tecnim@p Sevilla 2006
30 mayo a 2 junio
060.es
24x7x365

Comunicación

194

UTILIZACIÓN DE LOS REGISTROS ADMINISTRATIVOS CON FINES ESTADÍSTICOS. UNA EXPERIENCIA EN LA SEGURIDAD SOCIAL (MUESTRA ANÓNIMA CONTINUA DE VIDAS LABORALES)

María Ángeles Sevilla Escribano
Coordinadora de Estadísticas
MTAS. Gerencia de Informática de la Seguridad Social

Palabras clave

Estadística, microdatos, registros administrativos, recopilación sistemática, muestra anónima, muestra continua, extracción, vidas laborales.

Resumen de su Comunicación

En la mayoría de los casos la toma de decisiones sobre temas sociales, como la ocupación, población activa..., se producen a menudo con información muy parcial, o apoyándose en costosas encuestas que solo pueden dar datos aproximados.

De una manera u otra la Seguridad Social recibe información del ciudadano sobre la empresa en la que trabaja, con qué salario (cotización), que tipo de contrato tiene, por cuanto tiempo, si pasa a percibir una pensión por viudedad, por incapacidad, etc.

El objeto de esta comunicación es exponer la experiencia de la Seguridad Social en la explotación estadística de sus registros administrativos, creando para ello una muestra de información anónima, proveniente de los datos que recoge del ciudadano.

La muestra de información es la recopilación sistematizada y anónima de la información que la Seguridad Social conserva sobre 1,1 millones de personas, desde que se afiliaron por primera vez hasta la actualidad y utiliza los registros administrativos como materia prima para elaborar microdatos que sirvan al estudio y la investigación.

Esta muestra de información es generalista, pues ha sido elaborada sin pensar en un trabajo en concreto, sino en las diferentes perspectivas de análisis social y económico que se manejan en la actualidad.

María Ángeles Sevilla Escribano

UTILIZACIÓN DE LOS REGISTROS ADMINISTRATIVOS CON INES ESTADÍSTICOS. UNA EXPERIENCIA EN LA SEGURIDAD SOCIAL (MUESTRA ANÓNIMA CONTINUA DE VIDAS LABORALES)

1. Introducción

La Sociedad Actual, demanda información-datos, pero cada vez se exige más que el dato tenga mayor precisión, es decir, corresponda a la realidad.. Estamos inundados de encuestas que revelan las incidencias y tendencias en problemas sociales, de empleo, etc. Pero las encuestas siendo un medio tan utilizado, dan una visión aproximada de la realidad. Es claro que la Administración Pública recoge una cantidad ingente de datos de los ciudadanos, desde muy diferentes perspectivas, empadronamientos, registros civiles, de inscripciones a la Seguridad Social, de declaraciones de renta, etc., y si esta información se pusiera a disposición de los estudiosos, se conseguiría que la visión ya no fuera aproximada sino real.

Por supuesto que ya se usan datos procedentes de Seguridad Social en muchos estudios. La cifra de afiliación, por ejemplo, es un indicador de primer orden en el seguimiento de la coyuntura económica. El número de pensiones y su cuantía figura en todos los informes sobre la situación de la tercera edad, y son la base de las proyecciones sobre la sostenibilidad del sistema. Pero estos estudios siempre se han realizado enfocándolo a un proyecto en concreto.

La Seguridad Social ha decidido crear un medio para que los registros administrativos se puedan utilizar para explotaciones estadísticas de microdatos, y que sea general para que sirva para cualquier estudio. La decisión viene respaldada con la tecnología y recursos informáticos que la propia Seg. Social dispone, ya que toda extracción de información cuando se realiza sobre un volumen enorme de datos, supone un coste en tiempo de proceso y en mips de máquina.

2. Fuentes de datos

¿Quién es la Seguridad Social? .

“La Administración de la Seguridad Social es el conjunto de órganos Administrativos y Organismos Públicos con facultades y competencias reguladoras, directivas, planificadoras o gestoras encaminadas a garantizar a los ciudadanos españoles, y en su caso, a los extranjeros residentes en nuestro país, el conjunto de prestaciones asistenciales, sanitarias, económicas o de otro tipo que las normas han establecido.” (Fuente ¿Quiénes somos?. www.seg-social.es).

