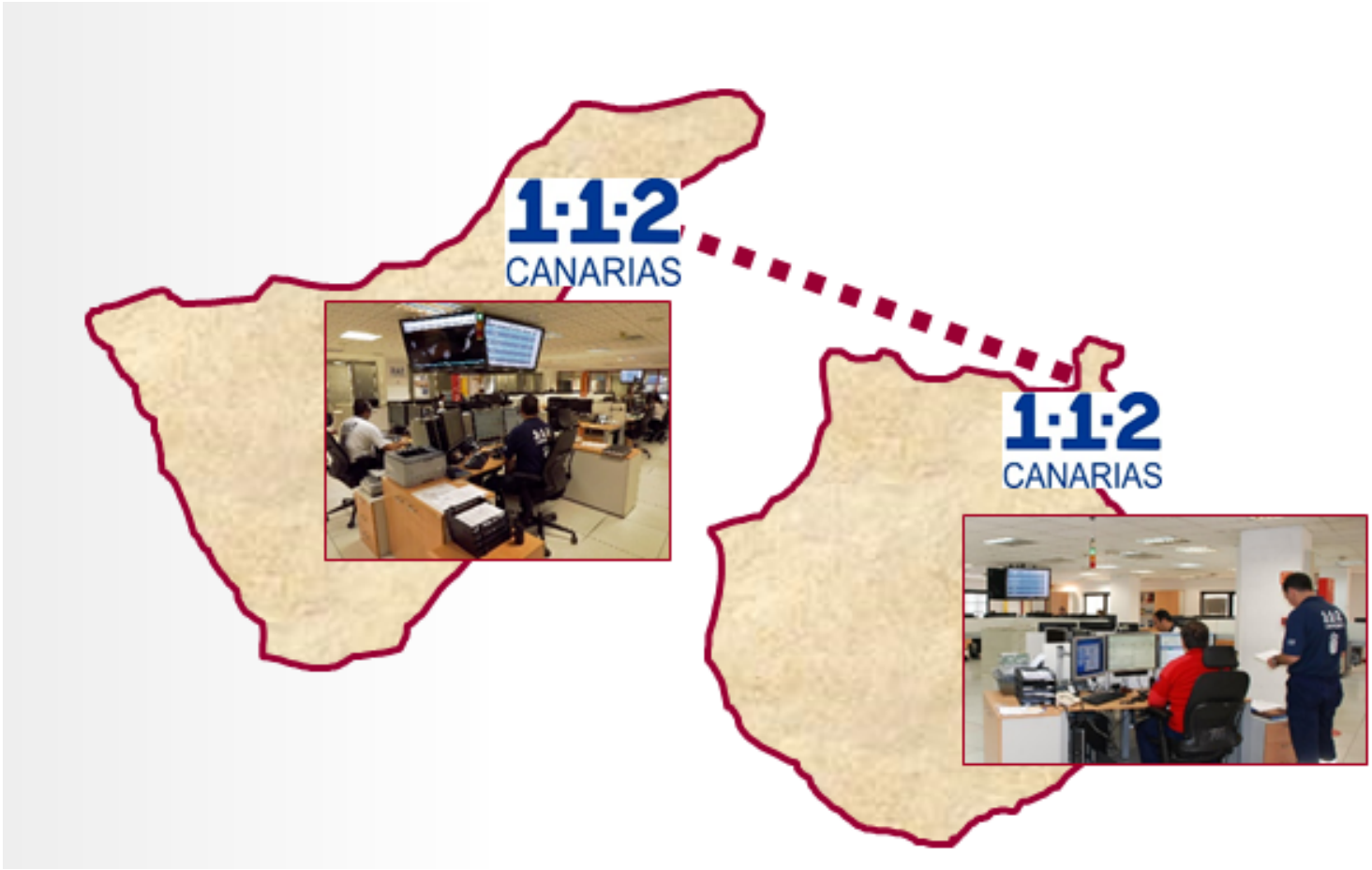
Estudio de las alertas del servicio de emergencias 112 del Gobierno de Canarias utilizando técnicas de ciencia de los datos

Ponentes: Carlos Rosa Remedios, Carlos Pérez González

El CECOES 1-1-2

- El CECOES 1-1-2 es el centro coordinador de emergencias y seguridad
  - Dos salas operativas: Las Palmas de Gran Canaria, Santa Cruz de Tenerife
  - Capacidad tecnológica para actuar como un solo centro



Sectores

Operación de Demandas

Coordinación Multisectorial

Sanidad

Seguridad

Extinción, Salvamento y Rescate

Atención a la mujer agredida

Integra otros servicios de emergencia

- Lograr la cooperación efectiva entre los diferentes servicios para ajustarse a la demanda ciudadana

Proceso de llamada

- Cuando un ciudadano llama al 1-1-2, al otro lado de la línea le responde un operador de demanda especializado en atender las peticiones de ayuda de los ciudadanos ante cualquier tipo de urgencia o emergencia

Mientras se toman los datos del incidente, la ayuda ya va en camino. Gracias a unos protocolos automáticos, se clasifica la llamada y se envía el recurso más adecuado para cada situación, ya sea sanitario, de seguridad, extinción de incendios, salvamento y rescate.

