"VIC",u"BADALONA",u"CASTELLDEFELS",u"CALLDETENES",u"MONTGAT",u"RUBI",u"SANT PERE DE RIBES",u"SANT BOI DE LLOBREGAT",u"BARCELONA",u"ARGENTONA",u"IGUALADA",u"ARENYS DE MUNT",u"MOLINS DE REI",u"TERRASSA",u"MANLLEU",u"ESPLUGUES DE LLOBREGAT",u"CORBERA DE LLOBREGAT",u"FOGARS DE LA SELVA",u"SANT FELIU DE LLOBREGAT",u"MALGRAT DE MAR",u"EL PRAT DE LL OBREGAT",u"CALDES D'ESTRAC",u"CASTELLOLI",u"MOLLET DEL VALLES",u"SANTA EUGENIA DE BERGA",u"BERGA",u"TORDERA",u"CERVELLO",u"MONTMANEU",u"JORBA",u"SANT MARTI DE TOUS",u"SANT CUGAT DEL VALLES",u"ABRERA",u"ELS HOSTALET DE PIEROLA",u"SANT MARTI SARROCA",u"CABRERA D'ANOIA",u"CABRERA D'ANOIA",u"MANRESA",u"PALAFOLLS",u"PACS DEL PENEDES",u"MATARO",u"LES MASIES DE VOLTREGA",u"SANTA SUSANNA"]
    madrid = ["MEJORADA DEL CAMPO","ALCALA DE HENARES","MADRID","COSLADA","POZUELO DE ALARCON","ARGANDA DEL REY","CAMARMA DE ESTERUELAS","SAN FERNANDO DE HENARES","ARANJUEZ","RIVAS-VACIAMADRID","DAGANZO DE ARRIBA","VILLALBILLA","LEGANES","VALDEOLMOS-ALALPARDO","TORREJON DE ARDOZ","VILLAR DEL OLMO","MOSTOLES","TORRES DE LA ALAMEDA","ALCORCON","VALDEOLMOS","CAMPO REAL","VELILLA DE SAN ANTONIO","LAS ROZAS DE MADRID","ESTREMERAS","ALCOBENDAS","PARACUELLOS DE JARAMA","VALDEAVERO","CHINCHON","MONTEJO","COLMENAR DE OREJA","LOECHES"]
    malaga = ["MARBELLA","FUENGIROLA","VELEZ-MALAGA","TORROX","ESTEPONA","VALLE DE ABDALAJIS","MALAGA","BENALMADENA","ANTEQUERA","MONDIA","RONDA","MIJAS","ALGARROBO","IZNATE","VILLANUEVA DE LA CONCEPCION","CARTAMA","YUNQUERA","ALHAURIN DE LA TORRE"]
    if municipio in coruna :
        return 1.0
    elif municipio in barcelona :
        return 2.0
    elif municipio in madrid :
        return 3.0
    elif municipio in malaga:
        return 4.0
    else :
        return 0
```

```
In [10]: buscarProvinciaUDF = F.udf(lambda z: buscarProvincia(z))
electricData = electricData.withColumn("PROVINCIA", buscarProvincia
UDF(F.col('DE_MUNICIP')).cast('float'))
```

```
In [11]: electricData = electricData.filter(F.col('PROVINCIA') == provincia)
```

Columna de tipo de día

```
In [12]: import datetime
def dayOfWeekElectric(fecha):
    fecha = str(fecha)
    diaSemana = datetime.datetime(int(fecha[:4]), int(fecha[4:6]),
int(fecha[6:8])).weekday()+1
    if (diaSemana<6):
        return '1'
    elif (diaSemana<7):
        return '2'
    else:
        return '3'
```

```
In [13]: dayOfWeekElectricUDF = F.udf(lambda z: dayOfWeekElectric(z))
electricData = electricData.withColumn("TIPO_DIA", dayOfWeekElectri
cUDF(F.col('DIA')).cast('float'))
```

```
In [14]: electricData = electricData.filter(F.col('TIPO_DIA') == dia)
```

Columna de estación del año

```
In [15]: import datetime
def buscarEstacion(fecha):
    fecha = str(fecha)
    md = int(fecha[4:8])
    if ((md >= 320) and (md <= 619)):
        s = 2.0 # primavera
    elif ((md >= 620) and (md <= 921)):
        s = 3.0 # verano
    elif ((md >= 922) and (md <= 1220)):
        s = 4.0 # otoño
    elif ((md >= 1221) or (md <= 319)):
        s = 1.0 # invierno
    else:
        raise IndexError(u"Fecha inválida")
    return s
```

```
In [16]: buscarEstacionUDF = F.udf(lambda z: buscarEstacion(z))
electricData = electricData.withColumn("ESTACION", buscarEstacionUD
F(F.col("DIA")).cast('float'))
```

```
In [17]: electricData = electricData.filter(F.col('ESTACION') == estacion)
```

```
In [18]: electricData = electricData.filter((F.col('ACTIVA_H1') > 0) & (F.col('ACTIVA_H2') > 0) & (F.col('ACTIVA_H3') > 0) & (F.col('ACTIVA_H4') > 0) & (F.col('ACTIVA_H5') > 0) & (F.col('ACTIVA_H6') > 0) & (F.col('ACTIVA_H6') > 0) & (F.col('ACTIVA_H7') > 0) & (F.col('ACTIVA_H8') > 0) & (F.col('ACTIVA_H9') > 0) & (F.col('ACTIVA_H10') > 0) & (F.col('ACTIVA_H11') > 0) & (F.col('ACTIVA_H12') > 0) & (F.col('ACTIVA_H13') > 0) & (F.col('ACTIVA_H13') > 0) & (F.col('ACTIVA_H14') > 0) & (F.col('ACTIVA_H15') > 0) & (F.col('ACTIVA_H16') > 0) & (F.col('ACTIVA_H17') > 0) & (F.col('ACTIVA_H18') > 0) & (F.col('ACTIVA_H19') > 0) & (F.col('ACTIVA_H20') > 0) & (F.col('ACTIVA_H21') > 0) & (F.col('ACTIVA_H22') > 0) & (F.col('ACTIVA_H23') > 0) & (F.col('ACTIVA_H24') > 0))
```

Descartar el 25% de los datos

```
In [19]: electricData.count()
```

```
Out[19]: 390784
```

```
In [20]: [trainingData, testData] = electricData.randomSplit([0.75, 0.25])
```

```
In [21]: from pyspark.sql.functions import monotonically_increasing_id
trainingData = trainingData.withColumn('ID', monotonically_increasing_id()+1)
```

```
In [22]: trainingData.select('ID', 'ACTIVA_H1', 'ACTIVA_H2').show(5)
```

```
+---+-----+-----+
| ID|ACTIVA_H1|ACTIVA_H2|
+---+-----+-----+
|  1|      51|      42|
|  2|      70|      51|
|  3|     149|     126|
|  4|     153|      57|
|  5|     190|     211|
+---+-----+-----+
only showing top 5 rows
```

```
In [23]: FEATURES_COL = ['ACTIVA_H1', 'ACTIVA_H2', 'ACTIVA_H3', 'ACTIVA_H4', 'ACTIVA_H5', 'ACTIVA_H6', 'ACTIVA_H7', 'ACTIVA_H8', 'ACTIVA_H9', 'ACTIVA_H10', 'ACTIVA_H11', 'ACTIVA_H12', 'ACTIVA_H13', 'ACTIVA_H14', 'ACTIVA_H15', 'ACTIVA_H16', 'ACTIVA_H17', 'ACTIVA_H18', 'ACTIVA_H19', 'ACTIVA_H20', 'ACTIVA_H21', 'ACTIVA_H22', 'ACTIVA_H23', 'ACTIVA_H24']
for col in trainingData.columns:
    if col in FEATURES_COL:
        trainingData = trainingData.withColumn(col, trainingData[col].cast('float'))
```

```
In [24]: trainingData.select('ID', 'ACTIVA_H1', 'ACTIVA_H2').show(5)
```

```
+---+-----+-----+
| ID|ACTIVA_H1|ACTIVA_H2|
+---+-----+-----+
|  1|      51.0|      42.0|
|  2|      70.0|      51.0|
|  3|     149.0|     126.0|
|  4|     153.0|      57.0|
|  5|     190.0|     211.0|
+---+-----+-----+
only showing top 5 rows
```

Creación de la columna de features

