Desarrollo de una base de datos integrada de Censo y encuesta mediante el uso de elementos de inteligencia de negocios y SIG

A development of census and survey integrated database using elements of business intelligence and SIG

Robert Cornejo[1] Mónica Navarrete[2] Ricardo Valdivia[1]
Patricio Aroca[3] Sebastián Aracena[1]

Recibido 23 de mayo de 2013, aceptado 13 de enero de 2014
Received: May 23, 2013 Accepted: January 13, 2014

RESUMEN

En este trabajo se plantea el desarrollo de una solución de Inteligencia de Negocios y Sistemas de Información Geográfica (SIG) para gestionar integradamente los datos generados por el Censo Nacional de personas, hogares y viviendas y la Encuesta de Clasificación Económica y Social (CASEN) como una alternativa a los métodos actuales de estimación en área pequeña (SAE) utilizados para obtener indicadores desagregados de bienestar y que estiman, por ejemplo, el ingreso a partir de los atributos del hogar, ignorando la ubicación geográfica de las observaciones de la encuesta.
En este trabajo proponemos un cambio en la lógica tradicional implícita de estos métodos, al estimar el ingreso medio en áreas pequeñas georreferenciando las observaciones de la encuesta mediante el método de emparejamiento conocido como *Matching Estimator* y luego extrapolando los datos con la técnica de predicción del Kriging. La propuesta plantea el uso de una base de datos integrada de ambas fuentes mediante tecnología ETL (*Extract, Transform and Load*), permitiendo obtener datos de la encuesta en niveles de desagregación que esta no provee originalmente y que le son transferidos por el emparejamiento con los datos del Censo. Utilizando los códigos de localización espacial incorporados a un SIG, se añaden herramientas de visualización cartográfica que facilitan la observación y análisis de las relaciones espaciales entre las unidades geográficas, así como la observación y análisis de las particularidades en áreas pequeñas. Como aplicación, se describe la desagregación espacial del ingreso per capita de los hogares en las Regiones XIII, VI y VII de Chile.

Palabras clave: Emparejamiento espacial, tecnología ETL, bases de datos espaciales, sistemas de información geográfica, servidores de mapas.

ABSTRACT

In this paper, we propose a solution development of Business Intelligence and Geographic Information Systems (SIG) to an integrated management of information generated from Census of population, households and dwellings, and the Survey of Economic and Social Classification (CASEN). This, as an alternative to current methods in small area estimation (SAE) that are used to obtain disaggregated welfare indicators and estimate, for example, the income from the household attributes, ignoring the geographical location of the observations in the Survey.

[1] Escuela Universitaria de Ingeniería Industrial, Informática y Sistemas. Universidad de Tarapacá. Casilla 6-D. Arica, Chile. E-mail: robert.cornejo.uta@gmail.com; rvaldivi@uta.cl; supersbx00@gmail.com
[2] Escuela Universitaria de Administración y Negocios. Universidad de Tarapacá. Casilla 6-D. Arica, Chile. E-mail: mnavarre@uta.cl
[3] Escuela de Negocios. Universidad Adolfo Ibáñez. Av. P. A. Hurtado 750. Viña del Mar, Chile. E-mail: patricio.aroca@uai.cl

In this paper, we propose a change in the traditional implicit logic in these methods, to estimate the average income in small areas, georeferencing the survey observations using the matching method known as Matching Estimator and then extrapolating data using the Kriging prediction technique. The proposal suggests the use of an integrated database of both sources allowing to obtain data from the survey of economic characterization at levels of disaggregation not originally provided and are transferred by pairing with census data. Using spatial location codes incorporated into a SIG, cartographic visualization tools are added, which facilitates the observation and analysis of spatial relationships among geographic units, as well as observation and analysis of particularities in small areas. As an application, we describe the per capita spatial disaggregation income of households in Regions XIII, VI and VII of Chile.

Keywords: Spatial matching, ETL technology, spatial databases, geographic information systems, map servers.

INTRODUCCIÓN

La información contenida en los Censos Nacionales y en las Encuestas de Hogar ha sido utilizada para la obtención y análisis de indicadores socioeconómicos que definen los niveles de pobreza o riqueza de los hogares [1] y buscan orientar el diseño de políticas públicas de superación de la pobreza, por ejemplo. Los métodos que mezclan información del Censo y Encuestas de hogar da cuenta de al menos tres variantes entre las cuales se destaca "The Empirical Best Predictor (EBP)", de Molina y Rao [2], "The M-quantile approach Chambers y Tzavidis" [3] y "The World Bank Method ELL Elbers" [4], siendo este último utilizado en más de 40 países para construir mapas de pobreza en pequeñas áreas [4-5]. ELL realiza mediante un modelo de predicción espacial para datos de encuesta utilizando un conjunto de otras variables disponibles también en el Censo. La estimación de la distribución conjunta se utiliza posteriormente para extrapolar la distribución de la variable de la encuesta, a cualquier subconjunto de los datos del Censo. Sin embargo, pese a la utilidad práctica del método ELL, hay algunas dudas en su aplicación y que en el caso particular de Chile quedan sin resolver: ¿Cómo opera el método ELL cuando hay áreas geográficas no cubiertas en la encuesta? ¿Cómo recoge el método ELL la ubicación de las observaciones y las interacciones entre ellas? Respecto de lo primero, nada hay al respecto, siendo este uno de los puntos abordados por nuestra propuesta. De lo último, recientes aplicaciones en SAE cuestionan las condiciones de homogeneidad y tratamiento del espacio que el método ELL contiene implícitamente [6].

En este trabajo se plantea el desarrollo de una solución de Inteligencia de Negocios [7] y Sistemas de Información Geográfica [8] para integrar datos del Censo y encuesta mediante el uso de tecnología ETL (*Extract, Transform and Load*), con el propósito de ofrecer datos de la encuesta en niveles de desagregación que esta no provee y que le son transferidos por información del Censo. La extrapolación de datos de encuesta en niveles microterritoriales se realiza mediante emparejamiento espacial y finalmente se hace uso de servidores de mapas web para su observación y análisis espacial. La integración y desagregación espacial de los datos permitirá analizar la información geográfica referenciada de la encuesta, con el fin de construir no solo indicadores de pobreza y/o riqueza, sino el uso de una base de datos desagregada espacialmente y con contenido socioeconómico de tantas variables como posea. Como aplicación, mostramos la desagregación espacial del ingreso per cápita de los hogares en las Regiones XIII, VI y VII de Chile. En el apartado de metodología se exponen las herramientas tecnológicas y econométricas utilizadas, mientras que en el apartado de aplicación se representa la cartografía microterritorial del ingreso per cápita de los hogares en las Regiones XIII, VI y VII de Chile. Las conclusiones se presentan al final de este artículo.

