Universidad de Valladolid

Diseño y Evaluación de Sistemas Interactivos

Análisis y Visualización de Datos MedioAmbientales de Castilla y León

Sergio García Prado

Seguimiento del trabajo en:
github.com/garciparedes/DatosMedioambientalesCYL

December 10, 2015

Contents

1 Introducción

1.1 Motivación

La visualización que se va a realizar consiste en la representación gráfica de datos medioambientales de la Comunidad Autónoma de Castilla y León. El objetivo de la práctica es descubrir aspectos poco obvios o destacar algo importante del conjunto de datos asignados.

1.2 Objetivos

La representación gráfica que se va a realizar pretende mostrar el flujo energético en Castilla y León durante los últimos años. Además una vez se pueda visualizar la implementación se pretenden obtener conclusiones sobre el impacto que ha tenido el cierre de la Centrar Nuclear de Santa María de Garoña (Burgos) en el año 2012 y representar cómo ha solventado la región ese déficit energético (si es que lo ha hecho). También se pretende analizar si este suceso está ha afectado al consumo energético en la zona.

2 Análisis de los datos

2.1 Descripción

La Junta de Castilla y León suministra conjuntos de datos referentes a la comunidad autónoma. Para esta práctica se nos han asignado los datos medioambientales.

2.2 Adquicisión

Los datos han sido obtenidos a partir de la iniciativa de Datos Abiertos de la Junta de Castilla y León.

El concepto "datos abiertos" es una filosofía y práctica que persigue que determinados tipos de datos estén disponibles de forma libre para todo el mundo, sin restricciones de derechos de autor, de patentes o de otros mecanismos de control. Tiene una ética similar a otros movimientos y comunidades abiertos, como el software libre, el código abierto y el acceso libre.

Los datos han sido obtenidos a partir del siguiente enlace: www.datosabiertos.jcyl.es/...

Una vez hemos obtenido los datos procederemos a realizar el análisis de los mismos.

2.3 Estructura de los datos

Los datos que vamos a manejar se encuentran en fichero con formato CSV (un tipo de documento en formato abierto sencillo para representar datos en forma de tabla, en las que las columnas se separan por comas (o punto y coma) y las filas por saltos de línea.), el cual, a pesar de tener una estructura eficiente para ser manejado por máquinas, complica la lectura y extracción de información para los humanos. Además la cantidad de información contenida en este grande (más de 5000 elementos).

A pesar de estas dificultades, fijándonos en la cabecera del fichero, la cual nos muestra información de la estructura, podemos conocer lo campos que nos encontraremos en cada una de las filas que contiene:

- **Indicador** Contiene información sobre a qué se refiere el dato. Representa una variable de tipo cualitativo nominal. Tras realizar una extracción de datos, hemos obtenido que hay 49 indicadores distintos. Debido a que la cantidad es numerosa habrá que tratar de seleccionar los que más se adecuen a lo que queremos mostrar debido a que mostrar demasiada información de manera inadecuada dificulta la visión de lo que se pretende representar.

- **Provincia** Provincia a la que está referida la información. Este campo es de tipo cualitativo nominal. Se refiere a la segmentación territorial de la comunidad autónoma. Como sabemos, Castilla y León está dividida en 9 provincias. Conviene señalar que hay indicadores que no poseen información para todas las provincias, como es el caso del "Tráfico aéreo de pasajeros", que suponemos que restringe la información a las provincias que disponen de aeropuerto ya que el resto tendrían valor 0 y no tiene sentido indicarlo. Este hecho es algo a tener en cuenta cuando representemos la información.

- **Fecha Validez** Fecha hasta la cual el dato se consideraba como actual. Representa una variable cuantitativa debido a que es una variable temporal. A pesar de ello toma únicamente valores discretos (años). Esto es algo a tener en consideración cuando determinemos la manera en que será representado.

- **Valor** Valor que toma el dato respecto de su unidad. Representa una variable cuantitativa de tipo continuo. Este campo es el que contiene la "información valiosa", ya que es lo que se utilizará como medida para compararla con los demás. Se representa en forma de número racional positivo.

- **Unidad** Este campo nos proporciona la unidad de medida del dato, lo cual es algo muy importante a la hora de mostrar información ya que siempre se deben señalar tanto la unidad de medida como la escala a la que corresponde la información representada.

