Universidad
Carlos III de Madrid

DATA SCIENCE EN EMPRESA

(Aplicación de técnicas de Inteligencia de Negocio y minería de datos sobre datos abiertos de consumo eléctrico)

TRABAJO FIN DE GRADO

Grado en Ingeniería en Tecnologías Industriales

ALUMNO: Adrián Fernández Hernández
TUTOR: Germán Gutiérrez Sánchez

ESCUELA POLITÉCNICA SUPERIOR
UNIV. CARLOS III DE MADRID

Septiembre´16

Resumen.

El análisis y manejo de grandes cantidades de datos juega un rol esencial en la empresa actual, tanto para la mejora de objetivos como para apoyar la toma de decisiones estratégicas. Un elevado conocimiento del dato permite mejorar la eficiencia del servicio, lo que se traduce en un mejor aprovechamiento de los recursos y aún más importante, en la reducción de costes. Para extraer este conocimiento, las organizaciones hacen uso de un conjunto de técnicas conocidas como Inteligencia de Negocio, o *Business Intelligence* en inglés, que permiten transformar los datos de la empresa en información de valor para mejorar su competitividad y comprender las necesidades de los usuarios.

Para la aplicación de estas técnicas, esta ciencia dispone de herramientas específicas diseñadas para asistir al análisis de la información y la presentación de los datos. La gran mayoría de herramientas requieren de licencia comercial, por lo que en el presente trabajo se ha decidido aplicar un modelo de Inteligencia de Negocio a un conjunto de datos abiertos mediante el uso de herramientas de software libre, con el fin de ofrecer una alternativa gratuita a un área de gran interés en el mercado actual.

Cabe aclarar que el trabajo no consiste en el diseño de la arquitectura de la herramienta, sino en la aplicación de algunas de las técnicas que ofrecen los productos comerciales existentes para Inteligencia de Negocio, pero implementadas desde software gratuito.

Como base de datos, se va a utilizar un archivo público con medidas relacionadas con el consumo eléctrico de un hogar almacenadas a lo largo de casi cuatro años, con el fin de estudiar el comportamiento del usuario y extraer información relevante que permita interpretar cómo ha sido el consumo y responder a preguntas tales como: cuáles son los dispositivos que más energía consumen o en qué momentos del día se produce mayor gasto energético.

Para ello se procederá a la generación de representaciones a través de una herramienta de visualización de datos conocida como Tableau, con el propósito de estudiar las variables más representativas a lo largo del período estudiado, y así establecer qué periodos tienen un comportamiento similar y cuales se desvían de un comportamiento normal. Esta herramienta permite conectarse a una base de datos y producir un gran número de soluciones interactivas enfocadas a aspectos de Inteligencia de Negocio de gran belleza y composición.

Dado que el análisis visual no ofrece suficiente conocimiento acerca de las características de las series temporales, además del análisis exploratorio y debido a que existen suficientes medidas, se va a proceder al análisis estadístico de series temporales para la aplicación de distintos modelos predictivos partiendo de la premisa de que cuanto más sencillo sea, mejor, por lo que se estudiará cómo de sencillo puede ser nuestro modelo para que haga una correcta predicción.

Debido a que existen diferentes posibilidades, en este proyecto, se plantean dos técnicas (modelos VAR y ARMA) en las que usaremos las series temporales al minuto y por hora, aprovechando, como posible solución para agrupar las horas, la utilización de técnicas de clúster. Nuestro objetivo será, por tanto, discutir si la mayor sofisticación mejora o no la

predicción con modelos de series temporales estándar. Para esta parte haremos uso del lenguaje de programación R y su entorno de programación RConsole y RStudio.

De esta manera, se realizará una aplicación de caso de Inteligencia de Negocio para realizar hallazgos interesantes sobre un servicio de interés general, como es el caso del consumo eléctrico, mediante la consecución de etapas y técnicas propias de la minería de datos y de la estadística.

Al trabajar con datos reales, se espera que la solución propuesta pueda servir de inspiración para cualquier empresa o grupo de investigación que maneje datos de consumo del mismo tipo para elaborar propuestas de eficiencia energética o diferentes planes de reducción de gastos para clientes, ya que se trata de una ciencia que ha tenido un auge importante en los últimos años y cuya aplicación presenta aún dudas e incertidumbres.

Palabras clave: análisis de datos, minería de datos, análisis exploratorio, Inteligencia de Negocio, series temporales, predicción, R, Tableau.

Abstract.

The analysis and management of large amounts of data plays an essential role in the current company, both for the improvement of objectives such as to support the strategic decision-making. A large knowledge of the data makes it possible to improve the efficiency of the service, which translates into a better use of resources and even more important, the reduction of costs.

For the application of these techniques, Business Intelligence has designed specific tools to assist the data analysis and presentation. The vast majority of tools require commercial license, so because of that in the present work has been decided to implement a model of Business Intelligence to a set of open data through the use of the open-source software, with the purpose of offering a free alternative to a high demand area of in the current market.

It must be clarified that the work does not consist in the tool architectural design, but in the implementation of the different techniques that use to be offered by commercial existing products in the Business Intelligence from free software choices.

To perform this study, we'll make use of a public file as database source filled with measures related to the household electricity consumption stored along almost four years, with the purpose of studying the behavior of the user, in this case of the family that lives in the house, and extract relevant information for interpreting their consumption to respond to such questions as: How much electricity do their devices really consume ? or at what time during the day is required higher energy demands.

To achieve this goal, we will proceed to generate different kinds of representations through a data-visualization tool known as Tableau. This tool allows us to connect to the database and produce a large number of interactive solutions focused on aspects of Business Intelligence of great beauty and composition, with the purpose of studying the most representative variables throughout the studied period, and establish what periods have a similar behavior and which divert from a normal behavior.

Given that the visual analysis does not offer sufficient knowledge about the characteristics of the time series, in addition to the exploratory analysis and because of there are enough number of measures, we will do a statistical analysis for building a forecasting model based on VAR and ARMA predictive models with the premise that the simpler the better, so we will explore how simple it can be our model to predict well. To reach this, we will group the time series model per minute, per hour and also seize a possible solution making use of clustering techniques for grouping the hours. Therefore, our main goal here will be to discuss whether the increased sophistication improves or not the prediction with standard time series models.

For this part we use the R programming language and its programming environment R Console and RStudio.

In this way we will have done a Business Intelligence case by reporting findings of a service of general interest as it is the electrical consumption, and by following stages and exploratory and statistical techniques own of Data Mining.

When working with real data, it is expected that the proposed solution might serve as inspiration for any company or investigation that handles consumption data of the same type for any kind of proposals in energy efficiency field since it's a science that has increased its interest in recent years and whose implementation still presents doubts and uncertainties.

Key-words: data analysis, Data Mining, exploratory analysis, Business Intelligence, time series, forecasting, R, Tableau.

Agradecimientos.

Primero agradecer a mis padres, por su apoyo incondicional en todo momento durante la realización del proyecto, sobre todo a mi madre que ha sido la que más me ha tenido que soportar en casa, día tras día, y la que me ha alimentado, física y espiritualmente, aportándome la energía que necesitaba para seguir trabajando.

En segundo lugar quiero agradecer a la Universidad Carlos III y a todos sus profesores la enseñanza impartida durante estos años de carrera, así como la oportunidad que me dieron de salir durante un año a realizar estudios en la Universidad de Estonia, lo que me permitió crecer, tanto académica como personalmente, consiguiendo que volviera una persona muy diferente, madura y mucho más segura de sí misma.

Agradecer también a mis compañeros de piso Gianmarco y Sebastián durante la estancia Erasmus, nunca olvidaré los momentos que compartí con ellos y que me hicieron sentir más cerca de casa. También a mis compañeros de la Universidad con los que realicé el intercambio: Almudena, María, Guille, Monro y Marco por supuesto.

Agradecer a mi amigo Iago, Diego, Jesús, Edu y a las chicas por todas las horas que hemos compartido en Leganés desde el primer día de Universidad.

Agradecer a la Universidad de California por disponer de sus datos en formato abierto y por fomentar la investigación y la libre utilización de las bases de datos:

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Agradecer especialmente la ayuda prestada en el último año por el profesor Carlos Santiuste, ya que sus tutorías y su manera de enfocar las asignaturas me ayudaron a superar las últimas asignaturas del Grado, las cuales se me habían resistido en primeras convocatorias. Y por supuesto una mención aparte para el tutor de este proyecto final de grado, Don Germán Gutiérrez Sánchez por su dedicación, apoyo e interés desde el primer día a la hora de aconsejarme y resolverme las dudas y obstáculos que han surgido durante su realización.

Índice general.

Índice de figuras y tablas.

---

1. Introducción.

1.1. Motivación.

Vivimos en la sociedad de la información. Gracias a Internet y a los avances tecnológicos, el mundo se ha convertido en un lugar digitalizado en el que cualquier persona puede acceder a información desde cualquier parte del planeta con sólo hacer uso de su teléfono móvil. La generación masiva de datos se produce de manera continua y a volúmenes cada vez mayores, ya que a medida que avanza la tecnología, la recopilación de datos y su almacenamiento se está convirtiendo en una tarea fácil y de bajo coste por parte de las empresas. A su vez, la aparición de redes sociales como Twitter, Facebook o LinkedIn, en las cuales cada usuario debe disponer de una cuenta personal con su correspondiente base de datos, ha supuesto una revolución para este movimiento, ya que en sus plataformas se comparte información en todo tipo de formatos, ya sean links de noticias, imágenes, videos, audios, etc.

Para poder hacer frente a esta nueva era del dato y a la información que conlleva, las organizaciones han tenido que adoptar la flexibilización como estrategia, con el objetivo de adecuarse a un mercado de creciente internacionalización e influenciado por el flujo y la cantidad de información que dispone de sus clientes.

Hoy en día, los directivos de las empresas pueden acceder a mucha más información que hace unas décadas, de más calidad y con mayor facilidad. Sin embargo, el tiempo de que disponen para acceder a esa información es cada vez menor. Esto convierte la toma rápida de decisiones en un auténtico reto, ya que los directivos se ven obligados a identificar y analizar sólo la información importante que permita conocer más allá de los datos que éstos manejan y descubrir conocimiento, para así guiar la toma de decisiones hacia objetivos que supongan una fuerte ventaja competitiva.

Debido a ello se ha decidido afrontar un problema de aplicación de diferentes técnicas propias de Inteligencia de Negocio sobre un software gratuito, con la motivación principal de aprender nuevas herramientas y métodos de análisis, así como para descubrir, por experiencia propia, los obstáculos y beneficios de esta nueva ciencia.

1.2. Presentación del problema.

A continuación se presenta la justificación y descripción del problema, i.e. dominio de aplicación de los conceptos, estrategias y técnicas que se aplican en este trabajo.

Uno de los sectores de la industria que más dudas genera en la población es el eléctrico, con un 46,6% de usuarios que asegura que el recibo de la luz es poco comprensible o incomprensible de acuerdo con la última encuesta publicada por la Comisión Nacional de los Mercados y la Competencia (CNMC)[1]. A su vez, los precios de la electricidad han subido un 33% en la UE desde 2008, situando a España como el cuarto país europeo con el precio

más caro de la electricidad, solo por detrás de Dinamarca, Alemania e Irlanda [2]. Este encarecimiento está vinculado a costes ajenos al suministro eléctrico (generación y distribución de luz) y son cargados al recibo por parte del Gobierno para penalizar entre otros factores, el posible derroche de energía.

La electricidad es un recurso necesario y sin el cual no podríamos llevar a cabo las tareas de nuestra vida cotidiana, por lo que para ser competentes en el mercado, las empresas energéticas se han visto en la necesidad de elaborar estrategias que promuevan el uso eficiente del consumo energético. El ahorro en el gasto de energía no sólo conlleva a una reducción del consumo, sino también permite mejorar la satisfacción con sus clientes por la reducción del importe de las facturas.

Llegados a este punto, surge la siguiente pregunta:

> ¿De qué elemento dispone principalmente la industria energética para poder desarrollar este tipo de estrategias?
>
> La respuesta a esta pregunta es: DATOS.

Las empresas energéticas disponen de todos los datos relacionados con el consumo de los edificios a los que dan servicio. Por tanto si se analizan esos datos en profundidad se podrá encontrar información de valor que permita ofrecer diversas soluciones, como por ejemplo posibles perfiles de suministro diseñados específicamente en función de las tendencias de consumo de cada cliente. Para poder hacer este tipo de análisis, las empresas utilizan técnicas pertenecientes a la Ciencia de datos: *Data Science.*

Uno de los principales dilemas del *Data Science*, es cómo extraer valor cuando se dispone de una gran cantidad de información en bruto, o mejor dicho, qué acciones llevar a cabo en esa información (preprocesamiento, análisis, etc.) para transformarla en valor y generar conocimiento.

En lo que respecta a las técnicas de *Data Science*, una de las cuestiones en las que ahonda este trabajo es sobre el grado de "complicación" técnica que ha de tener un modelo predictivo. Entendemos por complicación todo lo relativo a la elección del método, agregación de los datos, y tipo de relación de dichos datos entre ellos. Este tema no es baladí, puesto que autores seminales como Box y Jenkins (1970), Clements y Hendry (1998) y recientemente Gree y Amstrong (2015) debaten sobre las ventajas de modelos parsimoniosos y sencillos (poco parametrizados) respecto al trato de las relaciones de la serie con su propio pasado y con otras series.

En este trabajo nos centramos en dos tipos de dificultades: primeramente, el nivel de desagregación de las series. ¿Conviene predecir minuto a minuto para después agregar dicha predicción? ¿Es mejor hora a hora? Para ello, utilizaremos modelos univariantes con el correspondiente nivel de agregación. Por otro lado, nos preguntaremos si hay mejoras, agrupando las horas, pero complicando los modelos, y permitiendo estudiar la estructura de Feedback multivariante entre grupos (clúster) de horas.

Al respecto, los resultados parecen estar más cerca de la superioridad predictiva de modelos más sencillos en los tres aspectos presentados. De hecho, en este trabajo se encuentra que el modelo que podría considerarse el menos complejo (el univariante por horas) supera en capacidad predictiva a otros dos modelos con mayor complejidad (univariante por minutos y multivariante por horas).

Debido a estos factores, se ha decidido basar este trabajo **en el estudio de un consumo eléctrico sobre el que se aplicarán diversas técnicas relativas al *Data Science*** o Ciencia de datos. **La base de datos es de fuente abierta** y en ella se han almacenado medidas del consumo eléctrico de un hogar familiar anónimo a lo largo de un período de casi cuatro años. Dado que no se especifica en el repositorio fuente, se desconoce la localización del hogar a estudio.

Al tratarse de datos reales y similares a los que dispone cualquier empresa energética, se considera que las técnicas aplicadas para las distintas fases de la solución final (pre-procesado, análisis, regresión, etc.) podrían ser de interés, al disponer de pruebas relacionadas con modelos que permiten ajustar con la mayor fiabilidad posible la oferta del día después, atendiendo a la demanda pronosticada de cara a la a planificación de la red eléctrica.

1.3. Objetivos.

Los objetivos principales son:

1. Análisis exploratorio mediante gráficas interactivas para representar el conjunto de los datos y así analizar diferencias tendencias y hábitos de consumo a lo largo de diferentes niveles de granularidad (periodos de tiempo: anual, mensual, semanal y diario). Estas gráficas serán generadas mediante la herramienta interactiva de Inteligencia de Negocio Tableau (http://www.tableau.com/es-es).

2. Clasificación de perfil de consumo de la vivienda según los atributos estudiados en el análisis exploratorio. Para ello se proponen variables externas que hayan podido causar el comportamiento analizado, como por ejemplo factores temporales o periodos vacacionales por parte de la familia.

3. Análisis de series temporales de datos de corta frecuencia. Elección del grado de agregación de los datos evaluando la capacidad predictiva mediante técnicas de validación cruzada en *Data Mining*.

4. Utilización de la metodología Box-Jenkins para los modelos univariantes. Elección del orden de integración regular y estacional de las series, análisis del correlograma para la búsqueda del modelo óptimo. Implementación en R.

5. Utilización de modelos multivariantes VAR (Modelos de Vectores Autorregresivos) utilizando técnicas de componentes principales para buscar clúster de comportamiento horario común y así reducir la dimensión del problema.

6. Implementación en R de todos los modelos permitiendo analizar los errores cometidos para un conjunto de 100 días.

La consecución de los objetivos principales conlleva una serie de objetivos (de relevante interés académico y profesional para el autor de este trabajo), y entre ellos caben destacar los siguientes:

1. Comprensión de la teoría relacionada con la Inteligencia de Negocio y las diferentes soluciones que ofrece (minería de datos) y las herramientas disponibles para su aplicación.

2. Adquisición de fluidez en el manejo de algunas herramientas de software como Tableau y R.

3. Aprendizaje del lenguaje de programación R (versión GNU del lenguaje S) y su entorno de librerías y paquetes (CRAN R). Al tratarse de un lenguaje desconocido por el autor, se buscará documentación y sitios de ayuda donde poder encontrar apoyo frente a dudas que vaya surgiendo en el transcurso del trabajo.

4. Aprendizaje de técnicas de análisis y predicción de serie temporales. Búsqueda y comprensión de modelos en referencias bibliográficas y estudio de su aplicación para el conjunto de datos que manejamos.

1.4. Estructura de la memoria.

El trabajo se divide de la siguiente forma:

- En el capítulo 1 – **Introducción**: se presenta la principal motivación por la que se decide afrontar el trabajo, se define el planteamiento del problema y se comentan los objetivos que se pretenden llevar a cabo. Por último se describe el marco regulatorio a tener en consideración.

- En el capítulo 2 – **Estado del arte**: se explica qué es la Inteligencia de Negocio, sus antecedentes y características, así como los distintos niveles de solución que ofrece. También se comenta la importancia que tiene el dato como base para el proyecto y se desarrolla en profundidad la técnica de minería de datos, indicando sus etapas, escenarios y disciplinas.

  A continuación se define un concepto de gran relación con el área, como es el Big Data, donde además se expone un caso de uso real y se explican sus principales características y arquitectura. Se explica qué es la Ciencia de datos y los diferentes profesionales que presenta.

  Por último se hace una introducción acerca del lenguaje de programación utilizado para la aplicación de modelos R, con una breve definición, historia e indicación de los sitios donde nos apoyaremos para la búsqueda de código y documentación a lo largo del proyecto. También se exponen sus ventajas e inconvenientes, así como una propuesta de diferentes soluciones alternativas de programas que podríamos haber escogido. Haremos lo mismo para el otro software utilizado: Tableau con la intención de justificar la elección de estos.

  En el capítulo 3 – **Análisis exploratorio**: se procede a la descarga y descripción de los datos, indicando la fuente de los mismos y realizando su importación a Tableau para su tratamiento. Para discriminar las variables, tomaremos como interesante aquella que represente un mayor impacto en el consumo y por tanto influya en la factura eléctrica mensual, respecto de otras que simplemente resulten un ruido para el cálculo.

  Se estudian las variables más interesantes mediante la generación de visualizaciones que nos permitan responder a preguntas a medida que se generan nuevas gráficas, con el fin de descubrir cómo es el consumo promedio a lo largo de diferentes periodos (anuales, mensuales y horarios), así como el uso de los diferentes electrodomésticos.
  Para finalizar se realiza una predicción estimada del conjunto de los datos para observar como es su aspecto.

- En el capítulo 4 – **Etapa de modelos**: se explican los algoritmos para modelizar la correlación existente entre los datos de una serie temporal: modelos univariante y multivariante (ARIMA y VAR). Se especifica la metodología de identificación y estimación de los modelos sugerida por la literatura y cómo se ha implementado en R mediante la explicación de las estructuras y funciones utilizadas. Se añaden explicaciones adicionales sobre cómo se han agrupado las horas mediante el método de componentes principales (PCA).

Por último se analiza la capacidad predictiva utilizando validación cruzada y se representan los diferentes errores de predicción.

- En el capítulo 5 –**Rol de participantes, planificación y presupuesto**: se indica el rol de los participantes en el trabajo, la planificación, incluyendo las tareas críticas, y el presupuesto.

- En el capítulo 6 – **Conclusiones y líneas futuras**: se exponen las conclusiones extraídas en este trabajo, tanto las referentes a la visualización como a la etapa de modelos. También se indican conclusiones generales extraídas tras la finalización del proyecto y se enumeran una serie de trabajos futuros que sería interesante desarrollar.

- En el capítulo 7 – **Referencias**: se presentan las referencias consultadas, incluyendo enlaces web de artículos y noticias que han servido como inspiración para realizar el trabajo con solvencia y éxito.

- Al final de la memoria se puede encontrar un **Anexo** que contiene el código, el rellenado de valores faltantes y ayudas en la instalación de R y Tableau.

1.5. Marco regulador.

Antes de comenzar con el trabajo, es necesario conocer si existe alguna normativa legal que pueda limitar o prohibir la realización de nuestro estudio.

En el análisis de datos es de gran importancia conocer la legislación vigente ya que sus funciones están relacionadas con la extracción, manejo e incluso a veces publicación de resultados. Un desconocimiento de la ley podría causarnos problemas de violación de los derechos humanos y la privacidad de las personas, por ello habrá que consultar siempre la política reguladora del repositorio desde el que descarguemos los datos.

En nuestro caso, existe una política de cita en el caso de querer publicar el presente estudio. Esta refleja lo siguiente:

Política de citas del Repositorio de la Universidad de California en Irvine:

"En el caso de querer publicar material relacionado con bases de datos obtenidas desde este repositorio, se deberá mencionar en los agradecimientos del trabajo la colaboración recibida por parte del organismo. De esta manera esto podrá ayudar a otros a descargar el mismo conjuntos de datos y poder replicar sus experimentos. Como recomendación sugerimos el siguiente formato:

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science."

2. Estado del Arte.

Para desarrollar nuestro trabajo es importante enmarcar nuestra herramienta dentro del área relacionada. Para ello, es necesario hacer una breve introducción a la Inteligencia de Negocio, así como una explicación de las técnicas relacionadas con el proyecto, como son la minería de datos y sus disciplinas.

Asimismo, se van a enumerar tecnologías de gran relación con este campo, como son: la Ciencia de datos y el Big Data.

Por último realizaremos un repaso del software que se va a utilizar, así como de las diferentes alternativas existentes en el mercado como justificación de la elección para la solución final.

2.1. Introducción a la Inteligencia de Negocio.

"Es la década de los datos y de ahí vendrá la revolución"

Alex Pentland, director del programa de emprendores del 'Media Lab' del MIT

Con la aparición de Internet, la información se ha convertido en el activo más valioso para las empresas del siglo XXI. Esta información se obtiene a partir de los datos que maneja la empresa, la cual a su vez se divide en departamentos, cada uno de los cuáles lleva a cabo una actividad diferente que requiere de su propia fuente de datos. Debido a ello, surge la necesidad de integrar todos los sistemas de información de la empresa de manera eficiente para que trabajen de forma coordinada, y así la información pueda ser actualizada y compartida a las diferentes áreas funcionales y usuarios de la organización.

Este proceso se conoce como Inteligencia de Negocio (en inglés *Business Intelligence*), y trata de ofrecer soluciones mediante el análisis de los datos, a partir de los cuáles obtendremos información de valor que nos guie en la búsqueda de ideas y conocimiento. Las soluciones ayudarán a la empresa a conocer mejor tanto su ventaja competitiva como sus oportunidades a la hora de diseñar estrategias de negocio de éxito.

Figura 1. Ilustración del Business Intelligence

Pese a que el término haya aparecido con mayor fuerza en los último años, la Inteligencia de Negocio no es algo nuevo y si hacemos un repaso de su historia, podemos comprobar que se trata de un proceso que fue introducido por primera vez en el año 1958, año en el que el investigador Hans Peter Luhn hizo referencia a la Inteligencia de Negocio como: "La capacidad de comprender las interrelaciones de los hechos presentados en tal forma como para orientar la acción hacia una meta deseada".

Durante los años 60 y 70, fueron apareciendo los primeros software dedicados al tratamiento de las bases de datos. Estos sistemas permitían el almacenamiento de datos, pero a medida que la información aumentaba, el acceso a ella se hacía cada vez más lento y complejo. Hasta entonces la información había sido almacenada de forma manual y en papel.

En los años 80, Ralph Kimball y Bill Inmon introdujeron el concepto de *datawarehouse* o almacén de datos, junto con varios sistemas de reporte de datos. A pesar de la mejora en los sistemas de bases de datos, el número de herramientas que permitieran su explotación y manejo era muy pobre, además de necesitar de un personal altamente cualificado para la administración de dichas bases de datos.

En 1989, Howar Dresner vuelve a utilizar el término de Inteligencia de Negocio como: “Conceptos y métodos para mejorar la toma de decisiones en las organizaciones mediante el uso de sistemas basados en hechos de apoyo”. A finales de 1990, la centralización y organización de las bases de datos supuso un aumento de interés por parte de las empresas, dando lugar a la proliferación de aplicaciones y plataformas de Inteligencia de Negocio. No fue hasta el año 2000 cuando se produjo la verdadera expansión del *Business Intelligence*, gracias a la consolidación de herramientas de gran potencia, como Oracle, SAS, SAP e IBM.

Como resumen de sus principales características podríamos enumerar:

1. **Se trata de un proceso interactivo**. Un proceso de Inteligencia de Negocio debe ser continuado en el tiempo, ya que una de las vías para obtener conocimiento es el estudio de posibles patrones, cambios y tendencias en los datos, y para ello se requiere de una base de datos que contenga información almacenada durante un periodo suficiente de tiempo con la que poder hacer comparaciones y extraer conclusiones.

2. **No es una tecnología**. En algunos casos se hace referencia a la Inteligencia de Negocio como una tecnología, pero en realidad se trata de una estrategia empresarial basada en un conjunto de herramientas de tratamiento de datos.

3. **Accesible**. La accesibilidad de los sistemas debe estar garantizada por parte de todos los usuarios.

4. **Intuitiva**. Una herramienta de Inteligencia de Negocio útil, será aquella que ofrezca un manejo sencillo e intuitivo con el que poder manipular los datos sin necesidad de conocimientos científicos ni técnicos.

5. **Integrada y multidimensional**. La arquitectura de los sistemas de Inteligencia de Negocio es muy compleja, ya que sus funciones se extienden desde la extracción de los datos hasta su reporte, pasando por la aplicación de modelos y técnicas de análisis. Esto obliga a que la integración de los diferentes sistemas sea clave para que éstos trabajen de forma coordinada. Además la información debe ser recolectada desde todo tipo de fuentes. Por ejemplo un pronóstico de ventas de un nuevo producto en una región, requiere del estudio previo del historial de ventas en la zona para conocer el protocolo de actuación y el perfil del consumidor en esa determinada región.

6. **Completa**. La Inteligencia de Negocio requiere de técnicas relacionadas con la estadística y las matemáticas, por lo que sus herramientas de análisis deben ofrecer una amplia variedad para responder a problemas de predicción, clasificación, segmentación, etc.

En la Inteligencia de Negocio existen distintos niveles de complejidad de acuerdo a las soluciones que se pretendan llevar a cabo:

- **Informes:**
  - Informes predefinidos: son fijos y en ellos siempre se muestran parámetros predeterminados.
  - Informes a medida: se trata de informes específicos, en los que el usuario define los datos que desea tratar en función de su objetivo.

  - Query (consulta) / Cubos OLAP (On-Line Analytic Processing): es una búsqueda o pedido de datos sobre una base de datos en tiempo real. También se puede referir a alguna acción llevada a cabo sobre la base de datos, como actualizaciones o eliminaciones de información.

- **Análisis**:
  - Análisis estadístico: técnicas relacionada con la estadística para completar el análisis de los datos.
  - Pronósticos: predicción de resultados en base a datos ya estudiados.
  - Minería de datos: estudio de los datos para descubrir patrones y comportamientos con el objetivo de definir un perfil general.
  - Optimización.
  - Minería de procesos: permite el descubrimiento de procesos a partir del registro de eventos almacenados en el sistema.

De todas las soluciones descritas, se va a hacer especial hincapié en la minería de datos puesto que supone una de las bases de nuestro trabajo y se trata de una técnica que se alimenta de una gran diversidad de ciencias.

2.2. ¿Qué es la minería de datos?

2.2.1 La importancia del dato.

¿Cuál es la finalidad de generar información? Por lo general existen muchos motivos por los que interesa disponer de información, a partir de la cual examinar, investigar, organizar, optimizar o predecir factores que nos den una explicación a los hechos ocurridos y por tanto un conocimiento.

Las empresas son muy conscientes de la importancia que tiene la información. Más allá de los factores financieros y la gestión de capitales, el conocimiento supone una herramienta de supervivencia para toda empresa, ya que permite comprender el mercado y sus necesidades para poder enfrentarse a los desafíos del mercado, y mejorar la toma de decisiones para posicionarse en el mercado y mantener una ventaja competitiva.

Por tanto, para llegar a este conocimiento se tiene que analizar información, y para obtener esa información hay que estudiar datos. De forma general, los datos son la materia prima sin tratar. En el momento en que se le atribuye algún significado, el dato se convierte automáticamente en información. Si a partir de esa información, se encuentra un modelo que añade valor a esa información, ya podríamos hablar de conocimiento.

El siguiente ciclo representa un modelo jerárquico entre los tres términos: datos, información y conocimiento, y pretende reflejar por un lado, el valor que tiene cada concepto en la toma de decisiones, como por otro lado, el volumen que presenta cada estado:

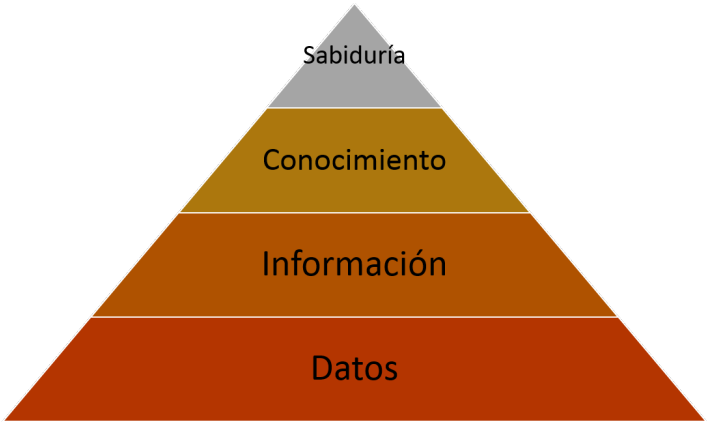


Figura 2. Jerarquía del conocimiento

Como se puede observar en la imagen, lo que interesa a las empresas es alcanzar la zona superior de la pirámide. Para poder ascender hasta el vértice, es necesario contar con

tecnologías que permitan explotar la base de la pirámide, ya que la unión entre dato e información no es igual de estrecha que la unión entre conocimiento e información.

Aquí es donde aparece el concepto de *Data Mining*.

2.2.2 Definición y escenarios.

La minería de datos o *Data Mining* como se conoce en inglés, es el proceso de análisis de la información desde diferentes perspectivas y con el principal propósito de detectar patrones, comportamientos, tendencias o grupos relevantes con los que generar algún modelo que nos permita conocer mejor los datos que manejamos, y así ayudar a comprender el contenido de la base de datos.

Las relaciones entre estos patrones no se pueden detectar mediante la exploración tradicional de los datos, bien por la complejidad de los mismos o porque hay demasiados datos. Por ello, para llevar a cabo esta cuestión, la minería de datos utiliza modelos científicos y sistemas que van más allá de los simples cálculos matemáticos y estadísticos que otros software de análisis podrían generarnos, motivo por el que sus aplicaciones abarcan cada vez escenarios más variados y orientados a todo tipo de negocio, como por ejemplo:

- **Segmentación de clientes**: análisis del perfil de consumo e historial de búsquedas de los clientes para hacer diferentes clasificaciones basadas en afinidades y orientadas a productos específicos, campañas publicitarias, aumento de ventas orienta a diferentes grupos de edades, etc.

- **Recomendación y sugerencia de productos o eventos**: análisis de búsqueda y consultas realizadas en Internet para conocer los intereses y gustos de la persona y así recomendar productos del mismo tipo o relacionados con su uso.

- **Evaluación del riesgo de fraude**: apoyo en la elaboración de programas antifraude con los que prevenir, identificar y remediar posibles alteraciones deliberadas de información que supongan un uso fraudulento de activos e ingresos.

- **Administración de recursos**: estudio de los hábitos de compra de los clientes para evitar posibles faltas de stock en compras masivas o para encontrar la mejor localización en las estanterías de un supermercado.

- **Detección de anomalías en los pacientes**: identificación del nivel de enfermedad en pacientes de un hospital, por ejemplo en la detección de tumores, pudiendo clasificar estos en malignos o benignos en función de su forma y tamaño**.**

Esta variedad de escenarios hace que los modelos de minería de datos supongan un auténtico desafío, tanto por la complejidad de comprensión y aplicación de las técnicas que requiere, como por la correcta elección de los mismos, ya que por lo general, cuando trabajemos con datos en una empresa, no vamos a disponer de todo el tiempo que necesitemos para probar modelos, sino que tendremos que saber con cierta seguridad qué modelo va a funcionar. Además las técnicas dependen del tipo de datos y de los factores que se quieren analizar, por lo que no existe una única vía para resolver un problema de minería de datos.

A pesar de esto, cualquier proceso comienza siempre con los datos en bruto y finaliza con la evaluación e interpretación con la que extraer conocimiento, por lo que es importante definir una metodología a seguir que proporcione respuestas a medida que se analizan los datos. Por lo general podríamos decir que un proceso típico de minería de datos consta de los siguientes pasos:

- **Selección del conjunto de datos**. Definir qué variables interesa llevar a estudio y comprender bien el escenario y sus conceptos.

