PROYECTOS DEL ÁREA DE METODOLOGÍA DE EUSTAT PARA 2004

Cristina Prado, Marina Ayestaran y Elena Goni

EUSKAL ESTATISTIKA ERAKUNDEA
INSTITUTO VASCO DE ESTADISTICA

Donostia-San Sebastián, 1
01010 VITORIA-GASTEIZ
Tel.: 945 01 75 00
Fax.: 945 01 75 01
E-mail: eustat@eustat.es
www.eustat.es

PROYECTOS DEL ÁREA DE METODOLOGÍA DE EUSTAT PARA 2004

Cristina Prado, Marina Ayestaran y Elena Goni

IV Congreso Vasco de Sociología 2004

RESUMEN

En esta comunicación se tratará de presentar los proyectos programados en el Área de Metodología de Eustat para el año 2004, subrayándose aquellos más novedosos. Estos proyectos podrían agruparse de la siguiente forma:

- Proyectos más nuevos
- Proyectos I+D
- Proyectos europeos
- Proyectos relacionados con la calidad aplicada a la producción de datos

Indice

Capítulo
1

Los proyectos más nuevos

Fusión de registros

Introducción

Eustat convoca ocasionalmente y actualmente para períodos de 2 años 2 becas de formación e investigación en metodologías de la estadística oficial. El proyecto de fusión de registros ha sido desarrollado en el marco de una de estas becas, desde octubre de 2002 hasta el momento.

El problema de relacionar registros de dos ficheros diferentes y que correspondan a la misma unidad de población ó unidad económica es algo muy habitual dentro de la dinámica de trabajo del Instituto. De hecho ya existen aplicaciones ad-hoc, que aplican unas reglas deterministas que asignan unos pesos según la coincidencia ó no en unos tipos de variables determinadas (p.e variables tipo nombre, apellido, fecha de nacimiento, sexo, …).

Los métodos de fusión de registros pueden servir también para localizar duplicados en un mismo fichero; y para procesos de validación de la calidad de los censos ó de grandes ficheros de datos a partir de otros ficheros de datos más actualizados provenientes de otras encuestas.

Otra aplicación posible es la fusión de dos ficheros de datos a través de unas variables comunes para obtener un fichero con mayor cantidad de información, al permitir relacionar las variables no comunes en los ficheros originales.

Objetivo

El objetivo del proyecto era aplicar métodos basados en un modelo teórico probabilístico al problema de fusionar registros de dos archivos distintos y referidos a la misma unidad.

Desarrollo

El método utilizado se basa en el modelo teórico desarrollado por Fellegi & Sunter en 1969, aunque ya las bases habían sido puestas por el trabajo de Newcombe unos años antes. Posteriormente diferentes trabajos se han realizado a partir de ese modelo inicial.

El planteamiento del problema es el siguiente: dados dos ficheros se trata de descubrir aquellos pares de registros, cada uno del fichero respectivo, que se refieren a la misma unidad, esto es, determinar el conjunto de matches.

A partir de los registros obtenidos de los ficheros respectivos, concretamente de un conjunto de variables comunes en ambos, se enfrentan las configuraciones de los mismos dos a dos y se obtienen unos vectores de comparación. A cada uno de los estados de estos vectores de comparación se les asigna un peso, en función de unas

probabilidades estimadas de obtenerse esos estados tanto en el caso en que estamos ante un match verdadero como en el caso en que no sea así.

Se define una regla de fusión, en la que 3 decisiones son posibles: que el par comparado sea un link, que sea un no-link y que no se pueda determinar. Al aplicar esa regla de fusión se pueden cometer 2 tipos de errores: uno, el de decidir que es un link cuando realmente no lo es, y el otro de decidir que no lo es cuando realmente lo es. La decisión se adoptará con el objetivo de minimizar ambos tipos de error. Ello equivale a fijar unos límites a los pesos que determinan la decisión adoptada, esto es, dado un peso se adoptará una decisión u otra según su relación con los límites establecidos.

Son aplicados criterios de blocking con el objeto de reducir el ámbito de comparación entre ambos ficheros. Mediante estos criterios se seleccionan dos subconjuntos de los ficheros respectivos y son los registros de esos subconjuntos los que se comparan dos a dos. Se reduce así el tiempo de ejecución y aumenta la eficiencia del proceso.