Proceso de llamada

- Cuando un ciudadano llama al 1-1-2, al otro lado de la línea le responde un operador de demanda especializado en atender las peticiones de ayuda de los ciudadanos ante cualquier tipo de urgencia o emergencia

Apoyado en un programa informatizado, el operador de demanda realiza un cuestionario que permite clasificar la alerta y asignar el recurso o recursos necesarios para resolver el incidente.

Objetivos planteados

- Desde el año 2005 se dispone de un histórico con las alertas o incidencias registradas en el sistema de gestión del 112 en ambas provincias. Como se puede observar, hay un número bastanto alto de incidencias cada año.

AÑO	Las Palmas	S/C. Tenerife
2005	345,035	313,875
2006	355,433	329,674
2007	355,070	330,162
....	.....	.....
2013	348,352	331,204
2014	359,035	346,127

- Fijémonos que, aunque el volumen de datos no es tampoco muy alto, lo que representa el auténtico problema de big data es el tipo de técnicas que queremos aplicar: comparación de incidencias entre sí (algoritmos de k-medias), técnicas de predicción en tiempo real (minería de datos), etc...

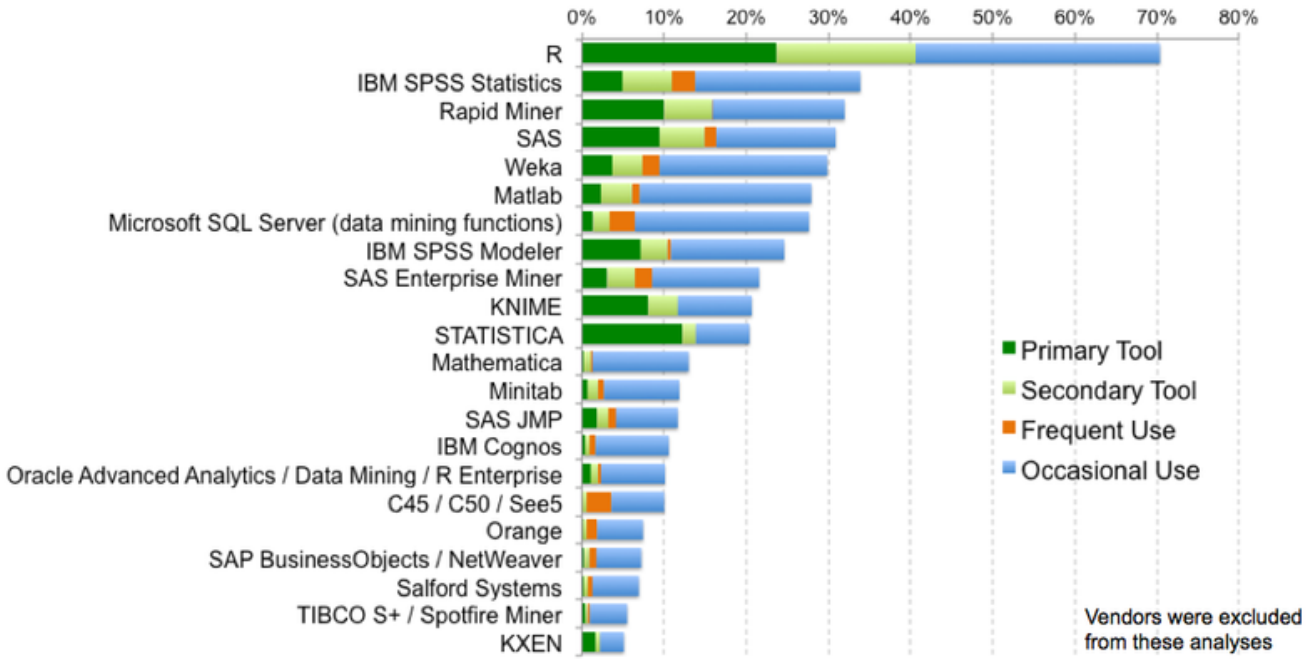
Conjunto de datos

- En la información relativa a las incidencias atendidas por el servicio se recogen variables como:
  - El sexo y la edad de la persona que es objeto de la alerta
  - Fecha y hora del inicio y el final de la alerta
  - El municipio donde tiene lugar
  - Los recursos que se han dispuesto para su resolución (policía, bomberos, etc..)
- Se han planteado varias fases de trabajo con los datos, de acuerdo a la siguiente planificación ('data science')
  - Adquisición de datos (R es el software que se ha considerado para este proyecto)
  - Preparación de los datos (depuración, normalización, etc. . .)
  - Análisis y presentación (herramientas dashboard tablas y gráficos estadísticos, aplicación de modelos estadísticos, etc..)

Adquisición de datos: uso de software R

- R es el lenguaje de programación que se ha convertido, en la última decada, en la herramienta más importante para estadística computacional, visualización y ciencia de los datos, tanto en el mundo académico como en la empresa y la industria.

- En todo el mundo, R se utiliza para resolver un gran número de problemas en campos que van desde la biología computacional hasta el marketing cuantitativo. De hecho, es una herramienta esencial para compañías como Google, Facebook y LinkedIn.