La Seguridad Social (en adelante S.Social) para poder responder a las funciones descritas, dispone de un “Sistema de Información” complejo para automatizar los procesos que realiza en su gestión, apoyándose en un sistema de bases de datos corporativas para almacenar la información necesaria. Por otra parte el compromiso de llegar al ciudadano desde la mayor variedad de canales posibles, atención presencial, canal telefónico o el caso de la Oficina Virtual La Oficina Virtual disponible en la página de Internet (www.seg-social.es), ofrece al ciudadano la posibilidad de solicitar entre otros servicios, la petición de informe de vida laboral, un programa de auto cálculo de la pensión, ¿cómo va mi pensión?, etc., que permiten al ciudadano realizar gestiones desde su domicilio sin tener que desplazarse a ninguna oficina. Estos son servicios para el ciudadano pero hay también otros servicios como el Sistema RED (Remisión Electrónica de Documentos) que ofrece la Tesorería General de la Seg. Social a empresas, agrupaciones de empresas y profesionales colegiados, cuyo objeto es permitir el intercambio de información y documentos entre las distintas enti-

dades a través de medios telemáticos. Este servicio abarca los ámbitos de actuación como la Cotización, Afiliación y Remisión de partes de Alta y Baja de Incapacidad Temporal.

Esquema del sistema de información

El esquema anterior refleja un esquema de los flujos de información entre los diferentes colectivos relacionados con la Seg. Social, trabajador, empresa, pensionista.. Así pues se dispone de mucha información de cada persona que entra en el Sistema, por muy diversas circunstancias, pero dicha información, no se debe olvidar que, está siempre enfocada a mejorar el proceso y la identificación de esa persona desde el punto de vista de su gestión. La información que se recoge es del tipo de:

- En qué empresa trabaja, si es un trabajador por cuenta ajena o un autónomo,
- Con qué contrato, en que categoría profesional, desde que fecha, en que categoría, por cuanto tiempo, si tiene baja por incapacidad temporal,
- En algunos casos si percibe alguna prestación social, tiene hijos minusválidos, si es viudo/a porque percibe una pensión de viudedad.
- Cuando pasó de situación de activo laboral a jubilado, etc.

Las bases de datos operativas/transacionales residen en un host central, con un SGBD ADABAS. Pero además esta información se vuelca a un sistema paralelo (tipo datawarehouse de grano fino), que se actualiza con una cadencia de un día, que reside en un host con procesamiento paralelo y SGBD Oracle, para poder hacer consultas masivas al sistema y proceso estadístico sin interferir en el proceso diario transaccional.

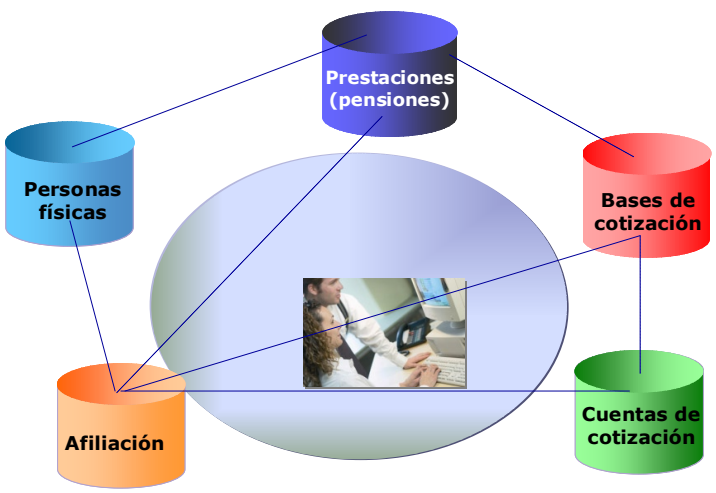
3. Cómo se ha construido la muestra

Cuando nació el proyecto de creación de una muestra continua de vidas laborales, que fuera lo más general posible para que pudiera servir a un extenso grupo de investigadores con diferentes intereses, a los responsables del Sistema Informático y a las personas que iban a formar parte de este grupo de trabajo (mixto entre gestores y estadísticos ajenos a la Administración de la Seg. Social), se les planteó un reto

importante, ¿Cómo seleccionar y adecuar la información que la Seg. Social tiene, para que la muestra sea útil y pueda servir para diferentes estudios?.

En primer lugar, hubo que seleccionar las variables de interés entre todas las que el sistema podía ofrecer y además establecer las relaciones entre ellas viendo como estaban organizadas. Lo primero fue construir un mapa conceptual de la información y sus relaciones.