Aplicaciones

Se han aplicado estas técnicas a la fusión de ficheros reales de Eustat, en concreto con registros referidos a personas y variables de tipo nombre, apellidos, sexo, fecha de nacimiento, etc. El primer caso fue la fusión de un fichero de matrimonios perteneciente al Movimiento Natural de la Población (MNP) con el fichero del Censo. Más del 90% de los registros de matrimonios se consiguieron fusionar con estas técnicas, cuando con las técnicas deterministas se venía consiguiendo la fusión de menos del 50% únicamente.

Otro caso práctico ha sido la fusión de un fichero de la Encuesta de Población en Relación con la Actividad (PRA) con el Censo. Este proceso sirve para obtener unos indicadores de calidad del Censo, y puede sustituir quizá a la Encuesta de Validación posterior al Censo y que tiene esa misma finalidad. Los resultados han sido igualmente buenos, con una fusión de registros próxima al 96%.

Conclusiones

Este método calcula unos niveles de error del proceso de fusión realizado, lo cual no es posible con métodos deterministas.

Hay que hacer una serie de asunciones a veces difíciles de estimar, en relación a los niveles de error aceptables y las probabilidades de error en los valores de las variables.

En la práctica se han obtenido muy buenos resultados.

Queda sin tratar aún en detalle el problema de utilizar para la fusión tipos de variables como direcciones, de gran interés para el problema de fusionar registros referidos a unidades económicas.

Estimación en áreas pequeñas

Introducción

La creciente demanda de información referida a áreas ó ámbitos pequeños ha dirigido la atención en los últimos años al desarrollo de metodología para la obtención de estimaciones en áreas pequeñas.

Los estimadores directos tradicionales no son adecuados debido a la falta de suficiente muestra en el dominio requerido. Por ello se han de utilizar métodos indirectos que se basan en información de otras áreas para construir los estimadores. Estos estimadores indirectos se basan en modelos, implícitamente en los métodos tradicionales (métodos demográficos, estimadores sintéticos, …) y de modo explícito en los métodos desarrollados en los últimos años.

Este proyecto se viene desarrollando en colaboración con el Departamento de Estadística e Investigación Operativa de la Universidad Pública de Navarra (UPNA), colaboración prevista hasta el año próximo.

Objetivos

El objetivo general es la aplicación de la metodología de estimación en áreas pequeñas en alguna de las operaciones estadísticas del Instituto.

Desarrollo

El proyecto tendrá varias fases. La primera fase será de formación de parte del personal de Eustat en estas metodologías, a través de diferentes sesiones impartidas por parte de la UPNA y estudio del material proporcionado. En esta fase también se empezarán a programar los diferentes estimadores estudiados así como sus varianzas, tomando como ejemplo los datos reales de la Encuesta Industrial de Eustat.

Otra fase consistirá en el análisis de los resultados obtenidos del proyecto Eurarea y la evaluación de su aplicabilidad a los datos de Eustat. La fase final será la aplicación práctica de estas técnicas en un caso de estimación en áreas pequeñas con datos reales del Instituto.

Aplicaciones

Una aplicación inicial del proyecto es la construcción de modelos para obtener estimaciones a nivel comarcal de las principales variables de la Encuesta Industrial junto con estimaciones de sus coeficientes de variación.

Una aplicación posterior y previsiblemente más compleja (debido por un lado al diseño multietápico de la propia encuesta así como al hecho de tratarse de una variable binaria la que está en juego) es la de obtener estimaciones de empleo a nivel comarcal a partir de la Encuesta de Población en Relación a la Actividad (PRA).

Conclusiones

La aplicación de estos métodos en la obtención de estimaciones para pequeñas áreas abre un importante área de estudio en la estadística oficial. Su importancia resulta clara debido al cada vez mayor requerimiento de información y a la necesidad de no incrementar los costes de obtención de esa información.

La colaboración en este proyecto entre la estadística oficial y la estadística académica abre también futuras vías de colaboración, sin duda de gran interés para ambas partes.

