Islas Canarias
Del 15 al 19 de noviembre de 2021

Sistema de georreferenciación para fines estadísticos

Jesús Alberto González Yanes
Instituto Canario de Estadística, ISTAC
jgonyanp@gobiernodecanarias.org

Rafael Betancor Villalba
Instituto Canario de Estadística, ISTAC
rbetvilb@gobiernodecanarias.org

Mª Salomé Hernández García
Ecosistemas Virtuales y Modulares, EVM
salome.hernandez@evm.net

Introducción 3

Objetivos 5

Metodología 5

Estructura del Registro de Portales 5

Normalización 8

Geocodificación 9

Georreferenciación 10

Proceso 11

Resultados 14

Principales referencias bibliográficas 21

Introducción

El artículo 32 de la Ley 1/1991 de Estadística de la Comunidad Autónoma de Canarias [1] indica que se constituirá un banco de datos administrativos para fines estadísticos, que se nutrirá prioritariamente de los ficheros administrativos de la Comunidad Autónoma de Canarias, a cuyo fin todos los departamentos deberán remitir al Instituto Canario de Estadística los ficheros administrativos de los que sean titulares y que sean necesarios para el ejercicio de la función estadística. Los referidos ficheros deberán adecuarse, para su remisión, a los requisitos técnicos que establezca el Instituto. Asimismo, se señala que el banco de datos administrativos para fines estadísticos deberá facilitar la fusión de los ficheros para fines estadísticos.

Este banco de datos se elabora a fin de mejorar la eficiencia de la actividad estadística de interés de la Comunidad Autónoma de Canarias y para su desarrollo durante la ejecución del Plan Estadístico de Canarias 2018-2022 se impulsará el Sistema de Datos Integrados (iDatos). Dicho sistema debe permitir la producción de estadísticas multifuentes mediante la elaboración de datos maestros organizados en directorios y registros que faciliten el enlazamiento de fuentes diversas.

A su vez el Sistema de Datos Integrados debe facilitar la obtención de estadísticas espaciales, mediante la georreferenciación de la información dentro de un Marco Geoestadístico de Canarias alineado con el Global Statistical Geospatial Framework elaborado por el Grupo de Expertos de Integración de Información Estadística y Geoespacial de Naciones Unidas (UN EG-ISGI. United Nations Expert Group on the integration of Statistical and Geospatial Information).

La Figura 1 muestra con detalle la arquitectura del banco de datos del sistema integrado y sus componentes. A continuación se explican las diferentes áreas que lo conforman.

El área de recopilación y preparación de datos (Staging Area) puede considerarse como un punto principal de inyección de datos donde se reciben estos. Los datos de entrada se pueden clasificar en datos estructurados, semiestructurados y no estructurados.

El entorno repositorio se organiza en esquemas-tipo, según la naturaleza de los ficheros que contiene, y estos esquemas se agrupan en tres tipos de niveles según almacenen cartografías, microdatos o macrodatos. Este es un repositorio de base de datos relacional (PostgreSQL) con extensión SIG (PostGIS), que permite tanto el almacenamiento de objetos cartográficos como la realización de procesamiento geográfico. Los esquemas-tipo sobre los que trabaja el Sistema de Datos Integrados son Raw Cartography (RC), Support Cartography (IGS), Geographic Information Reference (IGR), Raw Data (RD) y Master Data (ID).

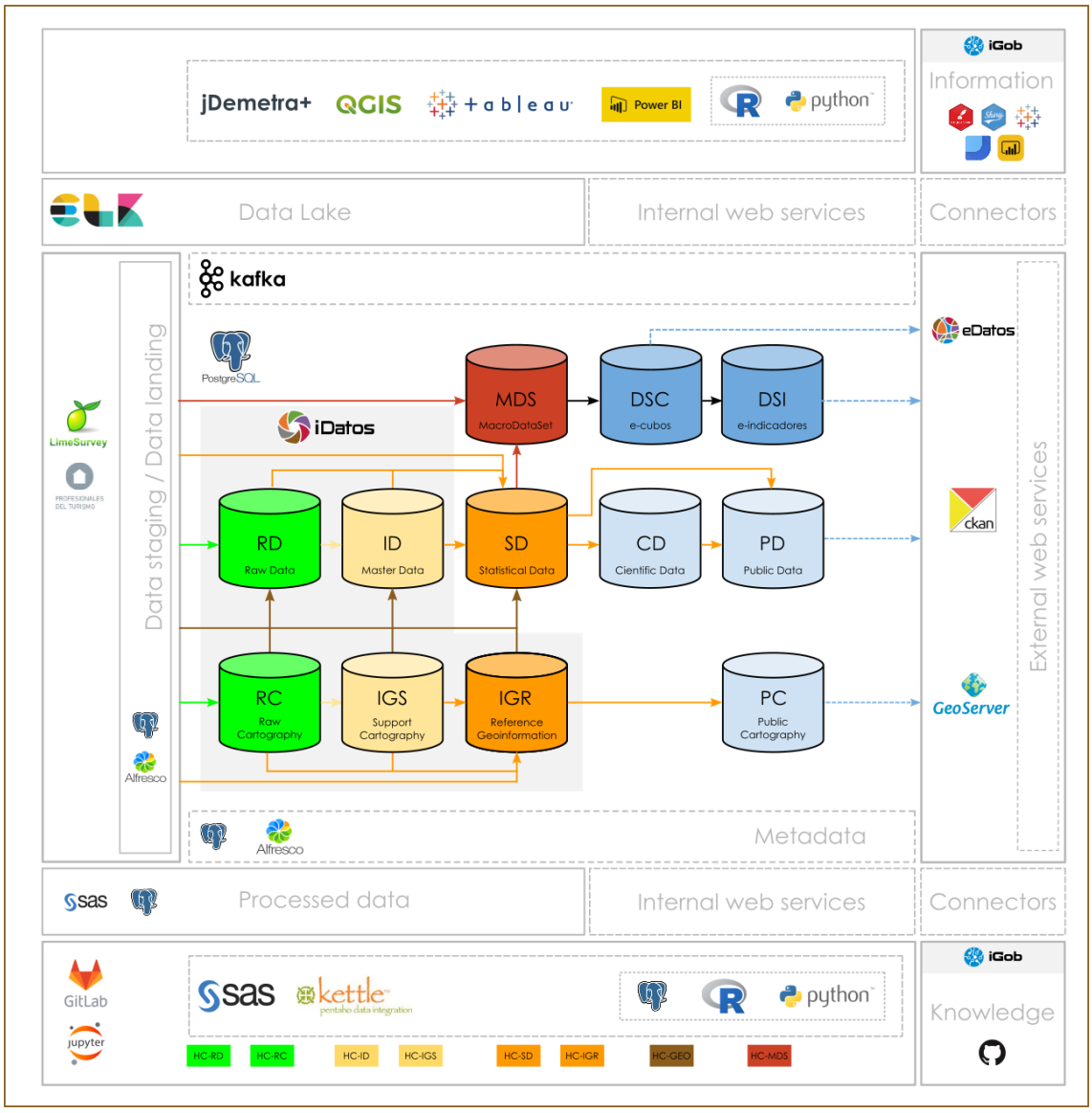

Figura 1: Modelo de arquitectura del banco de datos

El entorno computacional es el encargado del inicio, control, seguimiento y registro de los procesos en la capa de datos del Sistema Integrado.

