Universidad de Oviedo

Memoria del Trabajo Fin de Máster realizado por

Gonzalo Moreno Ávila

para la obtención del título de

Máster en Ingeniería de Automatización e Informática Industrial

VISTUG (Visualización del transporte urbano de Gijón).

Febrero de 2014

Título del Proyecto:
VISTUG (Visualización del transporte urbano de Gijón)

Laboratorio de desarrollo del Proyecto:
ISA INGENIERIA DE SISTEMAS Y AUTOMÁTICA DE LA UNIVERSIDAD DE OVIEDO

Autor del proyecto:
Gonzalo Moreno Ávila

Tutor del proyecto:
Ignacio Díaz Blanco

Indice

1.Introducción…… 4
1.1 El marco ……5
1.2 Visualización de datos ……8
1.3 Objetivos……...12
2 Métodos y técnicas ……13
2.1 Tipos de datos y su codificación: ……13
2.2 El color:……17
2.3 Geovisualización: ……18
2.4 Interactividad: ……..19
2.5 Datos temporales: ……22
2.6 Tecnologías de visualización ……23
3 Resultado ..……...31
3.1 Descripción general del proyecto ……31
3.2 Gestión de datos .. ……32
3.2.1 Adquisición de datos y parseado ……33
3.2.2 Filtrado de datos ….……..34
3.2.3 Cálculo de datos ….……..35
3.3 Visualización de datos ……..40
3.3.1 Codificación visual ….……..41
3.3.2 Interactividad ……...47
3.4 Caso de uso ……..……...49
3.4.1 Análisis de los retrasos medios producidos en Gijón ……..50
3.4.2 Análisis de los retrasos medios en las marquesinas ……51
3.4.3 Análisis de los perfiles de los retrasos de las líneas en relación al “rastro de Gijón” ……...53
3.4.4 Conclusiones en el análisis de la influencia de “El rastro” en las marquesinas de Gijón. ...……55
4 Conclusiones y líneas futuras……57
4.2Aportaciones ..……...57
4.1Conclusiones ……58
4.3 Líneas de futuro trabajo ……...58
5Cronograma ……60
6 Presupuesto ……62
7 Bibliografía ……63

1 Introducción

En la actualidad, nos encontramos ante lo que algunos denominan la "era del Big Data". Vivimos inundados por los datos. Además, entre las administraciones públicas se está produciendo un cambio de mentalidad. Antiguamente, se pensaba que los datos que generaban y custodiaban las administraciones, debían ser salvaguardarlos de toda persona ajena a las propias administraciones, y en la presente década, por el contrario, las administraciones públicas empiezan a publicar los datos de forma abierta (Open Data) a disposición de todo el mundo. De hecho esta filosofía es uno de los pilares de lo que se llama hoy en día "ciudades inteligentes" (Smart Cities), una de cuyas principales premisas es la publicación y transparencia de los datos de las administraciones públicas.

La disponibilidad de dichos catálogos de datos abiertos permite, cómo dice Lathrop D (2010), que "*El gobierno debe dejar de ser un proveedor de servicios para convertirse en una plataforma sobre la que poder crear servicios".* Es decir, que cualquier persona pueda proveer servicios a las administraciones y a los propios ciudadanos.

Para hacerse una idea, en España, gracias al catálogo abierto de las distintas administraciones públicas, se han creado aplicaciones que permiten que los ciudadanos puedan, por ejemplo, conocer el estado de los parkings de su ciudad (Barcelona), denunciar el estado de su ciudad advirtiendo a los otros ciudadanos de los problemas que hay y avisando a la administración de los mismos (Zaragoza), saber el estado de los distintos parkings de bicicletas de su ciudad (Barcelona), conocer que farmacias están abiertas en este momento (Zaragoza), conocer la posición de los autobuses en cada momento (Gijón), etc. Por lo tanto, el aprovechamiento de esta información puede proporcionar a la sociedad, a las propias administraciones y a los comercios, grandes ventajas tanto en el aspecto social como en el económico.

Tener conocimiento de cómo funcionan las administraciones públicas en cada momento, puede ofrecer la posibilidad de detectar las carencias que éstas tienen, y con ello dar el primer paso para su propia mejora. Por esta razón, se planteó la idea, de que si se dispone de la información de la posición de los autobuses de Gijón y de sus marquesinas, se podría hacer un estudio estadístico de distintos parámetros del servicio de transporte de Gijón. Y poder poner el foco en las principales quejas de los usuarios del servicio de transporte urbano de la ciudad, como en este caso, los retrasos de los autobuses entre las marquesinas.

1.1 El marco

Como ya hemos mencionado anteriormente, nos encontramos en la que algunos no dudan en denominar "*la era del Big Data*".

Pero ¿qué es el "Big Data"? Es difícil encontrar una definición a "Big Data", dependiendo del autor, nos encontramos con distintas definiciones del mismo.

Antes de definir el "Big Data", veamos la evolución histórica del concepto, para posteriormente, poder entender lo que esto significa.

Figura 1.1. Evolución del concepto de Big Data

Una vez vista la evolución histórica del concepto de "Big Data" se puede decir que:

- Inicialmente el término "Big Data" se utilizaba para describir al conjunto de datos que superaban la capacidad del software habitual, para ser capturados, gestionados y procesados en un tiempo razonable.
- Hoy en día, se emplea el término "Big Data" para describir el conjunto de procesos, tecnologías y modelos de negocio que están basados en datos y en capturar el valor

que los propios datos encierran. Esto se puede lograr a través de una mejora en la eficiencia gracias al análisis de los datos (una visión más tradicional). Hay que tener presente que es crítico encontrar la forma de dar valor a los datos para crear nuevos modelos de negocio o de ayuda a los existentes.

Los números del "Big Data":

A continuación se muestran algunas de las cifras que se barajan en torno al fenómeno "Big Data", tanto en cuanto al volumen de datos que se generan en el mundo (Figura 1.2), como al valor que potencialmente podrían tener estos datos según el informe elaborado por Mckinsey (2011), (Figura 1.3).

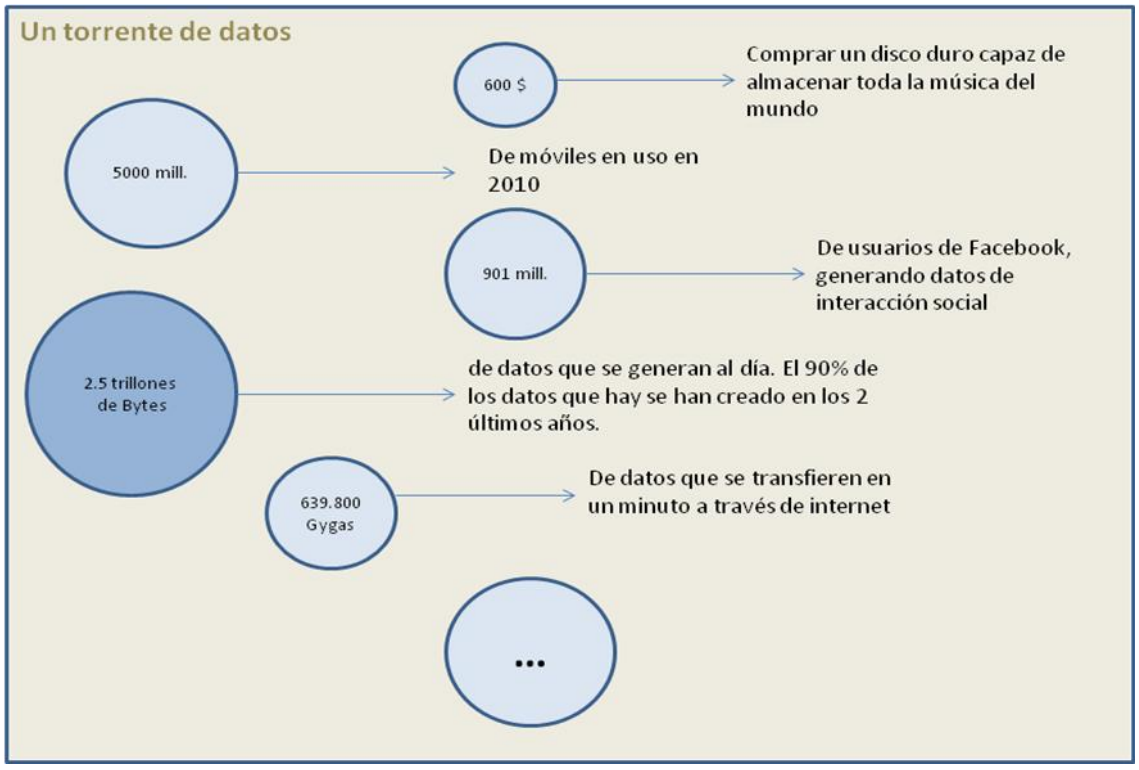

Figura 1.2. Los números del "Big Data"

Según el informe Mckinsey:

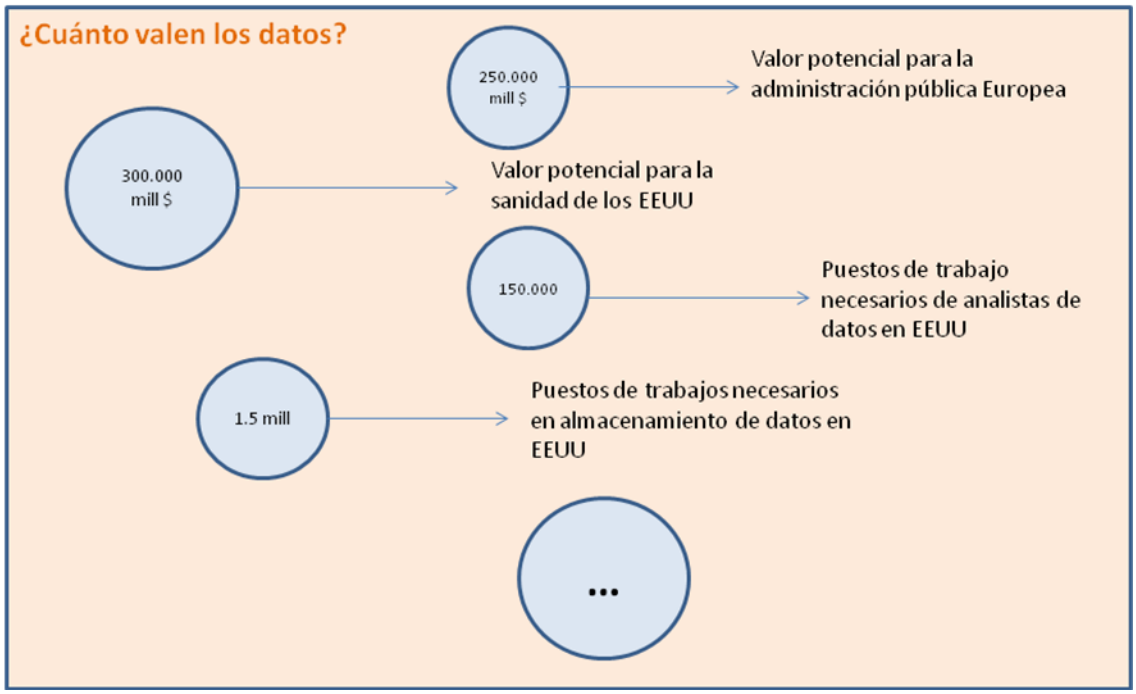

Figura 1.3. Cuantificación del valor de los datos

Dificultades del "Big Data":

Aún así, no todo es tan sencillo, si se producen tantos datos al minuto y esos datos tienen "de alguna forma" un valor económico, parece fácil preguntar ¿por qué no extraer ese valor de los datos?

Y la respuesta no es fácil, no es inmediato obtener conocimiento de los datos. Ya en los años 70 algunos científicos, pronosticaban la "maldición de los datos". Decían que al aumentar la capacidad de los sistemas informáticos para almacenar y cuantificar los distintos valores de infinitud de variables de los distintos procesos, pese a que en si mismo supondría un gran impulso para el avance de la ciencia, a su vez aparecería el problema de gestionar la cantidad tan ingente de datos que se obtendrían.

Si se quiere obtener conocimiento de los datos, se debe seguir una serie de pasos, según la metodología KDD (Knowledge Discovery in Databases) Fayyad et al (1996):

1. Recogida de datos: Debemos seleccionar las fuentes de información que nos pueden ser útiles y donde conseguirlas. Dificultad que surge: **Captura.**
2. Diseñar el esquema de un almacén de datos (Data Warehouse): se debe intentar unificar el almacenamiento de los datos. Se separan los datos de las fuentes que los

producen. Hacer bien esta fase, va a facilitar el análisis de los datos en tiempo real. Dificultad que surge: **Almacenado.**

3. Selección, Limpieza y Transformación de datos: Debemos eliminar los datos que sean inconsistentes (outliers) o que sean irrelevantes. Para llevarlo a cabo se suelen utilizar métodos estadísticos casi exclusivamente. Dificultad que surge: **Búsqueda.**
4. Análisis y Minería de datos: Se intentan descubrir patrones a partir de los datos, lo que se pretende es obtener información y conocimiento a partir de dichos datos, para poder extraer conclusiones. Se utilizan métodos de inteligencia artificial, aprendizaje automático y estadística. En la minería de datos se incluye la visualización de los datos, ya que permite transmitir y analizar mayor cantidad de datos en menos tiempo. Dificultades que surgen: **Análisis** y **Visualización.**

En la Figura 1.4, se ven las distintas dificultades que van apareciendo en las distintas fases a la hora de obtener conocimiento de los datos.

1.2 Visualización de datos:

La visualización de datos es una herramienta indispensable en la minería de datos, para poder encontrar patrones y poder obtener conocimiento a través de los datos. Pero también lo es para transmitir los conocimientos a la sociedad. La visualización de datos tiene dos funciones; la de explotar el potencial de los datos y la de transmitir los conocimientos obtenidos a las demás personas. La visualización de datos es una herramienta de ayuda en el análisis de los datos, no un sustituto del análisis de los datos.

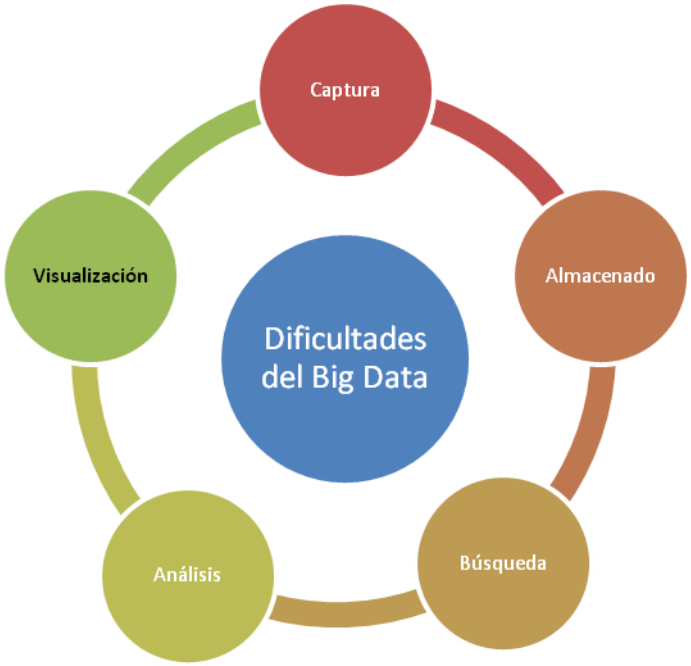

Figura 1.4. Dificultades que aparecen durante la extracción de conocimiento de los datos a partir de la metodología KDD.

La Visualización se puede definir como una representación gráfica de la información, con el objetivo de ofrecer al observador una compresión cualitativa de los contenidos de la información.

"La mejor visualización como aquella que muestra algo nuevo, en lo que a patrones y relaciones contenidas en los datos se refiere" Steele, J (2012).

Muchas veces las visualizaciones de datos pueden ser más precisas y reveladoras que las técnicas de estadística convencionales. Como se observa en el "cuarteto de Anscombe" Anscombe J (1973), que aparece para ilustrar el poder de las visualizaciones en el libro "*The visual display of quantitative information"* de Tufte, E.R(1983). Se disponen de cuatro conjuntos de datos (Figura 1.5).

I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Figura 1.5. Pares X, Y de cuatro conjuntos de datos distintos

Estos 4 conjuntos de datos tienen los mismos atributos estadísticos (Figura 1.6).

$N = 11$
mean of X's $= 9.0$
mean of Y's $= 7.5$
equation of regression line: $Y = 3 + 0.5X$
standard error of estimate of slope $= 0.118$

$t = 4.24$
sum of squares $X - \overline{X} = 110.0$
regression sum of squares $= 27.50$
residual sum of squares of $Y = 13.75$
correlation coefficient $= .82$
$r^2 = .67$

Figura 1.6. Atributos estadísticos de los cuatro conjuntos de datos de la Figura 1.5

Y, si se representan los cuatro conjuntos de datos como en la (Figura 1.7), se observa que no tienen nada que ver entre sí, aunque desde el punto de vista estadístico sean idénticos.

En el libro blanco de Harvard (http://www.sas.com/resources/whitepaper/wp_62194.pdf) se cita el valor del uso de las visualizaciones, gracias a que nuestro cerebro está acostumbrado a la percepción visual. Se muestra como ejemplo los mapas. Si alguien ve un mapa de EEUU (Figura 1.8) inmediatamente, sin que esté citado en un texto, puede saber que el estado de New York tiene en su frontera sur al estado de Pennsylvania, pero si en vez de un mapa se intenta describir esta información de otra forma, no hay ningún tipo de representación de datos que sea capaz de transmitir esa información tan rápidamente como los mapas. Esta es para Amanda Cox (editora gráfica del New York Times) la mayor virtud de las visualizaciones de datos, **la inmediatez**. Además, si alguien que jamás haya visto un mapa, ve por primera vez un mapa de EEUU, lo más seguro es que podrá saber la relación geográfica entre los distintos estados sin necesidad de que se le cómo funciona un mapa. Esta es la potencia de las visualizaciones, la capacidad de mostrar las relaciones de datos de una manera inmediata.

