Autores: Gonzalo Canales, Aida Centelles

1. Contexto

La información se ha extraído de la web www.embalses.net. Esta web contiene información actualizada semanalmente del estado de las reservas hídricas en cada uno de los embalses del territorio nacional de España. Esta web recoge la información del Ministerio de Transición Ecológica a través de su boletín hidrológico.

Esta página web permite un acceso fácil y directo a la información relativa a todos los embalses de España y ofrece diferentes formas de alcanzarla (por cuenca hidrográfica, provincia, comunidad, pluviómetros). De manera adicional, guarda un histórico a modo de gráfico de la evolución de todas las semanas del año de la cantidad de agua embalsada, que se puede comparar con los mismos datos del año anterior y de hace 10 años.

La razón por la que existe esta web es fundamentalmente porque ofrece un acceso muy sencillo e intuitivo a esta información, mucho más fácil de usar y localizar que la ofrecida en el Boletín Hidrológico del Ministerio de Transición Ecológica.

2. Título del dataset

Hemos decidido llamarlo Reservoirs/Embalses pues consideramos que es el nombre más adecuado habida cuenta de que contiene información de las reservas hídricas de los embalses así como otra información relevante a los mismos.

La limitación más importante de este dataset es que contiene información relativa a los embalses para una semana concreta del año. Si se quisiera tener una visión más holística del estado de los embalses a lo largo de un año, el script debería ejecutarse de forma semanal durante todo un año y agregar la información al dataset.

Los datos extraídos siempre son consistentes, y contienen los tipos de datos descritos en los apartados 3 y 5 de esta práctica. Una vez extraídos, los datos no necesitan limpieza posterior para su análisis. Las diferentes variables referentes al uso de los pantanos, sin embargo, son booleanas (0/1) y, según el tipo de estudio, podrían ser transformadas o agregadas en una única variable.

3. Descripción del dataset

El dataset contiene información acerca de las reservas hídricas de los diferentes embalses del territorio nacional español.

- Datos cuantitativos:

  Estos datos contienen la cantidad de agua embalsada media en hectómetros cúbicos y su porcentaje durante la semana del año en la que nos encontramos. Además, como referencia, también contiene la cantidad embalsada y su porcentaje respecto a la misma semana del año anterior, y la misma semana de hace diez años. Adicionalmente, contiene la capacidad total del embalse.

- Datos cualitativos:

Son datos referentes a la ubicación del embalse (Cuenca hidrográfica, provincia y municipio), río que abastece de agua al embalse, tipo de construcción y año de construcción.

- Información de uso:

  Estos datos indican cuales son los principales usos del embalse (abastecimiento, industrial, navegación, baño, pesca, etc..). Esta información se guarda en un valor booleano para cada una de las categorías.