Adquisición de datos: importación

- Los datos originales se encuentran en formato plano (*.txt), de tal forma que para cada uno de los 10 años, se tiene un archivo de un tamaño aproximado de 200 Mb. Cada uno de ellos tarda una media de 15 segundos en ser cargado

```
start <- proc.time()
datos_2005 <- read.csv("112_20XX.txt", sep=";", dec=",", stringsAsFactors=FALSE)
end<-proc.time() - start
```

```
## [1] "Fichero incidencias 2005: 14.15 segundos"
```

```
## [1] "Fichero incidencias 2006: 18.9 segundos"
```

```
## [1] "........."
```

```
## [1] "Fichero incidencias 2014: 15.9 segundos"
```

Adquisición de datos: importación

- El tiempo de carga total de los datos es muy elevado ($>$ 1min.) y el uso de memoria RAM se incrementa drásticamente. Por otro lado, al realizar ciertos análisis sobre los datos el tiempo de ejecución también es alto.

```
## [1] "Uso RAM fichero incidencias 2005: 156.7 Mb"
```

```
## [1] "Uso RAM fichero incidencias 2006: 178.9 Mb"
```

```
## [1] "........."
```

Adquisición de datos: uso de librerías de bigdata

- Para resolver el problema de la carga y tratamiento de datos utilizamos una librería de R ('ffbase') que nos permite trabajar con objetos y estructuras de datos masivas (por ejemplo, datos de secuenciación genómica).

- Dicha librería utiliza métodos de acceso rápido a los datos (mediante índices y paginación en memoria) y nos permite crear una estructura de datos óptima para trabajar con ella.

```
library(ffbase)
start <- proc.time()
datos_global <- load.ffdf(dir="data_ffdf")
end<-proc.time() - start
```

```
## [1] "Estructura ffdf: 0.13 segundos"
```

- Se aprecia una reducción significativa en el tiempo de carga de todo el conjunto de datos. Además, con respecto al uso de la RAM

```
## [1] "Uso RAM de estructura ffdf: 5.3 Mb"
```

Preparación de datos: librerías de resúmenes estadísticos

- Los tiempos de ejecución de los análisis se optimizan de forma similar haciendo uso de otra librería de R llamada 'dplyr'.

```
table.dia_by_muni_by_sexo<-112_ffdf %>%
                          filter(anio="2009") %>%
                          group_by(dia,sexo,municipio) %>%
                          summarize(n=n()) %>%
                          mutate(prop.catch=n/sum(n)) %>%
                          arrange(desc(prop.catch))
```

```
##     row           MES    SEXO MUNICIPIO TOTAL
##       1     01-ENERO HOMBRE     ADEJE     1
##       2     01-ENERO  MUJER     ADEJE     1
##       3     01-ENERO HOMBRE    AGAETE     1
##       4     01-ENERO  MUJER    AGAETE     1
##                 ...    ...       ...   NaN
##   69349 31-DICIEMBRE HOMBRE     YAIZA     1
##   69350 31-DICIEMBRE  MUJER     YAIZA     1
```

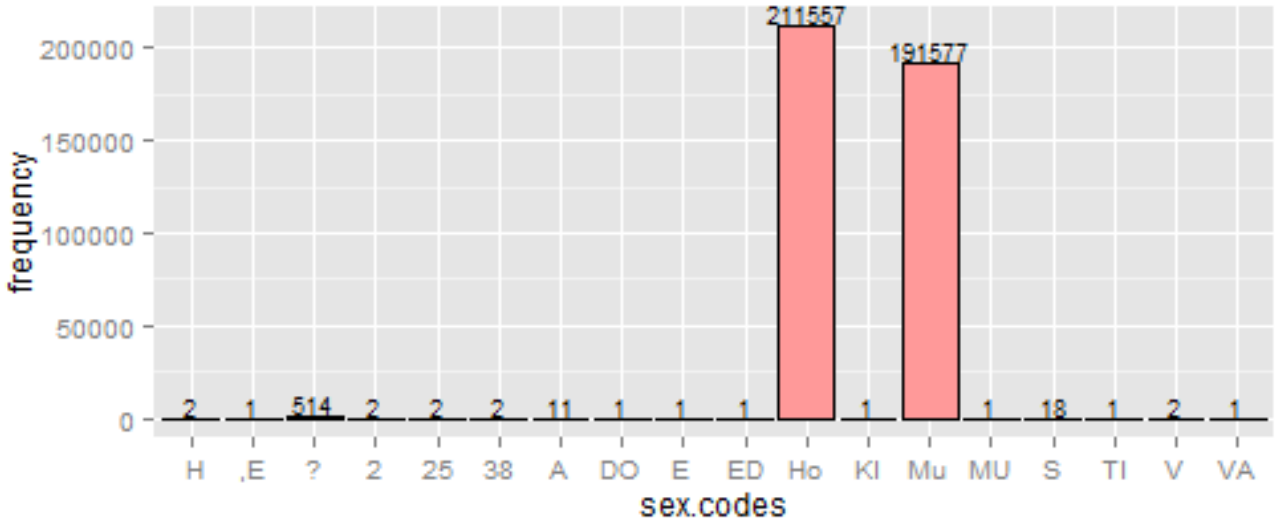
- La comparación de tiempos para obtener los resultados

```
## [1] "Consulta estad. data.frame: 37.6 segundos"
```

```
## [1] "Consulta estad. WITH ffdf y dplyr: 0.2 segundos"
```

Preparación de los datos: Depuración (I)

- Veamos algunas operaciones de depuración que, a través de R, resultan muy sencillas de llevar a cabo.
- Se necesitan depurar variables como el sexo, la unidad en la que se mide la edad (años, meses, semanas, etc..).