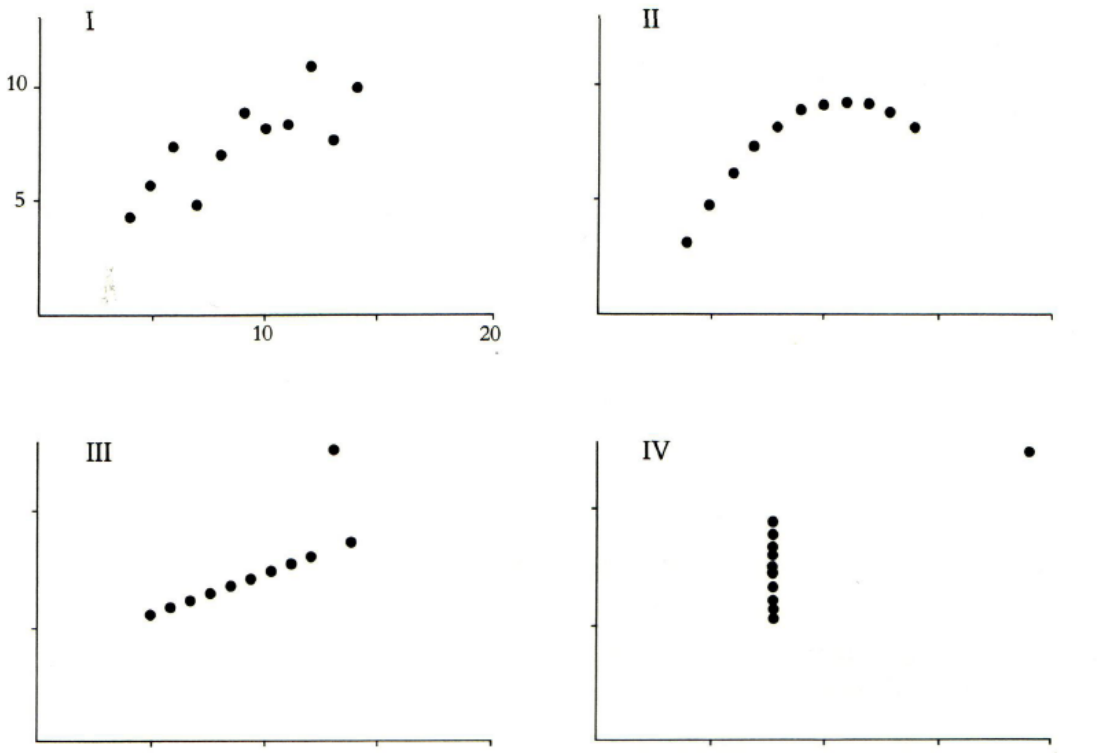

Figura 1.7. Representación gráfica de los cuatro conjuntos de datos

Figura 1.8. Mapa de los estados de EEUU, con detalle en rojo del estado de New York y Pennsylvania

1.3 Objetivos

En el presente proyecto se pretende que, gracias a los datos que aporta el ayuntamiento de Gijón a través del portal *datos.gijon.es,* se desarrolle una aplicación que permita supervisar el sistema de transporte urbano de Gijón, explorando las distintas técnicas de visualización y análisis de datos inteligentes para, con la combinación de algoritmos de análisis de datos, poder mostrar una forma visual, estadísticas e índices de calidad del servicio que nos permita identificar las deficiencias que pudiera tener el servicio de transporte urbano de Gijón y si fuera posible, gracias al conocimiento de las eventuales incidencias que pudiera tener el mismo, dar soluciones a los distintos problemas o en la toma de decisiones.

Se pretende a su vez que la disponibilidad de unos pocos índices específicos que resuman el comportamiento del sistema, permitan el análisis del mismo a largo plazo, facilitando el estudio de parámetros socioeconómicos mediante el análisis de la influencia de éstos en la calidad del servicio.

Lo que interesa en este proyecto es la posibilidad de disponer de forma visual, de estadísticas e índices de calidad del servicio. El principal interés de este trabajo es sobre todo el retraso que los ciudadanos soportan en las marquesinas de Gijón. Poder detectar e identificar a tiempo las deficiencias en el servicio, identificar en qué líneas se producen las deficiencias, en qué marquesinas y en qué días y horas, es decir tener una representación de información espacio-temporal del comportamiento de la flota de autobuses, para así disponiendo de información de los problemas al detalle, poder dar soluciones específicas.

Por último, gracias a la disposición de una flota de autobuses que recorre casi toda la ciudad, se podría tener una perspectiva del estado del tráfico de Gijón. Haciendo un análisis del estado del tráfico en Gijón en los distintos días de la semana a las distintas horas del día, se podría ayudar a las autoridades en la toma de decisiones y mejoras futuras.

2 Métodos y técnicas

2.1 Tipos de datos y su codificación:

La codificación visual es la vía mediante la cual los datos son transformados en estructuras visuales, a través de las cuales se crean las visualizaciones.

Cuando se realiza la codificación visual, lo primero que hay que analizar es qué tipo de datos tenemos.

Tipos de datos:

Categóricos: Son datos de categorías distintas, representan grupos de datos diferentes y no presentan un orden implícito. Se pueden utilizar atributos que permitan ordenarlos, como por ejemplo alfabéticamente. Ejemplos: Peras, manzanas, fresas. (Figura 2.1 a)

Ordenados: Estos datos tienen en sí mismos un orden implícito, dentro de este tipo de datos existen dos clases:

- *Ordinales*: Son datos que están claramente ordenados, pero sobre los cuales no es posible realizar una medida aritmética. Ejemplos: Tallas de ropa: Pequeño, mediano y grande. Sabemos que mediano está entre pequeño y grande. (Figura 2.1 b)

- *Cuantitativos*: Son datos, como la edad, que representan una medida de la magnitud que soporta una comparación aritmética. A su vez, en este tipo de datos se puede diferenciar entre:
    - Secuenciales: Existe un rango homogéneo entre un valor mínimo y uno máximo. Ejemplos: Habitantes: 140, 200, 3000. (Figura 2.1 c)
    - Divergentes: Son datos que pueden ser divididos en dos secuencias que van en direcciones opuestas y confluyen en un punto común (el 0). Ejemplos: Ingresos: -10€, 20€, 100€ (Figura 2.1 d)

Codificación:

Una vez que se sabe qué tipos de datos se van a visualizar, se debe tener en cuenta cuántos elementos distintos se quieren identificar, diferenciar y recordar. Se va a tener que definir las marcas y canales que se utilizarán.

Para ello, comenzaremos explicando qué son marcas y canales,

Marcas: Son elementos gráficos básicos en una imagen, elementos geométricos primitivos clasificados de acuerdo a las dimensiones espaciales que se necesitan. Ejemplos: puntos, líneas, áreas.

Conectar marcas entre sí, o contenerlas en un conjunto puede mostrar relaciones jerárquicas entre ellas.

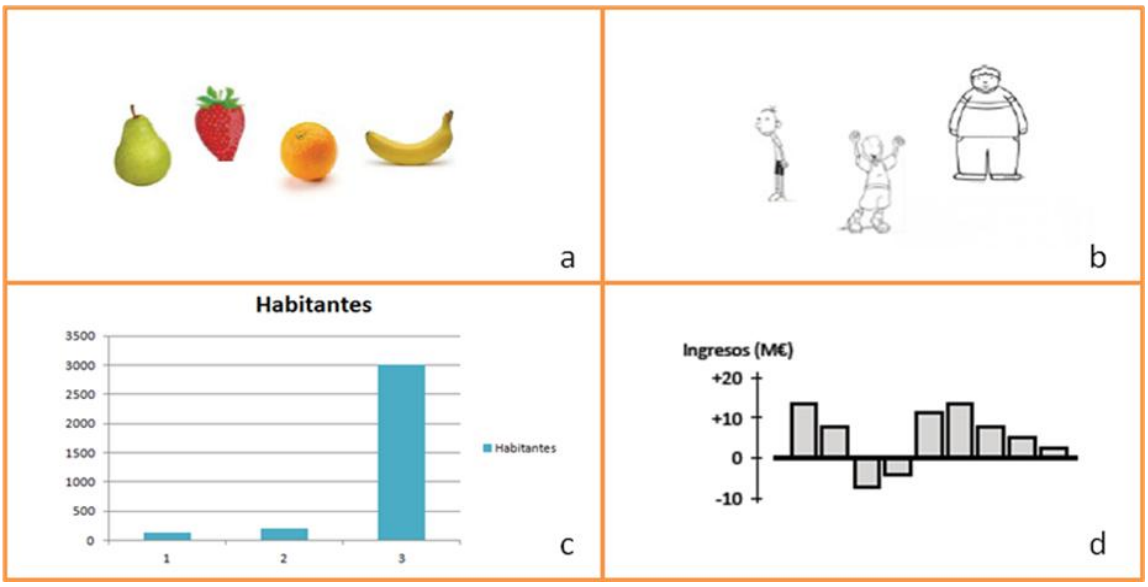

Figura 2.1: a. Datos categóricos (frutas), b. Datos Ordenados Ordinales (tamaño de ropa), c. Datos Ordenados cuantitativos secuenciales (Habitantes), d. Datos Ordenados cuantitativos divergentes (Ingresos). Fuente: Ignacio Díaz Blanco "VD 4"

Canales: Son todos los parámetros que controlan la apariencia de las marcas, son independientes de las dimensiones de las geometrías primitivas. Ejemplos: Tamaño, color, posición.

A continuación en la figura 2.2, se pueden ver distintos ejemplos de marcas y canales de codificación visual.

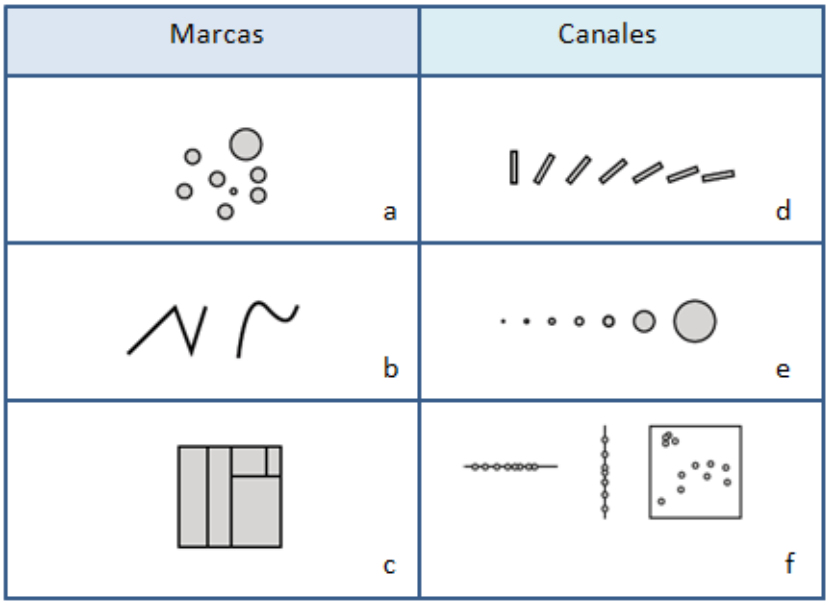

Figura 2.2: Marcas: a. Puntos, b. Líneas, c. Áreas. Canales: d. Orientación, e. Tamaño, f. Posición. Fuente: Tamara Muzner "Information Visualization: Principles, Methods and Practice"

Existen dos tipos de canales:

- "**Qué":** dan información sobre la identidad y la localización**.** Este tipo de canales son los adecuados para los atributos categóricos.
- **"Cuánta cantidad":** aportan información sobre la cantidad. Este tipo de canal es el más adecuado para los atributos ordenados, tanto ordinales como cuantitativos.

Los canales se rigen por dos principios:

- **Expresividad:** La codificación visual debe expresar toda la información contenida en los atributos (y sólo ésta).
- **Efectividad:** La efectividad del canal debe corresponderse con la importancia del atributo.

Cuando se va a seleccionar un canal, se debe tener presentes varios conceptos:

- **Precisión:** Los canales deben ser lo más precisos posible para realizar una visualización de datos eficiente. Es como se percibe la variación del canal con respecto a variación de intensidad física de la variable representada.
- **Discriminabilidad:** Es el número de intensidades que se pueden percibir como "diferentes" de manera eficiente en un canal. El hecho de que sean pocos no tiene por qué ser un problema, si los distintos atributos que tenemos que representar son escasos (Figura 2.3).
- **Separabilidad:** Es la capacidad de los canales de mostrar sus atributos de manera independiente del resto de canales, esto es especialmente importante cuando se visualizan varios atributos distintos, ya que todos los canales tienen cierto grado de dependencia.
- **Redundancia:** Se trata de utilizar varios canales para codificar el mismo atributo, esto provoca que el atributo que se codifica se perciba con mayor sencillez y, como contraprestación, disponemos de un canal menos para mostrar nuevos atributos.

En la Figura 2.4, se muestra un listado de canales, con sus ventajas e inconvenientes.

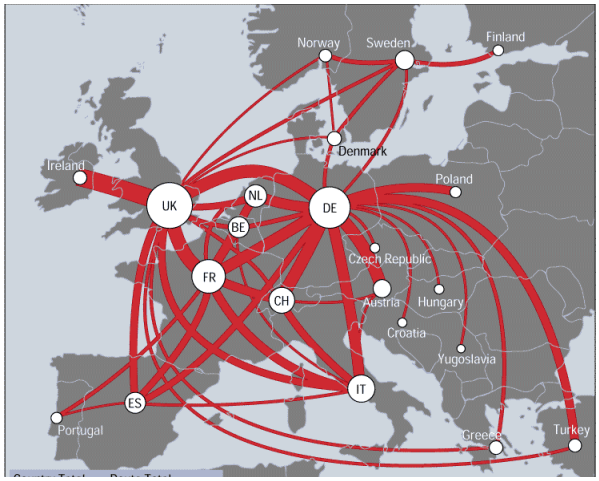

Figura 2.3. Las líneas tienen un grado limitado de grosores que se pueden identificar como diferentes. Fuente: Tamara Muzner "Information Visualization: Principles, Methods and Practice"

Ejemplo	Codificación	Ordenado	Cantidad de valores	Cuantitativo	Ordinal	Categórico	Relacional
	Posición	Sí	Infinitos	Bueno	Bueno	Bueno	Bueno
1,2,3, A,B,C	Campos de texto	Opcional, alfanuméricos	Infinitos	Bueno	Bueno	Bueno	Bueno
	Longitud	Sí	Muchos	Bueno	Bueno		
	Tamaño, Área	Sí	Muchos	Bueno	Bueno		
	Ángulo	Sí	Medio	Bueno	Bueno		
	Densidad	Sí	Pocos	Bueno	Bueno		
	Peso, Grosor	Sí	Pocos		Bueno		
	Brillo, saturación	Sí	Pocos		Bueno		
	Color	No	Pocos (<20)			Bueno	
	Forma, icono	no	Medio			Bueno	
	Textura	No	Medio			Bueno	
	Encapsulación, conexiones	No	Infinitos			Bueno	Bueno
	lineas	No	Pocos				Bueno
	Finales de linea	No	Pocos				Bueno
	Grosor de linea	Sí	Pocos				

Figura 2.4. Ranking de canales de codificación de atributos. Fuente: Noah Iliinsky; Julie Steele"Designing Data Visualizations"

2.2 El color:

El color es un canal rico y complejo, que debe ser utilizado para codificar alguno de los atributos más importantes de la visualización, siguiendo el principio de la efectividad.

El color puede servir para representar tanto atributos categóricos y subconjuntos, como atributos cuantitativos y ordinales.

- Atributos categóricos y subconjuntos de datos: Para este tipo de atributos, después del espacial (del que se habla más adelante), es el mejor canal. Se pueden discriminar hasta 12 colores distintos. Es un canal de tipo "**Qué".** En la Figura 2.5 se ve una sugerencia de colores categóricos.

Figura 2.5. Secuencia de colores adecuados para codificación de atributos categóricos

- Atributos cuantitativos y ordinales: se debe que tener cuidado con este canal, ya que en función de la escala de color puede no cumplir el principio de precisión, como la escala Rainbow. Pero existen otras escalas que sí cumplen este principio adecuadamente, en la página hEp://colorbrewer2.org/ hay disponible de manera gratuita, una herramienta que genera en función de las necesidades del usuario, la escala de colores más adecuada. En la Figura 2.6 hay 2 escalas de colores adecuadas para codificar atributos divergentes (a) y secuenciales (b)

  .

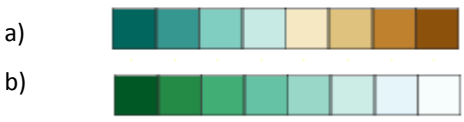

Figura 2.6 Codificaciones de colores

2.3 Geovisualización:

Como se ha descrito anteriormente, es de vital importancia decidir qué tipo de canal utilizar para representar cada tipo de información, el canal más poderoso a la hora de codificar la información es el espacio. Munzner,T (*pendiente de publicación*). La Geovisualización consiste en utilizar el canal espacial situando los datos sobre mapas, es decir, en codificar la latitud y la longitud en los ejes X e Y.

La geovisualización principalmente permite:

- Búsqueda de autocorrelación espacial.
- Conocimiento previo del usuario, pudiendo extraer conclusiones gracias al conocimiento previo de las áreas en las que se muestran los datos.

Búsqueda de autocorrelación espacial:

La primera ley de la geografía es*: "cualquier cosa está relacionada con cualquier cosa, pero las cosas cercanas están más relacionadas entre sí que las lejanas"*, Tobler (1970).

Una de las principales virtudes de las geovisualizaciones es la detección de fenómenos de autocorrelación espacial. Este término se refiere a la existencia de un patrón en la distribución espacial de una variable.

"La geovisualización permite facilitar el pensamiento, entendimiento y la construcción de conocimiento sobre las personas y el entorno físico que las rodea, aumentando la habilidad visual de los usuarios para comprender estructuras de complejidad elevada y la habilidad de detectar, explorar (con el uso de interacción) y explotar patrones significantes" Kim (2009). *"Aporta una poderosa contribución en la toma de decisiones"* Wright (2010).

Además a consecuencia de este fenómeno: *"La representación de datos en mapas, es una herramienta extremadamente útil para la identificación de valores atípicos"* (outliers) Anselín (1999). Esto es debido a que si un valor es muy distinto a sus vecinos , por la primera ley de la geografía, es posible que ese valor sea un outlier.

Conocimiento previo del usuario:

La representación geoespacial de datos, permite establecer un contexto geográfico adecuado (mapas políticos, orográficos, etc.), facilitando la interpretación y la extracción de conocimiento de los datos.