```
In [25]: from pyspark.ml.feature import VectorAssembler
```

```
FEATURES_COL = ['ACTIVA_H1', 'ACTIVA_H2', 'ACTIVA_H3', 'ACTIVA_H4', 'ACTIVA_H5', 'ACTIVA_H6', 'ACTIVA_H7', 'ACTIVA_H8', 'ACTIVA_H9', 'ACTIVA_H10', 'ACTIVA_H11', 'ACTIVA_H12', 'ACTIVA_H13', 'ACTIVA_H14', 'ACTIVA_H15', 'ACTIVA_H16', 'ACTIVA_H17', 'ACTIVA_H18', 'ACTIVA_H19', 'ACTIVA_H20', 'ACTIVA_H21', 'ACTIVA_H22', 'ACTIVA_H23', 'ACTIVA_H24']
vecAssembler = VectorAssembler(inputCols=FEATURES_COL, outputCol="features")
df_kmeans = vecAssembler.transform(trainingData).select('ID', 'features')
```

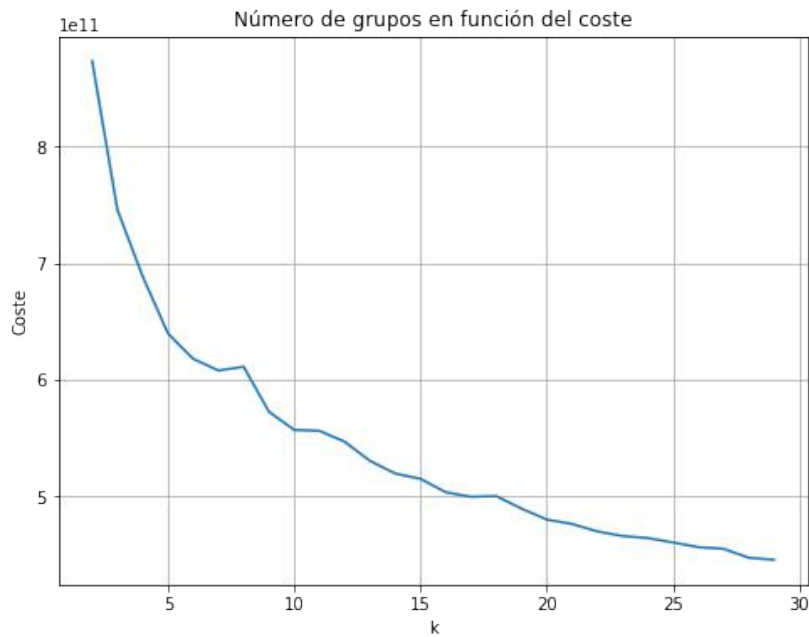
```
In [26]: df_kmeans.select('ID', 'features').show(5)
```

```
+---+-----+-----+
| ID|          features|
+---+-----+-----+
|  1|[51.0, 42.0, 33.0, 4...|
|  2|[70.0, 51.0, 16.0, 3...|
|  3|[149.0, 126.0, 114....|
|  4|[153.0, 57.0, 30.0, ...|
|  5|[190.0, 211.0, 255....|
+---+-----+-----+
only showing top 5 rows
```

Encontrar el número de clusters a emplear

```
In [81]: cost = np.zeros(30)
for k in range(2, 30):
    kmeans = KMeans().setK(k).setSeed(1).setFeaturesCol("features")
    model = kmeans.fit(df_kmeans.sample(False, 0.1, seed=42))
    cost[k] = model.computeCost(df_kmeans)
```

```
In [82]: import matplotlib.pyplot as plt
fig, ax = plt.subplots(1,1, figsize =(8,6))
ax.plot(range(2,30),cost[2:30])
ax.set_xlabel('k')
ax.set_ylabel('Coste')
plt.title(u'Número de grupos en función del coste')
plt.grid()
plt.show()
```



```
In [34]: cost[5]
```

```
Out[34]: 638230680293.3988
```

```
In [35]: cost[6]
```

```
Out[35]: 614313877823.0002
```

```
In [36]: cost[7]
```

```
Out[36]: 598946567693.5476
```

```
In [37]: (cost[6]-cost[5])/cost[5]
```

```
Out[37]: -0.03747360195753023
```

```
In [38]: (cost[7]-cost[6])/cost[6]
```

```
Out[38]: -0.025015404476798026
```

```
In [39]: (cost[8]-cost[7])/cost[7]
```

```
Out[39]: -0.032792073585222424
```

```
In [40]: (cost[9]-cost[8])/cost[8]
```

```
Out[40]: -0.030231438392055922
```

```
In [41]: (cost[10]-cost[9])/cost[9]
```

```
Out[41]: -0.017014408771334474
```

Calcular el centro del los clusters

```
In [27]: k = 8  
kmeans = KMeans().setK(k).setSeed(1).setFeaturesCol("features")  
model = kmeans.fit(df_kmeans)  
centers = model.clusterCenters()
```

Predicciones

```
In [28]: transformed = model.transform(df_kmeans).select('ID', 'prediction')  
rows = transformed.collect()  
print(rows[:3])  
df_pred = spark.createDataFrame(rows)  
df_pred.show()
```

```
[Row(ID=1, prediction=2), Row(ID=2, prediction=2), Row(ID=3, prediction=2)]
```

```
+---+-----+
| ID|prediction|
+---+-----+
| 1|          2|
| 2|          2|
| 3|          2|
| 4|          2|
| 5|          0|
| 6|          7|
| 7|          7|
| 8|          6|
| 9|          7|
|10|          7|
|11|          7|
|12|          6|
|13|          4|
|14|          2|
|15|          0|
|16|          2|
|17|          7|
|18|          3|
|19|          6|
|20|          6|
+---+-----+
```

only showing top 20 rows

In [29]: df_pred

Out[29]: DataFrame[ID: bigint, prediction: bigint]

In [30]: df_pred = df_pred.join(df_kmeans, 'ID').sort('ID')
df_pred.show(5)

```
+---+-----+-----+
| ID|prediction|      features|
+---+-----+-----+
| 1|          2|[51.0,42.0,33.0,4...|
| 2|          2|[70.0,51.0,16.0,3...|
| 3|          2|[149.0,126.0,114....|
| 4|          2|[153.0,57.0,30.0,...|
| 5|          0|[190.0,211.0,255....|
+---+-----+-----+
```

only showing top 5 rows

Procentaje de casos en cada grupo

In [31]: total = df_pred.count()

In [32]: total = float(total)

```
In [33]: porcentaje_g1 = (df_pred.filter(F.col('prediction') == 0).count() /  
total) * 100  
porcentaje_g1
```

```
Out[33]: 35.835989548420635
```

```
In [34]: porcentaje_g2 = (df_pred.filter(F.col('prediction') == 1).count() /  
total) * 100  
porcentaje_g2
```

```
Out[34]: 0.4059563058318171
```

```
In [35]: porcentaje_g3 = (df_pred.filter(F.col('prediction') == 2).count() /  
total) * 100  
porcentaje_g3
```

```
Out[35]: 35.52408377679585
```

```
In [36]: porcentaje_g4 = (df_pred.filter(F.col('prediction') == 3).count() /  
total) * 100  
porcentaje_g4
```

```
Out[36]: 6.607135528529802
```

```
In [37]: porcentaje_g5 = (df_pred.filter(F.col('prediction') == 4).count() /  
total) * 100  
porcentaje_g5
```

```
Out[37]: 1.5615808697793403
```

```
In [38]: porcentaje_g6 = (df_pred.filter(F.col('prediction') == 5).count() /  
total) * 100  
porcentaje_g6
```

```
Out[38]: 0.6610897549898084
```

```
In [39]: porcentaje_g7 = (df_pred.filter(F.col('prediction') == 6).count() /  
total) * 100  
porcentaje_g7
```

```
Out[39]: 4.276050287965635
```

```
In [40]: porcentaje_g8 = (df_pred.filter(F.col('prediction') == 7).count() /  
total) * 100  
porcentaje_g8
```