METODOLOGÍA

La solución propuesta utiliza herramientas de ETL [9], Inteligencia de Negocios [7], los Sistemas de Información Geográfica [8] y de econometría espacial. El proceso ETL es una tecnología de integración de datos que se utiliza en proyectos de implantación de Inteligencia de Negocios, el que

permite extraer datos alojados en diversas fuentes de información, transformarlos según las necesidades del analista y cargar estos en los entornos de destino [7, 9], siendo una de las tareas más significativas, el diseño y construcción de los almacenes de datos o Data Warehouse, conocidos como "una colección de datos orientados a un ámbito (empresa, organización), integrada, no volátil y variante en el tiempo, que ayuda al proceso de los sistemas de soporte de decisiones" [10], mientras que un Sistema de Información Geográfica brinda una representación gráfica de la información geográfica referenciada. Estos sistemas se basan en principios formales de matemáticas discretas, modelos de datos y geometría computacional; su desarrollo involucra nuevas tecnologías de la información: estándares e ingeniería de *software*, almacenes de datos, servidores web, metadatos, ambientes y lenguajes visuales, entre otros [7]. La mayor utilidad de un Sistema de Información Geográfico está intrínsecamente relacionada con la capacidad que este posee de construir modelos o representaciones del mundo real a partir de bases de datos, aplicando una serie de procedimientos específicos que generan información para el análisis [11].

Desde el punto de vista econométrico, la solución plantea la elaboración de un emparejamiento (o *matching*) espacial, el cual busca extrapolar los datos de la encuesta en niveles microterritoriales asignados por Censo. Este emparejamiento nace como resultado de la generación de la base de datos relacional y contemporánea entre el Censo 2002 y CASEN 2003, los cuales tienen diferentes niveles de agregación espacial y niveles de información. Por otro lado, debido a que la encuesta no cubre todo el territorio, utilizamos la técnica del Kriging (perteneciente a la Geoestadística), que permite extrapolar la información contenida en un punto a otros puntos cercanos utilizando la estructura de la dependencia espacial contenida en la información de la variable, cuyo valor se explica por su ubicación espacial. La Figura 1 muestra el diagrama de los procesos y subprocesos realizados en el desarrollo de esta propuesta, donde la construcción del modelo relacional entre ambas bases de datos cobra vital importancia.

MODELO RELACIONAL DE DATOS

Las bases de datos Censo 2002 y CASEN 2003 (en archivos SPSS), disponibles en las fuentes oficiales, carecían de la documentación respectiva del modelo que las sostiene, por cuanto se realizó un proceso de ingeniería inversa para la reconstrucción del modelo relacional, el que se obtuvo a partir del estudio de

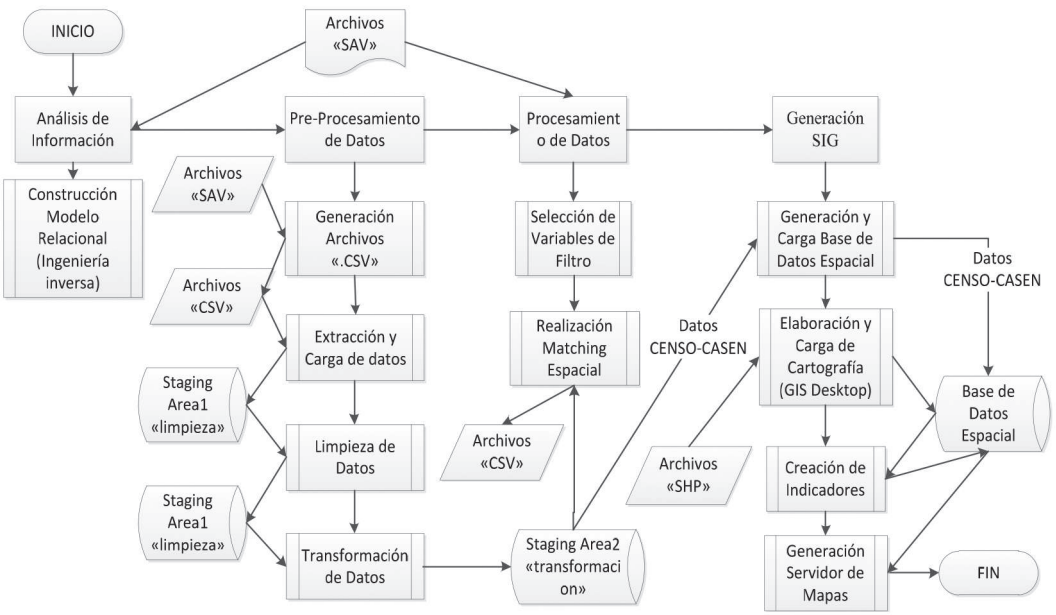

Figura 1. Diagrama de flujo de los procesos y sub-procesos.

