**Proyectos II, integración y preparación de datos**

**Primera presentación:**

**Captura de datos**

**Introducción**

Nombres y apellidos de los autores:

| Álvaro | Prado Expósito |
| --- | --- |
| Óscar | García Martínez |
| Aleixandre | Tarrasó Sorní |
| Víctor | Máñez Poveda |
| Jorge | Muedra Vela |

**1. Las Fichas de Configuración**

* 1. **Alcance (preliminar)**

De las distintas fuentes que vemos aquí queremos obtener las variables de la energía total demandada, el precio de la electricidad por día (o por mes, en función de cómo se hayan recogido los datos), la producción total de energía separada por sectores, el nivel de exportación y de importación energético, datos económicos de cada país tales como el IPC, el IPI para poder relacionar el consumo con la producción industrial.

| Identificador de la Ficha | **Búsqueda de fuentes** |
| --- | --- |
| *Relatad todas las fuentes que habéis investigado y consultado en el proceso de localización de las que vais finalmente a utilizar.* | ***España:***  [*Red Eléctrica Española*](https://www.ree.es/es/apidatos)  [*Instituto Nacional de Estadística (INE)*](https://www.ine.es/index.htm)  [*AEMET*](https://datosclima.es/index.htm)  ***Francia:***  [*Demanda Energética en Francia*](https://www.services-rte.com/en/download-data-published-by-rte.html?category=market&type=signal)  [*Ministère de la Transition Écologique et de la Cohésion des Territoires*](https://www.statistiques.developpement-durable.gouv.fr/)  [*INSEE (IPC)*](https://www.insee.fr/en/statistiques/serie/001759967#Telechargement)  [*Demanda Energética por Sectores*](https://odre.opendatasoft.com/explore/?disjunctive.theme&disjunctive.energie&disjunctive.maille-geographique&disjunctive.publisher&disjunctive.frequence-de-mise-a-jour&disjunctive.pas-temporel&sort=modified&refine.energie=%C3%89lectricit%C3%A9)  [*INSEE (IPI)*](https://www.insee.fr/en/statistiques/serie/010767574#Revision)  [IPI Francia](https://data.oecd.org/industry/industrial-production.htm)  [INSEE (Empleo)](https://www.insee.fr/en/statistiques/serie/010605852#Telechargement)  ***Portugal:***  [*Red eléctrica portuguesa (REN)*](https://datahub.ren.pt/pt/redes/)  [IPC Portugal](https://fred.stlouisfed.org/series/PRTCPIALLMINMEI)  [IPI Portugal](https://data.oecd.org/industry/industrial-production.htm)  [Empleo por sector en Portugal](https://data.oecd.org/emp/employment-by-activity.htm)  ***Italia:***  [*Red eléctrica italiana (TERNA)*](https://www.terna.it/en/electric-system/transparency-report/download-center)  [*Instituto de Estadística de Italia*](https://www.istat.it/)  ***Finlandia:***  [*Red eléctrica finlandesa (FINGRID)*](https://data.fingrid.fi/en)  [*Estadísticas Finlandia*](https://stat.fi/tup/tilastotietokannat/index_en.html)  ***Estonia:***  [*Red eléctrica estonia*](https://dashboard.elering.ee/assets/api-doc.html#/)  [*Estadísticas de Estonia*](https://www.stat.ee/en)  ***Materias primas:***  [*Investing*](https://es.investing.com/commodities/)  [*Yahoo Finance*](https://es.finance.yahoo.com/materias-primas/)  ***Balances energéticos:***  [*Eurostat*](https://data.europa.eu/data/datasets/information-on-energy-markets-in-eu-countries-with-national-energy-profiles?locale=es) |
| *Criterios seguidos para la selección de las fuentes que se van a usar para el proyecto.* | *Hemos obtenido datos de las redes eléctricas de distintos países de Europa (porque tenían datos abiertos). Los países de los que hemos podido extraer información son España, Francia, Italia, Portugal, Dinamarca, Finlandia y Estonia.*  *Con ellos buscamos obtener información a nivel general de todo el continente, así podremos separar por zonas y compararlas entre sí.*  *Los criterios de búsqueda que hemos seguido para seleccionar las fuentes de datos han sido principalmente que los ficheros de datos tuviesen suficientes valores como para poder trabajar con ellos. Para poder obtenerlos hemos tenido que descargar directamente los ficheros de datos, utilizar las APIs que nos daba la página web mediante peticiones GET… Además, cada acceso a la base de datos era distinto y muchas veces problemático, por ejemplo en la red eléctrica francesa, había que registrarse para acceder a la información.*  *La búsqueda y obtención ha sido difícil ya que muchas páginas tenían los datos, pero no se podían descargar, o había otro nivel de grado, o muchos valores faltantes, etc.*  *Finalmente, hemos elegido aquellos Data Frames que mejor se podían integrar bien con nuestros datos para poder cruzarlos en nuestros análisis.* |