```
Out[40]: 15.128113927687108
```

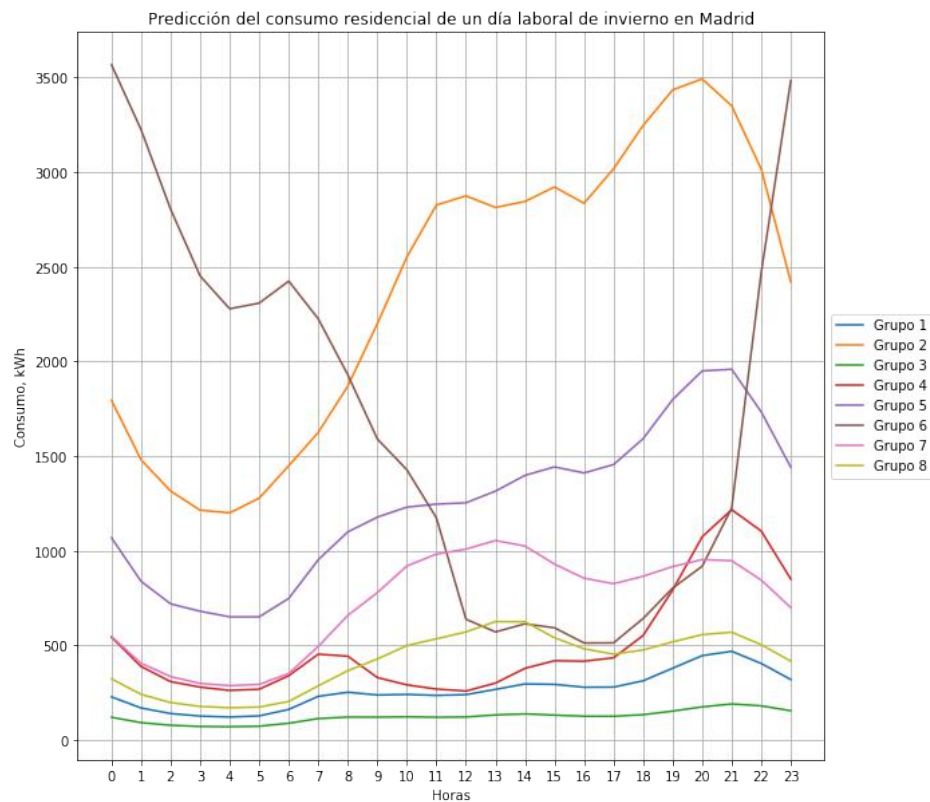
Gráfico de los clusters obtenidos

```
In [41]: import matplotlib.pyplot as plt

plt.figure(1,figsize = (10,10))

plt.plot(centers[0],'- ',label=u'Grupo 1',color='#1f77b4')
plt.plot(centers[1],'- ',label=u'Grupo 2',color='#ff7f0e')
plt.plot(centers[2],'- ',label=u'Grupo 3',color='#2ca02c')
plt.plot(centers[3],'- ',label=u'Grupo 4',color='#d62728')
plt.plot(centers[4],'- ',label=u'Grupo 5',color='#9467bd')
plt.plot(centers[5],'- ',label=u'Grupo 6',color='#8c564b')
plt.plot(centers[6],'- ',label=u'Grupo 7',color='#e377c2')
plt.plot(centers[7],'- ',label=u'Grupo 8',color='#bcbd22')

plt.title(u'Predicción del consumo ' + cnae_s + ' de un ' + dia_s +
' de ' + estacion_s + ' en ' + provincia_s)
plt.ylabel('Consumo, kWh')
plt.xlabel('Horas')
plt.xticks(np.arange(0, 24,1))
plt.grid()
plt.legend(loc='center left', bbox_to_anchor=(1, 0.5))
plt.show()
```



Desviación de los datos con respecto a su centroide

```
In [42]: grupo1 = df_pred.filter(F.col('prediction') == 0)
grupo1 = grupo1.select('features').toPandas()
centroide = centers[0]
horas_max_g1 = [centroide[0] + grupo1.features.map(lambda x: x[0]).
std()]
horas_min_g1 = [centroide [0]- grupo1.features.map(lambda x: x[0]).
std()]
for i in range (1,24):
    horas_max_g1.append(centroide[i] + grupo1.features.map(lambda x
: x[i]).std())
    aux = centroide[i] - grupo1.features.map(lambda x: x[i]).std()
    if aux >= 0:
        horas_min_g1.append(aux)
    else :
        horas_min_g1.append(0)
```

```
In [43]: grupo2 = df_pred.filter(F.col('prediction') == 1)
grupo2 = grupo2.select('features').toPandas()
centroide = centers[1]
horas_max_g2 = [centroide[0] + grupo2.features.map(lambda x: x[0]).
std()]
horas_min_g2 = [centroide [0]- grupo2.features.map(lambda x: x[0]).
std()]
for i in range (1,24):
    horas_max_g2.append(centroide[i] + grupo2.features.map(lambda x
: x[i]).std())
    aux = centroide[i] - grupo2.features.map(lambda x: x[i]).std()
    if aux >= 0 :
        horas_min_g2.append(aux)
    else :
        horas_min_g2.append(0)
```

```
In [44]: grupo3 = df_pred.filter(F.col('prediction') == 2)
grupo3 = grupo3.select('features').toPandas()
centroide = centers[2]
horas_max_g3 = [centroide[0] + grupo3.features.map(lambda x: x[0]).
std()]
horas_min_g3 = [centroide [0]- grupo3.features.map(lambda x: x[0]).
std()]
for i in range (1,24):
    horas_max_g3.append(centroide[i] + grupo3.features.map(lambda x
: x[i]).std())
    aux = centroide[i] - grupo3.features.map(lambda x: x[i]).std()
    if aux >= 0:
        horas_min_g3.append(aux)
    else :
        horas_min_g3.append(0)
```

```
In [45]: grupo4 = df_pred.filter(F.col('prediction') == 3)
grupo4 = grupo4.select('features').toPandas()
centroide = centers[3]
horas_max_g4 = [centroide[0] + grupo4.features.map(lambda x: x[0]).std()]
horas_min_g4 = [centroide [0]- grupo4.features.map(lambda x: x[0]).std()]
for i in range (1,24):
    horas_max_g4.append(centroide[i] + grupo4.features.map(lambda x : x[i]).std())
    aux = centroide[i] - grupo4.features.map(lambda x: x[i]).std()
    if aux >= 0:
        horas_min_g4.append(aux)
    else :
        horas_min_g4.append(0)
```

```
In [46]: grupo5 = df_pred.filter(F.col('prediction') == 4)
grupo5 = grupo5.select('features').toPandas()
centroide = centers[4]
horas_max_g5 = [centroide[0] + grupo5.features.map(lambda x: x[0]).std()]
horas_min_g5 = [centroide [0]- grupo5.features.map(lambda x: x[0]).std()]
for i in range (1,24):
    horas_max_g5.append(centroide[i] + grupo5.features.map(lambda x : x[i]).std())
    aux = centroide[i] - grupo5.features.map(lambda x: x[i]).std()
    if aux >= 0:
        horas_min_g5.append(aux)
    else :
        horas_min_g5.append(0)
```

```
In [47]: grupo6 = df_pred.filter(F.col('prediction') == 5)
grupo6 = grupo6.select('features').toPandas()
centroide = centers[5]
horas_max_g6 = [centroide[0] + grupo6.features.map(lambda x: x[0]).std()]
horas_min_g6 = [centroide [0]- grupo6.features.map(lambda x: x[0]).std()]
for i in range (1,24):
    horas_max_g6.append(centroide[i] + grupo6.features.map(lambda x : x[i]).std())
    aux = centroide[i] - grupo6.features.map(lambda x: x[i]).std()
    if aux >= 0:
        horas_min_g6.append(aux)
    else :
        horas_min_g6.append(0)
```

