Matching espacial para georreferenciar datos de encuestas de hogar*

Spatial matching for Georeferencing Data of Household Surveys

Mónica Navarrete**
Patricio Aroca***
Jorge Bernal****

Resumen

*Este trabajo desarrolla una metodología de inteligencia de negocios (*data warehouse*) y herramientas OLAP (*Online Analytical Processing*) para parear individuos de una encuesta de hogares con los del censo. Para georreferenciar los datos de los hogares, el método aprovecha la información geográfica del censo. Usando la Encuesta de Hogares de 2003 y el Censo de Población y Vivienda de Chile 2002, se estima el ingreso promedio familiar en niveles intramunicipales y los resultados se comparan con la metodología Elbers et al. (2003). Los resultados muestran que la metodología propuesta permite tener en cuenta la heterogeneidad intramunicipal y mostrar el impacto y la interacción espacial entre las unidades espaciales.*

Palabras clave: *Matching espacial, estimación en pequeñas áreas, encuestas de hogar, inteligencia de negocios.*

Clasificación JEL: *C5, C63, C81.*

Abstract

This paper develops a methodology using business intelligence (data warehouse) and OLAP tools (Online Analytical Processing) to match individuals from a household survey data to a census one. In order to geo-reference the household

* Los autores agradecen los aportes realizados por los árbitros anónimos del artículo, quienes contribuyeron a mejorar sustancialmente el trabajo. Mónica Navarrete y Jorge Bernal agradecen el financiamiento del Proyecto 8741-15 otorgado por el Fondo de Investigación Científica y Tecnológica de la Universidad de Tarapacá. Patricio Aroca agradece el apoyo de los proyectos CONICYT/FONDAP/15130009 and FONDECYT 1140082.

** [Corresponding autor] Centro de Estudios Regionales CEUTA. Universidad de Tarapacá, Chile. Email: mnavarre@uta.cl

*** Centro de Economía y Política Regional (CEPR). Universidad Adolfo Ibáñez, Chile. Email: patricio.aroca@uai.cl

**** Universidad de Tarapacá, Chile. Email: jbernal@uta.cl

data, the method takes advantage of the geographical information of the census. Using the 2003 Household Survey and the 2002 Chilean Census of Population and Housing, the average household income is projected to intra-municipality levels and the results are compared to Elbers et al. (2003) methodology. The results show that the proposed methodology allows for taking into account and display the impact of intra-municipality heterogeneity and the spatial interaction among the spatial units.

Key words: *Spatial Matching, Small Areas Estimation, Household Surveys, Household Income, Business Intelligence.*

JEL Classification: *C5, C63, C81.*

1. Introducción

Los modelos de estimación en pequeñas áreas que combinan datos de encuestas y censo, intentan responder a la creciente necesidad de indicadores de bienestar que muestren con mayor precisión las realidades microterritoriales, utilizando información de la encuesta de hogares que por lo general se ofrece en valores territoriales más agregados. Entre las metodologías existentes destacan los modelos bayesianos (Molina y Rao 2010), los modelos de regresiones cuantílicas (Tzavidis *et al.* 2008) o el modelo propuesto por el Banco Mundial conocido como ELL (basado en Elbers, Lanjouw and Lanjouw, 2003) y utilizado en más de 40 países para construir mapas de pobreza en pequeñas áreas.

El método ELL propuesto inicialmente en Hentschel, Lanjouw, Lanjouw y Poggi (2000) y perfeccionado en Elbers, Lanjouw y Lanjouw (2003), es un modelo de predicción de variables de la encuesta de hogar a partir del apoyo de variables en censo para superar las restricciones de muestreo de la encuesta. El modelo estima una medida de bienestar W calculada en función del ingreso o gasto de los hogares. En una primera etapa, el modelo se extrapola para predecir el ingreso o gasto y, en una segunda etapa, para obtener las medidas del bienestar W y los márgenes de error.

Formalmente el modelo ELL se expresa como la ecuación (1):

$$y_{ic} = x_{ic}\beta + \mu_c + \varepsilon_{ic} \tag{1}$$

Donde de y_{ic} representa los valores del ingreso o gasto de los hogares i en la unidad de *cluster* o subpoblación[1]; c, x_i es el conjunto de covariables restringidas a las disponibles también en el censo; μ_c *que* corresponde al componente de los residuos que resulta común a todos los hogares del *cluster*, mientras que ε_{ic}, es el residuo específico del hogar dentro del *cluster*. La posible correlación espacial de las perturbaciones causadas por unos potenciales efectos de localización se captura mediante *bootstraping* sobre distribuciones aproximadas de los componentes estocásticos del modelo (Elbers, Lanjouw y Lanjouw, 2003; Modrego

[1] Subpoblación asociada a las codificadas en censo como son los hogares, población urbana, población rural, población por etnia, entre otros.

et al., 2008). Para aproximarse a los efectos de localización, el modelo ELL sugiere incorporar el uso de agregados censales que están incorporados como variables en la encuesta, como el nivel de ruralidad de la comuna, la condición étnica, entre otros (Modrego el at 2009).

En Chile, la aplicación del método ELL ha contribuido con información útil para focalizar las políticas públicas de disminución de la pobreza, desagregando información socioeconómica de los hogares desde niveles regionales a niveles más locales como las comunas. Con los regresores y la estimación de la medida de bienestar *W* del modelo ELL, Agostini, Brown y Góngora (2008) identificaron diferencias en las tasas de pobreza por comuna, entre zonas urbanas y rurales. Adicionalmente, en Agostini *et al.* (2010), se identificó la diferencia en las tasas de pobreza de los pueblos indígenas respecto de la población no indígena por comuna. En Modrego *et al.* (2008, 2012), los investigadores utilizaron la encuesta de hogares llamada Encuesta Nacional de Caracterización Socioeconómica (CASEN) de 1992 y 2003, y los Censos de Población y Vivienda de 1992 y 2002 para mostrar la heterogeneidad territorial de las dinámicas de desarrollo del país al encontrar comunas de mayor desarrollo respecto de las rezagadas.

Pese a la potencialidad y utilidad práctica del modelo ELL, recientes aplicaciones en pequeñas áreas cuestionan las condiciones de homogeneidad y tratamiento indiferenciado del espacio[2]. Para Tarozzi y Deaton (2009), los regresores del modelo ELL pueden no recoger las relaciones espaciales dentro de un territorio, ya que puede existir heterogeneidad espacial en áreas pequeñas por diversas fuentes. Por ejemplo, aquellas que resultan de las diferencias propias de los instrumentos de recopilación de datos entre censos y encuestas (ej. preguntas, respuestas, tiempo de aplicación, etc.), las producidas por diferencias entre la relación de las variables predictoras con el espacio (ej. cuando el nivel educacional de la población responde a una política educativa local) o por diferencias relacionadas con la utilización de ciertos activos que dependen de las condiciones territoriales adecuadas para el uso o provisión (ej. uso de la bicicleta o televisión satelital, respectivamente).

Mohamed y Mohamed (2009), por su parte, muestran que el mismo conjunto de variables no explica de la misma forma las relaciones espaciales en el microterritorio, sino que varían según la región considerada, teniendo en algunas de ellas efectos marginales mayores sobre los gastos que en otras. Para Minot y Baulch (2005), el uso de modelos para la zona urbana y otro para la zona rural, identificando la ubicación específica de las personas (hogares viviendo en la costa norte, sur… etc.), es esencial, debido a la presencia de autocorrelación espacial en la distribución geográfica de la pobreza en Vietnam. Sowunmi *et al.* (2012) atribuyen el escaso resultado a la implementación de los programas de

[2] Consideraciones que se relacionan con el efecto que producen en casos de dependencia y la heterogeneidad espacial. El efecto de la dependencia se produce cuando el valor de una variable en un área espacial mantiene una similitud respecto del valor que asume la misma variable en las zonas vecinas, superior a lo que tendría lugar por casualidad. La heterogeneidad se relaciona con las características de multidireccionalidad de los datos espaciales, cuya relación contribuye a generar hipótesis acerca de la génesis de un evento como podría ser la ocurrencia de un proceso de "contagio", en el que el valor de la variable para un determinado lugar está afectado por el valor de "sus vecinos".

superación de la pobreza en Nigeria, a la no consideración de la heterogeneidad de la pobreza y la contigüidad espacial de las unidades geográficas en el diseño de estos programas. Los investigadores indican que los estudios de pobreza existentes tratan a una unidad geográfica como una entidad aislada independiente y no como una entidad rodeada de otras unidades geográficas con los que puede interactuar.

Este trabajo propone un cambio en la lógica tradicional implícita de los modelos de estimación en pequeñas áreas, al incorporar explícitamente la consideración espacial de los datos de la encuesta y del censo. Se parte del hecho que el Ministerio de Desarrollo Social en Chile entrega datos de ingreso o gasto de los hogares a nivel de comuna. Valor agregado que puede ser más o menos representativa del microterritorio en función del tamaño y ubicación de la comuna (o municipalidad), en particular; en aquellas comunas que tienen como vecinos a comunas consideradas hábitat de hogares de ingresos altos y por otro lado colindan con comunas con hogares de menores ingresos.

A diferencia del modelo ELL, no se utilizan las variables de la encuesta de hogares para calcular unos regresores y luego obtener una medida de bienestar al extrapolar estos coeficientes a las variables disponibles también en el censo. Se toman los microdatos de la encuesta de hogares por comuna y los microdatos del censo por comuna y se realiza un pareo entre ellos mediante el apoyo de las variables comunes entre ambos instrumentos. Se plantea que al existir 30 preguntas comunes entre instrumentos que han sido tomados en tiempos contemporáneos, es posible encontrar en censo a quien respondió la encuesta, de modo de transferir esta ubicación territorial dentro de la comuna para luego calcular las variables de la encuesta en unidades territoriales inferiores como son los distritos.