- **Establecer objetivos**. A qué queremos responder con el estudio de esos datos.

- **Análisis exploratorio**: Etapa de exploración en la que descubrir posibles tendencias o relaciones entre las variables. Esto se puede hacer con estudios estadísticos (análisis de correlación) o simplemente mediante la visualización de las variables (análisis visual).

- **Tratamiento**: puede llegar a absorber el 90% del tiempo que se dedica al proyecto. Engloba la importación, preparación y limpieza de los datos.

- **Estudio de modelos**: selección de las técnicas estadísticas y modelos predictivos que convendría aplicar en función de los objetivos deseados.

- **Generación de modelos**: se trata de la etapa más técnica y delicada, ya que en ella se llevan a cabo las operaciones sobre los datos, mediante la programación de código (en lenguajes como R o Python) y la estadística (modelos de predicción: ARMA, VAR, modelos clúster, etc.)

- **Exposición de resultados**: se comparan los modelos aplicados y se muestran los valores obtenidos.

- **Conclusiones**: difusión y comunicación de las ideas deducidas en base a los resultados.

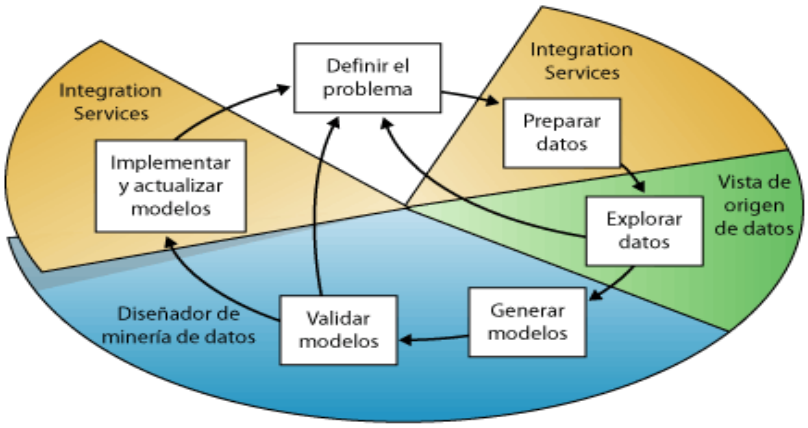


Figura 3. Ciclo común en la generación de modelos de minería de datos

Por lo tanto, debemos entender la minería de datos como un proceso compuesto por etapas que tratan de descubrir información oculta de gran relevancia para la consecución de estrategias en una gran diversidad de disciplinas y no como un gran software. Esto no quiere decir que durante su desarrollo, no sea necesario el uso de diferentes herramientas de software (como las utilizadas en este proyecto) que nos permitan realizar los cálculos estadísticos, la generación de gráficos o la propia obtención de los datos. En la minería de datos las herramientas suelen complementarse entre sí, con el fin de aprovecharse de las ventajas de cada una, y sus modelos proceden de todo tipo de áreas y disciplinas.

2.2.3 Disciplinas.

En el siguiente esquema se muestra la confluencia de disciplinas que se relacionan con el proceso de *Data Mining*:

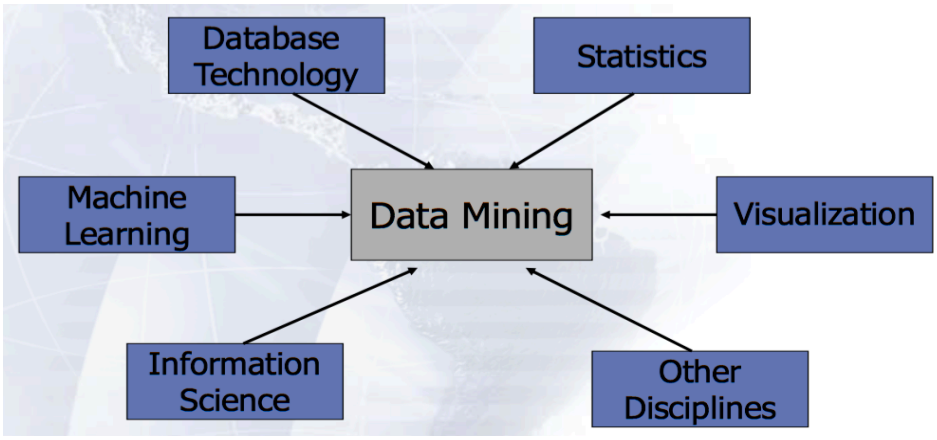


Figura 4. Confluencia de disciplinas en minería de datos

Bases de datos

La minería de datos puede aplicarse sobre cualquier tipo de dato: datos espaciales y temporales, series de tiempo, datos multimedia, textos, web, transacciones, etc.

Desde un punto de vista general, las técnicas de minería de datos más habituales distinguen entre dos tipos:

- **Datos categóricos**: (registran categorías o cualidades). Ejemplo: Positivo "P"
- **Datos numéricos** (registran conteos o numeraciones). Número de mediciones: 30

Estos datos a su vez se almacenan y organizan en bancos de información conocidos como bases de datos, las cuales pueden ser de diferente categoría dependiendo del tipo de información que contengan.

Entre las bases de datos más conocidas merecen una mención especial dos tipos:

- **Relacionales** (SQL): son las más utilizadas hoy en día, se componen de tablas que a su vez constan de un conjunto de campos (columnas) y registros (filas). Para manipular la información debe seguirse un esquema asociado a cada registro, en el que los datos siguen una estructura y son por lo tanto, estructurados (Structured Query Language).

  Destacan MySQL, PostgreSQL y Oracle.

- **No relacionales** (NoSQL): han aparecido con la llegada de Facebook, Twitter o Youtube y se diferencian de las relaciones en que este tipo no cumple con el esquema entidad-relación. Tampoco utilizan una estructura en forma de tabla, sino que para el almacenamiento hacen uso de otros formatos, como mapeo de grafos o formatos clave-valor.

  Su procesamiento supone un verdadero desafío para los analista y científicos de datos.

El siguiente esquema muestra la categoría en la que se incluyen algunas tecnologías:

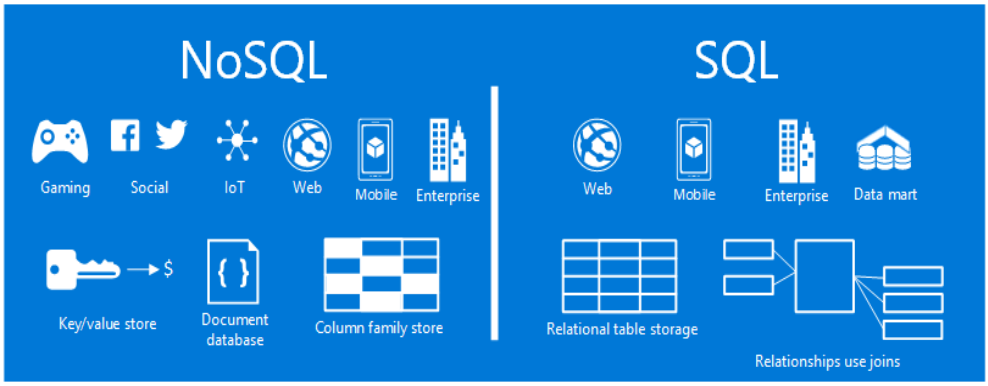


Figura 5. Clasificación de tecnologías según su base de datos requerida

Modelos

En la práctica, cuando disponemos de muchos datos, la minería de datos requiere de determinados modelos y técnicas para poder extraer información. Estos modelos por lo general provienen de campos como la estadística y la inteligencia artificial, y su elección dependerá de las necesidades y circunstancias de cada conjunto de datos o del tipo de datos.

Según el objetivo de su análisis, los modelos utilizados se clasifican en supervisados y no supervisados (Weiss y Indurkhya, 1998):

- **Predictivos o supervisados**: son capaces de construir un modelo o método por medio del entrenamiento de un conjunto de datos para poder predecir una variable de interés, una vez los datos son entrenados.

  Por ejemplo, si tuviéramos figuras de tres formas geométricas diferentes: círculo, triángulo y cuadrado, le diríamos al modelo:

  - Esto es un cuadrado 
  - Esto es un círculo 
  - Esto es un triángulo 

  Por tanto y una vez ha sido entrenado y supervisado el modelo, si le preguntamos, ¿qué figura es?  Su respuesta será círculo, porque ya ha aprendido de los datos anteriores para predecir una nueva figura. De ahí que también se conozcan como modelos de aprendizaje supervisado.

  Por tanto el objetivo de un modelo predictivo se basa en la búsqueda de normas de clasificación o de predicción en base a lo ocurrido en la práctica, para conocer el resultado de una muestra (variable dependiente) a partir de un conjunto de variables relacionadas con ella (variable independientes).

- **Descriptivos o no supervisados**: identifican patrones para interpretar y clasificar los datos. Por ejemplo, un modelo descriptivo puede utilizarse para segmentar clientes de un mercado en diferentes grupos de edad, que relacionen a estos por sus preferencias e intereses comunes y así descubrir vínculos desconocidos de antemano. Su objetivo no es predecir resultados, sino cuantificar las relaciones entre los datos, para observar comportamientos similares y así agruparlos de manera rápida.

A su vez, en cada uno de estos grupos, existen modelos que incluyen una larga lista de de algoritmos, siendo algunos ejemplos:

- **Regresión lineal (predictivo)**: se utiliza generalmente cuando tratamos dos escenarios, por ejemplo: altura y peso, nos suele interesar conocer y cuantificar la relación existente entre las dos variables. Puede ser simple (regresión simple) o múltiple (regresión múltiple) en función del número de variables estudiadas, y su resultado reproduce una recta de ajuste sobre todos los datos (normalmente dibujados en forma de puntos mediante un gráfico de dispersión) que hace mínima las distancias entre la recta y cada punto.

- **Árboles de decisión (predictivo)**: mecanismo basado en la construcción de un esquema lógico, muy similar a los modelos basados en reglas, que permite categorizar los datos para elegir el camino más adecuado. Se trata de un proceso de ramificación, de ahí que se conozca como árbol, en el que se busca generar diferentes alternativas que permitan ir poco a poco evaluando y diseccionando el conjunto de datos, con el propósito de evaluar la efectividad de las decisiones tomadas y así saber cómo de buena o mala es la solución final. Esto es de gran ayuda en las empresas para evitar malas experiencias en la toma de decisiones.

Aprendizaje automático (Machine learning y Deep learning)

A medida que disponemos de más datos, más ambicioso resulta la aplicación de nuevas técnicas que puedan abordarlos. En la actualidad, los investigadores consideran que para avanzar hacia el progreso es necesaria la aplicación de aspectos relacionados con la inteligencia artificial a nivel humano. Para ello se requiere de un conjunto de técnicas, conocidas como machine learning o aprendizaje automático, que van más allá de fundamentos matemáticos y estadísticos y con las que se pretende conseguir que las máquinas sean capaces de tener comportamientos y sentimientos propios del ser humano.

Para que una máquina pueda alcanzar tal nivel de desarrollo, gigantes tecnológicos como Amazon o Google llevan años apostando por el desarrollo de algoritmos capaces de llevar a cabo tareas como la conducción de vehículos sin conductor o el reconocimiento de voz e imágenes. Hoy en día, el uso del aprendizaje automático está tan generalizado que es muy común hacer uso de él sin ser conscientes de ello, y su crecimiento es tan grande, que muchas técnicas de machine learning se empiezan a considerar limitadas, dando lugar a la aparición de un nuevo área dentro del aprendizaje automático, conocido como Deep learning.

El Deep learning o aprendizaje profundo, surge según los expertos a raíz de dos razones:

- La reducción de costes de las tecnologías necesarias para investigar Inteligencia Artificial.
- La necesidad de entrenar a la máquina para que pueda representar información a partir de datos en bruto de manera automática.

Algunos resultados prometedores de su aplicación son el procesamiento del lenguaje natural para análisis de sentimiento y traducción de textos, el reconocimiento de rostro, texto y objetos en imágenes o la reconstrucción de circuitos del cerebro humano.

A continuación se muestra un test de procesamiento de imágenes para el que se han "traducido" una serie de imágenes en texto. Para ello se han utilizado Redes Neuronales Recurrentes (conocidas con las siglas RNN en inglés) tomando como entrada adicional, la

visualización de una muestra de imagen mediante el uso de Redes Neuronales de Circunvolución (conocidas con las siglas CNN en inglés):

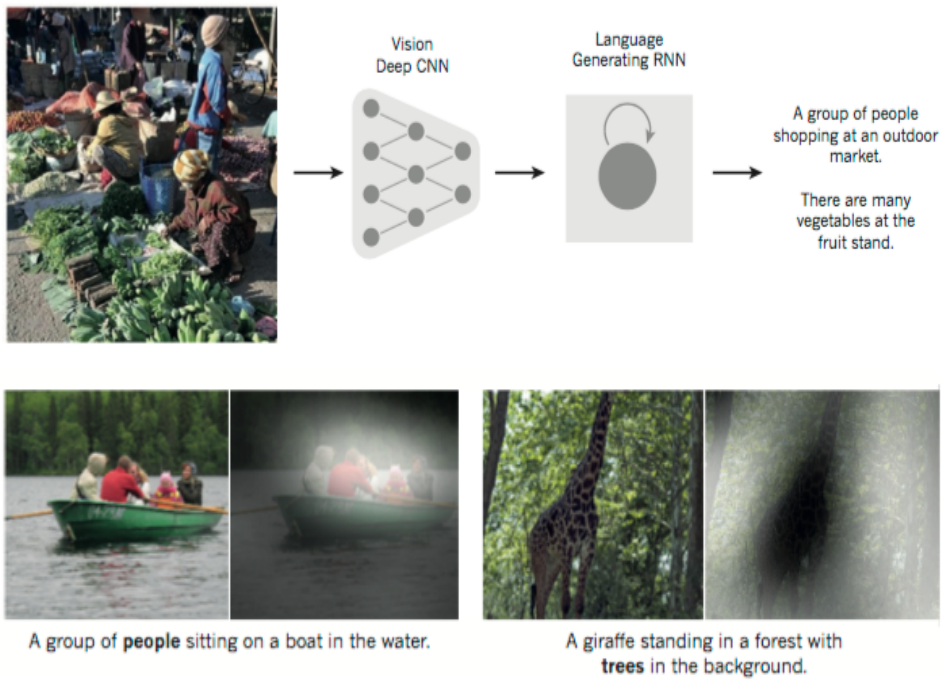


Figura 6. Aplicación de Deep Learning para el reconocimiento de imágenes

Según se explica en la investigación [4] la Red Neuronal Recurrente (RNN) es capaz de enfocar su atención en diferentes localizaciones de la imagen (zonas iluminadas) y generar una palabra (en negrita) que describa el objeto capturado.

2.2.4 Tendencias.

La minería de datos ha evolucionado en los últimos años debido a cambios en la tecnología, en los hábitos de compra de los usuarios, en el sector del marketing, etc.

Se van a enumerar recientes tendencias que han cobrado gran importancia para las organizaciones públicas y privadas de todo tipo:

- **Criminología**: Rastreo de pistas en investigaciones delictivas y criminales.
- **Comercio**: Predicción de tendencias de compra y recomendación de productos en línea en función de las búsquedas del cliente.
- **Marketing**: Análisis de impacto de campañas publicitarias y medición de satisfacción del cliente.
- **Marketing digital**: Publicidad de productos que han sido buscados recientemente.

- **Medicina**: investigaciones en la medicina para descubrir marcadores biológicos que aporten nuevos descubrimientos en el tratamiento de enfermedades.

- **Telecomunicaciones**: las compañías telefónicas pueden identificar llamadas inusuales que puedan estar relacionadas con actividades fraudulentas.

- **Investigación**: los científicos pueden aplicar más rápidamente sus hipótesis accediendo a los datos recopilados en experimentos anteriores.

Para poder llevar a cabo estas tareas, la minería de datos se enfrenta a problemas como:

- **Los tiempos de respuesta**: el procesamiento de grandes volúmenes de datos implica modelos que requieren de largos tiempos de procesamiento. Esto es un inconveniente para procesos que requieran respuesta en tiempo real.

- **La importancia de los datos no estructurados**: la información no estructurada potencia la capacidad de las empresas para obtener un mejor entendimiento y descubrir información en mayor profundidad de los conjuntos de datos. Tomando como ejemplo una empresa de atención al cliente, con el análisis de datos estructurados se podría conocer aspectos como: categorías donde se encuentra una queja, número de quejas, evaluación de servicio al cliente, rapidez en la atención a problemas, etc. Todos estos datos son de gran utilidad, pero no permiten dar respuesta a preguntas tales como:

    - ¿ Cuál ha sido la raíz del problema ?
    - ¿ Ha hablado el cliente con la persona más adecuada ?
    - ¿ Se podría haber resuelto la incidencia con mayor rapidez ?

Por tanto, aunque los datos estructurados seguirán siendo el pilar de todo análisis, la era de las redes sociales va a hacer que las organizaciones se vean en la necesidad de incorporar información no estructurada.

- **La integración de algoritmos en aplicaciones y páginas web**: para poder competir en el mercado, las aplicaciones deben de ofrecer el mayor número de funcionalidades posibles. Para esto, los desarrolladores tienen el desafío de integrar algoritmos cada vez más complejos con los que ofrecer un mejor servicio.

2.3. Conceptos relacionados.

2.3.1 Big Data.

La evaluación masiva de datos implica nuevos paradigmas y tecnologías que respondan a la necesidad de almacenamiento, procesamiento y análisis de los datos en escala masiva, en tiempo real y de la manera más automatizada posible.

El 90% de los datos almacenados en el mundo se generaron en los últimos 2 años, debido a factores como el crecimiento exponencial de los dispositivos conectados a Internet, el uso diario de redes sociales, los volúmenes de datos generados por sensores y sistemas de geolocalización o el número de transacciones llevadas a cabo por los sistemas de negocio. Esta revolución ha dado lugar a la aparición de una nueva tecnología capaz de capturar, gestionar y procesar en un tiempo razonable y de forma eficaz estos datos: el Big Data. También podemos encontrar el uso de este concepto para referirse a conjuntos de datos de gran volumen.

Como muestra del alcance que tiene la tecnología Big Data, se va a mostrar un ejemplo real que marcó un punto de inflexión en el mundo del análisis de datos:

- **La reelección de Obama**: el presidente de los Estados Unidos, Barack Obama, decidió aplicar Big Data en sus segundas elecciones en el año 2012. Para ello, un equipo de 100 personas se dedicaron a analizar datos relacionados con las elecciones a tres niveles:
    - Registro de datos de votantes convencidos.
    - Persuasión dirigida a votantes dudosos.
    - Certeza de que los partidarios tuvieran intención de ejercer el voto.

  El motor de trabajo fue una plataforma inteligente conocida como HP Vertica, entre cuyas acciones más destacadas estaban la realización de encuestas a pie de calle a partir de las cuales realizar un feedback rápido por parte del equipo o la detección de las franjas horarias en las que tendrían mayor impacto televisivo los anuncios publicitarios de campaña. Estas acciones permitieron segmentar el perfil de los votantes para así evitar el uso de tiempo, dinero y recursos en votantes que no tuvieran intención de ejercer el voto a su partido.

  Cuatro años después en España, el Partido Popular contrató los servicios de la misma compañía para intentar el voto táctico en las segundas elecciones a la presidencia de Gobierno organizadas en junio de 2016, y así realizar un esfuerzo por aumentar el número de escaños. En este caso el objetivo se centró en Facebook como herramienta publicitaria para localizar audiencias mediante el estudio de los temas y mensajes que más les interesaran.

Arquitectura Big Data

Los procesos de Big Data requieren de una compleja arquitectura que hace unos años era impensable desarrollar, debido al elevado presupuesto que suponían y a su baja fiabilidad.

Por lo general un sistema Big Data debe cumplir los siguientes requisitos:

- **Tolerancia a fallos**: partiendo de la premisa de que puede fallar o averiarse alguno de los nodos, el sistema debe seguir funcionando con normalidad.

- **Escalabilidad lineal**: aumento de la capacidad de procesamiento de un sistema para reaccionar y adaptarse al continuo crecimiento de la red sin que la calidad de su uso disminuya.

- **Modelos de despliegue de datos**: creación de servidores virtuales o Cloud para optimizar costes sin prescindir de escalabilidad y flexibilidad.

- **Almacenamiento local**: consiste en que los procesos analíticos se realicen lo más cercano posible del lugar de almacenamiento de los datos para así evitar traslados innecesarios de información.

- **Distribución**: evitar el almacenamiento centralizado mediante la distribución de las ejecuciones de los procesos.

Como principales componentes de la arquitectura Big Data cabe destacar:

- **Bases de datos**: [veáse "Bases de datos " apartado 2.2.3]

- **Apache Hadoop**: es un software de licencia libre, programado en Java e inspirado en el modelo de programación utilizado por Google conocido como MapReduce. Permite a las aplicaciones realizar tareas con grandes volúmenes de datos en paralelo mediante la división de los conjuntos de datos en clústeres que permiten distribuir bloques de datos a través sus diferentes nodos. Esto se conoce como escalabilidad, y sin ella sería imposible manejar las montañas de información que generan actualmente las empresas.

- **Complementos de Hadoop**: Spark, Hive, Pig o Impala, proporcionan a Hadoop un entorno para realizar trabajos de computación, gestión (lectura/escritura), almacenamiento de datos, etc.

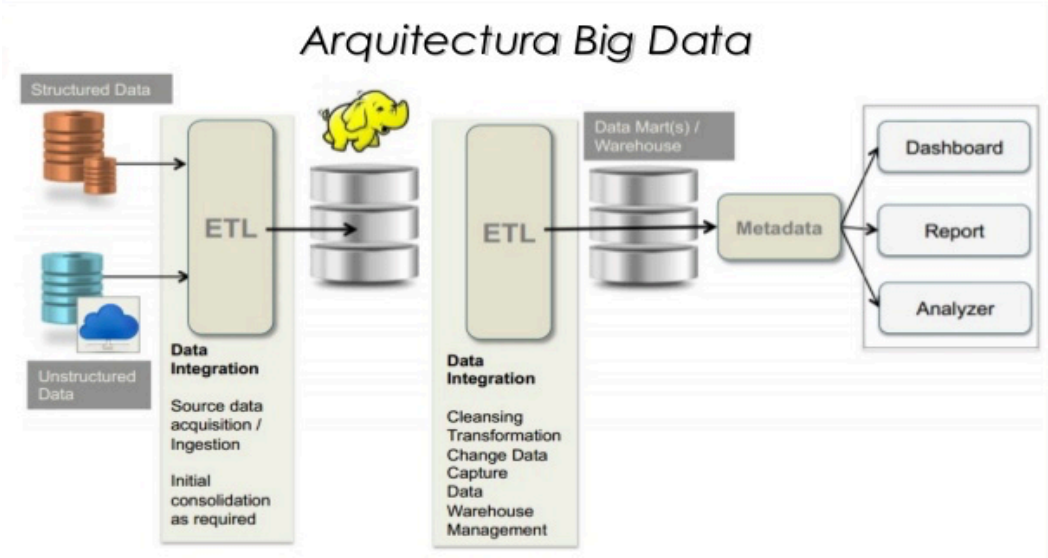


Figura 7. Componentes de la arquitectura de la tecnología Big Data

En la imagen se puede observar como opera la arquitectura Big Data. En primer lugar, la información es integrada en Hadoop (elefante amarillo), para posteriormente y una vez procesada, integrarla en las bases de datos de la compañía. A partir de este punto la información ya está disponible para ser analizada, visualizada o reportada a los diferentes departamentos.

Características del Big Data: Las cinco Vs

Desde el punto de vista académico, el concepto Big Data se define en torno a una serie de características conocidas como las 5 V, en relación a la primera inicial y al número de dimensiones que lo definen. Estas son:

- **Volumen**: probablemente la característica que mejor define la tecnología Big Data. Las estimaciones de generación de datos se han visto sobrepasadas por el movimiento de datos entre usuarios de móvil y redes sociales, lo que ha provocado una transformación en los sistemas de almacenamiento, pasando del terabyte (TB) al petabyte (1000 TB) y zetabyte (10^9 TB) de información. Esto implica una mayor complejidad en el análisis y almacenamiento de los datos para las tecnologías que soporten el Big Data.

- **Velocidad**: la rapidez con la que se generan los datos ha sufrido también un aumento considerable, obligando a los sistemas de procesamiento de datos a ofrecer una respuesta adecuada que permita procesar y analizar la información antes de que pueda ser obsoleta. Para ello el Big Data utiliza herramientas que permiten trabajar en paralelo y en tiempo real, lo que permite al analista de datos optimizar los resultados e identificar datos en función de su ciclo de vida.

- **Variedad**: como ya hemos comentado, en el análisis de datos es esencial poder trabajar con una gran diversidad de tipos de datos, así como de sus fuentes. Así se podrán procesar datos estructurados, semiestructurados o no estructurados desde fuentes como páginas web, tweets, sensores, videos, clips de audio, etc. Esta variedad determina la riqueza que conlleva el Big Data.

- **Veracidad**: se trata de la dimensión más importante para el analista, ya que se define como el grado de confianza aplicado sobre los datos a analizar y que por lo tanto determinará la calidad de los resultados y la autenticidad de los mismos.

- **Valor**: A medida que aumenta el volumen y la velocidad de generación de los datos, su valor disminuye, debido a la dificultad que supone la explotación de los mismos. Obtener valor a partir de los datos sigue siendo el principal objetivo para mejorar la toma de decisiones, tanto de la Inteligencia de Negocio como del Big Data. En otras palabras, la tecnología Big Data busca facilitar la explotación de los datos para que su valor marginal sea el mayor posible y de esta manera hacer que las empresas que usen estas tecnologías puedan adelantarse a la competencia.

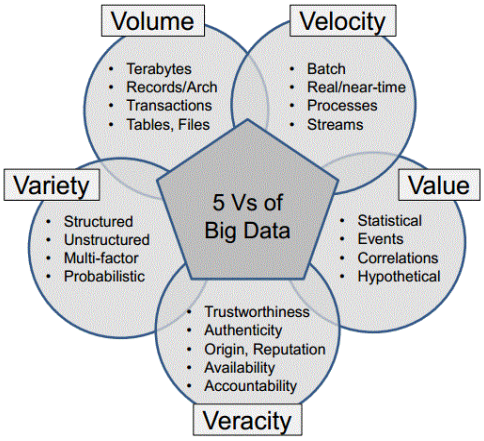


Figura 8. Pentágono de las cinco V's que definen la tecnología Big Data

Estas cinco dimensiones intentan reflejar que el objetivo final del Big Data no es sólo el de recopilar, combinar o procesar todos los datos, sino también aumentar su valor y eficacia. Esto significa que debemos evolucionar el Big Data a un análisis inteligente (*Smart Data*) ya que la eficacia de las empresas depende de la calidad de los datos.

2.3.2 Ciencia de datos.

Es un campo de continuación a los estudios relacionados con el procesamiento y la extracción de información como son la minería de datos y la analítica predictiva. Es muy habitual referirse al *Business Intelligence* como Ciencia de datos, de ahí que en este proyecto se

mencionen ambos conceptos para referirnos en general al análisis de datos para la obtención de conocimiento.

Cabe aclarar que pese a tratarse de una ciencia con un gran número de paradigmas por resolver, los expertos ya han comenzado a establecer diferencias entre ambos términos, siendo algunas de ellas:

Ciencia de datos	**Inteligencia de Negocio**
Trabaja con datos incompletos	Trabaja con datos completos
Fuentes de archivos desordenados	Fuentes de archivos limpios y ordenados
Grandes conjuntos de datos	Conjuntos de datos manejables
Los hallazgos impulsan nuevas decisiones	Sus hallazgos miden el rendimiento pasado
Algoritmos más avanzados (Deep learning)	Algoritmos más sencillos (exploración, estadística, etc.)

La necesidad de responder a este nuevo desafío ha provocado la aparición de nuevos expertos capaces de responder a las dudas y problemas que vayan apareciendo, estos son los científicos y analistas de datos.

Científico de datos vs analista de datos.

Aunque aún es pronto para distinguir diferentes roles en el sector del tratamiento de datos, existen una serie de diferencias en función de las habilidades y conocimientos que se tengan.

El siguiente diagrama de Venn muestran los aspectos en común y las divergencias de ambas definiciones:

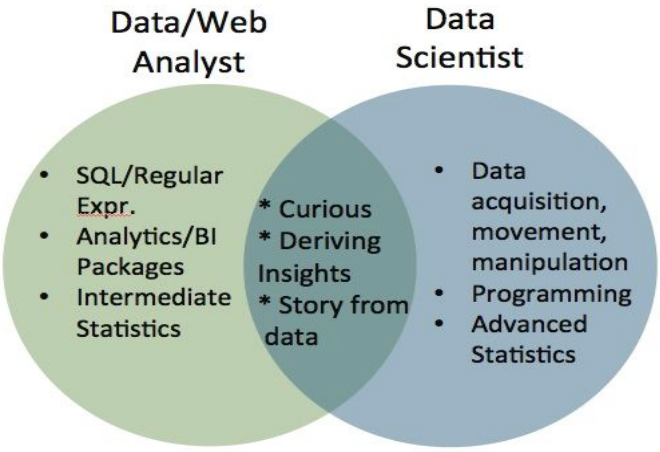


Figura 9. Diagrama de Venn sobre los dos roles de expertos en datos

Como se puede observar en la figura, el científico de datos suele tener un mejor dominio en técnicas estadísticas y conocimientos de programación que no son requeridos para el perfil de

analista. Los analistas de datos son generalmente requeridos para tareas de mantenimiento y gestión de las bases de datos, en general SQL.

Ambos perfiles deben ser capaz de comunicar resultados a proveedores y departamentos que puedan no tener conocimientos en la materia, así como poseer una gran capacidad de deducción y curiosidad académica, con la que examinar a los conjuntos de datos y fuentes para averiguar modelos que resuelvan los diferentes problemas de la empresa.

2.4. Herramientas de software

2.4.1 R

Definición: R (CRAN R) es un lenguaje y conjunto de módulos estadísticos que, mediante cualquiera de los interfaces de que dispone, permite realizar análisis de datos y generación de gráficos de los mismos. Además, cuenta con gran variedad de operadores para realizar un cálculo efectivo sobre variables indexadas, en particular, sobre vectores y matrices. Muy parecido a Matlab y Octave, y con sintaxis que recuerda a C/C++. Es software libre, donde el término software libre se refiere a la libertad de los usuarios para ejecutar, copiar, distribuir, estudiar, cambiar y mejorar el software. Están disponibles versiones de R para Windows de Microsoft, Unix, Linux y MacOS.

Figura 10. Logo de CRAN R

Historia

En 1976, la mayor parte la mayor parte del análisis y computación estadísticos se realizaba por medio de subrutinas en Fortran, algo que era bastante tedioso. Por eso, John Chambers, Rick Becker y Allan Wilks, pertenecientes a *Bell Laboratories* de AT&T, desarrollaron el lenguaje 'S', en referencia a "Statistical" o estadístico en castellano, el cual desarrollaron como un conjunto de bibliotecas de macros Fortran, que se convirtieron en su entorno de análisis estadístico interno y que salieron en 1979 por primera vez como un producto distribuible. El lenguaje era orientado a objetos e interpretado, por lo que permitía al usuario "interactuar" con la línea de comandos.

Mientras S se reescribía y evolucionaba en nuevas versiones, en 1992 los profesores Ross Ihaka y Robert Gentleman, del Departamento de Estadística de la Universidad de Auckland en Nueva Zelanda, decidieron implementar su propio entorno de programación, al que comienzan a llamar "R". La idea surgió durante una conversación en el pasillo, en la que ambos comentaban sobre la necesidad de encontrar un software, más adecuado y sencillo para sus estudiantes, los cuales no tenían apenas conocimientos en computación, con el que pudieran aplicar las estadísticas enseñadas para analizar datos y obtener algunas representaciones. En sus propias palabras:

"El resultado se llamó R en parte al reconocimiento de la influencia de S y en parte para hacer gala de sus propios logros. Buscamos que los usuarios puedan iniciar en un entorno interactivo, en el que no se vean, conscientemente, a ellos mismos como programadores."

En el año 1995 Martin Mächler, de la Escuela Politécnica Federal de Zúrich, convence a Ross y Robert a usar la Licencia GNU para hacer de R un software libre, y fue en febrero del año 2000, cuando se consideró lo suficientemente completo como para lanzar la primer versión, versión 1.0.

Pese a ser considerado muy limitado al principio, R inmediatamente ganó seguidores al tratarse de una herramienta de software libre, permitiendo a estadísticos, ingenieros y científicos poder mejorar el código del software o escribir variaciones para tareas específicas.

¿ Dónde encontrar documentación acerca de R ?

Documentación general

- Introducción a R (lectura recomendada): http://cran.es.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf
- Manuales de R: http://cran.ms.unimelb.edu.au/manuals.html#R-admin
- Librerías recomendadas entre investigadores relativos a diferentes campos de investigación: https://cran.r-project.org/web/views/
- Tutoriales en Youtube y ayudas en Internet (continua actualización)

En la web http://cran.es.r-project.org/other-docs.html se encuentra disponible una lista de documentos acerca de R que incluye recursos en castellano.