```
In [48]: grupo7 = df_pred.filter(F.col('prediction') == 6)
grupo7 = grupo7.select('features').toPandas()
centroide = centers[6]
horas_max_g7 = [centroide[0] + grupo7.features.map(lambda x: x[0]).
std()]
horas_min_g7 = [centroide [0]- grupo7.features.map(lambda x: x[0]).
std()]
for i in range (1,24):
    horas_max_g7.append(centroide[i] + grupo7.features.map(lambda x
: x[i]).std())
    aux = centroide[i] - grupo7.features.map(lambda x: x[i]).std()
    if aux >= 0 :
        horas_min_g7.append(aux)
    else :
        horas_min_g7.append(0)
```

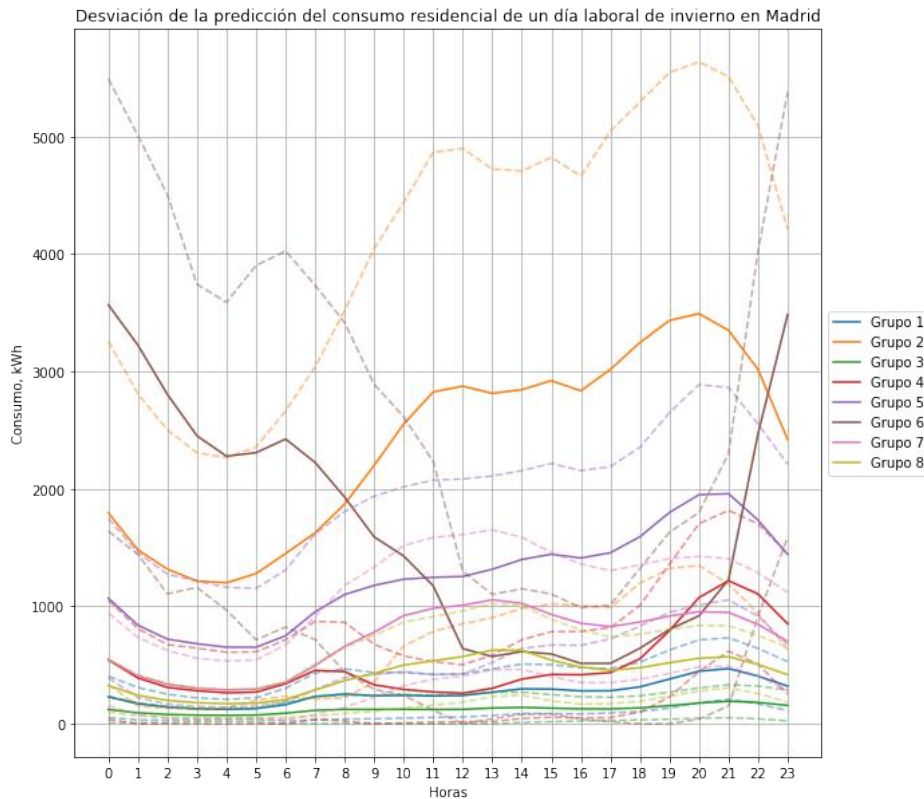
```
In [49]: grupo8 = df_pred.filter(F.col('prediction') == 7)
grupo8 = grupo8.select('features').toPandas()
centroide = centers[7]
horas_max_g8 = [centroide[0] + grupo8.features.map(lambda x: x[0]).
std()]
horas_min_g8 = [centroide [0]- grupo8.features.map(lambda x: x[0]).
std()]
for i in range (1,24):
    horas_max_g8.append(centroide[i] + grupo8.features.map(lambda x
: x[i]).std())
    aux = centroide[i] - grupo8.features.map(lambda x: x[i]).std()
    if aux >= 0:
        horas_min_g8.append(aux)
    else :
        horas_min_g8.append(0)
```

```
In [50]: import matplotlib.pyplot as plt

plt.figure(1,figsize = (10,10))

g1 = plt.plot(centers[0], '- ', label=u'Grupo 1', color='#1f77b4')
g1_max = plt.plot(horas_max_g1, '-- ', color='#1f77b4', alpha=0.5)
g1_min = plt.plot(horas_min_g1, '-- ', color='#1f77b4', alpha=0.5)
g2 = plt.plot(centers[1], '- ', label=u'Grupo 2', color='#ff7f0e')
g2_max = plt.plot(horas_max_g2, '-- ', color='#ff7f0e', alpha=0.5)
g2_min = plt.plot(horas_min_g2, '-- ', color='#ff7f0e', alpha=0.5)
g3 = plt.plot(centers[2], '- ', label=u'Grupo 3', color='#2ca02c')
g3_max = plt.plot(horas_max_g3, '-- ', color='#2ca02c', alpha=0.5)
g3_min = plt.plot(horas_min_g3, '-- ', color='#2ca02c', alpha=0.5)
g4 = plt.plot(centers[3], '- ', label=u'Grupo 4', color='#d62728')
g4_max = plt.plot(horas_max_g4, '-- ', color='#d62728', alpha=0.5)
g4_min = plt.plot(horas_min_g4, '-- ', color='#d62728', alpha=0.5)
g5 = plt.plot(centers[4], '- ', label=u'Grupo 5', color='#9467bd')
g5_max = plt.plot(horas_max_g5, '-- ', color='#9467bd', alpha=0.5)
g5_min = plt.plot(horas_min_g5, '-- ', color='#9467bd', alpha=0.5)
g6 = plt.plot(centers[5], '- ', label=u'Grupo 6', color='#8c564b')
g6_max = plt.plot(horas_max_g6, '-- ', color='#8c564b', alpha=0.5)
g6_min = plt.plot(horas_min_g6, '-- ', color='#8c564b', alpha=0.5)
g7 = plt.plot(centers[6], '- ', label=u'Grupo 7', color='#e377c2')
g7_max = plt.plot(horas_max_g7, '-- ', color='#e377c2', alpha=0.5)
g7_min = plt.plot(horas_min_g7, '-- ', color='#e377c2', alpha=0.5)
g8 = plt.plot(centers[7], '- ', label=u'Grupo 8', color='#bcbd22')
g8_max = plt.plot(horas_max_g8, '-- ', color='#bcbd22', alpha=0.5)
g8_min = plt.plot(horas_min_g8, '-- ', color='#bcbd22', alpha=0.5)

plt.title(u'Desviación de la predicción del consumo ' + cnae_s + '
de un ' + dia_s + ' de ' + estacion_s + ' en ' + provincia_s)
plt.ylabel('Consumo, kWh')
plt.xlabel('Horas')
#plt.ylim([0,20000])
plt.xticks(np.arange(0, 24,1))
plt.grid()
plt.legend(loc='center left', bbox_to_anchor=(1, 0.5))
plt.show()
```



Validación de predicciones con el 25% restante

```
In [51]: from pyspark.sql.functions import monotonically_increasing_id
testData = testData.withColumn('ID',monotonically_increasing_id()+1
)
```

```
In [52]: FEATURES_COL = ['ACTIVA_H1','ACTIVA_H2','ACTIVA_H3','ACTIVA_H4','AC
TIVA_H5','ACTIVA_H6','ACTIVA_H7','ACTIVA_H8','ACTIVA_H9','ACTIVA_H1
0','ACTIVA_H11','ACTIVA_H12','ACTIVA_H13','ACTIVA_H14','ACTIVA_H15'
,'ACTIVA_H16','ACTIVA_H17','ACTIVA_H18','ACTIVA_H19','ACTIVA_H20','
ACTIVA_H21','ACTIVA_H22','ACTIVA_H23','ACTIVA_H24']
for col in testData.columns:
    if col in FEATURES_COL:
        testData = testData.withColumn(col,testData[col].cast('floa
t'))
```

Creación de la columna features

```
In [53]: from pyspark.ml.feature import VectorAssembler

FEATURES_COL = ['ACTIVA_H1', 'ACTIVA_H2', 'ACTIVA_H3', 'ACTIVA_H4', 'ACTIVA_H5', 'ACTIVA_H6', 'ACTIVA_H7', 'ACTIVA_H8', 'ACTIVA_H9', 'ACTIVA_H10', 'ACTIVA_H11', 'ACTIVA_H12', 'ACTIVA_H13', 'ACTIVA_H14', 'ACTIVA_H15', 'ACTIVA_H16', 'ACTIVA_H17', 'ACTIVA_H18', 'ACTIVA_H19', 'ACTIVA_H20', 'ACTIVA_H21', 'ACTIVA_H22', 'ACTIVA_H23', 'ACTIVA_H24']
vecAssembler = VectorAssembler(inputCols=FEATURES_COL, outputCol="features")
df_kmeans_test = vecAssembler.transform(testData).select('ID', 'features')
```

```
In [54]: df_kmeans_test.select('ID', 'features').show(5)
```

```
+---+-----+
| ID|          features|
+---+-----+
|  1|[101.0, 89.0, 62.0, ...|
|  2|[802.0, 535.0, 287....|
|  3|[55.0, 41.0, 32.0, 3...|
|  4|[213.0, 130.0, 133....|
|  5|[318.0, 241.0, 251....|
+---+-----+
only showing top 5 rows
```

Predicciones

