Revista de Geografía Norte Grande, 81: 137-160 (2022)
Artículos

Recibido el 28 de julio 2020 / Aprobado el 17 de junio 2021 / Publicado el 1 de mayo 2022 ISSN 0718 3402

La importancia del espacio geográfico para minimizar el error de muestras representativas[1]

The importance of geographic space to minimize the error of representative samples

Ricardo Truffello[2,3,4], Mónica Flores[4], Matías Garretón[5,6] y Gonzalo Ruz[7,8,9]

RESUMEN

En el presente trabajo se discute la importancia del espacio geográfico en el contexto de la generación de marcos muestrales de encuestas, poniendo en tensión la premisa estadística tradicional de la aleatoriedad e independencia de las observaciones. Para esto se analiza el aporte de la geografía cuantitativa en la generación de metodologías de regionalización que permitan, de manera efectiva, mejorar el error muestral de las encuestas, enfocados principalmente en las áreas urbanas, en presencia de variables de estratificación con autocorrelación espacial.

Se testean de forma empírica algoritmos de regionalización con y sin procesos de optimización heurística, utilizando datos censales, para posteriormente definir el nivel de error y establecer comparaciones contra muestreos tradicionales de corte aleatorio y aleatorio bi-etápico, por medio de un procedimiento Montecarlo.

Los resultados obtenidos dan cuenta de una disminución de hasta un 20% en el error contra metodologías tradicionales o en su defecto la disminución de hasta 100 casos con el mismo nivel de error. Se concluye que las metodologías de muestreo espacializado con optimización heurística ofrecen ventajas evidentes en áreas urbanas, en presencia de autocorrelación espacial.

Palabras Clave: Regionalización, estratificación espacial, muestreo espacializado

ABSTRACT

This paper discusses the importance of geographic space in the context of generating a sample framework for surveys, questioning the traditional statistical premise of randomness and independence of the number of observations. The contribution of quantitative geography in the generation of regionalization methodologies is analyzed, since these

[1] Este artículo pertenece al proyecto ANID de iniciación N 11221028

[2] Instituto de Estudios Urbano y Territoriales, FADEU, UC. Correo electrónico: rtruffel@uc.cl

[3] El autor agradece al Centro de Desarrollo Urbano Sustentable (CEDEUS) ANID/FONDAP/15110020

[4] Observatorio de Ciudades UC, FADEU, UC. maflorescas@gmail.com

[5] Escuela de Diseño, Universidad Adolfo Ibañez. Correo electrónico: matias.garreton@uai.cl

[6] Centro de Estudios de Conflicto y Cohesión Social - COES - ANID/FONDAP/15130009

[7] Facultad de Ingeniería y Ciencias, Universidad Adolfo Ibáñez. Correo electrónico: gonzalo.ruz@uai.cl

[8] Center of Applied Ecology and Sustainability (CAPES), Santiago, Chile

[9] Agradecimiento ANID FONDECYT 1180706, ANID PIA/BASAL FB0002

allow the improvement of the sampling error of the surveys, focusing mainly on urban areas, and in the presence of stratification variables with spatial autocorrelation.

Regionalization algorithms with and without heuristic optimization processes are empirically tested, using census data, to subsequently define the level of error and establish comparisons against traditional random and two-stage random sampling, using a Monte Carlo procedure.

The results obtained show a decrease of up to 20% in error against traditional methodologies or alternatively, a reduction of up to 100 cases with the same level of error. It is concluded that spatialized sampling methodologies with heuristic optimization offer advantages in urban areas, in the presence of spatial autocorrelation.

Keywords: Regionalization, spatial stratification, spatial sampling.

Introducción

La diversificación de procesos y la intensificación de cambios en la sociedad otorga una importancia creciente al desarrollo de metodologías eficientes de recolección de datos relevantes para la geografía humana. Considerando la invalidación del censo de población 2012 (Bravo et al., 2013), el censo 2017 abreviado y los problemas de representatividad comunal de la encuesta de caracterización socioeconómica, se hace necesario plantear nuevas metodologías que permitan generar una alternativa competitiva a las metodologías de muestro tradicional.