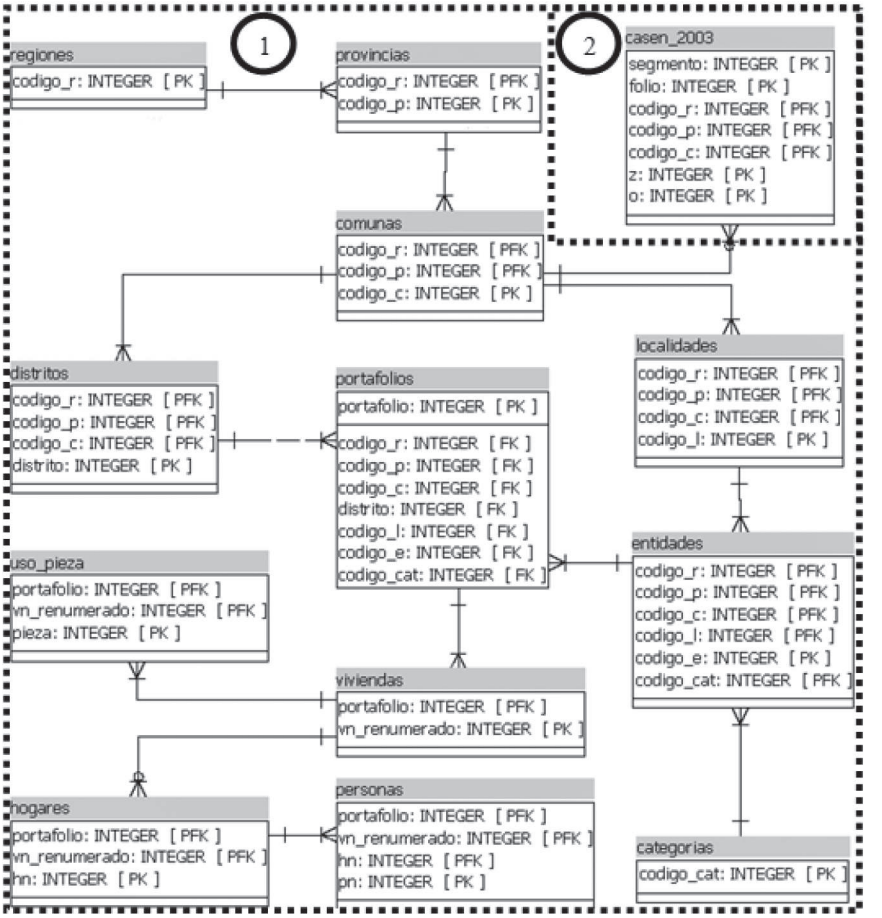

Figura 2. Modelado de datos relacional entre Censo 2002 y la encuesta CASEN 2003.

la semántica de la información contenida en cada base, respectivamente. Con lo anterior fue posible deducir las claves primarias y foráneas del modelo y de la asociación de cada uno de los archivos (regiones, provincias, comunas, distritos, entre otros), logrando con ello fusionar la información de ambos instrumentos en un único modelo, a fin de aprovechar la cobertura territorial del Censo con la información sobre la situación socioeconómica de los hogares que entrega la encuesta. El modelo relacional de la Figura 2 representa la estructura de datos obtenida de la integración entre ambos instrumentos de medición seccionada en dos partes, donde las entidades pertenecientes a la sección número '1' representan a Censo y la sección número '2' corresponde a CASEN.

Una vez integrada la base de datos se procedió a la preparación para el reprocesamiento de los datos utilizando el sistema PostgreSQL[4], reconocido como el sistema de gestión de bases de datos de código abierto más potente del mercado [12] y que además es una derivación libre, gratuita y que permite incorporar un módulo denominado PostGIS[5], que añade soporte a objetos geográficos, transformando una base de datos común en una base de datos espacial, que puede ser utilizada en Sistemas de Información Geográfica [13]. Con estas dos funcionalidades de la base relacional se cargan los datos con tecnología ETL en un almacén de datos intermedio, con el propósito de facilitar la extracción de datos y llevar a cabo su pretratamiento (limpieza y transformación de datos). La carga

[4] PostgreSQL, es un sistema administrador de base de datos de libre disposición sucesor de Ingres, que utiliza el estándar SQL.

[5] PostGIS, extensión de PostgreSQL, que lo convierte en un gestor de bases de datos espaciales.

de los archivos en la base de datos se realiza sin alterar su estructura original, generando una tabla por cada archivo a cargar, donde los datos fueron del tipo alfanumérico para evitar problemas de compatibilidad.

Una vez cargados los datos en el almacén de datos intermedio, se realizó el proceso de limpieza de datos, comprobando la calidad de los datos cargados (mediante test de valores), se eliminaron valores duplicados y se corrigieron valores erróneos. Posteriormente se generó un nuevo almacén de datos intermedio denominado "transformación", dejando a la base de datos inicial como respaldo en caso de que se requiriera volver al proceso de limpieza. En la base de datos "transformación" se efectuó la carga de la estructura del modelo de datos Censo-CASEN, que se pobló con los datos provenientes del subproceso de limpieza. El proceso de transformación consideró cambios de formato, sustitución de códigos y generación de valores derivados, entre otros. Por ejemplo, frente a la incompatibilidad en los códigos de identificación territorial se realizó un proceso de conciliación (coincidencia de códigos) de CASEN con una división regional del país en 13 regiones, a fin de hacer coincidir la división regional del Censo que en el 2007 registra 15 regiones.

Preparación de datos para el emparejamiento espacial

Nuestra propuesta plantea modelizaciones ajustadas a las particularidades de cada área, mediante la técnica de estimación de emparejamiento que busca observaciones en la encuesta que son estadísticamente similares con alguna observación del Censo [15]. Este emparejamiento entre estas observaciones transfiere la ubicación geográfica de los clones censales (observaciones con características similares) a las unidades de la encuesta, con lo que cada observación puede ser asociada a un lugar específico del área muestreada.

La aplicación del emparejamiento espacial mediante la integración de datos de Censo y encuesta requiere condiciones de homogeneidad respecto de la condición geográfica, de las características de la muestra y de las preguntas de recolección de datos [16]. En el primer caso se realizó un proceso de coincidencia de códigos entre la división regional del país considerada en el Censo (15 regiones) y en la encuesta (13 regiones). En el segundo caso, del Censo se excluyeron las observaciones correspondientes a "hogares compartidos", así como las viviendas vacías, al no tener ninguna probabilidad de emparejamiento con CASEN, ya que estas corresponden a viviendas particulares y a los hogares y personas que habitan en ellas. Por último, se definió un set de preguntas comunes entre Censo 2002 y CASEN 2003, así como de las opciones de respuesta y se efectuaron ajustes de sus códigos a valores comunes. La conciliación de las respuestas consistió en recodificar el valor de las variables escogidas y en algunos casos en generar nuevas variables a partir de las previamente existentes. Las variables escogidas se clasificaron en tres grupos de filtro:

–***Variables de identificación:*** dado por aquellas variables o preguntas que fueron consideradas como estructurales y obligatorias en el proceso de emparejamiento. Se optó por utilizar variables que contienen información sobre localización y variables relacionadas con la persona y no con el hogar, por cuanto estos últimos pueden cambiar en su estructura y composición al ganar o perder miembros, al dividirse o unirse.

–***Variables discriminadoras no estructurales:*** dado por aquellas variables con alta probabilidad de encontrar clones y con cierta estabilidad en el tiempo.

–***Variables discriminantes:*** dado por aquellas variables con mayor poder de filtro respecto de la situación socioeconómica de los hogares.