Además R cuenta con una amplia y dinámica comunidad de usuarios que facilita la consulta ante cualquier tipo de duda con sitios como:

Stack Overflow: http://stackoverflow.com/

Es un foro de dudas para programadores y profesionales que funciona con la fórmula de pregunta-respuesta. Permite tanto buscar la pregunta como formularla en el tablón para que el resto puedan ayudar a resolverla.

Github: https://github.com/

Se trata de una plataforma de desarrollo colaborativo para desarrolladores y emprendedores que permite desde almacenar proyectos personales como códigos o gráficos generados hasta la posibilidad de trabajar conjuntamente en proyectos y tecnologías de consulta abierta. Para ello cada usuario dispone de una cuenta personal

con la que poder compartir los trabajos con el fin de servir de inspiración para otros profesionales o guardarlos como portfolio privado.

Kaggle: https://www.kaggle.com/

Kaggle es una plataforma online que ofrece a los usuarios la posibilidad de participar en distintas competiciones cuyo principal tema es el análisis de datos. En algunos retos, las personas que obtengan los mejores resultados pueden recibir una remuneración económica en función de los resultados y la complejidad de los datos a tratar. Kaggle, hasta el momento ha creado más de 200 desafíos con más de 1,2 millones de dólares en premios.

Dado el nivel actual de los participantes, las grandes compañías prestan atención a las competiciones y los ganadores suelen entrar en las agendas de los cazadores de talento consiguiendo ofertas de trabajo. Al mismo tiempo es un sitio de gran ayuda para consultar código en lenguaje R y Python, ya que una vez finaliza la competición, las soluciones de cada usuario son publicadas de forma abierta.

R-Bloggers: https://www.r-bloggers.com

Blog muy similar al resto pero además de dudas, contiene noticias y artículos de actualidad que lo convierten en una de las mejores opciones para estar al tanto de las últimas novedades acerca del lenguaje. También presenta tutoriales de aprendizaje bien estructurados y enlazados con la plataforma de aprendizaje online Datacamp.

¿Por qué R se ha convertido en una opción tan interesante?

A continuación se enumeran algunas de las ventajas que han llevado a su elección:

1. *Lenguaje robusto*

   R es hoy día probablemente el entorno más usado por las universidades para investigaciones en estadística, lo cual ha garantizado su robustez. Pese a tener una curva de aprendizaje complejo, su lenguaje de programación está bien desarrollado y admite condicionales, bucles, funciones y posibilidad de entradas y salidas, lo que hace que la implementación sea efectiva y variada, así como la posibilidad de hacer uso reiterado de funciones existentes en sus librerías.

2. *Flexible*

   La salida que proporciona cualquier función se puede manipular a conveniencia y ver su implementación, ya que R guarda los resultados como objetos. Esto permite decidir, de toda la información que genera la ejecución de una función, qué es lo que realmente se desea mostrar. Además R puede trabajar con datos procedentes de todo tipo de archivos: ya sean formato texto (.txt) archivos con valores separados por comas (CSV) o un archivo Excel…

3. *Constante actualización y amplia literatura disponible*

   Al tratarse de uno de los softwares más utilizados para el análisis de datos, la comunidad de usuarios permanece activa y en continuo crecimiento, siendo cada vez más sencillo encontrar ayuda en Internet.

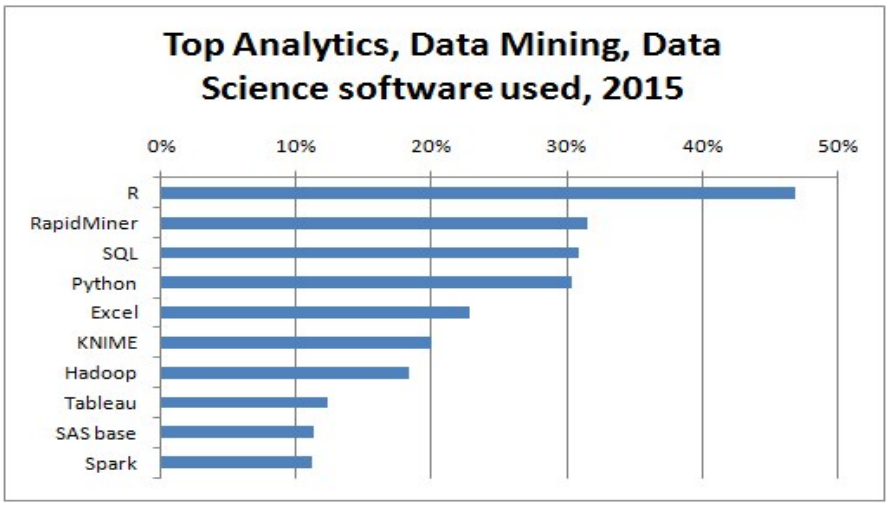


Figura 11. Herramientas más utilizadas en minería de datos en el año 2015

4. *Amplias facilidades de manipulación de bases de datos*

   Con frecuencia, la manipulación de los datos es igual o más laboriosa que el análisis subsiguiente. La facilidad de R para realizar una importación de datos desde una base externa, permite tratar ficheros de gran volumen y conectarlos en la misma o en otra máquina.

5. *Gratuita*

   R es Open Source (multiplataforma, libre, abierto, etc.)

6. *Visualizaciones*

   Aunque no posee la misma potencia gráfica que otros programas como Tableau, permite visualizar información compleja de una forma sencilla y que de esta manera resulte más sencillo comprenderla.

Pese a su extendido uso, R también presenta algunos inconvenientes:

1. R se ajusta a la RAM de la máquina ya que fue creado en el año 1996 cuando el espacio en disco duro y la memoria RAM eran mucho más costosos que hoy en día. Probablemente si fuera diseñado en la actualidad, respondería mejor a las arquitecturas modernas multi-core o CPUs múltiples.

2. A pesar de la versatilidad y funcionabilidad de R, este no dispone de un menú principal donde el usuario pueda acceder a submenús para la lectura de datos, la ejecución de procedimientos estadísticos o la generación de gráficos, sino que estas tareas se realizan mediante un lenguaje de comandos en línea. Esto es una desventaja con respecto a otros programas estadísticos ya conocidos, como son EViews, SPSS o Statgraphics, los cuales presentan un interfaz más amigable para un usuario que este comenzando a trabajar con ellos.

3. Los mensajes de error que R nos muestra, no especifican sobre los fallos que estamos realizando.

2.4.2 Alternativas al uso de R

La lista de herramientas que podríamos haber utilizado para el proyecto es muy variada, ya que como se ha comentado, la Inteligencia de Negocio y la minería de datos viven un momento de auge que ha hecho que desarrolladores de aplicaciones y de software vean en ello la oportunidad para el éxito. Aun así, resumimos algunas de las más conocidas entre los productos gratuitos o de fuente abierta:

Pentaho:

- Es un conjunto de programas libres para generar Inteligencia de Negocio y para la gestión y toma de decisiones empresariales.

  Su plataforma es capaz de ejecutar las reglas necesarias para tomar esas decisiones de negocio y de presentar la información adecuada en el momento adecuado.

- Ofrece soluciones para la preparación, gestión y análisis de la información, incluyendo análisis multidimensional OLAP. También está diseñado para la presentación de informes, minería de datos y creación de cuadros de mando para el usuario.
- Aunque es de código abierto y sus herramientas deben ser configuradas manualmente, una por una, ofrece una opción de pago en la que incluye automáticamente todas sus soluciones.
- Algunas de sus ventajas son:
  - **Últimas tecnologías**: creado bajo las plataformas libres de Java y mySQL, de gran auge y desarrollo en la actualidad.
  - **Comunidad amplia** de usuarios que permite la constante actualización del software y mejora de la aplicación mediante reportes de fallos y todo tipo de pruebas.
  - **Independiente**: se puede hacer uso independiente de cada programa, haciendo posible la integración de otros software ajenos a su desarrollo.

Weka:

- Programada en Java y muy reconocida para tareas de minería de datos y aprendizaje automático.
- Ofrece una amplia colección de herramientas de visualización y algoritmos para análisis de datos y modelado predictivo, integrados en una interfaz gráfica de fácil acceso para el usuario.
- Permite el pre-procesamiento de datos y la selección de todo tipo de características de clustering, clasificación, regresión y visualización.
- Sus técnicas se basa en la hipótesis de que los datos están disponibles en un solo archivo (fichero plano) o una relación, en la que cada registro de datos está descrito por un número fijo de atributos (normalmente numéricos o nominales). Para el aprendizaje automático, estas variables se corresponden con los atributos de entrada y salida.
- Algunas de sus ventajas son:
  - **Portabilidad**: al estar desarrollado en Java, puede ejecutarse sobre cualquier plataforma.
  - **Interfaz gráfica amigable**.
  - **Completa:** extensa colección de técnicas de pre-procesamiento y modelado de datos.

Rapid Miner:

- Programada en Java, produce sus resultados en archivos XML (eXtensible Markup Language, lenguaje utilizado para almacenar datos de manera legible).
- Permite el desarrollo de procesos de minería de datos mediante el encadenamiento de comandos a través de su entorno gráfico.
- Muy utilizada por científicos de datos en el campo de la investigación por su capacidad de análisis predictivo y su rapidez en la validación de modelos.
- También recomendada para el sector empresarial, permitiendo operaciones para la toma inmediata de decisiones y acciones para optimizar resultados de negocio.
- Algunas ventajas:
  - **Multiplataforma.**
  - **Potente**: Provee más de 500 operadores en relación a tareas de aprendizaje automático.
  - **Flexible y extensible**: Sirve tanto para un primer análisis como para una profunda minería de datos. Además permite usar algoritmos incluidos en Weka y un módulo de integración con R.

Python:

- Creado por Guido Van Rossum en el año 1991 e inspirado por el lenguaje de programación C, Python se ha convertido en el mayor competidor de R.

- Al igual que R, permite la aplicación de técnicas y métodos de todo tipo mediante la utilización de librerías.
- Lenguaje por excelencia de analistas y científicos de datos que provienen de áreas de la programación. Generalmente usado para tareas de integración de algoritmos en aplicaciones o cuando se quieren incorporar estos a nivel productivo.
- Ventajas:
  - **Sencillez**: su sintaxis es "limpia" y las funciones siempre se escriben de la misma forma.
  - **Flexible**: es ideal para crear nuevas librerías con las que probar funciones que no hayan sido aún desarrolladas, así como programaciones de script para web u otras aplicaciones.
  - **Fácil de aprender**: Python es fácil e intuitivo. Su buena lectura permite una curva de aprendizaje muy rápida y lineal.
  - **Extendido**: La comunidad online tiene una gran presencia en Internet y eso permite encontrar soporte y respuesta a todo tipo de dudas.

2.4.3 Tableau

Figura 12. *Logo del software Tableau*

"En 2020, el mundo generará 50 veces más cantidad de datos que en 2011. Y el número de fuentes de información se multiplicará por 75 (IDC, 2011). Con estos datos, se avecinan oportunidades enormes para el avance humano. Sin embargo, para que las oportunidades se conviertan en realidad, las personas necesitan tener el poder de los datos al alcance de la mano. Tableau está creando un software que proporciona exactamente eso."

Equipo de Tableau

Definición: Fundado en California, en el año 2003 por Chris Stolte, Christian Chabot y Pat Hanrahan, Tableau es un software de nueva generación orientado a la Inteligencia de Negocio para la visualización de datos interactiva mediante la generación de gráficos y mapas. A continuación se enumeran algunas de las ventajas que han llevado a su elección:

- **Belleza**: permite la creación de lienzos de gran belleza con la incorporación de todo tipo de colores, tipografías e iconografías acerca del diseño.
- **Conexión total**: permite la importación de datos desde todo tipo de fuentes: SQL, Oracle, archivos de texto, archivos Excel, etc.

- **Intuitivo y completo**: los controles presentan un formato de fácil manejo que invita a usar un mayor número de funcionalidades como filtros, agrupaciones y cálculos con las que mejorar la generación de gráficos y aumentar su impacto.
- **Rápido, ágil e interactivo:** para crear vistas sólo es necesario arrastrar los campos desde el conjunto de datos a los diferentes estantes que forman cada hoja de trabajo. El motor gráfico permite una velocidad incomparable a otros software a la hora de generar visualizaciones.
- **Comunidad y soporte:** peses a ser un programa relativamente nuevo, su comunidad de usuarios crece a gran ritmo, debido especialmente a que la propia compañía ha diseñado un sitio donde reunir a sus usuarios y socios para que éstos puedan compartir sus trabajos, estar al tanto de eventos de presentación de nuevos productos o buscar ideas y consejos en el foro.
- **Personalización y libertad:** admite un marco de extensibilidad eficaz para integraciones empresariales exhaustivas y complejas.

Como desventaja podríamos destacar su baja experiencia, ya que lleva poco tiempo en el mercado y esto puede hacer que las compañías no confíen en su efectividad y rendimiento.

2.4.4 Alternativas al uso de Tableau.

QlikView: herramienta de Data Discovery que ofrece soluciones intuitivas para guiar al usuario a una visualización y analítica de datos guiada de auto-servicio. Sus clientes (cerca de 36.000 hasta la fecha[3]) hacen uso de ella para extraer conocimiento y explorar relaciones ocultas entre los datos con las que impulsar la aparición de buenas ideas.

Principales características:

- **Big Data**: permite analizar grandes cantidades de información para la exploración y búsqueda de valor a través de todos los aspectos de negocio. Esto hace que sea una opción muy utilizada por empresas que requieren combinar acceso, con un nivel alto de volumetría.
- **Libertad**: tanto en las opciones de exploración como en la edición, ya que está diseñado para poder completarlo.
- **Experiencia**: lleva desde el 2005 en el mercado por lo que cuenta con un número importante de clientes a lo largo del mundo.
- **Buenas aplicaciones**: sus diferentes aplicaciones hacen que se pueda utilizar en cualquier lugar y en cualquier dispositivo, ya sea iPad, iPhone, PC, etc.

Debido a que su competidor directo en el mercado es Tableau, se van a citar ciertas desventajas en comparación con este:

- **Menos intuitivo y asequible**: esto hace que la dependencia del departamento de IT a la hora de realizar análisis sea mayor para otros departamentos como por ejemplo el Financiero o el de Marketing.
- **Más costoso**: no existe licencia académica y las licencias comerciales son de un coste sensiblemente menor a las de Tableau.

- **Más lento**: aunque su exploración también es totalmente libre, su respuesta a la hora de interactuar y generar visualizaciones es algo menor que en Tableau.

Microsoft Power BI: es un conjunto de aplicaciones de análisis de negocios de reciente aparición con los que Microsoft promete competir en muy poco tiempo en el mercado Big Data. Aunque muchas de sus características están aún en desarrollo, entre sus funciones más destacadas podemos mencionar:

- **Económico**: Su coste promete ser uno de los aspectos que atraiga a investigadores y empresas al presentar licencias por sólo 9.99 dólares al mes, lo que supone una fuerte reducción de precio en comparación con las dos herramientas anteriores.
- **Inteligente**: aunque aún está en desarrollo, se pretende que sea capaz de procesar peticiones en lenguaje natural. Es decir, los usuarios pueden preguntarle a la plataforma y ésta responde a las peticiones con gráficos.

3. Análisis exploratorio

3.1. Etapa de exploración

La etapa de exploración es sin duda una de las etapas más importantes del proyecto. El investigador debe conocer en profundidad las variables de las que dispone, para saber que información aporta cada una de ellas, ya que esto condicionará la selección de variables que interesa representar y la aplicación de los modelos estadísticos.

3.1.1 La fuente de datos

Figura 13. Logo del repositorio de la Universidad de California

Los datos que se utilizarán en este proyecto se encuentran en el repositorio de aprendizaje automático de la Universidad de California en Irvine. En él se pueden encontrar una gran colección de bases de datos, teorías de dominio y datos generados que son utilizados por la comunidad de machine learning o aprendizaje automático para el ensayo de algoritmos de este tipo. El archivo fue creado en 1987 como iniciativa de una serie de estudiantes de posgrado, y desde ese año ha sido usado por un gran número de estudiantes, educadores e investigadores, llegando a ser conocido por la comunidad de ciencia computacional como fuente primaria para abordar problemas de análisis de datos como predicciones, clasificaciones o series temporales.

Actualmente el proyecto se realiza en colaboración con Rexa.info en la Universidad de Massachusetts en Amherst, y cuenta con el apoyo financiero de la fundación de Ciencia Nacional.

3.1.2 El conjunto de datos. Descripción y atributos

El fichero de datos contiene medidas del consumo eléctrico de una vivienda a lo largo de un periodo de casi cuatro años (47 meses). Las medidas han sido tomadas con una frecuencia de un minuto por lo que el número de instancias es muy grande, concretamente se dispone de 2.075.259 mediciones, llevadas a cabo entre Diciembre del año 2006 y Noviembre del 2010. Los datos son reales y fueron donados el 30 de Agosto del 2012.

Por tanto, disponer de un gran número de datos continuos nos permitirá construir un modelo lo suficientemente sólido como para hacer observaciones, determinar si existe influencia de factores externos en determinados períodos o poder predecir el consumo de próximos meses.

A continuación se muestra una descripción de la información que representa cada una de los atributos del conjunto de datos:

1. **Date (**Fecha): corresponde con la fecha en formato día/mes/año en la que se realizó la medida.
2. **Time** (Hora): hora en formato hh:mm:ss en la que se realizó la medida.
3. **Global active power** (Potencia activa general, en kilovatio): cantidad de potencia activa consumida por el sistema eléctrico de la casa en el día y hora correspondiente.
4. **Global reactive power** (Potencia reactiva global, en kilovatio): cantidad de potencia reactiva consumida por el sistema eléctrico de la casa en el día y hora correspondiente.
5. **Voltage** (Voltaje, en voltios): demanda de voltaje de la casa
6. **Global intensity** (Corriente total, en amperios): corriente del sistema eléctrico
7. **Submetering 1** (Submedida 1, en vatio-hora): consumo correspondiente a la cocina, en la cual se encuentra un lavavajillas, un horno y un microondas (la cocina en sí no es eléctrica sino de gas).
8. **Submetering 2** (Submedida 2, en vatio-hora): consumo correspondiente a la otra parte de la cocina y que engloba el consumo de: lavadora, secadora, horno y luz.
9. **Submetering 3** (Submedida 3, en vatio-hora): consumo corresponde al calentador de agua y al aire acondicionado.

Asimismo, en la descripción del archivo se indica que hay algunos valores faltantes o perdidos, conocidos en la minería de datos como *missing values*, en torno al 1,25% de las filas e indicados con un signo de interrogación: "?". Esto suele ser un problema común en cualquier proyecto de tratamiento de datos relacionados con mediciones debido a fallos en los instrumentos de medida. Ignorar esto puede tener repercusiones graves que van desde la perdida de potencia del estudio hasta la aparición de anomalías en la tendencia de los gráficos, de ahí la necesidad de realizar etapas de limpieza y preparación de los datos antes de aplicar los modelos predictivos [veáse anexo para rellenado de valores faltantes].

3.1.3 Descarga y comprensión de los datos

El formato del conjunto de datos es de tipo texto (.txt), por lo que podremos descargarlo directamente desde el link del repositorio web [5], para importarlo posteriormente desde Tableau y R.

A continuación mostramos una vista previa de los primeros datos del archivo para comprobar su disposición:

household_power_consumption.txt

```
Date;Time;Global_active_power;Global_reactive_power;Voltage;Global_intensity;Sub_metering_1;Sub_metering_2;Sub_metering_3
16/12/2006;17:24:00;4.216;0.418;234.840;18.400;0.000;1.000;17.000
16/12/2006;17:25:00;5.360;0.436;233.630;23.000;0.000;1.000;16.000
16/12/2006;17:26:00;5.374;0.498;233.290;23.000;0.000;2.000;17.000
16/12/2006;17:27:00;5.388;0.502;233.740;23.000;0.000;1.000;17.000
16/12/2006;17:28:00;3.666;0.528;235.680;15.800;0.000;1.000;17.000
16/12/2006;17:29:00;3.520;0.522;235.020;15.000;0.000;2.000;17.000
16/12/2006;17:30:00;3.702;0.520;235.090;15.800;0.000;1.000;17.000
16/12/2006;17:31:00;3.700;0.520;235.220;15.800;0.000;1.000;17.000
16/12/2006;17:32:00;3.668;0.510;233.990;15.800;0.000;1.000;17.000
16/12/2006;17:33:00;3.662;0.510;233.860;15.800;0.000;2.000;16.000
16/12/2006;17:34:00;4.448;0.498;232.860;19.600;0.000;1.000;17.000
16/12/2006;17:35:00;5.412;0.470;232.780;23.200;0.000;1.000;17.000
16/12/2006;17:36:00;5.224;0.478;232.990;22.400;0.000;1.000;16.000
16/12/2006;17:37:00;5.268;0.398;232.910;22.600;0.000;2.000;17.000
```

Figura 14. Vista previa del conjunto de datos a analizar

Como se observa en la imagen, la información viene en el mismo orden y estructura que en la descripción, por lo que no tendremos que realizar ninguna modificación cuando las importemos en las herramientas. El archivo se ha guardado automáticamente en la carpeta de descargas del ordenador, pero lo moveremos a una nueva carpeta específicamente creada para el proyecto, a la que denominamos "DATOS", ubicada en el escritorio y que utilizaremos como directorio para las herramientas utilizadas en el trabajo.

¿ Con qué tipo de datos estamos trabajando ?

La información contenida viene separada por atributos mediante punto y coma, y es de dos tipos:

1. **Datos temporales**: fecha y hora a la que corresponde cada medida.
2. **Datos numéricos de tipo decimal**: correspondientes a las diferentes mediciones del consumo eléctrico del hogar.

Desde el punto de vista general, el conjunto de datos es una serie temporal:

> **Serie temporal**: se conoce así al conjunto de observaciones x_t recogidas a lo largo de un período continuo de tiempo t. Estas pueden ser regulares si presentan datos con una distribución igual en el tiempo, como en nuestro caso (cada minuto), o desiguales, como sería por ejemplo el peso de una persona en sucesivas mediciones en la consulta del médico.
>
> En una serie temporal, las observaciones sucesivas son dependientes entre sí, por lo que el análisis tiene que llevarse a cabo teniendo en cuenta que los valores en un instante estén en cierta medida determinados por los valores que ha tomado la serie el orden temporal de las observaciones. Esto va a ser de gran relevancia a la hora de estudiar los posibles modelos a aplicar, ya que todos aquellos que no estén basados en la dependencia de las observaciones, no serán válidos.
>
> En general los métodos más representativos a aplicar en series temporales buscan dos objetivos:
>
> 1. Estudiar los cambios que experimentan las variables respecto al tiempo.
> 2. Predecir sus valores futuros.
>
> En cuanto a clasificación, existen dos tipos de series temporales:
>
> 1. **Estacionaria**: cuando la media y la variabilidad se mantienen constante a lo largo del tiempo.
> 2. **No estacionaria**: cuando la serie muestra una tendencia, es decir, cambios a lo largo del tiempo. Además pueden existir efectos estacionales, en períodos que presenten un comportamiento parecido, o irregularidades.

¿ Qué variables nos interesa analizar ?

Antes de comenzar a manejar los datos en R, debemos hacer un estudio exhaustivo de cada variable, con el fin de conocer qué atributos van a jugar un rol más importante en el análisis.

Al tratarse de una base de datos almacenada a lo largo de un período de tiempo, las dos primeras variables, fecha y hora, van a ser esenciales a lo largo del estudio, ya que nos permitirán situar los datos cronológicamente y distribuir el resto de variables en intervalos de tiempo para estudiar posibles picos de consumo, así como para agrupar las variables en los modelos de agrupación o clustering.

En cuanto a los valores numéricos el atributo más importante va a ser la Potencia activa ("*Global active power*"), también conocida como Potencia real y que representa la energía neta transferida a lo largo del sistema eléctrico. Este valor será por tanto el que marque el importe de la factura final que pagará al usuario a final de mes.

Otras tres variables importantes serán las medidas de los diferentes electrodomésticos ("*Sub_metering_1, Sub_metering_2 y Sub_metering_3*"), ya que nos permitirán saber qué aparatos de la casa se utilizan con mayor frecuencia y por tanto, más gasto suponen. El resto de atributos: potencia reactiva, voltaje e intensidad van a ser omitidos en un primer momento, ya que para el estudio que queremos realizar no son de gran relevancia. Al final del trabajo se propondrán posibles trabajos futuros que podríamos llevar a cabo con estas variables.

Por tanto, las variables a estudio son:

Date, Time, Global_active_power, Sub_metering_1, Submetering_2 y Submetering_3.

3.2. Visualización (parte I)

No cabe duda de que la vista es el sentido que mayor información proporciona al ser humano. La introducción de un gráfico en una presentación, hace que el oyente preste más atención a lo mostrado y pueda interpretar datos complejos con menos esfuerzo y en cuestión de segundos. Cuando tratamos con series de datos de gran volumen como los correspondientes a nuestro proyecto, resulta muy difícil tener una visión general de los datos que nos permita comprender la información, describirla y establecer conclusiones.

La visualización de datos ofrece la posibilidad de mostrar la información relevante de un modo directo e intuitivo, apoyándose en múltiples gráficos que se consideren interesantes según las preferencias y objetivos de los usuarios. En este apartado se pretende reflejar la importancia que tiene la representación gráfica de los datos como etapa dentro del proceso de análisis como instrumento de estudio con el que comprender mejor el comportamiento de los datos para afrontar con mayor conocimiento la etapa de selección y aplicación de los modelos predictivos.

Para realizar las visualizaciones, los datos serán cargados en Tableau, en concreto en la última versión de su producto *Tableau Desktop*, versión 10.0. [véase instalación y primeros pasos en el anexo 3 al final de la memoria].

3.2.1 Extracción y preparación de los datos

Una vez realizada la conexión a los datos [véase anexo apartado 2], debemos comprobar si los datos se han importado en la disposición deseada y si estos contienen la información correcta.

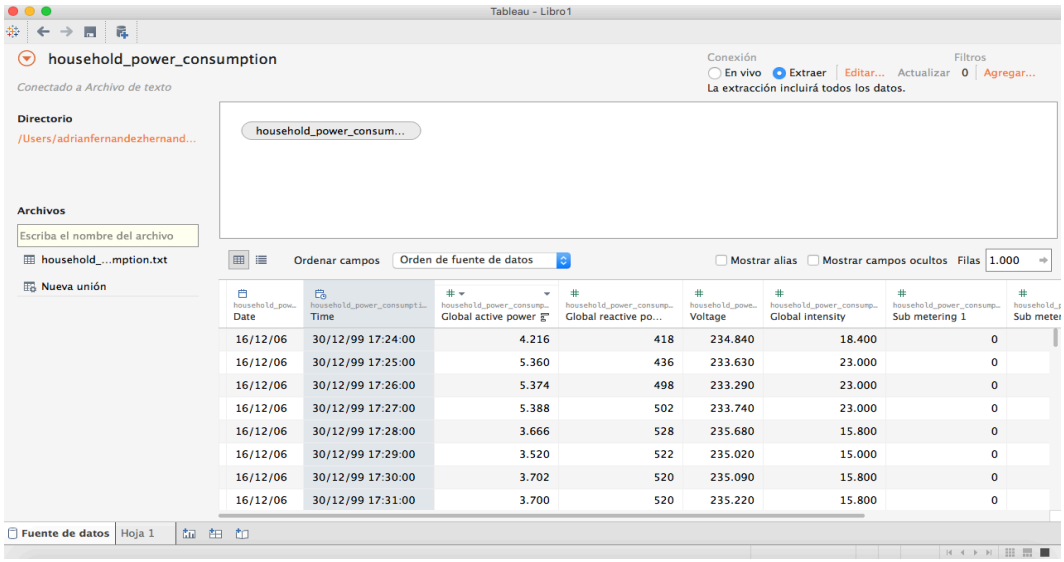


Figura 15. Fuente de datos conectada y extraída a Tableau (sin modificar)

Como vemos en la imagen, al importar nuestra fuente de datos se ha creado por defecto una fecha dentro de la columna *Time* de nuestro conjunto. Esto suele ocurrir con frecuencia en Tableau cuando datos relacionadas con fechas y horas se encuentran en columnas diferentes, por lo que habrá que conectar ambos campos antes de pasar al diseño de gráficos.

Para evitar este problema, Tableau permite calcular un campo calculado que combine las columnas *Date* y *Time*, y de esta manera eliminar esa fecha aleatoria y errónea que pudiera confundir al programa en la ubicación temporal de los gráficos.

Tras seleccionar ambas columnas, escribimos el cálculo deseado:

```
DATEADD('hour', DATEPART('hour', [Time]), DATEADD('minute', DATEPART('minute', [Time]), DATEADD('second', DATEPART('second', [Time]), [Date])))
```

Automáticamente aparece nuestro nueva columna *Date + Time*:

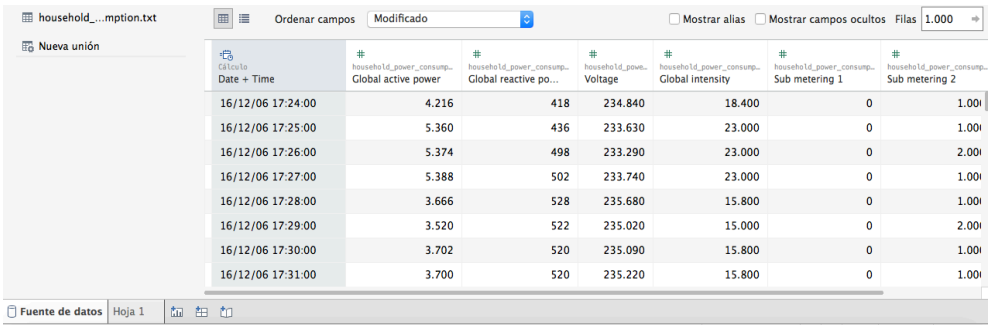


Figura 16. Fuente de datos conectada y extraída a Tableau (ya preparada)

El resto de columnas mantienen la misma información que el archivo original excepto la Potencia reactiva, Global reactive power, donde vemos que se dan valores de 418 kW frente a los 0.418 kW que reflejan nuestro archivo original. Esto ocurre porque Tableau ha leído los valores sin el cero de la izquierda, tomando la parte decimal como entera. Aunque no es una variable que vamos a utilizar, para solucionar esto podríamos crear un nuevo campo calculado:

[Global reactive power]/1000

La nueva columna generada ya tendría valores correctos: 0.418 kW, 0.436 kW, etc.

Como hemos podido comprobar Tableau no siempre importa las columnas del modo correcto, por lo que siempre que realicemos una nueva conexión de datos, es muy importante no pasar a la representación de gráficos sin antes comprobar que nuestros datos han sido correctamente importados.

Una vez conectado y modificado los errores de importación de nuestra fuente de datos, ya podemos dar paso a la visualización en la primera hoja generada: Hoja1.

3.2.2 ¿Qué tablas o gráficos son los más adecuados?

Para realizar una buena visualización es importante el uso de distintos gráficos, ya que cada tipo de gráfico representa la información de una forma diferente.

Por tanto, conocer qué gráfico es el adecuado va a ser esencial para obtener conocimiento en esta etapa de visualización.

Como en nuestro proceso trabajamos con una serie temporal de datos, los gráficos que nos interesan son aquellos que permitan representar las distintas variables a estudio a lo largo del periodo de tiempo.

Para este tipo de visualización, los gráficos más comunes son:

- **Histograma**: es un tipo de gráfico que representa la distribución y frecuencia de los datos. Se trata de una de las mejores opciones para hacer comparaciones entre valores numéricos, ya que separa los datos estudiados en barras de diferentes alturas, permitiendo de esta manera ver tendencias de un modo directo.
- **Gráfico de líneas:** es una forma simple de visualizar una secuencia de valores. Los gráficos de línea conectan puntos de datos numéricos almacenados en un periodo en el tiempo. Son muy recomendados para mostrar cambios en el tiempo y en algunos casos su combinación con gráficos de barras puede ser de gran utilidad para reforzar la investigación.
- **Diagrama de caja y bigote (box-plot):** es un gráfico que describe características importantes como la dispersión y la simetría de los datos.
  Está formado por un rectángulo (la caja) y dos brazos (los bigotes). La caja a su vez se divide en tres cuartiles, el segundo Q_2 se corresponde con la mediana y separa los cuartiles Q_1 y Q_3. Son muy recomendados para observar valores mínimos y máximos, medianas y sobre todo para descubrir posibles valores atípicos.

- **Gráficos de forma**: permiten contrastar distintas variables mediante la representación de cada una de ellas con una forma diferente. Son recomendados cuando se quieren visualizar variables con un comportamiento similar en el mismo gráfico como por ejemplo las mediciones de los diferentes equipos del hogar.
- **Mapas de calor**: revelan información mediante la graduación de color entre dos factores. Por ejemplo, tratando con datos de consumo, podríamos reflejar qué periodos de tiempo tienen mayor consumo mediante una transición de azul (bajo consumo) a rojo (alto consumo) Una vez definido y comprendido cada gráfico, el siguiente paso es saber a qué preguntas queremos tratar de responder con su representación.

A continuación se enumeran una serie de preguntas que nos pueden ayudar a encontrar el tipo de gráfico más adecuado :

- ¿ De qué tipo son las variables de que se dispone ?
- ¿ Qué información se desea visualizar ?
- ¿ Cuántas variables se desea representar ?
- ¿ A qué se pretende dar respuesta ?; ¿ Qué se desea mostrar o comparar ?
- ¿ Qué tipo de gráfico nos permite responder a esa pregunta ?
- De las opciones disponibles, ¿Qué tipo de gráfico se prefiere?