Las observaciones de Tarozzi y Deaton (2009) respecto de combinar datos de encuesta y censo relativas a la existencia de heterogeneidad espacial producto de la diferencia de instrumentos de recolección de información se abordan utilizando herramientas de inteligencia de negocios como un Data Warehouse y herramientas OLAP (Online Analytical Processing), realizando previamente un proceso de conciliación de códigos de preguntas, respuestas y unidades territoriales. En el trabajo de Cornejo *et al.* (2014) se explica con detalle el desarrollo de la solución informática respecto de la construcción de la base de datos espacial interconectada entre ambos instrumentos y sobre la que se realizan las consultas de emparejamientos para cada sujeto de la encuesta. El emparejamiento transfiere la ubicación geográfica de los clones censales a sus clones en la encuesta, de modo que estas puedan ser asociadas a un lugar específico del área muestreada en el evento (*matching* espacial). Esta transferencia se hace a nivel de distrito censal, que en promedio tiende a representar el diez por ciento de una comuna.

Con el *matching* espacial se responde a la sugerencia de Minot y Baulch (2005) en cuanto a identificar la ubicación específica del individuo de la encuesta de hogares. Ubicación que no es exacta, ya que irá al centroide del distrito censal, que de todas maneras será un avance significativo al de usar el centroide de la comuna. La identificación de la ubicación se realiza utilizando un set de variables comunes para cada comuna a diferencia de lo que hace el método ELL de aplicar un modelo por región, respondiendo de paso a la propuesta de Mohamed y Mohamed (2009) acerca de evaluar modelos para cada zona. Por ejemplo, los resultados muestran casos en que el *matching* no encuentra sujetos en la encuesta para zonas no habitables o que el ingreso medio de los hogares

de un distrito colindante entre dos comunas puede ser diferente o similar entre sí dependiendo de sus vecinos.

Además, para identificar la presencia de las relaciones espaciales en valores del ingreso medio de los hogares en los niveles intracomunales (distritos), se aplican herramientas de la geoestadística como el interpolador *kriging*, que hace uso del semivariograma para detectar e incorporar la autocorrelación espacial en la predicción de datos. También se hace uso del índice y diagrama de Moran utilizados por la econometría espacial para evaluar las relaciones de vecindad en el ingreso medio de los hogares obtenido por *matching* espacial a nivel de distritos.

Con este trabajo se espera adicionar a las técnicas actuales de interpolación o extrapolación de información desde encuestas a censos, los elementos de heterogeneidad e interacción espacial presentes en un proceso de estimación en pequeñas áreas. Los resultados muestran que ambos procesos están presentes en el ejemplo desarrollado y que la información generada será muy valiosa a la hora del diseño e implementación de políticas.

En el siguiente apartado se presenta una continuación del trabajo expuesto en Cornejo *et al*. (2014) acerca de la construcción de una base de datos espacialmente integrada entre datos de censo y encuesta, profundizando ahora en el aspecto metodológico del *matching* como técnica de desagregación espacial de datos. En la sección III se presentan los datos y el análisis que justifica la intervención de metodologías que consideran el espacio heterogéneo. En la sección IV se muestra el resultado de la aplicación metodológica propuesta en dos regiones del país colindantes y heterogéneas intra e interregionalmente: la Región Metropolitana y la VI Región del Libertador Bernardo O'Higgins. En la sección V se presentan las conclusiones y observaciones generales.

2. Metodología propuesta

2.1. *Matching* espacial

Utilizando los hogares comparables entre el CENSO del 2002 y los datos de la Encuesta de Caracterización Socioeconómica Nacional de hogares (conocida como CASEN) del 2003, se busca identificar en qué lugar de la comuna se ubican probablemente los hogares entrevistados en la encuesta, de modo que su georreferenciación permita observar territorialmente las diferencias (si las hubiera) en variables socioeconómicas, por ejemplo, en el ingreso de los hogares mediante un *matching espacial* y que es una adaptación de la técnica del *matching estimator*. El *matching estimator* (clasificado dentro de los métodos cuasiexperimentales) busca el emparejamiento de unidades muestrales pertenecientes a un grupo "tratado" y uno de "control" (Rubin, 1976) que son similares en términos de sus características observables, de modo que de su comparabilidad sea posible inferir las fuentes de diferencias que provengan del tratamiento[3]. Esta metodología considera que las diferencias en el tratamiento

[3] Ver Rubin (1973, 1976), Dehejia and Sadek (2002) y Paredes y Aroca (2008) para mayor información.

serían atribuidas en su totalidad a un conjunto de características o atributos y se asume que la distribución probabilística de estas es similar tanto para las unidades de control como para las unidades en tratamiento. Se le considera una técnica adecuada para valorar un tratamiento médico, el impacto de programas de vacunación o la evaluación de programas alimenticios.

En el área económica, Paredes y Aroca (2008) lo utilizan para construir un índice regional de precios de vivienda que tome en cuenta el efecto "territorial" del precio de la vivienda entre regiones. Para comparar el efecto "región" sobre el precio de la vivienda, tendría que valorarse una misma vivienda en dos regiones simultáneamente (una de control y la otra para medir el efecto de la región) y luego verificar si existen diferencias significativas en su precio por región. Sin embargo, como no es posible contar con la misma vivienda en dos regiones, seleccionan una vivienda con características similares en la región de control y luego verifican si existen diferencias significativas en sus precios.

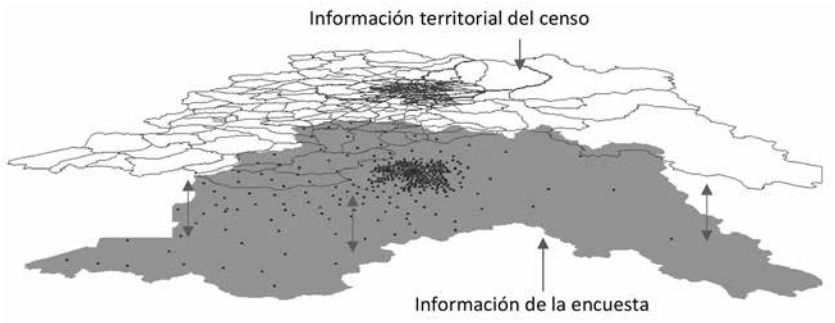
Siguiendo esta idea, el *matching* espacial busca observaciones del censo (unidad de control) con características similares a otra de la encuesta (unidad de tratamiento), para que se pueda imputar luego la ubicación espacial de los primeros a los segundos dentro de la comuna. La Figura 1 muestra la representación visual de esta idea con dos imágenes. La imagen inferior muestra un polígono correspondiente a una unidad territorial donde se toman los datos de la encuesta CASEN en alguna parte de ella. La imagen superior muestra la división intraterritorial de la unidad de observación dada por el Censo. Con el *matching* espacial los clones censales traspasan su ubicación a los datos de la encuesta, de modo que se puedan asociar a un área territorial particular, específicamente a un distrito censal, lo que es un mejoramiento considerable al estar asociado a una comuna.

2.2. Solución de inteligencia de negocios

La solución propuesta recurre al uso de tecnologías de integración de datos con herramientas de extracción, transformación y carga[4] (ETL), 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 negocio y cargar estos en almacenes de datos o *data warehouse* (DW) diseñados a propósitos del analista. En datos del Censo se realizó un proceso de ingeniería inversa entre los archivos de hogar, personas, viviendas y entidades (inicialmente separadas), con el fin de reconstruir la estructura de relaciones entre archivos para posteriormente realizar consultas de forma integrada mediante la incorporación del módulo PostGIS que le añade soporte a objetos geográficos, transformando la base a una base de datos espacial, para

[4] El proceso de ETL se aplicó a los datos del Censo 2002 y a la encuesta CASEN 2003 que disponen de sus datos en archivos tipo .sav y al no existir documentación relativa a su modelo de base, se utilizó un proceso de ingeniería inversa para reconstruir las relaciones de base entre los datos. Una vez generada la base de datos integrada entre censo y CASEN, se realiza una conversión de formato de los archivos del tipo ".SAV" a ".CSV", a fin de ejecutar sentencias SQL propias de PostgreSQL, reconocido como uno de los sistemas de gestión de objeto-relacional de código abierto más potente del mercado.

FIGURA 1
REPRESENTACIÓN VISUAL DEL *MATCHING* ESPACIAL



Fuente: Elaboración propia.

ser utilizada en Sistemas de Información Geográfica (SIG)[5], pertinente para visualizar la distribución de los clones censales en los niveles intracomunales[6].

2.3. Poblamiento de base de datos

El tratamiento o transformación y carga de los datos en el DW requirió unas condiciones de homogeneidad entre ambos instrumentos respecto de la división administrativa, de las características de la muestra y ante las preguntas y respuestas en la recolección de datos.

Concerniente a la división administrativa, se homogeneizaron los códigos de identificación territorial, ya que el Censo del 2002 contiene la información de acuerdo con la división administrativa del país vigente en el 2002, mientras que la encuesta CASEN del 2003 responde a la división administrativa de 1992. La Figura 2 muestra un ejemplo de conciliación de códigos de identificación para la comuna de Arica. El código 1101 de CASEN, identificador de la comuna de Arica perteneciente a la I Región de Tarapacá, se ajustó para ser equivalente al código 15101 asignado a la misma comuna en la base de datos de Censo asignado a la Región XV de Arica y Parinacota.