Las variables referidas a las personas en la muestra se toman de cinco bases de datos permanentes de que dispone la Seguridad Social:



Esquema conceptual de las bases de datos corporativas y sus relaciones.

En nuestro caso, el mapa se dividiría en información de: Empresas, Trabajadores (vida laboral: actual e histórica), Datos personales (personas físicas), Recaudación (Bases de cotización), Vida de las prestaciones....

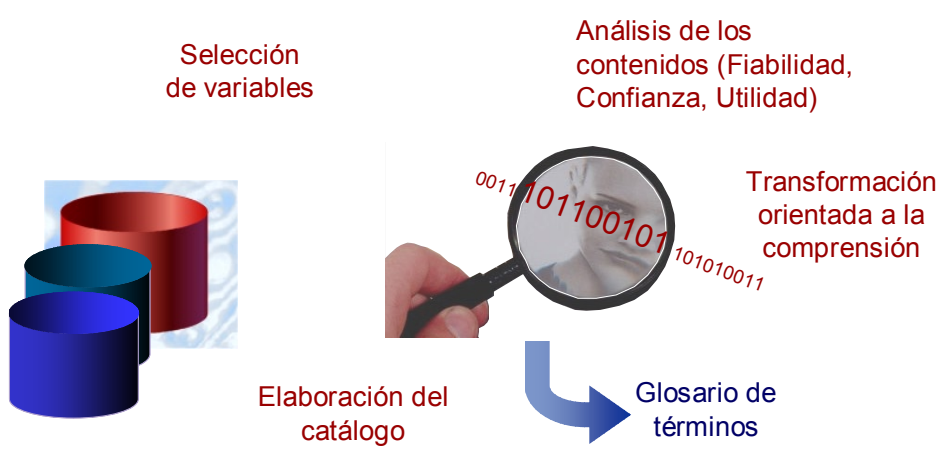
Se tenía como premisa la obtención de una muestra de población afiliada o percibiendo una prestación en cualquier momento del año 2004. Esto implica que había que tener en cuenta tanto la información actual como la histórica, puesto que una persona que estuviera en situación de alta al principio del año 2004, en el momento de la extracción podía continuar todavía en alta, estar en baja laboral y percibir una prestación o incluso estar en baja por haber fallecido.

Por otra parte había que elegir el identificador único de la persona en la muestra, ya que tiene que ser el mismo para las actualizaciones posteriores. Cuando una persona entra en el sistema de la Seg. Social, se le da de alta con un identificador propio del sistema, número de afiliación (NAF), que no coincide con el que estamos acostumbrados a manejar, que es el identificador de la persona física (NIF). La muestra estudia "personas", NIF's, que tienen relación con la Seg. Social en un momento dado, no NAF's. Se puede dar el caso de que una misma persona (NIF), tenga dentro del sistema varios NAFs, aunque lo normal es que solo tenga uno a lo largo de toda su vida laboral, o de su prestación. También se puede dar el caso de que a la hora de hacer la extracción una persona tenga varias relaciones laborales en el mismo periodo de tiempo, son los casos de pluriempleo (diferente régimen) y pluriactividad (mismo régimen pero diferente actividad económica). El identificador utilizado es el NIF, aunque, a la hora de reflejarlo en los datos que se facilitan, se le ha sometido a algoritmos para conservar el anonimato de la persona.

3.1. Selección de variables

De cada una de las personas que entran a formar parte del sistema de la S. Social se guardan una cantidad ingente de variables que son necesarias para la gestión, o que aún no lo siendo, se incluyen para recabar más información. El hecho de que se recoja información de una variable, es decir, que tengamos contenido, no significa que dicho contenido sea válido o en el mejor de los casos significativo. Pongamos por caso la provincia de nacimiento del afiliado, es un dato que se solicita en el alta en el sistema, pero para la gestión no es imprescindible, esto lo convierte por tanto en dato voluntario. Para los procesos de gestión no le supone ningún hueco, pero para los estudios de información de tipo estadístico implica un dato vacío en el que si el porcentaje de valores desconocidos es muy alto hace que se desprecie el dato como variable de estudio. Como este caso se dan otros muchos, sobre todo los que clasifican a la persona desde el punto de vista personal, como el domicilio de residencia, que en el caso del pensionista suele conocerse en un porcentaje prácticamente del 100%, pero no así en el caso del trabajador, en el que se puede conocer pero no se asegura que sea el último, ya que si no comunica su cambio de residencia en muchos casos podría ser el domicilio que tenía cuando se afilió por primera vez que incluso puede coincidir con el domicilio paterno.