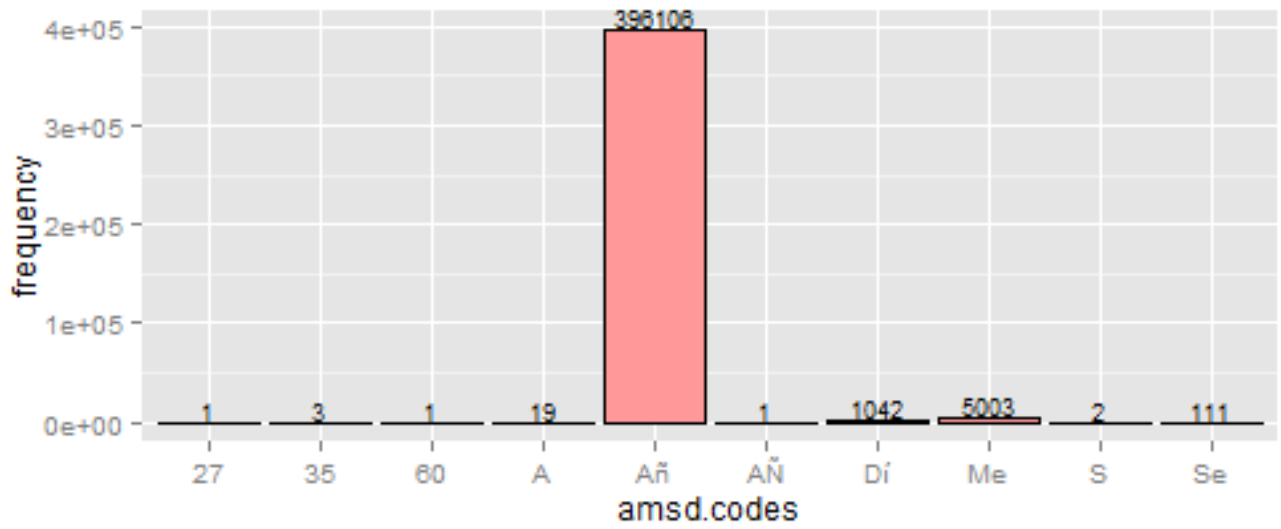


- Observamos que, en el tratamiento de las incidencias en un sólo año, la frecuencia de los valores atípicos es tan baja que se puede considerar que no representan un conjunto significativo para el análisis posterior.

```
SEXO[ SEXO %not in% c('Mu','Ho')]<-""
```

Preparación de los datos: Depuración (II)

- El tratamiento ha sido muy similar en el caso de otras variables. Por ejemplo, la variable edad se recoge considerando la unidad de medida (edad en días, semanas o meses para recien nacidos y niños y edad en años)



- En este caso, también se descartan unidades atípicas en la medida de la edad (se mantienen los años, meses, semanas y días).

Preparación de los datos: Normalización (I)

- La normalización también puede resultar muy sencilla a través de R. Se trata de eliminar posibles inconsistencias en los códigos de las variables.
- Por ejemplo, las variables de fecha y hora se normalizan siguiendo formato ISO

```
FECHA.INICIO<-
  strptime( paste(as.Date(DIA.INICIO,format= "%d/%m/%Y"), HORA.INICIO),
            format= "%Y-%m-%d %H:%M:%S", tz="GMT")

FECHA.FINALIZA<-
  strptime( paste(as.Date(DIA.FINALIZA,format= "%d/%m/%Y"), HORA.FINALIZA),
            format= "%Y-%m-%d %H:%M:%S", tz="GMT")
```

- De este modo, se pueden obtener variables derivadas

```
DURALERTA<-as.numeric(difftime(FECHA.FINALIZA,FECHA.INICIO,units="hours"))
```

Preparación de los datos: Normalización (II)

- En algunos casos, también se lleva a cabo una re-codificación de ciertas variables, como por ejemplo el nombre del municipio

```
## [1] "Adeje"                    "Arucas"
## [3] "Icod de los Vinos"        "Mar"
## [5] "San Bartolom? Tirajana" "Silos Los"
## [7] "Vallehermoso"
```

- Se trata de añadir el código de muncipio, con el objetivo de vincular información de referencia geográfica a las incidencias.