Acerca de las características de la muestra, la conciliación significó que se excluyeran las observaciones correspondientes a "hogares compartidos" de la base de censo así como las viviendas vacías, al no tener ninguna probabilidad de emparejamiento o *matching* con datos de CASEN, porque estas corresponden a viviendas particulares y a los hogares y personas que habitan en ellas. Por último, respecto de la conciliación de preguntas y opciones de respuesta; se realizaron ajustes a códigos a comunes buscando un *matching* exacto. Un ejemplo

[5] La mayor utilidad de un SIG está intrínsecamente relacionada con la capacidad que este posee de construir modelos o representaciones del mundo real a partir de las bases de datos digitales.

[6] Una explicación y descripción del diagrama del proceso implementado puede ser consultado en el trabajo de Cornejo *et al.* (2014).

FIGURA 2
CONCILIACIÓN DE CÓDIGOS ENTRE DATOS DE CENSO Y ENCUESTA

CENSO						CASEN				
C-R		C-P		C-C		C-R		C-P	C-C	
15000	XV de Arica y Parinacota	15100	Arica	15101	Arica	1	I De Tarapacá	11	1101	Arica
		15100		15102	Camarones				1102	Camarones
		15200	Parinacota	15201	Putre			12	1201	Putre
		15200		15202	General Lagos				1202	General Lagos

Fuente: Elaboración propia.

de cambio de código de respuestas a una pregunta viene dado en la Figura 3. La relación de parentesco con el jefe de hogar en la base de datos del censo diferencia cuando se responde esposo(a)/Cónyuge (código 2) o Conviviente/pareja (código 3), mientras que en la base de datos CASEN ambas respuestas se contabilizan como una sola (código 2). Lo mismo para la relación de hijo/a (código 4), Hijastro/a (código 5), respecto de la condición Hijo(a), hijastro(a) asignado con código 3 en CASEN. En casos como este, se reemplazó el código 4 y 5 por un código 3 en la base de censo.

Un caso similar se presenta en la Figura 4, respecto de su pertenencia a pueblos originarios. Mientras en el CENSO pertenecer al pueblo atacameño se codifica con un 2, en la encuesta CASEN se codifica con un 5. La homogeneización pasa por codificar de igual manera en la base de Censo.

En casos como el presentado en la Figura 5, se crearon variables de integración en preguntas excluyentes en un instrumento, pero colectivas en el otro. Por ejemplo, la consulta acerca de la disponibilidad de los artefactos o servicios en la vivienda, censo ofrece integradamente "conexión a T.V. Cable/Satélite" mientras que en CASEN esta opción (dada en alternativa j y k) es dicotómica SÍ o NO.

FIGURA 3
HOMOGENEIZACIÓN EN CÓDIGOS DE RESPUESTA BASE DE DATOS INTEGRADA. EJEMPLO 1

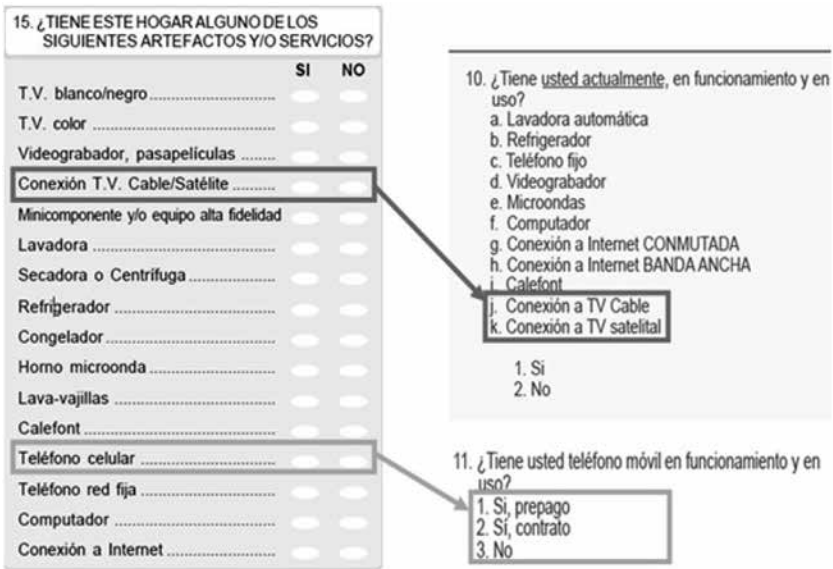
¿Cuál es su relacion de parentesco con el Jefe o jefa del hogar?				
Respuesta CENSO	Código CENSO	Cambio codigo en CENSO	Código CASEN	Respuesta CASEN
Jefe/a hogar	1		1	Jefe(a) de hogar
Esposo(a) conyuge	2	2	2	Conyuge o pareja
Conviviente/Pareja	3	2	3	Hijo(a) hijastro(a)
Hijo/a	4	3	4	Padre o madre
Hijastro/a	5	3	5	Suegro(a)
Yerno/nuera	6		6	Yerno o nuera
Nieto/a	7		7	Nieto(a)
Hermano/a	8		8	Hermano(a)
Cuñado/a	9		9	Cuñado(a)
Padres	10	4	10	otro familiar
Suegro/a	11	5	11	No familiar
Otro pariente	12	10	12	Servicio domestico puertas adentro
No pariente	13	11		
Servicio domestico puertas adentro	14	12		

Fuente: Elaboración propia.

FIGURA 4
HOMOGENEIZACIÓN EN CÓDIGOS DE RESPUESTA BASE DE DATOS INTEGRADA. EJEMPLO 2

Fuente: Elaboración propia.

FIGURA 5
HOMOGENEIZACIÓN EN CÓDIGOS DE RESPUESTA BASE DE DATOS INTEGRADA. EJEMPLO 3



Fuente: Cornejo et al (2014).

Debido a la combinatoria de respuestas se generó la variable "satel", que es negativa cuando las respuestas a las preguntas j y k son negativas. El mismo caso ocurre con la variable "conexión a internet" con código g y h de CASEN, donde se genera una nueva variable denominada "inter".

Por último, un *matching* inexacto se aplicó en preguntas que no tenían correspondencia exacta asignando un rango de variación. Por ejemplo, respecto

de la edad del entrevistado, la diferencia de año y meses en la aplicación de ambos instrumentos generó en promedio, un rango positivo de 24 meses al valor respondido en el censo.

2.4. Formulación del *matching spatial*

Con la base homogeneizada e integrada de censo y encuesta, se realizan consultas SQL en unidades territoriales a nivel de comuna que contienen variables con códigos comunes en ambos instrumentos. El set de variables fue dividido en dos bloques, asignadas de acuerdo con su importancia y función en el proceso de *matching*, el que busca, por una parte, identificar espacialmente las observaciones y, segundo, diferenciar entre ellas utilizando variables con capacidad de discriminar socioeconómicamente entre observaciones según las características y equipamiento de la vivienda como material de construcción, electrodomésticos, alcantarillado, luz, teléfono fijo, teléfono móvil, computador, entre otros, así como características personales como el nivel educativo y su condición laboral. Variables que en la metodología ELL son utilizadas como variables explicativas del ingreso de los hogares.

Sea P_0 el conjunto de datos correspondientes a las personas en el Censo 2002, considerado como Conjunto de Control, este a su vez formado por 13 conjuntos de datos $\{R_{0;i}\}^{13}_{i=1}$, donde $R_{0;i}$ representa el conjunto de datos correspondiente a las personas encuestadas en la región R_i. Cada dato del conjunto $R_{0;i}$ posee ciertas características que se pueden clasificar de acuerdo con la respuesta de las siguientes variables:

Variables de identificación, vector x_{id} formado por los siguientes componentes:
a) x^1_{id}: Código de región,
b) x^2_{id}: Código de provincia,
c) x^3_{id}: Código de comuna,
d) x^4_{id}: Zona (rural, urbana),
e) x^5_{id}: Parentesco con el jefe de hogar,
f) x^6_{id}: Género,
g) x^7_{id}: Pertenencia a algún pueblo indígena,
h) x^8_{id}: Rango de edad.

$$\text{De esta manera: } x_{(id)} = (x^1_{id}, x^2_{id}, x^3_{id,} x^4_{id,} x^5_{id,} x^6_{id,} x^7_{id,} x^8_{id})$$

Variables socioeconómicas, vector x_{or} formado por los siguientes componentes:
a) x^1_{or}: Material de construcción vivienda
b) x^2_{or}: Sistema de alumbrado
c) x^3_{or}: Sistema alcantarillado
d) x^4_{or}: Tipo de vivienda,
e) x^5_{or}: Propietario de la vivienda,
f) x^6_{or}: Refrigerador,
g) x^7_{or}: Teléfono de red _fija,
h) x^8_{or}: Horno microondas,
i) x^9_{or}: Computador,
j) x^{10}_{or}: Cálefont,
k) x^{11}_{or}: Teléfono celular,

l) x^{12}_{or}: Conexión a TV cable o satélite,
m) x^{13}_{or}: Nivel educativo cursado,
n) x^{14}_{or}: Condición laboral.