El entorno analítico contiene los almacenes de datos de forma optimizada para el procesamiento analítico en línea (OLAP), de forma que junto al repositorio de metadatos, contiene toda la información necesaria para las operaciones de análisis.

El entorno de difusión proporciona toda la funcionalidad necesaria para proporcionar los productos finales del entorno analítico a usuarios externos. Dicho entorno debe preservar tanto como sea posible el contenido informativo de los datos recopilados y procesados, garantizando al mismo tiempo que no se pueda volver a identificar a personas concretas, lo que se conoce como el problema del control de divulgación estadísticas (SDC).

La implementación de estos entornos dentro del Sistema de Datos Integrados (iDatos) facilita la fusión de ficheros con fines estadísticos. Para ello, los esquemas de microdatos dentro del Banco de Datos de la Infraestructura de Datos y Metadatos Estadísticos

organiza los distintos registros del sistema, tales como el de población, empresas, portales o viviendas.

El Registro de Portales almacena todos los datos relacionados con la georreferenciación como son las distintas direcciones, los distintos puntos geográficos y los enlaces entre direcciones y puntos y entre fuentes y puntos. Este registro permite identificar, reconocer y posicionar las direcciones que se han ido incluyendo provenientes de las distintas fuentes estadísticas.

Objetivos

El objetivo de este sistema es almacenar todas las direcciones, que provienen de las diferentes fuentes georreferenciadas y acceder a esta información de forma sencilla.

La georreferenciación es el proceso de referenciar datos contra un sistema de coordenadas geoespacial conocido, ajustándose a puntos conocidos en el sistema de coordenadas, de manera que los datos puedan ser visualizados, procesados, consultados y analizados junto con otros datos geográficos.

Con todas las fuentes georreferenciadas se puede tener la posición o ubicación de cada uno de los registros de las diferentes fuentes en el espacio. Lo que permite visualizar y consultar estos datos de una forma más visual y entendible.

Metodología

Al sistema de almacenamiento de las direcciones, puntos y enlaces del proceso de georreferenciación se le llama Registro de Portales.

El Registro de Portales se plantea como una infraestructura estadística integradora de las diferentes fuentes de datos, que permite identificar cada una de las diferentes direcciones existentes en los registros administrativos.

Estructura del Registro de Portales

En el registro se distinguen tres clases de tablas (Figura 2): tabla de versiones (IDF), tabla de puntos (IDT) y la tabla de relaciones (URD).

Figura 2: Componentes del Registro de Portales

El IDT tiene el propósito de almacenar puntos que identifican a direcciones. Para ello, se requieren campos que representen puntos. La Tabla 1 muestra el diseño del IDT, versión simplificada.

Variable	**Tipo**	**Descripción**
uuid	varchar (36)	Identificador Único Universal
luid	serial	Identificador Único Local
stid	varchar (61)	Identificador de esquema y tabla
marcatiempo	date	Sello de tiempo de creación de la observación
geom	geom	Localización del portal en el SRID4326
geohash	varchar (20)	Valor del geohash calculado con longitud y latitud (10 c.)
longitud	numeric	Longitud del punto en grados con decimales
latitud	numeric	Latitud del punto en grados con decimales
geocode_grid_250	varchar (18)	Geocode del grid de 250

Tabla 1: Diseño del IDT, versión simplificada.

El IDF tiene el propósito de integrar y almacenar fuentes de información heterogéneas. Para ello, se requieren campos normalizados y depurados que representen la dirección, que hace referencia al conjunto de características propias de una dirección y permiten diferenciarla frente a las demás. La Tabla 2 muestra el diseño del IDF, versión simplificada.

El URD tiene el propósito de integrar y almacenar las relaciones entre IDT-IDF e IDT-DAT. Permite identificar cuales son las versiones de las direcciones (IDF) que corresponden a cada punto (IDT) y a su vez que punto (IDT) le corresponde a cada registro de las diferentes fuentes (DAT). La Tabla 3 muestra el diseño del URD, versión simplificada.

Variable	Tipo	Descripción
uuid	varchar (36)	Identificador Único Universal
luid	serial	Identificador Único Local
stid	varchar (61)	Identificador de esquema y tabla
marcatiempo	date	Sello de tiempo de creación de la observación
tvia_nn	varchar (13)	Tipo de vía no normalizado
nvia_nn	varchar (100)	Nombre de vía no normalizado
numer_nn	varchar (10)	Número de portal no normalizado
tvia	varchar (20)	Tipo de vía
cvia	varchar (5)	Código de vía
nvia	varchar (100)	Nombre de vía
numer	varchar (5)	Número de portal
kmt	varchar (3)	Punto kilométrico
hmt	varchar (1)	Punto kilométrico, hectómetros
nomedif	varchar (50)	Nombre del edificio
codmun	varchar (5)	Código de municipio
nommun	varchar (35)	Nombre del municipio
direccion	varchar (255)	Dirección (tvia+nvia+numer+nommun)[1]

Tabla 2: Diseño del IDF, versión simplificada

Variable	Tipo	Descripción
uuid	varchar (36)	Identificador Único Universal
luid	serial	Identificador Único Local
stid	varchar (61)	Identificador de esquema y tabla
marcatiempo	date	Sello de tiempo de creación de la observación
uuid_a	varchar (36)	UUID de la observación que se complementa (idt)
stid_a	varchar (61)	STID de la observación que se complementa (idt)
uuid_b	varchar (36)	UUID de la observación que aporta información complementaria (idf o dat)
stid_b	varchar (61)	STID de la observación que aporta información complementaria (idf o dat)
rel_type	varchar (30)	Tipo de relación
rel_event	date	Fecha del evento
rel_period	date	Fecha de referencia
rel_description	varchar (255)	Descripción de la relación

[1] en la mayoría de los casos

link_type	varchar (30)	Tipo de enlazamiento según lista controlada (codelist)
link_description	varchar (255)	Descripción abierta del tipo de enlazamiento
link_group	varchar (36)	UUID de la relación que creó el grupo
link_quality	varchar (255)	Descripción abierta de la calidad del enlazamiento
link_quality_rank	varchar (10)	MUY_ALTA, ALTA, MEDIA, BAJA, MUY_BAJA
link_active	varchar (1)	Indicador de actividad del enlace (1 o 6)
link_active_in	date	Fecha de alta del enlace
link_active_out	date	Fecha de baja del enlace

Tabla 3: Diseño del URD, versión simplificada

Normalización

Antes de comenzar el proceso de georreferenciación hay un proceso de normalización [2]. Este proceso consiste en crear las variables necesarias de los registros que van en la tabla IDF. Y a su vez estos campos sirven para el proceso de georreferenciación.

Algunos de los procesos de normalización y obtención de variables se encuentran en la Tabla 4.