Todas estas preguntas (como mínimo) deberían ser formuladas cada vez que elaboremos un nuevo gráfico, ya que por ejemplo no es lo mismo representar localizaciones geográficas, para lo cual sería recomendable utilizar un gráfico de tipo mapa, que estudiar un conjunto de ventas producidas en un determinado período de tiempo, donde lo más adecuado sería usar un gráfico de línea o de barras.

3.2.3 Estudio preliminar

Una vez descritos los tipos de gráficos más recomendados para la representación de series temporales y la filosofía pregunta-respuesta(gráfico) con la que se va a trabajar, en este capítulo nos centraremos en su representación para analizar aspectos relacionados con el consumo (medida de potencia activa y de los diferentes electrodomésticos de la casa).

De esta manera vamos a poder analizar factores que no somos capaces de observar en un primer momento con los datos en el formato original y que darán respuesta a preguntas acerca de cómo se comporta la demanda eléctrica del hogar.

Para comenzar y como primera iniciación a Tableau, vamos a mostrar la configuración en vista completa de la primera hoja generada (Hoja 1) para reflejar la interfaz de trabajo a utilizar a lo largo de toda la etapa de exploración.

¿ Qué aspecto tiene el consumo a lo largo del período total ?

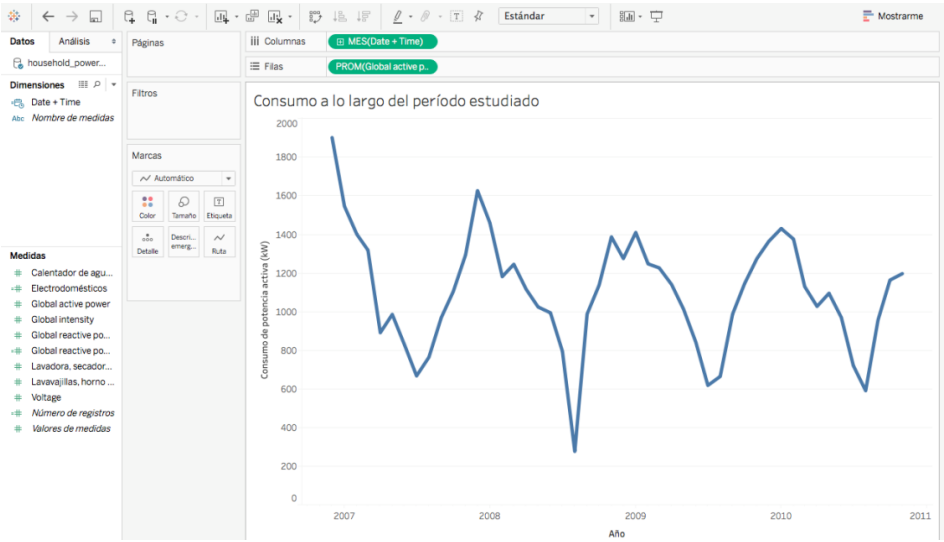


Figura 17. Consumo total de la serie temporal a lo largo de su periodo

A mayor potencia activa requerida, mayor consumo, por lo que como se puede apreciar en el gráfico:

- El gráfico presenta un comportamiento similar a lo largo de los cuatro años, viéndose cuatro descensos periódicos de consumo cada cierto tiempo, seguidos de respectivos aumentos con distintas pendientes.
- Las cifras de consumo más elevadas se producen en el primer mes (diciembre de 2006). A partir de esa fecha el consumo no vuelve a alcanzar dicha cifra en los cuatro años siguientes.
- Cada año presenta una serie de irregularidades en su comportamiento.
  Se pueden distinguir cambios bruscos de consumo frente a otros períodos en los que el consumo aumenta o disminuye de forma lineal.

Pese a que el gráfico representa toda la serie de un simple vistazo, no podemos establecer conclusiones fiables y concretas acerca del consumo, debido a que la gráfica es muy general y no permite realizar una buena lectura acerca de cuándo se producen esas fluctuaciones en el tiempo.

Llegados a este punto es donde tendremos que atacar a los datos, produciendo gráficas que vayan aportando cada vez más información y sobre todo que respondan siempre a alguna pregunta.

Al tratarse de datos relacionados con el uso de la electricidad, debemos tener claro desde el primer momento que la actividad de los habitantes de la casa va a influir de manera directa en el consumo. Asimismo, la energía no se utiliza de forma regular a lo largo del tiempo, ya que por ejemplo y centrándonos en nuestros datos, si la familia está de vacaciones, la actividad será menor, y por lo tanto el consumo medio bajará por todos esos días que la casa ha

permanecido inhabitada. Esto se va a traducir en gráficos con cambios bruscos entre meses de inactividad y aquellos en los que se haya hecho más vida en el hogar.

De esta manera, los meses de invierno deberían presentar un consumo elevado, ya que son períodos en los que los días son más fríos y las horas de luz se reducen en comparación a otras estaciones del año, obligando a la familia a incrementar el consumo eléctrico para poder iluminar y calentar las distintas zonas de la casa.

Por el momento, todo lo escrito en el anterior párrafo se basa en suposiciones. Es cierto que si preguntáramos a un grupo de diez personas acerca de por qué creen que el gráfico mostrado anteriormente presenta ese perfil, probablemente la mayoría coincidirían en lo que hemos interpretado nosotros. Con esto no se pretende reflejar que la representación de los datos de esta manera haya sido una pérdida de tiempo, sino todo lo contrario.

Cuando tratamos con datos en Tableau, debemos de empezar con la representación del gráfico más sencillo posible, para tener una primera idea de lo que está pasando y que esto desencadene el planteamiento de nuevas preguntas. Cada nuevo gráfico generado, permite descubrir nueva información que habrá que utilizar como pistas acerca de qué es lo siguiente que sería interesante analizar. De esta manera, se podrá establecer un análisis cada vez más completo y profundo de los datos que nos permita responder a preguntas que en un primer momento no éramos capaz de responder, y que probablemente ni siquiera se nos habría ocurrido preguntarnos.

Por tanto con el propósito de identificar mejor el consumo, volvemos al gráfico para pensar sobre aspectos que nos interesaría analizar.

Sabemos que el consumo general presenta una serie de fluctuaciones en el tiempo pero no sabemos a qué meses corresponde ni cómo se produce a lo largo de cada uno, por lo que vamos a distribuir los datos en lugar de por años, por meses, para ver a qué se puede deber esa forma en el perfil anteriormente dibujado, y sobre todo para comprobar si lo anteriormente planteado es cierto o no.

¿ Cómo es el consumo a lo largo del año ?

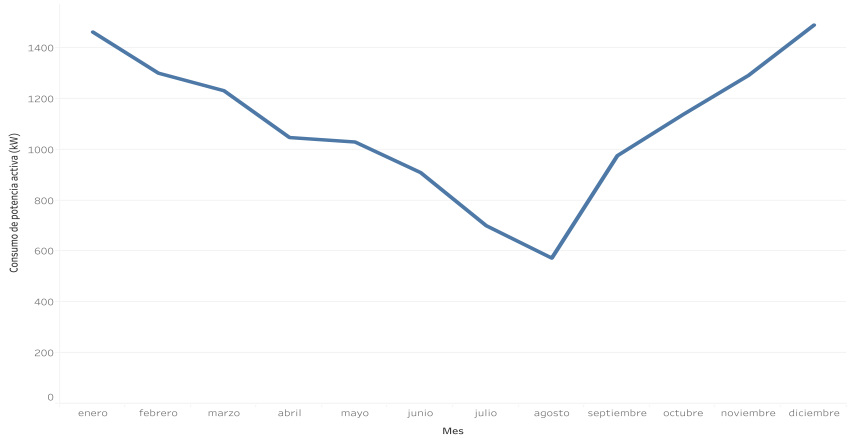


Figura 18. Consumo mensual a lo largo de un año promedio (gráfico de línea)

En este gráfico podemos ver de manera más directa lo que está provocando esas oscilaciones en nuestra serie temporal. Observamos como el consumo sufre una caída progresiva desde enero hasta agosto, seguramente a raíz de tratarse de un período vacacional en el que la familia permanece fuera de la casa. A partir de agosto se produce la mayor subida de consumo (mes de septiembre) y continua su ascenso lineal hasta diciembre.

Para tratar de descubrir aspectos más en profundidad, vamos a representar esto haciendo uso de las algunas de las opciones que ofrece Tableau. Para ello se han representa de nuevo la potencia activa promedio de todas las medidas recogidas, pero esta vez se han distribuido los datos en un diagrama de barras a lo largo de los doce meses que componen el año. A su vez utilizamos marcas de degradación de color para mayor impacto visual.

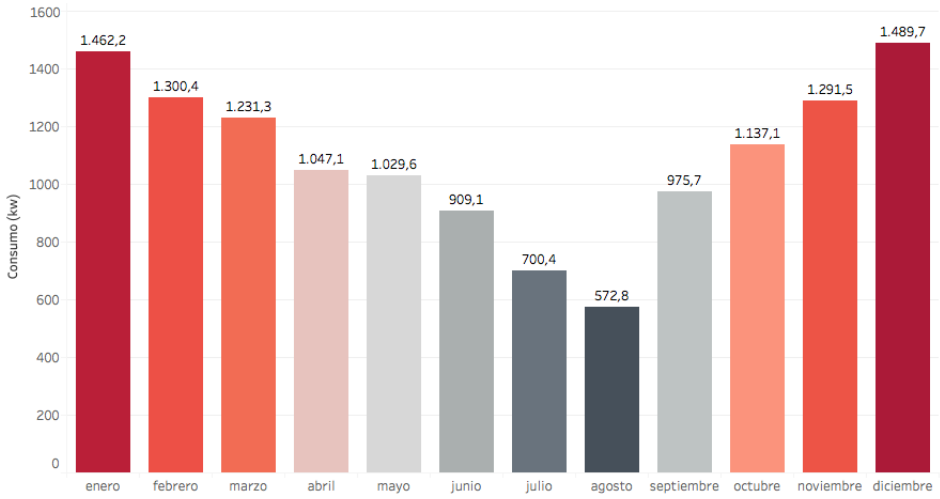


Figura 19. Consumo mensual a lo largo de un año promedio (gráfico de barras)

A primera vista podemos observar varios grupos:

- Los datos de consumo presentan un descenso progresivo desde el mes de Enero y hasta Agosto, y un aumento prácticamente lineal desde este mes y hasta final de año. Se confirman las suposiciones iniciales que habíamos considerado acerca del comportamiento de la primera gráfica.
- Meses de invierno: Enero y Diciembre, presentan el mayor consumo de electricidad, siendo Diciembre el mes de mayor gasto promedio con 1489,7 kW
- Febrero, marzo, octubre y noviembre presentan un consumo algo menor (1300-1100 kW)
- El resto de meses refleja un consumo en torno a los 1000kW, siendo agosto el mes de menor consumo eléctrico con casi la mitad de consumo (572,8 kW)

Si comparamos estas dos últimas gráficas entre sí, podemos ver que a pesar de estar representando la misma información (consumo a lo largo de cada mes), el hecho de utilizar diferentes estilos de visualización permite al analista interpretar factores diferentes en cada gráfico. Así por ejemplo, mientras que la primera gráfica nos permite leer mejor la tendencia y el comportamiento de los datos, la segunda nos muestra la segmentación de los meses en distintos grupos de diferentes alturas y colores en función del consumo medido.

Asimismo, estas nuevas gráficas nos indican que para analizar y poder sacar conclusiones sobre las características de nuestro conjunto de datos, va a ser una buena idea, enfocar el estudio en períodos más cortos y concretos en el tiempo en lugar de analizar los datos desde su visión general. Este nivel de visualización se conoce como granularidad de serie temporal y siempre que trabajemos con datos de este tipo va a ser de gran ayuda tenerlo en consideración, ya que de otra manera resulta muy difícil analizar correlaciones entre datos distribuidos en un periodo tan largo en el tiempo.

¿ Sucede esto en todos los años ?

Antes de pasar a reducir esa granularidad de la que hablábamos en el párrafo anterior, vamos a comprobar si el consumo siempre se produce de esta manera a lo largo de cada año o si la familia no presenta una regularidad en sus fechas. Dado que del año 2006 sólo disponemos de datos para un mes, con el fin de ofrecer una mayor precisión se decide excluirlo en la representación:

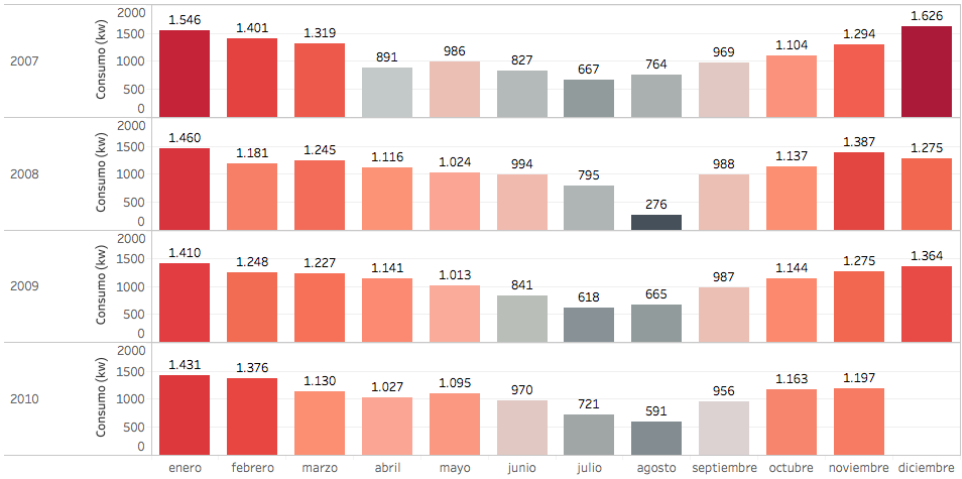


Figura 20. Consumo mensual de cada año de la serie temporal (gráfico de barras)

Observamos que:

- Julio y Agosto son los meses de cada año en los que menos electricidad se consume.
- Enero y Diciembre son los meses en los que más electricidad se consume a excepción de Noviembre de 2008 donde se dio un consumo mayor que para diciembre de ese mismo año.

Con esto se puede confirmar que la familia suele estar fuera en los meses de verano, lo que es de gran interés ante un posible ajuste de plan de ahorro energético, ya que si la familia no utiliza apenas su red eléctrica durante este período, se podría ahorrar en costes, tanto para la compañía en términos de generación como para la familia en costes relacionados con la potencia contratada.

¿ Cómo es el consumo durante los días del mes ?

Por lo general las familias tienden a consumir de manera irregular a lo largo de cada mes. Hay personas que tienden a consumir más a principio de mes, al haber cobrado el sueldo y sentir menos conciencia de ahorro económico, o puede ocurrir todo lo contrario, que durante las

primeras semanas de mes la familia disfrute de una mayor actividad social, realizando más vida fuera de la casa, y sea a finales de éste cuando mayor electricidad consuma.

Por tanto representamos el consumo medido durante los distintos días del mes

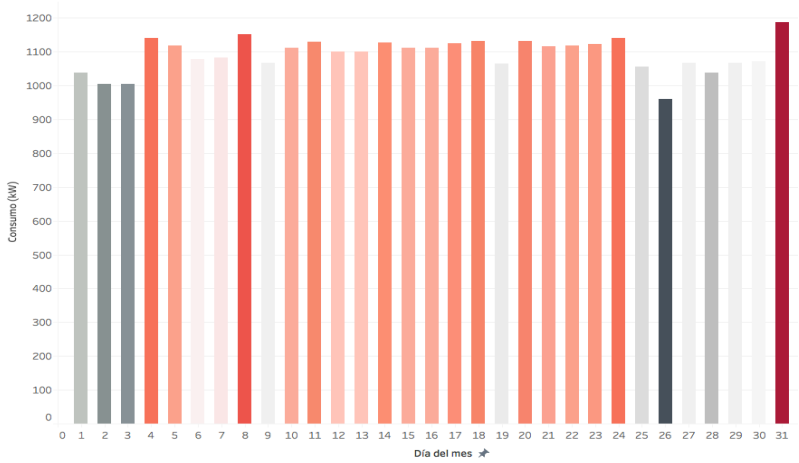


Figura 21. Consumo diario a lo largo de un mes promedio

Del gráfico podemos interpretar que:

- Los primeros y últimos días del mes el consumo es bajo.
- La curva de consumo es regular a medida que se avanza en el mes, siendo prácticamente constante desde el día 10 al 24 de mes.
- El día de mayor consumo es el día 8 con 1151,9 kW y el de menor el 26 con 922,1 kW
- El día 31 es de cierta sorpresa, ya que los días anteriores y posteriores el consumo es bajo.

Dado que resulta extraño que uno de los días de mayor consumo sea el 31, vamos a filtrar ese día para intentar descubrir a que se debe tal incremento

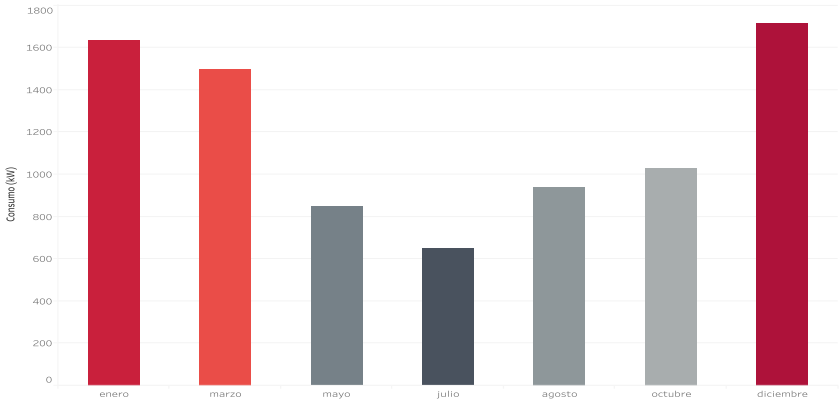


Figura 22. Consumo del día 31 de la serie temporal

Al filtrar todos los días 31 de nuestro conjunto de datos, sólo nos aparece información, cómo es lógico, de los meses que tienen 31 días. Justamente coincide que entre estos meses, se encuentran los meses de mayor consumo: Enero y Diciembre, por lo que aquí tenemos una de las posibles causas que hacen que el cálculo del consumo promedio se dispare para ese día.

Esto se debe tener en cuenta a la hora de trabajar con series temporales, ya que los meses tienen distinto número de días y al representar promedios totales puede influir en los resultados.

Otra causa podría estar relacionada con el aumento de consumo debido a los datos del mes de diciembre de 2006[1].

Día laborable vs fin de semana

A continuación, vamos a diferenciar entre los días laborables (en azul) y los días festivos (en verde) para saber si existe algún día especial en el que la familia realiza un mayor consumo, así como para conocer como se produce el consumo en los fines de semana.

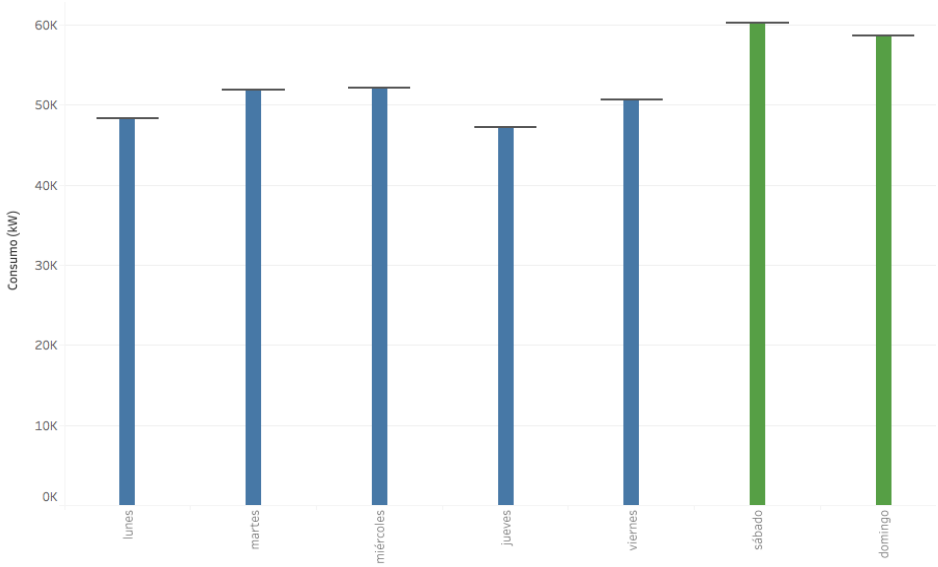


Figura 23. Consumo día laborable vs día fin de semana (gráfico de barras)

En un primer vistazo vemos que:

- Los Lunes y los Jueves son días de menor consumo con respecto al resto de la semana.
- Martes y Miércoles son días de mayor consumo semanal.
- El fin de semana el consumo es mayor, siendo el sábado el día de mayor consumo.

Para descubrir algo más, vamos a representar esto con un diagrama de cajas, donde cada punto representa el valor del consumo de cada una de las semanas medidas en cada día de la semana:

[1] Esto se comprueba filtrando el mes de diciembre de 2006, y efectivamente, el día 31 presenta un consumo algo menor si se decide no incluir en el gráfico.

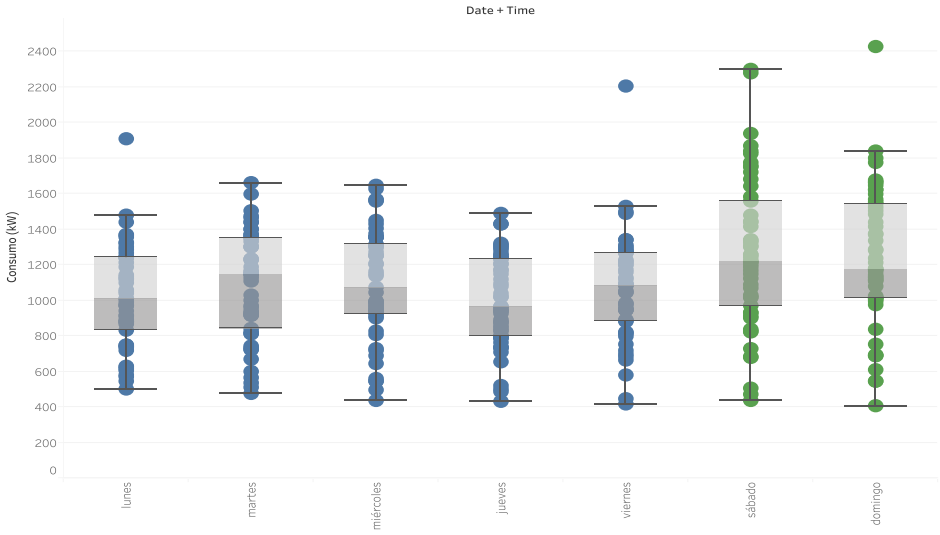


Figura 24. Consumo día laborable vs día fin de semana (gráfico de cajas)

Observamos que pese a algún punto disperso (lunes, viernes y domingo), la dispersión de los datos es uniforme entre el rango de valores (valor máximo y mínimo).

Por otro lado, los días laborables presentan una distribución de cuartiles más estrecha que para los días de fin de semana. Esto era de esperar, puesto que durante la semana los horarios de la familia están más limitados, bien por trabajo o por actividad escolares y universitarias, mientras que el fin de semana la familia no tiene obligaciones con las que cumplir, lo que se traduce en un consumo más aleatorio y difícil de predecir.

3.3 Visualización (parte II)

Llegados a este punto, se conocen ciertos comportamientos y rutinas que se repiten en el hábito de consumo y que nos han permitido sacar conclusiones respecto al perfil anual y mensual, lo que permite al analista conocer que la familia presenta un comportamiento de consumo más o menos esperado. Aunque esto nos servirá para trabajos futuros, es evidente que una empresa energética no requiere de un analista para explorar esta información, por lo que en los siguientes gráficos debemos reflejar aspectos que analicen más en profundidad los hábitos de consumo del hogar para poder encasillar y perfilar en mayor grado el tipo de consumo de la familia en su vida diaria y que de esta manera, se pueda llevar a cabo un futuro plan económico o estudio de eficiencia energética ajustado a una demanda más corta en el tiempo.

Por ello, vamos a focalizar nuestra exploración en en la búsqueda de respuestas a las dos preguntas planteadas en los objetivos, ya que son cuestiones que nos van a permitir conocer acerca del perfil y que por tanto van a influir en la toma de decisiones para la compañía. Estas son:

1. ¿ En qué momentos del día se produce mayor gasto energético y cómo se produce éste ? ¿Cómo es un día laborable frente a un día de fin de semana ?
2. ¿ Cuáles son los dispositivos que más energía consumen ?

3.3.1 Consumo diario total

Para responder a la primera pregunta debemos conocer cómo es la actividad de los habitantes de la casa a lo largo de las diferentes franjas horarias que componen un día (0-23h). De esta manera podremos agrupar las horas del día en las que se dispara el consumo, con el fin de saber cuáles son las horas en las que la familia realiza un mayor gasto energético. Estas horas se conocen como "horas pico" y son uno de los principales responsables del aumento de consumo eléctrico en las facturas.

Para ello vamos a generar gráficas de una granularidad más reducida en el tiempo que las anteriores, para así poder visualizar esto de manera directa.

¿ A qué hora se produce una mayor demanda energética ?

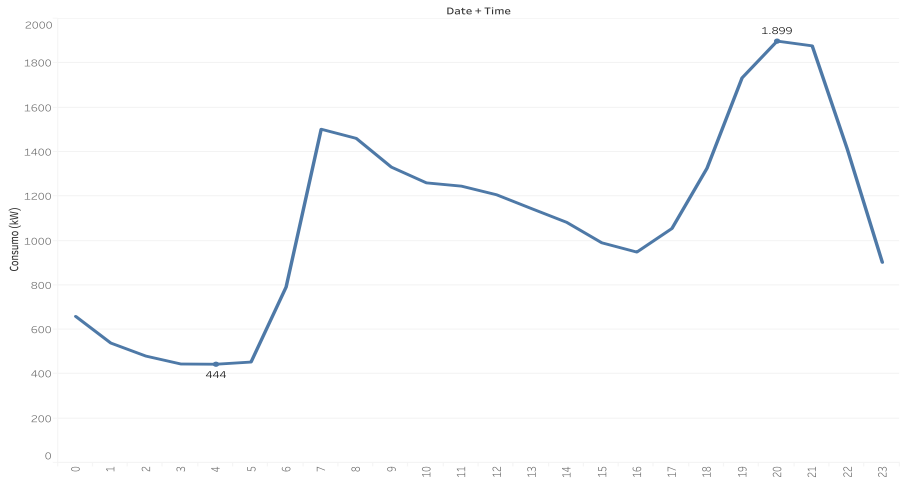


Figura 25. Consumo horario a lo largo de un día promedio (gráfico de línea)

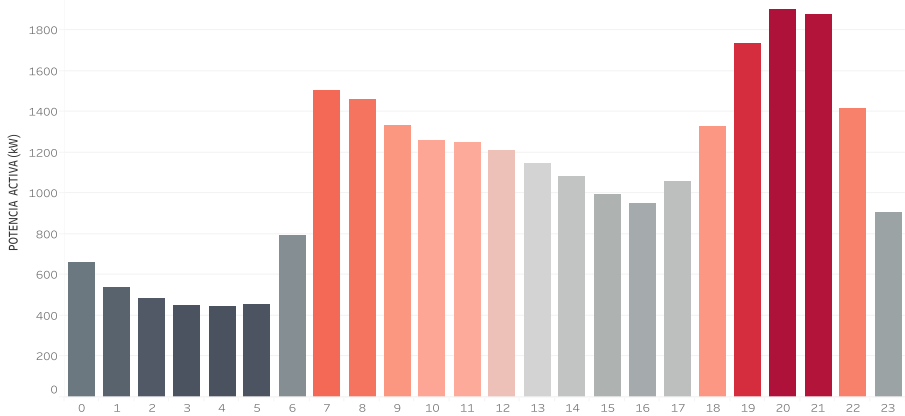


Figura 26. Consumo horario a lo largo de un día promedio (gráfico de barras)

Observamos que las horas de mayor consumo se corresponden con los periodos:

- Entre las 7:00 y las 8:00 horas
- Entre las 19:00 y las 21:00 horas

Por otro lado, durante la madrugada se produce el menor consumo energético:

- Entre las 2:00 y las 5:00 horas

También se puede observar un descenso a mitad de día:

- Entre las 12:00 y las 17:00 horas

Esto podría indicarnos que la mayor parte de la familia suele estar fuera de la casa en esas horas.

Tableau no sólo permite la visualización de los datos, si no que además ofrece la opción de copiar los datos como tabulación cruzada para poder exportarlos a otros programas como Excel, y de esta manera hacer una interpretación más detallada de los gráficos obtenidos.

Aunque el objetivo de la etapa no es la realización de cálculos si no la observación, a continuación, se va hacer un estudio más exhaustivo para determinar cuánto varía el consumo entre esas horas y así determinar cómo es el incremento/descenso de consumo entre horas.

Variación intrahoraria a lo largo del día

Tras importar los valores en Excel, se ha creado una nueva columna para calcular la variación porcentual o incremento que sufre cada nueva medida con respecto a su anterior valor. Así podemos comprobar si efectivamente las franjas horarias experimentan el mayor incremento de potencia en las horas destacas en al apartado anterior.

Como recordatorio, la variación porcentual se calcula de la siguiente forma:

$$Porcentaje\ de\ crecimiento = \frac{Medida\ actual - Medida\ anterior}{Medida\ anterior}$$

Tabla 1. Tabla de incrementos porcentuales de consumo por hora

Hora (0-24 h)	Potencia activa (kW)	Variación porcentual	Hora (0-24 h)	Potencia activa (kW)	Variación porcentual
0	659	-	12	1.207	-0,031300161
1	539	-0,182094082	13	1.145	-0,051367026
2	**481**	-0,107606679	14	1.083	-0,054148472
3	**445**	-0,074844075	15	991	-0,084949215
4	**444**	-0,002247191	16	949	-0,042381433
5	**454**	0,022522523	17	1.055	0,111696523
6	792	0,744493392	18	1.326	0,256872038
7	**1.502**	**0,896464646**	**19**	**1.733**	**0,30693816**
8	**1.461**	-0,027296937	**20**	**1.899**	0,095787651
9	1.332	-0,088295688	**21**	**1.878**	-0,011058452
10	1.261	-0,053303303	22	1.413	-0,247603834
11	1.246	-0,011895321	23	902	-0,361641897

En la tabla observamos como los valores de potencia más bajos, en kW y representados en color verde, se corresponden con las horas en las que las barras del gráfico 8 presentan una menor altura. Por otro lado los picos de consumo, representados en un tono más rojo,

coinciden con aquellos valores que experimentan un salto de altura importante con respecto a la hora anterior:

- De 6:00 a 7:00 la potencia sufre una subida del 89,6%, 710 kW, en la medida del contador eléctrico.
- A su vez, de 18:00 a 19:00 también se produce un salto considerable, esta vez del 30,7%, con una diferencia de 407 kW

¿ Los picos de consumo se dan a las mismas horas en los diferentes años ?

El análisis anterior se ha calculado con una estimación promedio de todos los datos almacenados durante los casi cuatro años de período pero si diferenciamos cada año, ¿el resultado sería el mismo?

Representamos los datos en un mapa de color para mejorar el impacto visual

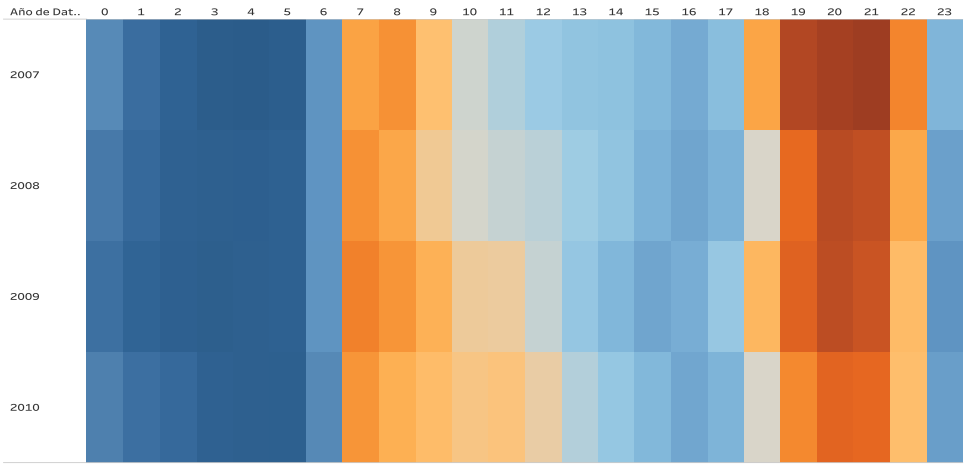


Figura 27. Consumo horario de cada año de la serie (mapa de calor)

Como se puede apreciar, en todos los años el pico de potencia se produce prácticamente en las mismas franjas horarias observadas en el anterior apartado.

¿ Y qué ocurre a lo largo de los meses del año ?