Tal que $x_{(or)} = (x^1_{or}, x^2_{or}, x^3_{or,} x^4_{or,} x^5_{or,} x^6_{or,} x^7_{or,} x^8_{or} x^9_{or,} x^{10}_{or,} x^{11}_{or,} x^{12}_{or,} x^{13}_{or,} x^{14}_{or})$

El primer objetivo del *matching* espacial es buscar, para cada dato de P_1, (grupo de tratamiento o personas encuestadas en CASEN), con características $y= (y_{id}, y_{or})$, uno o más datos de P_0 con características $x = (x_{id}, x_{or})$, tal que $x=y$, es decir, que las características del dato observado en P_1, sean iguales a las características de uno o más datos observados de P_0. Este primer objetivo se puede lograr mediante un algoritmo de selección para el que denotemos por $N= |P_0|$: cantidad de datos del conjunto P_0, y por $M= |P_1|$: cantidad de datos del conjunto P_1, se obtiene para cada $yk \in P_1$, con $k \leq M$, un conjunto $K(yk)$ formado por todas las observaciones de P_0, que poseen las mismas respuestas en las variables de identificación de yk, es decir, $K(yk)$ es el conjunto de "clones" de yk, respecto de las variables de identificación. Seguidamente para cada observación $yk \in P_1$, se emparejan sus respuestas con las variables socioeconómicas, para cada elemento de $K(yk)$, de manera de disminuir la cantidad de clones de yk.

El segundo objetivo del *matching* es maximizar la cantidad de observaciones de P_1 que poseen clones a nivel de comuna, por tanto se define $N' = |K(yk)|$ como la cantidad de clones encontrados en el censo, para la observación yk de la encuesta CASEN por comuna. Si x_α es uno de los clones encontrados, es decir, si $x_\alpha \in K(yk)$, entonces se procede a comparar sus respuestas respecto de cada variable socioeconómica (proceso uno a uno) de yk, con la misma característica de x_α. Para esto se define un nuevo conjunto $K'(yk)$ formado por todos los elementos de $K(yk)$ que tienen igualdad en las respuestas a las variables socioeconómicas de yk. Si la característica es distinta, entonces el elemento x_α se elimina del conjunto de clones. Luego se continúa con el siguiente dato, realizando otra vez la comparación de la misma característica hasta comparar todos los elementos (algoritmo descrito en el anexo 1). Después de esto se calcula la proporción de elemento de $K(yk)$ que se mantiene después de la comparación, utilizando el índice:

$$\phi_\lambda = \frac{|K'(yk)|}{|K(yk)|} \tag{2}$$

donde λ indica la característica $\lambda-$ésima de la observación.

Se consideran las variables cuyo índice fuese el mayor posible una vez evaluada cada variable socioeconómica. Este proceso se realiza de forma separada para cada comuna. Por ejemplo, para la comuna de Santiago perteneciente al área urbana de la Región Metropolitana se utilizaron las variables de identificación del vector $x_{(id)} =(x^1_{id}, x^2_{id}, x^3_{id,} x^4_{id,} x^5_{id,} x^6_{id,} x^7_{id,} x^8_{id})$ más el vector $(x^4_{or,} x^5_{or,} x^7_{or,} x^9_{or,} x^{12}_{or,} x^{13}_{or,} x^{14}_{or})$. Para comunas rurales como Melipilla, de la misma Región, fue necesario agregar las variables $(x^1_{or}, x^2_{or}, x^3_{or})$.

Una vez identificadas las observaciones muestreadas de la encuesta dentro de la comuna, se realiza la consulta SQL según distritos (división intracomunal)

y se obtienen los valores promedio de la variable solicitada; en este caso, para el ingreso medio de los hogares.

2.5. Herramientas de análisis espacial

Para detectar las relaciones espaciales entre los datos de la encuesta se utilizan dos herramientas de análisis espacial: el interpolador geoestadístico *kriging* y el diagrama de Moran. El primero porque en su concepción considera el carácter sesgado de la estadística convencional cuando no tiene en cuenta la correlación espacial de los datos en sus mediciones. La fundamentación teórica implícita en los modelos geoestadísticos incorpora el efecto de la posición geográfica de los datos y la interrelación espacial entre sus vecinos mediante el interpolador *kriging*. Propuesto por Matheron (1962), el *kriging* es una técnica basada en una media móvil ponderada que depende de la distribución de las observaciones (posibles agrupamientos), de la distancia o proximidad geográfica respecto del punto no muestral, del tamaño y calidad de los datos y de la estructuración de la variable[7] (Montero y Larraz, 2008). El *kriging* puede ser interpretado como una predicción de un valor desconocido $\hat{Z}(x)$ compuesto por las variables aleatorias $Z(x_i)$ que corresponden al valor que asume la variable x en las n localizaciones cercanas a x_o localización con valor desconocido, según:

$$\hat{Z}(x_0) = \lambda_1 Z(x_1) + \lambda_2 Z(x_2) + ... + \lambda_n Z(x_n) \tag{3}$$

O de forma más general (4):

$$\hat{Z}(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i) \tag{4}$$

Donde λ_i (con valores entre 0 y 1) es la función de ponderación[8] de los n puntos muestrales que intervienen en el cálculo de la distancia al punto no muestral $\hat{Z}(x_0)$ a partir de los puntos conocidos $Z(x_i)$. La estructura de la dependencia espacial es obtenida mediante el semivariograma γque establece una relación entre la semivarianza entre cualquier par de valores $Z(x_i)$ y $Z(x_i + h)$ y que se obtiene como el valor promedio de la diferencia al cuadrado de los valores de la variable (al utilizar la esperanza) en dos puntos separados por una distancia h donde (5):

$$\gamma(x, x+h) = \frac{1}{2} Var[Z(x+h) - Z(x)] = \frac{1}{2} \mathrm{E}[Z(x+h) - Z(x)]^2 \tag{5}$$

---

7 Factores que se modelizan para que produzcan un estimador óptimo y eficiente (Agterberg, 2004), diferenciando así al *kriging* de otras técnicas de interpolación determinísticas como el Inverso de la Distancia, las funciones polinómicas o del interpolador por mínimos cuadrados, por ejemplo y que han llevado a los investigadores a aplicar el método geoestadístico en diversas ramas de las ciencias naturales y sociales (Navarrete, 2012; Lam, 1983).

8 La influencia de los puntos muestrales disminuye con la distancia y es incorporada a la función de interpolación.

γes independiente de la localización, pero depende de valor y de la dirección del vector de distancia *h.* Los datos observados de la variable son utilizados para calcular el semivariograma empírico ajustado al semivariograma teórico o autorizado que mejor representa los datos. Para garantizar estimadores insesgados se impone que λ_i sume la unidad. La varianza del error de estimación se obtiene en función del semivariograma (6).

$$V\left[\hat{Z}(x_0) - Z(x_0)\right] = 2\sum_{i=1}^{n} \lambda_i \gamma(x_i - x_0) - \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(x_i - x_j) \tag{6}$$

Para el análisis de la dependencia espacial entre unidades territoriales la econometría espacial utiliza el índice y el diagrama de dispersión de Moran, que requiere la identificación de una matriz de contigüidad (W) que indica el grado de cercanía entre las observaciones, en donde cada elemento de W (7) refleja la intensidad de la interdependencia existente entre cada par de comunas *i* y *j* como consecuencia de su posición en la región[9].

$$\mathbf{W} = \begin{pmatrix} 0 & w_{12} & \dots & w_{1n} \\ w_{21} & 0 & \dots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \dots & 0 \end{pmatrix} \tag{7}$$

Si el ingreso de los hogares (x) presenta correlación espacial positiva, cabría esperar que para las comunas *i* y *j* cercanas $(x_i - x_j)^2$ fuese pequeño, por tanto, la suma de las diferencias cuadráticas $\forall\left(x_i, y_j\right) \middle| i \neq j$ sería pequeña en relación con un patrón no correlacionado espacialmente. Esta idea se puede medir de forma analítica o gráfica, destacándose entre las primeras el estadístico *I* de Moran, que permite contrastar la hipótesis nula de que un conjunto de valores muestrales de una variable se encuentran distribuidos de forma totalmente aleatoria en el espacio o si, por el contrario, existe una asociación significativa de valores similares o diferentes entre espacios vecinos (8). En ausencia de correlación espacial, el estadístico *I* de Moran es prácticamente nulo, mientras que el valor es positivo para correlación positiva y negativo en caso contrario.

$$I = \frac{N}{S_0} \frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2} \tag{8}$$

[9] Si el criterio de contigüidad física es de primer orden, será unitario cuando i y j son físicamente adyacentes y cero en caso contrario, mientras que los criterios conocidos como de la torre (*rook*), alfil (*bishop*) o reina (*queen*) aumentan el número de "vecinos".

Donde:

$S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$ Es la suma de los pesos espaciales, $\overline{X}$ es el valor medio de la variable y n el número de observaciones.

La representación gráfica de la dependencia espacial se realiza mediante el s*catterplot* de Moran que representa en el eje de abscisas el valor de la variable normalizada y en el de ordenadas, el retardo espacial que se obtiene del promedio estandarizado el resto de las localizaciones "vecinas", de modo que la nube de puntos permitirá comparar el valor de la variable que en este caso es el ingreso de los hogares de una comuna con el ingreso en otras comunas que se consideren como "vecinas".

2.6. Medidas de validación de observaciones georreferenciadas

Para evaluar la georreferenciación de las observaciones de la encuesta en niveles intramunicipales o intracomunales territoriales, se utiliza el Índice de Clasificación Socioeconómica (CSE) en 10 categorías (CSE_D) que fue realizado por el INE en una etapa de postcenso y que clasifica a los hogares del CENSO 2002 de acuerdo con un conjunto de variables socioeconómicas (variables relacionadas con el hogar, la vivienda, educación y ocupación del jefe de hogar entre otras) y de ubicación geográfica del hogar en distintos niveles geográficos, entre ellos el distrito.