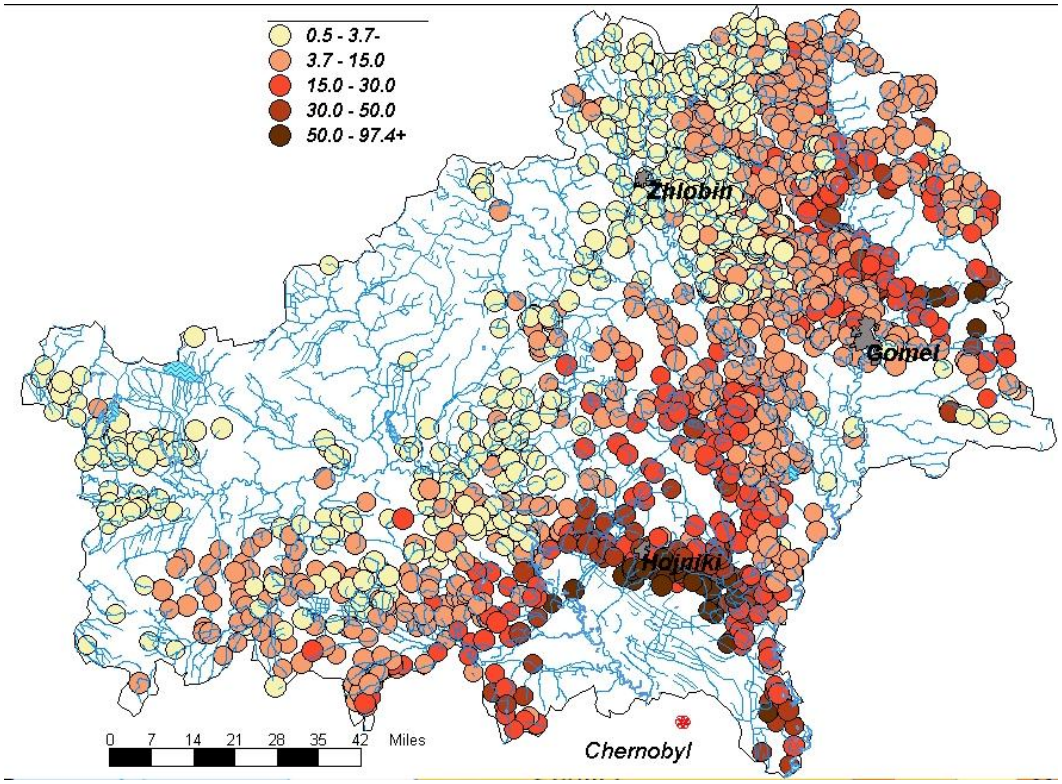

Figura 2.7. Efectos de la radiación en el área de Chernóbil en 1994, cantidad de Sr (kBq/m^2) *, tras el accidente de la central nuclear.*

"Cuando se realiza una representación apropiada del espacio, se permite explotar el conocimiento tácito de los datos asociados al espacio, permitiendo reconocer lugares conocidos y percibiendo relaciones y patrones espaciales asociados con conocimientos previos. La representación de datos en un buen mapa, no sólo transmite los datos en su contexto geográficos, sino que también como instrumento de generalización y síntesis, lo que permite revelar patrones invisibles que no eran evidentes en una representación no espacial de los datos." Andrienko et al (2008).

2.4 Interactividad:

¿Qué es interactividad?

Según la RAE: *"Dicho de un programa: que permite una interacción, a modo de diálogo, entre el ordenador y el usuario".*

Parece más acertada la definición de interactividad como: *"La condición de comunicación en la cual ocurren cambios constantemente, los cuales generan un aumento de la fuerza social".* Bronw and Yule (1983).

Y si se enlaza esta definición con la que dan Rafaeli, S. and Sudweeks, F. (1997), "*en una serie de intercambios de comunicaciones implica que el último mensaje se relaciona con mensajes anteriores, a su vez relativos a otros previos"*

En el caso de las visualizaciones, se puede entender interactividad como la herramienta que permite un diálogo entre el ordenador y el usuario, esta herramienta permitirá realizar cambios de manera constante en la aplicación, desarrollando de manera continuada visualizaciones nuevas. Que facilitará el aumento del conocimiento extraído de los datos a cada nueva visualización. Este aumento del conocimiento se verá retroalimentado de forma continuada, ya que la variación de las visualizaciones se hará de forma relacional con las visualizaciones anteriores, permitiendo que se incremente de forma constante la adquisición de conocimiento.

"*La interactividad es crucial para construir herramientas de visualización que manejen grandes cantidades de datos, y que estos cambien con el tiempo*.

La interacción en las visualizaciones de datos permite superar las limitaciones de los usuarios y de las pantallas. Ayudando a la investigación y el análisis de los datos a varios niveles de detalle.

El beneficio de la interactividad es que el usuario puede apreciar mucha más información que en una imagen estática, y su gran desventaja reside en que tiene un costo en tiempo mayor que una visualización en una sola imagen. En el caso de que fuese suficiente una visualización automática, sin necesidad de interactividad, habría que optar por esta última opción." Munzner, T (pendiente de publicación).

"La utilización de interactividad en las visualizaciones aumenta las opciones de análisis de los datos", Franks,B (2013).

En el Libro Blanco de Hardware citado anteriormente, se explica que, cuando el contexto de tus datos es el tiempo, la visualización debe tener movimiento e interacción, o al menos una de los dos. Según se cita, si se dota a la visualización de interactividad, se permite que el usuario participe en la historia que transmiten los datos, se permite que cada cual pueda contar su propia historia con los datos, pudiendo variar las perspectivas de la visualización. Aunque también se indica que el gran costo de la interactividad está en el diseño de la visualización, cuya complejidad aumenta a la par que aumenta la comprensión de datos complejos.

En el artículo "the value of visualization" Wijk (2005), se explica el funcionamiento de un modelo simple de visualización, en él se describe la importancia que tiene la interactividad en la adquisición de conocimiento.

En la Figura 2.9, se observa el modelo básico de visualización de Wikj, donde:

- Data (D): Son los datos que se analizan
- Specifications (S): Especificaciones que quiere ver el usuario
- Visualization (V): Es el proceso de la visualización, que, a partir de los datos y de las especificaciones del usuario, genera la Imagen "I".
- Imagen (I): Es la imagen que genera el proceso de visualización
- Perception (P): Es el proceso de percepción del usuario a partir de la Imagen
- Knowledge (K): Una vez que la imagen es percibida por el usuario, incrementa su conocimiento.

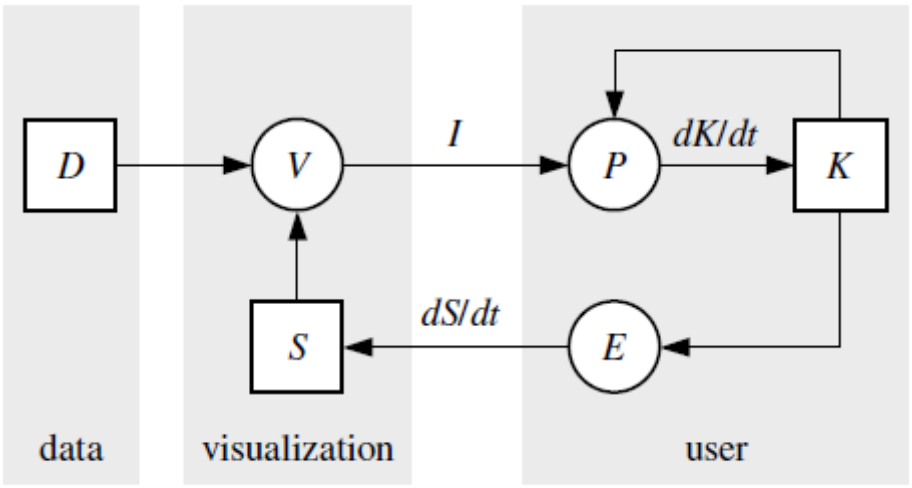

Figura 2.9. Modelo simple de visualización

En el modelo de visualización de la figura 2.9, los datos son continuamente transformados gracias a la especificaciones (S) del usuario, que se manifiestan gracias a la interactividad (E), definiendo el aumento del conocimiento (K) como dK/dt=P(I,K). Esto quiere decir que es función de las capacidades cognitivas del usuario y de la imagen proyectada (I), esta imagen es la que proyecta la herramienta de visualización a través de las especificaciones del usuario y los datos que se visualizan. La interactividad tiene una importancia marcada en el aprendizaje porque permite disponer en el momento que el usuario lo desea de una imagen nueva con tan sólo variar las especificaciones de los datos. Provocando un aumento de conocimiento que se retroalimenta, ya que cuanto más conocimiento se adquiere de lo visualizado, dicha interactividad proporciona nuevas herramientas para poder cambiar las especificaciones en la

dirección de interés, adaptando la visualización y permitiendo la obtención nuevos conocimientos.

a)	$K(t) = K_0 + \int_0^t P(I,K,t)dt$
b)	$S(t) = S_0 + \int_0^t E(K)$

Ecuación 2.1 . a. Ecuación de la extracción de conocimiento, b. Ecuación de la parametrización de los datos en una visualización

"La interacción en las visualizaciones de datos, permite a los usuarios realizar selecciones, deselecciones, rotaciones, y otras transformaciones de los datos que permiten ayudar en la búsqueda de estructuras y en el descubrimiento de patrones". (Buja et al 1996)

"La motivación principal para introducir interactividad en una aplicación de visualización de datos es introducir el factor humano de una manera más directa en la exploración de datos (explotando las capacidades inherentes al cerebro humano para la detección de patrones y estructuras), permitiendo que sea el usuario y no unos procedimientos estadísticos preestablecidos, el que determine qué acciones tomar. Es especialmente poderosa cuando hay una cantidad grande de datos y/o una gran cantidad de variables." Cook et al (1996).

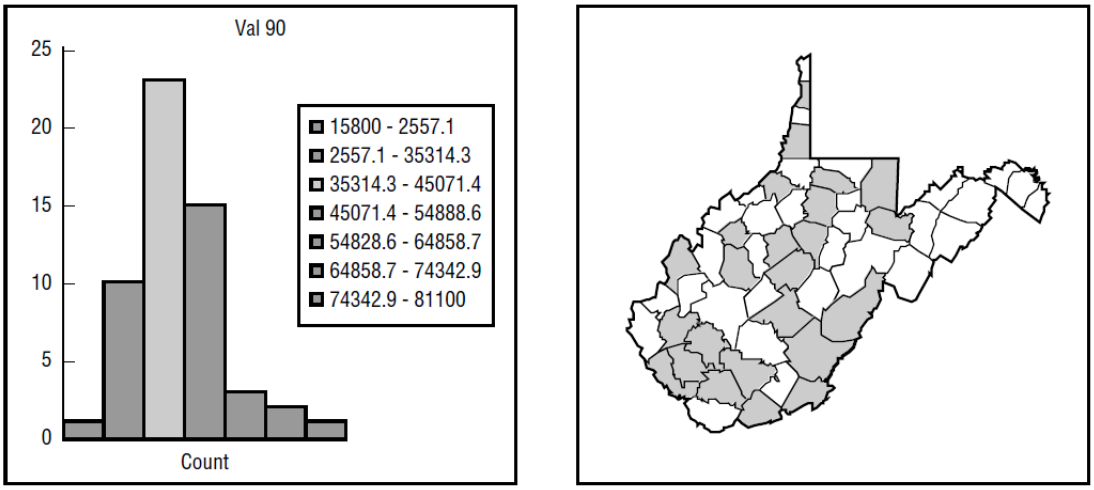

Figura 2.10. Histograma y mapa con el valor medio de los precios de las casas en West Virginia en 1997, desarrollado en ArcView-SpaceStat

2.5 Datos temporales:

Lo primero será definir qué son los datos temporales: "*datos temporales: son los datos que están relacionados de alguna forma con el tiempo"* Aigner et al (2008).

Para Aigner, escoger y parametrizar adecuadamente con respecto al tiempo es fundamental para hacer una representación visual adecuada que no dé lugar a falsas interpretaciones.

Este mismo autor, explica que la utilización de técnicas de interactividad es básica para realizar visualizaciones eficaces en series de datos temporales. En el mismo artículo se indica que, la interactividad, puede ser especialmente buena en datos orientados al tiempo, para: *"inicialmente crear una imagen global de los datos y permitir que, a través de dicha interactividad con el usuario, se exploten los detalles de los datos".*

Como puede parecer obvio, es especialmente importante el análisis correcto de la dimensión temporal de los datos, cuando lo que se busca son patrones temporales.

"Un patrón temporal es la aparición de unos parámetros (valores) o de la combinación de varios parámetros asociados a anotaciones temporales" Miksch et al (1998).

En la Figura 2.11, vemos un ejemplo de una visualización de datos temporales.

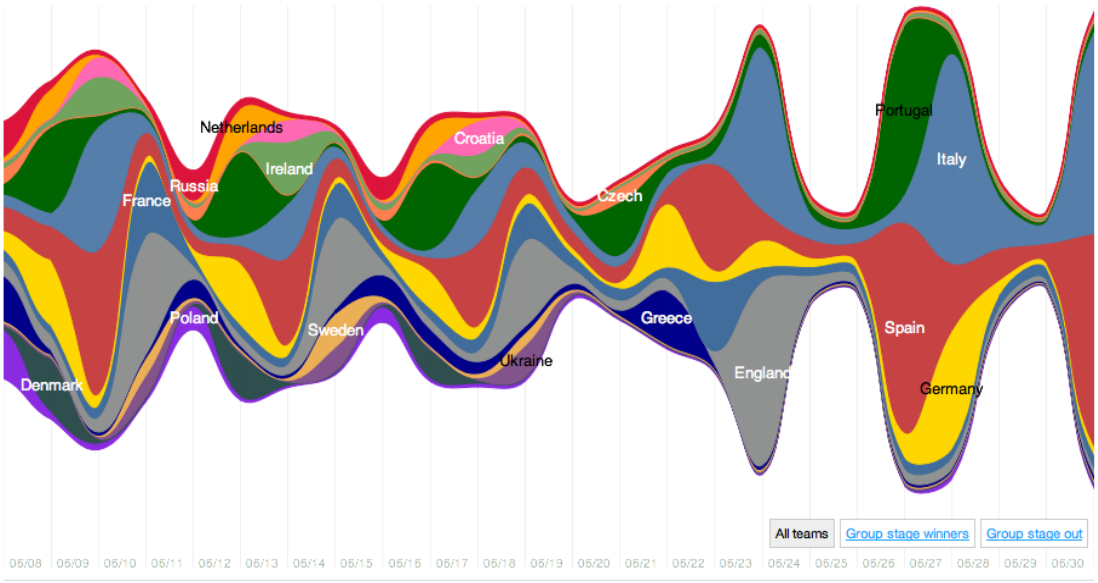

Figura 2.11. Evolución de los twiits durante la Eurocopa de 2012

2.6 Tecnologías de visualización

A continuación se desgranan algunas de las herramientas de creación de visualizaciones de datos que existen en la actualidad, destacando sus ventajas y desventajas. Evidentemente, sólo se van a describir las consideradas más significativas, ya que la descripción en detalle de las herramientas en sí, podría suponer la elaboración de otro proyecto distinto.

Excel:

Es la herramienta más popular para la generación de gráficos y visualizaciones, debido a la gran distribución del sistema operativo *"Windows"* y la de su paquete estrella *"Microsoft Office"*. Actualmente, se ofrece un gran catálogo de visualizaciones, como "mapas de calor" (heat maps), gráficos de barras, etc.

Ventajas:

- Sencillez y velocidad de creación

Desventajas:

- Es necesaria una licencia para poder utilizarla
- Capacidad en general limitada (colores, líneas, gráficos, etc.)
- No permite interactividad

Conclusiones:

No parece adecuado utilizar este software para la creación profesional de visualizaciones, además, se debe tener en cuenta la necesidad de disponer de una licencia de Microsoft Office para poder utilizarla.

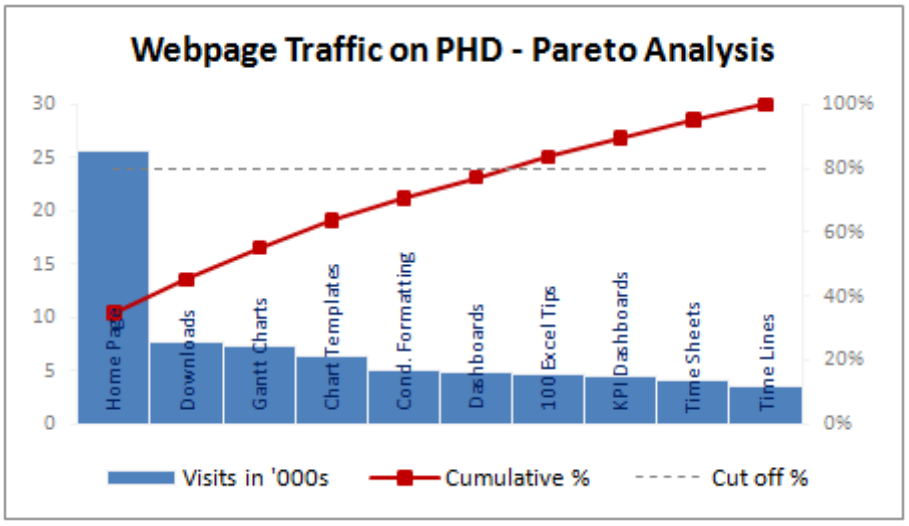

Figura 2.12. Ejemplo de visualización realizada con Excel sobre el tráfico en páginas web

Tableau Public:

Es una herramienta gratuita de visualización de datos mediante gráficos, desarrollada por Tableau, que dispone de otra herramienta de visualización de datos más potente, pero que a diferencia de esta, es de pago. Es un software desarrollado explícitamente para la realización de visualizaciones de datos.

Ventajas:

- Facilidad de uso
- Gran variedad de visualizaciones

Desventajas:

- Necesita Windows para funcionar
- No tiene las capacidades de visualización e interactividad que tienen las herramientas para desarrolladores

Conclusiones:

Es una herramienta de licencia libre, que no requiere conocimientos en programación para ser utilizada. Es muy superior a Excel para crear visualizaciones, pero es limitada en comparación con otras herramientas para desarrolladores.

