El reto del control de la calidad de la información estadística difundida mediante técnicas de Big Data

López Vizcaíno, M. Esther, Iglesias Patiño, Carlos L., Andión Hermida, Ana, Veiguela Fernández, Noa.

Instituto Galego de Estatística

Resumen

Los institutos oficiales de estadística cada vez ponemos una mayor cantidad de información a disposición del público. Este aumento en la cantidad de información hace necesario el disponer de herramientas que permitan asegurar la calidad de la información difundida. El objetivo de este trabajo es describir las herramientas de las que dispone el IGE para velar por la calidad de la información disponible en la web y describir los problemas que nos plantean las técnicas tradicionales para realizar los controles de calidad. En este último punto se abre el horizonte del uso de otro tipo de técnicas vinculadas al Big Data.

Palabras clave: difusión, control de calidad, series temporales, Machine Learning.

1.- Introducción

A lo largo de sus más de dos décadas de existencia el Instituto Galego de Estatística (IGE) ha trabajado para convertirse en una de las principales fuentes de información de datos de carácter socioeconómico de la Comunidad Autónoma de Galicia y de sus principales divisiones territoriales.

Esta circunstancia ha motivado un incremento notable de la cantidad de datos disponibles en el Instituto, disponiendo así, no sólo de los de producción propia, sino también de los de diversos organismos oficiales que ofrecen información estadística de nuestra Comunidad Autónoma (Iglesias, 1999, 2001).

Además de aumentar el volumen de información disponible en la página web, en los últimos años, se ha puesto también énfasis en mejorar la calidad, la accesibilidad y el intercambio de la información estadística entre distintos usuarios y/o plataformas.

En el IGE contamos con un banco de datos principal para difundir la información estadística disponible en nuestra web. Este banco tiene actualmente un tamaño de 4,1 Gigabytes. Es más, toda la información que el IGE tiene en sus bases de datos de producción y difusión ocupa poco más de un Terabyte. El servicio de Informática del IGE diseña y programa las distintas aplicaciones mediante las que se accede a las bases de datos, permitiendo realizar tareas de grabación, intercambio o publicación en la web.

La estructura del banco de datos permite, además del acceso mediante tablas multidimensionales, que es la forma más común de difusión de datos estadísticos en la web del IGE, la captura directa de información por parte de los usuarios, lo que evita la manipulación de

la información y facilita la actualización de la misma.

Con esta manera de estructurar la información respondemos a las siguientes líneas de trabajo (filosofía de trabajo):

- Eliminar, en la medida de lo posible, el trabajo mecánico y el copia-pega, tratando de minimizar los errores en el manejo de los datos.
- Almacenar un dato **una sola vez (filosofía del "único dato").**
- Depender del Servicio de Informática sólo cuando se desarrollan las aplicaciones, y luego independizarse.

Durante los últimos años se han incrementado de manera exponencial los procedimientos de carga automática de la información que se difunde, con el objetivo de reducir costes y errores manuales. Esta baja del número de errores manuales ha incrementado otro tipo de errores: de programación, de carga de la información, etc. Esto nos ha hecho reflexionar sobre el reto de establecer procedimientos que permitan evaluar la calidad de la información publicada, y es aquí en donde nos encontramos con la necesidad del uso de técnicas de Big Data.

Una de sus múltiples definiciones es "el Big Data excede el alcance de los entornos de hardware de uso común y herramientas de software para capturar, gestionar y procesar los datos dentro de un tiempo transcurrido tolerable para su población de usuarios". Otra definición muy acreditada es: "Big Data son los grandes conjuntos de datos que tienen tres características principales: volumen, velocidad y variedad". En síntesis, Big Data es un conjunto de datos tan grandes que rompe las infraestructuras tradicionales (Joyanes, 2014).

Por tanto, estrictamente, en el IGE no tendríamos información que puede ser "clasificada" como "Big Data", pero en cambio necesitamos herramientas de "Big Data" que nos ayuden a procesar la información.

Por otra parte, el Código de Buenas Prácticas de las Estadísticas Europeas (Eurostat, 2017) se basa en quince+1 principios, que abarcan el entorno institucional, los procesos de elaboración de estadísticas y la producción estadística. Uno de estos principios es el compromiso con la calidad y establece que "Las autoridades estadísticas están comprometidas con la calidad; identifican sistemática y regularmente los puntos fuertes y débiles para mejorar continuamente la calidad del proceso y del producto". La calidad en este código se evalúa a través de 4 indicadores:

4.1: Existe una política de calidad definida y está a disposición del público. Se dispone de una estructura y unas herramientas organizativas para gestionar la calidad.

4.2: Se han establecido procedimientos para la planificación y el seguimiento de la calidad del proceso de elaboración de estadísticas.