El índice CSE fue construido a partir de una muestra del 10% de las viviendas que arrojaron unos puntajes de las variables incluidas siguiendo el procedimiento PRINQUAL[10] o Componentes Principales Cualitativos, cuyos parámetros luego fueron aplicados a todos los hogares del censo, clasificando a cada hogar en una categoría de acuerdo con la escala generada por el puntaje Princals. El puntaje mínimo posible se asocia a un hogar con todas las condiciones socioeconómicas (medidas de acuerdo con el conjunto de variables dadas) desfavorables, mientras que si todas las condiciones le son favorables, se obtiene el caso contrario. El puntaje que se origina entre el rango de amplitud del puntaje más bajo y el más alto fue utilizado para la decilización (CSE_D) del puntaje, es decir, se ordenaron todos los hogares del Censo de mayor a menor puntaje Princals y se dividió en 10 partes iguales de acuerdo con el número de hogares. Así, cada decil contiene a un 10% de los hogares, siendo el primer decil compuesto por aquellos hogares con el 10% con los puntajes más bajos y el décimo al 10% con los puntajes más altos.

El CSE_D fue testeado con la clasificación socioeconómica obtenida directamente de la CASEN (en niveles comunales) con la variable ingreso y sin ella incluida en el modelo, la que arrojó clasificaciones altamente correlacionadas. Debido a que el CSE_D no incluye en el modelo la variable ingreso pero sí aquellas variables de censo que lo explican conjuntamente, hemos elegido este indicador para evaluar la desagregación espacial mediante la georreferenciación de los datos de la encuesta observando el grado de correlación entre el ingreso promedio del hogar obtenido a nivel de distrito y el puntaje promedio del CSE_D.

[10] Modelo de componentes principales que combina variables cuantitativas con variables cualitativas.

También se ha utilizado como tester cartográfico, la puntuación de factores que realiza la Biblioteca Nacional del Congreso para calificar a los distritos según la graduación de color en alto, medio alto, bajo, medio bajo, utilizando variables del Censo 2002 de niveles educacionales y económicos de cada distrito (BCN Informe, 2007, pág. 14) (http://siit2.bcn.cl/mapoteca/datos_tematicos).

3. Los datos

La Encuesta Nacional de Caracterización Socioeconómica (conocida como CASEN) se realiza en Chile cada dos años desde 1987. Tiene como objetivo realizar un diagnóstico de la situación socioeconómica de los hogares y de la población con el fin de ayudar en la evaluación del grado de focalización e impacto distributivo de los programas sociales de mejoramiento de las condiciones de vida de la población. La recolección de datos ha aumentado tanto en cobertura como en información pasando desde 48 comunas en 1987 a 302 comunas del país en el 2003, lo que equivale al 88% de ellas. La encuesta es representativa a nivel regional existiendo críticas respecto de su uso para inferencia estadística en el nivel comunal debido a los altos niveles de error asociado a dichas estimaciones (Agostini, Brown y Góngora, 2008). En la CASEN no es posible identificar a los hogares y viviendas en niveles de división intracomunales. Por otro lado, el Censo Nacional de Personas, hogares y vivienda corresponde al 2002. La división administrativa del Censo puede ser obtenida a nivel de región, provincia y comuna. En niveles de desagregación menor, el Instituto Nacional de Estadísticas utiliza niveles administrativos de planificación para el levantamiento de información censal dentro de las comunas que se definen como distritos y, dentro de estos se consideran áreas urbanas y rurales, segmentadas por zonas hasta llegar a manzanas y las viviendas con sus respectivos hogares. La Tabla 1 muestra algunas estadísticas descriptivas de ambos instrumentos.

TABLA 1
ESTADÍSTICAS DESCRIPTIVAS DE CASEN 2003 Y CENSO 2002

	Censo nacional de hogares, personas y viviendas. Año 2002	Encuesta Nacional de Caracterización Socioeconómica. Año 2003
Población	15.116.435	257.077
N° de hogares	4.112.838	68.153
N° de comunas	346	302

Para evaluar las interacciones espaciales entre comunas de diferentes regiones en la desagregación microterritorial se utilizan los datos de dos regiones heterogéneas respecto de sus estadísticas descriptivas (Tabla 2) y que además comparten límites administrativos. La Región Metropolitana (RM), que representa el centro urbano más poblado del país con más de 6 millones de personas (prác-

TABLA 2
ESTADÍSTICAS COMUNAS DE LA VI REGIÓN Y RM

Región	Estadísticas	Población Censo	Muestra CASEN
VI Del Libertador	Promedio	31.863	586
General Bernardo	Desviación estándar	44.3476	339
O'Higgins	Máximo	214.344	1.146
	Mínimo	4.221	153
	Número de comunas	33	21
	Nº de observaciones	780.627	12.295
RM Región	Promedio	116.561	1.018
Metropolitana	Desviación estándar	103.011	150
	Máximo	492.915	1.325
	Mínimo	4.435	606
	Número de comunas	52	52
	Nº de observaciones	6.061.185	52.931

ticamente el 40% del total del país) y que alberga a la capital de Chile (Comuna de Santiago). En esta región hay 52 comunas. Conectadas en el límite sur de la Región Metropolitana, la VI Región del Libertado Bernardo O'Higgins está compuesta por 33 comunas y una población promedio de 31.863 personas, con un máximo de 214.344 personas en su capital regional (comuna de Rancagua).

Las Tablas 3 y 4 presentan los valores del ingreso promedio per cápita de los hogares[11] de las 52 comunas de la RM con un ingreso promedio regional de 216.314 pesos y los valores del ingreso para las comunas de la VI Región con un ingreso regional de 93.367 pesos chilenos, respectivamente.

La aplicación del *kriging* a datos de ingreso medio de los hogares por comuna en ambas regiones, muestra una función continua dada por la graduación de colores que representa una segmentación del ingreso según los valores de la tabla. Esto es, las comunas con mayores ingresos se representan con colores más intensos, mientras que las comunas con menos ingresos medios del hogar para cada región es representado por colores más suaves. En la imagen de la izquierda de la Figura 6 aparece la representación cartográfica comunal del ingreso promedio de los hogares de la RM y de la VI Región. En la VI Región la comuna de Rancagua posee el ingreso medio más alto, con 131.014 pesos mientras que en la RM la zona más oscura es conocida como Santiago Oriente y está compuesta por las comunas de Providencia, Ñuñoa, La Reina, Peñalolén, Las Condes, Lo Barnechea y Vitacura, siendo esta última la comuna de mayores ingresos del país con un promedio de 1.124.076 pesos por hogar. Las comunas de la VI Región que no disponen de dato (en blanco) es porque no fueron encuestadas en CASEN.

[11] Se deja fuera del cálculo el servicio doméstico como parte del hogar.

TABLA 3
INGRESO MEDIO DE LOS HOGARES. COMUNAS RM. CASEN 2003
(en pesos chilenos)

Comuna	Ingreso per cápita	Comuna	Ingreso per cápita
Santiago	341.490	Recoleta	144.757
Los Cerrillos	153.675	Renca	134.973
Cerro Navia	110.436	San Joaquín	126.093
Conchalí	133.595	San Miguel	280.783
El Bosque	121.222	San Ramón	107.235
Estación Central	173.114	Puente Alto	166.364
Huechuraba	145.727	Pirque	271.530
Independencia	199.071	San José de Maipo	201.963
La Cisterna	226.518	Colina	176.319
La Florida	162.561	Lampa	119.281
La Granja	119.384	Tiltil	134.179
La Pintana	94.377	Buin	158.433
La Reina	481.642	San Bernardo	121.867
Las Condes	798.388	Calera de Tango	208.699
Lo Barnechea	613.445	Paine	114.638
Lo Espejo	96.652	Melipilla	172.431
Lo Prado	132.284	Alhué	105.382
Macul	277.529	Curacaví	108.816
Maipú	146.733	María Pinto	96.859
Ñuñoa	408.130	San Pedro	121.647
Pedro Aguirre Cerda	124.257	Talagante	170.048
Peñalolén	149.624	El Monte	81.588
Providencia	787.180	Isla de Maipo	106.042
Pudahuel	128.901	Padre Hurtado	119.810
Quilicura	156.256	Peñaflor	149.078
Quinta Normal	143.248	Vitacura	1.124.076

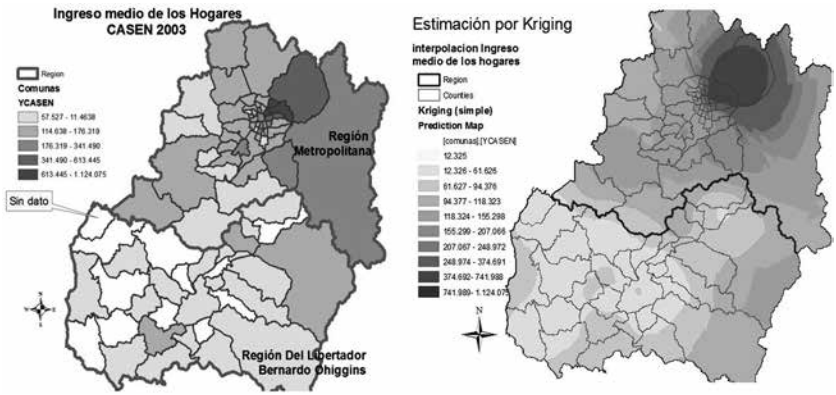
Fuente: CASEN 2003.

TABLA 4
INGRESO MEDIO DE LOS HOGARES. COMUNAS VI REGIÓN. CASEN 2003
(en pesos chilenos)

Comuna	Ingreso per cápita	Comuna	Ingreso per cápita
Chépica	64.634	Pichilemu	110.557
Chimbarongo	82.062	Quinta de Tilcoco	70.076
Coltauco	72.818	Rancagua	131.014
Doñihue	113.860	Rengo	92.067
Olivar	90.417	Requínoa	89.331
Graneros	94.598	San Fernando	110.783
La Estrella	76.162	Mostazal	85.503
Lolol	57.527	San Vicente	100.641
Nancagua	73.341	Santa Cruz	118.276
Machalí	128.063	Pichidegua	113.008
Peralillo	85.967		

Fuente: CASEN 2003.