Esta problemática se puede hacer extensible a todas las variables de estudio, y es por ello que fue necesario realizar un análisis previo de cada una de las variables de interés para determinar si una variable ofrecía la "confianza" suficiente para que las conclusiones que se puedan sacar de ella sean fiables o en cualquier caso significativas. Este proceso se ha realizado con cada una de las variables elegidas para la muestra, que en principio eran más de las que se incluyen en la misma, pero que por diversas circunstancias se han desechado, por lo menos en esta primera muestra.



Esquema de selección, transformación

Así se dieron dos pasos: Identificación de la información y conexión y estudio detallado de las variables. Ahora queda la traducción y conversión de los datos en legibles para los estudiosos ajenos a la Seg. Social. Para ello se ha elaborado un glosario de términos, en el que se hace una descripción exhaustiva de cada variable, anexando las tablas de posibles valores.... En algunos casos también fue necesario transformar o convertir los valores de las variables, como es el caso de la nacionalidad que ha sufrido conversiones que agrupan varios países, para una mejor organización de la información. En otros casos la transformación

se ha realizado para salvaguardar la confidencialidad del dato, como en el domicilio de residencia en el que se ha dado el valor completo del municipio siempre y cuando éste tenga una población mayor de 40.000 habitantes, y si no es así solo se ofrece la provincia.

3.2. Transformación y Confidencialidad

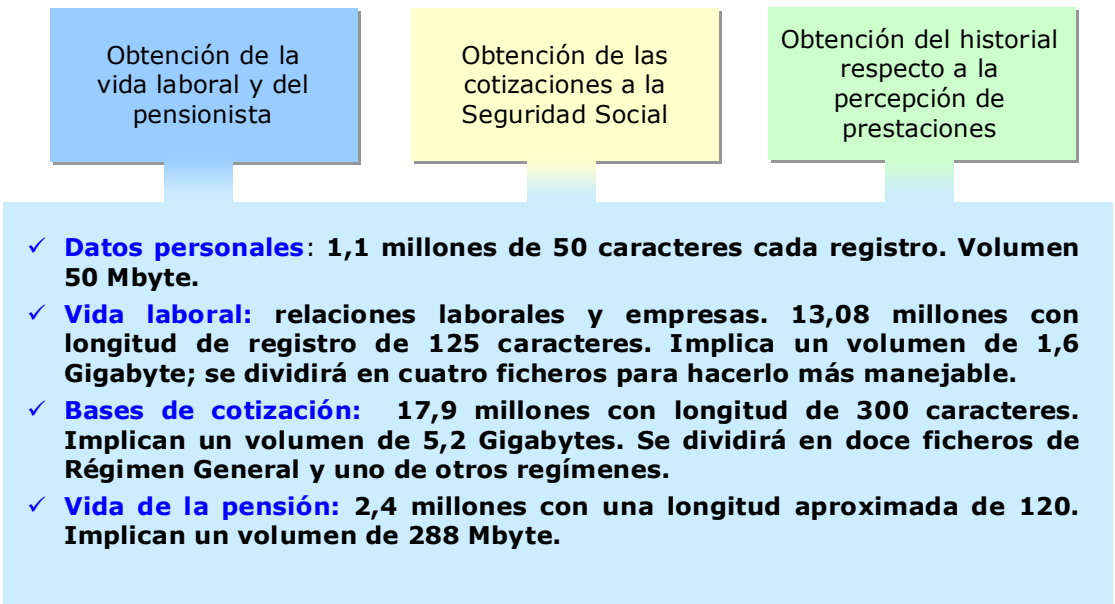
Las variables que necesariamente han sido transformadas, son los identificadores tanto de la persona como de la empresa, ya que como se ha citado antes hay que salvaguardar la identidad de las personas o empresas que han formado parte de la muestra y así asegurar que la muestra cumple con todos los requisitos de la Ley de Protección de Datos.