Variable	**Proceso**
cvia	Es el código de vía. En el caso de tener valor se le añade tantos 0s a la izquierda hasta alcanzar el tamaño 5. En el caso de no tener valor se le asigna _U.
tvia_nn	Es el tipo de vía no normalizado y que viene incluido en la fuente administrativa, puede ser vacío
tvia	Es el tipo de vía normalizado. Algunas de las funciones de normalización consisten en: ● Eliminar puntuaciones (.,:;?!|/`=*+-^_~”’[]{}()<>@#$) ● Eliminar espacios de más ● Reemplazar palabras abreviadas por su no abreviatura ● Eliminar tildes ● Poner en mayúscula
nvia_nn	Es el nombre de vía no normalizado y que viene incluido en la fuente administrativa, puede ser vacío.
nvia	Es el nombre de vía normalizado. Algunas de las funciones de normalización consisten en: ● Poner en mayúsculas

	● Eliminar S/N, C/ ● Eliminar espacios de más ● Eliminar puntuaciones ● Eliminar tildes ● Eliminar palabras reservadas (“ante”, “cabe”, “tras”, “la”, “las”, “el”, “los”, “del”, “urb”, “ur”, “avd”, “av”, “avda”, “bl”, “blq”, “ed”, “edif”, “edf”, “bloq”, “pta”, “ptal”, “urb”, “ur”, “local”, “apto”, “aptos”, “bloque”, “parcela”, “portal”, “apartamentos”, “porton”, “duplex”)
numer_nn	Es el número de portal no normalizado y que viene incluido en la fuente administrativa, puede ser vacío
numer	Es el número de portal normalizado. Algunas de las funciones de normalización consisten en: ● Eliminar NA ● Eliminar letras ● Eliminar S/N, SN ● Eliminar puntuaciones ● Eliminar tildes ● Añadir 0s hasta alcanzar el tamaño 5
codmun	Es el código del municipio. En el caso de tener valor se le añade tantos 0s a la izquierda hasta alcanzar el tamaño 5. En el caso de no tener valor se le asigna _U.
nommun	Es el nombre del municipio. Se obtiene a partir del código de municipio.
direccion[2]	Consiste en la unión de los campos tvia+nvia+numer+nommun en el caso de disponer de todos esos campos, hay ocasiones en la que la dirección viene completa. Algunas operaciones realizadas sobre este campo: ● Poner en mayúsculas el campo nommun ● Eliminar puntuaciones del campo nommun ● Eliminar tildes del campo nommun ● Eliminar espacios innecesarios del campo nommun ● Eliminar 0s de la izquierda del campo numer ● Eliminar espacios innecesarios

Tabla 4: Procesos de normalización de variables

Geocodificación

También se realiza un proceso de geocodificación, asignación de un cvia. Para este proceso se usa el paquete de R “fastLink” [3], el cual realiza un record linkage [4] de forma probabilística y en paralelo. Lo que hace que el proceso de geocodificación sea

[2] a partir de ahora se utilizará “direccion” para referirse a esta variable o campo de la tabla IDF

más rápido. Esta asignación nos permite que el proceso de georreferenciación sea más inmediato ya que permite relaciones mediantes códigos (codmun, cvia y numer).

Georreferenciación

El proceso de georreferenciación [5] se divide en diferentes tipos de georreferenciaciones: determinística, pseudodeterminística, probabilística y aleatoria.

La determinística consiste en las relaciones por códigos (codmun y/o cvia y/o numer), y las coincidencias por "direccion" o "tvia+nvia". Realizando un proceso de record linkage[3], que consiste en comparar estos campos "direccion" o "tvia+nvia" con otros registros ya georreferenciados y con los mismos códigos. Se le da un peso que nos indica el parecido entre ellos. El portal que se le asigna es uno ya existente y donde la "direccion" o "tvia+nvia" será igual o similar. En este caso el paquete de R elegido para la comprobación de la similitud es "RecordlLinkage" [6].

La pseudodeterminística sólo se aplica en la fuente de PMH. Consiste en asignar una aproximación del portal, es decir el portal que se le asigna no será el exacto sino uno que está a cierta distancia del mismo.

La probabilística consiste en la utilización de las APIs (ArcGIS, Bing, Cartociudad, GoogleMap, GooglePlace, Here, Komoot) para la georreferenciación. Cada "direccion" es llamada con cada una de las APIs. El resultado de esto son varios puntos para una misma "direccion", una api puede devolver cero, uno o varios puntos para un mismo portal. Esto implica que debemos seleccionar el punto más adecuado para cada "direccion". Para ello se ha implementado un método de selección, Tabla 5.

La aleatoria consiste en asignar un punto aleatorio a una "direccion" que no se ha podido georreferenciar. Este punto puede ser de la misma vía, del mismo municipio, misma provincia o de Canarias. La selección del punto aleatorio se realiza mediante el paquete "pps" [7] de R que consiste en obtener un candidato cuyo valor probabilístico está en proporción al tamaño. Los candidatos dependen de la fuente, por ejemplo para PMH es el número de personas que hay en cada portal; para demandantes y llamamientos, es la población activa que hay en cada portal; para cuentas de cotización es el número de cuentas coincidentes en los dos primeros dígitos de la CNAE en cada portal.

Orden	Método	Descripción

[3] https://en.wikipedia.org/wiki/Record_linkage

1º	Validaciones espaciales	distancia del punto al municipio y a la vía si esta se encuentra geocodificada
2º	Validaciones internas	distancia del punto al centroide calculado de los puntos asignados a la misma "direccion" (puntos a una distancia 'd' a la vía o puntos con un rango de calidad aportado por cada API)
3º	Cálculo de la gordura	se obtienen dos valores: ● la suma de los pesos que tiene cada api de las que han devuelto al menos un punto para esa "direccion", ● el número de portales que están en un buffer de distancia 10, de los devueltos para la misma "direccion".
4º	Cálculo de la verosimilitud	sería la gordura tipificada media elevada al error tipificado medio
5º	Proceso de selección	● seleccionar el de mejor verosimilitud, ● en el caso de igual verosimilitud se selecciona el de mejor gordura, ● en el caso de igual verosimilitud, igual gordura se selecciona el de la mejor API.

Tabla 5: Método de selección del mejor punto

Proceso

El proceso se realiza en su totalidad en código R[4], usando el entorno RStudio[5]. En la Figura 3 se explica el orden de cada proceso.