Conciliación de códigos de variables: El proceso de conciliación en la base de datos para las respuestas se presenta en los siguientes ejemplos. La pregunta de la derecha representa la variable escogida en la CASEN 2003 mientras que en la izquierda se encuentra su equivalente en el Censo del 2002:

- **Variables de identificación**

Caso 1: en la Figura 3 se pueden apreciar las preguntas de los instrumentos de medición en Censo y CASEN referentes al parentesco con el jefe de hogar. Aquí se destaca la equivalencia que existe entre las respuestas 2 y 3 de Censo y su homólogo (respuesta 2) en CASEN, el mismo caso ocurre con las respuestas 4 y Censo.

Como resultado del análisis de los datos en la Figura 3, surgen cambios en los códigos de respuesta en la base de datos del Censo, los cuales se pueden visualizar en la Tabla 1, conversión que tiene la finalidad del emparejamiento espacial.

Tabla 1. Cambios de código en Censo (relación Figura 3).

Respuesta Censo	Código Censo	Código de equivalencia CASEN
Conviviente/Pareja	3	2
Hijo/a	4	3
Hijastro/a	5	3
Padres	10	4
Suegro/a	11	5
Otro pariente	12	10
No pariente	13	11
Servicio doméstico puertas adentro	14	12

Fuente: Elaboración propia.

Caso 2: dentro de la conciliación de datos hubo casos aislados en los cuales no fue necesario alterar la codificación existente entre los instrumentos de medición, como es el caso de la pregunta "Sexo", cuyo resultado es idéntico en ambos instrumentos, lo cual se denominó "relación directa".

- **Variables discriminadoras no estructurales**

Caso 1: la pregunta referida al material predominante en el piso de la vivienda (Figura 4) también presenta un ejemplo de equivalencias entre las respuestas 2, 4, 5, 6 y 7 del Censo y el código 1 en CASEN.

- **Variables discriminantes**

Caso 1: un ejemplo de ajuste en este tipo de preguntas viene dado en la Figura 5 relativa a los artefactos y/o servicios que componen la vivienda. Aquí se destaca la relación que existe entre las preguntas "conexión a T.V. Cable/Satélite" de Censo y las preguntas j y k de CASEN. Debido a la diferencia en el tipo de respuestas se realizó una combinación de j y k de CASEN, cubriendo de ese modo los posibles valores en una nueva variable denominada "satel", la que consideró una respuesta negativa solo cuando la respuesta de ambas preguntas fue negativa. El mismo caso ocurre con la variable "conexión a Internet" de respuestas con código g y h de CASEN, donde se genera una nueva variable denominada "inter" (véase Tabla 2 con cambios de código). En la Tabla 3 se puede observar la conciliación de las preguntas realizadas para Censo y CASEN a base de la Figura 5; cabe destacar que los códigos de respuesta entre estas preguntas son idénticos para ambos instrumentos de medición.

PROCESAMIENTO DE DATOS

La consulta combinada a la base de datos busca realizar el emparejamiento de observaciones que comparten características similares entre ambos

Tabla 2. Generación de variables CASEN.

Fusión variables CASEN	Variable resultante
10j y 10k	satel
10g y 10h	inter

Fuente: Elaboración propia.

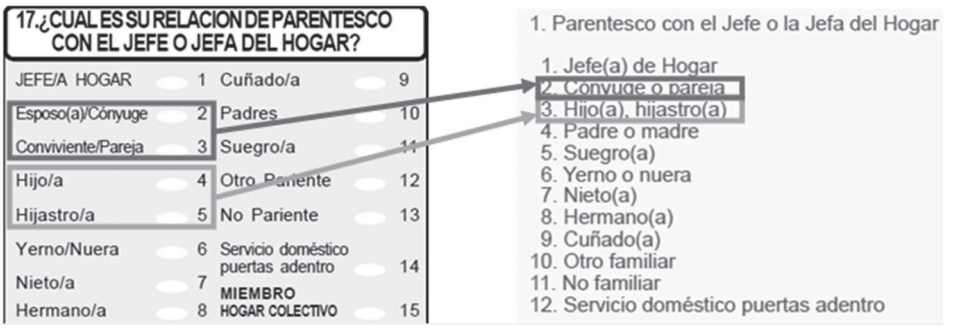

Figura 3. Pregunta referente al parentesco para Censo y CASEN.

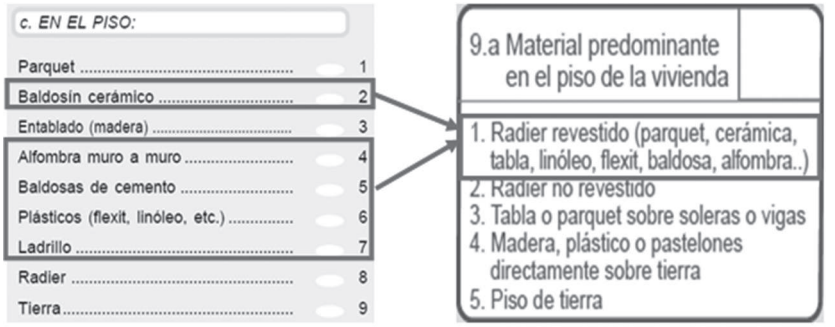

Figura 4. Pregunta referente al estado del piso en Censo y CASEN.

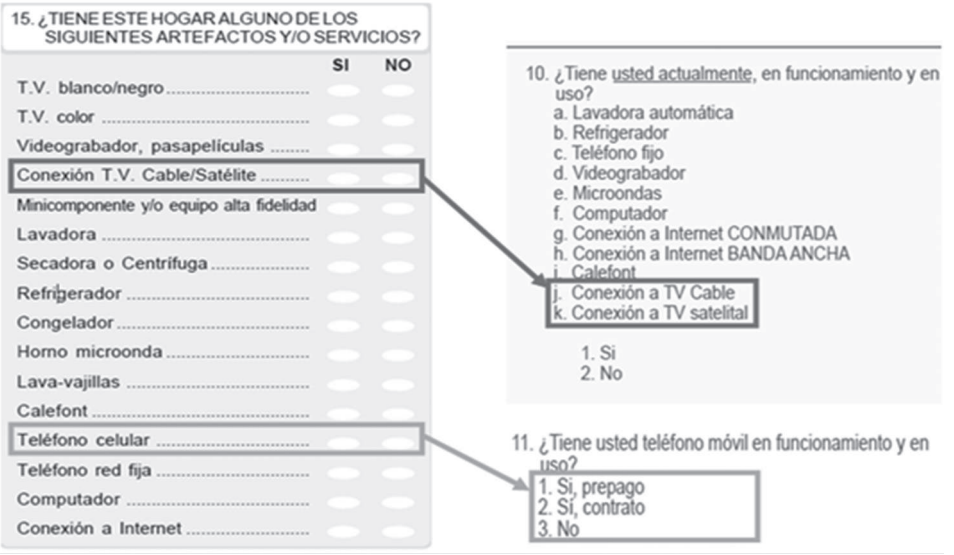

Figura 5. Pregunta referente a artefactos y servicios en Censo y CASEN.