Capítulo
2

Los proyectos I+D

Son los que incluyen alguna innovación o mejora en los métodos. Normalmente son proyectos que también se llevan a cabo en otros institutos europeos, debido a su directa influencia en la mejora de la calidad de las estadísticas.

El ajuste de muestras a la información auxiliar para el cálculo de elevadores, pesos o calibración.

Introducción

Cuando trabajamos con un fichero proveniente de una encuesta hay que calcular unos pesos para los casos encuestados que son los que nos permiten "elevar" los resultados a la población. Como primera fase, se calculan los pesos iniciales de cada uno de los casos o registros, que resultan ser la inversa de la probabilidad de selección en la muestra.

El inconveniente es que normalmente, la suma de esos pesos en la muestra no da como resultado el total poblacional conocido a partir de información externa a la propia encuesta. Un modo de resolver esto es acudir al ajuste o calibración de la muestra, o de los pesos, más propiamente, calculando para cada caso otro segundo peso, basado en y muy similar al anterior. Esto nos permite reflejar en la muestra los datos de la población (en sentido amplio) que ya conocemos, por ejemplo, la distribución de sexo y edad, la población de los TH, la distribución por sectores de los establecimientos, etc.

Objetivo

Obtener los pesos finales que nos permiten reflejar en la muestra información de la población ya conocida y que mejoran los pesos iniciales, que provienen de la probabilidad con la que se han seleccionado. Este proyecto se refiere principalmente al método de calibración cuando no disponemos de una información auxiliar de nivel máximo, es decir, cuando solo conocemos los marginales de las distribuciones en cuestión, por ejemplo, la distribución por sexo, por un lado, y la distribución por edad, por otro, y no la distribución conjunta.

Desarrollo

En los últimos años, se ha venido haciendo este trabajo con otras herramientas. Dependiendo del entorno informático, se ha venido utilizando el programa SPAD (procedimiento REDRE) para PC, y un programa en Pascal para grandes sistemas.

Actualmente, se está trabajando con la macro CALMAR, que es una macro SAS programada para esta finalidad por el INSEE (Instituto de Estadística Francés), con la ventaja principal de que funciona en el entorno SAS PC, y además permite ajustar a

variables numéricas y cualitativas al mismo tiempo, así como hacer ajustes simultáneos a varios niveles.

Qué es lo que hace la macro CALMAR?. Por métodos iterativos calcula los pesos finales para cada uno de los casos, según sus características, teniendo en cuenta los pesos iniciales (inversa de la probabilidad de inclusión en la muestra). Para este cálculo de nuevos pesos, se plantean las ecuaciones que deben satisfacer los pesos nuevos, de acuerdo con las marginales introducidas, con la condición de que sean lo más próximo posibles a los pesos iniciales.

Aplicaciones

Se aplica principalmente a las encuestas de tipo sociodemográfico, como la Encuesta de la Sociedad de la Información – Familias y la Encuesta de Presupuestos de Tiempo. Para estas encuestas se realiza un muestreo de viviendas en primer lugar, se calculan los pesos iniciales y después se hace el ajuste o calibración a la información auxiliar, ya sea a la población proyectada por TH, o a la proyección junto con otras variables, como la distribución de la población en las comarcas, proveniente del CPV2001.

Conclusiones

La macro en su versión actual ofrece más posibilidades que todavía no estamos utilizando como es el ajuste simultaneo a varios niveles, por ejemplo, de población y familias.

Una versión posterior próxima a difundirse tendrá otros aspectos en cuenta, como el ajuste debido a la no-respuesta.

El diseño y cálculo de errores muestrales para los principales resultados de las operaciones.

La mayoría de las operaciones estadísticas de Eustat, en particular las pertenecientes al Área Socio-Demográfica, siguen un diseño muestral complejo: en dos etapas, por clusters, etc. Por este motivo, los programas convencionales no hacen estimaciones del error.

En años anteriores se impulsó el cálculo del error de muestreo en estas encuestas, que se lleva a cabo por un método de replicación, concretamente por el Jacknife n) con el software americano WesVar. Este método consiste en extraer submuestras de la muestra total, para calcular el estadístico de interés en ellas. La estimación de la varianza de la muestra se realiza utilizando la variabilidad de estos estadísticos.