4.3: La calidad del producto se controla con regularidad, se evalúa con respecto a posibles compromisos entre componentes de calidad y es objeto de informes con arreglo a los criterios

de calidad de las estadísticas europeas.

4.4: Análisis periódico y minucioso de la producción estadística clave, recurriendo incluso a expertos externos cuando es necesario.

Como se puede observar la calidad es una variable importante a la hora producir datos estadísticos. En este trabajo nos centraremos en la calidad de la difusión de los resultados estadísticos, que entronca también con el principio 14 del Código de Buenas Prácticas, y en concreto, con el indicador 14.1 “Las estadísticas son coherentes y consistentes internamente (es decir, se observan identidades aritméticas y contables).

Teniendo en cuenta lo anterior, lo que resta de trabajo se estructura de la siguiente forma: en la siguiente sección se hace una descripción del problema que se nos plantea en la estructura organizativa de la difusión de la información. A continuación, se hace una descripción de las técnicas que utilizamos a día de hoy para efectuar los controles de calidad y finalizamos con los trabajos que son necesarios abordar en el futuro y con unas pequeñas conclusiones.

2.- Problema

En este apartado se hace una descripción del origen del problema a la hora de velar por la calidad de la información. En el IGE se difunden datos que proceden de múltiples fuentes de información muy variada:

- Estadísticas producidas en el IGE
- Información producida por las Consellerías de la Xunta de Galicia bajo el amparo del Plan Gallego de Estadística
- Estadísticas procedentes del INE y Eurostat
- Datos procedentes de los Ministerios
- Otra información

Los métodos para incluir toda esta información en nuestra base de datos también son variados y los podemos resumir en los siguientes puntos:

- La información producida en el IGE se carga en la base de datos mediante procedimientos estándares ya establecidos
- Programación a medida para cargar ciertos conjuntos de datos que se actualizan de manera periódica: paro registrado, contratos, matriculaciones, catastro, …
- Captura de información mediante programas hechos a medida: por ejemplo la captura de toda la información que está disponible en el formato de pc-axis o la captura de datos en formato .tsv de Eurostat

- Datos recibidos en formatos como Excel, pdf, Word, etc, donde se efectúa cargas manuales de la información

Toda la información, tanto la recibida del exterior, como la producida internamente, pasa controles de calidad antes de ser publicada, pero aun así, en los últimos años advertimos que se cometen errores que pueden ser de diversos tipos:

- Errores producidos en la actualización manual de tablas que vienen en formatos Excel, Word, pdf,…

- Errores de programación: por ejemplo se tiene una programación a medida para actualizar información municipal, si se fusionan o segregan municipios, es necesario actualizar la programación, y siempre puede quedar algo…

- Errores de la carga de información en la base de datos: errores en el servidor, etc….

En el IGE durante los últimos años se han incrementado sustancialmente los procedimientos de carga automática de la información disminuyendo la carga manual. Esta baja en el número de errores manuales ha incrementado los otros tipos de errores que motivó plantearse establecer controles de calidad justo en el momento anterior a la publicación de la información

3.- Herramientas de control de calidad disponibles

Cuando uno se plantea evaluar la calidad de la información estadística que se va a difundir piensa en hacer tres tipos de controles:

- Que los datos sean coherentes en la dimensión del espacio (la suma de provincias debe ser el total, etc.) o en otro tipo de dimensiones (sexo, sectores, etc.) (control de totales)

- Si es una actualización de datos ya existentes, que el último dato publicado sea coherente con los anteriores (coherencia del último dato)

- Que las series de tiempo que se publiquen sean consistentes a lo largo del tiempo (consistencia)

3.1.- Control de totales

La práctica totalidad de toda la información estadística que se publica tiene una base territorial. En el IGE podemos publicar información hasta 6 ámbitos geográficos en una misma tabla: Galicia, provincias, comarcas, municipios, distritos y secciones. Todos estos ámbitos geográficos siguen una jerarquía y es necesario asegurarse que la información que se publique cumple esta jerarquía, es decir que la suma de las provincias es Galicia, que la suma de las comarcas de Coruña es la provincia de A Coruña, y así sucesivamente. Esta situación sólo la podemos utilizar cuando el cálculo es para totales y no en el caso de indicadores. Lo que es

necesario que se cumpla para la dimensión del espacio, también lo es para otro tipo de dimensiones, como puede ser el sexo, es decir la suma de hombres y mujeres tiene que ser el total, u otros casos como la edad, la actividad económica, etc.

Para hacer el control de totales en todas las dimensiones de la tabla que se pretende publicar nos apoyamos en una librería del software R (R Core Team, 2018) desarrollada en el IGE y con nombre "libhip" (Gómez, 2018) que permite, entre otras funcionalidades, leer las tablas multidimensionales y almacenarlas en objetos de una clase S3 de R (hip); manipular la información contenida en la clase "hip" mediante operadores como filtrado, permutaciones de ejes, …; crear tablas de la clase "hip" y grabar tablas multidimensionales en la base de datos.