Tabla 3. Conciliación entre los instrumentos del Censo y CASEN (relación Figura 5).

Pregunta Censo	Pregunta Censo	Pregunta equivalente CASEN
Videograbador, pasapelículas	H15_3	10d
Conexión T.V. Cable/Satélite	H15_4	satel
Lavadora	H15_6	10a
Refrigerador	H15_8	10b
Horno microonda	H15_10	10e
Calefont	H15_12	10i
Teléfono celular	H15_13	11
Teléfono red fija	H15_14	10c
Computador	H15_15	10f
Conexión a Internet	H15_16	inter

instrumentos. Se busca encontrar a personas en el Censo que tengan características similares a otras que respondieron la encuesta, según su respuesta a un set de preguntas comunes entre ambos instrumentos. La forma de hacer esta asociación se realiza en dos etapas: la primera consiste en la aplicación de un pareo exacto que busca para cada observación *i* perteneciente a la encuesta con características *x* definidas por un vector de variables comunes, una observación *j* perteneciente al Censo que posea las mismas características. Es decir, $x_i = x_j$, transfiriendo así la ubicación espacial de *j* a *i*, de tal forma que los datos de la encuesta (con el ingreso por hogar incluido) pueden ser asociados a una localización determinada dentro de la unidad territorial.

El vector de variables *x* fue dividido en tres subvectores: el subvector x_{vi} permite ubicar y

diferenciar espacialmente a la persona, el subvector x_{vo} contiene aquellas variables con mayor probabilidad de encontrar clones y con cierta estabilidad en el tiempo, y finalmente el subvector x_{vd}, compuesto por variables discriminantes relacionados con la tenencia de activos o de características personales que poseen un mayor poder de filtro respecto del ingreso del hogar.

El subvector x_v ingresa obligatoriamente en las ecuaciones del emparejamiento, mientras que x_{vo} y x_{vd} lo hacen mediante un proceso de entrada y salida que busca maximizar un patrón de cobertura y precisión. El proceso de cobertura se logra si los clones se distribuyen en todos y cada uno de los distritos de la comuna (menor división intracomunal), mientras que la precisión viene dada por la identificación de la mayor cantidad de observaciones de CASEN que encuentran sus clones entre las observaciones del Censo, evento que se ha llamado *matching* espacial y que se obtiene encontrando las n regresiones de variables para los m clones espaciales. El número de observaciones de j para el que se cumple $x_i = x_j$, dependerá del grado de heterogeneidad espacial que presente la variable.

A partir de los resultados entregados producto del emparejamiento, se generaron indicadores de eficiencia basados en el cálculo de la mayor cantidad de observaciones de la CASEN con la menor cantidad de clones en el Censo. Estos permitieron analizar y escoger el set de variables que mejor se adaptaban al territorio, debido a que se detectó que cada territorio reacciona de forma diferente al conjunto de variables utilizadas. Por ejemplo, entre las zonas urbanas y rurales el mismo set de preguntas mostró diferencias de efectividad. Ejemplo de ello es el uso de electrodomésticos en algunas zonas rurales, debido a la carencia de tendido eléctrico, situación contraria a lo que sucede en el caso de las áreas urbanas.

Una segunda etapa viene dada por la obtención del valor de la variable (en este caso, del ingreso de los hogares) en pequeñas áreas, que se obtiene como el promedio del ingreso a nivel de distritos (menor división administrativa) de las observaciones de la encuesta georreferenciada por sus clones de Censo. Esta etapa permite pasar de un "Único" ingreso medio representativo de los hogares de esa comuna, a un valor del ingreso que cambia de acuerdo con las condiciones económicas de cada distrito dentro de la comuna. Por tanto, con el *matching* espacial se logra la desagregación de los datos de encuesta en el microterritorio, dando de baja el supuesto de la distribución homogénea de los hogares en el territorio, al mostrar una distribución urbana segregada en algunos casos, de acuerdo con su condición socioeconómica. La Figura 6 muestra el diagrama de flujo del *matching* espacial, con

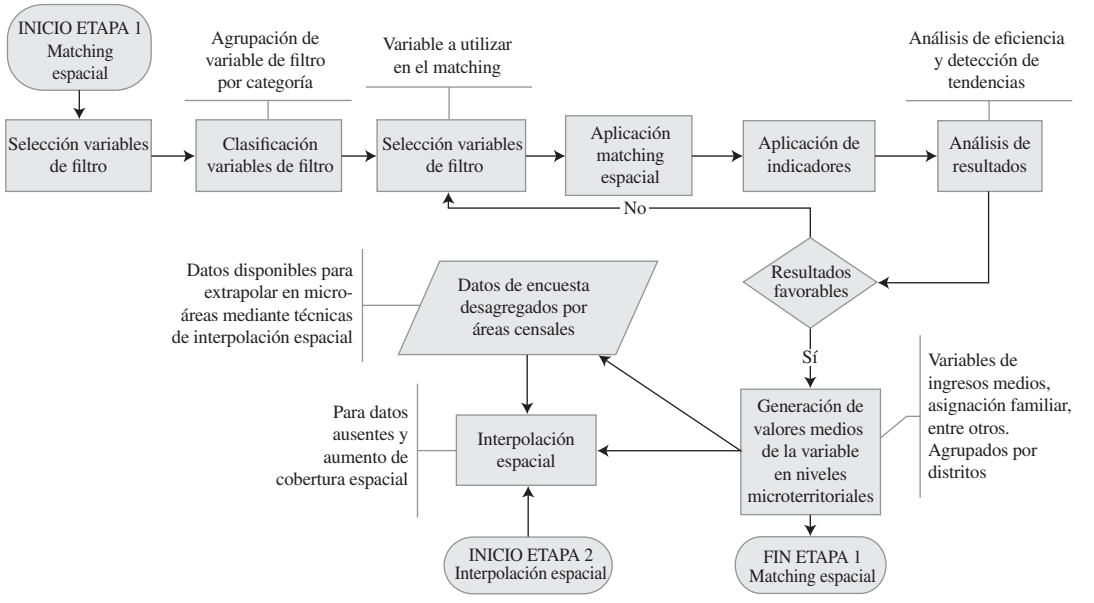

Figura 6. Diagrama de flujo de procesos para la elaboración del calce espacial.

las etapas mencionadas previamente, referentes al tratado de los datos.