Esto bajo el entendido de que una gran cantidad de las encuestas actuales utilizan marcos muestrales basados en censos de población o bien métodos aleatorios convencionales, lo que en un territorio con fenómenos de autocorrelación espacial puede producir sesgos y duplicación de información (Griffith, 2005).

De esta forma el objetivo de la presente investigación es verificar el rendimiento de técnicas de estratificación espacial (muestro espacializado) y compararlas con metodologías tradicionales de corte aleatorio de manera de verificar cómo la incorporación del espacio en los diseños muestrales permite mejorar de manera significativa los niveles de error.

Para esto se trabajará con un marco de referencia muestral basado en el Censo de 2017 para el área urbana consolidada de Santiago (1,8 millones de hogares), usando una variable de segmentación socioeconómica y testeando tres metodologías de clusterización con y sin procesos de optimización heurística, estas son: Max-p, Grouping Análisis y REDCAP.

Los resultados serán contrastados con dos métodos tradicionales de muestreo: muestro aleatorio simple, sin la incorporación del espacio y un muestro bi-etápico considerando las zonas censales como unidad geográfica muestral (disponible como unidad mínima para el censo 2017).

Los resultados serán analizados a partir del porcentaje de diferencia sobre el error muestral y verificando la disminución de casos a iguales niveles de error. A partir de esto se explorará cuáles son los factores de diseño muestral determinantes en las variaciones expuestas en los resultados.

Importancia y desafíos del levantamiento de datos sociodemográficos

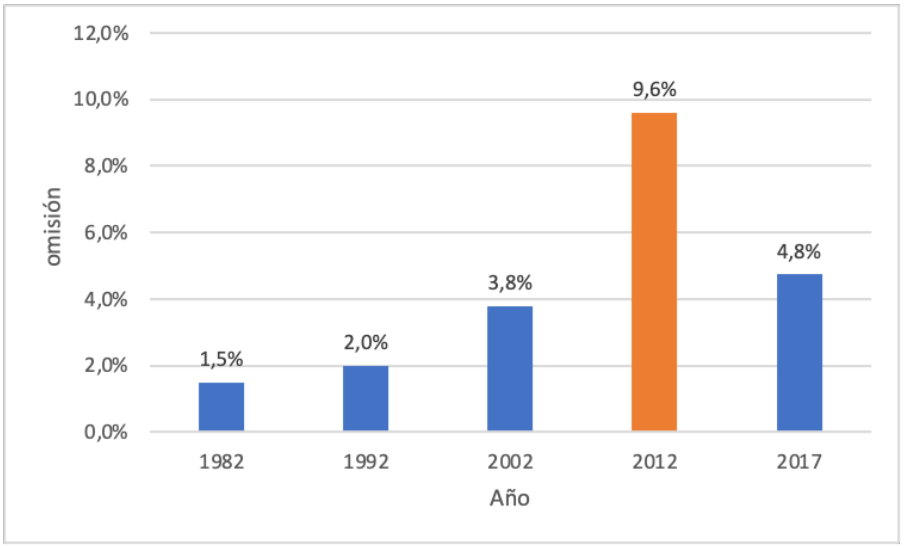
En el siglo XX, la necesidad de planificar y focalizar los recursos económicos, en conjunto con el desarrollo de las ciencias sociales y el auge del modelo capitalista, sentaron las bases para que los censos de población se convirtieran en una herramienta fundamental para los Estados en su esfuerzo para orientar planes y proyectos urbanos y territoriales (Yates, 1946).

En particular para las áreas urbanas los censos son, probablemente, el instrumento con mayor validación y uso en la focalización de políticas públicas en Latinoamérica, siendo para algunos países la única información disponible (Arretx, 1989).

No obstante, la metodología de conteo totalizante supone altos costos monetarios y largos tiempos de levantamiento y procesamiento. Más aún, la información censal, por sus características como instrumento de conteo demográfico exhaustivo, puede presentar errores relevantes, situación que se acentúa en las ciudades más pobladas y áreas metropolitanas (Cohen, 2006).