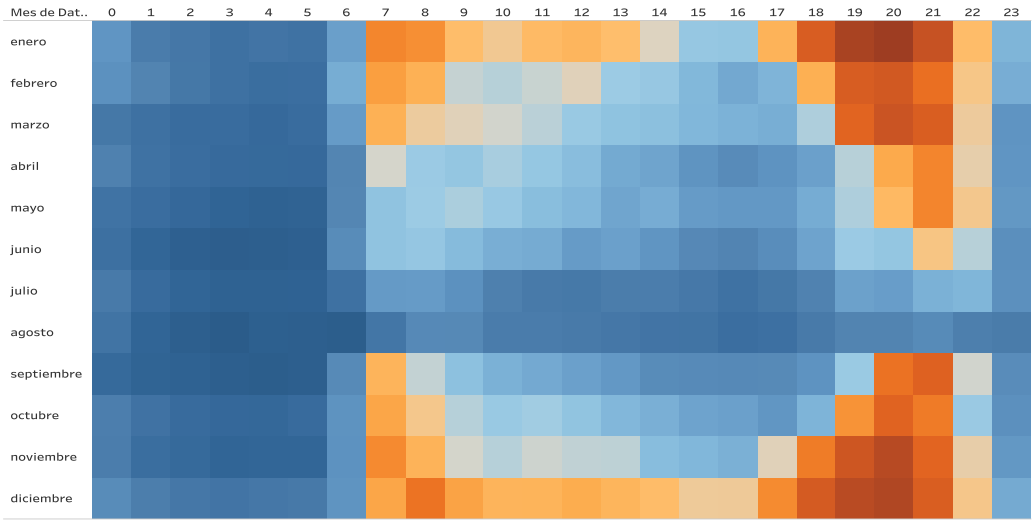


Figura 28. Consumo horario por mes en un año promedio (mapa de calor)

Lo mismo. Esto nos confirma que la familia tiene como hábito realizar sus actividades domésticas en estas horas.

3.3.2 Consumo diario de los diferentes electrodomésticos

Al igual que resulta interesante conocer las horas de mayor consumo del hogar y dado que tenemos datos de las medidas de algunos electrodomésticos de la casa, podría ser interesante conocer como es el consumo de estos, ya que su uso va a estar directamente relacionado con el consumo general diario calculado en el anterior apartado.

En este caso, la información nos viene dada en vatio hora, por lo que elaborar una representación apropiada de los datos, puede aportar al analista conocimiento importante acerca de comportamientos y hábitos de uso que contribuyan a una mayor conciencia en el uso eficiente de los aparatos y por tanto a una reducción en la factura.

Representamos los tres grupos de electrodomésticos

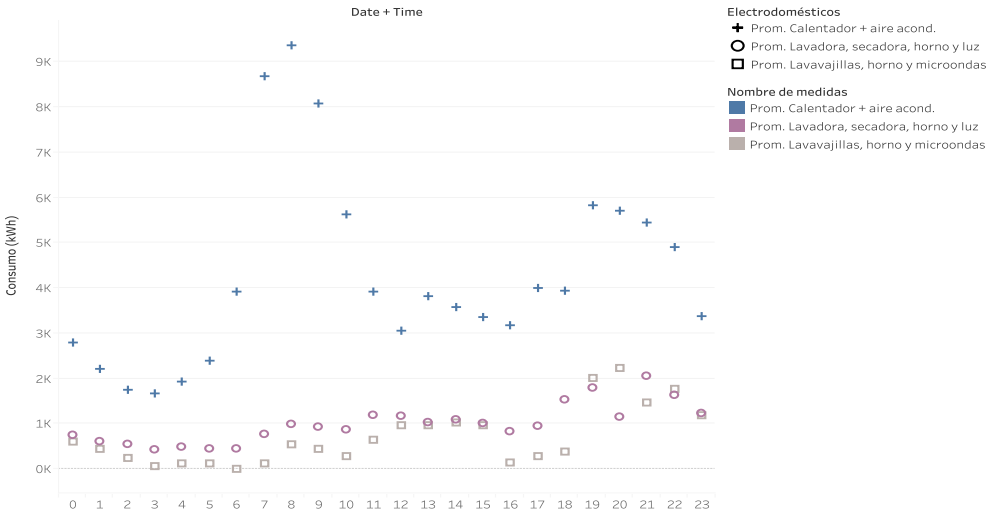


Figura 29. Consumo de los tres grupos de electrodomésticos a lo largo de un día

Del gráfico podemos observar:

- Sub_metering 3: El calentador de agua y el aire acondicionado (símbolo + en color azul), presentan un consumo considerablemente por encima del resto de electrodomésticos. Además, el consumo se dispara en los picos de consumo de primeras horas de la mañana (7-8 h), lo que nos da otra pista acerca de por qué el consumo aumentaba en esta franja horaria.
- Sub_metering 1 y Sub_metering 2: El uso de los otros dos grupos (círculo púrpura vs cuadrado gris claro) muestra un uso muy similar a lo largo del día, sufriendo un aumento a medida que avanza el día.
  Aunque no se aprecia con exactitud, se pueden apreciar ciertas diferencias a partir de las 15:00 h entre ambos grupos.

Dado que en el anterior apartado no se han podido leer esas diferencias entre las submedidas 1 y 2 correspondientes a diferentes aparatos de la casa [véase apartado 3.1.2]. Representamos estas de manera independiente para ver cuando se produce su mayor consumo a lo largo de los diferentes días promedio de cada mes:

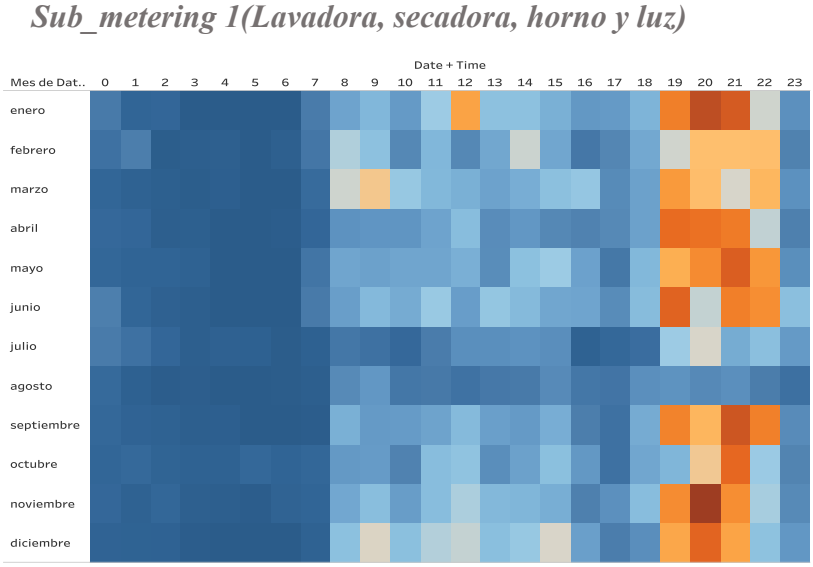


Figura 30. Consumo de un grupo de electrodomésticos (Submetering1) por mes a lo largo de un año promedio

Observamos que las horas de mayor consumo se corresponden con los periodos:

- Entre las 19:00 y las 21:00 horas

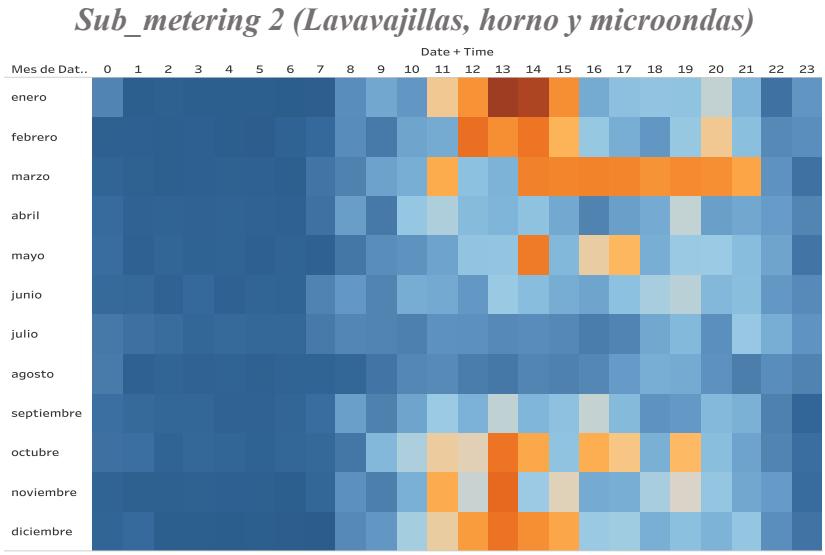


Figura 31. Consumo de un grupo de electrodomésticos (Submetering2) por mes a lo largo de un año promedio

En este caso, las horas de mayor consumo se corresponden con los periodos:

- Entre las 11:00 y las 14:00 horas

Calentador de agua vs Aire acondicionado

Los datos del aire acondicionado y el calentador del agua han sido medidos en la misma variable, por lo que no somos capaces de observar cómo es la relación entre ellos.

Actualmente existen aires acondicionados que disponen de bomba de calor y que permiten al usuario tanto enfriar como calentar su casa. Al mismo tiempo se espera que el calentador de agua se use algo más durante los meses de invierno que el resto de estaciones.

En este caso, lo que tratamos de averiguar no es como influye la variable en el consumo horario (eso ya lo estudiamos con el anterior gráfico), sino determinar cuanta energía consumimos en los dos meses en los que más hacemos uso de la variable: Enero (mes más frío) y Agosto (mes más caluroso), para estudiar si el aire acondicionado influye en nuestro consumo.

Representamos esto en un gráfico de forma para comparar su uso entre los dos meses: enero (círculo azul) y agosto (cruz naranja)[2]. Se filtran algunos años para comprobar la regularidad, obteniendo:

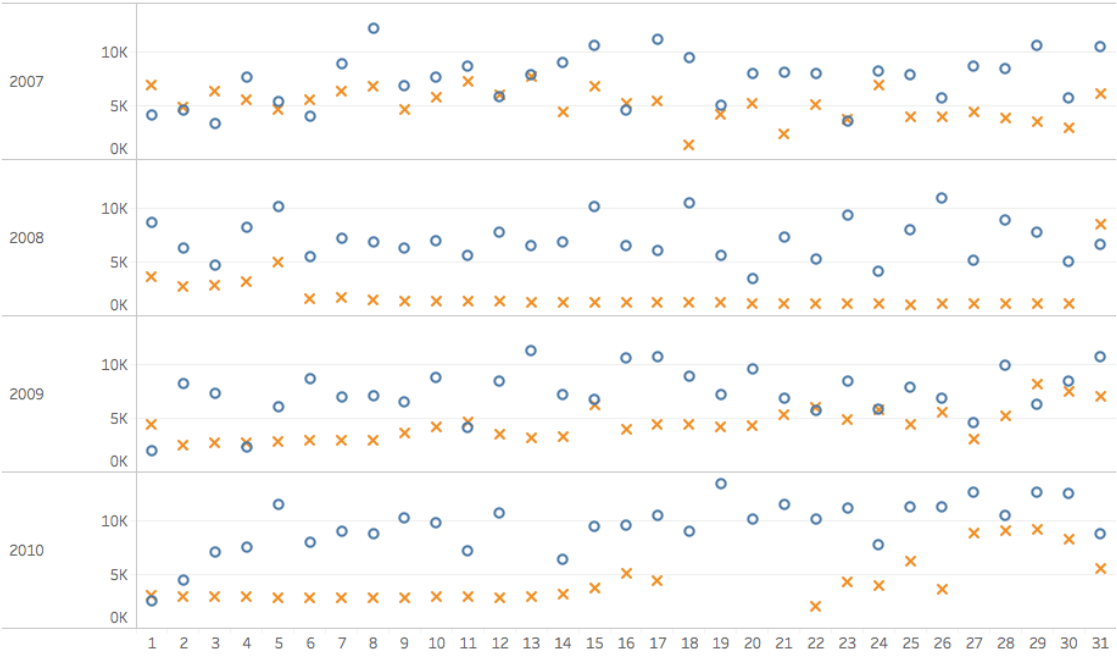


Figura 32. Consumo de calentador de agua + aire acondicionado en los meses de enero y agosto

Ante la distribución obtenida observamos que el consumo en verano no es superior a ninguno de los días de los meses de invierno.

Por otro lado, mientras en invierno el consumo es más irregular a lo largo de los días, en general en verano la dispersión se mantiene constante.

[2] En este gráfico se ha decidido no incluir la leyenda para que se observe con mejor precisión la distribución temporal. De ahí que lo especifiquemos antes de su representación.

3.4 Estimación del consumo eléctrico para el próximo año

Tableau ofrece la posibilidad de análisis de pronóstico de series temporales basadas en suavización exponencial lineal si el número de datos es lo suficientemente grande[3].

Para poder realizar una predicción en Tableau, el programa requiere al menos de una dimensión temporal y un tipo de medida. En nuestro caso, como variable temporal se ha elegido la variable *Fecha* en modo MES, para representar los valores de cada mes a lo largo de los años, y como medida, la variable de estudio del consumo activo global.

Como tipo de gráfico usaremos el de línea, ya que permite representar con gran impacto, cantidades de datos que tienen lugar durante un período continuado de tiempo

De esta manera obtenemos la siguiente representación:

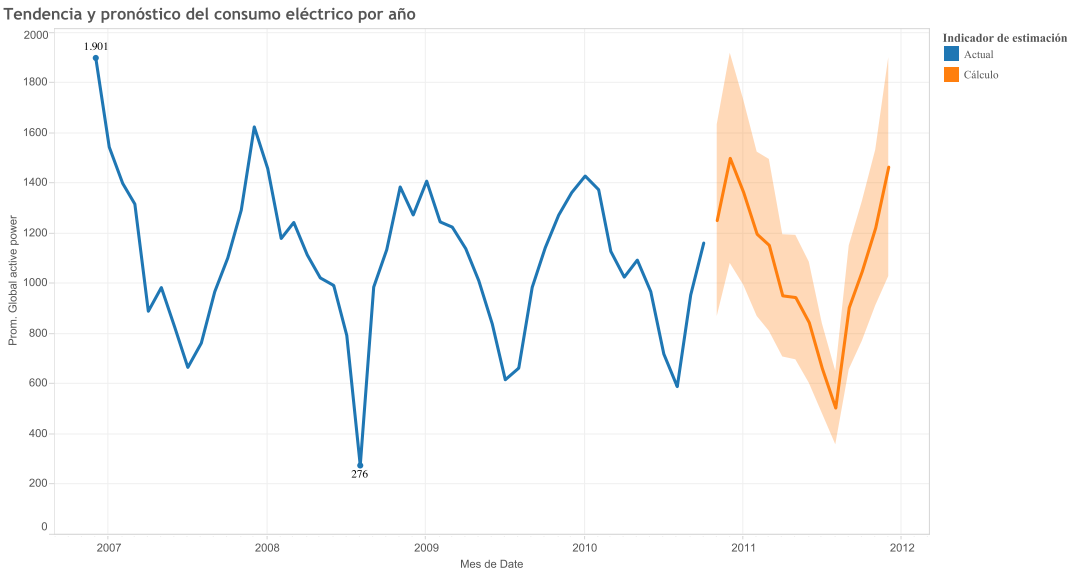


Figura 33. Predicción de consumo estimado para la serie temporal

Como podemos observar, la tendencia del consumo estimado (en color naranja) presenta una evolución similar al resto de los años. La predicción refleja un consumo ligeramente mayor al año anterior (2010).

Aunque este cálculo nos permite ver una posible opción del comportamiento futuro de nuestros datos, la franja del indicador de estimación es demasiado ancha como para establecer conclusiones fiables. Como indicamos al inicio del trabajo, en el mercado energético es clave saber con la mayor exactitud posible los valores que se manejan, por lo que en el siguiente apartado se van a aplicar una serie de modelos para la predicción con el fin de obtener un mejor resultado.

[3] Dado que vamos a llevar a cabo una etapa completa de análisis de modelos predictivos, no nos vamos a detener a explicar en que se basa la suavización exponencial lineal, ya que además no vamos a considerar esta técnica para nuestro trabajo.

4. Etapa de modelos

4.1. Modelos Univariantes y Multivariantes

En este apartado analizamos los modelos utilizados para discutir la capacidad predictiva. Para ello, nos basaremos en un conjunto de estos, los univariantes y multivariantes, que analizan de forma distinta las correlaciones temporales de las series.

Como variable de estudio, en esta etapa haremos sólo uso de la potencia activa global (Global_active_power), ya que es el atributo que determina la medida total de consumo.

4.1.1 Modelo Univariante

A continuación explicamos la metodología general de los modelos univariantes de series temporales (modelos ARIMA) siguiendo como referencia básica Peña, 2005.

Definamos x_t una serie temporal como una realización concreta del proceso estocástico $\{X_t\}_{t=0}^{\infty}$ tal que $x_t \in X_t$. Un modelo univariante (con componente estacional "s"), el cual pretende recoger las correlaciones temporales más sobresalientes, se define en forma compacta (esto es, usando el polinomio característico[4]) como:

$$\emptyset(L)\Phi(L)\ \Delta\Delta_s\ x_t = \theta(L)\Theta(L)\epsilon_t$$

Donde se pueden obtener los siguientes componentes:

- $\boldsymbol{\emptyset(L)}$ es el polinomio autorregresivo regular. Puede escribirse de forma menos compacta como $(1 - \phi_1 L - \phi_2 L^2 - \cdots)$ de tal forma que las raíces de este polinomio característico son todas, en módulo, mayores que 1 como condición básica de estacionariedad.
- $\boldsymbol{\Phi(L)}$ es el polinomio autorregresivo estacional de periodo "s". De la misma manera que en el caso anterior, este polinomio se desarrolla como sigue: $(1 - \Phi_1 L^s - \Phi_2 L^{2s} - \cdots)$ donde ha de cumplirse, igualmente, que las raíces de este polinomio característico sean mayores que la unidad en módulo.
- $\boldsymbol{\Delta\Delta_s}$ representan los operadores de diferencias. En nuestro caso es común necesitar, a lo sumo, una diferencia regular $\Delta = 1 - L$ y una diferencia estacional $\Delta_s = 1 - L^s$.
- $\boldsymbol{\epsilon_t}$ representa el término de error del modelo, el cual por hipótesis será independiente e idénticamente distribuido (i.i.d.), con media cero y varianza constante σ^2.
- Los términos finales aluden a la parte media móvil del modelo. $\theta(L)$ se corresponde con la parte media móvil regular: $\theta(L) = (1 + \theta_1 L + \theta_2 L^2 + \cdots)$ y, por otro lado, $\Theta(L) = (1 + \Theta_1 L^s + \Theta_2 L^{2s} + \cdots)$ se corresponde con la parte media móvil estacional.

[4]Resulta destacado notar que el operador de retardos L funciona de esta manera sobre variables con subíndice temporal: $L^r x_t = x_{t-r}$

De nuevo, los requisitos para estos polinomios son que sus raíces asociadas al polinomio característico sean mayores, en módulo, que la unidad. De esta manera, se permite la característica de la invertibilidad útil en la predicción de procesos estacionarios de una manera parsimoniosa.

4.1.2 Modelo Multivariante

Un modelo multivariante es una generalización m-dimensional de un modelo ARIMA. Se puede definir utilizando la siguiente notación. Llámese Y_t al siguiente vector:

$$Y_t = \begin{bmatrix} x_{1,1} \\ \dots \\ x_{1,T} \\ x_{2,1} \\ \dots \\ x_{2,T} \\ \dots \\ x_{M,1} \\ \dots \\ x_{M,T} \end{bmatrix}$$

Donde se agrupan M series temporales en un vector columna. Se define un modelo VAR(p)[5] como un modelo Vectorial Autorregresivo de forma que:

$$\phi(L)Y_t = \epsilon_t$$

Donde $\phi(B)$ es un polinomio matricial de p retardos retardos con la misma estructura que en caso AR(p), salvo que ahora ϕ es un conjunto de matrices que representan los pesos con el que el pasado conjunto de todas las variables se afectan entre sí. Por otro lado, ϵ_t es un proceso de ruido aleatorio M-dimensional donde:

$$cov(\epsilon_t) = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \dots & \sigma_{1,M} \\ \sigma_{1,2} & \dots & \dots & \sigma_{2,M} \\ \dots & \dots & \dots & \dots \\ \sigma_{1,M} & \sigma_{2,M} & \dots & \sigma_M^2 \end{bmatrix}$$

Es decir, se permite que tengan correlación contemporánea entre ellos y matriz de varianzas - covarianzas no escalar. Por otra parte, la condición de estacionariedad, implica:

$$|I - \phi_1 L - \phi_2 L^2 - \cdots| = 0$$

Es decir, el determinante de este polinomio matricial ha de ser cero (lo que implica que las raíces deben ser mayores, en módulo, que la unidad).

Es decir, un modelo VAR es una representación lineal de la distribución conjunta de un proceso en el pasado. Utiliza, por tanto, las interrelaciones entre variables para mejorar la

[5] Un modelo VAR(p) es un modelo deliberado que ajusta un caso más general denominado VARMA(p,q), heredado de la notación ARMA. Sin embargo, en la práctica son más utilizados (dada la sencillez) los modelos VAR puros. (Hamilton, 1994)

habilidad predictiva de los modelos ARMA (que sólo se basan en su propio pasado). Del mismo modo que en los modelos SARIMA (los estacionales), se pueden diferenciar las series con el objeto de tener una representación estacionaria del proceso[6].

$$\phi(L)\Delta\Delta^s Y_t = \epsilon_t$$

El interés que tiene, en nuestro caso, es el uso de la interacción entre momentos temporales para la mejora de la previsión del consumo futuro. Esta interacción reducirá error de previsión al suponer un mayor incremento del conocimiento de la dinámica del consumidor.

Sin embargo, los modelos VAR tienen el inconveniente de requerir un alto número de parámetros para su estimación. Como solución a esto, se deben buscar estructuras sencillas que permitan obtener un compromiso entre sencillez paramétrica (parsimonia) y mejoras en el proceso.

Modelos Multivariantes: Utilización del método de componentes principales (PCA)

Es por ello que, en este caso, utilizaremos el método de los componentes principales para poder determinar conjuntos de horas con comportamiento similar. Si no hiciéramos esto, nos encontraríamos que el modelo VAR tendría 24 variables por lo que, en caso de estimar un modelo VAR(1), la matriz de parámetros tendría 24x24. Esta matriz elimina grados de libertad a las estimaciones del modelo y representa una estructura "sobreparametrizada". Para evitar este problema, que afectara a la precisión con que se obtienen resultados, trataremos de reducir el número de parámetros necesarios.

El método de componentes principales se basa en la matriz de correlación que surge al computar:

$$corr(Y_t) = corr\left(x_{it}, x_{jt}\right) = corr\begin{bmatrix} x_{1,1} \\ \dots \\ x_{1,T} \\ x_{2,1} \\ \dots \\ x_{2,T} \\ \dots \\ x_{M,1} \\ \dots \\ x_{M,T} \end{bmatrix} = \begin{bmatrix} 1 & \rho_{1,2} & \dots & \rho_{1,M} \\ \rho_{1,2} & \dots & \dots & \rho_{2,M} \\ \dots & \dots & \dots & \dots \\ \rho_{1,M} & \rho_{2,M} & \dots & 1 \end{bmatrix}$$

De tal forma que obtenemos una medida estandarizada del grado de relación entre las M unidades temporales (en este trabajo se considerarán horas, en vez de minutos).

[6] Nótese que hay una amplia literatura sobre cointegración y, por tanto, representaciones estacionarias de series que no lo son, pero cumplen la propiedad de equilibrio a largo plazo conjunto. Como nuestro objetivo es predecir, seguimos la literatura propuesta por (Clements y Hendry (1995)) en la que se muestra las propiedades positivas de estabilidad en la capacidad predictiva de modelos, quizás, sobrediferenciados.

La idea del método consiste en construir una serie de "componentes" de tal forma que resuma la información facilitada de entrada.

$$z_1 = v_{11}x_{1t} + \cdots + v_{1M}x_{Mt}$$

$$\ldots$$

$$z_M = v_{M1}x_{1t} + \cdots + v_{MM}x_{Mt}$$

Sin embargo, la utilización de este método (basado en los autovalores de la matriz de correlación) permitirá restringir eso a un conjunto de ecuaciones con "p" componentes donde p<M:

$$z_1 = v_{11}x_{1t} + \cdots + v_{1M}x_{Mt}$$

$$\ldots$$

$$z_p = v_{p1}x_{1t} + \cdots + v_{pM}x_{Mt}$$

De tal forma que podamos obtener las series temporales que deben unirse mediante estos componentes para reducir la dimensión del problema de estimación[7].

4.2. Estrategia de modelización

La estrategia de modelización habitual de estos procesos estocásticos sigue las ideas obtenidas desde Box-Jenkins (Box y Jenkins, 1970 hasta Box et al. 2015) pasando por la inclusión de test que aseguren ciertas características como la necesidad de diferenciar (Dickey y Fuller, 1981). Primeramente mostramos el diagrama de flujo en la lógica de la modelización de series temporales univariantes. Este esquema fue desarrollado por Box-Jenkins:

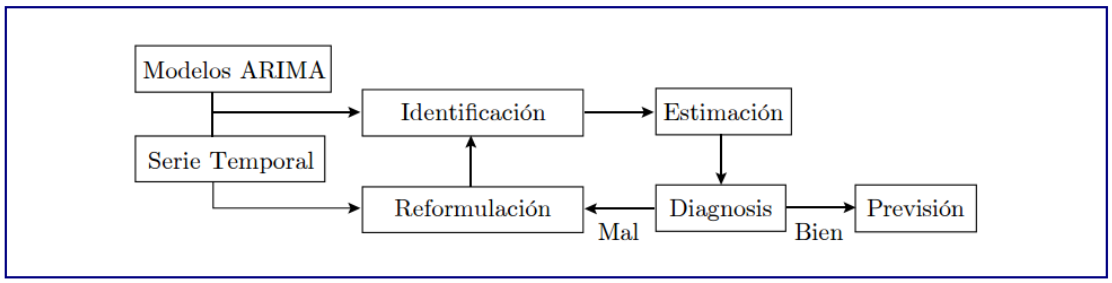


Figura 34. Proceso de construcción de un modelo ARIMA univariante

El esquema, básicamente insiste en la necesidad de establecer una identificación del modelo más adecuado. En dicha identificación es preciso indagar sobre las propiedades más

[7] Posteriormente se mostrará un ejemplo con los datos del caso.

destacadas de las series temporales: orden de integración[8], existencia de heterocedasticidad (generalmente multiplicativa), tipo de estacionalidad y modelo tentativo.

Para detectar el orden de integración es aconsejable la realización de test de raíz unitaria. Dichos test, desarrollados en Dickey y Fuller (1981), se basan en esencia en la siguiente ecuación de regresión:

$$\Delta x_t = \rho x_{t-1} + \epsilon_t$$

Donde la hipótesis nula de raíz unitaria (esto es, la serie necesita una diferencia regular adicional) , o lo que es lo mismo, el proceso tiene una tendencia estocástica (o es integrado de orden 1, es decir, I(1)) se testea mediante:

$$Ho: \rho = 0$$

En cambio, si se rechaza dicha hipótesis, la alternativa implica estacionariedad del proceso estocástico[9]:

$$H1: \rho < 0$$

Este test se lleva a cabo mediante un ratio similar al que se utiliza en los modelos de regresión para testar la significatividad individual de una variable concreta:

$$\tau = \frac{\hat{\rho}}{\widehat{dt(\rho)}}$$

Sin embargo, dicho ratio no se distribuye de manera estándar y requiere acudir a la distribución Dickey-Fuller generada para dicho test.

Sin embargo, hay más versiones del test, dependiendo del tipo de tendencia de la serie temporal: con deriva (o constante) y con tendencia lineal. El primero se utiliza si la serie contiene asimismo un crecimiento determinista. El segundo si este crecimiento tiene un comportamiento no lineal.

En el caso de los modelos multivariantes, el procedimiento es similar (véase, por ejemplo, Hamilton, 1994). Sin embargo, en este caso, el orden del modelo VAR se determina estimando un conjunto de modelos con diferentes retardos y eligiendo, según criterios que posteriormente veremos, el retardo óptimo.

[8] Dicho orden indica el número de diferencias tanto regulares como estacionales que la serie necesita para ser estacionaria.

[9] Siempre que asumamos que el proceso, a lo sumo, puede ser I(1).

4.2.1 Etapa de identificación: modelo univariante

Identificación(I): La serie global

La serie temporal en estudio presenta el siguiente aspecto en el que, sobre todo, destaca el comportamiento estacional de esta, con ciclos perceptibles a lo largo del tiempo (presentamos la serie original en consumo por minuto y el logaritmo de dicha serie)

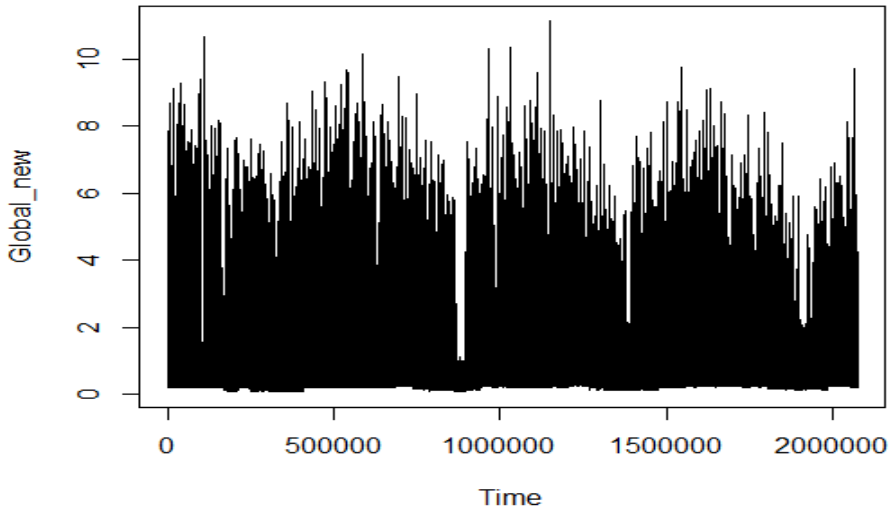


Figura 35. Serie temporal original en consumo por minuto

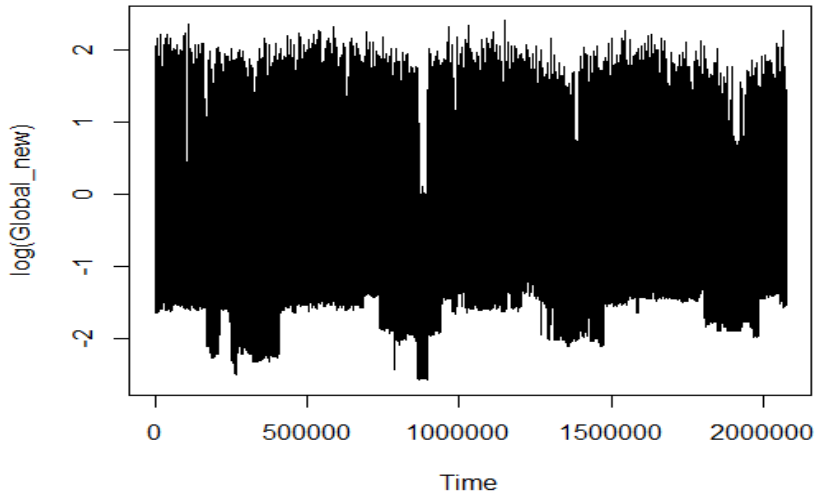


Figura 36. Serie temporal original en consumo por minuto (logarítmica)

Por un lado, resulta más complicado analizar la estacionariedad de dicha serie. Esta, al ser de frecuencia muy corta (al minuto) presenta la dificultad añadida de separar el componente estacional del resto. En lo que respecta al correlograma (figura siguiente):

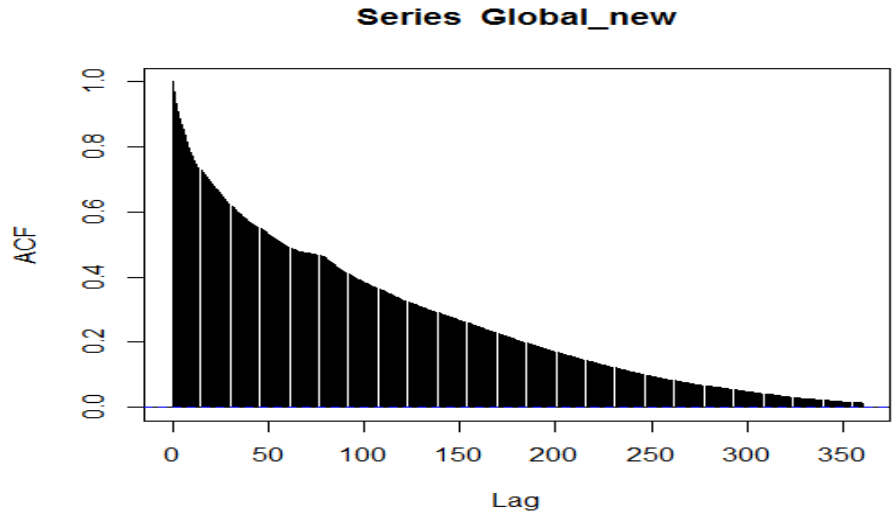


Figura 37. Función de autocorrelación simple (ACF) de la serie en función del retardo

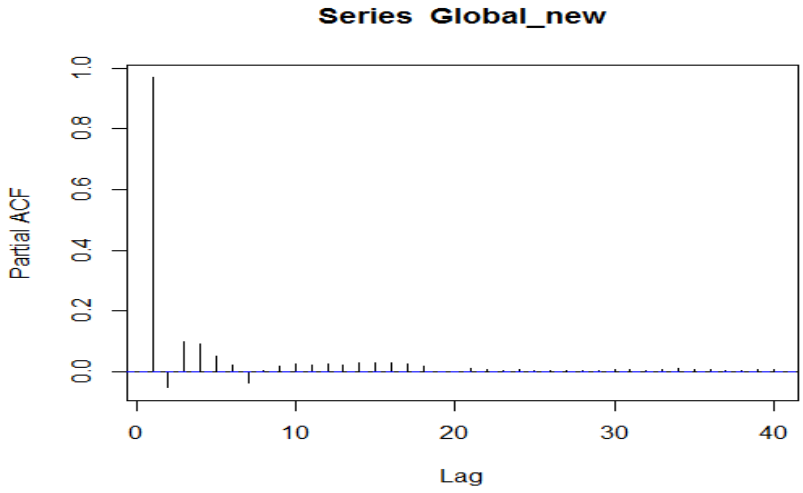


Figura 38. Función de autocorrelación parcial (PACF) de la serie en función del retardo

Observamos las características de una serie no estacionaria: la ACF muestra un decaimiento lento a lo largo de los retardos (todavía en el retardo 350 este no se ha anulado) y una PACF con un primer valor cercano a la unidad.