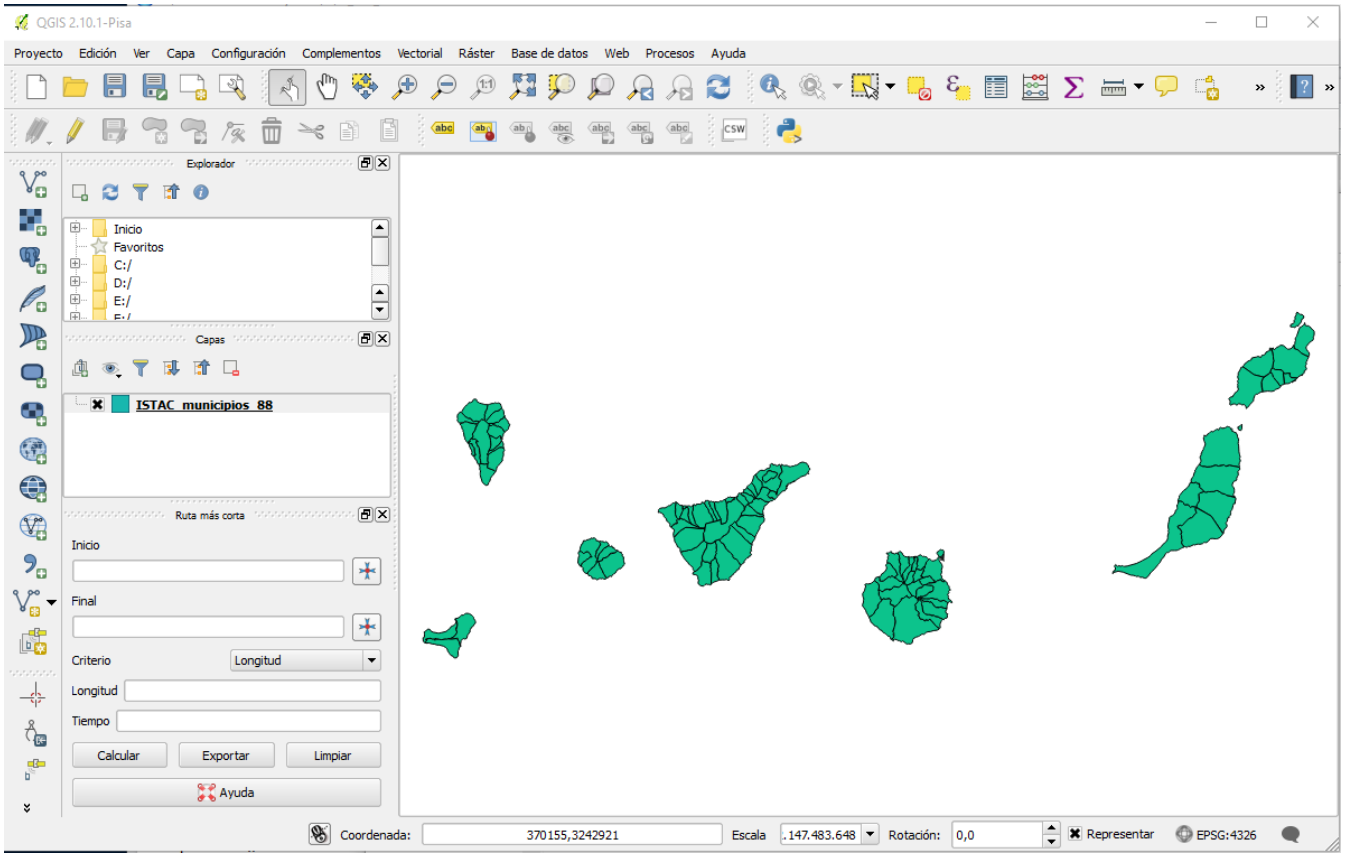
- El problema es asociar un código a una denominación de municipio que no es la oficial

```
#Custom R function
compare.linkage(Nombre.Municipio, Nombre.Oficial.INE)
```

```
## [1] "Operación de comparación: 0.1 segundos"
```

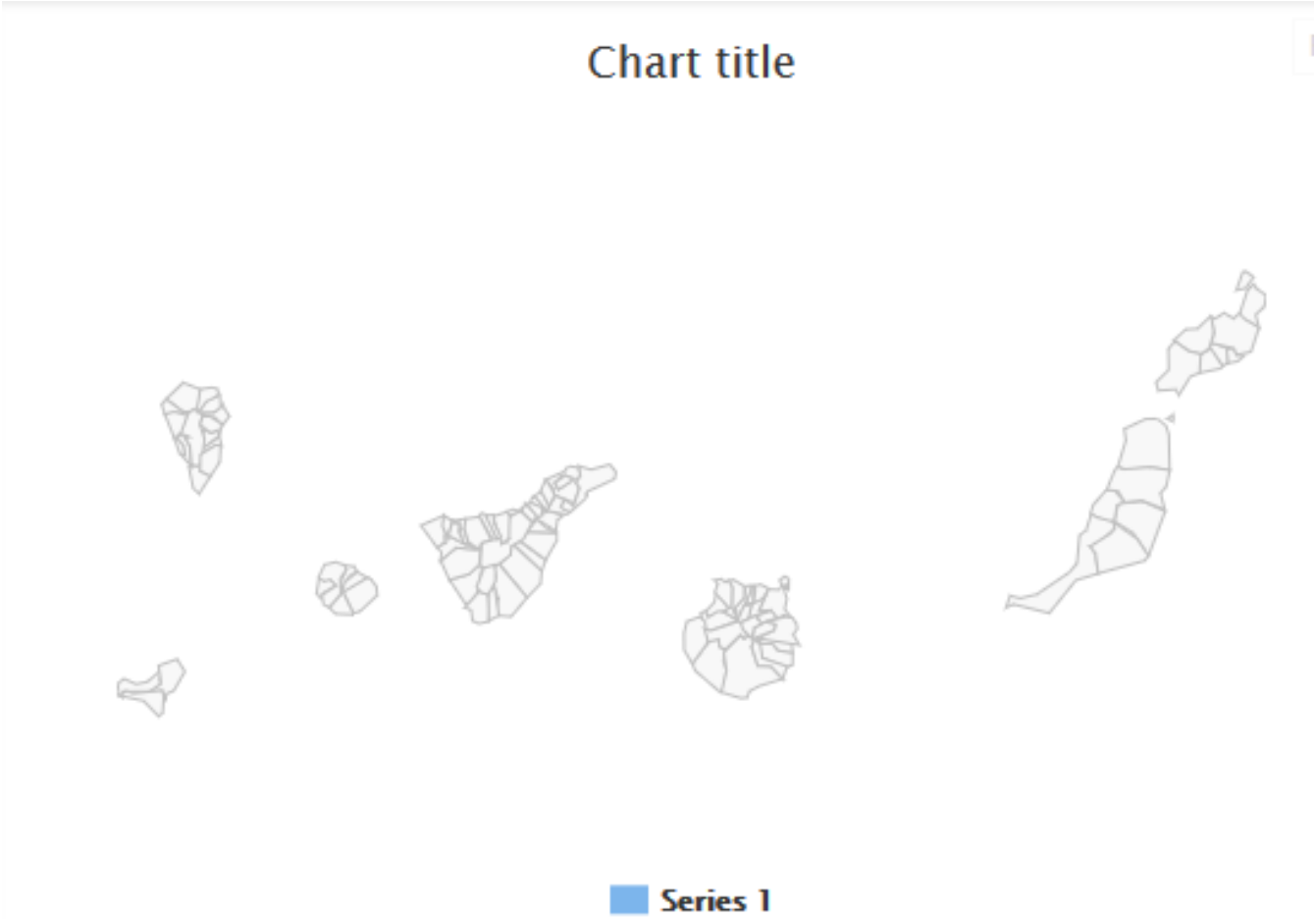
Preparación de los datos: Creando un mapa base (I)