[4] https://www.r-project.org/
[5] https://www.rstudio.com/

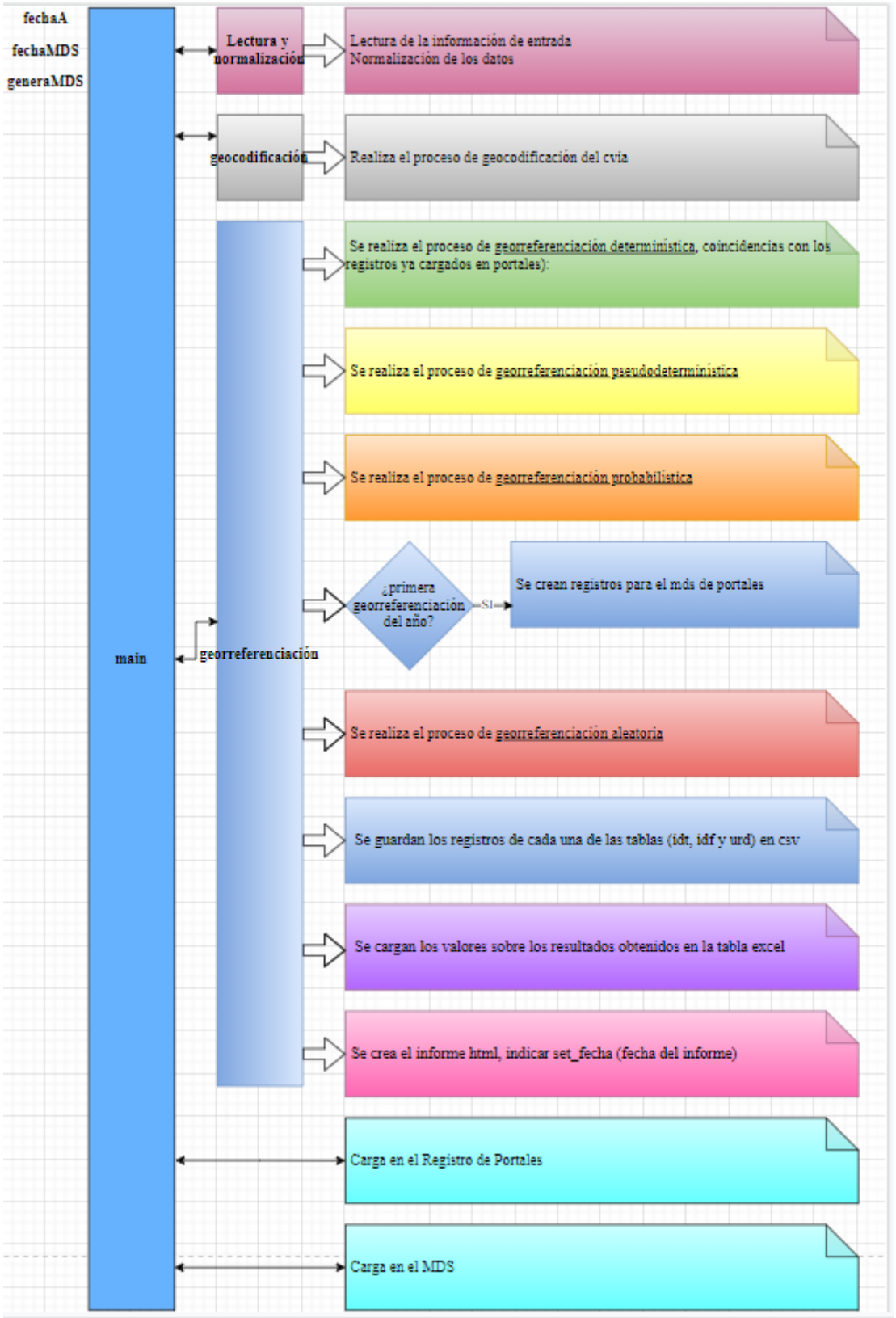

Figura 3: Diagrama de flujo del proceso

Fase 1: Lectura de datos y normalización de las direcciones, ver Figura 4.

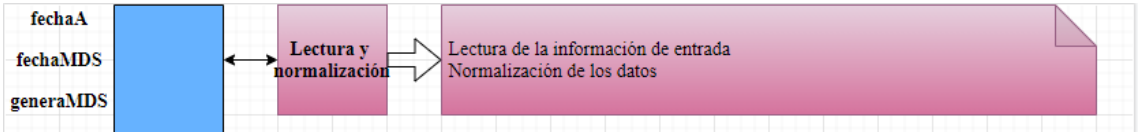

Figura 4: Lectura y normalización

Fase 2: Geocodificación de las direcciones, asignación de cvia. Ver Figura 5.

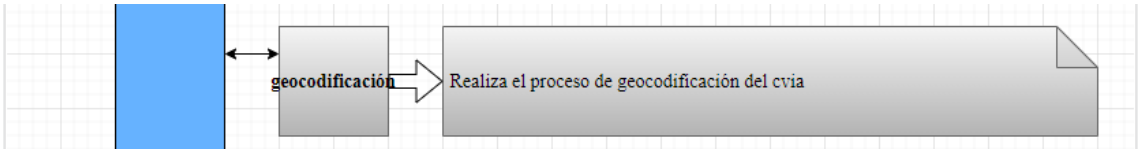

Figura 5: Geocodificación de direcciones

Fase 3: Georreferenciación determinística. Ver Figura 6.

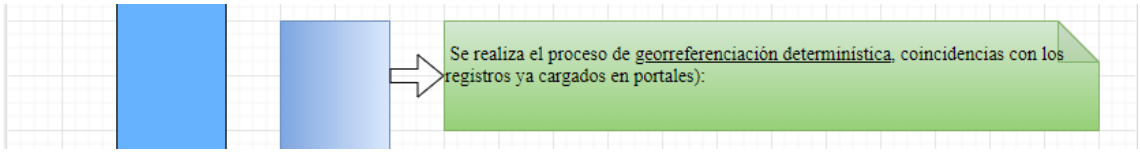

Figura 6: Georreferenciación determinística

Fase 4: Georreferenciación pseudodeterminística, sólo se realiza en el caso de PMH. Ver Figura 7.

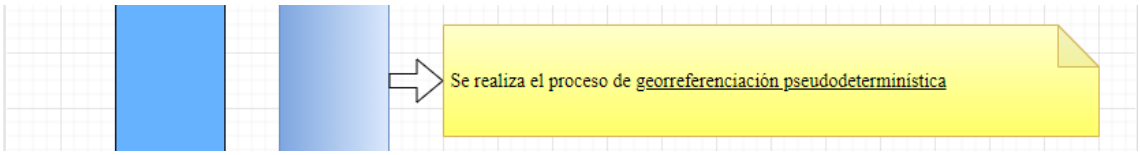

Figura 7: Georreferenciación pseudodeterminística

Fase 5: Georreferenciación probabilística, mediante APIs. Ver Figura 8.

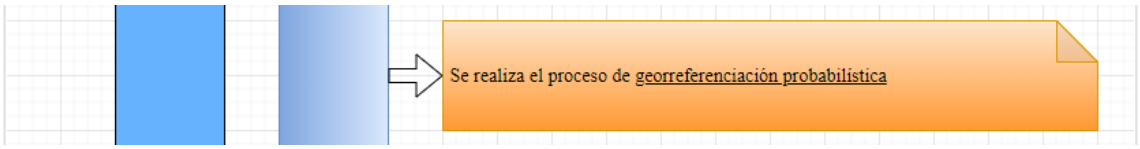

Figura 8: Georreferenciación probabilística

Fase 6: En el caso de ser la primera georreferenciación del año se crean registros para el MDS de portales[6]. Ver Figura 9.