El proyecto en la actualidad consiste en conseguir la disponibilidad de esta información de forma generalizada en estas encuestas, siguiendo este método, por un lado; y el cálculo de los errores de muestreo en otras operaciones del área económica, que normalmente siguen un diseño de muestra estratificado con otras peculiaridades.

En este momento, se están estudiando los estimadores más próximos a los que se llevan a cabo en la práctica para los establecimientos muestrales de la Encuesta

Industrial, y sus correspondientes cálculos de la varianza. Este proyecto se está viendo muy ayudado por el de áreas pequeñas, ya comentado.

Otros proyectos de I + D: los métodos automáticos de imputación para variables cuantitativas y cualitativas

La imputación de la información en un tema de interés central en todos los institutos de estadística. Del mismo modo, estos proyectos están presentes en el Plan Vasco de Estadística 2005-2008, dando continuación a los ya existentes en Planes anteriores.

Se trata de hacer una revisión continua de los métodos de imputación utilizados en las operaciones, contrastarlos con los métodos propuestos y recomendados por Eurostat y medir su influencia en la calidad de los datos ofrecidos. La evaluación de la variabilidad introducida por la imputación dependerá del tipo de diseño de la encuesta y del método de imputación empleado, por lo que se requiere hacer un análisis de cada caso específico.

En este año 2004 están programados trabajos de imputación en las operaciones de Investigación y Desarrollo, por un lado, y en la Encuesta de la Construcción.

Capítulo

3

Los proyectos europeos

En los cuales el Instituto puede tener una participación directa o indirecta. Actualmente están a punto de finalizar los dos siguientes.

El proyecto ASSO, de análisis de datos simbólicos oficiales

Este proyecto ASSO (Analysis System of Symbolic Objects) surge como continuación de una iniciativa enmarcada en 4º Programa Marco Europeo (999-1996) y del que también formaba parte EUSTAT: el proyecto SODAS. El objeto de ambos proyectos consiste en el desarrollo de técnicas estadísticas para el análisis de datos complejos. Es lo que se denomina Análisis Simbólico de Datos y supone una extensión del análisis estadístico clásico.

Los 'objetos' simbólicos son información agregada que representa a grupos de individuos o unidades de la población, que van a ser tratados como una nueva unidad estadística. Cualquier técnica estadística en el análisis clásico puede tener su correspondiente versión simbólica.

Los ámbitos de aplicación de estas técnicas dentro de la estadística oficial son muy variados. Desde la mejora de la calidad en determinadas fases de la producción estadística, hasta la fusión de encuestas, pasando por la preservación de la confidencialidad y la posibilidad de tratar con pesos muestrales e intervalos de confianza.

Dentro de este proyecto se desarrolla un software modular específico para este tipo de análisis: Sodas 2.5. El objetivo final es la implementación y normalización del uso de esta herramienta como una parte más del proceso de producción y análisis de datos en la estadística oficial.

El proyecto CASC, de protección de datos y confidencialidad estadística

Al igual que el anterior proyecto, CASC (Computational Aspects of Statistical Confidentiality) es la continuación de un anterior proyecto englobado en el 4º Programa Marco Europeo. EUSTAT ha mantenido el contacto con este proyecto desde su inicio y actualmente trabaja de forma conjunta con el Instituto de Estadística de Cataluña (IDESCAT), participante oficial del proyecto, a través de un convenio de colaboración.

La necesidad de preservar el secreto estadístico ha llevado al desarrollo de técnicas y métodos de protección de datos que impiden la identificación directa de los individuos o entidades presentes en una encuesta, ofreciendo a la vez a los usuarios e investigadores una información lo suficientemente desagregada y detallada. En el proyecto CASC estás técnicas se implementan en el software modular ARGUS, capaz de proteger de forma eficiente tanto tablas como ficheros de microdatos.

La introducción de este software como paso necesario en la producción de estadísticas y previo a su difusión, es de nuevo el objetivo principal a conseguir. Diversas aplicaciones con este software se han realizado ya en el último Censo de población.