```
In [55]: transformed = model.transform(df_kmeans_test).select('ID', 'prediction')
rows = transformed.collect()
print(rows[:3])
df_pred_test = spark.createDataFrame(rows)

[Row(ID=1, prediction=2), Row(ID=2, prediction=6), Row(ID=3, prediction=0)]
```

```
In [56]: df_pred_test = df_pred_test.join(df_kmeans_test, 'ID').sort('ID')
df_pred_test.show()
```

```
+---+-----+-----+
| ID|prediction|      features|
+---+-----+-----+
|  1|         2|[101.0, 89.0, 62.0, ...|
|  2|         6|[802.0, 535.0, 287. ...|
|  3|         0|[55.0, 41.0, 32.0, 3. ...|
|  4|         0|[213.0, 130.0, 133. ...|
|  5|         3|[318.0, 241.0, 251. ...|
|  6|         7|[1085.0, 474.0, 296. ...|
|  7|         2|[94.0, 62.0, 61.0, 7. ...|
|  8|         7|[493.0, 415.0, 230. ...|
|  9|         2|[20.0, 68.0, 169.0, ...|
| 10|         2|[140.0, 91.0, 59.0, ...|
| 11|         2|[174.0, 181.0, 119. ...|
| 12|         2|[33.0, 15.0, 77.0, 9. ...|
| 13|         0|[375.0, 247.0, 129. ...|
| 14|         7|[386.0, 434.0, 94.0. ...|
| 15|         6|[562.0, 351.0, 208. ...|
| 16|         7|[146.0, 91.0, 74.0, ...|
| 17|         2|[168.0, 112.0, 74.0. ...|
| 18|         7|[217.0, 138.0, 112. ...|
| 19|         0|[316.0, 211.0, 140. ...|
| 20|         7|[347.0, 293.0, 260. ...|
+---+-----+-----+
only showing top 20 rows
```

Porcentaje de cada grupo

```
In [57]: total = df_pred_test.count()
```

```
In [58]: total = float(total)
```

```
In [59]: porcentaje_g1_test = (df_pred_test.filter(F.col('prediction') == 0)
      .count() / total) * 100
porcentaje_g1_test
```

```
Out[59]: 35.64459080375656
```

```
In [60]: porcentaje_g2_test = (df_pred_test.filter(F.col('prediction') == 1)
      .count() / total) * 100
porcentaje_g2_test
```

```
Out[60]: 0.4004553400821238
```

```
In [61]: porcentaje_g3_test = (df_pred_test.filter(F.col('prediction') == 2)
      .count() / total) * 100
porcentaje_g3_test
```

```
Out[61]: 35.3386591860796
```

```
In [62]: porcentaje_g4_test = (df_pred_test.filter(F.col('prediction') == 3)
        .count() / total) * 100
        porcentaje_g4_test
```

```
Out[62]: 6.570923283327234
```

```
In [63]: porcentaje_g5_test = (df_pred_test.filter(F.col('prediction') == 4)
        .count() / total) * 100
        porcentaje_g5_test
```

```
Out[63]: 1.6434931089157212
```

```
In [64]: porcentaje_g6_test = (df_pred_test.filter(F.col('prediction') == 5)
        .count() / total) * 100
        porcentaje_g6_test
```

```
Out[64]: 0.6301581493678091
```

```
In [65]: porcentaje_g7_test = (df_pred_test.filter(F.col('prediction') == 6)
        .count() / total) * 100
        porcentaje_g7_test
```

```
Out[65]: 4.328779932512095
```

```
In [66]: porcentaje_g8_test = (df_pred_test.filter(F.col('prediction') == 7)
        .count() / total) * 100
        porcentaje_g8_test
```

```
Out[66]: 15.442940195958856
```

Media de las variaciones

```
In [67]: variacion_g1 = abs(porcentaje_g1 - porcentaje_g1_test)
        variacion_g2 = abs(porcentaje_g2 - porcentaje_g2_test)
        variacion_g3 = abs(porcentaje_g3 - porcentaje_g3_test)
        variacion_g4 = abs(porcentaje_g4 - porcentaje_g4_test)
        variacion_g5 = abs(porcentaje_g5 - porcentaje_g5_test)
        variacion_g6 = abs(porcentaje_g6 - porcentaje_g6_test)
        variacion_g7 = abs(porcentaje_g7 - porcentaje_g7_test)
        variacion_g8 = abs(porcentaje_g8 - porcentaje_g8_test)
```

```
In [68]: variaciones = [variacion_g1, variacion_g2, variacion_g3, variacion_g4,
        variacion_g5, variacion_g6, variacion_g7, variacion_g8]
        np.mean(variaciones)
```