* 1. **Técnicas de obtención de datos y extracción**

Explica las técnicas utilizadas en la obtención de las fuentes de datos. Especialmente, explica si has utilizado y para qué:

* + Descarga de ficheros .csv
  + Descarga desde una URL
  + Lectura de tablas incrustadas en HTML
  + Conversión de .JSON
  + Conversión de .XML
  + Recoger datos de Twitter y limpieza sobre expresiones regulares
  + Recoger datos de Google y limpieza sobre expresiones regulares
  + Web scraping
  + Descargas en tiempo real durante varios ciclos

Para recopilar los datos necesarios para nuestra base de datos, realizamos una investigación exhaustiva a través de las diversas páginas web de las redes eléctricas de cada país. Este proceso implicó enfrentarnos a una variedad de formatos de datos, los cuales describimos a continuación:

* **Descarga de archivos en formato .csv:** En algunos casos, encontramos que los datos estaban disponibles para su descarga en archivos con extensión .csv. Este fue el caso, por ejemplo, en Francia, Italia y Dinamarca. Descargamos estos archivos y los utilizamos como parte de nuestra base de datos. Esto nos ayudó a que la descarga fuese más rápida y el almacenamiento más ligero que al usar otro tipo de datos.
* **Descarga de archivos en formato .xlsx:** En otras ocasiones, los datos estaban presentados en archivos de Excel con extensión .xlsx. Esto fue común en países como Finlandia, Portugal y en algunos conjuntos de datos de Francia. Descargamos y procesamos estos archivos para incluir la información relevante en nuestra base de datos.
* **Obtención de datos a través de APIs:** En ciertos casos, los datos estaban disponibles a través de APIs. Utilizamos esta metodología para obtener datos de países como España y Estonia. Desarrollamos un programa en Python para interactuar con estas APIs y recuperar todos los datos relacionados con el país respectivo. Los datos obtenidos de esta manera se integraron también en nuestra base de datos.
* **Extracción de datos de tablas HTML**: Cuando encontramos tablas en páginas web con información que no está disponible en archivo descargable o de manera web scraping, como el IPI de Portugal, pasamos esa tabla a otro fichero CSV, trabajamos con él y lo formateamos mediante Python, R o Excel. Requiere analizar el HTML para extraer la información de manera eficiente.
  1. **Análisis de las fuentes: interpretación de los datos, valoración de su utilidad en el proyecto.**

Las fuentes de las que descargamos los datos se pueden dividir en dos grupos, fuentes relacionadas a las redes eléctricas y fuentes ajenas a ellas. De las redes eléctricas extraemos el precio del MWH, la demanda, la generación energética y su origen, y el saldo (importación y exportación) internacional. Además, en determinadas redes eléctricas también hemos extraído datos sobre la demanda del sector industrial. Consideramos que estas variables son las más útiles a la hora de analizar el funcionamiento del mercado energético de un país, ya que definen todos los factores que afectan al precio y la disponibilidad de la energía.

Por otro lado, de las fuentes ajenas a las redes eléctricas obtenemos datos económicos y meteorológicos que ayudan a discernir las necesidades energéticas de cada región. Estos datos son imprescindibles para nuestro estudio porque sin ellos es imposible explicar las causas que moldean el mercado energético.

Además hemos recopilado información sobre materias primas, como el precio del gas, el precio del petróleo, el precio del uranio, y el precio del carbón, para poder tener un análisis más completo.

* 1. **Análisis de los campos, formatos y tipo de información de cada fuente y valoración del cruce de datos de distintas fuentes para nuestro proyecto.**

España

* Fecha: Formato yyyy-mm-dd
* Precio: €/MWH, horario
* Demanda: MWH, diario
* Generación : MWH, diario
  + Hidráulica
  + Turbinación de bombeo
  + Nuclear
  + Carbón
  + Diésel
  + Turbina de gas
  + Ciclo combinado
  + Eólica
  + Solar Fotovoltaica
  + Solar Térmica
  + Otras Renovables
  + Cogeneración
  + Residuos no renovables
  + Residuos renovables
  + Generación Total
* IPI (Índice de Producción Industrial): mensual
* Empleo: miles de personas, trimestral
* Porcentaje de empleo industrial, trimestral
* PIB per cápita: €, trimestral