- Utilizando la herramienta QGis se elabora una mapa base de Canarias para utilizarlo en las representaciones gráficas



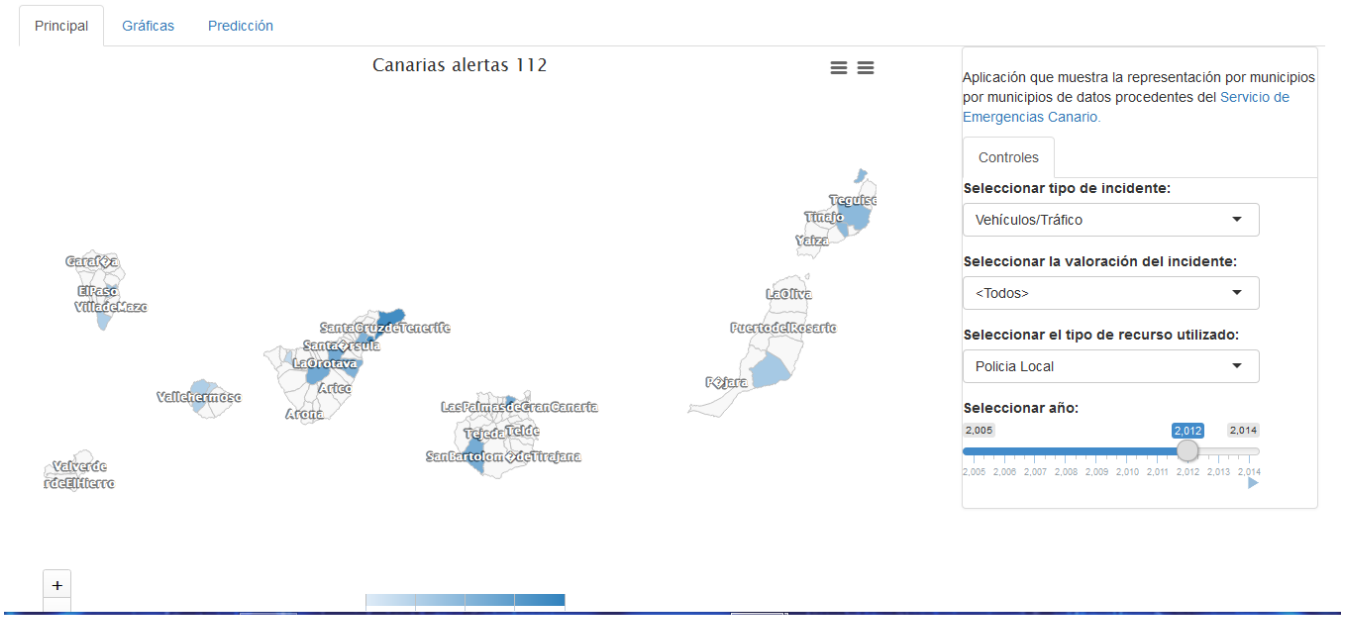
Preparación de los datos: Creando un mapa base (II)

- El mapa se genera en formato GeoJSON, que es un formato muy ligero para su utilización en aplicaciones web en combinación con librerías javascript como Highmaps.