Además, se percibe más dinámica autorregresiva (tal y como se puede extraer del comportamiento de la PACF).

Por otro lado, se puede realizar el test de Dickey Fuller para dicha serie. Desafortunadamente, la dimensión de la estacionalidad de la serie y la propia serie hace que el test aumentado (con retardos) necesite más de 200 (que, en este caso equivalen a un desfase lógico de dos o tres horas) y esto implica que el propio R, en su función fUnitRoots de error de ejecución.

Asumiremos, por tanto, que la serie global tiene raíz unitaria y estimaremos un modelo ARIMA potencial utilizando la primera diferencia del logaritmo de la serie.

Identificación(II): Las series por minuto

Una de las complejidades mayores en este análisis de datos es el alto componente estacional que podemos encontrarnos (intra minuto, intra hora, semanal, mensual, etc…) con datos de tan corta frecuencia.

Una manera de atacar dicho problema puede consistir en "desagregarlo", es decir, generar series temporales alternativas y predecir cada una de estas series para, posteriormente, agruparlas en la frecuencia necesaria.

En este primer caso, separaremos las series por minuto, teniendo, entonces un conjunto de 1440 series temporales.

Código	**Explicación**
Ct1$Time<-as.numeric(Ct1$Time) Ct1$Global_active_power<- as.numeric(as.character(Ct1$Global_active_power)) D<-Ct1[Ct1[, 2] == 1,3] D<-ts(D, frequency=7)	La variable a predecir se denomina Global_active_power. Esta variable se encuentra en el set de datos "Ct1" que, asimismo, contiene información sobre la fecha y la hora. Transformamos la hora a un número, de tal manera que quede etiquetada de 1 a 1440 (minutos en un día) Inicializamos D, que será un acumulador por columnas de cada minuto. Inicializamos con el minuto "1".

for (i in 2:1440) {
 	
x<-Ct1[Ct1[, 2] == i,3]
 	
x<-ts(x, frequency=7)
 	
D<-cbind(D,x)
 	
}
	Este bucle va del minuto 2 al minuto 1440, acumula los minutos (en vectores columna) utilizando "cbind" que anexiona vectores.

Las 1440 series por minuto obtenidas en este código tendrán propiedades diferentes a las observadas en la serie global. A continuación, representamos algunos de esos minutos para que se puedan, a modo de ejemplo, ver dichas propiedades de las series:

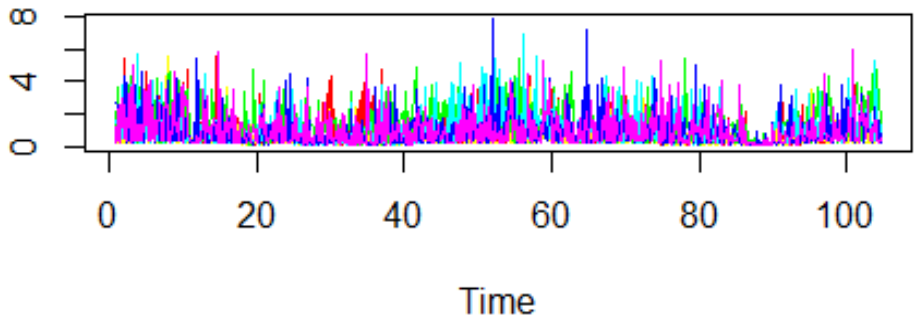


Figura 39. Algunos minutos del conjunto de series temporales alternativas generadas

Representamos el minuto 1, el minuto 100, el minuto 500, el minuto 800, el 1000 y el 1400 de todos los días con diferentes colores. Encontramos que, en este caso, parecen procesos estacionarios en media (aunque no en varianza, por lo que necesitarán el logaritmo). Como vemos, tras esta transformación, las series son estacionarias en varianza:

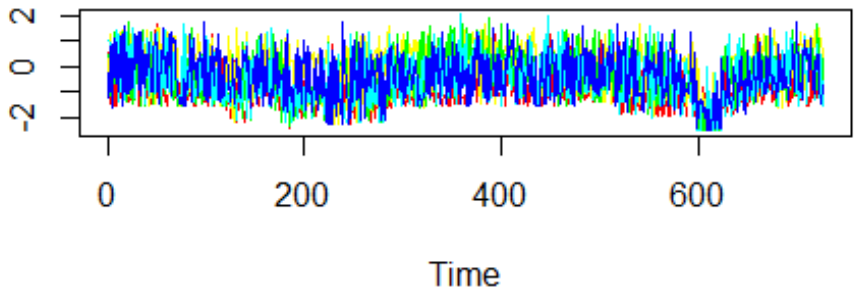


Figura 40. Algunos minutos del conjunto de series temporales alternativas generadas (logaritmo)

En lo que respecta al correlograma, observamos una pauta que se repite de forma habitual:

- Poca estructura en la parte regular (es decir, autorregresivos débiles, implicando una inercia ligera entre minutos)

- Mayor estructura en la parte estacional de periodo 7 (es decir, los consumos por minutos se suelen repetir en el mismo día de la semana). Mostramos, a continuación, los correlogramas de algunos de estos minutos (minuto 800 y minuto 1).

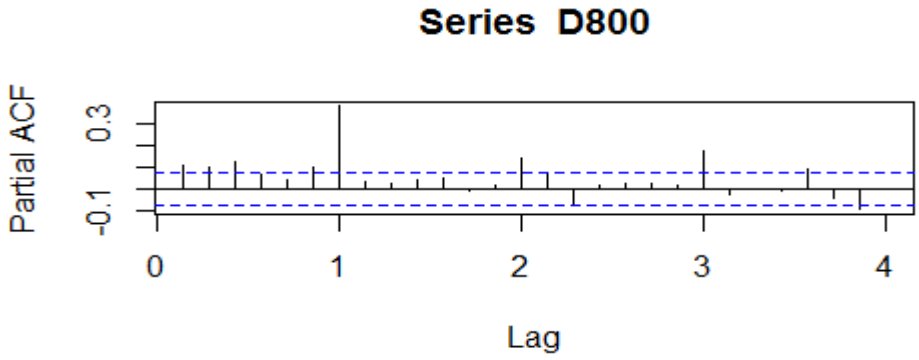


Figura 41. PACF del minuto 800 del conjunto de series temporales alternativas (generadas) en función del retardo

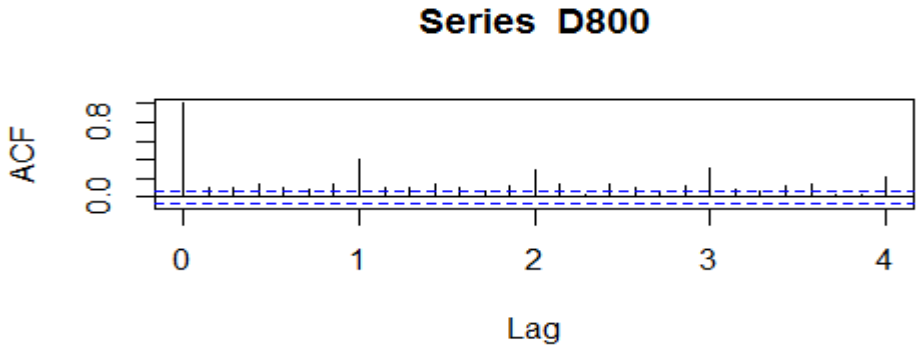


Figura 42. ACF del minuto 800 del conjunto de series temporales alternativas (generadas) en función del retardo

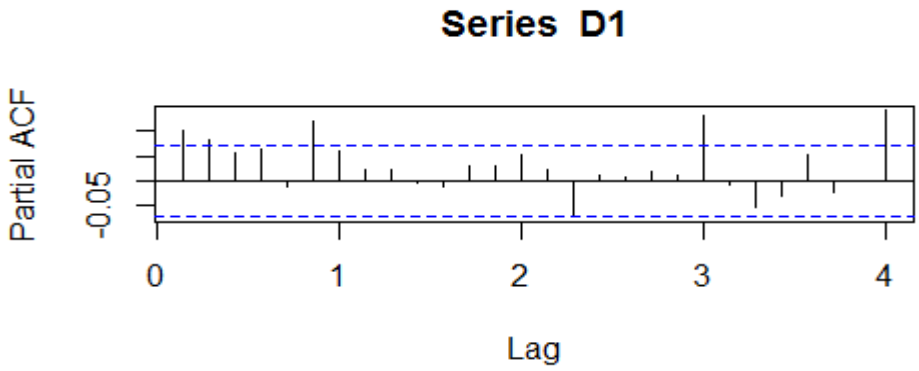


Figura 43. PACF del minuto 1 del conjunto de series temporales alternativas (generadas) en función del retardo

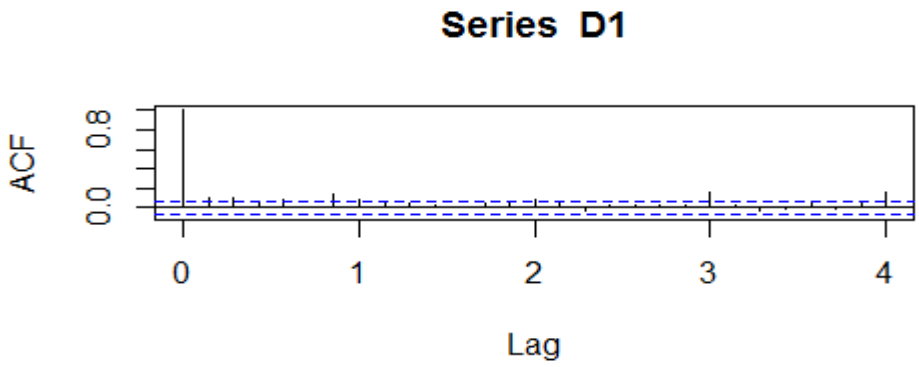


Figura 44. ACF del minuto 1 del conjunto de series temporales alternativas (generadas) en función del retardo

Identificación(III): Las series por hora.

Por otro lado, dado que nuestra discusión también se centra en la dicotomía "complejidad-capacidad predictiva", también analizamos otra agrupación más sencilla: el consumo por horas:

Código en R	**Explicación**
hr.means <- aggregate(data2["Global_active_power"], list(hour = cut(data2$date, breaks="hour")), mean, na.rm = FALSE) t<- strftime(hr.means$hour,format="%H:%M:%S")	Esta función (aggregate) junta la información de minutos a horas, agregando de 60 en 60. Adicionalmente, generamos un vector de "tiempos" por horas, ya que en el set inicial lo teníamos al segundo.
D2<-cbind(hr.means,t) D2$t<-as.numeric(D2$t) D2$Global_active_power<- as.numeric(as.character(D2$Global_active_power))	Generamos el nuevo data set y cambiamos el vector de tiempos por una ordenación de 1 a 24.
Dd<-Ct2[Ct2[, 3] == 1,2] Dd<-ts(Dd) for (i in 2:24) {	Este bucle, al igual que en el código minuto a minuto, organiza la información de tal

x<-Ct2[Ct2[, 3] == i,2] x<-ts(x) Dd<-cbind(Dd,x) }	manera que las columnas son las horas disponibles.

Etapa de estimación: modelo univariante.

Estimación (I) Las series por minutos.

La estrategia de estimación se realiza, como se puede ver en la figura 34 de procedimiento, a continuación de la identificación. En este caso, consiste en estimar, mediante métodos de máxima verosimilitud, el modelo ARIMA x SARIMA adecuado para obtener residuos ruido blanco (objetivo básico de la diagnosis).

En nuestro caso, estos procedimientos los automatizaremos, de tal forma que el modelo se estime siguiendo un criterio *stepwise*, esto es, eligiendo mediante un rastreo de modelos aquel que mejor ajusta la idea de "un compromiso" entre pocos parámetros y capacidad de ajuste. Uno de esos estadísticos es el Akaike (AIC) el cual se define como.

$$AIC(k) = 2k - 2\log(L)$$

Donde k representa el número de parámetros del modelo y L es la evaluación en la función de verosimilitud. Como puede verse, el criterio penaliza por modelos sobre parametrizados ya que, cuando menor sea el valor del estadístico, más preferible es el modelo.

A modo de ejemplo, y antes de explicar la implementación, podemos mostrar la estimación de uno de los segundos representados por el procedimiento automático.

Por ejemplo:

```
ARIMA(2,1,1)(1,0,1)[7]

Coefficients:
          ar1     ar2      ma1     sar1    sma1
      0.0779  0.0290  -0.9749  -0.6188  0.6438
s.e.  0.0423  0.0422   0.0110   0.4169  0.4051

sigma^2 estimated as 0.8027:  log likelihood=-782.99
AIC=1577.98   AICc=1578.12   BIC=1604.35
```

Figura 45. Estimación de uno de los segundos representados por el procedimiento automático

Este modelo, siguiendo la notación anterior, podría representarse como:

$$(1 - 0.08\mathrm{L} - 0.03\mathrm{L}^2)(1 + 0.61\mathrm{L}^7)\ln(D1_t) = (1 - 0.97L)(1 + 0.64L^7)\epsilon_t$$

Presentamos, a continuación, la estructura del código desarrollado para la estimación automática de los modelos.

Código en R	Explicación
k <- 600 n <- 720 st <- k error<-0 errorf<-0 real<-0 realf<-0	Declaro las variables iniciales: k ➔ histórico mínimo para poder estimar modelos n➔dimensión del histórico total disponible st➔ punto de inicio error,real➔ inicializo errores (error y errorf) así como un almacenador de datos reales (real, realf)

for (i in 1:1440) {
 	
p<-log(D[,i])
 	
for (j in 1:(n-k-1))
{
 	
xshort <- p[1:st+j]
xnext <- p[st+j+1]
xshort<-ts(xshort,frequency=7)
 	
fit2 <- auto.arima(xshort, d=1,max.p=10, max.q=10,max.P=5,max.Q=5, ic=c("aicc","aic", "bic"), stepwise=TRUE, allowdrift=TRUE)
fcast2 <- forecast(fit2, h=1)
	Bucle principal, "i" representan los minutos del día.
	Se obtiene el logaritmo de la demanda.
	Entramos en Bucle de estimación. J representa el histórico por el que moverse.
	Xshort coge el histórico hasta el punto st+j
	Xnext evalúa en el histórico el punto j+1 desde xshort.
	Se estima un modelo ARIMA automático con parte estacional.
	Se evalúa la previsión dada por el modelo ARIMA.

error[j]<-(exp(xnext[1])-exp(fcast2$mean[1])) real[j]<-(exp(xnext[1])) }	Se evalúa el error y se almacena así como el valor real consumido
errorf=cbind(errorf,error) realf=cbind(realf,real) }	Se almacena una matriz con los errores de tal forma que luego se puedan agregar por día.

4.2.2 Modelos Multivariantes: Identificación y Estimación.

Como antes se explicó, con la idea de estimar modelos que aprovechen la información conjunta proporcionada por la inercia a consumos similares en ciertas horas, tratamos de analizar, usando las correlaciones entre horas, cuáles de ellas son comunes. Primeramente, ofrecemos una matriz de correlaciones "visual":

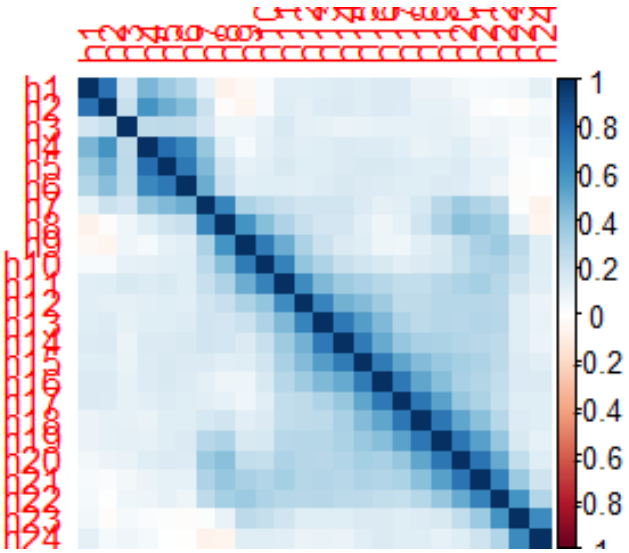


Figura 46. Matriz de correlaciones entre horas para el modelo multivariante

Los colores azules muestran las correlaciones de las horas. Esta matriz podría asimilarse a una matriz "banda", es decir, con elementos distintos de cero en la diagonal principal y en las dos diagonales contiguas. El resto de correlaciones parecen mucho más insignificantes.

El estudio de los posibles componentes principales denota la existencia de 12 clúster. Hemos reducido a la mitad el conjunto de variables para el modelo multivariante.

El dendograma siguiente muestra los clúster que pueden generarse mediante estos componentes:

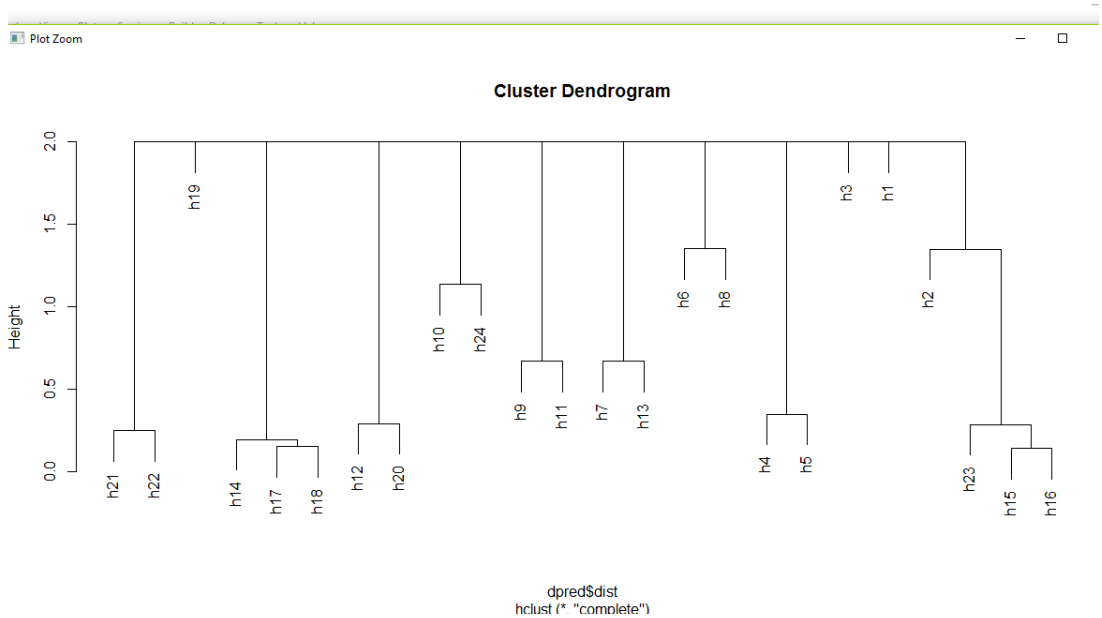


Figura 47. Dendograma de clúster que pueden generarse para el modelo multivariante

Lo cual nos permite agrupar las horas de esta manera:

Grupo 1: H21 y H22

Grupo 2: H19

Grupo 3: H14,H17, H18

Grupo 4: H12, H20

Grupo 5: H10,H24

Grupo 6: H9 H11

Grupo 7. H7 y H13

Grupo 8. H6 y H8

Grupo 9: H4 y H5

Grupo 10: H3

Grupo 11: H1

Grupo 12: H2, H23, H15 y H16

A continuación mostramos la manera en que hemos programado la estimación de los modelos multivariantes para, además, mostrar la estimación de uno de ellos:

Código en R	Explicación
dfmin<-data.frame(Dd)	Organizamos por columnas las series temporales por horas y estimamos, mediante el método de

colnames(dfmin)<- c("h1","h2","h3","h4","h5","h6","h7","h8","h9","h10","h11","h12","h13","h14","h15","h16","h17","h18","h19","h20","h21","h22","h23","h24") pca_a<-princomp(na.omit(dfmin), cor = FALSE)	componentes principales, la matriz de coeficientes para organizar dichos componentes
diffs<-rep(1,ncol(dfmin)) logs<-rep(TRUE,ncol(dfmin)) dpred<- diss(na.omit(dfmin),"PRED",h=1,B=1000,logarithms=logs,differences=diffs, plot=TRUE) hc.dpred <- hclust(dpred$dist) plot(hclust(dpred$dist))	Estas sentencias nos permiten obtener el dendograma que organiza los componentes principales en cluster
S1<-ts(dfmin["h19"]) S2<- rowSums(ts(data.frame(dfmin["h22"],dfmin["h18"],dfmin["h21"]))) S3<- rowSums(ts(data.frame(dfmin["h16"],dfmin["h17"]))) S4<- rowSums(ts(data.frame(dfmin["h13"],dfmin["h14"]))) S5<- rowSums(ts(data.frame(dfmin["h12"],dfmin["h15"]))) S6<- rowSums(ts(data.frame(dfmin["h9"],dfmin["h10"]))) S7<- rowSums(ts(data.frame(dfmin["h11"],dfmin["h8"],dfmin["h20"]))) S8<- rowSums(ts(data.frame(dfmin["h6"],dfmin["h7"]))) S9<- rowSums(ts(data.frame(dfmin["h5"],dfmin["h23"]))) S10<-ts(data.frame(dfmin["h3"]))	Este código permite obtener la suma de horas de acuerdo con los resultados obtenidos en el análisis por componentes principales

S11<- rowSums(ts(data.frame(dfmin["h2"],dfmin["h4"]))) S12<- rowSums(ts(data.frame(dfmin["h1"],dfmin["h24"]))) dataVAR=ts(data.frame(S1,S2,S3,S4,S5,S6,S7,S8, S9,S10,S11,S12))	

for(j in 1:(n-k))
{
xshort <- dataVAR[1:st+j,]
xnext <- dataVAR[st+j+1,]
 	
selec<-VARselect(xshort, lag.max = 15, type =
"both")
 	
fit2 <- VAR(xshort, p = selec$selection[1])
fcast2 <- predict(fit2, n.ahead=1)
 	
error1[j]<-(exp(xnext[1])-
exp(fcast2$fcst$h19[1]))
error2[j]<-(exp(xnext[2])-
exp(fcast2$fcst$s2[1]))
error3[j]<-(exp(xnext[3])-
exp(fcast2$fcst$s3[1]))
error4[j]<-(exp(xnext[4])-
exp(fcast2$fcst$s4[1]))
error5[j]<-(exp(xnext[5])-
exp(fcast2$fcst$s5[1]))
error6[j]<-(exp(xnext[6])-
exp(fcast2$fcst$s6[1]))
error7[j]<-(exp(xnext[7])-
exp(fcast2$fcst$s7[1]))
error8[j]<-(exp(xnext[8])-
exp(fcast2$fcst$s8[1]))
	Este código estima el modelo VAR y predice un paso adelante. Para ello, utiliza la función selec que, de acuerdo con el criterio AIC obtiene el mejor modelo VAR en cada momento. Los errores de previsión se obtienen por hora, para luego agregarse en la media diaria.

```
        error9[j]<-(exp(xnext[9])-
exp(fcast2$fcst$s9[1]))

        error10[j]<-(exp(xnext[10])-
exp(fcast2$fcst$h3[1]))

        error11[j]<-(exp(xnext[11])-
exp(fcast2$fcst$s11[1]))

        error12[j]<-(exp(xnext[12])-
exp(fcast2$fcst$s12[1]))

        real1[j]<-exp(xnext[1])

        real2[j]<-exp(xnext[2])

        real3[j]<-exp(xnext[3])

        real4[j]<-exp(xnext[4])

        real5[j]<-exp(xnext[5])

        real6[j]<-exp(xnext[6])

        real7[j]<-exp(xnext[7])

        real8[j]<-exp(xnext[8])

        real9[j]<-exp(xnext[9])

        real10[j]<-exp(xnext[10])

        real11[j]<-exp(xnext[11])

        real12[j]<-exp(xnext[12])


        }
```

4.3. Etapa de predicción: (Uso de los modelos para predicción y test de capacidad predictiva a un paso (un día)).

Uno de los objetivos de la elaboración de estos modelos es la predicción. Dado que las ecuaciones planteadas pertenecen al campo de las ecuaciones en diferencias estocásticas, resulta sencillo plantearlas, de forma recursiva, para realizar predicciones condicionales. Para simplificar la explicación, mostraremos la previsión con un modelo ARIMA(1,1,1) sin parte estacional, el cual resultará muy ilustrativo:

$$(1 - \emptyset L)(1 - L)x_t = (1 + \theta L)\epsilon_t$$

Este modelo, despejando la variable a predecir, se puede escribir como :

$$x_t = (1 + \emptyset)x_{t-1} + \emptyset x_{t-1} + \epsilon_t + \theta\epsilon_{t-1}$$

Y, finalmente, obtener la ley de recurrencia. En este caso, nos interesará siempre el análisis a un paso adelante, es decir, con horizonte temporal $h = 1$:

$$x_{t+h} = (1 + \emptyset)x_{t-1+h} + \emptyset x_{t-1+h} + \epsilon_{t+h} + \theta\epsilon_{t-1+h}$$

Denotaremos I_t el conjunto de información (esto es, la información disponible el sistema sobre x en el momento temporal "t").

Definiremos la previsión mediante el operador esperanza condicional:

$$E(x_{t+h}|I_t) = (1 + \emptyset)E(x_{t-1+h}|I_t) + \emptyset E(x_{t-2+h}|I_t) + E(\epsilon_{t+h}|I_t) + \theta E(\epsilon_{t-1+h}|I_t)$$

De tal forma que, usando estos resultados (véase, por ejemplo, Peña, 2005), para h=1:

$$E(x_t|I_t) = x_t$$

$$E(x_{t-1}|I_t) = x_{t-1}$$

$$E(\epsilon_{t+1}|I_t) = 0$$

$$E(\epsilon_t|I_t) = \hat{\epsilon}_t$$

Por lo que, podemos usar la ecuación recursiva para realizar predicciones:

$$E(x_{t+1}|I_t) = (1 + \emptyset)x_t + \emptyset x_{t-1} + \theta\hat{\epsilon}_t$$

Esta línea de razonamiento se puede aplicar de acuerdo a los diferentes modelos ARIMA que necesiten ser especificados (incluyendo la parte de estacionalidad como un añadido pero cuya álgebra es similar a la aquí expuesta).

Por otro lado, la previsión con modelos VAR es similar a lo expuesto anteriormente, puesto que la notación puede verse como un proceso vectorial. La predicción de la variable de interés, $E(Y_{t+1}|I_t)$ requiere del conocimiento del estado de dicho vector en el momento anterior:

$$Y_t = \begin{bmatrix} x_{1,t} \\ x_{2,t} \\ \dots \\ x_{M,t} \end{bmatrix}$$

De la misma forma que se ha explicado en los modelos univariantes.

Lo siguiente que debemos tener en cuenta es la evaluación de la capacidad predictiva del modelo. Utilizaremos procedimientos sencillos de validación cruzada (véase Arlot y Celisse, 2010) de tal forma que tratemos de utilizar la muestra con fines de aprendizaje-evaluación.

De esta manera, dado que la muestra tiene suficientes observaciones, podemos utilizar parte de dichos valores (ordenados cronológicamente) para estimar el modelo y, otra parte para cotejar la previsión que haría el modelo frente a lo que en realidad ocurrió.

En este trabajo sólo estamos preocupados en la previsión "un paso adelante" es decir, conociendo el valor de un día, tratar de predecir el siguiente.

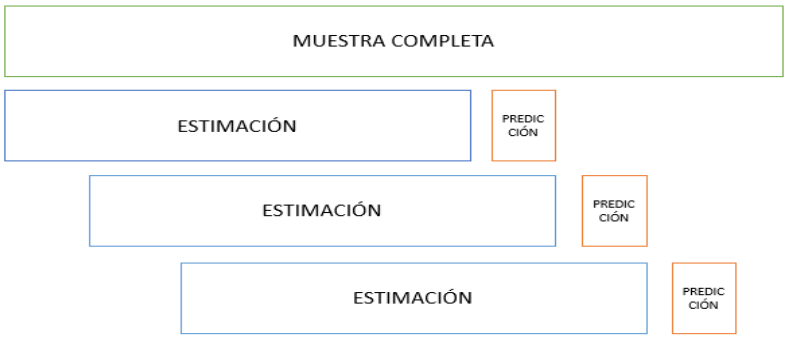


Figura 48. Esquema de validación cruzada de los modelos

De acuerdo con esta figura, la manera de comprobar el modelo será mediante el muestreo de los datos disponibles. Del total de la muestra, se seleccionará un conjunto de información para hacer la previsión y otro para validarla (no utilizado en el proceso de estimación previo). A continuación se computará el error de previsión a un paso adelante:

$$\epsilon_{t+1} = x_{t+1} - E(x_{t+1}|I_t)$$

Acumularemos esta información de tal forma que, a posteriori, podamos tener una distribución del error para diferentes submuestras (contextos). Computaremos, asimismo, el error porcentual absoluto medio (MAPE) con el objeto de facilitar una medida relativa:

$$MAPE = \frac{\sum \frac{|\epsilon_{t+1,i}|}{x_{t+1,i}}}{n}$$

Donde hay "n" submuestras para las que se calcula dicho estadístico. Finalmente, podremos hacer un estudio de error de previsión para el propio modelo y entre los modelos correspondientes.

4.4. Resultados.

Analizamos, primeramente, la distribución de los errores por cada uno de los métodos y conjuntamente, a posteriori. Para que sean comparables, los métodos muestran los errores de predicción agregados por horas de tal manera que la magnitud final de interés es la predicción de demanda diaria.

4.4.1 Análisis del error de predicción (por cada uno de los métodos)

Método univariante minuto a minuto, colapsado por horas.

En este caso, el error parece ligeramente asimétrico a la derecha, esto es, tendiendo a infrapredecir (esto es, el error se calcula como REAL-PREDICCIÓN, por lo que valores

positivos del error indican una infrapredicción y viceversa). El valor medio del error de predicción está en torno a 425 y el valor mediano, algo inferior: 395.

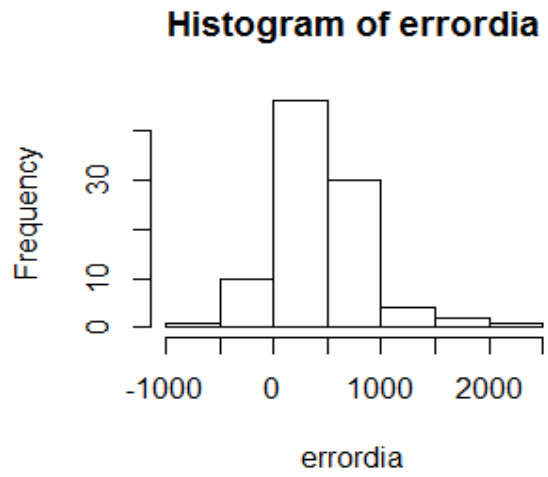


Figura 49. Error medio de predicción para el método univariante por minuto

El mínimo error es -598 (es decir, en una hora previó en exceso 598 unidades), mientras que el máximo es de 2154 unidades. El valor del coeficiente de asimetría es 0.96, el cual habrá que poner en comparación con el resto de métodos.

Modelo univariante hora a hora

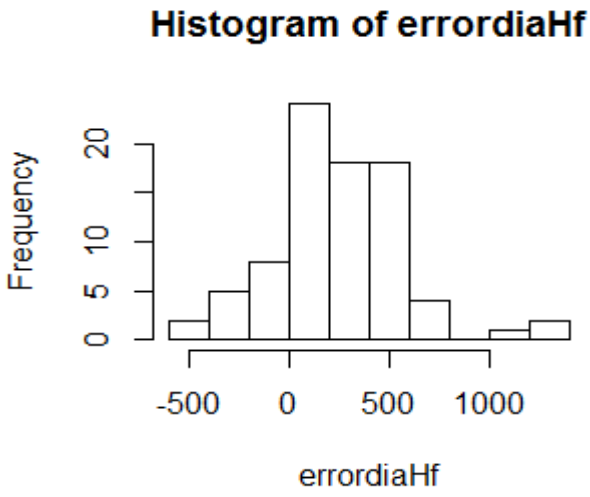


Figura 50. Error medio de predicción para el método univariante por hora

La estimación de un univariante con datos agregados a la hora, parece mejorar los resultados en torno a la varianza del error de predicción, puesto que los valores máximos y mínimos son menores. Ahora el valor mínimo es -462.66 y el máximo es 1309. El coeficiente de asimetría se ha reducido a 0.70 mientras que el error medio es 232 y el mediano 238 unidades. Parece, además, que hemos reducido asimetría (es decir, sesgo en el error) utilizando este método.