Un buen ejemplo de esto es lo que ha sucedido en Chile a partir de la aplicación del Censo de 2012, el cual fue invalidado por tener un margen de error muy alto (9,6%). Como consecuencia natural de la evolución demográfica de Chile, los censos de población han ido aumentando progresivamente su error por omisión (ver Figura N°1). Esta situación, en países desarrollados, ha fomentado el recambio del instrumento por metodologías complementarias (encuestas focalizadas, encuestas generales, panel y metodologías de corrección) o en su defecto por el uso de registros administrativos (Borchsenius, 2001; Cook, 2004)

Figura N°1.
Omisión Censos de Población en Chile

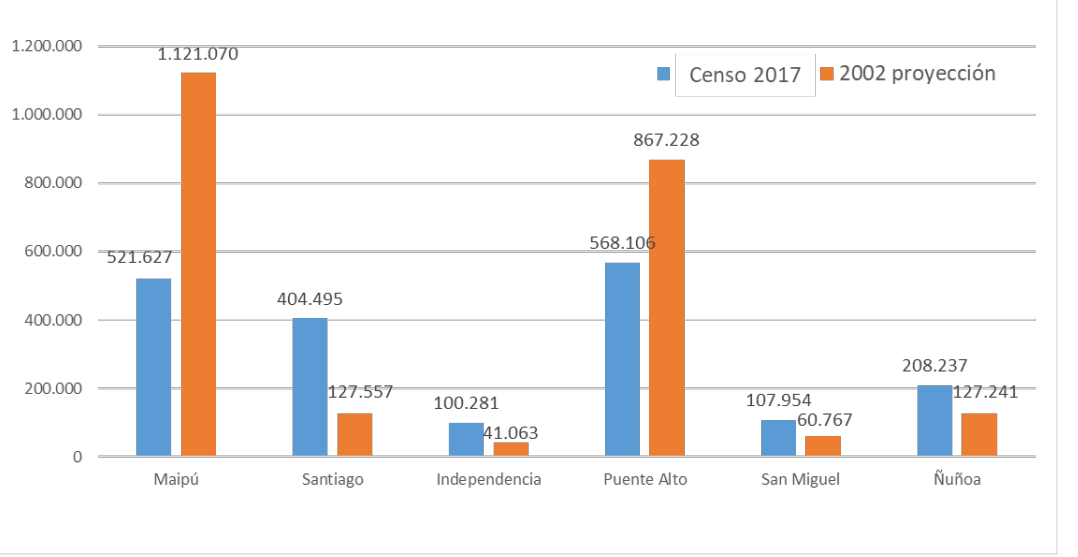


Fuente: Elaboración propia en base a Villalón & Vera, 2011; Bravo, Larrañaga, Millán, Ruiz, & Zamorano, 2013

El impacto de los errores en los censos de población no queda aislado en el instrumento, sino que es heredado a todas aquellas encuestas que ocupan como base su marco muestral. De hecho, en muchos casos estas encuestas son justamente aquellas de mayor relevancia y con muestreos masivos y por lo tanto costosos, como es el caso de la encuesta CASEN[10] en Chile.

Para el caso de la encuesta CASEN, con la invalidación del censo de 2012, se tuvo que continuar utilizando como base, el marco muestral del censo de 2002 y sus proyecciones de población al año 2017, en base al período intercensal 1992-2002. Esto a nivel comunal produce serias discordancias invalidando en muchos casos su representatividad territorial.

Figura N°2.
Población comunal - Habitantes



Fuente: Elaboración propia en base a Censo 2017, INE; CASEN 2017.

Como se observa en la Figura 2, la brecha demográfica entre la proyección poblacional y la población censada a nivel comunal puede ser relevante y esto es especialmente significativo en casos como la comuna de Santiago, en donde se estimaba una población de alrededor de 130.000 habitantes y fenómenos demográficos e intervenciones de planificación territorial, como el proceso de regeneración urbana, dan cuenta de más de 400.000 habitantes efectivamente censados, generando una obvia subestimación de la muestra CASEN con la siguiente invalidación de la representatividad del instrumento para dicho territorio.