Análisis y presentación: Aplicación web

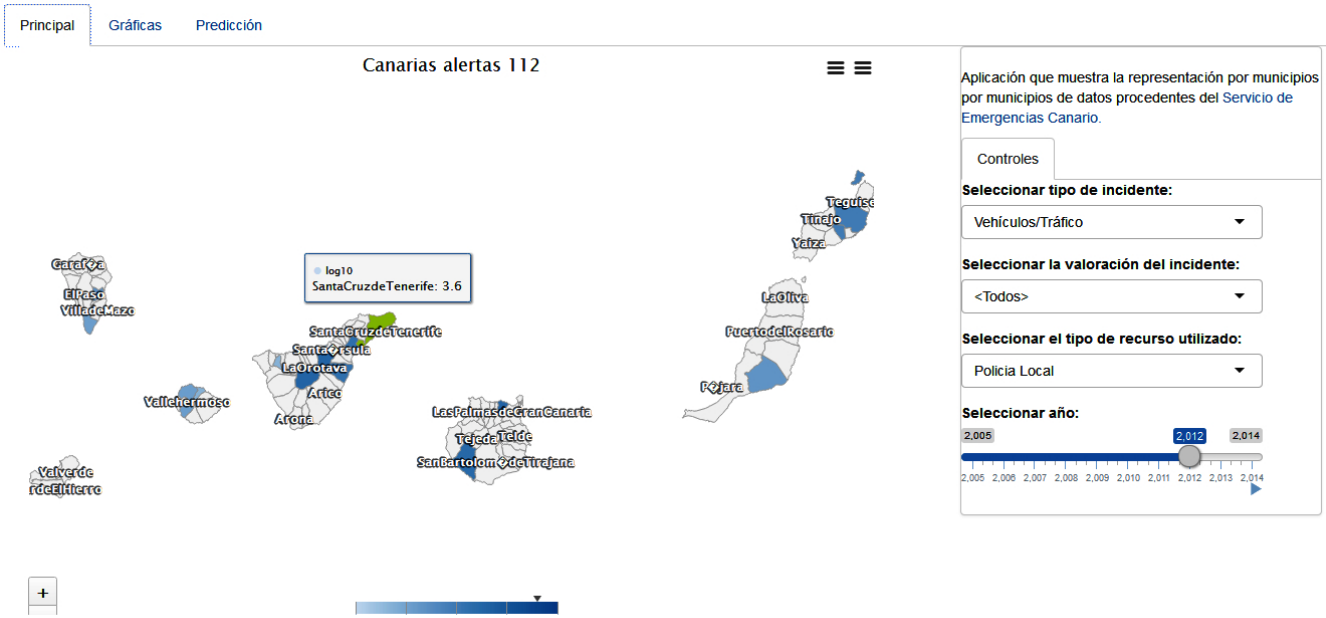
- En esta fase del trabajo, se ha creado una aplicación web para el análisis y consulta de los datos utilizando una librería de R+Rstudio conocida como Shiny



- Como se puede observar, la aplicación permite ofrecer un cuadro de mando donde el usuario realiza una selección de valores en los que filtrar su consulta y obtiene una distribución geográfica en municipios de los incidentes que corresponden a la misma.

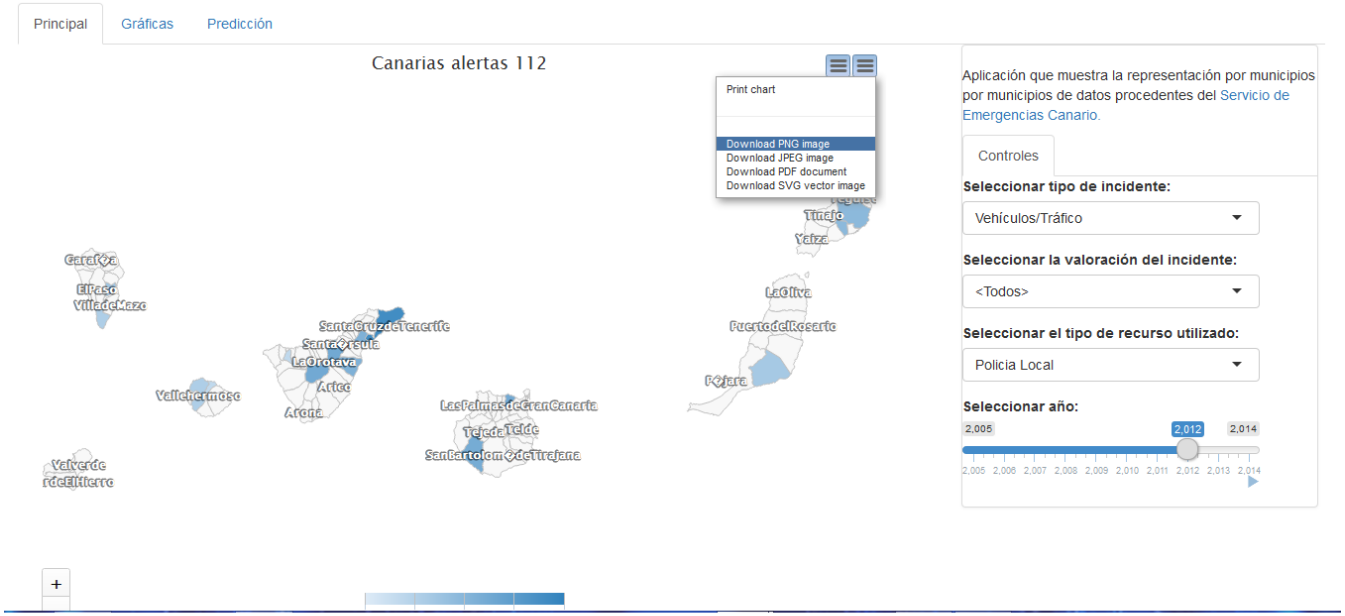
Análisis y presentación: Aplicación web

- En particular, se puede seleccionar un municipio y ver los detalles del número de incidentes.