SISTEMAS DE INFORMACIÓN GEOGRÁFICA

Una vez obtenido el emparejamiento de datos se implementó una solución de SIG con la herramienta "gvSIG" con tecnología Internet Map Server (IMS, Servidor de Cartografía Digital), a fin de observar y analizar la información socioeconómica georreferenciada tanto en modo vectorial como ráster, realizado con un servidor de mapas denominado MapServer con el Framework Pmapper.

La carga de la cartografía en la base de datos espacial se llevó a cabo mediante códigos SQL. La visualización en gvSIG de indicadores con su cartografía respectiva fueron alojados en tablas de atributos, las cuales se relacionaron con las tablas de mapas mediante las claves utilizadas en el territorio (ej.: código de comunas o distritos, entre otros). Por ejemplo, la Figura 7 muestra a la Región del Maule dividida en cuatro secciones. La sección 1 corresponde al mapa de los distritos de la Región del Maule y en la sección 2 su respectiva tabla de atributos. La tabla de atributos se asocia con los indicadores o variables de interés en la cartografía mediante la herramienta "cálculos" (sección 3). Una vez cargada la tabla con los indicadores, mediante las funcionalidades de gvSIG se seleccionan las tablas a asociar (sección 4) con el código equivalente entre estas (línea entre sección 2 y 3).

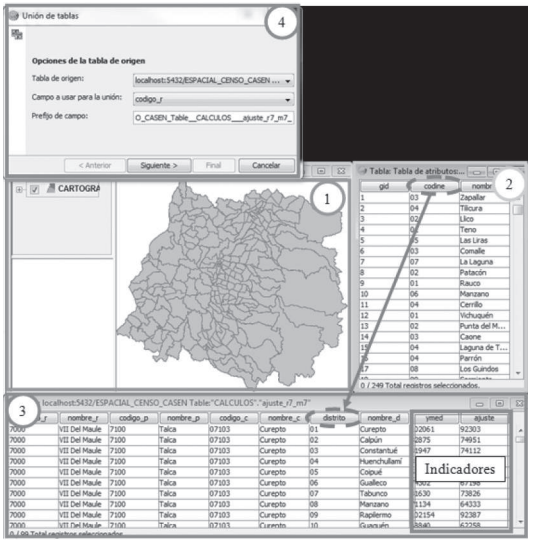

Figura 7. Incorporación de indicadores en gvSIG.

Durante el proceso de generación de indicadores en el esquema "cálculo" fue necesario la conciliación de códigos equivalentes entre las tablas de indicadores y las de mapas, debido a la incompatibilidad del tipo de atributo, ya que estos últimos eran caracteres mientras que las primeras eran valores enteros. También se presentaron inconvenientes con el valor almacenado dentro de estos atributos, debido a que por ejemplo en el esquema "Censo_CASEN", para el distrito "Intendencia" perteneciente a la comuna de "Concepción", Región del "Biobío", se manejaba el código de identificación "0810101", mientras que en la tabla de mapa correspondiente en el esquema "cartografía", el mismo distrito se manejaba con el código "1". Para solucionar este problema se recodificaron las tablas de mapas, específicamente los códigos equivalentes, dejando para el caso de los distritos solamente los dos últimos dígitos, en el caso de las tablas de indicadores, estas fueron generadas con el código de equivalencia de tipo carácter, de esta manera ambas tablas se lograron asociar sin ningún inconveniente. Para el caso de los indicadores de Regiones, Provincias y Comunas, se implementaron soluciones similares.

Uno de estos indicadores es el ingreso per cápita de los hogares, por ejemplo. La Figura 8 muestra los intervalos naturales para el ingreso per cápita del hogar a nivel de distrito en la Región Metropolitana, Sexta y Séptima Región. En el mapa las regiones se analizan de forma independiente, donde los colores más oscuros representan los ingresos más elevados, mientras que los espacios sin color corresponden a un dato ausente en la encuesta. Esta representación gráfica permite analizar la distribución de ingresos en el territorio y distinguir los diversos estratos sociales existentes en las regiones.

La obtención de valores del ingreso en áreas pequeñas también es el resultado de una consulta a la base de datos. Un ejemplo de georreferenciación y desagregación espacial de los datos de la encuesta viene dado por el valor estimado para la comuna de Rancagua en la VI Región, que según la Encuesta CASEN tiene un ingreso medio de los hogares de 131.114 pesos chilenos. En la Tabla 4 se puede ver que hay diferencias significativas en el ingreso per cápita de los hogares en los 17 distritos de la comuna de Rancagua.