Ante este tipo de situaciones, la evolución natural de las metodologías de levantamiento demográfico, en especial en los países del primer mundo, ha cambiado gradualmente por el desarrollo de encuestas espacialmente representativas y sistemas estadísticos basados en registros (Borchsenius, 2001; Cook, 2004; Wallgren & Wallgren, 2016). Ambas metodologías, sobre todo las

[10] Encuesta de caracterización socioeconómica nacional.

basadas en registros, permiten mejorar la consistencia y coherencia de los datos demográficos, dando además mayor disponibilidad temporal de los mismos (Wallgren & Wallgren, 2007, 2016)

No obstante, estos métodos implican grandes costos monetarios y largos plazos de implantación, con metodologías muchas veces complejas de abordar por parte del estado. En la práctica esto implica: mejorar los registros y las plataformas tecnológicas que sustentan la generación de métodos de gestión de Big Data, capacitación de personal especializado, gestión continua de datos (Jin et al., 2015) e implementación de Infraestructuras de Datos Espaciales (IDE) para asegurar la consistencia espacial y coherencia entre fuentes, así como también su correcta representación geográfica (Williamson et al., 2006).

La calidad y representatividad de los resultados demográficos, por medio de instrumentos alternativos a los censos, se convierte entonces en una prioridad. La generación de metodologías que incluyan un marco muestral basado en el espacio, no sólo como contenedor de distribución de los datos, resulta fundamental, ya que muchas de estas metodologías carecen de la especificidad territorial (Rodríguez-Iglesias & Teresa, 2010), generando sub-representatividad estadística de parte de la población objetivo.

El muestreo espacializado aparece entonces como un cambio en las estrategias de diseño, estratificando la muestra para optimizar la generación de un marco muestral con menores niveles de error. Esto debido a que modula el principio de aleatoriedad del muestreo asumiendo que éste puede estar sesgado por la configuración natural del espacio geográfico y sobre todo urbano (fenómenos de autocorrelación espacial) y que por ello la independencia entre las observaciones no es efectiva. En contrapartida, el muestreo espacializado agrega una componente de incerteza estadística relevante (J. Wang et al., 2010), por lo que aparece como una necesidad establecer cuál es el mejor acercamiento metodológico para reducirla y verificar si efectivamente el espacio puede aportar al diseño muestral en zonas urbanas con alta complejidad y heterogeneidad sociodemográfica.

Ley de los grandes números versus procesos de aglomeración espacial

Encuestas y enfoque tradicional estadístico

Uno de los instrumentos complementarios a los censos de población son las encuestas. Éstas permiten a partir de una muestra acotada, obtener estadísticas representativas de un grupo de población mayor o bien de la población total. Comparándolas con los métodos de enumeración exhaustiva tienen grandes ventajas, tales como costos reducidos, mayores velocidades de aplicación, permiten focalizarse en temas más específicos y, por lo general, tienen menos porcentajes de omisión (Cochran, 1977).

Dicha representatividad depende de factores como la escala, el número de individuos y los métodos de muestreo que se utilicen. De esta forma, el principal tópico que ha abordado la estadística, en el proceso de optimización de las encuestas, es el de la representatividad, equilibrando efectividad, rapidez (Yates, 1946) y valor monetario.

En tal sentido, el paradigma que ha dirigido la representatividad en las encuestas es la aleatoriedad, bajo la premisa fundamental que las observaciones son "independientes e idénticamente distribuidas" (Brus & De Gruijter, 1997; Stock, Ward, Thorson, Jannot, & Semmens, 2019: 1), lo que no es necesariamente cierto en la distribución espacial de los datos. Bajo esta premisa es necesario escapar de la poco realista inferencia de independencia, para entrar a lo que se conoce como modelos dependientes, ya sea con correlaciones intraclase o bien estructuras seriales de correlación, en donde dicha dependencia puede estar presente en todas las direcciones, pero probablemente sea más débil en cuando más dispersos sean las observaciones (Cressie, 1993).

Espacio y autocorrelación espacial

Las relaciones espaciales y el espacio como sujeto de estudio, es una de las concepciones principales de la Geografía. Su alcance es transversal a las diferentes subáreas de conocimiento como la geografía económica, geografía física y geografía humana. En ese contexto, la "autodenominada" primera ley de la geografía, "Primera ley de Tobler define lo siguiente: Todo está relacionado con todo, pero las cosas cercanas están más relacionadas que las cosas distantes" (Tobler, 1969). Esto es una referencia directa a los principales conceptos del análisis espacial y el modelamiento más específicamente derivado de la estadística espacial, relacionado de forma unívoca con la autocorrelación espacial y las relaciones dadas por la conectividad topológica (Lin & Wang, 2019; Miller, 2004).