Es bueno antes de nada fijar una serie de conceptos que se repetirán a lo largo de la ponencia:

- Dimensiones o ejes: cada una de las características clasificatorias de los datos (sexo, edad, espacio, …)

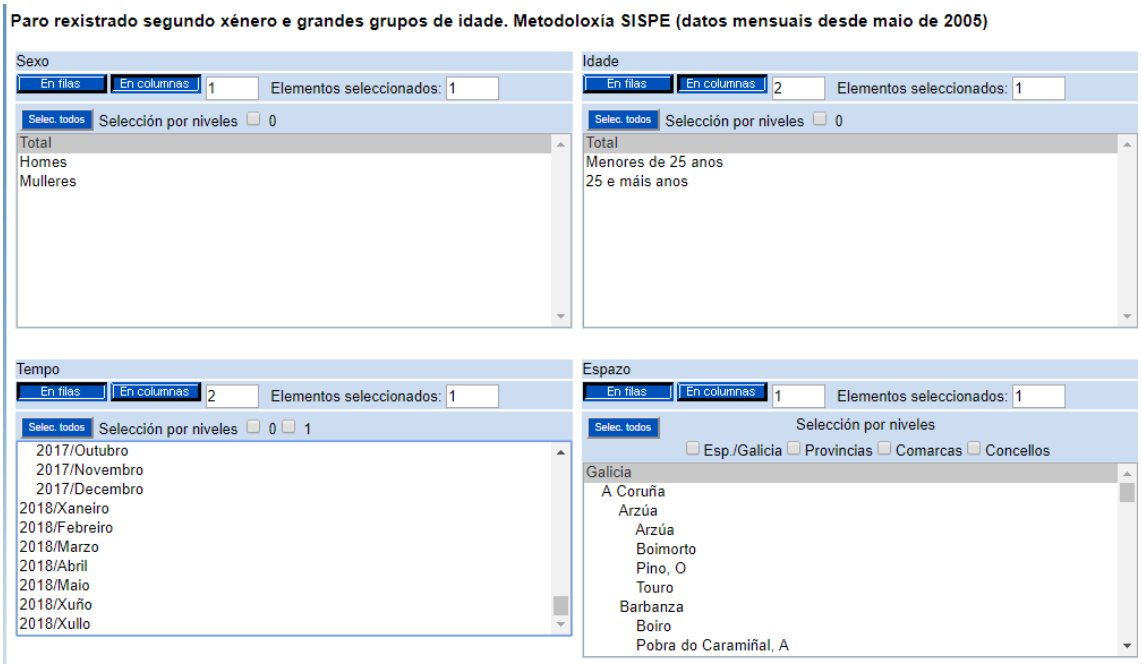
- Miembros: cada una de las modalidades que puede tomar una dimensión: en el caso del sexo, hombres, mujeres y total.

- Niveles: las tablas multidimensionales pueden tener una jerarquía de miembros presentando unos niveles según la profundidad en ella: Galicia, provincia, comarca, municipio, son los niveles de la dimensión espacio.

Aprovechando las funcionalidades que nos ofrece la librería "libhip" crearemos una nueva función en R, con nombre `controlTotal`, que evalúe la coherencia de la información en todas las dimensiones y a lo largo de todos los años. Esta función tendrá como argumentos la tabla multidimensional, los números de las dimensiones de los ejes diferentes al espacio en donde se quiere comprobar la coherencia, y las relaciones de los miembros dentro de los ejes. Por ejemplo, con la tabla disponible ahora en la web sobre *Paro registrado según género y grandes grupos de edad para Galicia, provincias, comarcas y municipios*:

(http://www.ige.eu/igebdt/selector.jsp?COD=744&paxina=001&c=0204036001):

Figura1.- Pantalla de selección de la tabla Paro registrado según género y grandes grupos de edad para Galicia, provincias, comarcas y municipios disponible en la web del IGE



Se invoca la función creada con la siguiente sintaxis:

```
resultado=controlTotales(tabla,dimespacio,ejes,rel,pos,rele)
```

tabla: es el código de la tabla, en este caso la 744.

ejes: dimensiones (diferentes del espacio) donde se quiere evaluar la coherencia, las dimensiones 1 (Sexo) y 2 (Edad): c(1,2).

rel: es necesario aportarle las relaciones dentro de los miembros, por ejemplo en el Sexo el primer miembro (Total) es la suma de los otros dos (Hombres y Mujeres) y se le indica que rel=c(1,2,2).

pos: parámetro que indica los niveles que se quiere comparar, en este caso el 1 con el 2, c(1,2)

rele: por último es necesario especificarle las relaciones entre los niveles del espacio *rele*=list(c(1,0),c(2,1),c(3,2)). En este caso se le está diciendo que Galicia es la suma de provincias, las comarcas suman las provincias y los municipios dentro de cada comarca, la comarca.