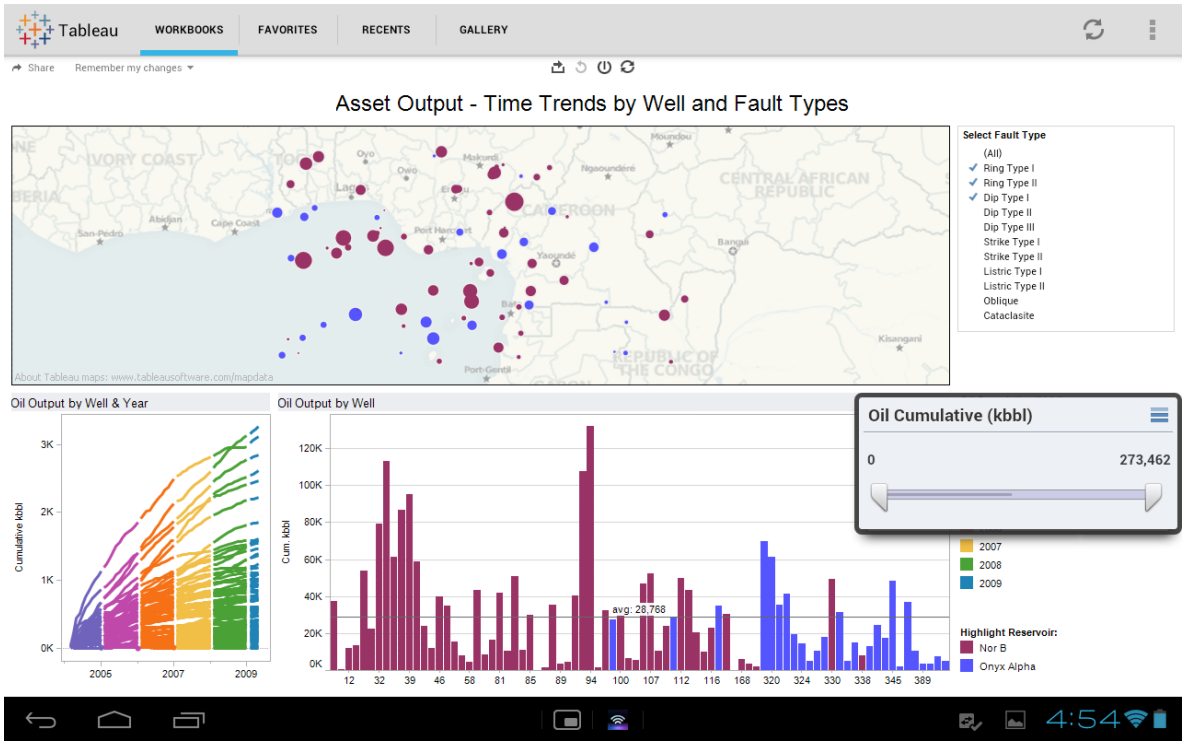

Figura 2.13. Ejemplo de visualización realizada con Tableu Public

The R Project for Statistical Computing:

R es un lenguaje y entorno de programación para gráficos y computación estadísticos, de distribución libre y de código abierto. Es un lenguaje basado en comandos, que permite crear gráficos a medida, no sólo gráficos estandarizados, y también nuevos gráficos adecuados al conjunto de datos y problema que sea necesario abordar.

Ventajas:

- Fuerte comunidad de desarrolladores
- Gran diversidad y cantidad de librerías desarrolladas
- Es gratuito

Desventajas:

- Curva de aprendizaje elevada
- No permite su exportación a los navegadores Web

Conclusión:

Es una herramienta que si es conocida por el desarrollador, ya que tiene una amplia implantación sobre todo en estadística, puede aportar soluciones en el desarrollo de visualizaciones. Pero si se desconoce este lenguaje, puede resultar complicada de utilizar, ya que su curva de aprendizaje es bastante elevada y es necesario un alto nivel de conocimiento para poder hacer trabajos de cierta complejidad.

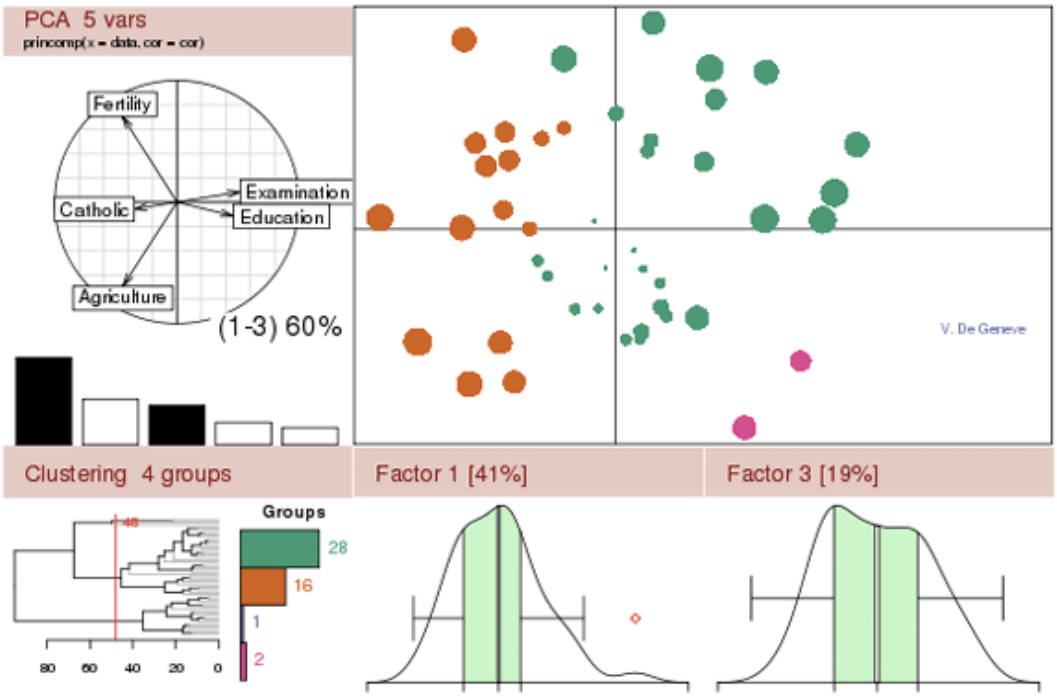

Figura 2.14. Ejemplo de visualización realizada con R

Gephi:

Plataforma de exploración y visualización interactiva para todo tipo de redes y grafos complejos, dinámicos y jerárquicos.

Ventajas:

- Permite visualizar relaciones entre datos y su evolución
- Maneja gran cantidad de datos

Desventajas:

- No es la adecuada para otros tipos de visualizaciones
- Curva de aprendizaje elevada

Conclusiones:

Es la herramienta perfecta si queremos representar redes con hasta 50000 nodos y más de 1.000.000 de aristas, pero en caso contrario, no será la solución adecuada

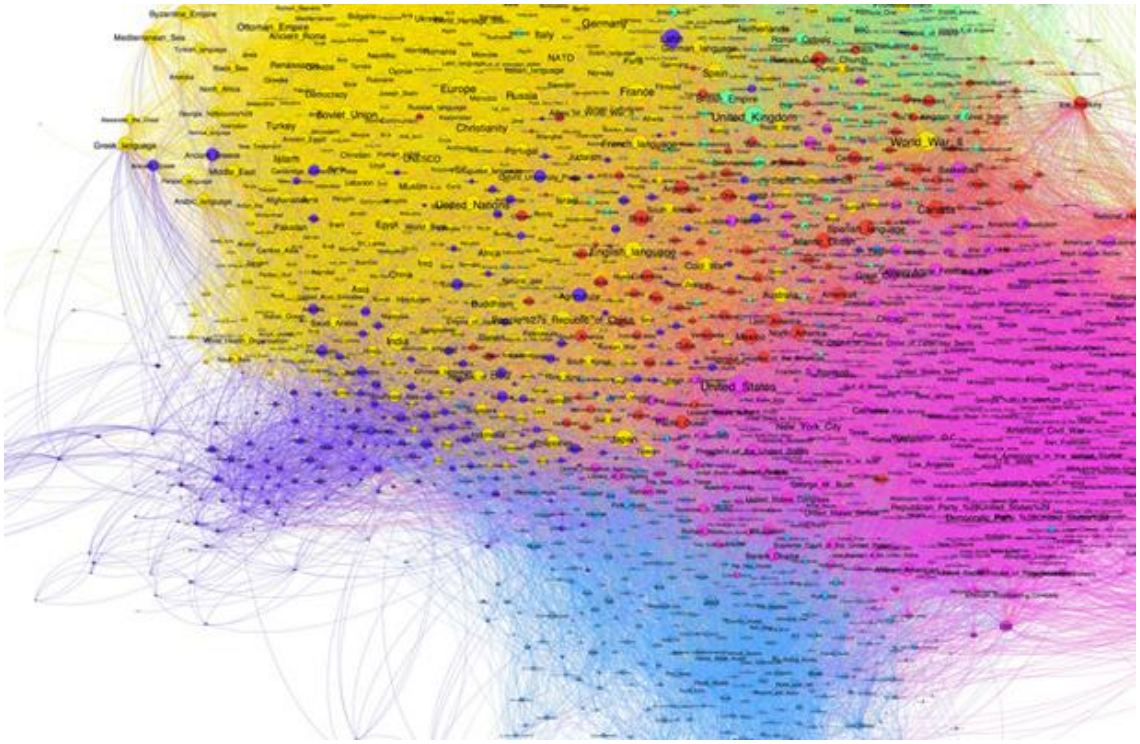

Figura 2.14. Ejemplo de visualización realizada con Gephi

Processing:

Es un lenguaje de programación creado por Ben Fry y Casey Reas en 2001, para la creación de visualización de datos interactivos. Es código abierto, permite la creación de librerías propias y la utilización un amplio abanico de librerías ya creada por una comunidad muy extensa de desarrolladores.

Ventajas:

- Gran capacidad de interactividad
- Puede gestionar gran cantidad de datos a la vez
- Velocidad de ejecución
- El límite en la creación de la visualización está, casi exclusivamente, en el desarrollador
- Es gratuita

- Utilización de librerías ya creadas (mapas, etc.).

Desventajas:

- No permite su exportación a los navegadores Web, aunque existe un proyecto hermano llamado “processingjs” que si dispone de esta posibilidad, aunque tiene un capacidad de cálculo y gestión de datos inferior.

Conclusiones:

Es una de las herramientas de visualización de datos más potentes que existen, es especialmente importante si se pretende una representación de colecciones de datos muy grande y si uno de los pilares de la visualización es la interactividad.

Figura 2.15. Ejemplo de visualización hecha con Processing

D3js:

Es una librería de JavaScript desarrollada por Michael Bostock que permite la creación de visualizaciones complejas y gráficos interactivos. Sus desarrollos son abiertos y pueden ser reimplementados por otros desarrolladores.

Ventajas:

- Se pueden hacer visualizaciones de todo tipo
- Es gratuita
- Sus posibilidades son tan amplias como la geometría misma

- Permite utilizar librerías ya creadas (mapas, etc.)
- Permite usar google maps, etc.
- Permite uso Web

Desventajas:

- Tiene una curva de aprendizaje elevada
- Para hacer cosas sencillas, puede ser lenta

Conclusiones:

Es una librería muy potente para los proyectos de visualización; si bien es verdad que no tiene un aprendizaje tan rápido como pueda tener Processing, permite a su vez el uso de interactividad, transiciones y un sinfín de posibilidades.

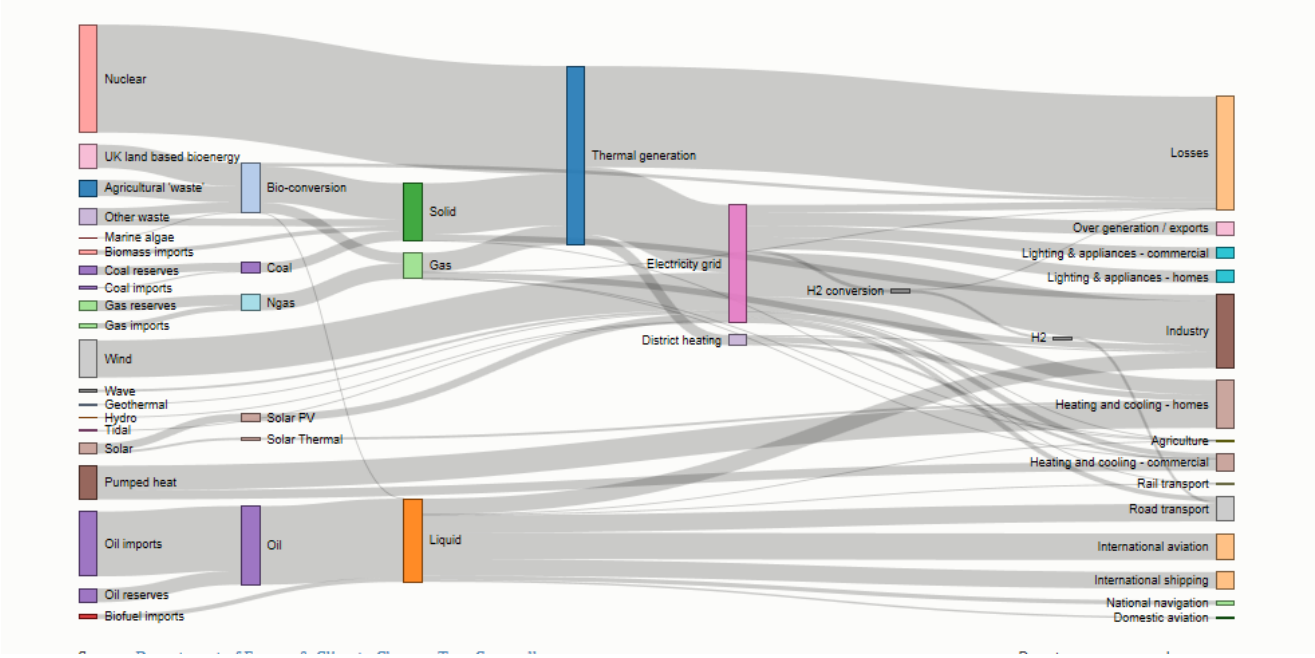

Figura 2.16. Ejemplo de visualización realizada con D3js

Herramienta	Categoría	Visualización multipropósito	Tipo	Tecnología	Licencia	Plataforma	Almacenamiento de datos	Publicación
Excel	Hojas de cálculo	No	Aplicación de escritorio	Windows	Office	Windows	Local	Como imagen
Tableau Public	Aplicación/ Servicio de visualización	Si	Aplicación de escritorio	Windows, JavaScript	Libre	Windows	Servidor público externo	Si
The R Project for Statistical Computing	Análisis estadístico	Si	Lenguaje de programa-ción	R	GPL	Linux, Mac OS X, Unix, Windows XP	Local	No
Gephi	Análisis de grafos	No	Aplicación de escritorio	Windows, Linux, MacOS X, Java	CDDL, GPL3	Equipos de escritorio que ejecutan Java	Local	Como imagen
Processing	Servicio de visualización	Si	Lenguaje de programa-ción	Java	Libre	Linux, Mac OS X, Unix, Windows XP	Servidor local o externo	No
D3.js	Librería	Si	Biblioteca	JavaScript	BSD	Editor de código y navegador	Servidor local o externo	Si

Tabla 2.1. Tabla resumen de herramientas de visualización

3 Resultado

Los objetivos fundamentales de este proyecto fin de Máster, son:

1. La elaboración de una aplicación que permita supervisar los retrasos que se producen en las distintas marquesinas de Gijón, en las distintas horas y los distintos días de la semana, en las distintas líneas del sistema de transporte urbano. Pudiendo analizar los perfiles de los retrasos.

En el capítulo 2 se revisaron numerosas técnicas de visualización de datos, identificando las que se consideraban más adecuadas para la identificación de patrones y extracción de conocimiento en autocorrelación espacial y series temporales.

Durante este capítulo vamos a describir las soluciones adoptadas tanto en la gestión de los datos como en la elaboración de la aplicación de visualización de datos.

3.1 Descripción general del proyecto

El desarrollo del proyecto consta de varias fases (Figura 3.1), que se han agrupado en:

- Gestión de datos
- Visualización de datos
- Conclusiones

En la gestión de datos, incluyo la adquisición, parseado, filtrado y los distintos cálculos que se han realizado para obtener los retrasos que se producen en las distintas marquesinas de Gijón, los distintos días de la semana y a cada hora del día.

En el apartado de la visualización, se incluye la codificación visual que se ha utilizado, así como las distintas herramientas de interactividad que se han incluido en la aplicación.

En las conclusiones se incluyen las discusiones generales, las aportaciones que se han hecho con el proyecto y las líneas de futuro trabajo.

En el presente capítulo se aborda el apartado de Gestión y Visualización de datos. El de conclusiones se incluye en el siguiente capítulo de la memoria.