Fase 7: Georreferenciación aleatoria. Ver Figura 10.

[6] El MDS de portales consiste en el peso de cada uno de los portales.

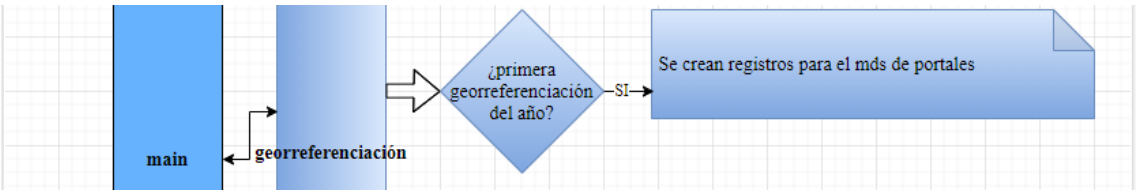

Figura 9: Creación de registros para el MDS de portales

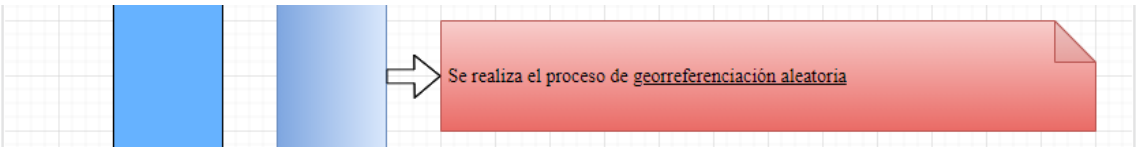

Figura 10: Georreferenciación aleatoria

Fase 8: Se guardan los registros de cada una de las tablas (idt, idf y urd) en csv. Se cargan los valores sobre los resultados obtenidos en una tabla excel. Estos valores son útiles para la creación del informe final. Ver Figura 11.

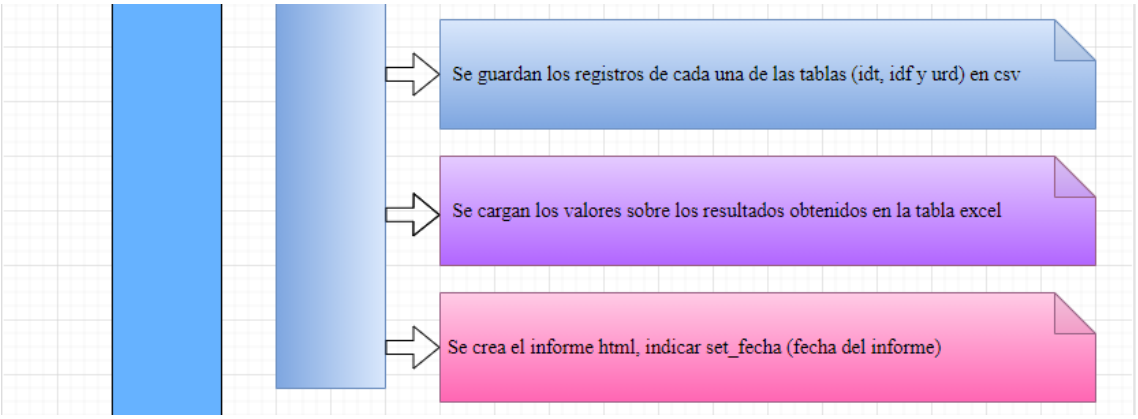

Figura 11: Procesos de almacenamiento y creación de informe

Fase 9: Carga de los registros en el Registros de Portales, y en el caso de haber creado registros para el MDS de portales estos también se cargan. Ver Figura 12.

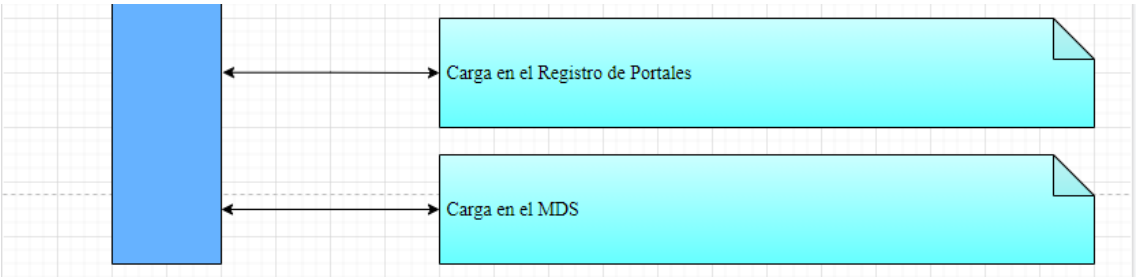

Figura 12: Carga en Postgres

Resultados

Este método se ha implementado en su totalidad en código R, donde las principales librerías utilizadas son dplyr, stringr, RPostgreSQL, RecordLinkage, jsonlite, RCurl. Como interfaz de desarrollo se ha utilizado RStudio, como almacenamiento de datos se dispone de una base de datos en PostgreSQL.

La integración de las diferentes fuentes en el Registro de Portales se ha llevado a cabo de la siguiente manera:

La primera integración se ha realizado con el año piloto 2017, siendo la fuente principal los registros proporcionados por el Sistema de Información Territorial de Canarias (SITCAN), y una segunda fuente el Padrón Municipal de Habitantes (PMH).

Fichero	N_Ficheros	N_DAT	N_IDF	% DAT_IDF	N_IDT	% DAT_IDT	N_URD
SITCAN	1	466.668	466.668	79,89	458.813	97,32	933.336
PMH	2	4.383.451	117.453	20,11	12.633	2,68	4.500.904
TOTAL	3	4.850.119	584.121	100	471.446	100	5.434.240

Tabla 6: Ficheros incluidos en el Registro de Portales en el año piloto 2017

La Tabla 6 muestra el número de ficheros que se han incluido para cada una de las fuentes, siendo un sólo fichero para SITCAN y dos ficheros de PMH (enero y julio). Muestra el número de registros de cada fuente (N_DAT) así como el número de registros incluidos en las tablas IDF, IDT y URD (N_IDF, N_IDT, N_URD). Por último el tanto por ciento en cada una de las tablas IDF, IDT y URD. El 79,89% de las versiones (registros del IDF) y 97,32 % de los puntos (registros del IDT) han sido creados por la fuente SITCAN.