Esta función devuelve una lista de cada una de las dimensiones con los posibles errores. En la siguiente figura se presenta una lista de errores para la dimensión del sexo. En las tres primeras columnas se pueden observar las posibles combinaciones de todas las dimensiones excepto el sexo, y en la última columna están las diferencias entre el total y las sumas por sexo.

Tabla 1: Extracto de las diferencias existentes en la dimensión del sexo entre las suma de los miembros del sexo (hombres y mujeres) y los totales.

	Espazo	Idade	Tempo	Diferencia
156957	Entrimo	Menores de 25 anos	2016	-3
597	Baltar	Menores de 25 anos	2005	-2
649	Chandrexa de Queixa	Menores de 25 anos	2005	-2
10719	San Amaro	Menores de 25 anos	2006	-2
11006	Negueira de Muñiz	25 e máis anos	2006	-2
25153	Coirós	Menores de 25 anos	2007	-2
25298	Samos	Menores de 25 anos	2007	-2
25344	San Amaro	Menores de 25 anos	2007	-2
39808	Moeche	Menores de 25 anos	2008	-2
39997	Arnoia, A	Menores de 25 anos	2008	-2
40024	Chandrexa de Queixa	Menores de 25 anos	2008	-2
40041	Laza	Menores de 25 anos	2008	-2
54578	Xunqueira de Ambía	Menores de 25 anos	2009	-2

Como se puede observar se encuentran pequeñas diferencias en los años para los distintos municipios. La causa de que aparezcan estos errores sólo en los años viene determinado porque en el paro registrado los datos anuales son medias de los doce meses. Entonces, primero se calculan las medias de los meses para todas las posibles combinaciones y luego al evaluar la coherencia en las dimensiones nos encontramos con estas pequeñas diferencias en los datos anuales. Este es un problema general en este tipo de tablas que entroncaría con otro tema y es la calibración de los datos a los totales utilizando técnicas de programación matemática, que se alejan del objetivo de esta ponencia, pero que es un tema importante a tratar.

3.2.- Coherencia del último dato

Al igual que en el caso de la dimensión del espacio, la práctica totalidad de toda la información estadística que se publica tiene una base temporal, son datos referidos a un intervalo de tiempo y con una periodicidad determinados: anual, trimestral o mensual (en la mayor parte de los casos). Por tanto, como dimensión importante que es, es interesante estudiarla para controlar adecuadamente la calidad.

En el IGE, y en general en los institutos de estadística, se está continuamente añadiendo períodos de tiempo a nuestra información. Por tanto, el primer control que tendremos que hacer es verificar que el último dato añadido es coherente con los datos disponibles para períodos anteriores. En un primer momento se planteó calcular la tasa de variación (interanual, intertrimestral, intermensual,..) para todas las posibles combinaciones y establecer un umbral que no debería superar. Esto nos ocasionó bastantes problemas con datos inferiores a 100

unidades, por ejemplo, que fluctúan mucho y tienen tasas de variación que superaban los umbrales considerados. Desechada esta posibilidad, se echó mano de la teoría de las series de tiempo. Por tanto, la solución que se tomó fue:

- Ajustar la serie temporal para todas las posibles combinaciones disponibles en la tabla, sin el último período, utilizando la metodología ARIMA y los ajustes automáticos que tiene R. Se utiliza el paquete *seasonal* (Sax, 2017) y la función de R *auto.arima ()* que busca el modelo ARIMA más adecuado para la serie utilizando como criterio de selección el BIC (que penaliza el número de parámetros de la serie).

- Calcular el intervalo de predicción para el período siguiente (que es el que se va a publicar).