Modelo multivariante con clúster de horas

Este modelo es el que proporciona errores más simétricos pero, nuevamente, con mayor dispersión. El mínimo son -1244 unidades y el máximo 1384. La asimetría es la menor (en torno a 0.24) y la media es -13.15 (con mediana muy similar).

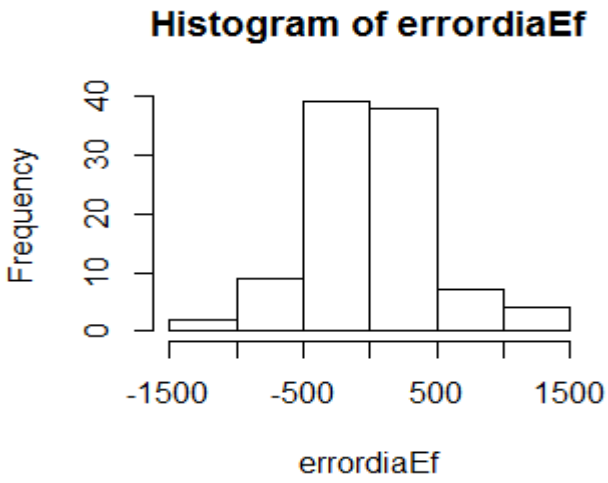


Figura 51. Error medio de predicción para el método univariante con clúster de horas

4.4.2 Valoración conjunta del error de predicción.

A continuación, analizamos los errores cometidos de forma conjunta. En rojo representamos los errores del método 1, es decir, el modelo univariante por minutos, en verde representamos el método univariante por horas y en azul representamos el modelo multivariante con horas agrupadas.

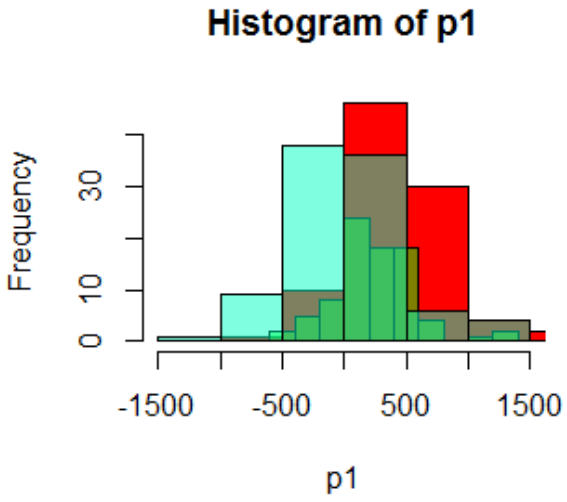


Figura 52. Errores medios cometidos de forma conjunta en los tres modelos

La impresión que obtenemos de este gráfico es que el modelo univariante por minutos tiene ciertas diferencias con respecto al resto. La asimetría a la derecha (en nuestro caso, interpretable como el sesgo a infrapredecir) parece ser, claramente, mayor que la del resto de modelos. El modelo multivariante y el univariante por horas podrían solaparse, aunque el univariante por horas parece predecir con menor dispersión, lo cual sería preferible al reducir el riesgo.

	Univariante por minutos	Univariante por horas	Multivariante horas agrupadas
MAPE	0.2964	0.2052	0.2075

Con respecto a una magnitud "resumen" del error de predicción, el univariante por minutos (que podría considerarse un modelo demasiado costoso en tiempo de cómputo y con demasiada desagregación) tiene el mayor error absoluto porcentual (cerca de un 30% de error diario). Por otro lado, el univariante por horas (más sencillo que el multivariante por horas) tiene un error ligeramente inferior al multivariante (en torno al 20.5% versus el 20.75%) sin embargo, puede concluirse que la utilización de un modelo más sofisticado no facilita una ganancia predictiva clara (y , además, proporciona errores de predicción más dispersos, lo que incrementa las medidas de riesgo).

5. Roles, planificación y presupuesto.

5.1. Rol de los participantes.

A continuación se van a enumerar los roles de trabajo que se han desarrollado durante el trabajo teniendo en consideración las funciones desempeñadas:

Análisis:

- **Analista de datos**: responsable de la etapa de exploración. El análisis requiere de una serie de conocimientos de manejo de software, en este caso Tableau, así como de una cierta capacidad de análisis para su interpretación. Estas habilidades son desempeñadas con éxito por un analista de datos.
- **Científico de datos**: responsable de la etapa de aplicación de modelos y código en R. Debido a que este proceso requiere de cierta experiencia y conocimientos estadísticos, sería apropiado contar con un científico de datos para que los resultados obtenidos sean correctos y de esta manera se pueda confiar en ellos

Inteligencia de Negocio:

- **Ingeniero industrial**: responsable del departamento de Inteligencia de Negocio y Analytics y persona a cargo del proyecto de la empresa. Este profesional desempeña dos tareas principales:
    - Estudio de los resultados obtenidos para la toma de conclusiones acerca del perfil que han generado los datos.
    - Planteamiento de trabajos futuros.

5.2. Planificación.

En este apartado se va a mostrar la planificación seguida para la elaboración del proyecto así como las tareas que la componen, indicando su secuenciación y temporización.

Debido a que se trata de un proceso de análisis de datos, resulta muy difícil planificar desde el inicio ciertas tareas críticas como pueden ser la etapa de generación de gráficos o la programación de código en R para la etapa de modelos. A su vez, la planificación se ha visto en algunos casos influenciada por su compaginación con prácticas extracurriculares en empresa.

Por ello cabe destacar que la distribución inicial de tareas se ha ido modificando a lo largo de su desarrollo, dependiendo de las necesidades que han ido surgiendo. Más adelante mencionaremos las tareas que han sido críticas en la consecución de las diferentes etapas.

Respecto a la organización, desde el inicio del proyecto se hizo saber al tutor la distribución general del trabajo, así como la consulta de cambios y modificaciones.

Como una propuesta para mejorar el diagrama de planificación, se ha decidido dar un color a cada profesional que ha participado en el trabajo [véase apartado roles], asignándole un color diferente, siendo:

- Ingeniero industrial: color naranja
- Analista de datos: color azul
- Científico de datos: color verde

A continuación se muestra la planificación seguida[10]:

Tabla 2. Planificación del proyecto

Tarea	Comienzo	Duracion (días)	Fin
Decisión de la idea y objetivos principales	1-7	7	8-7
Búsqueda e interpretación del conjunto de datos	9-7	3	12-7
Memoria: Estado del arte	13-7	18	31-7
Conexión a Tableau y preparación de los datos	20-7	3	23-7
Diseño de gráficos de visualización (Tableau)	24-7	17	10-8
Documentación acerca de modelos predictivos	1-7	4	5-7
Lectura e interpretación de los gráficos (Tableau)	11-8	10	21-8
Conexión a R y relleno de valores faltantes	6-7	2	8-7
Interpretación de modelos (código R)	9-7	10	19-7
Estimación de modelos (código R)	20-7	11	31-7
Evaluación de la capacidad de predicción	1-8	14	15-8
Análisis de resultados	16-8	9	25-8
Conclusiones y lineas futuras	29-8	20	19-9

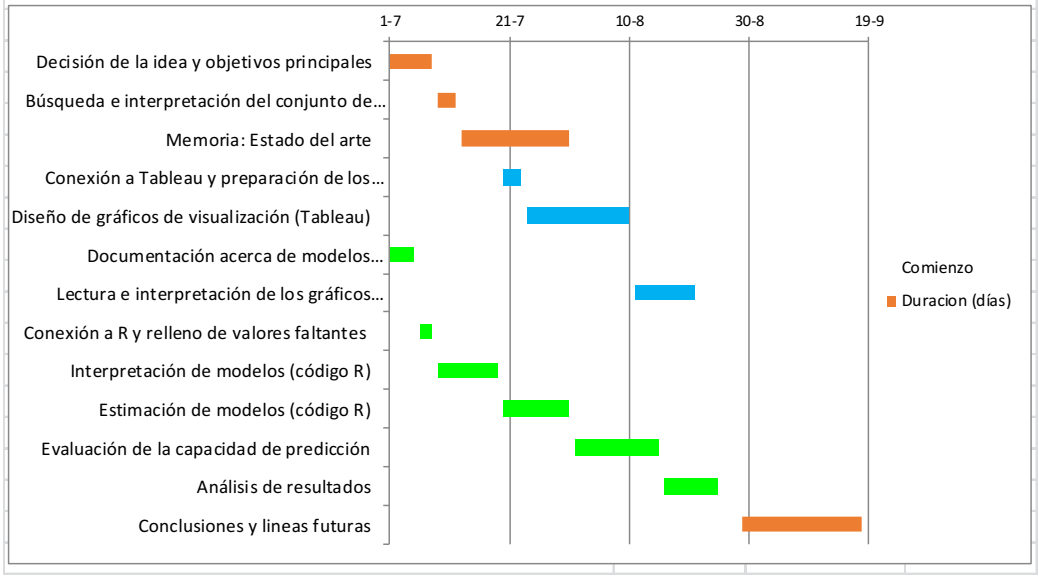


Figura 53. Planificación del proyecto

[10] La fecha se representa en formato día-mes, siendo por ejemplo 1-7 el día 1 de Junio. No se indica el año para favorecer la claridad en el gráfico (se sobreentiende que se conoce el período de trabajo).

En total, se han empleado 80 días a tiempo parcial (una media aproximada de 3-4 horas diarias). Como se puede observar, los tres profesionales realizan su respectiva tarea, llegando a trabajar de manera paralela (como por ejemplo el 21 de Julio). Esto es usual en el análisis de datos ya que se trata de un proceso llevado a cabo por un equipo de trabajo, en el que cada integrante se centra en una determinada tarea para que al final el jefe de proyecto, en este caso nuestro ingeniero industrial, saque conclusiones y líneas futuras acerca de lo observado.

A pesar de que todas las actividades se han realizado con éxito, hubo alguna etapa delicada que ha requerido aumentar el número inicial de horas estimado.

Tareas críticas:

Decisión de los modelos a implementar: debido al gran número de opciones que ofrece la aplicación de modelos predictivos.

Interpretación y estimación de modelos (código R): en esta etapa se produjeron errores de ejecución de código que hubo que consultar en los diferentes sitios que se especifican en el estado del arte del presente trabajo.

Este aumento del coste computacional ha requerido realizar un aumento en el número de días contratados del científico de datos de 30 a 50 días.

5.3. Presupuesto.

En el siguiente apartado, se va a hacer detalle del coste estimado que supone cada una de las partes necesarias para justificar el presupuesto final:

- **Hardware:** se ha utilizado el ordenador **Macbook Pro**. Especificaciones:
  - Memoria RAM 4 GB DDR3 a 1.600 MHz
  - Disco duro 500 GB.
  - Procesador 2,5 GHz Intel Core i5
  - Sistema Operativo: OS X El Capitan
  - Antigüedad: medidados de 2012
  - Vida útil: 4 años

  PRECIO IVA incluido: 1200 €

  Coste material proporcional al tiempo de ejecución y la vida útil relativa al equipo:

  1200 € / 4 años / 12 meses = 25 € por mes

  Coste total: 25 €/mes * 3 meses = 75 €

- **Software**:

  Consola de R y su interfaz gráfica RStudio: 0 € (licencia gratuita)

  Tableau Desktop[11]: 0 € (no es necesaria la compra de licencia del programa ya que se va a usar bajo versión académica)

[11] En caso de no poseer carnet académico, el coste del producto asciende a 999 dólares, en el caso de la edición personal, y 1999 dólares, la edición profesional [7].

Excel: 0 € (licencia académica)

- **Personal contratado**:

  Para estimar los costes del personal, se fijan tres costes:

  - 25 euros por hora por cada profesional que realice tareas de visualización y análisis exploratorio (Analista de datos).
  - 35 euros por hora por cada profesional que realice tareas de estadística y aplicación de modelos (Científico de datos).
  - 40 euros por hora por cada profesional que realice tareas de gestión de proyecto e Inteligencia de Negocio (Ingeniero industrial).

  En este caso el trabajo sólo es realizado por una persona con diferentes roles [véase apartado 5.1], por lo que se dividen los costes teniendo en cuenta la planificación de horas [véase punto 5.2] necesarias para la realización de cada etapa:

  Coste del analista de datos: 25 €/h * 4 h/día * 30 días = 3000 €

  Coste del científico de datos: 35€ /h * 3 h/día * 50 días = 7000 €

  Coste del ingeniero industrial: 40€ /h * 4 h/día * 48 días = 7680 €

- **Costes adicionales:**
  Debido a que todo el proyecto se ha realizado en la máquina, se van a calcular los costes materiales y de amortización que han supuesto su manejo:

  Coste relativo al consumo eléctrico [6]:

  (300 horas de encendido) * 0.102 €/kWh * 0.06 kW = 1,84 €

  Coste extra:

  Durante el desarrollo del proyecto, la batería dio fallos por lo que fue necesario llevar la máquina a la Genius bar de la tienda de Apple de la puerta del Sol para su revisión. Esto supuso un coste extra a nuestro proyecto de 150 €.

  Coste extra: 150 €

Por lo tanto, el presupuesto total del proyecto es el siguiente:

Coste total por hardware

Tabla 3. Coste total por hardware

Hardware	Precio
Equipo (parte proporcional)	75,00 €
Total	78,00 €

Coste total por software

Tabla 4. Coste total por software

Software	Precio de licencia
MAC OS X	Incluido en el precio del portátil
R	0 €
Tableau	0€ (Licencia académica)
Excel	0€ (Licencia académica)
Total IVA incluido	0 €

Coste total por el personal contratado

Tabla 5. Coste total por personal contratado

Personal contratado	Horas de trabajo por dia	Precio por hora (€)	Dias trabajados	Coste
Analista de datos	4	25	30	3.000,00 €
Cientíífico de datos	3	35	50	5.250,00 €
Ingeniero industrial	4	40	48	7.680,00 €
Total				15.930,00 €

Costes adicionales

Tabla 6. Costes adicionales

Concepto	Coste
Electricidad	1,84 €
Extra (reparación batería)	150,00 €
Total	151,84 €

PRESUPUESTO FINAL

Tabla 7. Coste del presupuesto final

Concepto	Coste
Software	0,00 €
Hardware	78,00 €
Personal	15.930,00 €
Otros costes	151,84 €
Total IVA incluido	**19.553,41 €**

6. Conclusiones y líneas futuras.

6.1. Conclusiones.

A la vista de los resultados obtenidos se pueden extraer las siguientes conclusiones:

En la primera parte del trabajo se ha realizado un estudio de exploración a través de la visualización del conjunto de series temporales con el objetivo de estudiar el comportamiento del consumo. A partir de la experiencia con el programa (Tableau) y de las diferentes lecturas de los gráficos, podemos extraer las siguiente conclusiones:

- Se ha conseguido obtener información acerca de los hábitos de consumo que realiza el hogar estudiado a lo largo del período de que se dispone. Esto nos ha confirmado que para hacer un estudio de serie temporal es necesario aplicar distintos niveles de granularidad o estudio en el tiempo para facilitar la lectura de tendencias y anomalías.
- Se ha mejorado en el manejo del programa así como en su familiarización con el alumno, a través del manejo de marcas, gráficos y funciones que no habían sido usadas con anterioridad.
- Tal y como se dijo a principio del análisis, el estudio de exploración se mejora si se conocen bien las variables y si estas son representadas mediante distintos tipos de gráfico. También se ha podido comprobar la importancia de realizar un pretratamiento de los datos antes de pasar a la generación interactiva de gráficos.
- Aunque el camino por el que se ha optado en cuanto a la granularidad y generación de gráficas ha proporcionado resultados óptimos, se podría haber llegado a la misma conclusión mediante el uso de otros gráficos u ofreciendo otro tipo de visualizaciones. Esto nos confirma la flexibilidad de la herramienta de la que hablábamos en su descripción, así como de su personalización y libertad.

En resumen se puede decir que el fruto de esta etapa ha sido bueno y nos ha confirmado la causa de que Tableau se encuentre entre los mejores programas para Inteligencia de Negocio.

Con respecto al consumo:

- Las horas de mayor consumo se dan de 7 a 8 de la mañana y de 19 a 21 de la noche.
- A medida mañana, período en el que la electricidad suele ser de menor coste que en otras franjas horarias, el consumo es muy bajo.
- El calentador de agua y el aire acondicionado son los electrodomésticos que suponen un mayor consumo.
- Lavadora, secadora, horno y luz son utilizados a la hora pico de consumo por lo que esto potencia el consumo de electricidad y por tanto el aumento de la factura.
- Se hace un buen uso de lavavajillas, horno y microondas (Sub_metering 2) dado que se utilizan en los períodos de menor consumo diario.

El resultado de estas lecturas se incluirá en el balance anual que la empresa energética envía cada año a la familia para hacer conciencia de un mejor uso energético.

En la segunda parte se ha realizado un estudio de habilidad predictiva de un conjunto de modelos de series temporales ampliamente utilizados en la literatura, implementadas desde R:

- Uno de los primeros fines ha sido el de evaluar su capacidad de predicción un paso adelante (un día) utilizando para ello una serie temporal de micro consumo eléctrico.

  Las particularidades de esta serie, datos de un hogar representativo con información al minuto ha hecho necesario cuestionarse no sólo qué modelo emplear sino qué nivel de agregación de los datos permite una mejora predictiva.

  Para ello, este estudio se ha centrado en la exploración de métodos univariantes (ARIMA x SARIMA) de series temporales y multivariantes (VAR). La desagregación va desde el nivel minuto al nivel horario pasando por un análisis por conglomerados de horas (estos conglomerados se han hecho con técnicas de análisis clúster.

- La capacidad predictiva se ha analizado utilizando validación cruzada, esto es, separando la muestra en entrenamiento y validación permitiendo ver cómo los modelos se van adaptando a los errores cometidos según se va conociendo información.
- El mejor modelo predictivo ha sido el univariante (ARIMA x SARIMA) referido a las series agregadas en horas.
- El correspondiente multivariante con las horas agrupadas no mejora sustancialmente la predicción.
- El univariante para la serie por minutos parece ser demasiado desagregado: el error de predicción es mayor, muestra más sesgo (infrapredicción en este caso) y la varianza de dicho error es la más alta (generando, por tanto, incertidumbre).

  Estos resultados cuadran con la teoría predictiva consultada que aboga, fundamentalmente, por un compromiso entre sencillez funcional y reducción del error.

  Por lo tanto, nuestra contribución trasversal con respecto a la metodología es una evidencia más a favor de que mayor desagregación de los datos o modelos complejos (generalmente multivariantes) no parecen mejorar ganancias predictivas adicionales.

  Quizás una solución óptima podría implicar combinar los mejores modelos predictivos (el univariante por horas y el VAR por horas agrupadas).

Por último y como conclusiones generales extraídas una vez finalizado el proyecto:

- Las herramientas de Inteligencia de Negocio y *Data Science* ofrecen soluciones reales para la búsqueda de conocimiento a partir del dato.
- La programación de los modelos ha sido la parte más compleja del proyecto, por lo que con esta experiencia se entiende la razón por la que la aplicación de modelos de *Deep Learning* y *Machine Learning* resulta aún un reto para las empresas.

- El manejo de un mayor número de variables hubiera hecho más completo nuestro trabajo, por lo que siempre que se disponga de tiempo, se intentarán analizar el mayor número de atributos posibles, aunque haya que buscar en otras fuentes.
- Aprendizaje de diferentes librerías y del entorno del software estadístico R. Gracias a la generación de código necesario para la construcción de los modelos se ha adquirido un nivel medio en el uso del programa, el cual no se había utilizado hasta la fecha.
- Aprendizaje de técnicas de análisis y aplicación de series temporales. La variedad de bibliografía consultada ha permitido asentar conocimientos sobre el uso de series temporales.

6.2. Líneas futuras

El resultado del análisis de este trabajo abre un sinfín de propuestas para futuros trabajos.

A continuación se van a enumerar posibles estudios que han quedado fuera del análisis y que sería de gran interés incorporar:

- **Estudio económico**: dado que en la fuente de datos no disponíamos de ninguna información acerca del coste del kilovatio, no hemos podido realizar un plan de ahorro específico para el hogar. Por tanto se propone como ampliación del estudio, la posibilidad de añadir al conjunto de datos, una nueva columna que indique el coste del kilovatio relativo a cada hora y día determinado, y de esta manera poder reflejar datos de interés para la familiar como importes de facturas, o tablas que mostrasen en que franjas horarias del día, el precio de la electricidad es más barato.
- **Estudio de factores externos**: por lo general, en la predicción de electricidad se debe tener en cuenta la más que posible presencia de perturbaciones externas a la hora de aplicar los modelos. El clima de la región, los cambios bruscos en el tiempo, la fauna o la actividad solar serían algunos ejemplos de factores que pueden añadir incertidumbre y hacer menos efectivos a nuestros modelos, lo que conduce a resultados menos fiables.

  Dado que en el proyecto no disponíamos de información acerca de la localización del hogar en el que se han recopilado los datos, no se ha podido estudiar la influencia de factores meteorológicos y medioambientales, pero resultaría muy interesante tenerlo en cuenta para comprender más a fondo el comportamiento del consumo y cuantificar la perturbación que puede estar sufriendo.
- **Estudio de control sobre la red**: En el caso de que quisiéramos realizar un estudio más profundo de la red eléctrica, como por ejemplo para conocer el nivel de seguridad de sobrecarga de nuestro sistema, podríamos hacer un análisis de control tomando como variables de entrada el voltaje y la potencia reactiva, ya que el flujo de potencia reactiva influye considerablemente en los niveles de voltaje, y de esta manera asegurarnos que no se sobrepase de los límites contratados.
- **Desarrollo de una aplicación**: esta aplicación debe permitir al usuario conocer en todo momento datos de la red eléctrica como el consumo actual, el consumo acumulado hasta la fecha y una predicción que refleje un importe aproximado de

factura. De esta manera el cliente puede reducir el consumo en caso de tener problemas económicos o simplemente hacer una gestión más responsable en lo que quede de mes. Además la aplicación debe ser intuitiva e interactiva, por lo que debe presentar una interfaz que permita al usuario introducir los atributos que desea conocer y estudiar posibles modificaciones como por ejemplo cuánto puede ahorrar si apaga algún electrodoméstico prescindible, como el horno.

En el ámbito de la visualización, disponer de más datos, como por ejemplo datos de temperatura, nos permitiría realizar gráficos más completos con los que poder responder a nuevas preguntas relacionadas con hábitos de consumo o con el uso de aparatos eléctricos como el aire acondicionado. En la siguiente ilustración se muestra un gráfico descargado desde el repositorio público de Tableau, en el que se compara la influencia de la temperatura en el consumo de potencia. A partir de él se podría estudiar cómo influye ésta en el consumo a medida que aumenta o disminuye, y así conocer aspectos como por ejemplo si la casa dispone o no de aire acondicionado, y cuanto influye su uso en el consumo sin necesidad de estudiar otras variables:

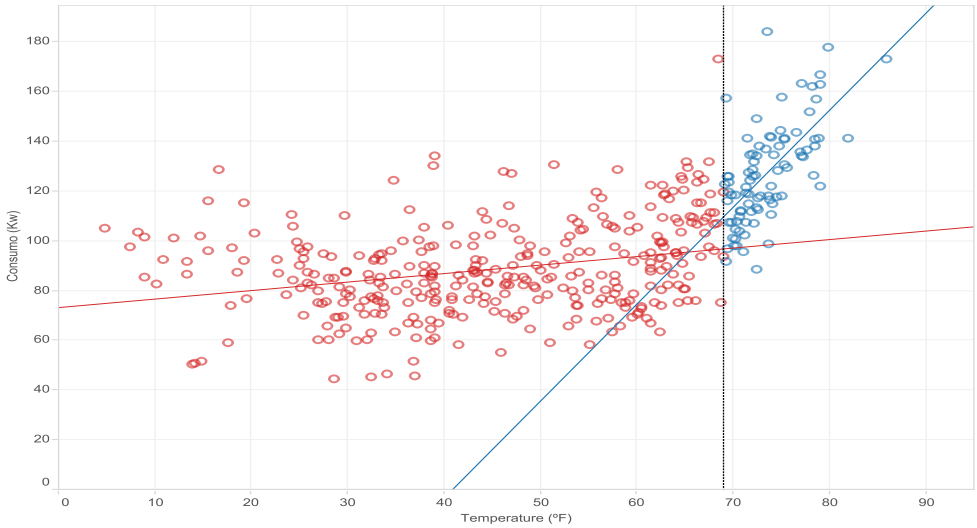


Figura 54. Comportamiento del consumo eléctrico en función de la temperatura.

En la imagen podemos observar la dispersión de puntos relativa a las diferentes medidas de consumo. El color azul se ha utilizado para representar el punto en el que se estima que se ha encendido el aire acondicionado.

Por último y en relación al análisis predictivo, se podrían haber aplicado otros modelos con los que contrastar resultados como:

- El uso de un modelo FVAR (factorial VAR) que estime el modelo multivariante y los clúster de agrupación por horas de forma simultánea. Esto, seguramente, reduzca incertidumbre al realizarse todo en una sola etapa y permita mejorar error predictivo (al menos, en varianza).
- El uso de modelos de redes neuronales y de aprendizaje automático: podrían ser buenos modelos para comparar con los más sencillos. Sin embargo, el coste

computacional que conlleva el trabajo diseñado (por ejemplo modelos para cada uno de los 1440 minutos disponibles en un día) desaconsejaron su uso

- Introducir modelos para la varianza y tratamiento de atípicos (fundamentalmente, modelos GARCH). Primeramente requeriría una rutina de análisis de *outliers* y, posteriormente, procesar la posible heterocedasticidad condicional de los datos. Esto podría mejorar la eficiencia predictiva.

7. Referencias.

Bibliografía.

Peter J. Brockwell, Richard A. Davis (Second Edition) – "Introduction to Time Series and Forecasting"

Sean Patrick Murphy (Marzo 2016) – "Data and electric power - From deterministic machines to probabilistic systems in traditional Engineering" O'Reilly

Josep Lluís Cano (No fig. fecha), "Business. Intelligence: Competir con información" ESADE

Jean-Louis Monino y Soraya Sedkaoui (2016), "Big Data, Open Data and Data Development", Volume 3

Green, K. C., & Armstrong, J. S. (2015). "Simple versus complex forecasting: The evidence. Journal of Business Research", 68(8), 1678-1685.

Box, G. E., & Jenkins, G. M. (1970). "Time Series Analysis Forecasting and Control". WISCONSIN UNIV MADISON DEPT OF STATISTICS.

Clements, M., & Hendry, D. (1998). "Forecasting economic time series". Cambridge University Press.

Peña, D. P. S. (2005). "Análisis de series temporalis". Alianza Editorial.

Hamilton, J. D. (1994). "Time series analysis (Vol. 2). Princeton: Princeton university press".

Clements, M. P., & Hendry, D. F. (1995). "Forecasting in cointegrated systems.Journal of Applied Econometrics", 10(2), 127-146.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). "Time series analysis: forecasting and control". John Wiley & Sons.

Dickey, D. A., & Fuller, W. A. (1981). "Likelihood ratio statistics for autoregressive time series with a unit root". Econometrica: Journal of the Econometric Society, 1057-1072.

Arlot, S., & Celisse, A. (2010). "A survey of cross-validation procedures for model selection". Statistics surveys, 4, 40-79.

Durbin, J., & Koopman, S. J. (2012). "Time series analysis by state space methods" (No. 38). Oxford University Press.

·Tableau: Información sobre gráficos. Disponible online (último acceso 20/09/2016)

https://www.tableau.com/sites/default/files/media/Whitepapers/which_chart_v6_es-final_0.pdf

·KDnuggets: Información sobre minería de datos. Disponible online (último acceso 25/08/2016): http://www.kdnuggets.com

·Artículo de Luis Carlos Molina Félix: “Data mining: torturando los datos” Disponible online (último acceso 15/07/2016):

http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html

·Vídeo sobre la historia de la inteligencia de negocios. Disponible online (último acceso 10/07/2016):

: https://www.youtube.com/watch?v=_1y5jBESLPE

·Noticias de actualidad sobre el mundo del análisis de datos. BBVA. Disponible online (último acceso 15/07/2016):

http://www.centrodeinnovacionbbva.com

https://bbvaopen4u.com/es

[1] *El País*. Datos de la estadística publicada por el periódico *El País* y llevada a cabo por Comisión Nacional de los Mercados y la Competencia (CNMC). Disponible online (último acceso 18/07/2016): http://economia.elpais.com/economia/2016/05/13/actualidad/1463138230_999456.html

[2] *El Mundo*. Datos publicados por el periódico *El Mundo*. Disponible online (último acceso 20/07/2016):

http://www.elmundo.es/economia/2015/10/20/5626187fca474195608b45c7.html

[3] Qlikview. Datos de Qlikview. Disponible online (último acceso 22/07/2016):

http://global.qlik.com/es

[4] Yann LeCun, Yoshua Bengio y Geoffrey Hinton: “Artículo sobre Deep learning” Disponible online (último acceso 16/07/2016):

https://www.cs.toronto.edu/~hinton/absps/NatureDeepReview.pdf

[5] Base de datos del Repositorio UC Irvine. Disponible online (último acceso 01/09/2016): https://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption

[6] Apple. Información acerca del consumo del equipo Macboook pro. Disponible online (último acceso 10/09/2016) :

https://support.apple.com/kb/SP649?locale=es_ES&viewlocale=es_ES

[7] Tableau. Productos de Tableau Desktop. Disponible online (último acceso 01/09/2016) :

http://www.tableau.com/products/desktop

Anexo.