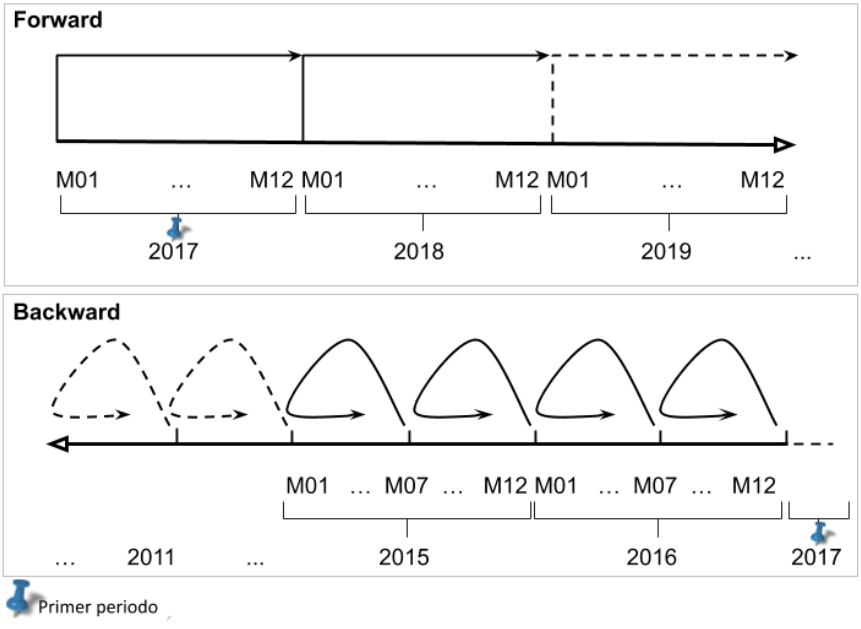

Figura 13: Integración de ficheros según periodo

La segunda integración fue todo el histórico de PMH, desde 2010 hasta el 2020. El orden para la integración de estos registros fue PMH20160701, PMH20160101, PMH20150701, PMH20150101, PMH20140701, PMH20140101, PMH20130701, PMH20130101, PMH20120701, PMH20120101, PMH20110701, PMH20110101,

PMH20100701, PMH20100101, SITCAN2018, PMH20180101, PMH20180701, PMH20190101, PMH20190701, PMH20200101, PMH20200701 (Figura 13).

Se ha intentado que la distancia entre un periodo y otro fuese el mínimo, para que el incremento de versiones y puntos no fuera muy grande sino pausado.

Fichero	N_ Ficheros	N_DAT	N_IDF	% DAT_IDF	N_IDT	% DAT_IDT	N_URD
SITCAN	2	469.690	469.690	72,57	461.835	96,14	939.380
PMH	22	48.277.695	177.518	27,43	18.523	3,86	48.455.213
TOTAL	24	48.747.385	647.208	100	480.358	100	49.394.593

Tabla 7: Ficheros del histórico de PMH incluidos en el Registro de Portales

En este caso (Tabla 7) ya tendríamos dos ficheros de SITCAN (2017 y 2018), y 22 ficheros de PMH (2 semestres por año desde 2010 hasta 2020). SITCAN sigue aportando la gran mayoría de versiones y puntos siendo el 72,57% y el 96,14% respectivamente.

A continuación se muestra una gráfica (Figura 14) donde se puede visualizar el comportamiento de la fuente de PMH en el Registro de Portales.

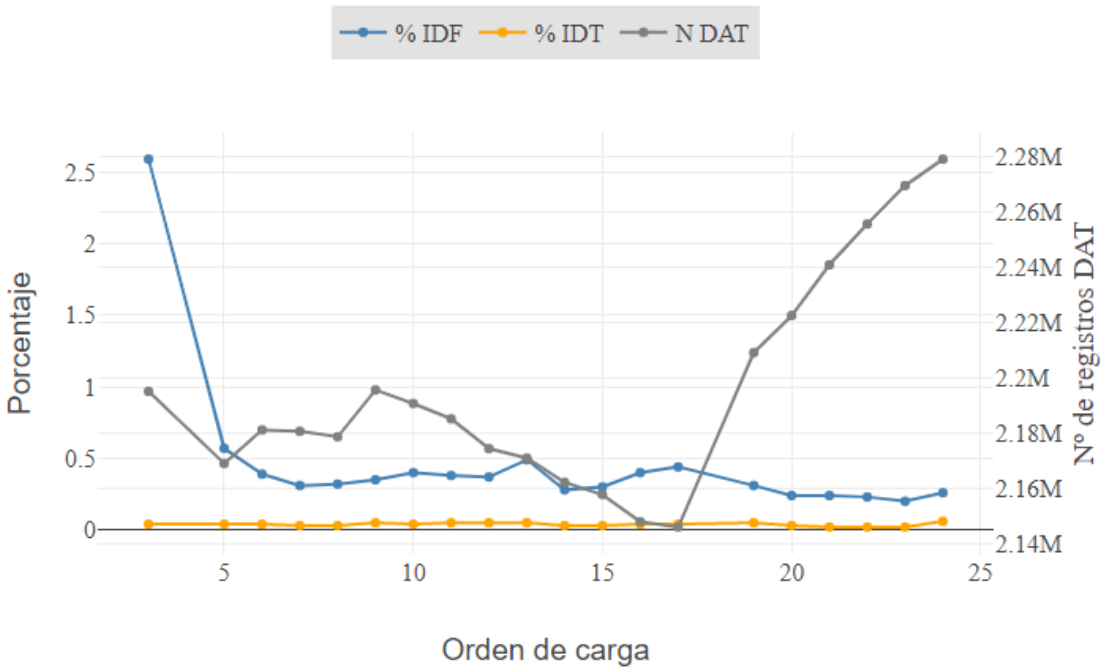

Figura 14: PMH en el Registro de Portales

Se puede observar que el número de versiones que se obtiene es muy bajo comparado con el obtenido en el primer periodo de PMH. El número de puntos nuevos obtenidos con esta fuente es muy bajo.

A continuación, Figura 15, los tipos de enlace que se produce en cada periodo de PMH.

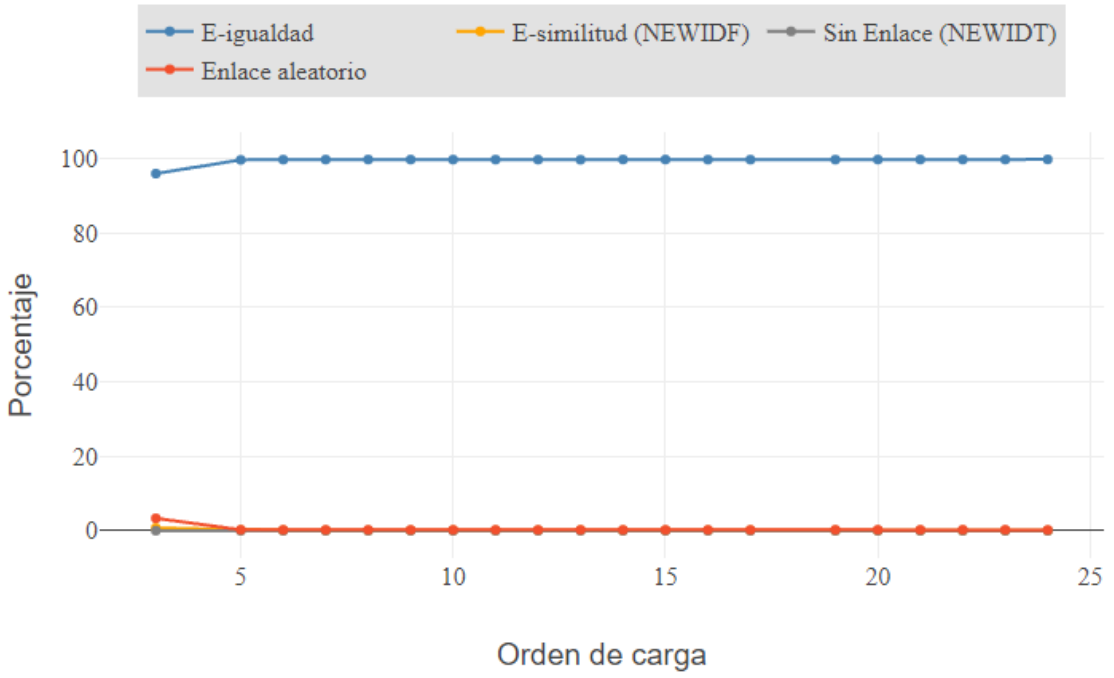

Figura 15: Tipos de enlace en PMH

Vemos que la mayoría de enlaces se producen por igualdad, siendo del 99%.