- **Frecuencia** Este campo representa la periodicidad con la que se ha obtenido el dato. La mayoría de los datos se han obtenido anualmente, aunque también existen algunos que se han obtenido bianualmente, como es el caso de los "Espacios Naturales".

2.4 Calidad

A mi juicio la calidad de los datos es buena a pesar de que existen algunos puntos mejorables.

Me llama la atención la columna de "meses" que, entiendo que está ahí porque es necesario especificar la unidad de medida de la frecuencia, pero creo que se debería haberse indicado en la cabecera del fichero. Además también he encontrado valores atípicos en algunos datos, posiblemente producidos por un error en las mediciones o al transcribirlo al documento.

2.5 Selección de datos

Ahora determinaremos los datos que se utilizarán para la visualización. Como se dijo al principio del informe, el objetivo de la visualización será visualizar cómo a repercutido el cierre de la central nuclear en la región. Para ello nos vamos a apoyar en los siguientes indicadores:

- Producción de energía con carbón
- Producción de energía eólica
- Producción de energía hidráulica
- Producción de energía nuclear
- Producción de energía primaria
- Producción energía solar en Castilla y León
- Consumo de energía final
- Consumo de energía del sector del transporte
- Consumo de energía del sector
- Consumo doméstico de electricidad
- Consumo doméstico de gas natural
- Consumo doméstico de G.L.P.
- Consumo doméstico de productos petrolíferos

De cada uno de estos indicadores se pretenten utilizar tanto la información de todas las provincias como fechas de validez.

2.6 Transformaciones

Para simplificar la visualización realizaremos varias transformaciones en los datos:

Fusionaremos estos indicadores en uno nuevo que denominaremos "Producción de energía final" cuyo valor será el sumatorio de todos los métodos de producción (Carbón, Eólica, Hidráulica, Nuclear, Primaria y Solar). La unidad de medida de todos ellos es "Toneladas Equivalentes de Petróleo" por lo cual no habrá que realizar ningún cambio de unidad para poder fusionarlos.

Para clarificar el tipo de energía que se utiliza se ha decidido crear dos otros dos nuevos indicadores denominados "Producción de Energía Renovable" y "Producción de Energía No Renovable" que se formarán a partir del sumatorio de sus respectivos métodos de energía.

También fusionaremos los indicadores relacionados con el consumo doméstico en uno nuevo que denominaremos "Consumo Doméstico" cuyo valor será el sumatorio de estos (Electricidad, Gas Natural, G.L.P. y Productos Petrolíferos). En este caso la información no tiene la misma unidad de medida para todos ellos, por lo que habrá que adaptar cada uno de ellos a la misma (Toneladas Equivalentes de Petróleo) para luego crear el nuevo indicador a partir de estos. El motivo de convertirlos en Toneladas Equivalentes de Petróleo nos servirá para luego compararlos con los indicadores de producción energética.

Por último convertiremos todas las variables que no están dadas en "Toneladas Equivalentes de Petróleo" para así poder compararlas fielmente entre sí.

3 Planificación de la información

3.1 Propósito de la visualización

El proyecto de visualización se va a enfocar en el sector energético. El objetivo de la visualización es mostrar la relación entre la producción y el consumo energético de la comunidad autónoma. Además de mostrar la proporción tanto de los diferentes métodos de producción como de las formas de consumo y cómo estas han variado durante el tiempo.

También se pretende obtener conclusiones acerca de si el déficit producido por el cierre de la central nuclear de Santa María de Garoña (Burgos) en la producción energética de la comunidad autónoma ha sido compensado con el incremento de la producción mediante otros métodos alternativos. También queremos analizar si este déficit a afectado al consumo

energético, que se intuye que no ya que probablemente se habrá importado de comunidades colindantes.

3.2 Factores que afectan al proyecto

El consumo energético no está únicamente influenciado por la producción energética ya que depende muchos otros factores como la situación económica, el estado atmosférica, la actividad industrial, etc.

4 Diseño

Primeramente se hablará sobre los colores escogidos para representar la información ya que es un tema muy importante. Dado que lo que se quiere resaltar es la dualidad entre producción y consumo de energía una primera idea fue la de utilizar los colores rojo y verde, pero esta idea se descartó debido a que en muchas culturas los tonos rojo y verde no tienen el mismo significado que en la nuestra. Es por ello que se ha decidido utilizar la metáfora de temperatura. Esto se debe a que es algo que en todas las culturas se entiende que el naranja significa calor y el azul frío. Además estos colores son complementarios entre sí, esto hace que se puedan distinguir fácilmente entre sí. Por estos motivos utilizaremos el azul para representar la producción y el naranja para el consumo.