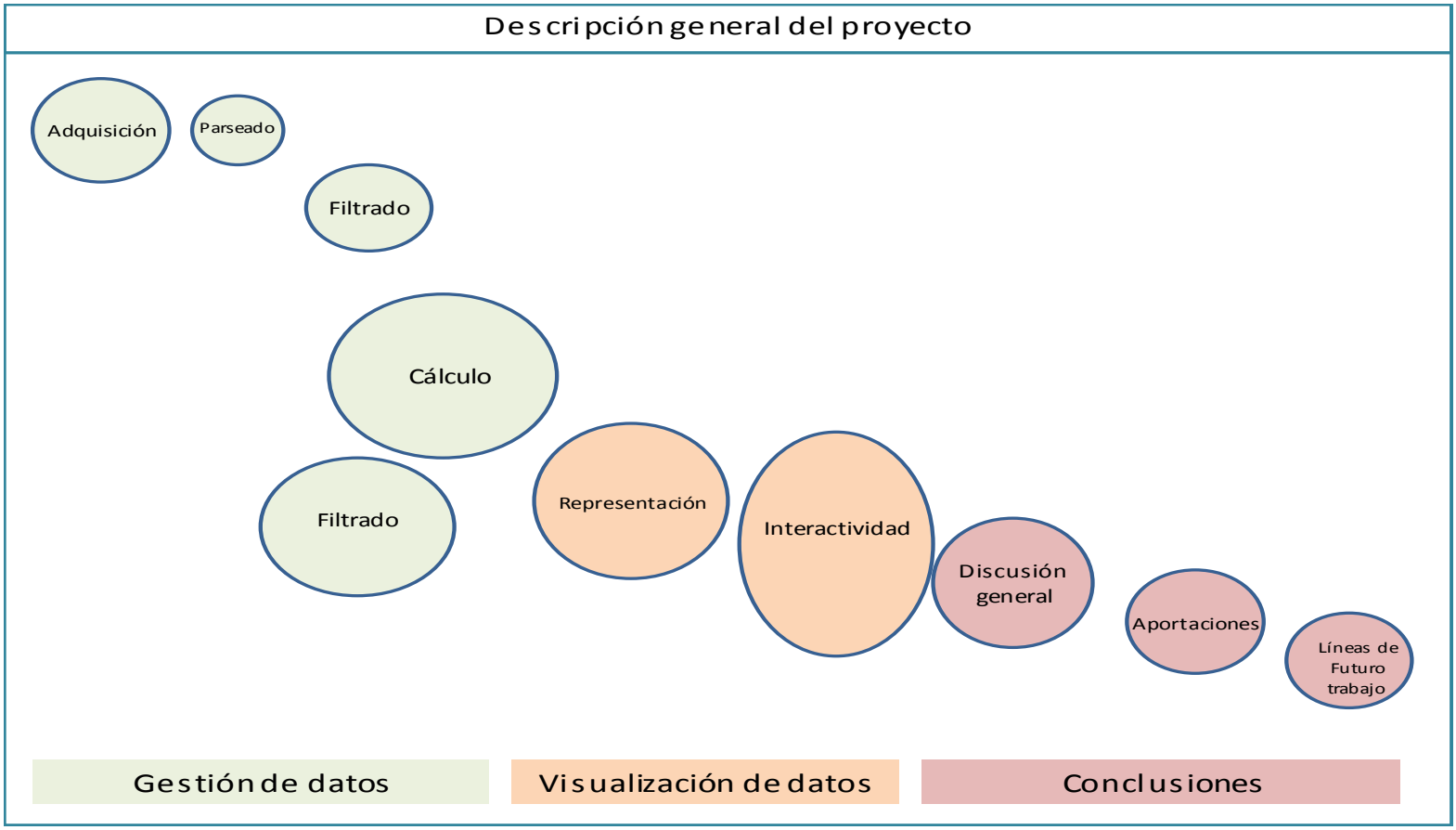

Figura 3.1. Descripción del funcionamiento de la aplicación

3.2 Gestión de datos

El objetivo fundamental de la Gestión de datos es la de recoger (Adquisición de los datos), manipular (parsear), seleccionar (filtrar) y procesar (cálculo) los datos para posteriormente, hacer el análisis de los mismos.

Cómo se ha mencionado anteriormente esta fase consta de cuatro apartados:

- Adquisición de datos

- Parseado de los datos
- Filtrado de los datos
- Cálculo

3.2.1 Adquisición de datos y parseado

El ayuntamiento de Gijón a través del portal web http://datos.gijon.es, dispone cada intervalo de 30''-1'30'' en formato XML en http://datos.gijon.es/doc/transporte/busgijontr.xml, los siguientes datos de las marquesinas (Figura 3.2):

- Autobús que va a llegar a la marquesina
- La matrícula del autobús
- El modelo del autobús que va a llegar
- La línea a la que pertenece el autobús
- La trayectoria dentro de la línea a la que corresponde el autobús
- La marquesina a la que se refieren los datos
- El tiempo que falta para que llegue el autobús a la marquesina
- La distancia entre el autobús y la parada
- Hora a la que se ha producido la última actualización
- Fecha en la que se ha producido la actualización
-

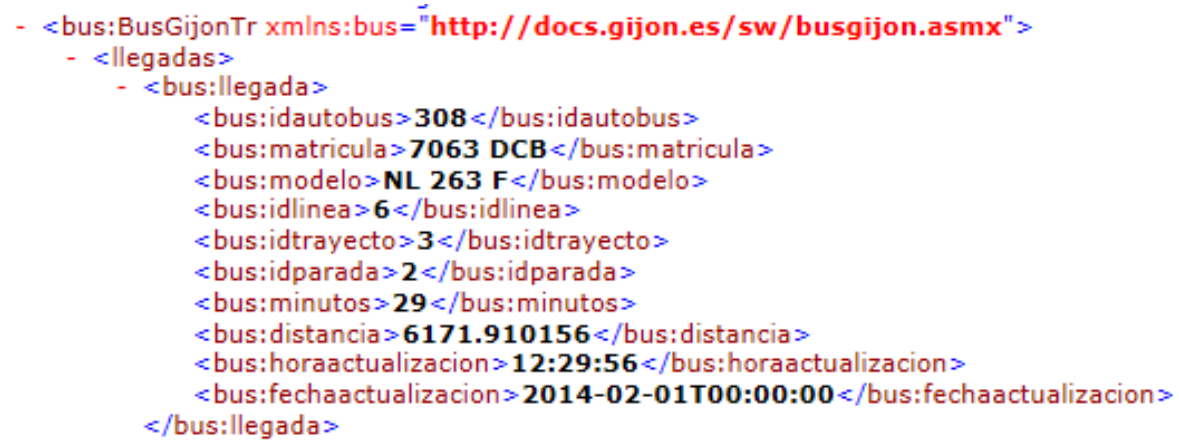

Figura 3.2. Información sobre la marquesina 2, publicada en el portal datos.gijón.es

Lo que se ha hecho en este proyecto, es almacenar los datos que publica el ayuntamiento de Gijón en una base de datos a través de un script de Python. Cada vez que se publica un dato

nuevo de cada marquesina, el script verifica que ese dato no esté ya incluido en la base de datos, y lo almacena tal y como se publica; se guardan todos los datos en crudo (sin modificar), siguiendo uno de los principios fundamentales en los sistemas de gestión de la información, que consiste en almacenar los datos tal y como se recopilaron. Posteriormente se transforma la fecha parseada a timestamp UNIX.

En el proceso de transformación del par fecha-hora en timestamp UNIX, hubo que gestionar una excepción que se producía en la publicación de datos. En el intervalo entre las 00:00 y 04.30 am, los autobuses y marquesinas de Gijón no actualizaban la fecha, manteniendo la fecha del día anterior.

3.2.2 Filtrado de datos

Una vez capturados los datos, se filtraron. Este proceso se hizo fundamentalmente con dos fines:

- Utilizar sólo los datos que se consideraban útiles para el proyecto
- Eliminar los datos que provocaban outliers en la aplicación

El filtrado de los datos que se consideraban útiles para el proyecto, se realizó previamente a la aplicación de Processing. Se rechazaron los datos que por su naturaleza, se consideró que no aportaban información útil para el proyecto.

Por otro lado, el filtrado de los datos que provocaban outliers en la visualización, se hizo posteriormente a la aplicación de Processing. Gracias a la geovisualización, tal y como se había mencionado en el capítulo 2, se identificaron varios outliers, que obligaron a llevar a cabo una investigación en los datos en crudo, que permitió detectar y solventar las anomalías en la emisión de los datos que provocaban estos outliers.

Utilizar los datos que se consideran útiles para el proyecto:

Se decidió inicialmente suprimir los datos que se publicasen en las marquesinas que no estuviesen dentro de un margen temporal, en el presente proyecto el margen se fijó en 30 minutos, se valoró que el tiempo útil de espera de un usuario no excedería los 30 minutos, ya que en líneas residuales se llega a informar de la llegada de autobuses con tiempos de espera superiores a 2 horas. Pero parece lógico pensar que ningún usuario estará esperando un autobús 2 horas antes de su llegada.

Eliminar los datos que provocaban outliers en la aplicación:

Gracias a la geovisualización, en la visualización de datos se detectaron dos tipos de anomalías distintas que provocaban outliers, ya que no mantenían el principio de vecindad en ningún sentido del trayecto, ya sea en procedencia o en destino:

- Las marquesinas publican periódicamente la información del autobús más cercano de cada línea y trayecto que pasa por ella, pero algunas veces, se produce un error y en vez de publicar la información del autobús mas cercano, publica la información de otro autobús de esa misma línea y trayecto, dando lugar a errores que modifican los valores en las medidas. En la Figura 3.3, se muestra un ejemplo de esta anomalía.

Hora	13:31:10	13:31:40	13:32:10	13:32:40
Idautobus	101	101	145	101
Idlinea	1	1	1	1
Idtrayecto	2	2	2	2
Idparada	5	5	5	5
Minutos	10	9	29	8
Distancia	2500	2200	6854	1750

Figura 3.3. Error que se produce en la información publicada por las marquesinas de Gijón

- Las marquesinas pueden dejar de publicar información sin previo aviso, perdiendo el servicio de publicación. Sucedería algo parecido al ejemplo de la Figura 3.4. Esto provocaba que los datos que se estaban publicando cuando se perdía el servicio, se contabilizasen con los datos que se publicaban cuando este volvía a funcionar. Este problema se manifestaba con valores muy elevados en puntos muy concretos, por lo que produjo los outliers que se identificaron más rápido gracias a la Geovisualización.

Hora	13:31:10	13:31:40	17:21:10	17:21:40
Idautobus	137	137	145	145
Idlinea	4	4	4	4
idtrayecto	6	6	6	6
idparada	5	5	5	5
minutos	10	9	16	15
distancia	2500	2200	3322	3200

Figura 3.4 Error que se produce cuando se pierde el servicio en la publicación de la información

3.2.3 Cálculo de datos

En la aplicación se visualizan cuatro datos distintos, el cálculo de ninguno de ellos es de una complejidad elevada. Son los retrasos que se producen en las marquesinas de Gijón a cada hora en un día determinado o en un grupo de varios días (si así lo requiere el usuario), los retrasos que se producen de cada línea en cada marquesina y los retrasos que se producen en todas las marquesinas de Gijón en cada hora.

Como se observa en la Figura 3.2, en el catálogo de datos que se publican no se encuentra ninguno de los datos con los que trabaja la aplicación, por lo que se deberán calcular a partir de los datos que se publican en el portal web. En la Figura 3.4, se describe el proceso que se ha seguido en el cálculo de los datos que utiliza la aplicación.

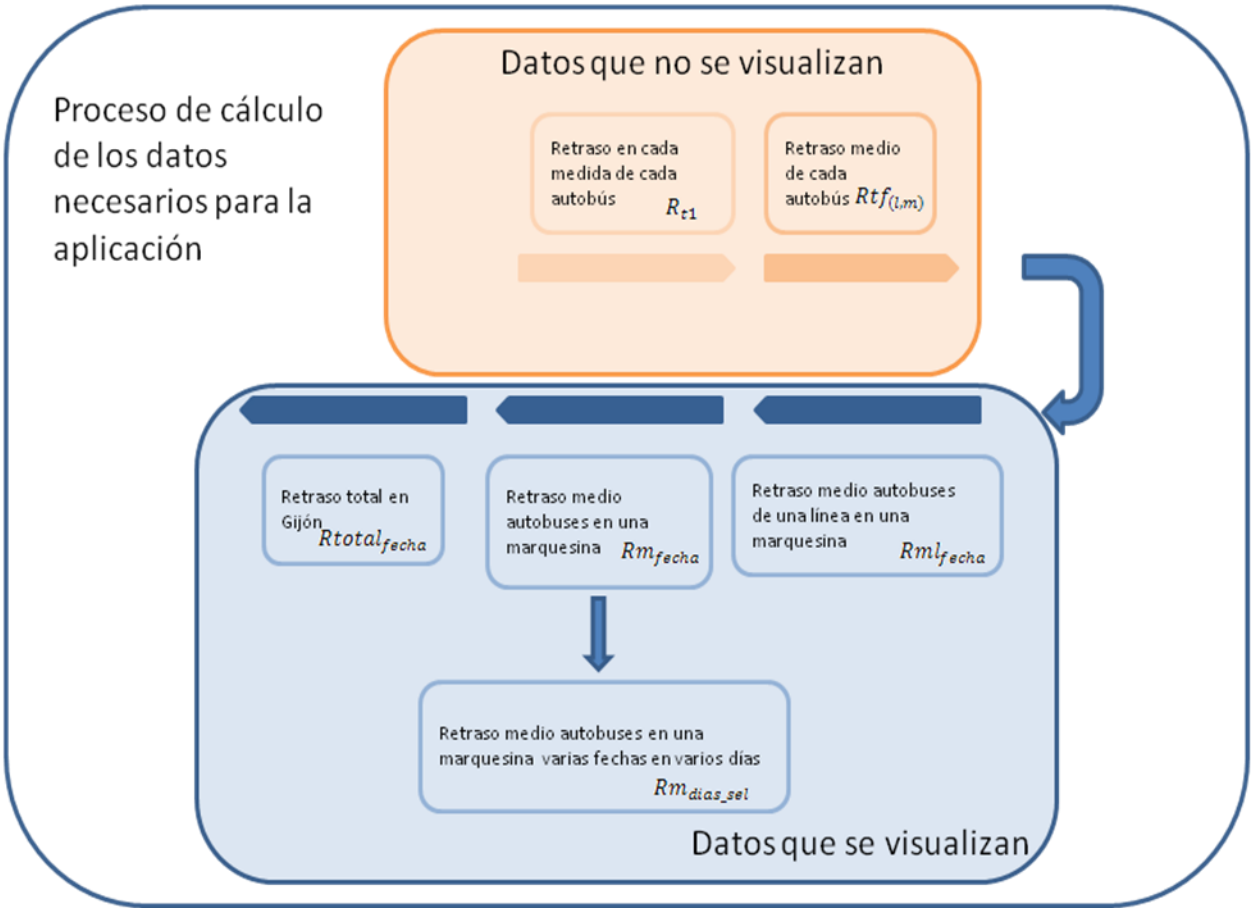

Figura 3.4. Proceso de cálculo de datos de la aplicación

Retraso producido en cada publicación de la marquesina:

Para el cálculo del retraso en cada publicación, se sigue la Ecuación 3.1, donde:

- R_{ti}, es el retraso en la publicación i
- $\Delta h_{llegada-ti}$, es la diferencia en minutos de la hora a la que llega el autobús a la marquesina y de la hora a la que se realizó la publicación i. Es decir el tiempo real transcurrido desde la publicación hasta la llegada
- T_{ti}, tiempo esperado en minutos que falta para que llegue el autobús en el momento de la publicación i

$$R_{ti} = \Delta h_{llegada-t1} - T_{ti}$$

Ecuación 3.1 Ecuación que calcula el retraso en el timestamp i

Retraso medio de un autobús en una marquesina:

Una vez que se calculan todos los retrasos acumulados hasta que el autobús llega a la marquesina, se calcula su media siguiendo la Ecuación 3.2, donde:

- $Rtf_{(l,m,t)}$, es el retraso medio producido por el autobús de la línea (l) en la marquesina (m) en el momento (t). El t se corresponde con la hora a la que el autobús pasa por la marquesina (m)
- $\sum_{i=0}^{N-1} R_{ti}$, es el sumatorio de retrasos desde el primer registro i=0, hasta el penúltimo, ya que el último por la naturaleza de la Ecuación 3.1 será 0
- N, es el número de registros que tenemos

$$Rtf_{(l,m,t)} = \sum_{i=0}^{N-1} R_{ti} \Big/ N$$

Ecuación 3.2 Ecuación que calcula el retraso medio del autobús

Retraso medio de los autobuses de una misma línea en una marquesina a una hora determinada en un día:

Ahora se calcula el retraso que producen todos los autobuses de una línea, en una marquesina a una hora determinada. Este cálculo se hace siguiendo la Ecuación 3.3, donde:

- Rml_{fecha}, es el retraso medio en cada marquesina producido por los autobuses de una misma línea en una hora determinada
- $Rtf_{(l,m,t)}$, es el retraso medio calculado en la Ecuación 3.2
- N, es el número de $Rtf_{(l,m)}$ que tenemos en cada hora

$$Rml_{fecha} = \sum_{i=0}^{N} Rtf_i \Big/ N$$

Ecuación 3.3 Ecuación que calcula el retraso medio producido en la marquesina por una línea en una fecha (por horas)

Retraso medio en una marquesina:

Otro dato que se utiliza en la aplicación es el retraso medio de los autobuses que llegan a cualquier marquesina a cada hora. Para calcular la media de estos retrasos, se sigue la Ecuación 3.4, donde:

- Rm_{fecha}, es el retraso medio en cada marquesina a cada hora
- Rml_{fecha}, es el retraso medio producido en la marquesina m por los autobuses de la línea l, calculado en la Ecuación 3.3
- N, es el número de Rml_{fecha} que tenemos en la marquesina m, en cada hora

$$Rm_{fecha} = \sum_{i=0}^{N} Rml_{fecha} \Big/ N$$

Ecuación 3.4 Ecuación que calcula el retraso medio producido en la marquesina en una fecha a una hora determinada

Retraso medio en una marquesina a una hora determinada en varios días:

Una vez que se ha hecho el cálculo para un día determinado, si el usuario de la aplicación desea visualizar el retraso medio de un grupo de varios días, la aplicación calcula una media con los retrasos medios que se hayan producido en esos días, tal y como muestra la Ecuación 3.5, donde:

- Rm_{dias_sel}, es el retraso medio en cada marquesina a cada hora en varios días
- Rm_{fecha}, es el retraso medio producido en la marquesina en una hora determinada, calculado en la Ecuación 3.4
- N, es el número de Rm_{fecha} que tenemos en los días seleccionados

$$Rm_{dias\_sel} = \sum_{i=0}^{N} Rm_{fecha} \Big/ N$$

Ecuación 3.5 Ecuación que calcula el retraso medio producido en la marquesina en una hora determinada, en los días seleccionados

Retraso medio de todas las marquesinas de Gijón en cada hora:

El último dato que se utiliza en la aplicación, es el retraso medio que se produce en todas la marquesinas Gijón en cada hora. Para ello, se sigue la Ecuación 3.6, donde:

- $Rtotal_{fecha}$, es el retraso producido en todas las marquesinas en una hora determinada
- Rm_{fecha}, es el retraso medio en cada marquesina en una hora determinada, calculado en la Ecuación 3.5
- N, es el número de Rm_{fecha} que tenemos, en una hora determinada

$$Rtotal_{fecha} = \sum_{i=0}^{N} Rm_{fecha} \Big/ N$$

Ecuación 3.6 Ecuación que calcula el retraso medio producido en todas las marquesinas en una fecha en una hora determinada

Siguiendo el proceso de la Figura 3.4, tenemos calculados los valores que se utilizan en la aplicación:

- Rml_{fecha}
- Rm_{fecha}
- Rm_{dias_sel}
- $Rtotal_{fecha}$

3.3 Visualización de datos

Una vez obtenidos los datos que se van a visualizar, hay que diseñar la aplicación; se debe tener especial cuidado con la codificación visual que se vaya a utilizar y con las herramientas de interactividad que se deseen integrar para la correcta explotación de los datos.