```
Out[68]: 0.11236703798864643
```


F. Consumo anual de los municipios

DE_MUNICIP	CONSUMO ANUAL TOTAL (kWh)
MADRID	16980998236
SANT BOI DE LLOBREGAT	10044053219
RUBI	8431887781
MAIRENA DEL ALCOR	8072565136
L'HOSPITALET DE LLOBREGAT	7332389656
SANTA COLOMA DE GRAMENET	6990004513
TORROX	6837054973
BAILEN	6809571341
SANT PERE DE RIBES	6535862191
SANT CARLES DE LA RAPITA	5396616001
SANTANDER	5243358736
HUELVA	4734805892
SALAMANCA	3610684500
ZARAGOZA	3600735112
UBRIQUE	3546214269
PEÑARROYA-PUEBLONUEVO	3350463711
PALMA	3262265004
BILBAO	3128266077
ALBACETE	3113836188
ARENYS DE MAR	2897858078
GANDIA	2835287154
ARGENTONA	2690644240
BURGOS	2595033201
VINAROS	2553210870
SAN FERNANDO	2451607729
CAZORLA	2384621102
TORREDELCAMPO	2365304892
CASTELLON DE LA PLANA	2218999180
MALAGA	1967083756
GRANADA	1867328700
CORDOBA	1723448354
SEVILLA	1663495445
VALVERDE	1640552619
S/C DE TENERIFE	1627438685
BENIDORM	1479005877

PORTUGALETE	1477231434
BEAS DE SEGURA	1427029068
BREÑA ALTA	1424128327
BARCELONA	1398914563
CACERES	1325614730
CUERVO (EL)	1290940393
MONTIJO	1267993699
VILLARREAL/VILA-REAL	1253141965
ARENYS DE MUNT	1214146874
PAMPLONA	1212822013
ANDRATX	1201491528
LORCA	1187515055
GRANADILLA DE ABONA	1134945764
LAS PALMAS DE G.C.	1125420426
A CORUÑA	1125395686
FORMENTERA	1069858979
VALENCIA	1062407064
BENALMADENA	1055675360
FRONTERA	1018879039
ALCALA DE HENARES	956699798
TORRELAVEGA	931724597
BADAJOS	928593735
PUERTO DEL ROSARIO	900201769
SAN BARTOLOME	817737816
CARTAGENA	766532137
RUTE	765833824
JAEN	748689401
AVILA	741739054
EIVISSA	727806269
SANTOÑA	725224269
REUS	724514345
LA UNION	713781515
ADEMUZ	678180219
TELDE	660020595
SES SALINES	646852743
LAREDO	643200309
CHICLANA DE LA FRONTERA	640775710
SANT LLUIS	634584811
EL PINAR DE EL HIERRO	634358347
SAN BARTOLOME DE TIRAJANA	610395658
ANTEQUERA	591917674
PUERTO DE SANTA MARIA (EL)	585881825
ARUCAS	580923264
MEDIO CUDEYO	578886423
CAMARMA DE ESTERUELAS	550712687

VIGO	543365764
AGUIMES	537659902
CAMARGO	536055198
POZUELO DE ALARCON	532764881
ESTEPONA	528062275
ARRECIFE	517035386
VILLARES DE LA REINA	509215993
PUERTO DE LA CRUZ	505939273
MARTOS	500083345
MAIRENA DEL ALJARAFE	472292381
PALMA DEL CONDADO (LA)	465864050
MANACOR	463179861
BADIA DEL VALLES	462068849
POLLENÇA	461508675
ADEJE	461004593
SANTA MARIA DE GUIA	451278261
BREÑA BAJA	446251819
CABEZON DE LA SAL	445524038
VIC	443794097
PIELAGOS	433431780
SUANCES	429366756
ALMENARA	427302584
SAN MIGUEL DE ABONA	421626115
ALMONACID DE LA SIERRA	413161313
TAUSTE	400634204
HERMIGUA	382976758
PEÑISCOLA	370959323
TARRAGONA	370240700
SANTA CRUZ DE BEZANA	361117466
JAVEA	353287785
ARONA	348203977
ALCALA DE XIVERT	341204066
CASTELLON	338174735
LOJA	327134935
VILA-REAL	324893908
ROTA	324263864
TARAZONA DE LA MANCHA	318432677
MIENGO	313828358
SARIÑENA	306470329
MONTELLANO	297474307
SANTANYI	293983086
OURENSE	286736018
FUENGIROLA	282038309
REINOSA	268688214
MARBELLA	266289162

TERRASSA	264963114
ALGECIRAS	261150289
BURRIANA	256954072
MARINA DE CUDEYO	256874578
VALDALIGA	256853030
TORREJON DEL REY	255441001
DENIA	251111967
POLANCO	250272801
JEREZ DE LA FRONTERA	230397544
SAN VICENTE DE LA BARQUERA	229087048
GUISSONA	225822332
VELEZ-MALAGA	225180018
HUESCA	219430767
ALMERIA	218254210
ARNUERO	212848336
SANT JOAN DE LABRITJA	212118723
LIERGANES	203553651
MAÓ	202811805
MORON FRONTERA	201586096
PUENTE VIESGO	201021262
SAN JAVIER	198209844
CHILCHES	194253055
COSLADA	192908384
ENTRAMBASAGUAS	187713181
MOTRIL	184489487
SANT JOSEP DE SA TALAIA	183808978
CAPDEPERA	181657817
LLEIDA	177642112
DOS HERMANAS	174845242
CARTES	174048643
SANTILLANA DEL MAR	172008845
BADALONA	168798017
SANT ANDREU DE LA BARCA	168644119
MIERES	165265324
CIUTADELLA DE MENORCA	162072748
UCEDA	158633258
BORMUJOS	158332916
CASTELLANOS DE MORISCOS	158063243
POZO CAÑADA	157839298
RENEDO	150172052
SAN FELICES DE BUELNA	149815823
CORVERA DE TORANZO	149559555
ARICO	148233723
SAN CRISTOBAL DE LA LAGUNA	147534551

MONT-ROIG DEL CAMP	147315055
CASTELLDEFELS	146307424
GUADALAJARA	143494859
HERRERA DE PISUERGA	143464671
SAN VICENTE DE ALCANTARA	141722984
TORREDONJIMENO	141007097
LA GINETA	140990878
S/C DE LA PALMA	140897390
VALDES	139128075
SESTAO	138233880
SANTURTZI	137994775
QUART DE POBLET	135309819
MOGAN	133430258
ESPARTINAS	128832902
SABADELL	126864458
AGRAMUNT	126300274
TOLOSA	119617099
CAMPOO DE ENMEDIO	119550748
CHINCHILLA DE MONTE-ARAGON	119464460
PAMPLONA/IRUÑA	119086835
PUNTA UMBRIA	116113936
VILLANUEVA DE HUERVA	115713802
DAIMIEL	114188107
TORDERA	111804135
PONTEVEDRA	109721301
VALLADOLID	109617712
BARAKALDO	106250355
EL VENDRELL	105137999
LA POLA DE GORDON	103994413
ONDA	103748415
VALDEAVERO	100359015
ROQUETAS DE MAR	100353748
ILLESCAS	100125297
RIONANSA	99401514
SANTA MARGALIDA	98577196
ELS POBLETS	98006604
SANTA EULARIA DES RIU	97927829
MIGUEL ESTEBAN	97594459
BAREYO	96723283
FISTERRA	95174525
MURIEDAS	94853806
ANTIGUA	94750308
LENA	94225061
RINCONADA (LA)	93698469
ARTEIXO	93312477

FUENTE-OBEJUNA	90560770
SANTA LUCIA DE TIRAJANA	89778853
SAN FERNANDO DE HENARES	89629482
CORBERA DE LLOBREGAT	88157793
IGUALADA	86468888
TORRALBA DE CALATRAVA	85231073
CASTAÑEDA	84923282
RIVAS-VACIAMADRID	84384745
VILLACARRILLO	83160570
ES MERCADAL	82982298
LOS CORRALES DE BUELNA	82736579
ICOD DE LOS VINOS	82617984
PENAGOS	82457293
MERUELO	82285189
IBARRA	81661728
CIUDAD REAL	80841500
MANLLEU	80087377
SANTAS MARTAS	79670179
GALAPAGOS	78177437
SADA	77292106
VERA	75962638
ESPLUGUES DE LLOBREGAT	75001851
CULLEREDO	74772273
CALANDA	74548953
SANTA MARGARIDA DE MONTBUI	74526659
L'ESCALA	74034632
CARBALLO	72791373
ARANJUEZ	72667840
DONOSTIA-SAN SEBASTIAN	71784903
MONTERRUBIO DE ARMUÑA	70912104
MIGUELTURRA	70673612
SAN CRISTOBAL DE LA CUESTA	70266315
LOS REALEJOS	69738289
RUBAYO	69650098
CALVIA	69605550
SAN ILDEFONSO	67417865
SEGOVIA	67410267
LUCENA	67169233
LEIOA	66257756
S.SEBASTIAN DE LA GOMERA	66228514