La visualización tiene va a ser un componente que se acoplará en una página web para poder visualizar la relación entre producción y consumo, así como el desglose de estas dos variables para conocer la procedencia de cada una de ellas. Se proponen varias alternativas para la visualización.

4.1 1ª Alternativa

4.1.1 Introducción

La estructura del componente será de 3 pestañas (Relación entre Producción y Consumo, Producción y Consumo). Se ha escogido esta estrategia por que mostrar toda la información a la vez además de ser menos atractiva visualmente, puede dificultar el entendimiento de la información. A continuación procederemos a detallar cada una de las pestañas.

4.1.2 Visualización de la relación entre producción y consumo

Esta visualización será la que englobe toda la información. Para ello se utilizarán los indicadores "Consumo de energía final" y "Producción de energía final" (anteriormente hemos descrito cómo obtenerlo). La visualización se basará en dos gráficos:

- **Mapa de Coropletas** (Choropleth Map) La visualización se compondrá de un mapa de coropletas en el cual se muestrará la relación entre producción y consumo en cada una de las provincias. Este mapa además de para representar información se utilizará también para seleccionar las provincias de las que se quiere obtener información unitaria. Estás se resaltarán al clickar sobre ellas y el resto de representación del componente cambiarán. Este mapa deberá ir acompañado de su respectiva leyenda que indique la diferencia proporcional entre producción y consumo. También mostrará un texto con el nombre de cada provincia al colocar el puntero encima de él, ya que no podemos deducir que las personas que van a ver la representación se conozcan la geografía española.

- **Gráfico líneas** (Line Chart) Con este gráfico se representar la evolución temporal de la diferencia entre producción y consumo. Al colocarnos encima de el se mostrará el valor de las dos variables en ese punto.

4.1.3 Visualización de producción

- **Mapa de Coropletas** (Choropleth Map) Al igual que con el mapa de coropletas de la pestaña anterior, este serviría para representar la diferencia entre provincias en la producción de energía. Pero en este caso la únicamente la medida es la producción por provincia.

- **Gráfico apilado de areas**(Stacked Area Chart) Con este gráfico se representa el peso de cada método de producción durante el tiempo. Nos muestra una visión fiel de cómo han variado cada uno de los métodos de producción en el tiempo y cual es su peso en el aporte global de la producción. Los colores escogidos para cada una de las métodos de producción pertenecerán a una paleta de colores equidistantes.

4.1.4 Visualización de consumo

Esta pestaña tendrá el mismo contenido que la visualización de producción solo que mostrando los respectivos datos de consumo. También cambiaría el color del mapa de cloropletas a tonos naranjas, para distinguirse de la anterior pestaña y además simbolizar que el consumo es algo que se debe reducir en la media de lo posible.

Se ha creado un boceto correspondientes a esta alternativa. Este se encuentra en la figura 2 y esta compuesto por las 3 pestañas de la representación.

4.2 2ª Alternativa

4.2.1 Introducción

La segunda alternativa que se propone es no depender de pestañas para mostrar la información, es decir, mostrar todo de un único vistazo. Esta representación tiene la desventaja de que a pesar de mostrar la relación entre producción y consumo con todos sus componentes en una sola vista exige "ocultar" la medida temporal. Con ocultar nos referimos a que no se puede ver de forma estática sino que tenemos que ir avanzando el tiempo mediante un control de reproducción al estilo de un reproductor multimedia.

4.2.2 Composición

La visualización estará compuesta por los siguientes gráficos:

- **Mapa de Coropletas** (Choropleth Map) Al igual que en la anterior alternativa, se mostrará un mapa de la comunidad autónoma, una vez más para ilustrar la relación entre producción y consumo. Este tendrá las mismas características que el anterior.
- **Diagrama Sankey** (Sankey Diagram): En este caso la descomposición de los métodos de producción y las formas de consumo se ilustrarían en este mismo gráfico. La información se dividiría de la siguiente forma. A la izquierda se localizará la cantidad de producción interna y externa. Seguidamente se descompondría la producción interna con los indicadores que se poseen (eólica, carbón, nuclear, etc.) hasta llegar al centro en el que se fusionarán todos estos indicadores en consumo, el cual, de la misma manera se irá descomponiendo hacia la derecha en los indicadores que se tienen (transporte, industrial, etc.). Esto se ha representado mediante el boceto correspondiente a la figura 1.