Francia

* Fecha (Días): Formato dd-mm-yyyy
* Demanda: MWH, diario
* Fecha (Meses): Formato yyyy-mm
* Precio: €/MWH, semestral
* Generación (Todas las fuentes) GWH, mensual
* Importación-Exportación Energía: GWH, mensual
* IPI (Índice de Producción Industrial): Mensual
* Inflación (IPC, IPC Energía): Mensual
* Empleo: Miles de personas, trimestral

Portugal

* Fecha: Formato dd/mm/yyyy, diario
* Hidráulica: MWH, diario
* Eólica: MWH, diario
* Solar: MWH, diario
* Biomasa: MWH, diario
* Ondas: MWH, diario
* Gas natural ciclo combinado: MWH, diario
* Gas natural: MWH, diario
* Carbón: MWH, diario
* Otra térmica: MWH, diario
* Importación: MWH, diario
* Exportación: MWH, diario
* Turbinación bombeo: MWH, diario
* Demanda: MWH, diario
* Inflación Portugal:(IPC) Índice, mensual
* IPI PORTUGAL: Índice, mensual
* EMPLEO AGRICULTURA: Miles de personas, trimestral
* EMPLEO INDUSTRIA: Miles de personas, trimestral
* Empleo servicios: Miles de personas, trimestral
* Nivel Pantanos: hm^3, semanal
* Temperatura Máxima: ºC, diario

Estonia

* Fecha: Formato dd/mm/yyyy, mensual
* Precio de compra electricidad: MWH, mensual
* Precio de venta electricidad: MWH, mensual
* Precio de compra de gas: MWH, mensual
* Precio de venta de gas: MWH, mensual
* Importación total: MWH, mensual
* Importación de Letonia: MWH, mensual
* Importación de Finlandia: MWH, mensual
* Importación por intercambio: MWH, mensual
* Exportación total: MWH, mensual
* Exportación a Letonia: MWH, mensual
* Exportación a Finlandia: MWH, mensual
* Exportación por intercambio: MWH, mensual
* Exportación por contratos bilaterales: MWH, mensual
* Total energía renovable: MWH, mensual
* Energía eólica:MWH, mensual
* Energía hidroeléctrica: MWH, mensual
* Energía por biomasa: MWH, mensual
* Total energía no renovable: MWH, mensual
* Energía renovable solar: MWH, mensual
* Nivel Pantanos: hm^3, semanal
* Temperatura Máxima: ºC, diario

Dinamarca

* Fecha: Formato dd/mm/yyyy
* Precio: MWH, horario
* Demanda: MWH, horario
* Generación (todas las fuentes): MWH, horario
* Saldo energético con varios países: MWH, horario
* Demanda industrial: MWH mensual
* IPI: Índice base 100, mensual
* Inflación: (IPC) Índice base 100, mensual
* PIB per cápita
* Nivel Pantanos: hm^3, semanal
* Temperatura Máxima: ºC, diario

Finlandia

* Fecha: Formato yyyy-mm-dd
* Precio: €/MWH, horario
* Demanda: MWH, cada 15 minutos
* Generación (todas las fuentes): MWH, cada 15 minutos
* Saldo energético internacional: MWH, cada 15 minutos
* IPI: Índice base 100, mensual
* Inflación: (IPC) Índice base 100, mensual
* PIB per cápita
* Nivel Pantanos: hm^3, semanal
* Temperatura Máxima: ºC, diario

Italia

* Fecha: Formato dd/mm/yyyy
* Demanda: MWH, cada 15 minutos
* Generación (todas las fuentes): MWH, cada 15 minutos
* Saldo energético internacional: MWH, cada 15 minutos
* Índice de Demanda Industrial: Índice base 100, mensual
* IPI: Índice base 100, mensual
* Inflación: (IPC) Índice base 100, mensual
* PIB per cápita
* Nivel Pantanos: hm^3, semanal
* Temperatura Máxima: ºC, diario

Eurostat

* Energía: Mtoe, anual
  + Solid fossil fuels
  + Oil and petroleum products
  + Natural gas
  + Nuclear
  + Renewables and biofuels
  + Others
  + Total energy supply
  + Available for final consumption
  + Final energy consumption
  + Transformation Losses
  + Energy Sector
  + Distribution losses
  + Final non-energy consumption
  + Industry
  + Transport
  + Residential
  + Services
  + Agriculture and fishing

Debido a la magnitud del estudio y a la restricción temporal hemos tenido que acotar nuestro estudio y no nos ha sido posible abarcar todos los países planteados en el hito, nos centraremos en los datos respectivos a España. No obstante, pese a que hubiese sido excitante introducir los datos de otras regiones, los datos de los que disponíamos para el territorio español nos han conseguido alcanzar nuestro objetivo principal que era el de entender el funcionamiento de los sistemas energéticos y su estructura de generación, junto al de crear un modelo que nos ayude a predecir los valores de precio futuros.