- Codificación visual
- Interactividad

3.3.1 Codificación visual

La aplicación consta de varias técnicas de visualización, que están vinculadas entre sí y permiten la explotación detallada de los datos. Estas herramientas son:

- Mapa, donde se proyectan los retrasos medios de cada marquesina en una hora determinada, ya sea en un día o en varios días seleccionados. Rm_{fecha} y Rm_{dias_sel}
- Calendario, donde se proyectan los retrasos medios en Gijón en una hora determinada. $Rtotal_{fecha}$.
- Diagramas de barras (Barchart), donde se proyectan los perfiles de retrasos medios horarios de cada línea que pasa por la marquesina, ya sea de un día o de varios días seleccionados. Rml_{fecha}

Con la combinación de estas tres herramientas se puede acceder a información detallada cuando lo requiera el usuario. A continuación se detallan estas tres herramientas y se especifican las distintas codificaciones visuales que se han utilizado.

Mapa donde se proyectan los retrasos medios producidos en cada marquesina en una hora(en un día concreto o en varios días):

Codificación espacial:

Al proyectar los puntos sobre un mapa, se codifica la localización de la marquesina (latitud, longitud). Además, tal y como se puede ver en la Figura 3.6, se utiliza un mapa de carreteras de Gijón que permite contextualizar las marquesinas en el recorrido que siguen los autobuses de cada línea.

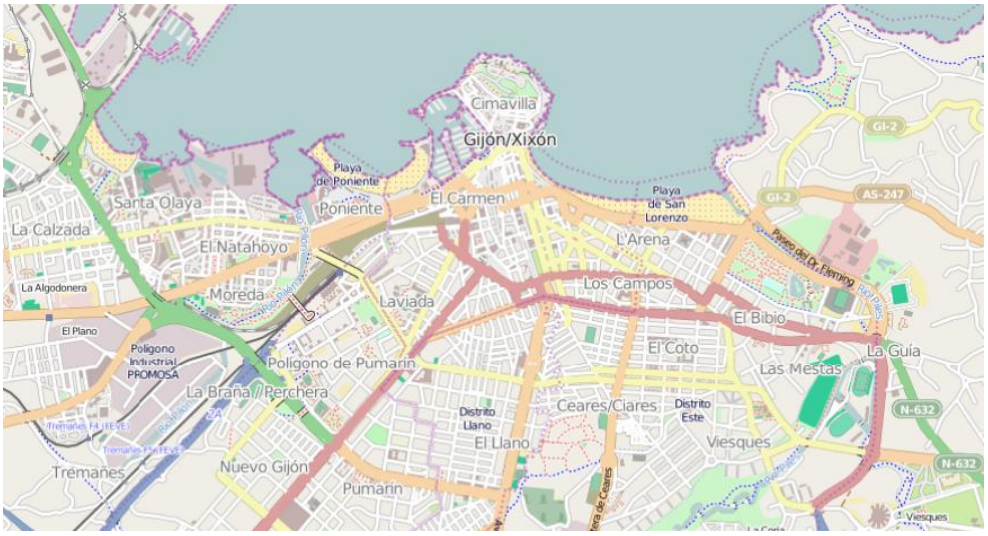

Figura 3.6. Mapa de Gijón donde se proyectan los retrasos medios en cada marquesina y hora

Codificación Área:

El valor absoluto de los retrasos viene codificado por el área del círculo, en la Figura 3.7 se puede observar la relación entre el área y la información que se codifica. Se asigna un área determinada al valor absoluto de cada retraso. Esto significa que -1' y +1´ de retrasos, tendrán la misma codificación de área, pero no así de color, como se explica en el siguiente apartado.

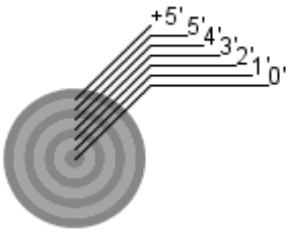

Figura 3.7 Áreas de círculos y el valor absoluto en minutos que codifican. (Escala 1:1)

Codificación de color:

El valor numérico del adelanto o del retraso, viene codificado por una escala de color divergente. Tal y como se explicó en el capítulo 2, los atributos que pueden ser divididos en dos secuencias que confluyen en un punto común (el 0) son divergentes, por esta razón se ha utilizado una codificación de color adecuada para este tipo de información, tal y como se observa en la Figura 3.8.

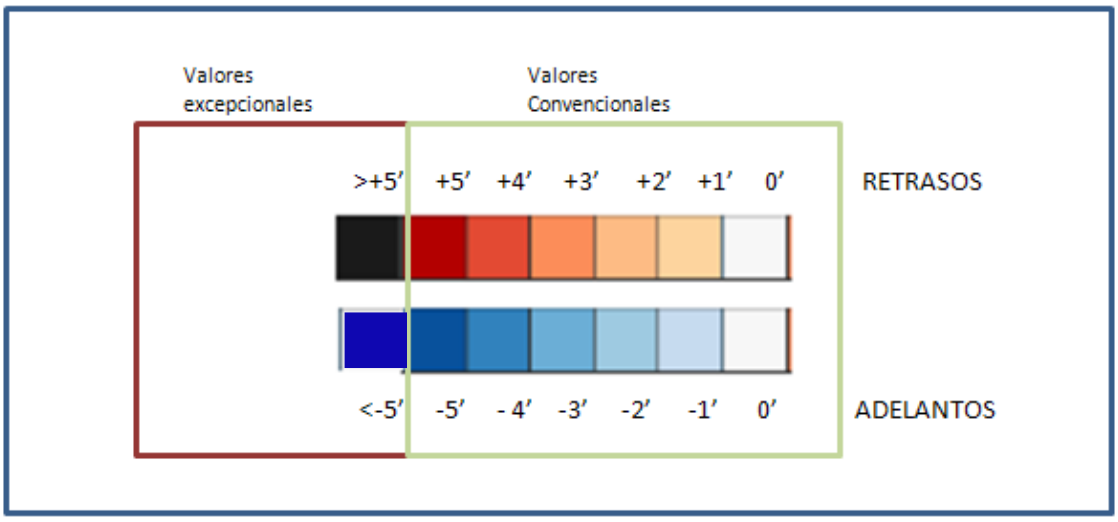

Figura 3.8 Codificación de color con el valor de los retrasos que codifican

Se utiliza para los retrasos, una codificación de color que va desde el blanco hasta el rojo. El negro se utiliza en este caso para un retraso excepcional.

Por otro lado los retrasos negativos o adelantos, utilizan una codificación de azules.

En la imagen 3.9, se observa la aplicación funcionando. En ella se puede ver como los retrasos están codificados tal y como se ha explicado en este apartado.

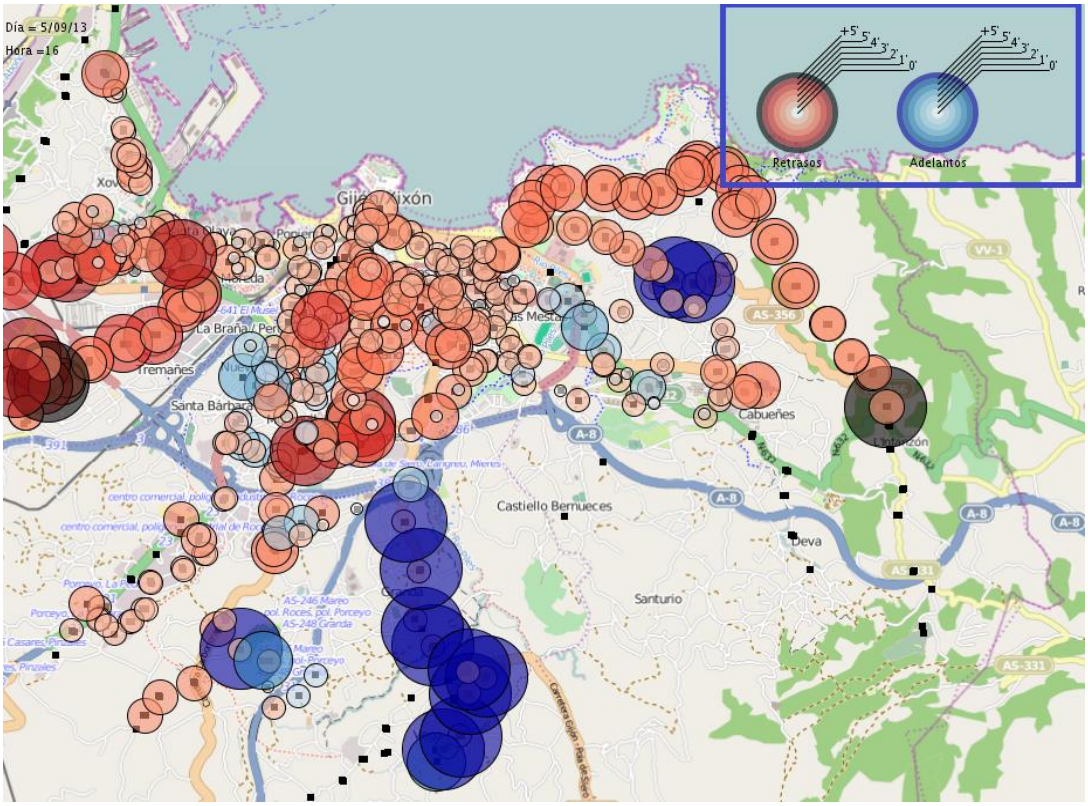

Figura 3.9 Aplicación funcionando, con detalle en cuadrado azul donde aparecen la leyenda de los retrasos y los adelantos

<u>Calendario donde se proyectan los retrasos medios horarios de Gijón:</u>

Codificación espacial:

En el calendario, el espacio se usa para situar en orden cronológico los días del mes (Figura 3.10). Sobre cada día se proyecta el valor medio de los retrasos en Gijón a la hora seleccionada.

Figura 3.10 Calendario donde se proyectan los retrasos medios horarios de Gijón

Codificación de área y de color:

Se siguen los mismos principios que en la proyección en el mapa. La única diferencia radica en que el retraso/adelanto 0 y retraso/adelanto +1, utilizan el mismo área para su codificación ya que si el retraso/adelanto 0 utilizaba una codificación de área menor, interfería con la identificación del día. Se ha seguido el mismo criterio que en la proyección en mapas, persiguiendo el principio de consistencia, que facilita el entendimiento del usuario. En la Figura 3.11, se muestra el calendario con los retrasos producidos en el mes de septiembre a las 5 de la mañana.

Figura 3.11 Calendario con la proyección de los retrasos medios a las 5.00 am en Gijón, durante el mes de septiembre

El calendario también informa de los días que están seleccionados en ese momento, aumentando el tamaño y coloreando en rojo el día en el calendario (Figura 3.12).

Figura 3.12 Calendario, detalle en circulo azul del día seleccionado

Diagramas de barras (Barchart) donde se proyectan los perfiles de retrasos medios horarios de cada línea que pasa por la marquesina.

Este tipo de representación se ha incluido para añadir información a la que se muestra en los mapas, pudiendo ver los perfiles de retrasos de las distintas líneas en las 24 horas del día en cada marquesina. Se pueden mostrar hasta 7 días distintos.

Codificación espacial:

Se utiliza el eje **X**, para codificar la hora a la que se produce el adelanto o el retraso, y el eje **Y** identifica de qué línea se trata y, dentro de cada línea, codifica el retraso o adelanto que se produce (Figura 3.13).

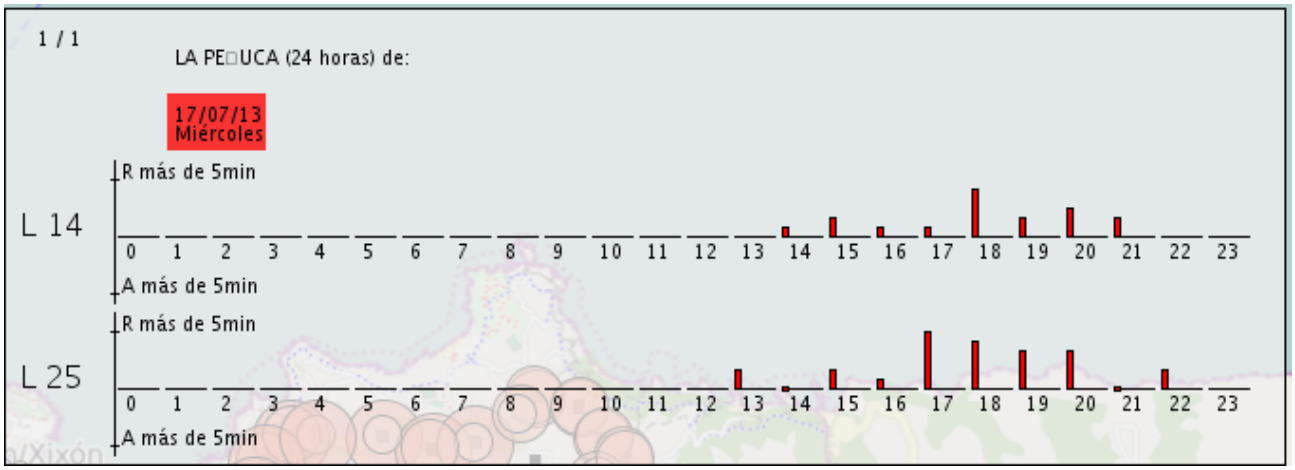

Figura 3.13 Diagrama de barras en el que se muestran los perfiles de retrasos de la marquesina La Peñuca el 17/07/13

Codificación de color:

En este caso, lo que se pretende codificar es el día al que corresponden los retrasos o adelantos, y cómo el usuario puede seleccionar hasta 7 días distintos. Lo que se ha decidido es utilizar una escala de colores categórica.

Figura 3.14 Codificación categórica en función del día que representa

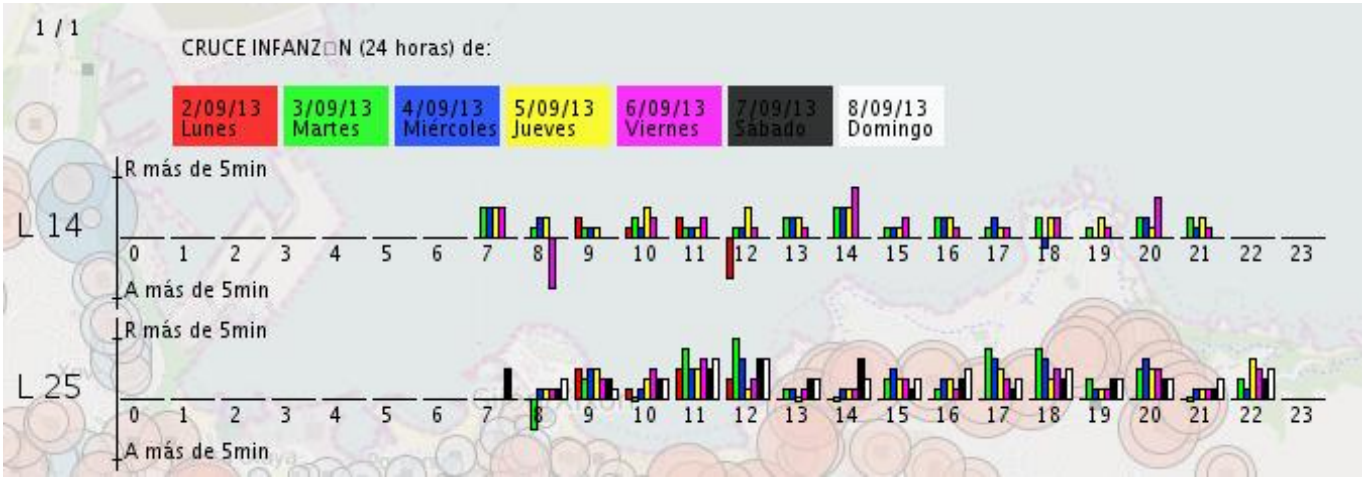

Figura 3.15 Diagrama de barras mostrando el perfil de retrasos de 7 días distintos

3.3.2 Interactividad

Una vez descritas las distintas técnicas de visualizaciones de datos que permite la aplicación, falta describir las herramientas de interactividad disponibles al servicio del usuario en la aplicación. La interactividad es posiblemente el punto fuerte de esta aplicación, ya que posee varias herramientas que permiten al usuario obtener información más detallada en el momento que lo desee.

Las herramientas de interactividad de la aplicación, son:

- Barra de Scroll
- Calendario
- Selección de marquesina
- Selección de retrasos específicos

Barra de Scroll:

Es una barra de Scroll doble, que permite cambiar el día y la hora de manera dinámica. Al mover la barra se puede generar una animación que muestra cómo van evolucionando los retrasos en las marquesinas de Gijón (Figura 3.16). A la derecha dispone de una leyenda que informa del día y la hora seleccionadas; en el caso de la Figura 3.16, aparece "8/09/13 + …", esto es porque se están visualizando la media de varios días.

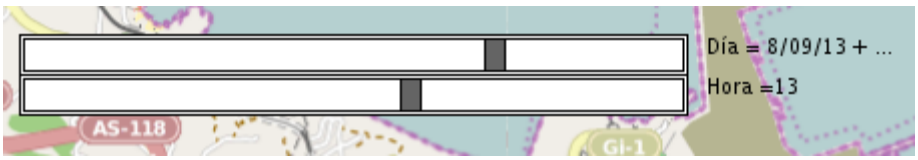

Figura 3.16 Barra de Scroll que permite la navegación temporal de los datos.