OLEIROS	65736383
HERRERIAS	65153797
CILLORIGO DE LIEBANA	64869003
CUARTE DE HUERVA	64048576
RUILOBA	63411179

ABRERA	63298101
VILLARRUBIA DE LOS OJOS	62680747
SA POBLA	60519348
ALLER	59365959
ODENA	58956713
ALMENDRALEJO	58250778
JAVEA/XABIA	57782983
CONIL DE LA FRONTERA	57314611
TACORONTE	56919929
L'ESCALA	56706834
TERUEL	56063895
AJO	55840494
SAN VICENTE DE TORANZO	55632202
VALDEOLMOS-ALALPARDO	55615601
CASTELLOLI	55486501
MELGAR DE FERNAMENTAL	55455054
GUADIX	55398250
GOMECELLO	55236241
MEJORADA DEL CAMPO	54153655
CARRION DE LOS CONDES	51859539
SANT CUGAT DEL VALLES	51683806
GUADALUPE	51562487
RODA DE ANDALUCIA (LA)	51401243
PEÑARRUBIA	50510818
SOCUELLAMOS	49909450
SAN CRISTOBAL DE BOEDO	49428192
LA LLOSA	48971972
LUENA	47990633
GIRONA	47587970
HUETOR TAJAR	47289997
ARREDONDO	46937486
SALT	46839942
ALHAURIN DE LA TORRE	45627314
LUGO	45233886
ALMANSA	45124955
PAJARA	44903433
CABANES	44518765
ESPORLES	44478631
PIERA	44397129
CESPEDOSA DE TORMES	43934516
TOMELLOSO	43324382
SOBA	43030635
TOMARES	42874950
VILOBI D'ONYAR	42520014
ANOETA	41934015

VALDECILLA	41782991
PESAGUERO	41181766
GRADEFES	41055745
VILLADIEGO	40881698
ROSES	39681316
LLUCMAJOR	39360428
LOS SANTOS	39339460
MATARO	39302612
TORRE DEL CAMPO	39010672
ARTA	38474186
PUEBLA DE GUZMAN	38021226
ZAFRA	37588060
CALLDETENES	37202407
TORRE BAJA	36915340
LANTADILLA	36735642
LAGUNA DE NEGRILLOS	36651482
VILLAESCUSA	36337054
ALCUDIA	35720100
ARGANDA DEL REY	35204942
MURCIA	35027218
LA ALMUNIA	35009323
BENICARLO	34361488
VICAR	34102303
SASAMON	34028016
OSORNO LA MAYOR	33992341
LINARES	33895632
SIGUENZA	33824632
SOTRESGUDO	33613569
CHOZAS DE ABAJO	33353786
CUBILLAS DE RUEDA	33273901
ALCAÑIZ	33086216
ALCAZAR DE SAN JUAN	33027312
MATAMOROSA	32972195
PEDRERA	32844081
PEDROSA DE LA VEGA	32710073
CALZADA DE LOS MOLINOS	32447872
TORREJON DE ARDOZ	32350281
FIGUERES	31823954
BANYALBUFAR	31428461
VITORIA-GASTEIZ	31421741
LA BAÑEZA	31380053
VILLANUEVA DE GALLEGO	31229655
LEBRIJA	30939387
UGENA	30655770
RUESGA	30509595

MOLLERUSSA	30470936
GUIA DE ISORA	29955432
SANT ANTONI DE PORTMANY	29768239
ROIZ	29742225
ALAMEDA DE LA SAGRA	29644605
UBEDA	29608550
SANTA MARINA DEL REY	28945934
SITGES	28513411
SESEÑA	28255342
TORREVIEJA	28045109
BUSTILLO DE LA VEGA	28010906
SAN MIGUEL DE MERUELO	27997128
CALATAYUD	27594751
ALBURQUERQUE	27529050
LA VELLES	27486081
CALACEITE	27366342
ELEXALDE	27324676
ALCANAR	26557131
LA CODOSERA	26332419
TARRIO	26251037
GIBRALEON	26089689
MONTGAT	26067210
SOLORZANO	25859495
BASCONES DE OJEDA	25788311
CAMBRE	25552490
AITONA	25372590
MURO	25187168
VALVERDE DEL CAMINO	24840511
VILLASARRACINO	24483892
LAMASON	23939366
INCA	23922560
SANTILLANA	23684563
CIEZA	23367866
IZNAJAR	22807857
LAS PALMAS DE GRAN CANARIA	22580386
CANDELARIA	22309953
TARREGA	22274199
VILLAMERIEL	22158495
BOROX	22088036
MOLINS DE REI	22034665
MANSILLA MAYOR	21962510
EL TANQUE	21848077
TIVISSA	21806489
TAMA	21733660
CEA	21636257

MIERES DEL CAMIN	21559441
SALOU	21406659
TOPAS	21329542
AGUILON	21315833
LA LASTRILLA	21291345
NIJAR	21074715
VILANOVA D'ALCOLEA	21030703
POIO	21019587
MANRESA	21017435
VILLALBILLA	20734641
EL ROSARIO	20679233
CASTELLON LA PLANA/CASTELLO PL	20620641
ALQUERIAS DEL NIÑO PERDIDO	19656978
SANTA MARIA DE CAYON	19651698
TINEO	19525157
CORIA DEL RIO	19410107
BARBADAS	19341239
SANTA EUGENIA DE BERGA	19338836
GARLITOS	18689365
PALAFOLLS	18622338
GARCIA	18121701
ELCHE	18058993
MORISCOS	17870923
CABEZON DE LIEBANA	17783516
PILOÑA	17422050
ESCALANTE	17368267
VALLE DE VALDEBEZANA	16892879
CABEZAS DE SAN JUAN (LAS)	16741598
CAMBRILS	16713172
SANTA MARIA DEL PARAMO	16636134
BENASAU	16153111
PURULLENA	16131943
SAN ILDEFONSO O LA GRANJA	16106671
VALLE DE ABDALAJIS	16093458
CRETAS	16067392
MONDA	15831735
RENEDO DE LA VEGA	15550135
EL ESPINAR	15407893
ESPINOSA DE VILLAGONZALO	15116570
OSORNO	14838632
VILLALUENGA DE LA VEGA	14775559
RIVERO	14717104
VILCHES	14598303
BETXI	14563685
O CARBALLIÑO	14331924

CAMPO DE CRIPTANA	14230611
NOGAL DE LAS HUERTAS	14169310
VEGUILLA	14086603
UMBRETE	13856695
MIÑO	13389077
LA CABEZA DE BEJAR	13310961
LOGROÑO	13292121
SANT FELIU DE LLOBREGAT	13290217
SON SERVERA	13278977
SETLA	13221715
LA SERNA	13149374
ALLARIZ	13124361
SANT JULIA DE VILATORTA	12969149
VERGER (EL)	12703361
CHINCHON	12675808
MURO DEL ALCOY	12432623
BENILLOBA	12382337
BEAS	12199474
ALFOZ DE SANTA GADEA	12163970
VILLANUEVA DE LA CONCEPCION	12083321
MONCOFA	11907028
YAIZA	11901809
MORA D'EBRE	11856717
BOAL	11674354
YELES	11671449
PUENTENANSA	11668675
SANLUCAR DE BARRAMEDA	11597954
GARRUCHA	11558571
CERVELLO	11546713
UTEBO	11478958
SELVA	11389645
POMALUENGO	11339379
LEON	11275068
LA TALA	11249571
MOCLIN	11120281
ESPINO DE LA ORBADA	10986823
FELANITX	10982655
TRESVISO	10806935
SAN CRISTOVO DE CEA	10726825
SANLUCAR LA MAYOR	10693184
PEDROSA DEL PARAMO	10625946
AGUADULCE	10549703
BUBION	10517345
AMER	10474113
XUNQUEIRA DE AMBIA	10285841

SANT MARTI SARROCA	10087648
SADABA	9927825
PLASENCIA	9544444
GALDAR	9508055
MURO DE ALCOY	9501371
BAIDES	9470402
SANT MARTI DE TOUS	9456952
VILLARTA DE SAN JUAN	9433696
LEPE	9387117
CARMONA	9338467
TEROR	9283147
FUENTE EL FRESNO	9118115
MARUGAN	9101752
EL VERGER	8931626
MIERA	8850509
LLORET DE MAR	8818545
BENALUA	8790900
CABRERA D'ANOIA	8789292
DALIAS	8770208
UZTARROZ	8746143
VALDEVIMBRE	8617839
ROPERUELOS DEL PARAMO	8540669
LOS CORRALES	8443079
CALAMONTE	8236936
PESQUERA	8112901
BENQUERENCIA DE LA SERENA	8068492
ALARO	8024697
EJEA	7891539
RIAÑO	7878881
BARRIO DE BRICIA	7799385
VALDE-UCIEZA	7751700
VILLAREJO DE ORBIGO	7704729
MASIDE	7655444
MONFERO	7654192
BORRIANA/BURRIANA	7594448
MONACHIL	7535192
VILLAHERREROS	7506948
TARIFA	7454226
ARIJA	7300839
SANTA COLOMA DE FARNERS	7208911
ZARZOSA DE RIO PISUERGA	7198451
OLIVA	7177576
ONDARROA	7159578
VILAOBISPO DE OTERO	7142438
CALAHORRA DE BOEDO	7041722

ALICANTE/ALACANT	7015045
ELCHE/ELX	6860468
LOECHES	6838529
SANTA MARTA DE TORMES	6797224