La tercera integración ha sido el resto de fuentes desde 2011 a 2020. Las siguientes georreferenciaciones corresponden a restos de fuentes que fueron necesarias su inclusión en el Registro de Portales para la realización de publicaciones: afiliaciones a la seguridad social y afiliaciones a muface (AFI), demandantes (DEM) y contratos (CON). También para concluir con todos esos periodos se han incluido cuentas de cotización a la seguridad social (CC) y demandantes (DEM) en su totalidad.

La tabla resumen por fuente de datos, Tabla 8, muestra la totalidad de las fuentes de PMH, DEM, CC, SITCAN del periodo 2011-2020 y restos necesarios de las fuentes de AFI y CON. En este caso DEM proporciona una gran cantidad de versiones, un 34,01% de los registros incluidos en el IDF. Esto es debido a que en la primera carga de DEM las direcciones que llegan no coinciden con las que ya tenemos almacenadas en el Registro de Portales, la forma de escribir la dirección es muy diferente a las fuentes que ya están cargadas. Aún así, SITCAN sigue siendo el que incorpora el mayor caso de versiones un 34,94% y el mayor caso de puntos un 76,05%.

Fichero	N_ Ficheros	N_DAT	N_IDF	% DAT_IDF	N_IDT	% DAT_IDT	N_URD
SITCAN	2	469.690	469.690	34,94	461.835	76,05	939.380
PMH	22	48.277.695	177.518	13,20	18.523	3,05	48.455.213
AFI-							

DEM-CON	3	919.412	145.419	10,82	88.881	14,64	1.064.831
CC	45	3.799.145	94.494	7,03	13.412	2,21	3.893.639
AFI	1	292	6	0,00	0	0,00	298
DEM	119	40.594.687	457.212	34,01	24.596	4,05	41.051.899
TOTAL	192	94.060.921	457.212	100	607.247	100	95.405.260

Tabla 8: Ficheros del 2011-2020 incluidos en el Registro de Portales

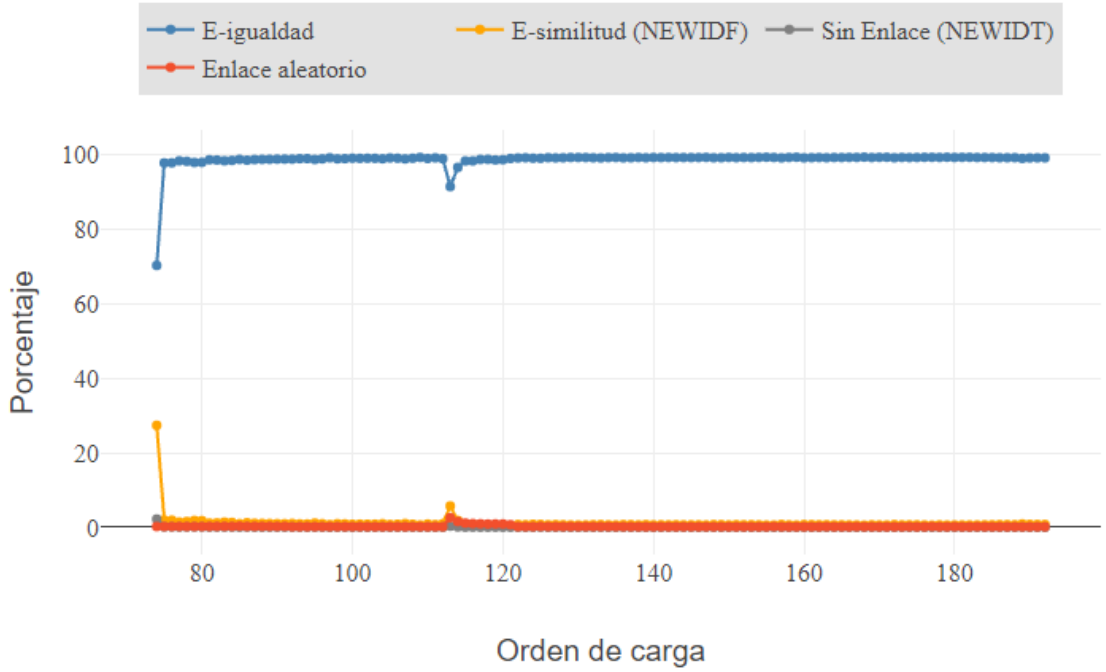

Figura 16: Tipos de enlace de DEM

La gráfica de la Figura 16 muestra los enlaces en la fuente de DEM. Para la fuente de demandantes el proceso de integración se ha realizado en el siguiente orden, teniendo en cuenta que es mensual: 2017-2020, 2016-2011. En este caso podemos ver como el número de versiones es mayor y el número de igualdades es menor con respecto a otros periodos en dos ocasiones. La primera ocasión es el inicio de la georreferenciación de esta fuente, como es lo normal cada fuente proporciona nuevas versiones debido a que la toma de los datos siempre es diferente. La segunda ocasión coincide con el periodo de abril de 2020, sería el primer periodo de demandantes en el inicio del COVID-19.

La cuarta integración es la que se está haciendo actualmente, implica todas las fuentes antes mencionadas del año en curso. Podemos ver el calendario en la Figura 17.