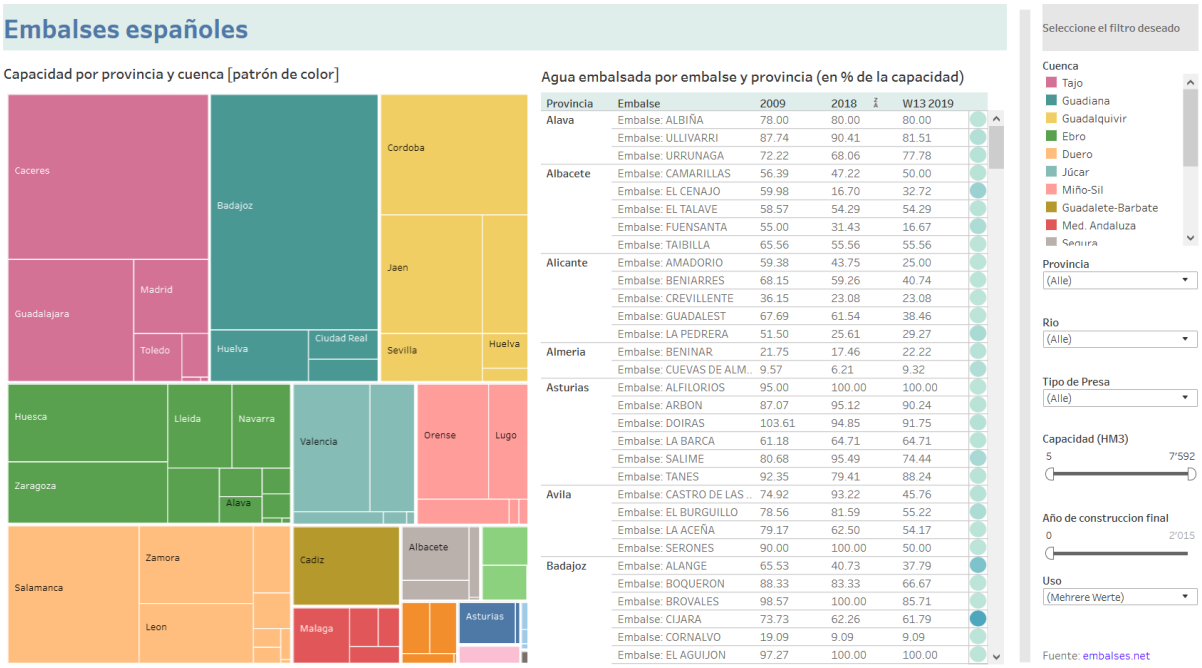
4. Representación Gráfica

Si se tiene en consideración el volumen de datos del dataset, se puede representar de diversas formas.

En nuestro caso hemos decidido crear gráficos interactivos en Tableau que permiten al usuario seleccionar diferentes filtros como por ejemplo: provincia, cuenca, rio, tipo de presa o uso del embalse. La publicación en Tableau se puede encontrar bajo este enlace: https://public.tableau.com/profile/aida.c3420#!/vizhome/EmbalsesReservoirsSpain/Maindashboard

De entre todas las gráficas, hemos decidido incluir ésta en la memoria ,ya que consideramos que es la que mejor explica el proyecto y el dataset:

- Diagrama de árbol, con información sobre la capacidad de los embalses, según las cuencas y provincia (con un patrón de color).
- A la derecha, una lista con el porcentaje de agua embalsada por provincia y embalse, en % respecto a la capacidad total, en 2009 (hace 10 años), en 2018 (hace 1 año) y recientemente en la semana 13 del 2019. Es fácil ver la evolución de la capacidad en el tiempo.



5. Contenido del dataset

Los campos contenidos en el dataset son los siguientes:

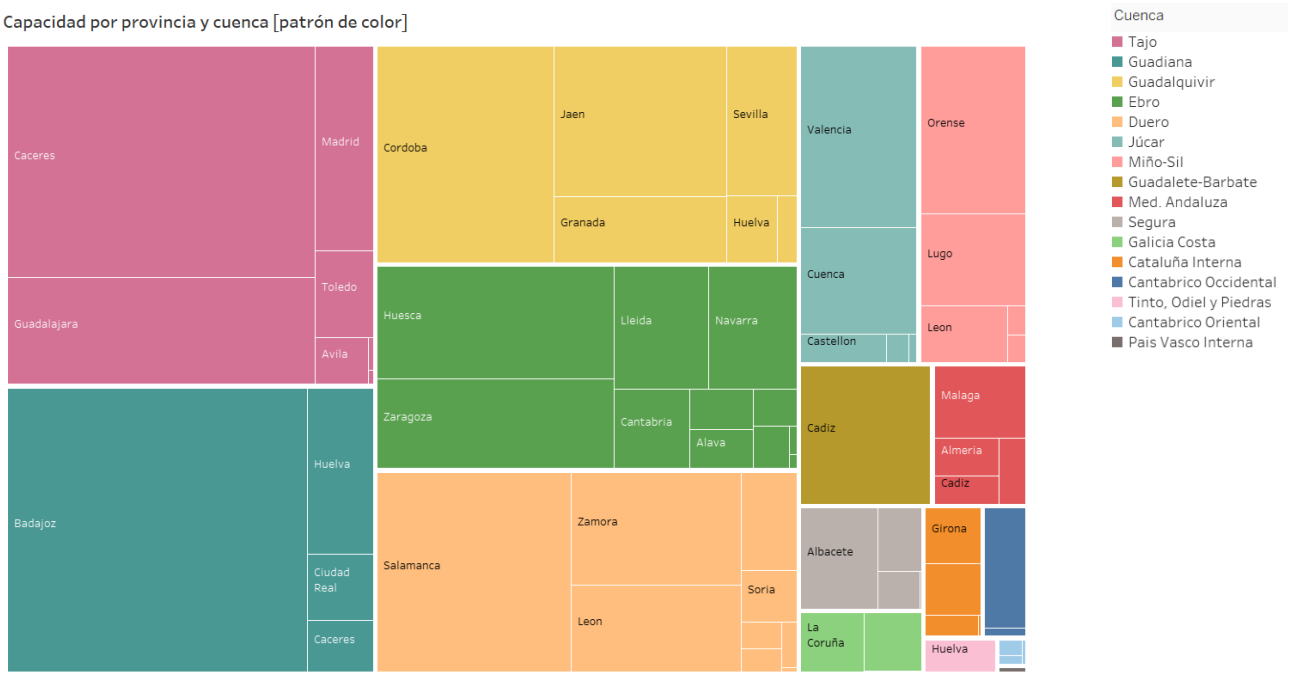
- **Nombre del embalse:** Atributo que distingue unívocamente el embalse.
- **Agua embalsada (HM3):** Cantidad de agua media embalsada durante la última semana en hectómetros cúbicos.
- **Porcentaje:** Porcentaje de agua embalsada durante la última semana con respecto a su capacidad total.
- **Variación respecto a semana anterior (HM3):** Diferencia de agua embalsada con respecto a la semana anterior en hectómetros cúbicos.
- **Porcentaje Capacidad:** Mismo dato que el anterior, pero en porcentaje con respecto a la capacidad total
- **Capacidad:** Capacidad total del embalse en hectómetros cúbicos.
- **Misma Semana (año anterior) (HM3):** Cantidad de agua embalsada en la misma semana del año anterior, en hectómetros cúbicos.
- **Porcentaje Cuenca :** Misma información que en el campo anterior, mostrada en porcentaje con respecto a la capacidad total.
- **Misma Semana (Med. 10 Años) (HM3):** Cantidad de agua embalsada en la misma semana de hace 10 años, en hectómetros cúbicos.
- **Porcentaje:** Misma información que en el campo anterior, mostrada en porcentaje con respecto a la capacidad total
- **Cuenca:** Cuenca hidrográfica a la que pertenece el embalse.
- **Provincia:** Provincia en la que se encuentra el embalse.
- **Municipio:** Municipio al que pertenece
- **Río:** Río que abastece de agua al embalse.
- **Tipo de Presa:** Tipo de construcción de la presa (gravedad, bóveda, materiales sueltos,etc…)
- **Año de construcción:** Año en el que la presa fue construida.
- **Superficie:** Superficie en hectáreas.

Adicionalmente, se guarda información del uso del embalse. Se incluye un valor booleano para indicar si el embalse tiene alguno (o varios) de los siguientes usos:

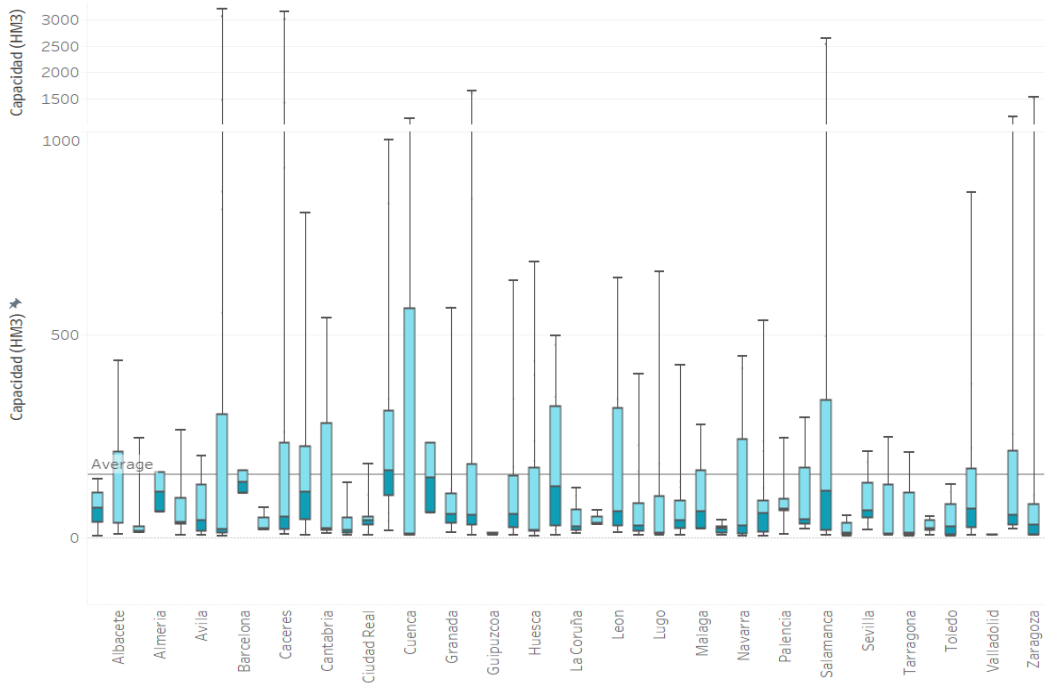
- **Abastecimiento**
- **Riego**
- **Electricidad**
- **Industrial**
- **Pesca**
- **Navegación**
- **Baño**
- **Pic-nic**
- **Restaurantes**

El periodo de tiempo representa de una semana, ya que la web de la que se han extraído los datos actualiza la información semanalmente (cada domingo). Los datos presentados pertenecen a la semana del 25 al 31 de Marzo de 2019.

Un diagrama de árbol o treemap nos muestra los datos estructurados del dataset en forma de árbol, por jerarquía, como un conjunto de rectángulos anidados. Al existir una correlación entre el color (que representa cada cuenca) y el tamaño (que representa la capacidad en HM3), el diagrama ayuda a identificar patrones que, de otro modo, serían difíciles de ver.



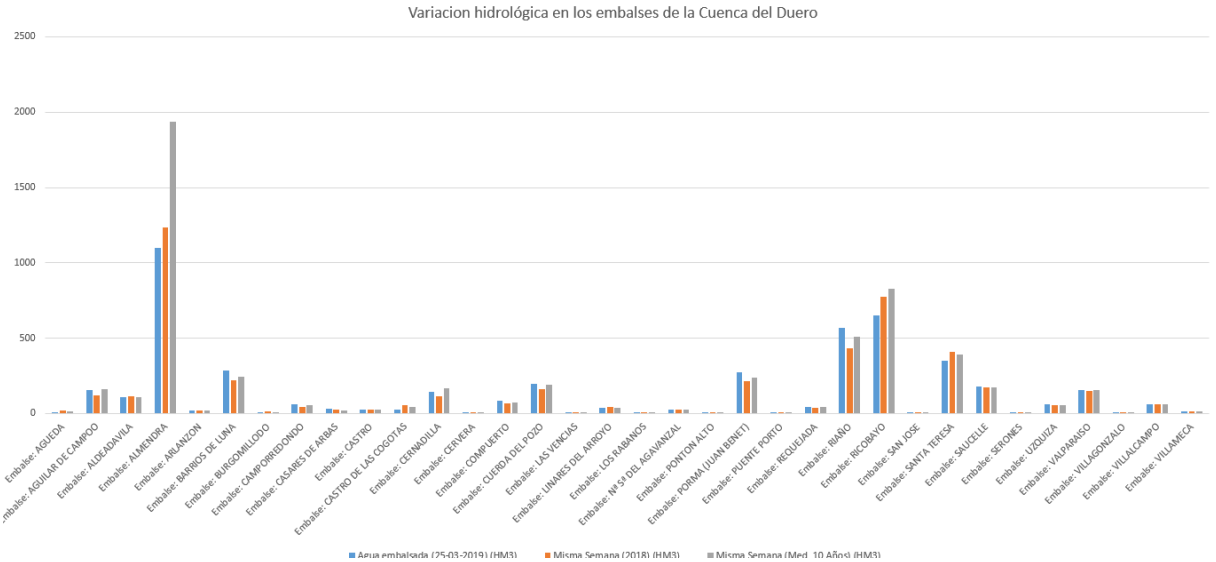
También podemos visualizar la capacidad por provincia, y ver la distribución de valores:



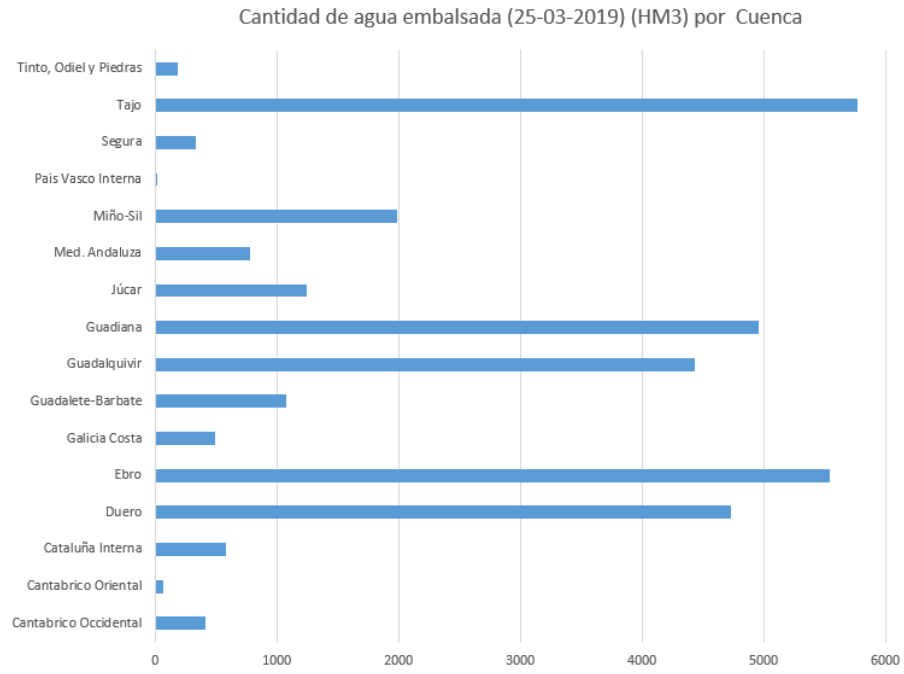
La media se sitúa en torno a los 150HM3, con provincias como Cuenca o Cantabria por encima de la media.

Igualmente, podemos representar los datos en Excel. Una manera de representarlos es mediante un diagrama de barras que muestre el agua embalsada con respecto al mismo dato la semana anterior de hace un año y respecto a hace 10 años. Esta información puede servir como referencia para indicar la tendencia.