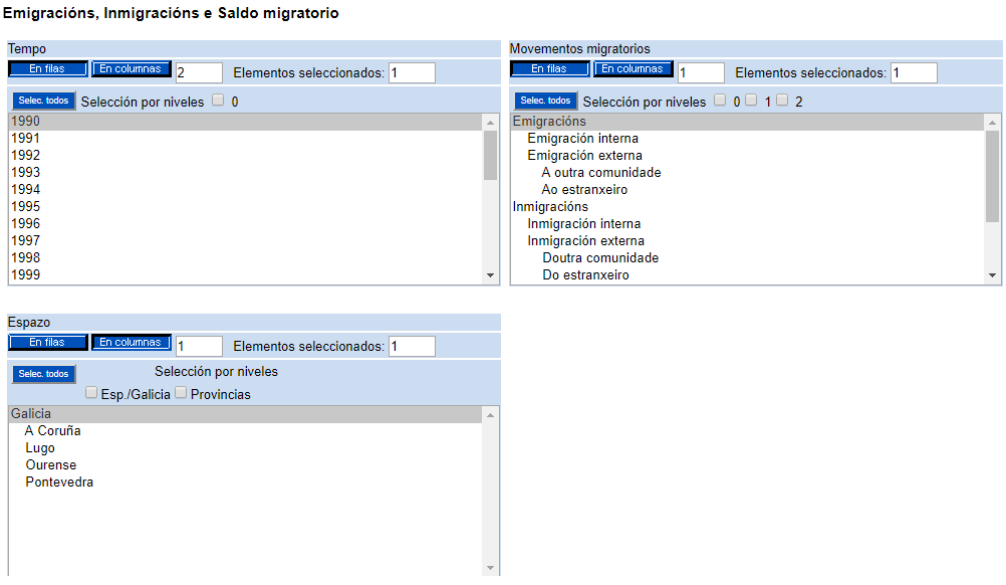
- Comprobar que el dato que se va a introducir está en el intervalo de predicción con un 95% de confianza

Al igual que para hacer el control de totales en esta ocasión también nos apoyamos en la librería "libhip". En este caso se construyó una función en R a la que se invoca aportándole dos argumentos, el código de la tabla y el lugar que ocupa la dimensión del tiempo

```
resultado=coherenciaUltimoDato(tabla, dimtempo)
```

Por ejemplo con la siguiente tabla disponible ahora en la web sobre *Emigraciones, inmigraciones y saldos migratorios para Galicia y provincias* (http://www.ige.eu/igebdt/selector.jsp?COD=1254&paxina=001&c=0204036001):

Figura 2.- Pantalla de selección de la tabla Emigraciones, inmigraciones y saldos migratorios para Galicia y provincias disponible en la web del IGE



Invocaríamos a la función con :

```
resultado=coherenciaUltimoDato(1254, 1)
```

donde, 1254 es el código de la tabla y 1 es el lugar que ocupa la dimensión del tiempo.

En este caso la función ajusta 75 series de tiempo (15 miembros en la dimensión Movimientos migratorios * 5 miembros en la dimensión del espacio)

La salida de esta función es un fichero con las series cuyo último dato está fuera del intervalo de predicción, se muestra un ejemplo para el año 2017 en la Tabla 2:

Tabla 2: Series cuyo último dato está fuera del intervalo de predicción.

	Espazo	Movementos migratorios	dato 2017	Predicción puntual	Lo.80	Hi.80	Lo.95	Hi.95
282	A Coruña	Emigración externa	11136	13170.777	12097.7634	14243.791	11529.7443	14811.810
285	Pontevedra	Emigración externa	11688	9094.000	8062.2049	10125.795	7516.0056	10671.994
1402	A Coruña	Saldo migratorio	5369	1387.018	-430.8897	3204.926	-1393.2321	4167.269
1405	Pontevedra	Saldo migratorio	-404	2409.638	1344.1498	3475.126	780.1144	4039.162

En la Tabla 2 se obtienen 4 series que están fuera del intervalo de predicción del 95% (Lo.95, Hi95). Estas serían series que habría que revisar por si hubiera errores o confirmar que el dato está correcto.

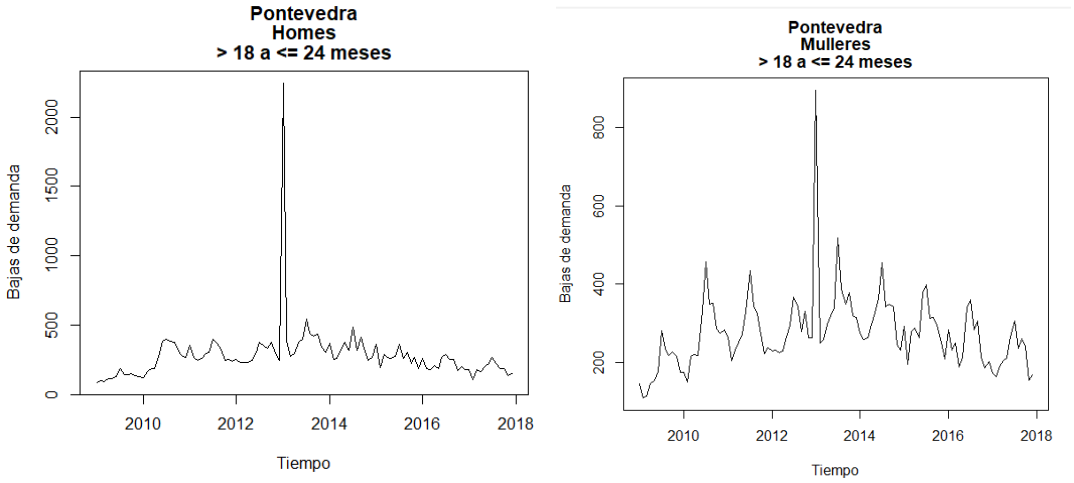
Esta herramienta además de ofrecer información sobre posibles errores también aporta información sobre cambios que es necesario resaltar en nuestra notas de prensa o resúmenes de resultados.