CARTAMA	6781278
SANTA SUSANNA	6773370
ALCUBIERRE	6744122
CARTAYA	6742011
CABANAQUINTA/CABAÑAQUINTA	6585008
CALZADA DE VALDUNCIEL	6523356
VILLANUEVA DE LAS MANZANAS	6502622
LA GUANCHA	6494920
SALOBREÑA	6449419
VILLASILA DE VALDAVIA	6421579
COLLAZOS DE BOEDO	6401773
HIENDELAENCINA	6389877
TARADELL	6386544
TEGUESTE	6328019
ASTORGA	6202904
PARAMO DE BOEDO	6177166
TARDAGUILA	6116497
ARAHAL	6094332
TORRENT	6045058
CASAS ALTAS	6032320
DAGANZO DE ARRIBA	6027525
HOSPITAL DE ORBIGO	5972523
CHILCHES/XILXES	5952792
ESTEPA	5871824
VILLAVERDE DE GUADALIMAR	5775699
SAN PEDRO DE BREÑA ALTA	5773388
GINESTAR	5764067
CAÑAMERO	5758638
VILABLAREIX	5673392
EL PUERTO DE SANTA MARIA	5658477
DESTRIANA	5589568
TEGUISE	5583745
LA TAHA	5482153
SANTA CRUZ DE TENERIFE	5400750
EL PINAR	5310460
VILLAYUSO	5291650
RAMALES DE LA VICTORIA	5251561
FUENTE TOJAR	5199142
RONCAL	5155228
ORERA	5153224

HORNILLOS DEL CAMINO	5137965
ALJARAQUE	5125467
VILLABUENA DEL PUENTE	5072236
ROBLEDO DE CORPES	5023793
SANTA EUFEMIA	5012855
ESQUIVIAS	4987224
HERNIALDE	4962725
MANCHA REAL	4938258
CONDEMIOS DE ABAJO	4863141
RIOTORTO	4839058
MONTMANEU	4831534
LA POLA	4809160
ALCOBENDAS	4793950
MARIA DE HUERVA	4786835
LA VALL D'UIXO	4764786
MAHORA	4729991
VIVEIRO	4642942
PALENCIA DE NEGRILLA	4614676
PINA DE EBRO	4599066
EL CUBILLO DE UCEDA	4579055
CAMPO REAL	4562799
VILLAPROVEDO	4558552
ILLORA	4521004
TORRES DE LA ALAMEDA	4480731
MIRAVET	4427139
PEDRO MUÑOZ	4402988
MOSTOLES	4391079
GRIJALBA	4384088
LLANSA	4380969
EL CUERVO DE SEVILLA	4370369
TOCINA	4365082
BERCIANOS DEL REAL CAMINO	4361666
REOCIN	4186881
BARRAX	4165880
L'ALFAS DEL PI	4104230
MARRATXI	4091482
CARBONERAS	4089936
REVILLA DE COLLAZOS	4055137
SAN MIGUEL DE LUENA	4038276
TIAS	3997971
MORON DE LA FRONTERA	3976581
VALDELCUBO	3957537
PACS DEL PENEDES	3864774
COLMENAR DE OREJA	3856889
ALBOLOTE	3849796

VALLDEMOSSA	3799012
CAÑIZAL	3783262
MAELLA	3770211
ENCINILLAS	3707105
VILLASANDINO	3691532
ALFORJA	3646407
EL PRAT DE LLOBREGAT	3639956
LA IGLESIA	3622037
VILLAFRANCA DE LOS BARROS	3568958
VELILLA DE SAN ANTONIO	3552414
RUENTE	3543088
TORREMEJIA	3487643
IZNATORAF	3473572
TORRE ENDOMENECH	3437157
LA HUERCE	3425720
SAN JUAN DE AZNALFARACHE	3424474
SOTO DE LA VEGA	3414233
CONSELL	3403120
CARRANQUE	3365939
CASTILBLANCO DE LOS ARROYOS	3353978
POLACIONES	3331879
GUIJUELO	3305346
SAN JOSE	3296278
VEGAS DE MATUTE	3246804
BUENAVISTA DE VALDAVIA	3199602
CALAÑAS	3179031
IZNATE	3156293
BENIARRES	3127897
VILLANUEVA DEL RIO Y MINAS	3124771
ALGARROBO	3104995
CASTROJERIZ	3085291
JORBA	3059827
ARMILLA	3015954
ROSSELL	3001470
SANTA MARIA DEL CAMI	2928555
BERGONDO	2925585
BUSTARES	2883758
LOS SILOS	2883387
ALCORCON	2874459
LES PLANES D'HOSTOLES	2803026
LA CONCHA	2725005
JIRUEQUE	2701473
PALMA DE MALLORCA	2688598
EL GOLFO	2685956
CASAS BAJAS	2660986

VALDERREDIBLE	2624283
LEGANES	2622526
ELS HOSTALET DE PIEROLA	2585560
SAN CRISTOBAL DE SEGOVIA	2566322
LES MASIES DE VOLTREGA	2514975
ACEUCHAL	2502580
CENDEJAS DE LA TORRE	2468753
PALAFRUGELL	2401685
ECIJA	2391426
ALMAZORA	2359339
URDIALES DEL PARAMO	2356309
LA PALMA DEL CONDADO	2345567
TINAJO	2334415
TUDELA	2294761
NAVIA	2237616
YUNCOS	2232989
ARGANDA	2220756
LA VECILLA	2194259
TALAVAN	2182697
FUENTE VAQUEROS	2167525
JADRAQUE	2149698
MEDRANDA	2126423
PRIEGO DE CORDOBA	2099010
RONDA	2084591
TORREBLANCA	2059842
PIZARRAL	2056012
HORCAJO MEDIANERO	2052597
AGÜMES	2025637
POSADA DE VALDEON	2013150
BURON	2000679
ARENILLAS DE RIOPISUERGA	1998879
OSORNILLO	1987028
HITA	1975953
SANTIAGO DE COMPOSTELA	1961000
MONTEJO	1939836
MARAÑA	1932000
LA MUELA	1915854
ZUERA	1885596
ALCAUDETE	1853959
LA TORRE D'EN DOMENEC	1850490
SANTA CRUZ DE BOEDO	1816390
VALDEOLMOS	1797096
SANT FELIU DE PALLEROLS	1778672
VELEZ BENAUDALLA	1751119
GUARDAMAR DEL SEGURA	1740767

SAN LEONARDO DE YAGÜE	1567589
BUENAVISTA DEL NORTE	1555622
PORTUGOS	1554443
FOGARS DE LA SELVA	1547640
BELORADO	1526955
SOTOBANADO Y PRIORATO	1525498
PARACUELLOS DE JARAMA	1523177
REZMONDO	1488643
ARGAMASILLA DE ALBA	1484672
VILLANUÑO DE VALDAVIA	1459840
LASARTE-ORIA	1442806
GUILLENA	1426453
MOGUER	1413995
AYORA	1403001
ARRONIZ	1398235
VILAFRANCA DE BONANY	1396957
PECHINA	1392961
PILES	1372664
ALDEATEJADA	1368826
NAVALVILLAR DE IBOR	1312271
ALDEASECA DE ALBA	1291619
MURCHANTE	1241845
BENICASIM	1234418
LA GRANJA D'ESCARP	1215218
ACEBEDO	1200911
BEARIZ	1182813
MENGIBAR	1161748
LEITZA	1155897
POBLADURA DEL VALLE	1127563
GARROBO (EL)	1078570
ALBOCASSER	1055526
SABIÑANIGO	1034571
BAEZA	1025359
CUADROS	1020544
ALCALA LA REAL	1018110
FUSTIÑANA	1008073
SIGÜENZA	984869
PALENCIA	961497
SAN MAMES DE CAMPOS	934577
PUENTE DEL CONGOSTO	932420
XIRIVELLA	932276
PALMACES DE JADRAQUE	915874
ALBENDIEGO	910742
BERIAIN	870202

LOMA DE UCIEZA	855012
LUESIA	844366
FOZ	839392
MOLLET DEL VALLES	833752
ESTREMERÁ	793888
NUMANCIA DE LA SAGRA	790950
VALDEPIELAGO	781422
LA TOBA	775414
CASTRILLO DE VILLAVEGA	728578
MIRALRIO	716212
MAZALEON	699617
VILLAR DEL OLMO	658781
CALAFELL	653591
MIJAS	653555
CALDES D'ESTRAC	653020
CABRERA D'ANOIA	631013
GALVE DE SORBE	567867
HARO	522246
CASTELLANOS DE VILLIQUERA	505670
MARCILLA DE CAMPOS	496061
TARODA	451268
AÑOVER DE TAJO	425687
CENDEJAS DE ENMEDIO	384524
ADEJE CASCO	373980
FUENTES DE BEJAR	367955
MALGRAT DE MAR	351407
SAN AMARO	326162
ABIA DE LAS TORRES	321654
VILLARRABE	287466
CENIZATE	282892
BROTO	249873
TAFALLA	220584
VALDEPRADO DEL RIO	216876
SANTA GADEA DEL CID	200517
FONELAS	187464
CIHUELA	166351
SARRION	153161
VILLAZANZO DE VALDERADUEY	146444
LUESMA	119061
CASTRILLO DE RIOPISUERGA	117782
SANT MATEU	109671
TORREDEMBARRA	98267
CIUDAD RODRIGO	93434
CAPILEIRA	69248
ITERO DEL CASTILLO	39186

HUERMECES DEL CERRO	29000
YUNQUERA	15713
LAS ROZAS DE MADRID	6883
VILLAJOSYOSA/LA VILA JOIOSA	2181
VALSECA	2136
SANT FELIU DE GUIXOLS	715
RONQUILLO (EL)	125
ALFOZ DE BRICIA	0
HERRERA	0
BERGA	0
ANDORRA	0

Tabla F.2.: Consumo total anual de todos los municipios