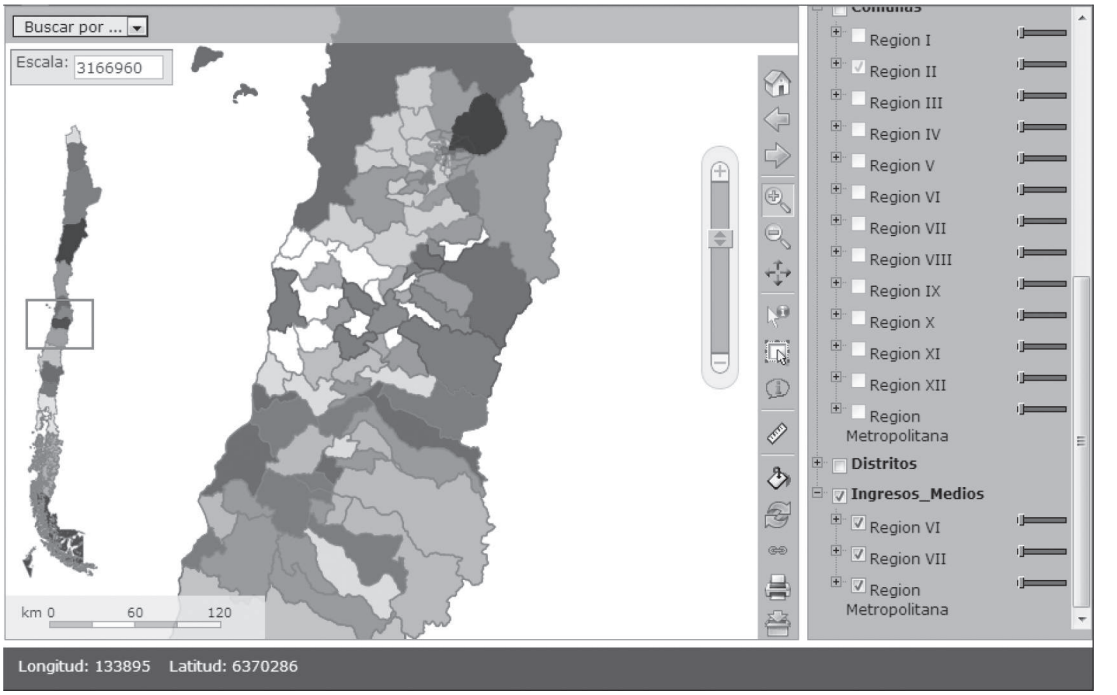

Figura 8. Ingreso medio per cápita del hogar por comuna para las Regiones Metropolitana, VI y VII.

Tabla 4. Ejemplo de desagregación espacial del ingreso per cápita de los hogares.

Distrito	Ingreso del hogar
Intendencia	198.510
Estación	150.487
Cementerio	225.491
Regimiento	194.884
Estadio	149.341
Centenario	140.687
La Capilla	121.572
San Pedro	185.474
La Gamboina	111.345
Los Quilos	116.955
Santa Leonor	110.941
Primavera	75.779
Punta de Cortés	44.967
La Feria	110.280
Medialuna	95.102
Ruta Cinco Sur	155.603
La Moranina	39.819

Otro ejemplo de desagregación en áreas pequeñas por emparejamiento espacial pero desde una vista cartográfica se ve en la Figura 9. Nos centraremos en la comuna de Lo Barnechea en la Región Metropolitana. La tabla de atributos indica que el

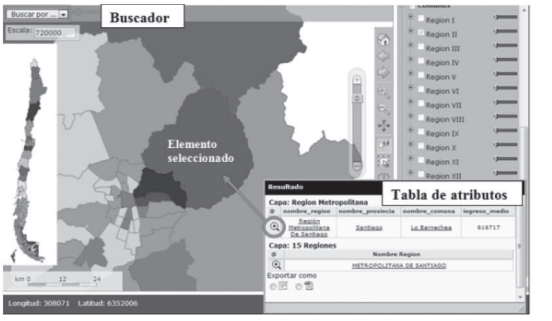

Figura 9. Buscador y tabla de atributos.

ingreso medio de los hogares de la comuna según la CASEN es de 616.717 pesos chilenos. Dicha comuna tiene cuatro distritos, dos de los cuales están más próximos a la comuna de Las Condes (comuna con rentas muy altas), mientras que otro es más próximo a una comuna de ingresos bajos.

Pues bien, la Figura 10 muestra la desagregación espacial de ingreso en esta comuna, mostrando las diferencias internas entre distritos. Los colores más oscuros representan un mayor valor del ingreso de los hogares. Estos valores se obtienen como resultado de la georreferenciación de las observaciones de la encuesta.

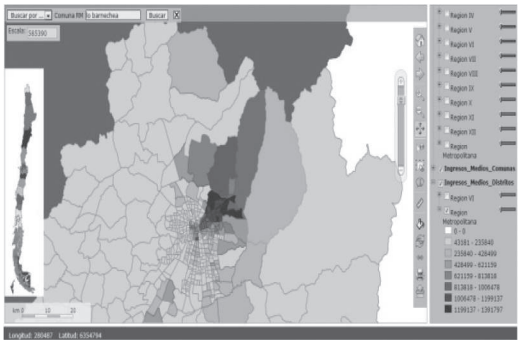

Figura 10. Desagregación espacial. Vista cartográfica.

Estimación de datos ausentes por Kriging

Con la representación de la distribución espacial del ingreso medio de los hogares en las Figuras 8, 9 y 10, es posible ahora mejorar la precisión de la predicción del valor ausente de la encuesta mediante técnicas de interpolación espacial como el Kriging proveniente de la Geoestadística [17], ya que se dispone de un mayor nivel de granularidad o desagregación espacial de los datos en el microterritorio. En virtud de la existencia de una relación inversa entre densidad, la distribución muestral y el error en la predicción, la integración de datos del Censo con la encuesta por medio del emparejamiento espacial resulta adecuada para pasar de datos del ingreso per cápita de los hogares de las 301 a datos para 2.749 distritos de estas comunas en Chile, o en el caso de la VI Región, pasar de 21 comunas a 140 distritos en esas comunas.

Evaluación de resultados

Para la validación de resultados, hemos calculado el error cuadrático medio (RMSE), el error absoluto medio (MAE), el error absoluto medio normalizado (NMAE) y el error sistemático del modelo (BIAS) entre los valores estimados con el emparejamiento espacial de nuestra propuesta y los valores predichos por la metodología ELL, y en cada uno de ellos nuestra propuesta de emparejamiento espacial obtiene mejores aproximaciones (véase Tabla 5).

Tabla 5. Medidas del error.

	RSME	BIAS	MAE	NMAE
Método ELL	184.365	-87.229	87.318	0,2135
Emparejamiento espacial	121.669	40.203	53.597	0,1528

Según las estadísticas anteriores, el método ELL comete un sesgo de subestimación en la estimación del ingreso de los hogares de 87 mil pesos, aproximadamente. Por ejemplo, la Tabla 6 muestra el ingreso estimado por ambos métodos en comunas con valores conocidos. Si comparamos el ingreso dado por la CASEN versus nuestra propuesta de emparejamiento o *matching* espacial con el método ELL, vemos que en la mayoría de los casos nuestra propuesta se aproxima mejor al valor real. En el caso de la comuna de Santiago, el método ELL subestima el ingreso, al igual que en el caso de Doñihue, Rancagua y Cerrillos.

Tabla 6. Comparación del ingreso.