Código (R)

```
#localizo el directorio y el conjunto de datos
getwd()
setwd("~/desktop/DATOS")
#en PC o Portátil
HPC <- read.csv("household_power_consumption.txt", sep=";")
attach (HPC)
install.packages("quantmod")
library("quantmod", lib.loc="~/R/win-library/3.1")
#convierto el tiempo en un vector numérico
#HPCdf<-data.frame(HPC)
#HPCdf$Time<-as.numeric(HPCdf$Time)
plot(Global_new)
acf(Global_new,360)
pacf(Global_new,40)
plot(log(Global_new))
acf(log(Global_new),360)
pacf(log(Global_new),40)
install.packages("fUnitRoots")
x<-log(Global_new)
k = trunc((length(x)-1)^(1/3))
adfTest(x, lags = 12, type = "nc", title = NULL,
        description = NULL)
##################################################################  MÓDULO  POR MINUTOS ###################
#en este módulo dividiremos el dataset por minutos, de tal forma que predigamos en bloques por minutos.
#Tendremos tantas filas como días y tantas columnas como minutos un día (1440).
#hago un subconjunto (subset) que comience en Tiempo 1
Ct1 <- HPC[397:1047276,]
Ct1$Time<-as.numeric(Ct1$Time)
Ct1$Global_active_power<-as.numeric(as.character(Ct1$Global_active_power))
```

```
D<-Ct1[Ct1[, 2] == 1,3]

D<-ts(D, frequency=7)

#D<-t(D)

#ahora tengo que generar una matriz donde cada columna sea un minuto

for (i in 2:1440) {



  x<-Ct1[Ct1[, 2] == i,3]

  x<-ts(x, frequency=7)


  D<-cbind(D,x)




}

#gráfico de un subset

D1<-D[,1]

D100<-D[,100]

D500<-D[,500]

D800<-D[,800]

D1000<-D[,1000]

D1400<-D[,1400]

minutes<-ts(data.frame(D1,D100,D500,D800,D1000,D1400))

ts.plot(minutes,gpars= list(col=rainbow(6)))

lminutes<-log(minutes[,2:6])

ts.plot(lminutes,gpars= list(col=rainbow(6)))

############################################################ MODELO UNIVARIANTE POR MINUTOS##############

#en este módulo vamos a hacer un modelo ARIMA automático. Para ello, haremos un subsample y testaremos

#la capacidad predictiva

install.packages("forecast")

library("forecast", lib.loc="~/R/win-library/3.1")
```

```
k <- 620 # mínima longitud para ajustar el modelo
n <- 720
st <- k
error<-0
errorf<-0
real<-0
realf<-0


for (i in 1404:1440) {


              p<-log(D[,i])



                for(j in 1:(n-k-1))
                     {
                            xshort <- p[1:st+j]
                            xnext <- p[st+j+1]
                            xshort<-ts(xshort,frequency=7)
                            fit2<-arima(xshort,order=c(4,0,0),
seasonal=list(order=c(1,0,0),period=7))
                           #auto.arima(xshort)
                                #fit2         <-         auto.arima(xshort,max.p=10,
max.q=10,max.P=5,max.Q=5,
                                #                         ic=c("aicc","aic",  "bic"),
stepwise=FALSE,
                                #                  allowdrift=TRUE)
                                fcast2 <- forecast(fit2, h=1)

                                    error[j]<-(exp(xnext[1])-exp(fcast2$mean[1]))
                                    real[j]<-(exp(xnext[1]))
                                  }

                 errorf<-cbind(errorf,error)
                realf<-cbind(realf,real)
```

```
        }


errordia<-rowSums(errorf)

errordia<-na.omit(errordia)

realdia<-rowSums(realf)

realdia<-na.omit(realdia)

MAPEuni<-mean(abs(errordia)/realdia)

#agrupar por horas#

#genero las horas en un formato legible para R

data1 <- data.frame(date = seq(from = ISOdatetime(2006, 12, 16, 17, 24, 00),

                               length.out = 2075259, by=60))

#genero un set intermedio

data2<-cbind(data1,HPC)

data2$Global_active_power<-as.numeric(as.character(data2$Global_active_power))

#agrupo por horas la información al minuto

hr.means <- aggregate(data2["Global_active_power"],

                      list(hour = cut(data2$date, breaks="hour")),

                      sum, na.rm = FALSE)

#obtengo las horas del vector de fechas-horas

t<-strftime(hr.means$hour,format="%H:%M:%S")

#y las pego en el archvo

D2<-cbind(hr.means,t)

#ecodifico las horas a números

D2$t<-as.numeric(D2$t)

D2$Global_active_power<-as.numeric(as.character(D2$Global_active_power))

#me quedo con la muestra con hora desde la 1 y que acabe en la 24

Ct2 <- D2[8:34539,]

Dd<-Ct2[Ct2[, 3] == 1,2]

Dd<-ts(Dd)

#ahora tengo que generar una matriz donde cada columna sea una HORA

for (i in 2:24) {
```

```
  x<-Ct2[Ct2[, 3] == i,2]

  x<-ts(x)


  Dd<-cbind(Dd,x)


}
#Ahora estimo los modelos univariantes
k <- 1340 # mínima longitud para ajustar el modelo
n <- 1440
st <- k
errorH<-0
errorHf<-0
realH<-0
realHf<-0
for (i in 23:24) {


       p<-log(Dd[,i])



          for(j in 1:(n-k))
            {
                xshort <- p[1:st+j]
                xnext <- p[st+j+1]


                fit2 <- auto.arima(xshort, d=1,max.p=10,
max.q=10,ic=c("aicc","aic", "bic"), stepwise=TRUE,allowdrift=TRUE)
                fcast2 <- forecast(fit2, h=1)


                  errorH[j]<-(exp(xnext[1])-exp(fcast2$mean[1]))
                  realH[j]<-exp(xnext[1])
                }


          errorHf<-cbind(errorHf,errorH)
```

```
            realHf<-cbind(realHf,realH)


    }

    errordiaHf<-rowSums(errorHf)

    realdiaHf<-rowSums(realHf)

    MAPEuni<-mean(abs(na.omit(errordiaHf))/(na.omit(realdiaHf)))

    ############################################## MODELO MULTIVARIANTE

    #Modelo de Componentes principales

    install.packages("corrplot")

    library("corrplot", lib.loc="~/R/win-library/3.2")

    dfmin<-data.frame(Dd)

    colnames(dfmin)<-
c("h1","h2","h3","h4","h5","h6","h7","h8","h9","h10","h11","h12","h13","h14","h15","h1
6","h17","h18","h19","h20","h21","h22","h23","h24")
    corrplot(cor(na.omit(dfmin)),is.corr=TRUE, method="color")

    pca_a<-princomp(na.omit(dfmin), cor = FALSE)

    require(graphics)

    plot(pca_a)

    #plot pca

    biplot(pca_a)

    #plot scores with labels

    plot(pca_a$loadings[,1:2],type="n", main="Title", sub="A subtitle")

    text(pca_a$loadings[,1],pca_a$loadings[,2],c("Var1","Var2","..."))

    install.packages("TSclust")

    library("TSclust", lib.loc="~/R/win-library/3.2")

    diffs<-rep(1,ncol(dfmin))

    logs<-rep(TRUE,ncol(dfmin))

    dpred<-diss(na.omit(dfmin),"PRED",h=1,B=1000,logarithms=logs,differences=diffs,
plot=TRUE)
    hc.dpred <- hclust(dpred$dist)

    plot(hclust(dpred$dist))

    #obtenemos que las horas clúster son

    #h19

    #h22 h18 h21

    #h16 h17
```

```
#h13 h14
#h12 h15
#h9 h10
#h11
#h8 h20
#h6 h7
#h5 h23
#h3
#h2 h4
#h1 h24


################################ MODELO MULTIVARIANTE AGRUPADO POR HORAS
#agrupo las horas
S1<-ts(dfmin["h19"])
S2<-rowSums(ts(data.frame(dfmin["h22"],dfmin["h18"],dfmin["h21"])))
S3<-rowSums(ts(data.frame(dfmin["h16"],dfmin["h17"])))
S4<-rowSums(ts(data.frame(dfmin["h13"],dfmin["h14"])))
S5<-rowSums(ts(data.frame(dfmin["h12"],dfmin["h15"])))
S6<-rowSums(ts(data.frame(dfmin["h9"],dfmin["h10"])))
S7<-rowSums(ts(data.frame(dfmin["h11"],dfmin["h8"],dfmin["h20"])))
S8<-rowSums(ts(data.frame(dfmin["h6"],dfmin["h7"])))
S9<-rowSums(ts(data.frame(dfmin["h5"],dfmin["h23"])))
S10<-ts(data.frame(dfmin["h3"]))
S11<-rowSums(ts(data.frame(dfmin["h2"],dfmin["h4"])))
S12<-rowSums(ts(data.frame(dfmin["h1"],dfmin["h24"])))


dataVAR=ts(data.frame(S1,S2,S3,S4,S5,S6,S7,S8,S9,S10,S11,S12))
dataVAR[is.na(dataVAR)] <- 0
install.packages("vars")
library("vars", lib.loc="~/R/win-library/3.2")
selec<-VARselect(dataVAR, lag.max = 15, type = "both")
k <- 1339 # mínima longitud para ajustar el modelo
```

```
n <- 1439
st <- k
error1<-0
error2<-0
error3<-0
error4<-0
error5<-0
error6<-0
error7<-0
error8<-0
error9<-0
error10<-0
error11<-0
error12<-0
real1<-0
real2<-0
real3<-0
real4<-0
real5<-0
real6<-0
real7<-0
real8<-0
real9<-0
real10<-0
real11<-0
real12<-0

  for(j in 1:(n-k-1))
  {
    xshort <- dataVAR[1:st+j,]
    xnext <- dataVAR[st+j+1,]

    selec<-VARselect(xshort, lag.max = 15, type = "both")
```

```
        fit2 <- VAR(xshort, p = selec$selection[1])
        fcast2 <- predict(fit2, n.ahead=1)

        error1[j]<-(xnext[1])-(fcast2$fcst$h19[1])
        error2[j]<-(xnext[2])-(fcast2$fcst$S2[1])
        error3[j]<-(xnext[3])-(fcast2$fcst$S3[1])
        error4[j]<-(xnext[4])-(fcast2$fcst$S4[1])
        error5[j]<-(xnext[5])-(fcast2$fcst$S5[1])
        error6[j]<-(xnext[6])-(fcast2$fcst$S6[1])
        error7[j]<-(xnext[7])-(fcast2$fcst$S7[1])
        error8[j]<-(xnext[8])-(fcast2$fcst$S8[1])
        error9[j]<-(xnext[9])-(fcast2$fcst$S9[1])
        error10[j]<-(xnext[10])-(fcast2$fcst$h3[1])
        error11[j]<-(xnext[11])-(fcast2$fcst$S11[1])
        error12[j]<-(xnext[12])-(fcast2$fcst$S12[1])
        real1[j]<-(xnext[1])
        real2[j]<-(xnext[2])
        real3[j]<-(xnext[3])
        real4[j]<-(xnext[4])
        real5[j]<-(xnext[5])
        real6[j]<-(xnext[6])
        real7[j]<-(xnext[7])
        real8[j]<-(xnext[8])
        real9[j]<-(xnext[9])
        real10[j]<-(xnext[10])
        real11[j]<-(xnext[11])
        real12[j]<-(xnext[12])

        }

    E<-
cbind(error1,error2,error3,error4,error5,error6,error7,error8,error9,error10,error11,e
rror12)
```

```
      R<-
cbind(real1,real2,real3,real4,real5,real6,real7,real8,real9,real10,real11,real12)
      errordiaEf<-rowSums(E)

      realdiaRf<-rowSums(R)



      MAPEuni<-mean(abs(na.omit(errordiaEf))/(na.omit(realdiaRf)))



      cyl<-cbind(errordia,errordiaHf[1:94],errordiaEf[1:94])



      # plot densities

      p1 <- errordia                  # centered at 4

      p2 <- errordiaHf[1:94]

      p3<-errordiaEf[1:94]# centered at 6

      hist(p1, xlim=c(-1500,1500), col="red")


      hist(p2, add=T, col=rgb(0, 1, 0, 0.5) )

      hist(p3, add=T, col=rgb(0, 1, 0.75, 0.5) )
```

Anexo 1. Rellenado de valores ausentes mediante el filtro de Kalman

Debido a la existencia de observaciones faltantes en el set de datos original, hemos utilizado el filtro de Kalman (Durbin, J., & Koopman, S. J. (2012)) como un método óptimo (en entornos lineales) para atribuir dichos datos.

Este filtro, básicamente, contempla la existencia de una ecuación de "medida":

$$y_t = Z\alpha_t + \epsilon_t$$

Y otra de "transición":

$$\alpha_t = T\alpha_{t-1} + u_t$$

Donde asumiremos que ϵ_t y u_t son perturbaciones de media cero, matriz de varianzas y covarianzas escalar e incorreladas entre ellas y que α_t es no observable con condición inicial aleatoria.

Este filtro tiene unas ecuaciones recursivas que permiten obtener, a cada paso, el valor de los no observables basado en las medidas de y_t. Realiza un conjunto de evaluaciones del estilo "predicción-corrección". Es decir, asume valores para los parámetros desconocidos y realiza una predicción óptima para y_{t+1}. Es entonces cuando evalúa la predicción y trata de corregirla (donde emplea la ganancia de Kalman, ver Durbin, J., & Koopman). Cuando se minimizan dichos errores para toda la serie, el filtro ha convergido. En caso de haber observaciones faltantes, aprovecha las estimaciones realizadas con las observaciones no faltantes y las aplica de igual modo, pero sin poder efectuar una "corrección".

Anexo 2. Instalación de R.

Para la instalación de R habrá que realizar los siguientes pasos:

1. Navegar hasta la página web de R en el siguiente enlace:
   https://cran.r-project.org
2. Seleccionar dentro del apartado **Download and Install R**, el sistema operativo de la máquina que manejamos (Windows, Mac OS X o Linux), en nuestro caso Mac OS X
3. Una vez dentro del directorio de R para Mac OS X, encontraremos una serie de ediciones del programa. En tono gris, se nos indica la última versión disponible hasta la fecha:
   R 3.3.1 "Bug in Your Hair" released on 2016/06/21
   Para comenzar la descarga, seleccionamos esta versión de R dentro del apartado Files. Comprobamos que se inicia la descarga.
   NOTA: Se nos indica que el uso de X11 (paquetes especiales de R como su interfaz gráfica R-commander, Rcmdr o paquetes de representación de gráficos) requiere de la instalación de XQuartz, una aplicación que se instalará aparte en la máquina y que tendrá que ser reinstalada cuando existan nuevas versiones de R.
4. Tras finalizar la descarga del programa, iniciamos la instalación en nuestro equipo, aceptamos licencias y seleccionamos el sitio donde se desea guardar la aplicación.
5. Comprobamos que tenemos la aplicación instalada en nuestro sistema y la abrimos haciendo click sobre el icono de R. Debería aparecer la siguiente ventana:

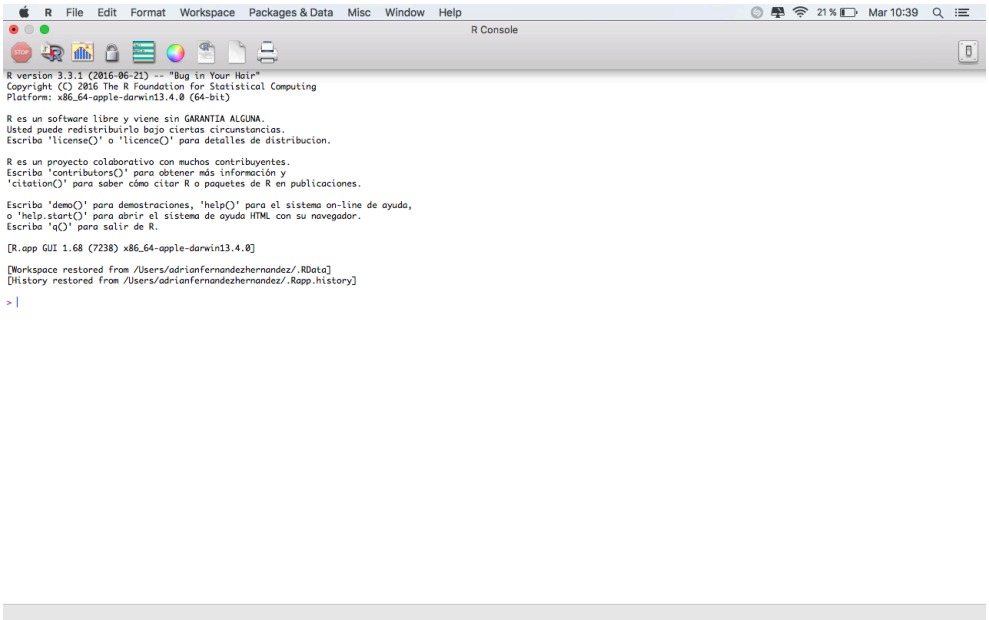


Figura 55. Interfaz de la consola de R en sistema operativo MAC OS

En la interfaz de R podemos diferenciar tres partes:

- **Barra de tareas**: En la zona de arriba como en todo programa que utilicemos en Mac, se encuentra la barra de tareas cuyos menús principales son:
  - Archivo (File): nos permite abrir un nuevo archivo, así como cargar directamente trabajos de R que podamos haber descargado o escrito en otro momento. También será el sitio donde guardar el proyecto cuando se desee.
  - Edición (Edit): sirve para realizar modificaciones en el código, como copiar, cortar, pegar o eliminar algunas líneas.

- Formato (Format): nos permite realizar cambios en el formato y fuente del texto escrito.
- Directorio de trabajo (Workspace): muestra el directorio de trabajo donde se guardan los trabajos. El directorio debe localizarse en el sitio donde guardemos nuestros archivos de datos, para que de esta manera podamos leerlos desde el programa.
- Paquetes y datos (Packages & Data): en esta opción podemos cargar librerías de paquetes para apoyarnos en funciones ya creadas.
- Misc: pestaña específica para cargar el servidor X11 (XQuartz) y cambiar el directorio de trabajo.
- Ventana (Window): permite acceder al modo pantalla completa, minizar o hacer zoom sobre ella y cambiar de vista entre las diferentes ventanas abiertas.
- Ayuda (Help): pestaña para buscar soporte en caso de dudas. Contiene un buscador donde escribir el concepto a preguntar, además de un enlace a preguntas frecuentes.

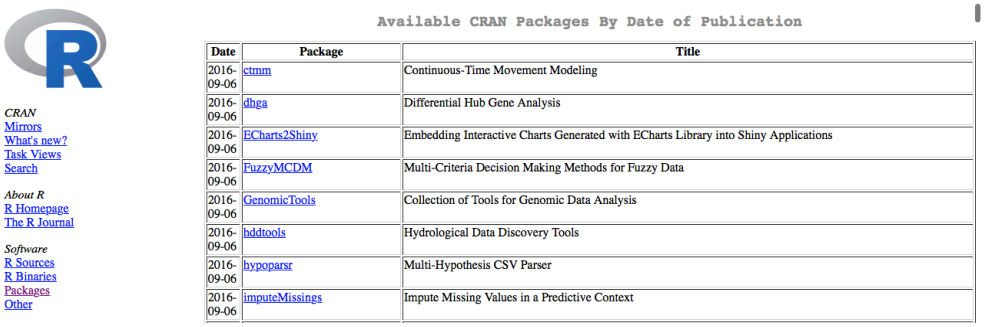
- **Zona de iconos**: Debajo de esta barra aparecen una serie de iconos de acceso rápido con los que podremos realizar algunos comandos sin necesidad de entrar en la barra de tareas. Cabe destacar por su utilizad el icono "stop", muy utilizado en momentos en los que pueda colapsar el programa como por ejemplo cuando se realiza la lectura de un archivo que presenta varios millones de entradas. Al hacer click sobre él se para automáticamente la ejecución del código.
- **Consola de R**: zona en la que escribir las sentencias de código y donde irán apareciendo los resultados de los mismos. Cada nueva instrucción se representa mediante el símbolo: >

  Para ejecutar un comando, no tenemos más que introducirlo y pulsar Intro. Un comando puede extenderse varias líneas. En caso de error se escribirá un mensaje en rojo en la pantalla indicándonos brevemente el motivo del mismo. Si se desea cerrar la consola y salir en cualquier momento, usar el comando quit() o q().

  En el caso de querer disponer de un script donde escribir de una manera más limpia todas las sentencias para ejecutarlas a la vez, R permite realizar esto si seleccionamos: Archivo -> Nuevo script

Instalación de paquetes

R permite hacer uso de paquetes de librerías ya creadas cuyo listado podemos encontrar en la página principal: http://cran.us.r-project.org/web/packages/, en el apartado Software: **Packages** (Paquetes). En esta lista además del nombre del paquete, aparece el titulo de la librería donde se especifica la función principal de la misma, así como la fecha en la que se ha añadido al repositorio:



CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Available CRAN Packages By Date of Publication

Date	Package	Title
2016-09-06	ctmm	Continuous-Time Movement Modeling
2016-09-06	dhga	Differential Hub Gene Analysis
2016-09-06	ECharts2Shiny	Embedding Interactive Charts Generated with ECharts Library into Shiny Applications
2016-09-06	FuzzyMCDM	Multi-Criteria Decision Making Methods for Fuzzy Data
2016-09-06	GenomicTools	Collection of Tools for Genomic Data Analysis
2016-09-06	hddtools	Hydrological Data Discovery Tools
2016-09-06	hypoparsr	Multi-Hypothesis CSV Parser
2016-09-06	imputeMissings	Impute Missing Values in a Predictive Context

Figura 56 . Repositorio CRAN de paquetes de librerías disponibles en R

Cuando se quiera instalar un paquete en R, tendremos que ir al menú Paquetes en la barra de tareas y seleccionar instalar paquete. Al hacer click, se abre una pantalla que nos pide seleccionar la localización del servidor https de CRAN donde nos encontramos. En nuestro caso, seleccionamos **Spain (Madrid)**:

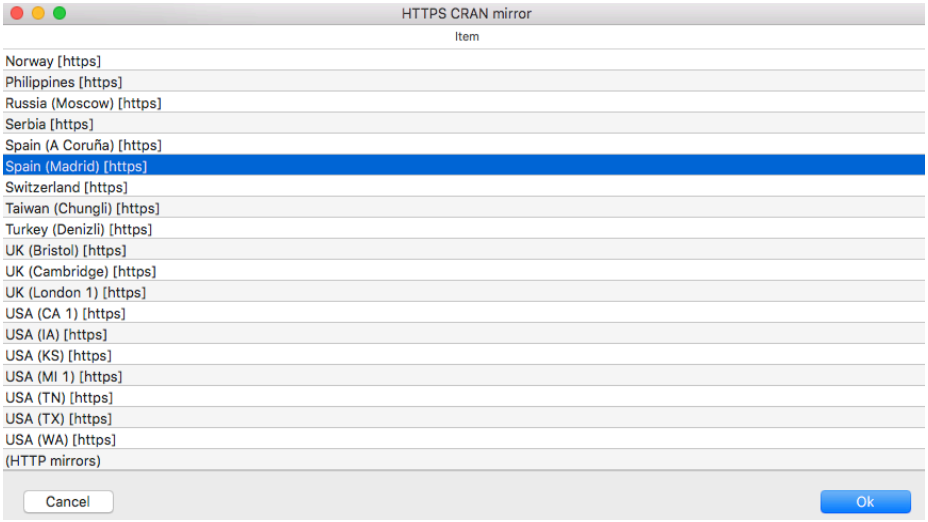


Figura 57 . Configuración del servidor [https] de CRAN

Guardamos la configuración seleccionada para que no nos vuelva a preguntar acerca de ello en la siguiente instalación de paquetes. De esta manera ya tendríamos cargada la lista de librerías disponible:

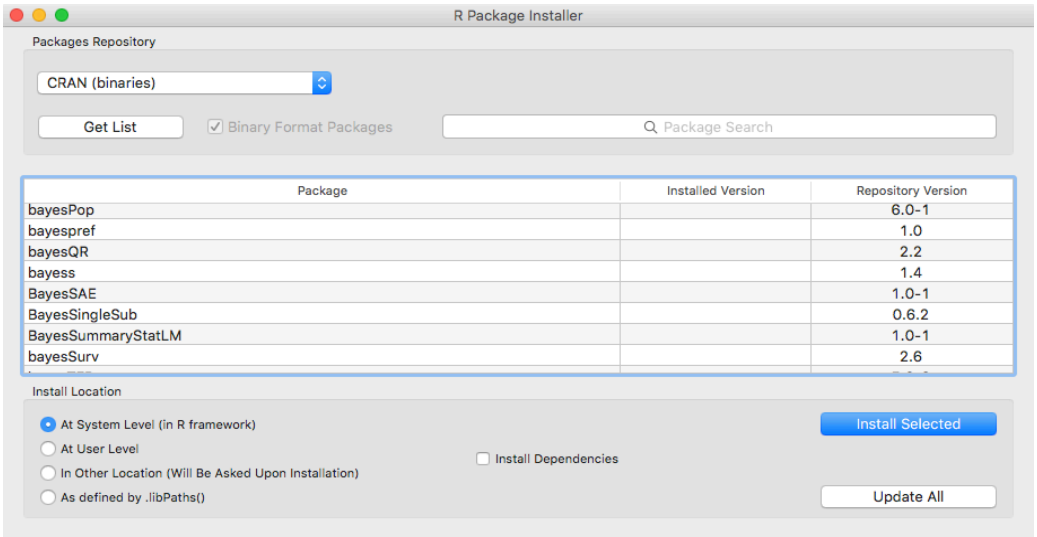


Figura 58. Instalador de paquetes de R

Una vez localizada el paquete necesario, seleccionamos **Install selected** (instalar selección) y automáticamente se nos generará un mensaje en la consola de R indicando acerca de la instalación del paquete.

Además de la forma anteriormente citada, R permite la instalación de paquetes desde la consola si escribimos la sentencia:

```
install.packages("Nombre de la librería")
```

Importante: no olvidar las comillas porque en caso contrario nos aparecerá un mensaje de error en la consola.

Configuración del directorio de trabajo

En la primeria línea por defecto que aparece en la consola [véase figura 55], podemos leer la versión de R con la que estaremos trabajando: versión 3.3.1, así como el directorio actual de trabajo: /Users/adrianfernandezhernandez/.Rapp.history. En nuestro caso y dado que nuestra fuente de datos se ha almacenado en una carpeta específica para el trabajo, carpeta DATOS, no nos permitiría importar el archivo desde el programa, por lo que tendremos que cambiar el directorio de trabajo.

Para ello disponemos de las funciónes getwd() y setwd(). La primera nos indica el directorio actual de trabajo y la segunda nos permite decirle a R, donde queremos fijar nuestro directorio, donde por lo general, se deberían de encontrar los conjuntos de datos a tratar.

Escribimos en la consola el siguiente código, obteniendo como resultado:

```
> getwd()
[1] "/Users/adrianfernandezhernandez"
> setwd("~/desktop/DATOS")
> getwd()
[1] "/Users/adrianfernandezhernandez/Desktop/DATOS"
```

Ya hemos cambiado el directorio de trabajo por lo que ahora ya sí, podemos comenzar la importación.

Anexo 3. Instalación de Tableau, conexión a los datos y descripción de las hojas de trabajo.

Instalación de Tableau desktop.

Para la instalación de la herramienta Tableau debemos realizar los siguiente pasos:

1. Acceder a la página oficial del producto: http://www.tableau.com/es-es
2. Seleccionamos “Probar ahora” en el recuadro naranja situado en la esquina superior derecha. Automáticamente nos pedirá una dirección de correo electrónico que servirá como cuenta de usuario para acceder posteriormente a registros de otros productos, tutoriales o a la comunidad de usuarios de Tableau. Tras introducirlo, comenzará la descarga del ejecutable.
3. Una vez descargado el ejecutable, nos aseguramos que aparece el icono del programa en nuestra máquina y haciendo click sobre él, ya podremos comenzar a utilizarlo.

Conexión de Tableau al conjunto de datos.

Una vez abierta la aplicación y como primer paso, debemos conectar el conjunto de datos con el software, indicando el tipo de archivo que se desea cargar.

Al abrir la aplicación aparece el siguiente panel de mandos:

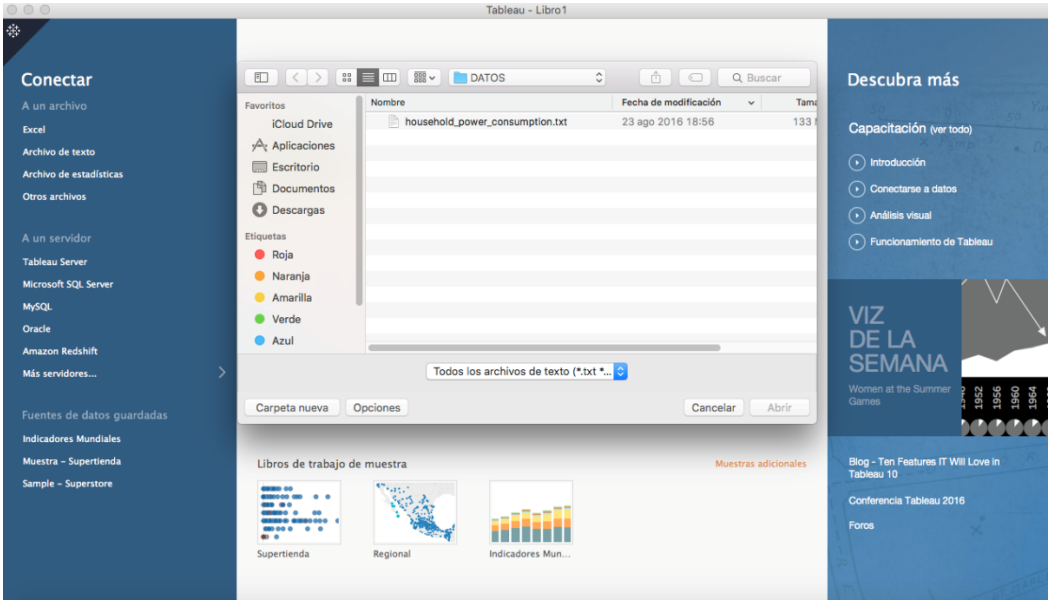


Figura 59. Interfaz de conexión a Tableau

Como se puede observar, Tableau permite una gran variedad de conexiones: archivo de texto, Excel, conexiones a servidores MySQL u Oracle. En nuestro caso, al tratarse de un archivo de tipo texto, seleccionamos **Conectar a un archivo de texto**, y abrimos nuestro fichero, el cual teníamos almacenado en la carpeta DATOS de nuestro escritorio.

Una vez realizada la conexión, se abre un Libro en el que podemos observar la fuente de datos ya importada, con los diferentes atributos y valores y en forma de tabla, muy similar a lo que obtendríamos al cargar un archivo de bases de datos en otros programas, como por ejemplo Excel:

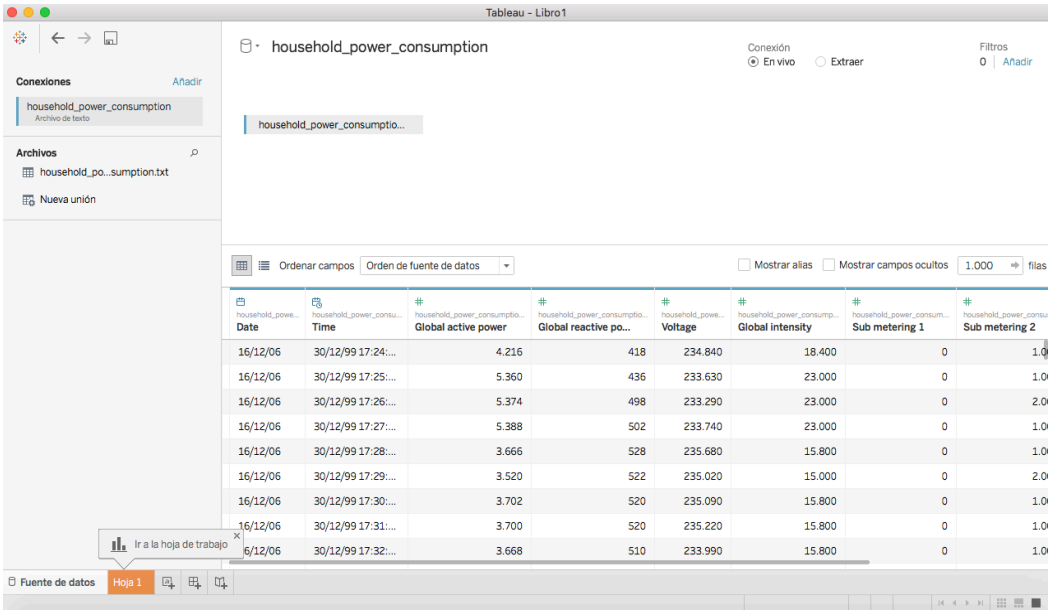


Figura 60. Hoja de configuración de fuente de datos en Tableau

Tableau nos permite realizar las visualizaciones en vivo o en modo extraer. El primer caso se usa cuando los datos fuente sufren variaciones continuas en vivo, por ejemplo si estamos tratando con una base de datos que almacena opiniones de twitter sobre un determinado tema. Estas opiniones se generan a medida que la gente escribe en la red social, por lo que interesa estar conectado a la misma y de esta manera no tener que actualizar los datos cada cierto tiempo.

Debido a que nuestros datos no van a sufrir cambios, seleccionaremos la opción extraer para así poder trabajar con ellos en todo momento, estemos conectados o no a la red.

Hojas de trabajo y diseño de gráficas.

Las hojas de trabajo son el lugar donde crear las visualizaciones.

La interfaz de comandos de las hojas de trabajo es siempre la misma:

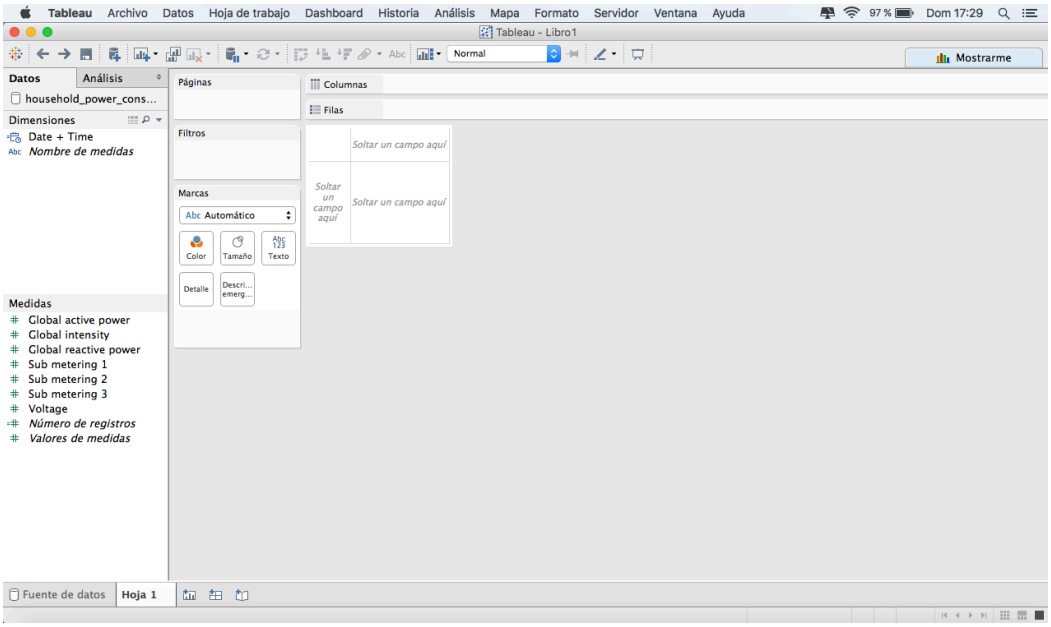


Figura 61. Configuración de la hoja de trabajo (Hoja1) en Tableau

En ella podemos diferenciar cuatro zonas principales:

Parte superior: encontramos los diferentes menús disponibles: análisis, formato, ayuda, etc. Además de comandos importantes como guardar o deshacer.

Parte izquierda: ventana de datos. Se divide en dimensiones y medidas. En ella aparecen todos los datos de que se dispone y a partir de los que cuáles se podrán crear las visualizaciones mediante la selección de los mismos con la fórmula "arrastrar y soltar" hasta los diferentes estantes.

Parte central: zona de estantes donde se soltarán los elementos seleccionados. Las visualizaciones sufrirán variaciones en función de la configuración seleccionada. Existen cinco estantes: *Columnas*, *Filas*, *Filtro*, *Páginas* y *Marcas*.

Parte inferior: pestañas de hojas, dashboards e historias. Se usa para renombrar las hojas, duplicarlas, cambiar su orden y muchas otras funciones.

Tableau funciona de un modo interactivo, por lo que generar un gráfico es tan sencillo como arrastrar con el ratón el elemento hasta el estante deseado. La dificultad por lo tanto, no reside en el manejo del programa, si no en la correcta elección de los gráficos.

Videos de aprendizaje

Como recomendación, Tableau nos ofrece un breve tutorial formado por una serie de vídeos en los que podremos consultar dudas acerca del uso de la herramienta en la dirección: http://www.tableau.com/es-es/learn