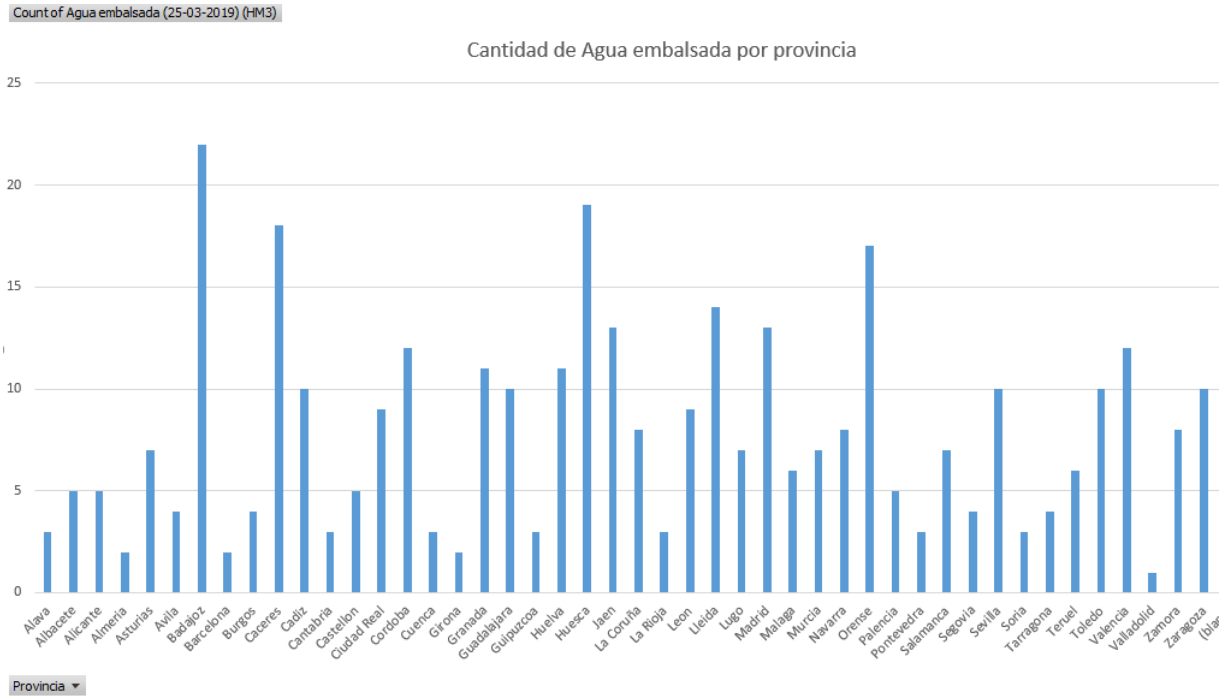
Debido al gran número de embalses, en el diagrama que se muestra debajo únicamente se han incluido los embalses de la cuenca del Duero.



Pero también se puede agrupar de otras formas para, por ejemplo, obtener la cantidad de agua total embalsada por cuenca hidrográfica:



O por provincia



Para obtener los datos, el script de scraping (embalses.py), se conecta a la página principal de la web www.embalses.net, busca la página que contiene todas las cuencas y a partir de éstas obtiene los links a cada uno de los embalse, a continuación, accede a cada uno de estos y extrae la información.

El script esta íntegramente codificado en lenguaje Python y hace un chequeo del fichero robots.txt para evaluar las restricciones (ninguna afecta a las paginas accedidas). Además, emula la conexión desde un navegador mediante el remplazo de la cabecera user-agent.

Para la extracción de la información de las páginas se ha utilizado la librería Python BeautifulSoup. Esta librería permite el *parseo* del fichero HTML extrayendo los *tags* relevantes para el dataset mediante la ejecución de un análisis vertical y horizontal del fichero. Los datos, una vez extraídos, se guardan en una matriz. La primera entrada incluye las cabeceras y el resto una entrada por cada embalse con toda la información descrita anteriormente.

Finalmente, se hace una limpieza y ajuste de los datos y se vuelcan al fichero embalses.csv

6. Agradecimientos

La información se ha obtenido de la web www.embalses.net que obtiene la información del Ministerio de Transición Ecológica. Dado que no hay ninguna referencia al propietario de la web embalses.net, nuestro agradecimiento va dedicado al Ministerio de Transición Ecológica y a su boletín hidrológico.

En España existen numerosos estudios e investigaciones que trabajan con bases de datos similares. En el apartado 7 se describen estudios que, a nivel nacional o internacional, se

enfocan en el análisis de diferentes datos referentes al estado de embalses (capacidad, estado, ecosistema, calidad del agua, etc.)

7. Inspiración

Consideramos esta dataset interesante porque puede ser usado en diversas áreas.