El sistema para mantener el anonimato no reside en almacenar los identificadores reales y los convertidos en una tabla y guardarla en un sitio seguro, sino que se ha optado por confeccionar unos algoritmos de conversión reversibles (más fáciles de conservar que tablas con millones de valores) de tal manera que con los datos de la muestra y la aplicación del sistema de conversión automáticamente se consigue el identificador en claro. Esto supone que en el caso de perder la fuente inicial de obtención de muestra siempre a partir de los resultados se podría continuar con las actualizaciones anuales, y así mantener el concepto de "muestra continua de vidas laborales". Por supuesto hay que prestar atención especial en mantener seguros los algoritmos de conversión, así como el algoritmo de selección de los individuos que forman parte de la muestra.

La transformación de los datos es en muchos casos imprescindible ya que en algunas variables el dato que se guarda es un código que por si mismo no tiene sentido sino va asociado a otro concepto y que al ser puramente de gestión, para los estudiosos puede resultar incomprensible o inútil.

3.3. Extracción

De los pasos descritos anteriormente queda por mencionar el "más penoso" en cuanto a tiempo de proceso de máquina, que es la extracción real de la información y su validación. En el caso de la Seg. Social, huelga decir que el volumen de información que se ha tenido que manejar para obtener la muestra ha sido enorme. Solo a título de ejemplo la información de vidas laborales supone alrededor de 510 millones de registros y la de bases de cotización contiene alrededor de 500 millones de registros con alrededor de 4.000 millones de cotizaciones. El proceso de obtención de la muestra ha supuesto primero buscar la población, seleccionar la muestra (1,1 millones de personas), el 4% de los que han tenido recientemente una relación con la Seg. Social, ya sea como cotizante en alta laboral, con prestaciones de desempleo o con un Convenio Especial, ya sea como beneficiario de pensiones contributivas y posteriormente completar la historia que cada persona tiene en la Seg. Social, tanto como afiliado como pensionista. Al final el conjunto de ficheros que comprende la muestra se ha grabado en un DVD y comprimidos ocupan alrededor de 1 Gigabyte.



3.4. Validación de los resultados

Una vez obtenida la muestra queda el proceso de validación de la misma. Esto no implica solo que las variables escogidas sean las correctas y se hayan colocado en el orden indicado sino que hay que ver que haciendo un seguimiento de un individuo dentro de la muestra la "historia" que la muestra nos cuenta acerca de esta persona sea comprensible y se saquen conclusiones que aunque puedan resultar sorprendentes responden a la realidad. Esta labor se ha realizado en dos fases, primero una validación técnica realizada por el equipo de desarrollo informático y otra realizada por las personas que formando parte del equipo de trabajo son ajenas a la Seg. Social y actúan como futuros estudiosos, de tal manera que pueden dar el punto de vista de usuario final de la información.

3.5. Entorno tecnológico

Como se citó en el apartado de fuentes de datos, la información residía en ADABAS y en Oracle. En este caso se ha optado por realizar todo el proceso de extracción de la información utilizando programación en PL/SQL sobre Oracle. Los procesos de transformación y elaboración de la muestra final se han realizado en entorno Host centralizado baja programación COBOL, manejando ficheros secuenciales transferidos desde los entornos abiertos, auque también se podría haber realizado todo el proceso en entorno Oracle con PL/SQL. La información resultante, se transfirió a ficheros tipo .txt para que pudieran ser accesibles por cualquier plataforma que pudieran utilizar los investigadores para tratar la información.

3.6. Conclusiones finales

Para la obtención de esta "Muestra Anónima y Continua de Vidas Laborales", se ha realizado un gran esfuerzo para darle una orientación polivalente y aislarla lo máximo posible de los conceptos de gestión para los que fueron concebidas cada una de las variables incluidas en la misma. Se ha reconstruido la vida en la Seg. Social en torno al concepto de persona, que no es el centro de identificación de la Seg. Social, el afiliado.

Se llama continua porque esta previsto actualizarla periódicamente, añadiendo la información nueva de las personas que ya están en la muestra y ampliando con personas que entren en contacto con la Seg. Social en períodos sucesivos.

Cabe esperar que según se vayan encontrando nuevas aplicaciones a la muestra, la relación entre el beneficio y el coste irá aumentando a medida que aumente el beneficio puesto que aunque hay que tener en cuenta el coste de la actualización este nunca será igual al coste inicial.

Por otra parte sería un gran éxito que esta experiencia animara a otras Administraciones y además ayudara a allanar el camino a la hora de intercambiar información con fines estadísticos entre las diferentes Administraciones.