	Año 2021											
	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
Padrón Mun.de Hab. (PMH)		RD PMH 2021 0101						RD PMH 2021 0701				
Cuentas Cotización (CC)	RD CC 2020 1231	RD CC 2021 0131	RD CC 2021 0228	RD CC 2021 0331	RD CC 2021 0430	RD CC 2021 0531	RD CC 2021 0630	RD CC 2021 0731	RD CC 2021 0831	RD CC 2021 0930	RD CC 2021 1031	RD CC 2021 1130
Afiliados a la SS (AFI)	RD AFI 2020 1231	RD AFI 2021 0131	RD AFI 2021 0228	RD AFI 2021 0331	RD AFI 2021 0430	RD AFI 2021 0531	RD AFI 2021 0630	RD AFI 2021 0731	RD AFI 2021 0831	RD AFI 2021 0930	RD AFI 2021 1031	RD AFI 2021 1130
Afiliados a MUFACE (MUF)	RD MUF 2020 1231			RD MUF 2021 0331			RD MUF 2021 0630			RD MUF 2021 0930		
Demandantes (DEM)	RD DEM 2020 1231	RD DEM 2021 0131	RD DEM 2021 0228	RD DEM 2021 0331	RD DEM 2021 0430	RD DEM 2021 0531	RD DEM 2021 0630	RD DEM 2021 0731	RD DEM 2021 0831	RD DEM 2021 0930	RD DEM 2021 1031	RD DEM 2021 1130
Contratos (CON)	RD CON 2020 1231	RD CON 2021 0131	RD CON 2021 0228	RD CON 2021 0331	RD CON 2021 0430	RD CON 2021 0531	RD CON 2021 0630	RD CON 2021 0731	RD CON 2021 0831	RD CON 2021 0930	RD CON 2021 1031	RD CON 2021 1130
LLamamientos (LLAM)	RD LLAM 2020 1231	RD LLAM 2021 0131	RD LLAM 2021 0228	RD LLAM 2021 0331	RD LLAM 2021 0430	RD LLAM 2021 0531	RD LLAM 2021 0630	RD LLAM 2021 0731	RD LLAM 2021 0831	RD LLAM 2021 0930	RD LLAM 2021 1031	RD LLAM 2021 1130

Figura 17: Calendario de ficheros administrativos

tvia	cvia	nvia.idf	numer	kmt	hmt	nomedif	codmun	nommun
CALLE	00425	GERARDO ESTEVEZ	00001				35014	La Oliva
CALLE	00426	PROFESOR BARTOLOME HERNANDEZ ESTUPIÑAN	00020				35011	Ingenio
CALLE	00426	PROFESOR BARTOLOME HERNANDEZ ESTUPIÑAN	00017				35011	Ingenio
CALLE	00426	PROFESOR BARTOLOME HERNANDEZ ESTUPIÑAN	00007				35011	Ingenio
CALLE	00426	PROFESOR BARTOLOME HERNANDEZ ESTUPIÑAN	00022				35011	Ingenio
CALLE	00426	PROFESOR BARTOLOME HERNANDEZ ESTUPIÑAN	00004				35011	Ingenio
CALLE	00426	PROFESOR BARTOLOME HERNANDEZ ESTUPIÑAN	00021				35011	Ingenio
CALLE	00426	PROF BARTOLOME HDEZ EST	00017				35011	Ingenio
CALLE	00426	PROF BARTOLOME HERNANDEZ	00007				35011	Ingenio
CALLE	00426	PROFESOR BARTOLOME HERNAN	00022				35011	Ingenio
CALLE	00426	PROF BARTOLOME HDEZ EST	00004				35011	Ingenio
CALLE	00426	PROF BARTOLOME HERNANDEZ	00021				35011	Ingenio
CALLE	00426	PROF BARTOLOME HDEZ EST	00020				35011	Ingenio

Figura 18: Ejemplo de muchas versiones

Seguidamente se muestran algunos ejemplos del porqué se ha ideado esta estructura del Registro de Portales y lo que implica tener todas las versiones de una misma dirección.

En la Figura 18 podemos ver 4 formas diferentes de llamar a una vía. La variedad de formas se puede deber a que provienen de diferentes fuentes o simplemente quien lo escribe lo hace como el ciudadano se lo dicta o como más fácil o rápido le resulte.

tvia	cvia	nvia	numer	kmt	hmt	nomedif	codmun	nommun
CALLE	00239	JOSE ANTONIO PRIMO RIVERA	00037				35011	Ingenio
CALLE	00239	JOSE ANTONIO PRIMO RIVERA	00060				35011	Ingenio
CALLE	00239	JOSE ANTONIO PRIMO RIVERA	00065				35011	Ingenio
CALLE	00239	JOSE ANTONIO PRIMO RIVERA	00053				35011	Ingenio
CALLE	00239	ALONSO MATOS	00037				35011	Ingenio
CALLE	00239	ALONSO MATOS	00060				35011	Ingenio
CALLE	00239	ALONSO MATOS	00065				35011	Ingenio
CALLE	00239	ALONSO MATOS	00053				35011	Ingenio
AVENIDA	00257	TABABAIRE	00015				35014	La Oliva
AVENIDA	00257	TABABAIRE	00019				35014	La Oliva
CALLE	00257	GENERALISIMO FRANCO	00015				35014	La Oliva
CALLE	00257	GENERALISIMO FRANCO	00019				35014	La Oliva

Figura 19: Ejemplo de ley de memoria histórica

La Figura 19 contiene un ejemplo de como las direcciones pueden cambiar en su totalidad el nombre y el tipo de vía. En este caso es debido a la ley de memoria histórica, pero el renombramiento de vía puede ser por cualquier otro motivo.

El hecho de guardar todas las versiones nos facilita la tarea de georreferenciar esa misma dirección cuando nos llega de la misma fuente en periodo diferente o de otra fuente donde se ha guardado de la misma manera.

Conclusión

El principal motivo para disponer del Registro de Portales es que nos permite tener un sistema “sencillo” de almacenaje de la georreferenciación de cualquier fuente estadística.

Este sistema almacena versiones de todas las direcciones proveniente de diferentes fuentes, el hecho de almacenar esta información hace que nuestro registro sea una base de conocimiento que se nutre a medida que llegan nuevos periodos o fuentes. Y a su vez sea más fácil la georreferenciación debido a que cada vez será más probable que la dirección que llegue ya se encuentre almacenada en el registro.

Los puntos geográficos ya almacenados permiten posicionar a cualquier registro de cualquier fuente en el espacio permitiendo de manera fácil la representación o visualización de la información que proporcionan estas fuentes en mapas. Este tipo de visualizaciones está teniendo bastante auge hoy en día debido a que hace más visible y entendible toda información.

El hecho de contemplar los diferentes enlaces entre puntos y versiones y entre puntos y registros de fuentes, hacen de esta estructura una herramienta muy manejable.

Principales referencias bibliográficas

[1] «BOE» núm. 68, de 20 de marzo de 1991, páginas 8892 a 8897 (6 págs.)

[2] Christen, P. (2006) "A Comparison of Personal Name Matching: Techniques and PracticalIssues", Sixth IEEE International Conference on Data Mining, pp. 290-294.

[3] Paquete "fastLink": https://cran.r-project.org/web/packages/fastLink/fastLink.pdf

[4] Ivan P. Fellegi & Alan B. Sunter (1969) A Theory for Record Linkage, Journal of the American Statistical Association, 64:328, 1183-1210, DOI: 10.1080/01621459.1969.10501049

[5] JA González Yanes, R Betancor Villalba, E Torres Medina, M Sanz Gil, MS Hernández García. (2018, octubre). Marco de estadística espacial de Canarias. Instituto Canario de Estadística. XX JECAS 2018 La Rioja.

[6] Paquete "Record Linkage": https://cran.r-project.org/web/packages/RecordLinkage/RecordLinkage.pdf

[7] Paquete "pps": https://cran.r-project.org/web/packages/pps/pps.pdf