Región	Comuna	Ingreso del hogar. CASEN 2003	Ingreso del hogar. *Matching* espacial	Ingreso del hogar. Método ELL
VI	Rancagua	131.014	129.353	126.435
VI	Coltauco	72.818	81.356	89.177
VI	Doñihue	113.860	100.305	98.229
VI	Graneros	94.598	99.789	97.397
VI	Machalí	128.063	122.340	122.857
VI	Mostazal	85.503	84.477	88.761
RM	Santiago	341.490	379.420	197.627
RM	Cerrillos	153.675	151.394	128.853
RM	Cerro Navia	110.436	131.207	96.313
RM	Conchalí	133.595	151.392	118.424

Respecto de la técnica de predicción Kriging y la desagregación en pequeñas áreas, comparamos la estimación para la comuna de Rancagua por ejemplo, con un Kriging con valores del ingreso per cápita de los hogares en niveles de distritos con un Kriging a nivel de comunas. Para el primer caso se aplica un Kriging (con tendencia) tomando datos del ingreso per cápita para los 830 distritos de las tres regiones colindantes, se obtiene un ingreso estimado de 129.353 pesos, mientras que si se toman del ingreso a nivel de comunas (113 comunas), el ingreso estimado es de 95.140 pesos chilenos, muy por debajo de los 131.114 pesos entregados por la encuesta CASEN.

Por otro lado para evaluar la calidad del emparejamiento espacial también hemos calculado el Índice Global de la Vivienda (IGV), siguiendo la propuesta de CELADE (1996) en los datos de Censo y CASEN. El Índice de Calidad Global

distingue a viviendas de calidad aceptable (A), recuperable (R) e irrecuperables (I). Se construye en función de tres variables: el Índice de Materialidad, Índice de Saneamiento y el Tipo de Vivienda. Si el emparejamiento de datos es el adecuado, los clones encontrados debieran coincidir en el IGV, lo que ocurre en 95,3% de los casos para clasificación de 209.706 pares de clones encontrados entre ambas bases de datos (véase resultados en Tabla 7).

Tabla 7. Validación IGV.

Región Metropolitana	
Observaciones Censo 2002	6.061.185
Observaciones CASEN 2003	52.931
Nº pares de clones	209.706
Coincidencia clones IGV (A,R,I)	199.997
% coincidencia	95,30%

CONCLUSIONES

Esta investigación referente a la estimación de datos con contenido socioeconómico ha girado en torno a la construcción de una base de datos espacial entre Censo y Encuesta, que son dos fuentes de información que comparten características de ubicación territorial y cuya integración permitirá aprovechar las ventajas de la desagregación espacial, de una, y la riqueza de la información de los hogares, en la otra.

Durante el período de elaboración del trabajo se han abordado problemas de diversa índole, como el alojamiento de los datos propios de los instrumentos de medición provenientes de archivos SPSS en una base de datos centralizada, por lo que se investigaron las fuentes de información (objeto en estudio), las técnicas y las herramientas que permitieran realizar esta labor. Como resultado de la investigación asociada de las fuentes de datos se llevaron a cabo procesos de ingeniería inversa para la elaboración del modelo relacional de Censo, la generación de un almacén de datos intermedio en PostgreSQL para el almacenamiento de los datos con miras a la creación futura de un Data Warehouse, la implementación del proceso de ETL para la extracción, limpieza, transformación y carga de los datos mediante sentencias propias de PostgreSQL, y la selección de variables a utilizar en el calce espacial entre los instrumentos de medición.

Con la desagregación territorial de los datos de la encuesta mediante el emparejamiento espacial, la implementación del Sistema de Información Geográfica facilita la posibilidad de mejorar la precisión en las estimaciones de datos inexistentes sobre la condición socioeconómica de sus habitantes en el territorio, así como la observación de zonas de rezago económico, a fin de generar una serie de medidas que permitan mejorar las condiciones socioeconómicas, como es el caso de los programas de asistencia pública.

REFERENCIAS

[1] INE. Instituto Nacional de Estadísticas. Fecha de Consulta: 13 de mayo de 2012. URL: http://www.ine.cl/

[2] I. Molina and J. Rao. " Small area estimation of poverty indicators". Canadian Journal of Statistics. Vol. 38, Issue 3, pp. 369-385. 2010.

[3] R. Chambers and N. Tzavidis. "M-quantile models for small area estimation". Biometrika. Vol. 93, Issue 2, pp. 255-268. 2006.

[4] C. Elbers, J.O. Lanjouw and P. Lanjouw. "Micro-Level Estimation of Poverty andInequality". Econometrica. Vol. 71, Issue 1, pp. 355-364. 2003.

[5] N. Minot and B. Baulch. "Spatial patterns of poverty in Vietnam and their implications for policy". FoodPolicy, Vol. 30, Issues 5-6, pp. 461-475. 2005.

[6] A. Tarozzi and A. Deaton. "Using Census and Survey data estimate poverty and inequality for small areas". The Review of Economics and Statistics. November Vol. 91, Issue 4, pp. 773-792. 2009.

[7] J.Ll. Cano. "Business Intelligence: competir con Información". Fundación Cultural Banesto. Madrid, España. 2007.

[8] P.L. Luna. "Sistema de Información Geográfica para la ayuda de toma de decisiones en políticas sociales". Tesis para optar al grado de Maestría en Ciencias en Computación. Instituto Politécnico Nacional. México. 2010.

[9] J.C. Caralt y J.C. Díaz. "Introducción al Business Intelligence". UOC. Barcelona, España. 2010

[10] W.H. Inmon. "Building the Data Warehouse". Wiley. Third edition. 2002.

[11] GEOINFO. Fecha de Consulta: 15 de noviembre de 2012. URL: http://www.geoinfo.cl/pdf/sig.pdf

[12] PostgreSQL. Fecha de Consulta: 11 de octubre de 2012. URL: http://www.postgresql.org.es/sobre_postgresql

[13] PostGIS. Fecha de Consulta: 11 de octubre de 2012. URL:http://postgis.refractions.net/

[14] R. Dehejia and S. Wahba. "Propensity score-matchingmethods for nonexperimental causal studies". The Review of Economics and Statistics. Vol. 84, Issue 1, pp. 151-161. 2002.

[15] D. Paredes. "A Methodology to Compute Regional Housing Index Price using Matching Estimator Methods". The Annals of Regional Science. Vol. 46, pp. 139-157. 2011.

[16] D. Paredes y P. Aroca. "Metodología para estimar un índice regional de costo de vivienda en Chile". Cuadernos de Economía. 2008.

[17] M. Navarrete. "Modelos Geoestadísticos del Precio de la Vivienda: aproximación al conocimiento microterritorial". Tesis Doctoral. Universidad Autónoma de Madrid. Madrid, España. 2012.