Capítulo

4

Proyectos relacionados con la calidad aplicada a la producción de datos

Los indicadores de calidad en la producción estadística y la creación de una base de datos documental de operaciones estadísticas

Como continuación del Plan de Calidad impulsado por el Instituto para sus operaciones, en este año se va a trabajar en los indicadores de calidad y en la creación de una base de datos documental de operaciones.

En un primer momento se trabajó en la creación de un documento uniforme para todas las operaciones, el proyecto técnico, cuyos apartados fundamentales son: la documentación, los factores de calidad de la operación a medir con los indicadores respectivos y las responsabilidades en todas las etapas ó procedimientos de la operación.

Por un lado, la parte relativa a la documentación supone la recopilación de todos los documentos necesarios para la operación y su referencia exhaustiva en el proyecto técnico. Se archivan en un lugar común y quedan registrados en una base de datos documental construida al efecto.

Por otro, los factores de calidad y la fijación de sus indicadores es un tema clave que luego será objeto de revisión más exhaustiva por los coordinadores de las operaciones.

Los indicadores ya definidos y susceptibles de ser adoptados para todas las operaciones en los que sean aplicables son: el tiempo de elaboración del resultado, el coste monetario, del desfase temporal (fecha de referencia y fecha de publicación), el número de accesos a la Web, las repercusiones en los medios, el número de peticiones de datos, la comparabilidad con resultados de Eurostat o del Ine, la satisfacción de los clientes, el error de muestreo de los principales resultados, el tiempo de cumplimentación del cuestionario, la tasa de imputación y la tasa de cobertura de la encuesta.

El plan de formación dirigida a la Organización Estadística Vasca,

En la actualidad, se está llevando a cabo el Plan de Formación 2003-2005, elaborado por Eustat y dirigido al propio Instituto y a la Organización Estadística Vasca. En este Plan se establecen sus objetivos, que son: definir un marco de conocimientos mínimos del personal de Eustat; analizar las necesidades de formación en función de los mínimos anteriores; establecer los medios necesarios para su consecución; establecer las acciones formativas y conseguir la adaptación del personal a un entorno tecnológico cambiante y que posibilite la realización de un trabajo de mayor calidad.

La formación podrá impartirse a través de las siguientes modalidades: cursos organizados por Eustat o el IVAP, autoaprendizaje, talleres, reuniones de expertos y cursos no organizados por Eustat y solicitados a demanda por su personal.

Los Seminarios Internacionales de Estadística.

Eustat viene organizando estos seminarios a fin de acercar expertos internacionales y los más recientes desarrollos en metodologías estadísticas a nuestro entorno.

El Seminario que está programado para la próxima sesión, que va a tener lugar en abril, es una introducción a la Estadística Bayesiana. Lo va impartir el profesor Jose Miguel Bernardo, de la Universidad de Valencia, que es una autoridad internacional en esta materia.

También se está ultimando el seminario que va a tener lugar en otoño. Un tema que se está barajando ahora mismo con más posibilidades es el de la recogida de datos, las nuevas tecnologías y su influencia en la calidad de los datos.

Capítulo

5

Referencias

Se puede encontrar más información de algunas de estas materias en la página de Eustat, entre otras:

- PATRICIA CALVO, CRISTINA PRADO, YOLANDA PÉREZ, MARINA AYESTARÁN, Creación de objetos simbólicos a partir de encuestas almacenadas en bases de datos relacionales http://www.eustat.es/document/datos/ct_ME_c.pdf

- MARTA MAS, Estudio comparativo para diferentes umbrales en tablas de frecuencias: un ejemplo de funcionamiento de t-Argus, http://www.eustat.es/document/ct_3_c.html

- E. BUENO, A. ZARRAGA y A. IZTUETA, Ajuste de muestras con información auxiliar, http://www.eustat.es/document/datos/ajuste1_c.pdf

- JESÚS MANCHO CORCUERA, Técnicas de estimación en áreas pequeñas http://www.eustat.es/document/datos/ctjesús_c.pdf

- AITOR PUERTA GOICOECHEA, Imputación basada en árboles de clasificación, http://www.eustat.es/document/datos/ctaitor_c.pdf