Análisis y presentación: Aplicación web

- La aplicación tiene aspectos muy interesantes, como el que permite exportar el mapa resultante en un archivo gráfico para su inclusión en informes.



Análisis y presentación: Aplicación web

- También hemos generado un apartado para el análisis temporal de los incidentes, lo que permite realizar un estudio detallado de los datos

- En este caso, se puede seleccionar una isla y/o municipio y obtener una representación del número de incidentes registrados a lo largo del tiempo en función, por ejemplo, el tipo de incidente más frecuente, los recursos más demandados, las valoraciones más habituales, etc. . .

Análisis y presentación: Aplicación web

- Por ejemplo, en un municipio determinado podemos seleccionar sobre el mismo gráfico diferentes recursos (como ambulancias, bomberos, policía, etc.) y observar la evolución del número de incidentes donde dichos recursos son utilizados.

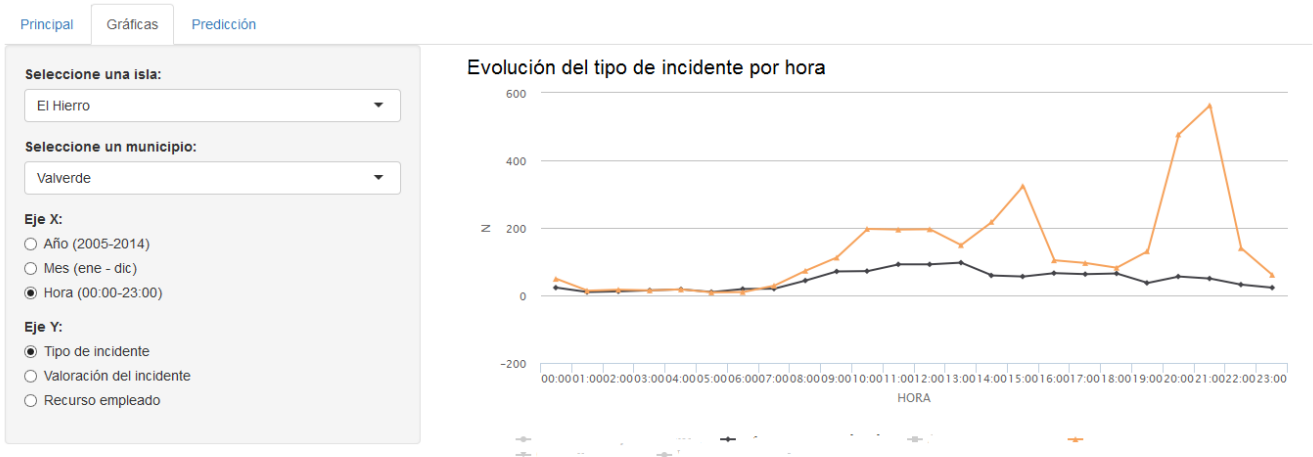
- De nuevo, el gráfico nos permite observar los detalles en un momento determinado.

Análisis y presentación: Aplicación web

- Observemos que en la dimensión de análisis temporal se puede escoger analizar los resultados por años, meses u horas del día.

Análisis y presentación: Aplicación web

- Vemos también un ejemplo del gráfico en el caso del tipo de incidente (desvanecimientos, consulta médica, etc..)



Análisis y presentación: Aplicación web

- Y también un ejemplo del gráfico en el caso de la valoración del incidente (incidente indemne, leve, etc. . .)

Análisis y presentación: Conclusiones (I)

- Hay otros objetivos que podemos analizar, y que tenemos aún en desarrollo:
  - La posibilidad de llevar a cabo análisis mediante modelos de predicción
  - Estudiar la relación con factores soci-económicos (variaciones en los incidentes registrados con respecto a la factores de empleo y/o paro en los municipios)

- Por ejemplo, aplicando un modelo de redes nuronales:

```
#Custom R function
library(nnet)
mod1<-multinom(TRECURSO~SEXO+EDAD+MES+TIDE, data=112.datos)
```

- En este caso, hay algunos problemas para combinar las estructuras de datos 'ffdf' con 'nnet'.

Análisis y presentación: Conclusiones (II)

- Si utilizamos 'nnet' con las estructuras 'data.frame' usuales de R tenemos que el tiempo de entrenamiento del modelo es muy alto

```
## [1] "# weights:  181"

## [1] "initial  value 2091.871880 "

## [1] "iter  10 value 5.870648"

## [1] "iter  20 value 2.942765"

## [1] "........"

## [1] "iter  1000 value 0.273944"

## [1] "stopped after 1000 iterations"

## [1] "Tiempo total: 5.3 minutos"
```

- Creemos que si estudiamos atentamente estas librerías para resolver los problemas encontrados al utilizar estructuras de datos óptimas, se pueden obtener modelos de predicción muy interasantes en tiempos bastante aceptables.

Referencias

- Este trabajo ha sido posible con la colaboración del servicio del 112, quienes tienen un histórico bastante amplio de datos registrados de incidencias, y los expertos que conocen las técnicas y herramientas para analizarlos.

- RStudio Support: web de soporte de RStudio (https://support.rstudio.com).
- RStudio Training: página de formación de RStudio (https://www.rstudio.com/resources/training/).
- Google's R Style Guide: guía de estilo de programación en R según Google (https://google.github.io/styleguide/Rguide.xml).
- Shiny: web de Rstudio shiny (http://shiny.rstudio.com/).