En este caso, el ejemplo expuesto sólo tiene 75 series y su ajuste es inmediato, el problema viene cuando es necesario ajustar series para 10.000 posibles combinaciones. En este caso la función necesita mucho tiempo para ofrecer resultados, con lo cual es necesario acudir a otras técnicas que nos ofrezcan resultados en un tiempo más razonable.

3.3.- Consistencia

Si se hubiera hecho el control del último dato desde que el IGE empezó a funcionar, el control de consistencia no tendría sentido. El caso es que partimos de una situación en donde ya tenemos una cantidad importante de datos y tablas publicadas a las cuales es necesario hacerles un control de consistencia de series, para detectar si hay datos erróneos, como los que se pueden ver en la Figura 3. En esta figura se observan dos series de tiempo sobre las *Bajas de demanda de empleo para la provincia de Pontevedra para la duración de la demanda de 18 a 24 meses y por sexo* y en ellas se presentan dos datos atípicos que, a priori, parecen erróneos.

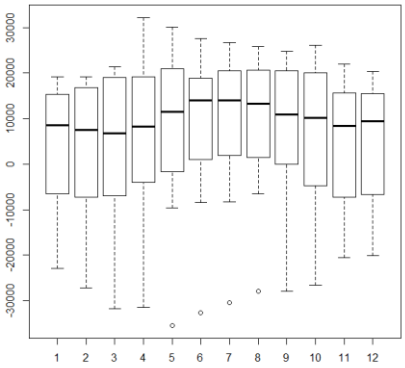
Figura 3: Series de bajas de demanda de empleo en la provincia de Pontevedra por sexo y para el grupo de duración de la demanda de 18 a 24 meses



Por tanto, como no partimos de una situación inicial, es necesario validar la coherencia de la información publicada en las tablas disponibles en la web del IGE. Lo haremos siguiendo la metodología que se expone a continuación y que tiene por objetivo detectar datos atípicos (outliers) en las series que están publicadas:

- Calcular las series de todas las posibles combinaciones de las dimensiones
- Determinar si las series son anuales o con períodos inferiores al año
- Calcular las series en primeras diferencias
- Para las series anuales se considera valor atípico todo dato que esté fuera del intervalo [Q1-3*(Q3-Q1), Q1+3*(Q3-Q1)], donde Q3 es el tercer cuartil y Q1 es el primer cuartil.
- Para las series con período inferior al año se evaluarán los boxplot estacionales (como el que se presenta en la Figura 4) y se buscarán los valores atípicos para cada mes. El criterio para detectar los valores atípicos es el mismo que para las series anuales, pero se aplicará para cada mes.

Figura 4: Boxplot estacional para la serie de afiliaciones a la seguridad social en diferencias



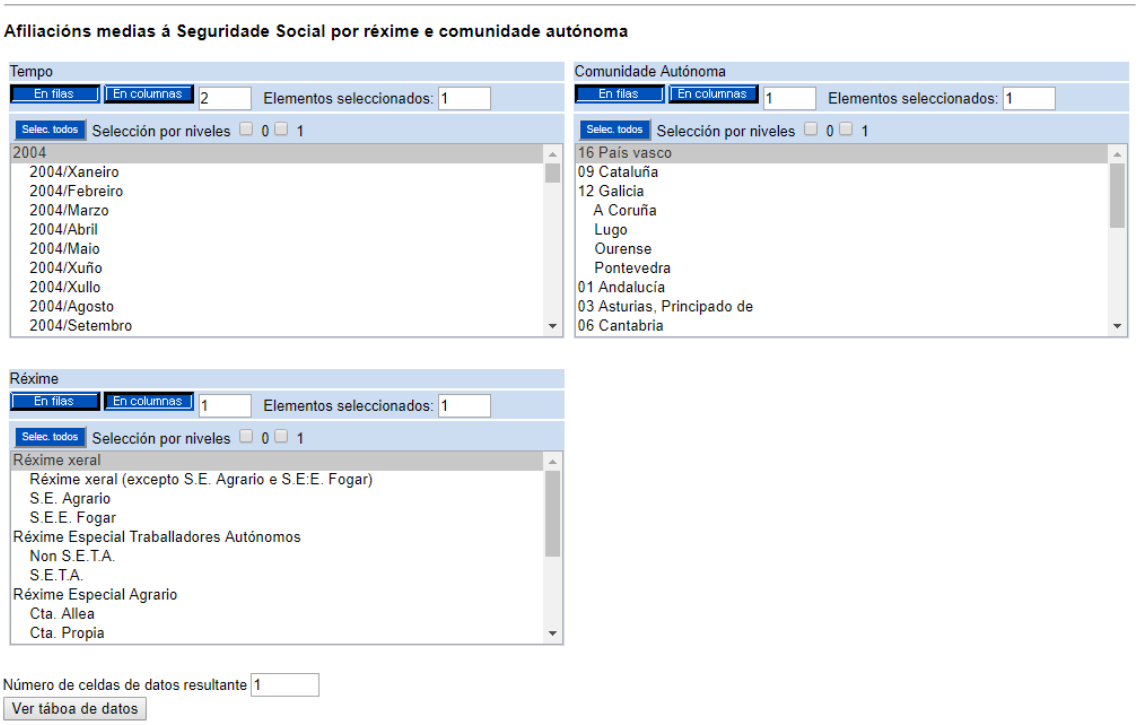
Al igual que en los casos anteriores en esta ocasión también nos apoyamos en la librería "libhip" desarrollada en el IGE. En esta situación se construyó una función en R, con nombre `consistencia`, a la que se invoca aportándole dos argumentos, el código de la tabla y el lugar que ocupa la dimensión del tiempo.

```
resultado=consistencia(tabla,dimtempo)
```