FIGURA 6
DISTRIBUCIÓN ESPACIAL DEL INGRESO PROMEDIO DE LOS HOGARES
RM Y VI REGIÓN. CASEN 2003



Fuente: Elaboración propia.

La imagen de la derecha de la Figura 6 muestra la estimación por *kriging* del ingreso medio de los hogares por comuna en ambas regiones. La tendencia de la serie muestra que existen diferencias en el valor del ingreso medio al interior de las comunas y que varía por el espacio en diferentes direcciones e intensidades. Se observa un "contagio" en el ingreso medio, ya que aparecen datos clusterizados por sectores espaciales. Comunas donde el ingreso medio es más alto, tienen vecinos con similares características, según la orientación de estos "vecinos", como se observa claramente en el sector noreste de la región metropolitana. La visualización de este evento permite inferir diferencias en los valores del ingreso medio de los hogares en los niveles intramunicipales.

4. Georreferenciación por *matching* espacial

4.1. Distribución espacial del emparejamiento por distritos

El paso de datos del ingreso a nivel de comuna, a datos del ingreso en distritos, dependerá del número de observaciones en CASEN que encuentren emparejamiento entre las unidades del Censo. Lo anterior estaría relacionado con la condición de selección de las unidades primarias de muestreo, que en la encuesta actúa según la probabilidad proporcional del número de viviendas particulares obtenidas en el Censo de 2002. El emparejamiento resultante de la comuna de Padre Hurtado es un ejemplo de esta condición. Esta comuna se ubica en la periferia urbana del sector suroeste de Santiago a lo largo del eje camino que conecta la RM con la zona costera del país. La comuna tiene 38.768 habitantes según el Censo del 2002, se levantaron 920 encuestas en CASEN. Respecto del ingreso, la CASEN entrega un valor medio de 119.810 pesos por hogar (Tabla 5). El ingreso medio de la comuna obtenido como el promedio de los ingresos medios por *matching* espacial de los hogares en cada distrito fue

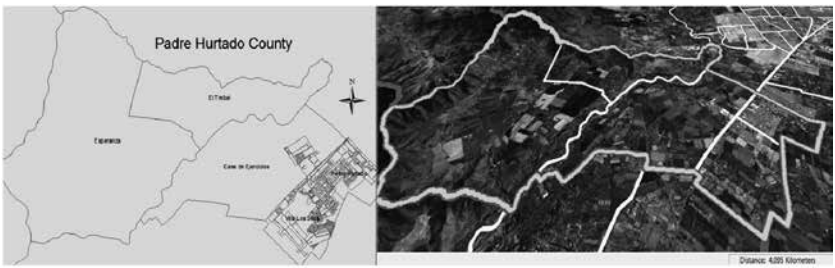
de 106.086 pesos, mientras que la metodología ELL arroja un valor de 114.392 (disponible en Modrego, Ramírez y Tartakowsky, 2008; pag. 63), de la que solo se dispone el valor por comuna y a nivel urbano/rural.

TABLA 5
INGRESO MEDIO DEL HOGAR. COMUNA DE PADRE HURTADO. RM

Población Censo 2002	Muestra CASEN	Ingreso medio hogares CASEN 2003	Ingreso medio hogares *matching* espacial	Ingreso medio hogares ELL
38.768	920	119.810	106.086	114.392

La Figura 7 muestra dos imágenes de esta comuna dividida en cinco distritos. La imagen de la izquierda muestra la cartografía distrital con las secciones por manzana de los distritos Padre Hurtado y Villa Los Silos, que observados satelitalmente en la imagen de la izquierda se comprueba que contienen la zona urbana de la comuna. El emparejamiento mostró que los mayores clones se encuentran efectivamente en los distritos Padre Hurtado y Villa Los Silos con 44% y 42% de los hogares viviendo en ellos, mientras que los distritos Esperanza, El Trebal y Casa de Ejercicios son distritos menos densos y característicos de zonas más rurales.

FIGURA 7
OBSERVACIÓN DE DISTRITOS CON EMPAREJAMIENTO ESPACIAL.
COMUNA DE PADRE HURTADO, RM



Fuente: Elaboración propia.

En particular, el distrito Casa de Ejercicios está compuesto por parcelas de agrado ubicadas a lo largo de la carretera que une a RM con la costa de la V Región. De los 604 hogares identificados en el censo, 115 se emparejaron con 65 jefes de hogar identificados en la encuesta CASEN entregando un ingreso promedio de 159.437 pesos (ver Tabla 6), siendo el distrito de mayor ingreso de la comuna, lo que es coincidente con el más alto CSE_D (6,642).

En el distrito Padre Hurtado, 121 observaciones de CASEN encontraron 1.055 clones en censo equivalente a un factor de expansión de 8,7 personas. Para

el distrito de Villa Los Silos, este valor es de 138 observaciones de la encuesta que equivalen a 851 clones censales. En estos distritos el ingreso promedio de los hogares fue de 133.006 y 137.847 pesos, respectivamente. El ingreso promedio de los hogares obtenido por *matching* entre 373 clones de CASEN y 2.102 clones censales, es de 106.086 pesos.

TABLA 6
INGRESO MEDIO DEL HOGAR POR DISTRITO Y CLASIFICACIÓN CSE_D.
COMUNA DE PADRE HURTADO. RM

Nombre del distrito	Nº de hogares Censo	Clones censales	Clones CASEN	Ingreso *matching* espacial	CSE_D
Casa de Ejercicios	604	115	65	159.437	6.642
Esperanza	522	68	37	56.957	3.116
Padre Hurtado	4.338	1.055	121	133.006	5.424
Villa Los Silos	4.190	851	138	137.847	5.475
El Trebal	167	13	12	43.181	3.514
	9821	2.102	373	106.086	
				Coef. Corr.	0,971187

Otro ejemplo de este proceso es la desagregación espacial del ingreso para la Comuna de Rancagua en la VI Región, que según la CASEN tiene un ingreso medio de los hogares de 131.114 pesos chilenos, mientras que el ingreso estimado por *matching* es de 148.844 pesos. La metodología ELL otorgó un ingreso de 122.005 (pág. 43, Modrego *et al.* 2008). Esta comuna está compuesta por 18 distritos, de estos se encontraron emparejamientos para 17 de ellos con una correlación de 0,817 entre el CSE_D y el valor del ingreso por distrito (Tabla 7). El proceso de emparejamiento no encontró clones para el Distrito El Carmen.

Un acercamiento satelital de la comuna con la división administrativa por distritos (imagen inferior de la Figura 8), en el que se destaca el distrito de El Carmen, permite ver que este distrito está ubicado dentro de un cordón montañoso y cuya planicie muestra algunas zonas agrícolas con muy baja densidad habitacional. Aun cuando esta imagen corresponde al 2015, es posible pensar que en el 2003 había aún menos densidad habitacional.

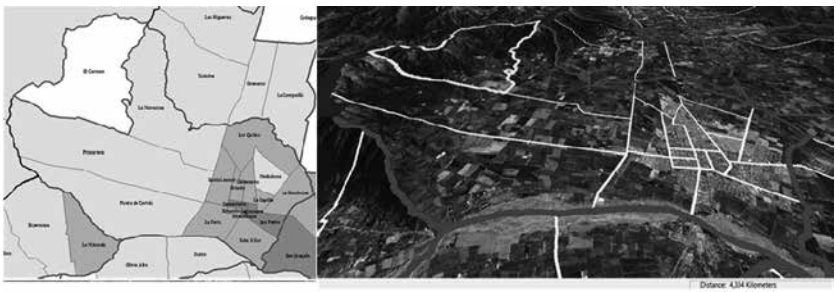
La comuna de Machalí, colindante a la comuna de Rancagua, es un caso que muestra dos eventos:

a) El *matching* espacial encontró que para tres distritos de los 9 que componen la comuna no hay emparejamientos de CASEN con censo. El Pangal, Caletones y El Teniente destacados en la imagen central en la Figura 9 muestra su ubicación en zonas cordilleranas. En los dos de los tres distritos el CSE_D es cero, lo que efectivamente indica que no hay viviendas habitadas en la zona. El grado de correlación entre los valores del ingreso distribuidos en distritos y el CSE_D es del 0,7. En promedio, esta comuna estimada por *matching* espacial da un ingreso promedio de los hogares de 132.449 pesos, mientras que la encuesta CASEN informa de un ingreso

TABLA 7
INGRESO MEDIO DEL HOGAR POR DISTRITO Y CLASIFICACIÓN CSE_D.
COMUNA DE RANCAGUA. VI R

Comuna	Nombre distrito	Ingreso de los hogares *matching* espacial	CSE_D
Rancagua	**El Carmen**	**S/D**	**2.646**
	Intendencia	240.036	7.090
	Estación	101.063	5.536
	Cementerio	200.457	5.590
	Regimiento	269.127	7.188
	Estadio	209.521	6.940
	Centenario	225.351	7.011
	La Capilla	100.365	5.513
	San Pedro	224.273	8.129
	La Gamboina	118.612	5.986
	Los Quilos	155.160	7.469
	Santa Leonor	129.733	6.303
	Primavera	65.879	2.973
	Punta de Cortés	65.221	3.059
	La Feria	133.350	5.663
	Medialuna	77.773	5.457
	Ruta Cinco Sur	166.276	5.616
	La Moranina	48.149	2.781
		148.844	**5.608**
	Coef. Corr.	0,817	

FIGURA 8
OBSERVACIÓN DE EMPAREJAMIENTOS ESPACIAL. COMUNA DE RANCAGUA. VI REGIÓN.