También tendrá un "control de reproducción" para selección la fecha sobre la que queremos obtener la información

Se han creado dos bocetos correspondientes a esta alternativa. Estos son las figuras 3 y 4. Los colores que se muestran en el Diagrama Sankey de estos bocetos no se sabe si serán los definitivos, ya que esta es una imagen de ejemplo obtenida de la red. Se utilizará una paleta de tonos equidistantes, pero que no varíe mucho el diseño final, es decir, que resalte los colores de manera sutil para así no perder la estética obtenida con los tonos naranja y azul.

Dado que con esta nueva alternativa se ha perdido la visión directa de evaluación temporal a continuación se propone una nueva alternativa que solucione este defecto.

4.3 3ª Alternativa (Definitiva)

Esta tercera alternativa se compone de los mismos elementos que la segunda, pero además se le ha añadido un elemento que se utilizaba en la primera representación pero fue descartado.

Se ha añadido un **Gráfico de Líneas** (LineChart) que se además de presentar la evolución temporal de las variables producción y consumo servirá para controlar el momento temporal del resto de la representación, es decir, actuará como el "control de reproducción" de la alternativa anterior. Por lo que ya no tiene sentido que este siga en la representación.

Se ha creado un boceto correspondientes a esta alternativa. Este se encuentra en la figura 5.

5 Implementación

5.1 Bocetos

A continuación se muestran los bocetos utilizados para diseñar la representación:

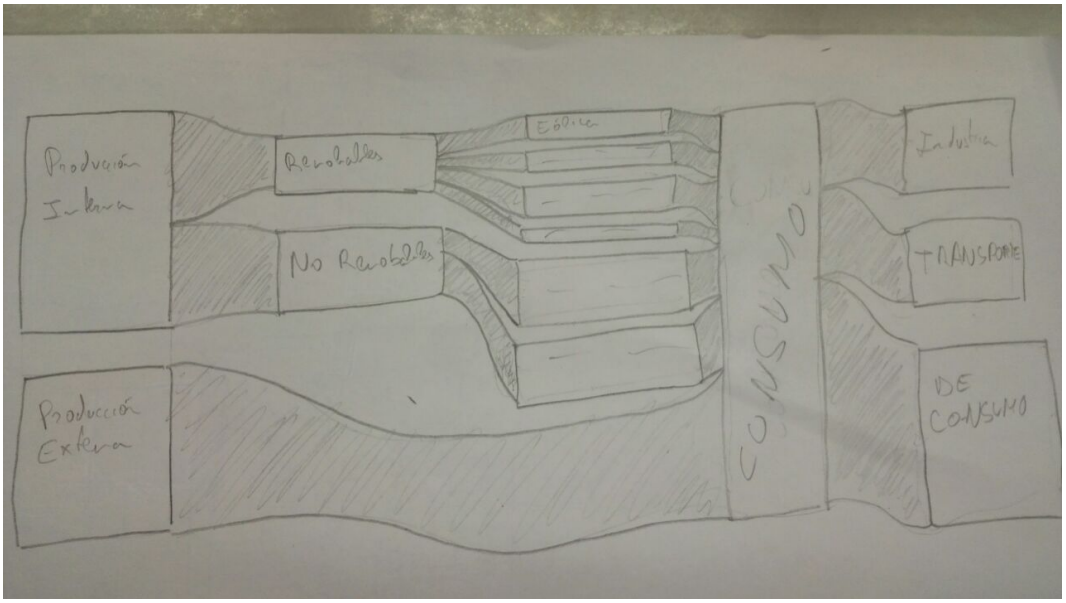

Figure 1: Alternativa 2 - Boceto Sankey

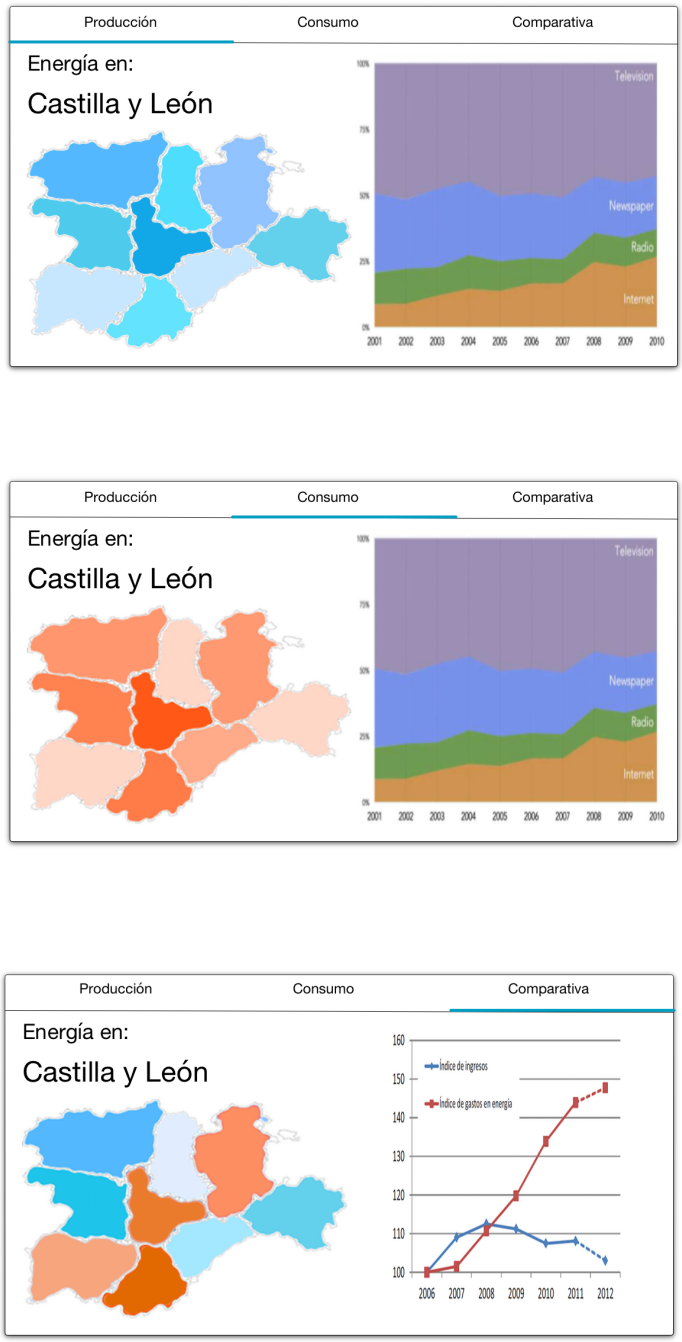

Figure 2: Alternativa 1 - Representación 1

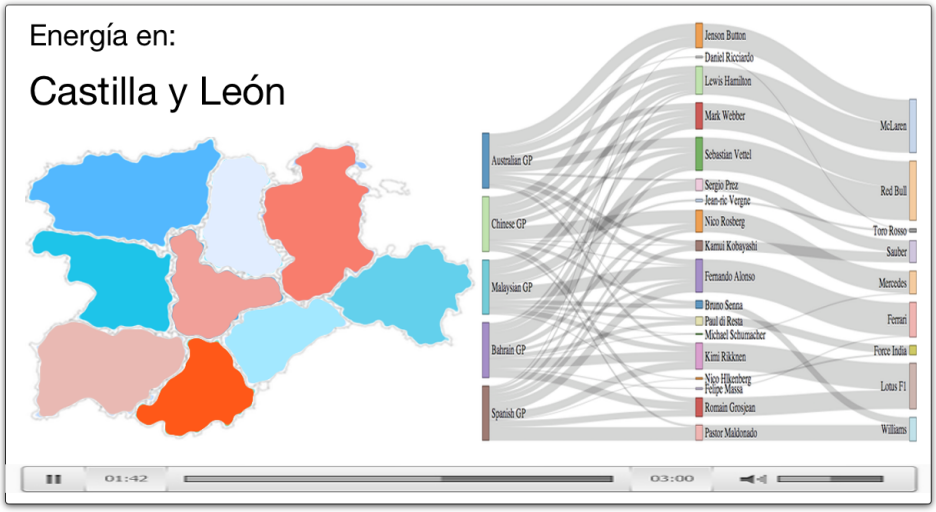

Figure 3: Alternativa 2 - Representación 1

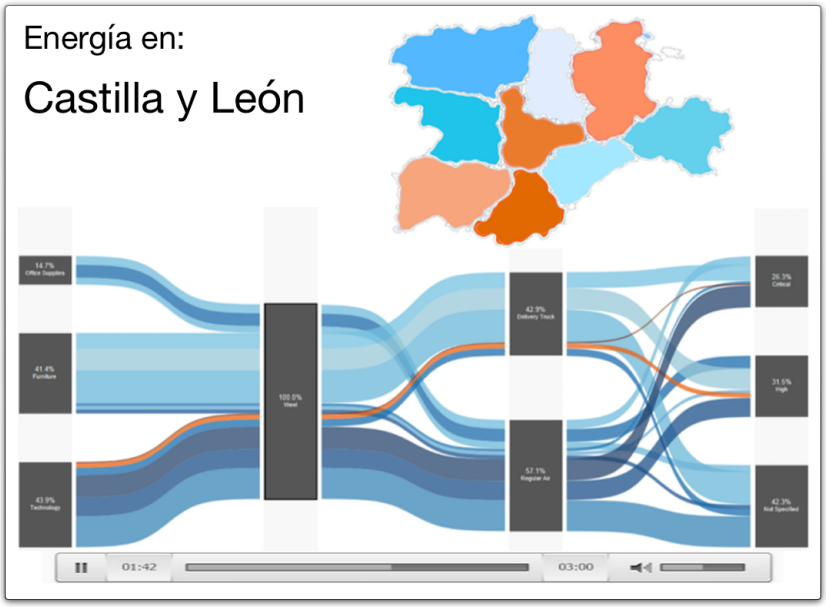

Figure 4: Alternativa 2 - Representación 2

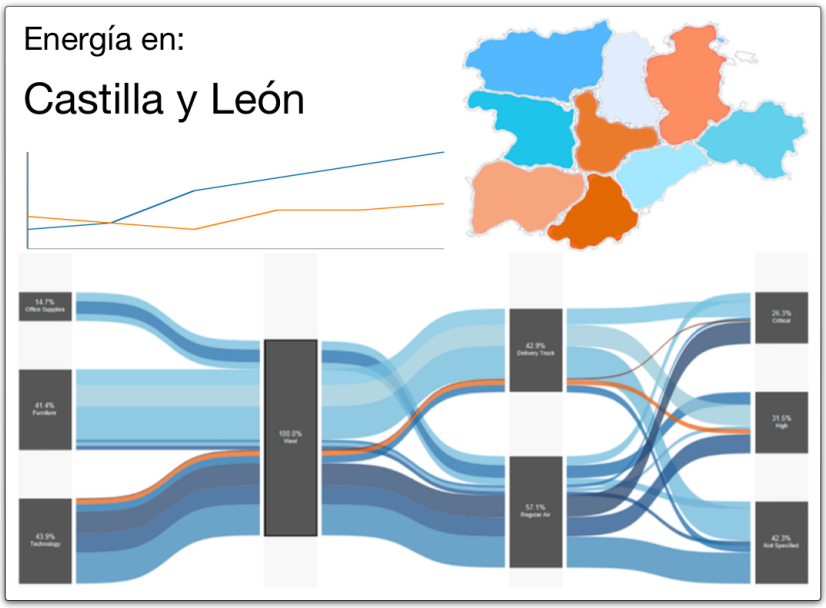

Figure 5: Alternativa 3 - Representación 1

6 Conclusiones

Tras analizar las alternativas propuestas se seleccionará la tercera alternativa, ya que bajo mi punto de vista mejora todos los aspectos de las anteriores.

Se ha escogido esta alternativa frente a la primera porque permite "de un solo vistazo" conocer el flujo de energía de la comunidad autónoma sacrificando la visión directa del tiempo.

La representación de grandes cantidades de datos no es una tarea sencilla, ya que depende de muchos factores como el tipo de dato que se quiere mostrar, variables temporales, geográficas, etc. Es una tarea compleja tratar de perder la mínima información posible sin sacrificar el diseño visual y la claridad de la información representada.

7 Bibliografía

- Diseño y Evaluación de Sistemas Interactivos (ETSII Valladolid): Apuntes de la asignatura
- Fundamentos de estadística / D. Peña Sánchez de Rivera.
- Wikipedia: CSV
- Converme: TEP
- Tableu: Sankey Diagram
- Internationa Energy Agency: Sankey Balance
- Mike Bostock: Let's make a map
- Vis4: Goodbye RedGreen Scales