Por ejemplo si consideramos la siguiente tabla disponible ahora en la web sobre *Afiliaciones medias a la Seguridad Social por régimen y comunidad autónoma*

(http://www.ige.eu/igebdt/selector.jsp?COD=4854&paxina=001&c=0306002)

Figura 5: Pantalla de selección de la tabla Afiliaciones medias a la Seguridad Social por régimen y comunidad autónoma disponible en la web del IGE



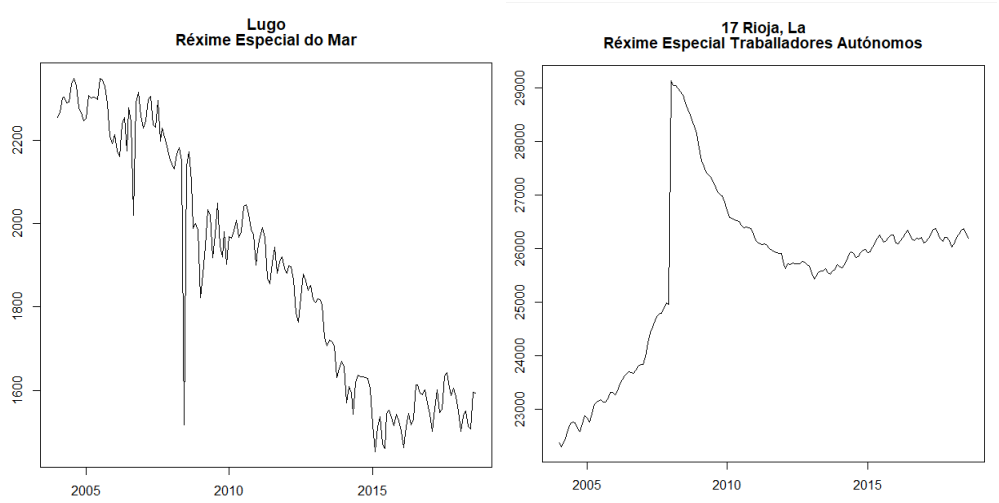
Para su control se invocaría a la función con:

```
resultado=consistencia(4854, 1)
```

donde 4854 es el código de la tabla y 1 es el lugar que ocupa la dimensión del tiempo.

La salida de esta función es un fichero con las series que tienen datos atípicos y los gráficos de estas series. Mostramos en este caso alguno de los gráficos:

Figura 6: Series afiliaciones al régimen especial del mar en Lugo (izquierda) y afiliaciones al régimen especial de autónomos en La Rioja (derecha)



En el caso de la Figura 6 (izquierda) que muestra la evolución de las afiliaciones medias a la seguridad Social en el Régimen del Mar en la provincia de Lugo parece que el dato de junio/2018 es un dato erróneo, aunque se ha revisado la fuente de información y nos ha confirmado que el dato es conforme a la fuente. En la figura de la derecha se muestra la evolución de las afiliaciones al régimen especial de autónomos en La Rioja, aunque el método detecta un atípico, en este caso, sabemos que está correcto debido a que hubo un cambio metodológico en esta estadística, de tal manera que el régimen especial agrario por cuenta propia se integró en el régimen especial de autónomos, de ahí ese cambio.

4.- Herramientas de Big Data

El control de coherencia del último dato nos conduce a hacer ajustes masivos de series de tiempo. En concreto, el problema que se plantea es como ajustar en un tiempo razonable entre 10 y 100.000 series de tiempo.