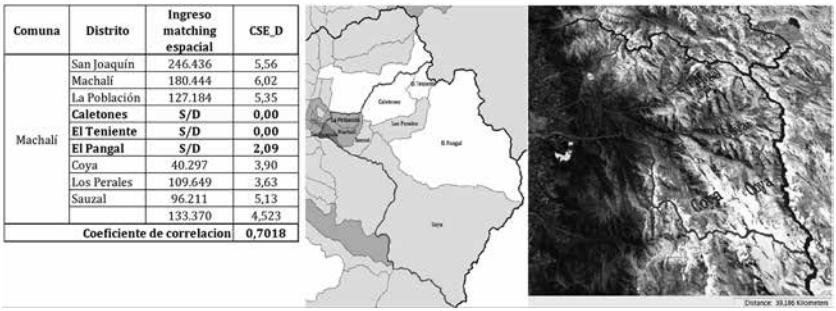
Fuente: Elaboración propia.

comunal de 128.063 pesos y el modelo ELL un ingreso de 120.772 pesos (pag. 44, Modrego *et al.* 2008)

b) El *matching* espacial captura relaciones de contigüidad espacial o autocorrelación espacial entre distritos de estas comunas. El diagrama de Moran dividido en cuadrantes, de la Figura 10, muestra un índice global de autocorrelación de Moran de 0,42, lo que indica que la distribución espacial de los valores

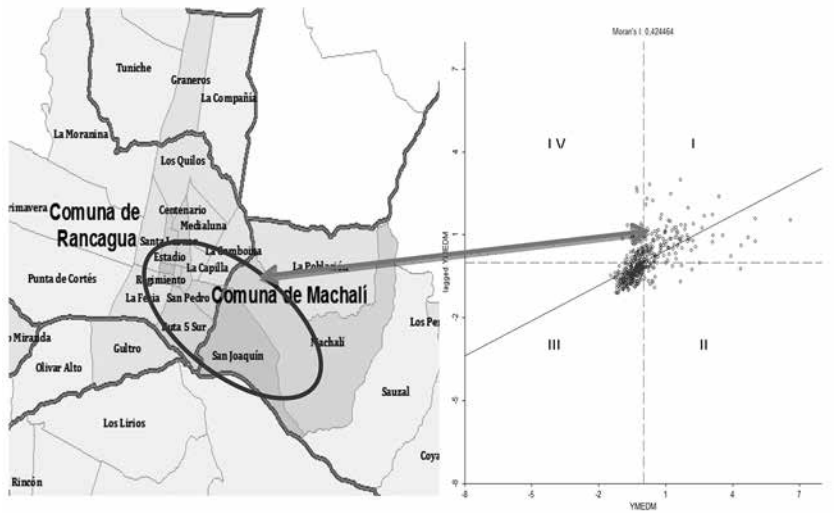
FIGURA 9
OBSERVACIÓN DE DISTRITOS SIN EMPAREJAMIENTO ESPACIAL.
COMUNA DE MACHALÍ. VI REGIÓN

Comuna	Distrito	Ingreso matching espacial	CSE_D
Machalí	San Joaquín	246.436	5,56
	Machalí	180.444	6,02
	La Población	127.184	5,35
	Caletones	**S/D**	**0,00**
	El Teniente	**S/D**	**0,00**
	El Pangal	**S/D**	**2,09**
	Coya	40.297	3,90
	Los Perales	109.649	3,63
	Sauzal	96.211	5,13
		133.370	4,523
Coeficiente de correlacion			**0,7018**



Fuente: Elaboración propia.

FIGURA 10
AUTOCORRELACIÓN ESPACIAL EN DISTRITOS ENTRE COMUNAS DIFERENTES.
VI REGIÓN



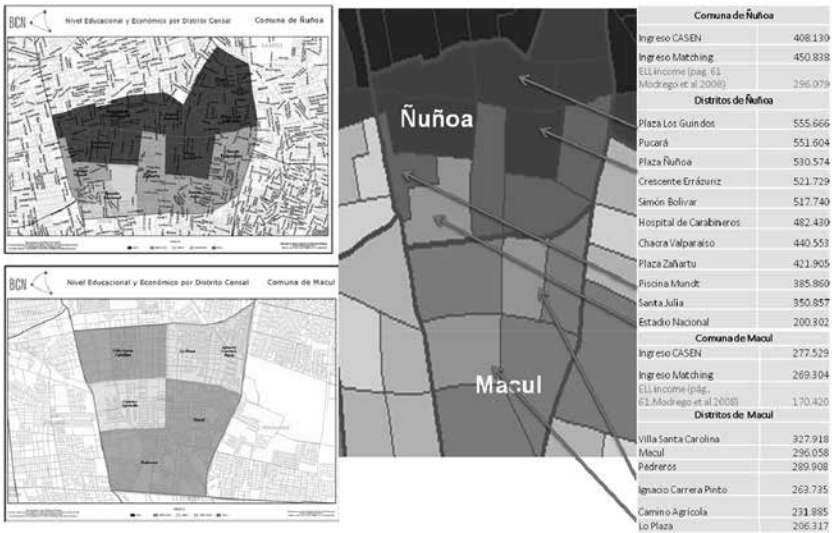
Fuente: Elaboración propia.

altos y bajos del ingreso medio de los hogares está más agrupado espacialmente de lo que se esperaría si los procesos espaciales subyacentes fueran aleatorios. El distrito San Pedro con un ingreso de 224.273 pesos chilenos y Ruta Sur ($ 166.276), Estadio ($ 209.521) y Regimiento ($ 269.127) de la comuna de Rancagua y los distritos de San Joaquín ($ 246.436) y Machalí ($ 180.444) de la comuna de Machalí. Estos distritos pertenecen al cuadrante I de Moran. Distritos que muestran una asociación espacial positiva entre vecinos de altos ingresos medios por hogar.

Un ejemplo aplicado a las comunas de Ñuñoa y Macul ubicadas en la RM se muestra en la Figura 11. En él se ha superpuesto en el extremo izquierdo dos imágenes extraídas desde la Biblioteca Nacional del Congreso que muestran los niveles educacionales y económicos de cada distrito censal de la comuna de Ñuñoa y de la comuna de Maipú, obtenidos mediante análisis factorial[12] (BCN Informe, 2007, pág. 14) con datos del Censo 2002 (http://siit2.bcn.cl/mapoteca/datos_tematicos). La puntuación de los factores califica a los distritos según la graduación de color en alto, medio alto, bajo, medio bajo, bajo.

Nótese que el *matching* espacial aplicado a la obtención del ingreso medio de los hogares desde la encuesta CASEN captura la distribución distrital de los niveles educacionales y económicos recogido con datos de censo (BCN Informe, 2007). Para ambas comunas, todos los distritos parecen coincidir, salvo para la comuna de Macul, que el *matching* reconoce una diferencia en los ingresos medios del hogar entre el distrito de Lo Plaza ($ 206.317) e Ignacio Carrera Pinto ($ 263.735).

FIGURA 11
OBSERVACIÓN DE DISTRITOS COMUNAS VECINAS. RM



Fuente: Elaboración propia.

[12] Los factores que explican el 75,66% de la varianza fueron: Factor 1: Nivel Educacional y Económico; Factor 2: Empleados Urbanos de Nivel Medio y Bajo, Factor 3: Adultos Mayores, Factor 4: Vivienda en altura; Factor 5: Estado Civil Casados y Factor 6: Identidad Alternativa.

4.2. Emparejamiento espacial y su ajuste por comuna y región

Una comparación de la estimación del ingreso medio por hogar obtenido como promedio del ingreso medio por distritos mediante *matching* espacial y el ingreso obtenido por la metodología ELL muestra que las mayores diferencias de estimación se dan en zonas de heterogeneidad espacial con comunas con dependencia espacial positiva entre ellas y próximas a comunas de ingresos bajos. Ejemplo de comunas en esta situación están en la Tabla 8.

El *matching* espacial se aproxima mejor frente a valores atípicos de la serie, como el ingreso de la comuna de Vitacura. Ñuñoa, al ser vecina de Providencia (comuna de ingresos altos) y de Macul (comuna de ingresos bajos), captura el efecto espacial de la autocorrelación positiva (cuadrante I). En el caso de Pirque, Macul y Calera de Tango, son comunas que le rodean otras de ingresos más bajos pero el diagrama de Moran detecta dependencia espacial con otras vecinas.

TABLA 8
COMPARACIÓN INGRESO MEDIO DEL HOGAR POR COMUNAS. RM. CASO 1

Código comuna	Nombre comuna	Ingreso CASEN 2003	Ingreso *matching* espacial	Ingreso ELL
13132	Vitacura	1.124.076	1.168.722	832.970
13123	Providencia	787.180	858.016	455.624
13114	Las Condes	798.388	946.565	521.910
13115	Lo Barnechea	613.445	907.381	527.287
13113	La Reina	481.642	654.659	375.574
13120	Ñuñoa	408.130	450.838	296.079
13101	Santiago	341.490	401.749	212.364
13202	Pirque	271.530	284.128	181.103
13118	Macul	277.529	269.304	170.420
13130	San Miguel	280.783	298.599	213.556
13403	Calera de Tango	208.699	308.483	173.800
	Promedio	**508.445**	**595.313**	**360.062**

Por otro lado, las mejores aproximaciones de ELL se dan en comunas con ingresos muy similares con sus vecinos, en que el método ELL comete los menores errores (Tabla 9), es decir, en situaciones de homogeneidad espacial. El *matching* espacial sobreestima en promedio el valor de estas comunas, pero con un menor error respecto del cometido por ELL en zonas heterogéneas. Todas estas comunas comparten vecindad hasta en segundo grado (vecino de mi vecino) de comunas con ingresos similares.