Calendario:

Es una de las herramientas más versátiles de la aplicación, permite seleccionar el día del que se quiere que se representen los retrasos, pulsando encima del día, pero además permite la opción de seleccionar varios días, lo cual variará la información que se representa en las herramientas de la visualización:

- El mapa representará la media de retrasos entre los días seleccionados en las distintas marquesinas a la hora seleccionada.
- El diagrama de barras mostrará los perfiles de los distintos días seleccionados

Si se pulsa con el botón derecho del ratón dentro del área del calendario, se verán semitransparentes los días y de color oscuro, los retrasos o adelantos (Figura 3.17).

Figura 3.17 Calendario con los retrasos más oscuros y los días semitransparentes

Selección de la marquesina:

Cuando se pulsa con el botón secundario al lado de una marquesina, la aplicación muestra los diagramas de barras de la marquesina seleccionada (Figura 3.15 y Figura 3.13).

Selección de retrasos específicos:

Si se selecciona con el ratón alguno de los círculos de la leyenda, se visualizarán en verde fosforito las marquesinas que tienen ese mismo retraso (Figura 3.18).

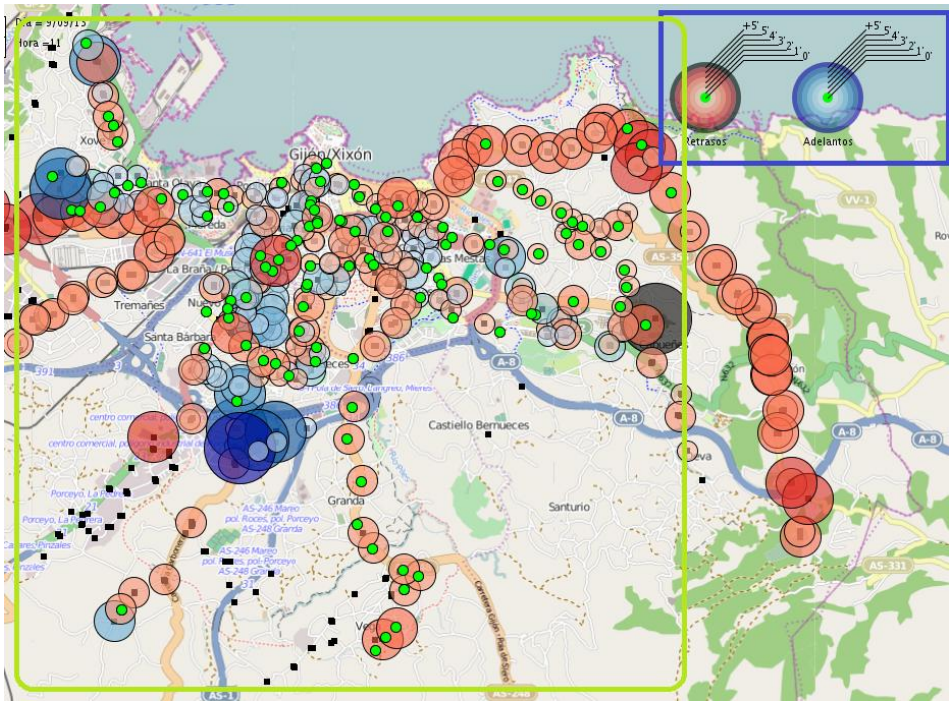

Figura 3.18 Aplicación, detalle recuadro azul la leyenda seleccionada y en el detalle recuadro verde se ve las marquesinas con retraso/adelanto 0 en verde fosforito.

3.4 Caso de uso

A continuación, se muestra un caso práctico del uso de la aplicación, en el que se muestra cómo influye la celebración de un evento como "el rastro" en los retrasos de los autobuses de Gijón. Para ello, lo más adecuado es realizar un análisis de los retrasos producidos durante dos días, un día en el que no se celebra el citado mercadillo (por ejemplo, el miércoles 4 de septiembre) y otro día en el que sí que se celebra (por ejemplo, el domingo 8 de septiembre). El análisis se centra en la zona este de la ciudad (Figura 3.19). Los barrios que se van a analizar en más detalle son:

- Periurbano rural
- La arena
- El bibio-parque
- El coto
- Las Mestas
- Viesques

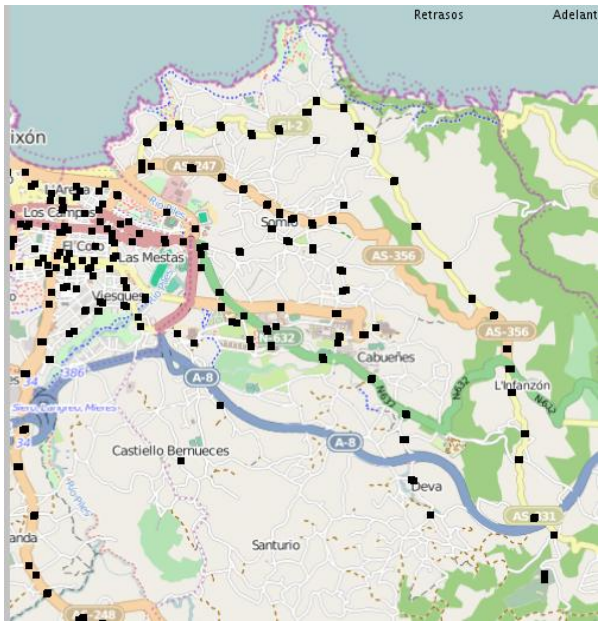

Figura 3.19 Mapa y zona de estudio

Este caso ejemplo consta de tres partes. Inicialmente, se hace un análisis de los retrasos medios en Gijón en los primeros días de septiembre, para intentar establecer una relación entre los retrasos que se producen en los autobuses y las actividades diarias que se realizan en la ciudad; después se estudiarán con mas detalle los retrasos que se producen en las marquesinas de la zona de estudio los días 4 y 8 de septiembre y, por último, se hará el análisis de los perfiles de los retrasos en las líneas que pasan por los alrededores del mercadillo.

3.4.1 Análisis de los retrasos medios producidos en Gijón

Para llevar a cabo este análisis, hay que prestar atención a los datos que se proyectan en el calendario, lo que permitirá hacerse una idea global de los retrasos que se producen en Gijón a una hora determinada. En este caso se han seleccionado tres horas de la mañana, tres horas de la tarde y otras tres de la noche, tal y como se puede observar en la Figura 3.20, pudiendo estudiarse la evolución de los retrasos en los distintos días del mes.

Como muestra la figura 3.20, los retrasos que se producen durante los días lectivos son mayores que los del fin de semana. Si se observa detenidamente la figura, los retrasos medios son de aproximadamente 1 minuto los días lectivos y nulos los fines de semana.

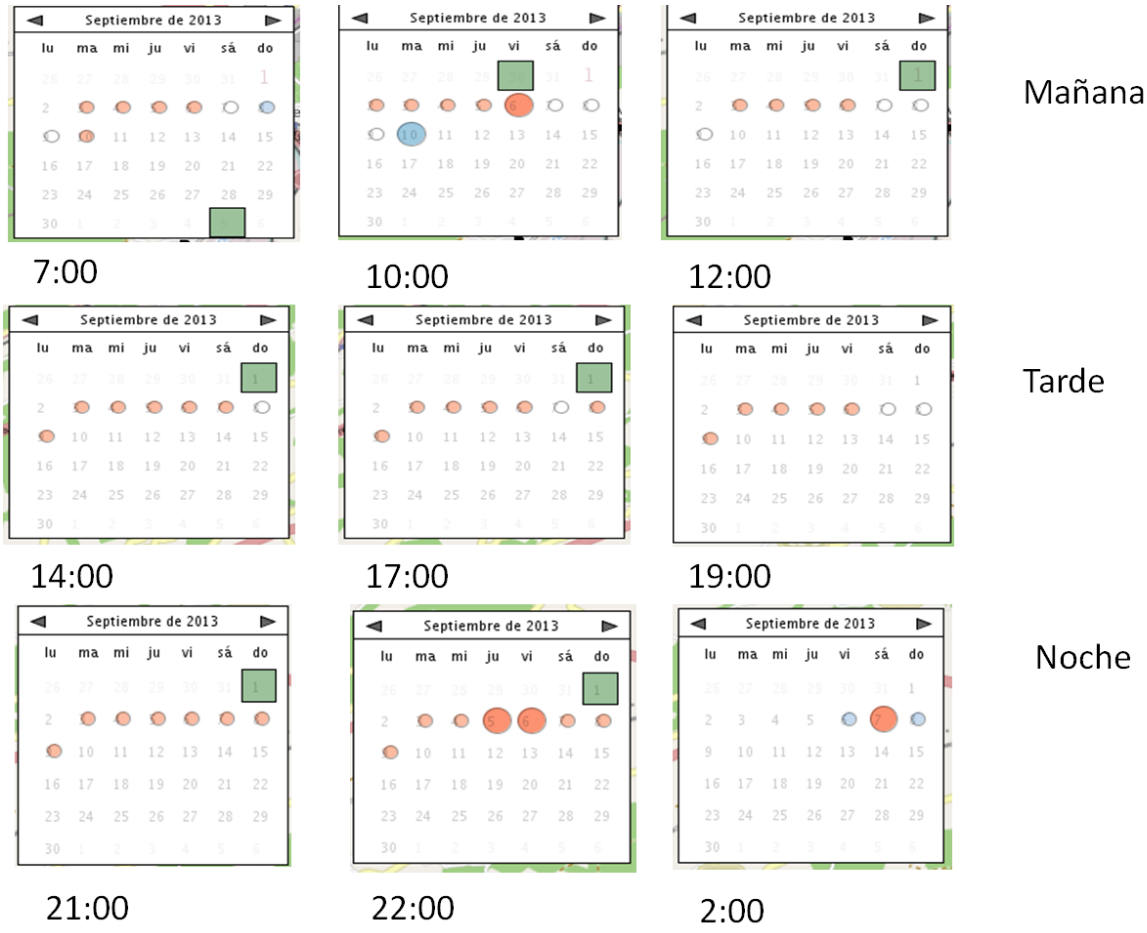

Figura 3.20 Retrasos medios en Gijón a lo largo del día

Se aprecia que los días lectivos mantienen retrasos de un minuto a lo largo de todo el día. Sin embargo durante el fin de semana, a lo largo de la mañana no se producen retrasos, es al llegar la tarde, sobre las 14:00 horas cuando aparece el primer retraso del fin de semana. Es especialmente interesante ver la evolución de las noches de los fines de semana, pudiendo ser significativo de la actividad nocturna de la ciudad, los retrasos de dos minutos que se producen a las 22:00 el viernes y a las 2:00 el sábado.

3.4.2 Análisis de los retrasos medios en las marquesinas

Ahora, para intentar dar respuesta a la influencia que tiene el rastro de Gijón en los retrasos de los autobuses, se estudian los retrasos en las marquesinas de la zona de este. En la figura 3.21,

aparecen los retrasos que se han producido en las horas más significativas del día en las dos fechas seleccionadas. A partir de estas visualizaciones, se puede deducir que:

- El miércoles, a las 7:00 horas de la mañana, ya hay bastante actividad, tanto en las salidas de Gijón como en el centro. En cambio, el domingo a la misma hora, tal y como parecía predecible existe mucha menos actividad, y los datos de los retrasos/adelantos presentan una varianza mayor. Esto se debe a que en Gijón el domingo, la mayor parte de los servicios de autobús comienzan a esta hora y la varianza en estos datos se debe a que los conductores no son extremadamente precisos con la hora a la que comienza el servicio, pudiendo salir con unos minutos de adelanto o retraso.
- El miércoles a las 16:00 horas se aprecian retrasos elevados y aunque a las 21:00 horas pudiera parecer que no existen demasiados retrasos, se debe a que a principios de septiembre (que son las fechas que se están analizando) aún no había comenzado el curso universitario. Pero si se observa el cuadrado azul, en el que se han seleccionado otras entradas y salidas de Gijón, sí que se producen retrasos elevados a esta hora en los días laborables. Tiene sentido que en los días laborables a estas horas se produzcan retrasos en los autobuses, porque se corresponden con las horas de la comida y de fin de jornada.
- Por otro lado, el domingo mantiene en líneas generales pocos retrasos en la zona este de la ciudad, lo que también era previsible, ya que la actividad de la ciudad es menor el domingo que los días laborables.

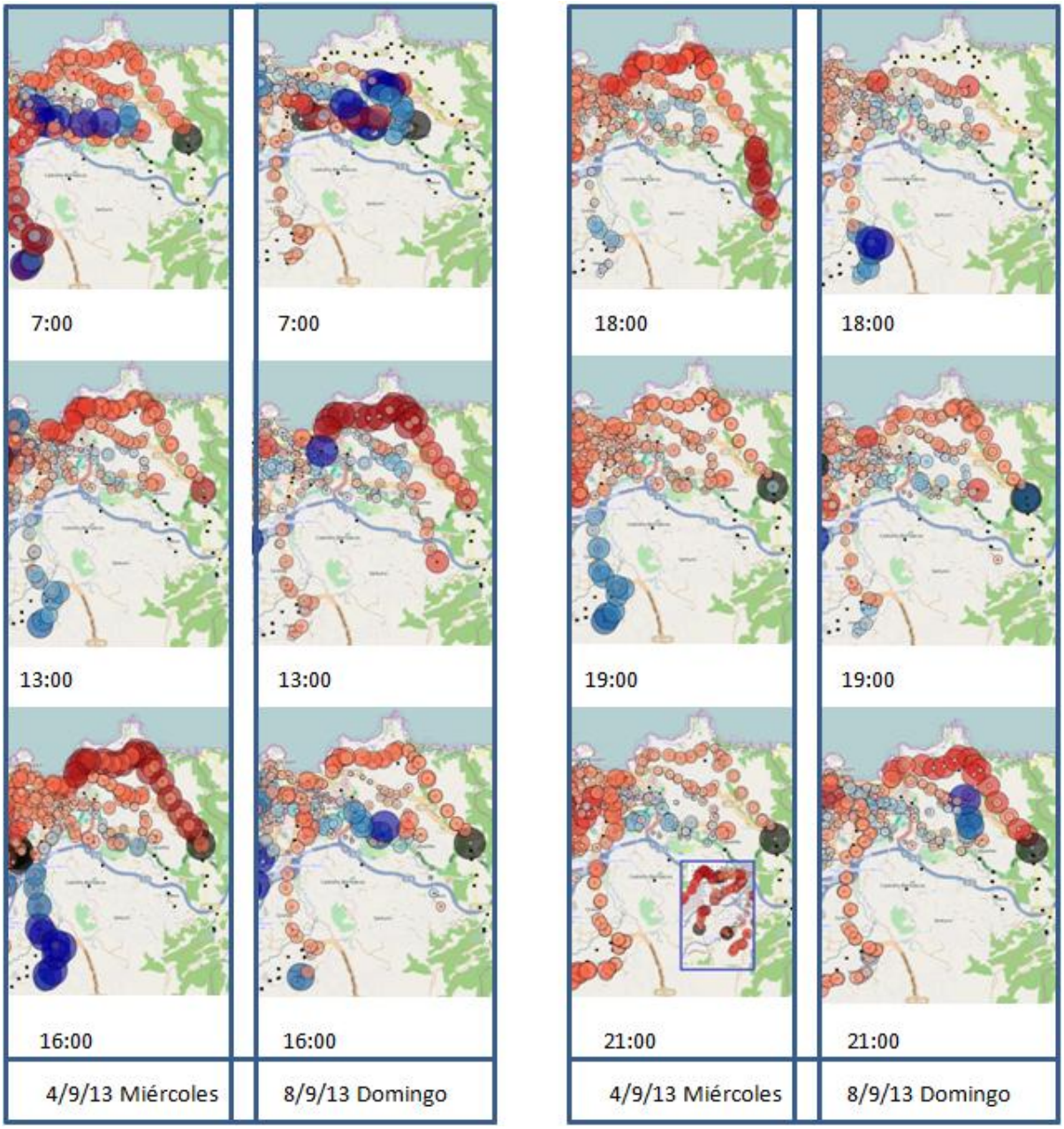

Figura 3.21 Retrasos en las marquesinas de Gijón en la zona de estudio en los días 4-8/9/13

3.4.3 Análisis de los perfiles de los retrasos de las líneas en relación al "rastro de Gijón"

Se ha descrito de manera muy simplificada la evolución de los retrasos a lo largo de los dos días, pero para ver la influencia que tiene el rastro de Gijón en los retrasos de los autobuses, se va proceder al análisis de los perfiles de las distintas líneas que pasan por la zona.

El rastro de Gijón se celebra todos los domingos de 7:00 a 15:00 horas, siendo la hora punta las 12:00 horas. El rastro dispone de su propia marquesina “Rastro”, ya que los domingos a las horas en las que está abierto el mercadillo, el ayuntamiento de Gijón dispone de unos trayectos especiales de las líneas 1 y 4, que paran en esta marquesina. Está situado al lado del estadio de futbol “El Molinón”.

Para hacer este análisis se estudiaron la Línea 1, la Línea 4 y la Línea 10, tanto el día de rastro (8 de septiembre), como el día que no había rastro (4 de septiembre).

En la Figura 3.22, se pueden observar los perfiles de 4 marquesinas de la Línea 1 durante el miércoles 4(rojo) y el domingo 8 (verde). En dicha figura se ven 4 imágenes que están numeradas, y cuya numeración está ordenada según el orden en el que pasa el autobús de cada línea. En el perfil en rojo, no se aprecia ningún tipo de patrón, ni de incidencia cuando el autobús atraviesa la posición de la marquesina “Rastro”. Esto no es así en el perfil en verde, el rastro afecta ligeramente al retraso que se produce a las 12:00 horas (hora punta) en la marquesina de “Las Mestas” que es la inmediatamente posterior a la marquesina “Rastro”, pero no afecta al resto de retrasos.