También es importante señalar que la aplicación mecánica de la metodología ARIMA puede conducir a la elaboración de modelos inadecuados que, en muchos casos carecerán de sentido

(Uriel, 1992). Por tanto, el proceso de identificación de los modelos ARIMA necesita supervisión y en esta situación es inviable hacerla, debido a la elevada cantidad de series que es necesario ajustar.

El problema que tenemos es similar, por ejemplo, al que tiene una empresa que necesita predecir la venta de un conjunto muy alto de productos y para una cantidad importante de centros de venta de esos productos. Otro aspecto importante de la tarea de predecir está representado por el número de períodos a predecir. Si la predicción de un período en una serie de tiempo ya es una tarea desafiante, realizar una previsión de múltiples períodos es más difícil debido a complicaciones adicionales, como la acumulación de errores, la precisión reducida y una mayor incertidumbre. Este segundo problema no lo tendríamos porque el objetivo en este trabajo es hacer la predicción para un único período.

En las últimas dos décadas, los modelos de aprendizaje automático se han empezado a utilizar mucho y se han convertido en serios competidores de los modelos estadísticos clásicos utilizados para predecir. Estos modelos, también llamados modelos de *caja negra* o basados en datos (*data-driven*), son ejemplos de modelos no paramétricos que usan solo datos históricos para aprender la dependencia estocástica entre el pasado y el futuro.

En este trabajo, por tanto, no podemos utilizar las técnicas tradicionales y se abre el horizonte del uso de este tipo de técnicas de aprendizaje automático que:

- Eviten la supervisión
- Se disminuya el tiempo de ajuste de las series

Una posible solución para este problema es acudir a la predicción de series de tiempo con aprendizaje automático. Para ello es necesario:

- Crear un modelo de aprendizaje automático para pronosticar datos
- Utilizar una técnica de predicción recursiva para la predicción de series de tiempo a corto plazo

La previsión de series de tiempo mediante el aprendizaje automático es más compleja que el aprendizaje automático estándar porque la componente temporal de los datos agrega una dimensión extra al problema. La metodología consiste en utilizar un conjunto (un periodo de tiempo) como entrenamiento y otro conjunto como test para validar el modelo. Para la validación del modelo se suele utilizar el Error cuadrático medio relativo (RMSE). Una vez que el modelo está validado se utiliza una técnica de predición recursiva para predecir los períodos siguientes.

En este momento se están evaluando diferentes técnicas de aprendizaje automático y una de las posibilidades que se abre es el uso del paquete “xgboost” (Chen et al., 2018) del software R.

4.- Conclusiones

En este trabajo se aborda el problema de la importancia de efectuar un buen control de calidad sobre la información que se difunde en un instituto de estadística. Se describen varias herramientas que hoy en día se están utilizando a modo de prueba en el IGE.

Desde que se ha empezado a utilizar estas herramientas se han detectado errores de control de totales, de consistencia y también en la inclusión del último dato, que tenían su origen, tanto en errores manuales de carga de información, como de fallos en los programas de automatización de la carga de la información.

Todos estos errores detectados refuerzan la idea de que es necesario realizar un control de calidad de la información antes de publicarla que asegure que los datos que se están publicando son correctos.

Por otra parte, también tenemos problemas que aun no hemos resuelto y que viene del ajuste masivo de series de tiempo, que esperamos que las técnicas de aprendizaje automático nos ayuden a resolver en el futuro.

5.- Referencias

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y, Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y. and Li, Y. (2018). xgboost: Extreme Gradient Boosting. R package version 0.71.2. https://CRAN.R-project.org/package=xgboost

Eurostat (2017). European statistics Code of Practice — revised edition 2017. Recuperado de http://ec.europa.eu/eurostat/web/products-catalogues/-/KS-02-18-142?inheritRedirect=true&redirect=%2Feurostat%2Fpublications%2Fleaflets

Gómez, J. (2018). Tablas multidimensionales en R. XX Jornadas de Estadística de las Comunidades Autónomas. Logroño

Iglesias, C. L y Arias, A. (1999). A aplicación de consulta de series no web do IGE: un instrumento para o estudio da conxuntura económica da C. A de Galicia. Boletín de Series estatísticas de Galicia Nº 47.

Iglesias, C. L. (2001). La nueva política de Difusión del Instituto Galego de Estatística. Jornades europees d’estadística. Palma

Joyanes, L. (2014). Big Data. Análisis de grandes volúmenes de datos en organizaciones. Alfaomega Grupo Editor.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for

Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Sax, C. (2017). seasonal: R Interface to X-13-ARIMA-SEATS. R package version 1.6.1. https://CRAN.R-project.org/package=seasonal

Uriel, E. (1992). Análise de series temporales. Modelos ARIMA