Por último, una estimación de ingresos medios del hogar en valores regionales muestra que la metodología ELL subestima el ingreso de CASEN para RM (Tabla 10). Esto ocurre porque en RM hay mayor heterogeneidad espacial en el valor del ingreso y el método del *matching* espacial captura mejor estas heterogeneidades (comunas como Vitacura, Las Condes, Providencia) a diferencia de comunas de mayor homogeneidad espacial en valores del ingreso medio del hogar como lo es la VI Región; situación en que las diferencias de estimación del método ELL y *matching* espacial son menores.

TABLA 9
COMPARACIÓN INGRESO MEDIO DEL HOGAR POR COMUNAS. RM. CASO 2

Código comuna	Nombre comuna	Ingreso CASEN 2003	Ingreso *matching* espacial	Ingreso ELL
13602	El Monte	81.588	56.678	84.127
13119	Maipú	146.733	156.498	142.915
13116	Lo Espejo	96.652	101.069	90.910
13603	Isla de Maipo	106.042	61.150	114.163
13401	San Bernardo	121.867	140.202	127.588
13504	María Pinto	96.859	68.699	80.740
13121	Pedro Aguirre Cerda	124.257	120.849	118.839
13131	San Ramón	107.235	111.042	104.107
13503	Curacaví	108.816	83.503	121.695
13112	La Pintana	94.377	111.499	87.707
13105	El Bosque	121.222	134.841	108.565
13103	Cerro Navia	110.436	133.272	92.375
13117	Lo Prado	132.284	117.168	112.021
13126	Quinta Normal	143.248	130.854	132.010
13111	La Granja	119.384	112.239	102.363
13104	Conchalí	133.595	137.926	112.029
	Promedio	**115.287**	**117.093**	**108.260**

TABLA 10
COMPARACIÓN INGRESO MEDIO DEL HOGAR POR REGIÓN

Nombre región	Ingreso CASEN 2003 por comunas	Ingreso *matching* espacial promedio distritos	Ingreso ELL por comuna
Región Metropolitana	$ 216.314	$ 235.783	$ 169.295
VI Región	$ 93.367	$ 98.016	$ 78.435

5. Conclusiones

Esta investigación ha propuesto una metodología variante de los modelos de estimación en pequeñas áreas, que proyecta valores de una encuesta de hogares a la población censal, aprovechando la información geográfica recolectada en el Censo y las variables similares que son recolectadas en ambos instrumentos, poniendo especial énfasis en municipios para los cuales la información existente es muy agregada y escasa.

Para realizar el aprovechamiento de la información geográfica que está en el Censo y las variables de la encuesta que no están en él, se ha utilizado un proceso que consiste en el apareamiento entre las observaciones que están en la encuesta con clones encontrados en el Censo mediante una metodología denominada *matching* espacial, que permite localizar las observaciones de la

encuesta en los distritos del Censo. El trabajo muestra un ejemplo utilizando el ingreso medio de los hogares disponibles en la encuesta de hogares.

Los resultados pueden ser visualizados en mapas georreferenciados y permite apreciar la heterogeneidad al interior de las comunas, lo que es ignorado por la técnica de estimación en pequeñas áreas donde los resultados se han contrastados. Los resultados de esta metodología permiten visualizar también que existen áreas internas a las comunas que se comportan más parecido a sus vecinos de la comuna adyacente que como los vecinos de su propia comuna, lo que es conocido como autocorrelación espacial y que la técnica utilizada toma en cuenta, mientras que esta interacción espacial intermunicipios tiende a ser ignorada por las técnicas tradicionales.

En resumen, el aporte de este trabajo consiste en adicionar a las técnicas actuales de interpolación o extrapolación de información desde encuestas a censos, elementos espaciales presentes en el proceso. Estos elementos son la heterogeneidad e interacción espacial. Los resultados muestran que ambos procesos están presentes en el ejemplo desarrollado y que la información generada puede ser muy valiosa a la hora del diseño e implementación de políticas públicas.

Un trabajo futuro consistirá en evaluar la ganancia de información al incorporar estos elementos espaciales en la información generada, proponiendo un procedimiento que permita estimar los datos ausentes en CASEN al contrastar los resultados derivados de las técnicas tradicionales como ELL. Se plantea como trabajo futuro, mostrar la robustez estadística de la existencia de diferencias significativas entre la estimación por *matching* espacial y la metodología ELL en zonas de heterogeneidad espacial con comunas de dependencia espacial positiva entre ellas y cercanas a comunas de menores ingresos.

6. Referencias

Agostini, C.; Brown, P. y Góngora, D. (2008). Distribución Espacial de la Pobreza en Chile. *Estudios de Economía*. 1(35), 79-110.

Agostini, C.; Brown, P. y Román, A. (2010). Estimando Indigencia y Pobreza Indígena Regional con Datos Censales y Encuestas de Hogares. *Cuadernos de Economía* (47), 125-150.

Agterberg, F. (2004). "George Matheron: Founder of Spatial Statistics". Earth Sciences History, 23 (2), 325-335.

BCN Informe (2007). Caracterización Educacional y Económica de Chile a partir del Censo 2002". Biblioteca del Congreso de Chile, Área Sistema Integrado de Información Territorial, Asesoría Parlamentaria BCN. 19 pág.

Cornejo, R.; Navarrete, M.; Valdivia, R.; Aroca, P. y Aracena, S. (2014). Desarrollo de una base de datos integrada de Censo y encuesta mediante el uso de elementos de inteligencia de negocios y SIG. *Ingeniare. Revista chilena de ingeniería*, *22*(2), 205-217.

Dehejia, R. and Wahba, S. (2002). Propensity score- matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, 84(1), 151-161.

Elbers, C., J.O. Lanjouw, P. Lanjouw (2003). Micro-Level Estimation of Poverty and Inequality. *Econometrica*, 71(1), 355-364.

Hentschel, J., Lanjouw, J.; Lanjouw, P. y Poggi, J. (2000). Combining Census and Survey Data to Trace the Spatial Dimensions of Poverty. *The World Bank Economic Review* 14(1), 147-65.

Lam, N. (1983). Spatial Interpolation Methods: A Review, *The American Cartographer*, 10 (2), 129-149.

Minot, N. and Baulch, B. (2005). Spatial patterns of poverty in Vietnam and their implications for policy. *Food Policy*, 30 (5-6), 461-475.

Modrego, F., Ramírez, E. y Tartakowski, A. (2008). La heterogeneidad espacial del desarrollo económico en Chile: radiografía a los cambios en bienestar durante la década de los 90 por estimaciones en pequeñas áreas. Documento de Trabajo Nº 9. Programa Dinámicas Territoriales Rurales Rimisp-Centro Latinoamericano para el Desarrollo Rural.

Modrego, F., Ramírez, E., Tartakowski, A. y Jara, E. (2012). "La heterogeneidad territorial del desarrollo en la década de oro de la economía chilena", en Modrego, F. y J.A. Berdegué (Eds.), *Los Dilemas Territoriales del Desarrollo en América Latina*. Editorial Teseo. Buenos Aires, Argentina (En prensa).

Mohamed, A. y Mohamed, A. (2009). Spatial Patterns And Geographic Determinants Of Welfare And Poverty In Tunisia. Working Papers , 478.

Molina, I. y Rao, J. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 38, 369-385.

Montero, J. y Larraz, B. (2008). Introducción a la Geoestadística Lineal, Netbiblo, España. 142 p.

Moran, P. (1948). "The interpretation of statistical maps". *Journal of the Royal Statisical Society*. Series B (Methodological). 10 (2), 243-251.

Navarrete, M (2012). Métodos geoestadísticos del precio de la vivienda. Una aproximación al conocimiento urbano de la ciudad de Madrid. Tesis Doctoral Universidad Autónoma de Madrid.

Paredes, D. y Aroca, P. (2008). "Metodología para estimar un índice regional de costo de vivienda en Chile". *Cuadernos de Economía*, 45, 129-143.

Rubin, D. (1976). "Matching Methods that are Equal Percent Bias Reducing: Some Examples". *Biometrics*, 32, 109-120.

Rubin, D. (1973). "Matching to Remove Bias in Observational Studies," *Biometrics*, 29, 159-183.

Sowunmi, F., Akinyosoye V., Okoruwa V. and Omonona B. (2012). The Landscape of Poverty in Nigeria: A Spatial Analysis Using Senatorial Districts-level Data. *American Journal of Economics*, 2(5): 61-74.

Tarozzi, A. and Deaton, A. (2009). Using Census and Survey data estimate poverty and inequality for small areas. *The Review of Economics and Statistics,* November 91(4), 773-792.

Tzavidis, N., Salvati, N., Pratesi, M. y Chambers, R. (2008). "M-quantile models with application to poverty mapping". *Statistical Methods and Applications*, 17, 393-411.

Anexo 1

Algoritmo *matching* espacial

BEGIN
{
$k = 1$, $K(yk) = \phi$;
While ($k \leq M$):
{
While ($\lambda \leq N$):
{
$j = 1$;
If($(y^j_{id})_k=(x^j_{id})_\lambda$ y $j \leq 8$), then:
{
If ($j = 8$), then:
{
$K(yk) = K(yk) \cup \{x_\lambda\}$;
}
$j = j + 1$;
}
$\lambda = \lambda + 1$;
}
$k = k + 1$;
}
}
END