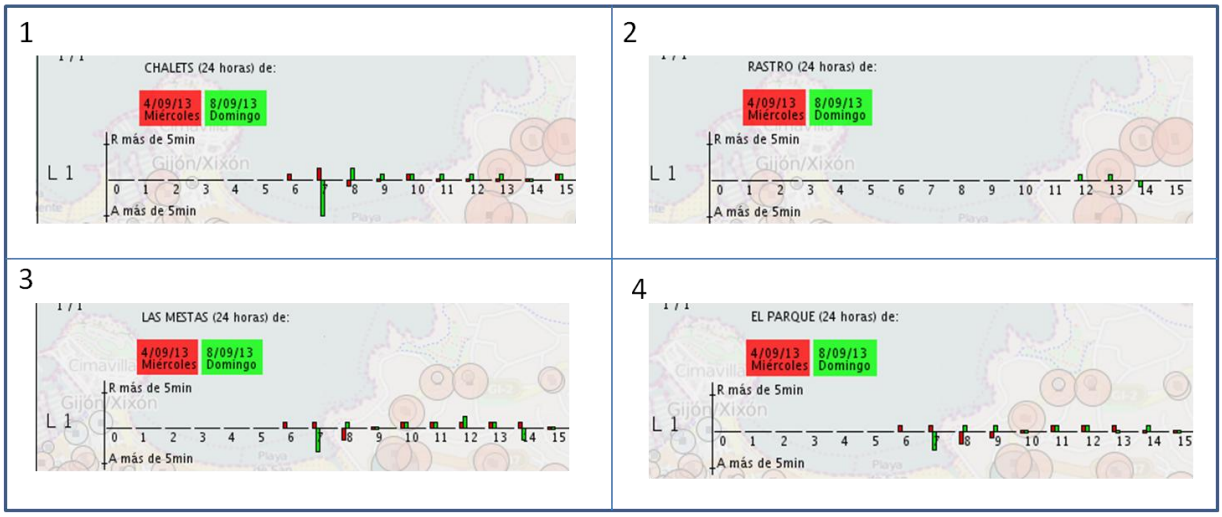

Figura 3.22 Retrasos en las Línea 1en cuatro marquesinas de Gijón en los días 4-8/9/13

En la Figura 3.23, se analizan los perfiles de la Línea 4 el miércoles 4 (rojo) y el domingo 8 (verde), y muestra cuatro imágenes numeradas al igual que en el caso anterior siguiendo el orden de paso del autobús. En este caso, a diferencia de lo que ocurría en la Línea 1, sí que se observa la influencia del rastro en todos los retrasos, siendo la hora punta (las 12:00 horas) la hora en la que aparece el mayor retraso en esta línea.

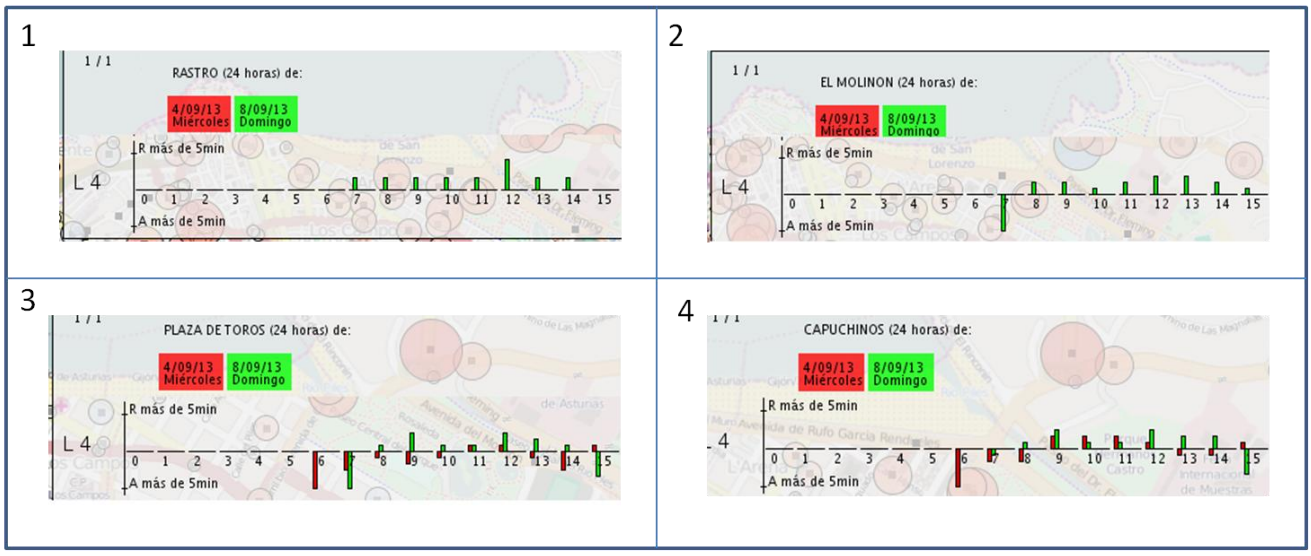

Figura 3.23 Retrasos en las Línea 4 en cuatro marquesinas de Gijón en los días 4-8/9/13

Por último, se hace el mismo análisis que en los casos de las líneas 1 y 4, para la Línea 10 (Figura 3.24). En este caso la línea no pasa por la marquesina "Rastro", así que el análisis se hace en las marquesinas que están más próximas a la zona donde se celebra el mercadillo. En los perfiles se observa que no hay ninguna influencia del rastro en los retrasos de la línea, habiendo más retraso el miércoles que el propio domingo.

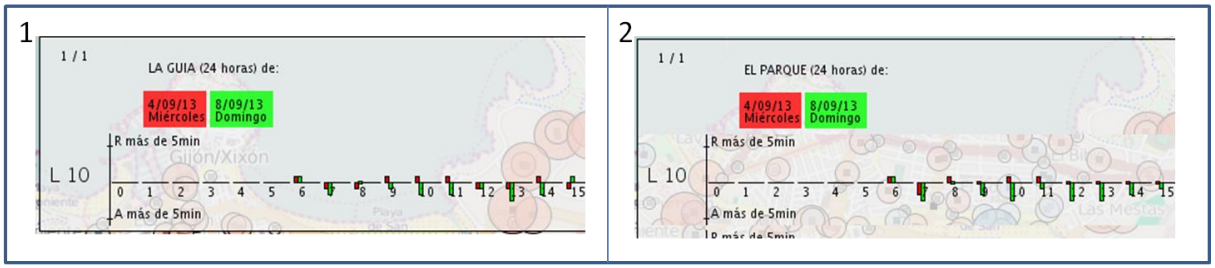

Figura 3.24Retrasos en las Línea 10 en cuatro marquesinas de Gijón en los días 4-8/9/13

3.4.4 Conclusiones en el análisis de la influencia de "El rastro" en las marquesinas de Gijón.

Tras el análisis realizado, se deduce que la celebración de "El rastro", no afecta en absoluto a las líneas que no pasan por la marquesina "Rastro", aunque la línea pase cerca a la zona donde se celebra el citado mercadillo. Si bien, este evento sí que tiene una ligera influencia en las

horas punta sobre las líneas que pasan por la marquesina “Rastro”, pero no parece que tenga demasiada importancia. Donde se puede concluir que sí existe una influencia marcada de la celebración del mercadillo, es en la trayectoria de la Línea 4, esto se debe a que el inicio de esta línea está en la propia marquesina “Rastro”.

4 Conclusiones y líneas futuras

En el presente capítulo se exponen las conclusiones que se obtuvieron con este proyecto, a las aportaciones resultantes del mismo, y las posibles líneas de trabajo futuras que se desprenden del proyecto.

4.1 Conclusiones

Para la elaboración de este proyecto se propone la utilización de datos abiertos (Open data provenientes de las administraciones públicas, para analizar, supervisar y, en consecuencia, obtener conocimientos (Data mining) sobre el propio funcionamiento de dichas administraciones, que pueden ayudarles de manera significativa en la toma de decisiones en cuanto a los servicios que se prestan a los ciudadanos. Por lo tanto, la publicación de datos abiertos se presenta como una prioridad para las administraciones públicas, ya que su utilización ayuda a que los propios ciudadanos puedan elaborar servicios útiles para ellos mismos y, que a su vez, ayuden en la mejora de los servicios que les prestan las administraciones.

En el presente proyecto, las herramientas de visualización se han utilizado como técnicas que ayudan a comprender el comportamiento del sistema de transporte público de Gijón. En especial se ha incidido en la gran capacidad que tienen las técnicas de visualización de datos, como la Geovisualización, para la extracción de conocimiento (Data Mining) y la búsqueda de patrones socio-culturales, en particular si se combinan con herramientas interactivas que faciliten que usuarios con conocimientos previos de las zonas de estudio, puedan extraer conclusiones sobre esos datos.

Las herramientas de interactividad en las visualizaciones de datos se presentan como un mecanismo especialmente poderoso a la hora de mostrar los datos a sus usuarios, ya que permiten que éstos designen y varíen las especificaciones concretas que les interese resaltar del caso de estudio.

La interactividad permite la selección de los datos que se consideren interesantes para cada usuario, sin descartar la posibilidad de estudiar otros datos posteriormente, que puedan considerase interesantes, gracias al conocimiento adquirido con el análisis de los datos.

Se ha podido comprobar, como un análisis meticuloso de los datos es fundamental antes de realizar técnicas de minería de datos, presentándose la Geovisualización como una técnica extremadamente poderosa en la detección de outliers.

4.2 Aportaciones

Las principales aportaciones que, a juicio del autor, ofrece este trabajo son:

- Un enfoque de utilización de los datos abiertos (Open Data) que publican las administraciones para la elaboración de visualizaciones que permiten la supervisión de los servicios prestados.
- La utilización de los datos abiertos para analizar las costumbres socio-culturales de los habitantes de las ciudades, como se explicó en el ejemplo práctico en que, el análisis a través de la herramienta Processing, mostró la influencia de "el rastro" en los retrasos que se producen en la zona este de la ciudad, deduciéndose que sólo son significativos en la Linea 4, que empieza su recorrido en la marquesina situada en el propio mercadillo.
- Se ha desarrollado una aplicación interactiva utilizando Processing para el análisis y supervisión de los retrasos o adelantos que sufren los autobuses del sistema de transporte urbano de Gijón.

4.3 Líneas de futuro trabajo

A partir de este proyecto se abren varias líneas futuras de trabajo sobre la utilización de datos abiertos en el análisis de los servicios de las administraciones públicas y de los hábitos socio-culturales de los ciudadanos:

- Añadir a la aplicación de Processing, proyectar los retrasos de las líneas sobre el plano de Gijón, individualmente.
- Desarrollar una aplicación en D3js, que permita el análisis de las velocidades de los autobuses en la ciudad de Gijón, gracias a los datos que se han utilizado para este proyecto o con los referentes a los autobuses que se publican en la misma página Web. Combinar el análisis de las velocidades con los retrasos que se producen en las marquesinas, desarrollando en la misma aplicación en D3js las dos opciones de análisis.
- Realizar una aplicación que permita explotar otro conjunto de datos del portal datos.gijon.es (como por ejemplo la ocupación de las plazas de parking de la ciudad y del parque de bicicletas), pudiendo extraer información y combinándola con está hacer un análisis más amplio de los hábitos de los ciudadanos.

- Desarrollar una aplicación que permita visualizar los retrasos que se producen en Gijón en tiempo real y que ayude a los usuarios a decidir que líneas de autobús son las más adecuadas para desplazarse a un determinado punto desde su ubicación actual.

5 Cronograma

En este capítulo se describen las fases del proyecto y la duración de cada una de ellas.

FASE 1. Estudios preliminares. En esta fase se realizó un estudio detallado de los datos y la documentación disponible en los servicios web de datos.gijon.es, incluyendo estructura y campo disponible de los datos, formatos de las llamadas a los servicios, etc. Asimismo, se hizo un estudio del estado de la técnica en cuanto a la caracterización de los sistemas de transporte urbano.

FASE 2. Extracción de datos. Esta fase incluye la realización de los scripts que permitieron extraer los datos que se consideraron relevantes del portal web datos.gijon.es y el diseño de las bases de datos en las que estos se almacenaron.

FASE 3. Diseño de algoritmos e interfaz. En esta fase se realizó el diseño de los algoritmos de cálculo y preparación de datos para su posterior explotación, así como el filtrado inicial de los datos. Además se hizo el diseño de la interfaz que se pretendía implementar en Processing, y la elección de las librerías más adecuadas para este fin.

FASE 4. Programación. Esta fase está divida en dos partes: programación de los algoritmos de cálculo y programación de interfaz.

FASE 4.1 Programación de los algoritmos de cálculo. En esta fase se programaron los algoritmos de cálculo de los datos necesarios para la posterior visualización. Se realizaron en Python, filtrando los datos que se consideraban in adecuados para el análisis, y creando los scripts que filtraban los datos que provocaban outliers.

FASE 4.2 Programación Interfaz (Processing). En esta fase, se incluye la programación de toda la aplicación en Processing, incluyendo la versión de Multitouch.

FASE 5. Puesta a punto y revisión. Se corrigieron todos los bugs que aparecieron, se depuraron algunas partes de la aplicación en Processing y se añadieron algunas funcionalidades.

FASE 6. Manual de usuario. En esta fase, se realizó el manual de usuario.

FASE 7. Memoria. Las últimas semanas se utilizaron para desarrollar la memoria del proyecto.

	1ª Semana	2ª Semana	3ª Semana	4ª Semana	5ª Semana	6ª Semana	7ª Semana	8ª Semana	9ª Semana	10ª Semana	11ª Semana	12ª Semana	13ª Semana
Horas	25	50	75	100	125	150	175	200	225	250	275	300	325
Fase 1	Estudios perliminares												
Fase 2		Extracción de los datos											
Fase 3				Diseño de algoritmos e Interfaz									
Fase 4					Programación								
Fase 4.1					Programación cálculo								
Fase 4.2						Programación Interfaz (Processing)							
Fase 5								Puesta a punto y revisión					
Fase 6										Manual de Usuario			
Fase 7												Memoria	

6 Presupuesto

En este capítulo se detallan todos los gastos que conlleva la ejecución del proyecto obteniendo como resultado su coste total.

El desglose del presupuesto se ha realizado separando los costes en dos partes, los gastos correspondientes a materiales necesarios y los referentes a la mano de obra que necesaria para llevar a cabo el proyecto.

Con motivo de cubrir una serie de gastos de carácter no específicos, se incluye en el coste total un 15% de costes generales.

PRESUPUESTO (VISTUG)					
Nº	NÚMERO DE UNIDADES	DESIGNACIÓN DE LAS OBRAS	PRECIOS POR UNIDAD	IMPORTES	
				PARCIALES	TOTALES
				Euros	Euros
		Materiales			
1	1	HP ProLiant G7 Micro Server AMDTurion II N54L/4GB	249	249	
2	1	PcCom Experience i5-3350P/8GB/1Tb/GTX 660 OC	659	659	
3	2	Genius SlimStar 8000M Teclado + Ratón Wireless	14,75	29,5	
4	1	BenQ GL2450HT 24''	175	175	
5	1	Packard Bell Viseo 193 DXb 18.5'' LED	77,95	77,95	
6	1	Ovislink EVO W322AR Router	29	29	
		TOTAL PARCIAL		1219,45	1219,45
		Mano de obra			
7	25	Horas de estudios previos	50	1250	
8	200	Horas de diseño y programación	50	10000	
9	25	Puesta a punto y revisión	50	1250	
10	30	Horas elaboración manual de usuario	50	1500	
		TOTAL PARCIAL		14969,45	15219,45
		Subtotal			16438,9
		Gastos generales (15%)			2465,835
		Total sin I.V.A.			18904,735
		I.V.A. (16%)			3024,7576
		TOTAL PRESUPUESTO			21929,4926

El coste total del proyecto asciende a VEINTIUNMIL NOVECIENTOS TREINTA euros.

7 Bibliografía

1. Aigner W, Miksch S, Müller W, Schumann H, Tominski C. *Visual Methods for Analyzing Time-Oriented Data.* IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 14, NO. 1, JANUARY/FEBRUARY 2008.
2. Andrienko G, Andrienko N, Dykes J, Fabrikant S, Wachowicz M. *Geovisualization of dynamics, movement and change: key issues and developing approaches in visualization research.* Information Visualization. 2008
3. Anscombe F. *Graphs in Statistical Analysis*, 1973
4. Anselín L. *Interactive techniques and exploratory spatial data analysis.* 1999
5. Anselin L, Bao S, *Exploratory spatial data analysis*. 1997
6. Arentze T. *Spatial Data Mining, Cluster and Pattern Recognition*. International Encyclopedia of Human Geography, 2009
7. Ben F. *Visualizing Data: Exploring and Explaining Data with the Processing Environment.* O'reilly, 2008
8. Brown, G. and Yule, G. *Discourse Analysis*. Cambridge , UK : Cambridge University Press, 1983.
9. Buja A, McDonald J A, Michalak J, Stuetzle W. *Interactive data visualisation using focusing and linking*. In Nielson G M, Rosenblum L (eds) Proceedings of Visualisation 91. Los Alamitos, IEEE Computer Society Press: 155–62, 1991.
10. Cook D, Majure J, Symanzik J, Cressie N. *Dynamic graphics in a GIS: a platform for analysing and exploring multivariate spatial data*. Computational Statistics 11: 467–80, 1996.
11. Fayyad U, Piatetsky-Shapiro, Gregory; Padhraic S (1996). *From Data Mining to Knowledge Discovery in Databases*. 2008.
12. Kim C. *Spatial Data Mining, Geovisualization.* Rob Kitchin and Nigel Thrift, 2009
13. Lathrop D, Ruma L. *Open Government:Collaboration, Transparency, and Participation in Practice,* O'reilly Media, 2010.
14. Mckinsey and Company, *Big data: The next frontier for innovation, competition, and productivity,* [En linea] http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation, 2008

15. Martino M, Bertone A, Albertoni R, Hauska H, Demsar U, Dunkars M. *Technical Report of Data Mining*, Information Visualisation for Site Planning, WP No2: Technology Analysis, D2.2, 28.2.2002
16. Miksch S, Kosara R, Shahar Y, PD Johnson. *AsbruView: Visualization of Time-Oriented, Skeletal Plans.* AIPS, 1998
17. Muzner T. <*Information Visualization: Principles, Methods and Practice*>
18. Thrift N, Kitchin R . *Understanding the Changing Planet: Strategic Directions for the Geographical Sciences* . National Research Council. 2010
19. Iliinsky N, Steele J. *Designing Data Visualization.* O`reilly, 2011.
20. Rafaeli S, Sudweeks *F. Networked Interactivity*. Journal of Computer-Mediated Communication, 1997.
21. Steele J, *Why data visualization matters.*[En linea] http://strata.oreilly.com/2012/02/why-data-visualization-matters.html, February, 2012
22. Tufte, E. *The Visual Display of Quantitative Information.* Graphics Press, 1983
23. Wijk . *The value of visualization*". 2005