Por ejemplo, puede ser usado por investigadores para demostrar el efecto del cambio climático en las reservas hídricas en España. Si se guarda un histórico de datos durante un intervalo de tiempo, se podría observar la evolución de forma muy precisa.

También puede tener uso empresarial con el objetivo de encontrar nuevas oportunidades de negocio: considerando los campos relativos al uso de los embalses, se puede observar que hay varios de ellos dedicados al baño, pesca y/o navegación que carecen de áreas de pic-nic o restauración.

De hecho, hay sitios web que ofrecen una información muy similar como la ofrecida por el Departamento de Recursos Hídricos del Estado de California. En este site se mantiene y opera una extensiva cantidad de información hidrólogica (estado de embalses, ríos, nieve acumulada) que permite controlar y prever sequías e inundaciones.

En España existen numerosos estudios e investigaciones que trabajan con bases de datos similares (datos referentes a la capacidad, agua embalsada, uso y estado de los embalses) como el proporcionado por embalses.net.

Un ejemplo reciente es el estudio sobe la predicción del impacto del cambio climático sobre la cuenca de ríos y vida útil de los embalses en España, llevado a cabo por un equipo de investigadores del Instituto de Ingeniería del Agua y Medio Ambiente (IIAMA) de la Universitat Politècnica de València en 2018 (Fuente: https://www.iiama.upv.es/iiama/es/sala-prensa/noticias/tecnica-que-determina-el-impacto-del-cambio-climatico-sobre-el-ciclo-del-agua-y-de-sedimentos.html).

Existen otros ejemplos de estudios enfocados en la evaluación del potencial ecológico de los ecosistemas en cuencas y pantanos, como el llevado a cabo por la Confederación Hidrográfica del Ebro en 20143 (Fuente: https://www.iagua.es/noticias/biodiversidad/13/09/02/un-estudio-identifica-especies-de-peces-en-embalses-donde-su-existencia-no-era-conocida-35741). En este estudio se realizó una evaluación cuantitativa de las poblaciones de peces de los embalses de la Cuenca. En el estudio se cita explícitamente embalses.net como una de las fuentes de datos utilizadas (link al estudio https://studylib.es/doc/5018624/estudio-de-las-poblaciones-de-peces-en-embalses-de).

Otro ejemplo, en este caso, que aunque no se sirvió de datos de embalses.net se sirvió de técnicas de *web scraping* para obtener datos de embalses, es el estudio “Web scraping for automated water quality monitoring system: A case study of PDAM Surabaya” (Fuente: https://ieeexplore.ieee.org/document/7828735). En este caso el tipo de datos extraídos son datos sobre la calidad del agua, que son medidos por sensores. Para el estudio se desarrolló una aplicación de data logger que recolecta los datos de estos sensores mediante técnicas de web scraping. Esto es necesario para recolectar datos de forma periódica, ya que los sensores no almacenan los datos.

En general, la página web embalses.net está considerada como la mejor fuente para conocer en tiempo real el estado de los embalses de las comunidades autónomas en España.

La base de datos que hemos creado en este proyecto puede ser de gran valor para estudios que necesiten disponer de este tipo de datos de manera periódica y actual, ya que extrae los valores semanalmente, cada domingo.

8. Licencia

La licencia que seleccionaríamos sería: *Released Under CC0: Public Domain License* (Fuente: https://creativecommons.org/publicdomain/zero/1.0/deed.es)

El Ministerio de Transición Ecológica ofrece acceso a estos datos al dominio público, por ello suponemos que renuncia a todos los derechos a la creación de ellos. Intuimos por ello que es permitido copiar, modificar, distribuir e interpretar la obra, incluso para propósitos comerciales, sin pedir permiso.

9. Código

El código se encuentra bajo:

https://github.com/Acent1/UOC_Webscraping_reservoirs/tree/master/src

10. Dataset

El data set se encuentra bajo:

https://github.com/Acent1/UOC_Webscraping_reservoirs/tree/master/data

Contribuciones

Contribuciones	Firma
Investigación Previa	AC, GC
Redacción de las respuestas	AC ,GC
Deasarrollo del código	AC, GC