La autocorrelación espacial entonces, es un fenómeno transversal a todo el espacio geográfico, comúnmente asociada con las relaciones ecológicas, comportamiento personal y fenómenos urbanos. Este fenómeno representa un problema respecto a la representatividad de los instrumentos muestrales de medición demográfica, debido a que los datos autocorrelacionados violan la premisa de independencia de las observaciones en la mayor parte de los procedimientos estadísticos (Legendre, 1993). No podemos desconocer que esta transgresión estadística sucede de forma efectiva y particularmente con el espacio entendido no sólo como el lugar o contenedor en donde habita y se relaciona el hombre, sino que el espacio concebido como la trialéctica definida por Henry Lefebvre. De hecho, la misma concepción del espacio como práctica y lugar que puede auto reproducirse (Lefebvre, 1991), da cuenta de la relación con el concepto de la autocorrelación espacial.

Esta contraposición o paradoja del concepto de autocorrelación, entre problema y cualidad da cuenta de la necesidad de profundizar en las discusiones interdisciplinarias relacionados con el levantamiento de muestras estadísticas, de forma de permitir el desarrollo de métodos consistentes que aprovechen dichas cualidades y que module los problemas de transgresión de la independencia estadística. Desde el punto de vista metodológico se debe profundizar entonces el rol de la estratificación espacial y el muestreo espacializado.

Muestreo espacializado y regionalización

La estructura espacial de la información involucra a la vez fenómenos de autocorrelación - cuando existe una similitud mayor entre observaciones cercanas - y de heterogeneidad - cuando

los parámetros estudiados varían dependiendo de la localización en que se miden (Brus & De Gruijter, 1997; Griffith, 2005; J.-F. Wang et al., 2013; A Review of Spatial Sampling, 2012). A mayor autocorrelación espacial, mayor es la cantidad de información duplicada y menor es la varianza entre observaciones próximas, por lo que el tamaño muestral efectivo - entendido como el número de observaciones independientes - puede ser menor que la muestra obtenida (Griffith, 2005; Vallejos & Osorio, 2014). Además, al existir heterogeneidad espacial, es imprescindible tener una buena cobertura geográfica para garantizar la representatividad de la muestra, lo que puede obtenerse mediante un diseño que complemente este objetivo con el de representatividad demográfica (Griffith, 2005; Pettitt & McBratney, 1993; Wang *et al*., 2010).

En general, el objetivo del muestreo espacializado es generar un diseño experimental óptimo, considerando a la vez las características de una población y la distribución espacial de ésta, para generar el máximo de información relativa al campo de análisis (Lindley, 1956). Este objetivo puede formularse en términos de un muestreo con máxima entropía, que permita capturar en la muestra la mayor variabilidad posible de la población en estudio (Shewry & Wynn, 1987) y minimizar la redundancia de datos en presencia de autocorrelación espacial.

Para esto, diversas disciplinas han abordado el muestro espacializado como una técnica que permita mejorar la representatividad de las encuestas. Desde el ámbito territorial, es imprescindible considerar el sesgo natural que impone el espacio en las condiciones que permiten aproximarse al principio de la aleatoriedad (Griffith, 2005; J.-F. Wang *et al*., 2013; Wang *et al*., 2012). Este sesgo ha sido relevado y probado desde las ciencias de la tierra, principalmente a partir de la geostadística – estrategias basadas en modelo - (Brus & De Gruijter, 1997; de Gruijter & ter Braak, 1990), pero también como estrategias basadas en diseño (estadística espacial). Para las ciencias sociales esto sólo ha sido demostrado a nivel empírico simulado, dada la complejidad de los territorios en donde importa verificarlo, las ciudades, específicamente las grandes áreas metropolitanas; no obstante, la teoría clásica, como la empírica derivada de la observación de los procesos de autocorrelación en las ciudades, sustentan esta hipótesis (Lefebvre, 1991; Miller, 2004; Tobler, 1969).

Estratificación

La estratificación contribuye a parcelar una población en subpoblaciones con menor varianza, bajo el supuesto de que existe una estructura en el universo estudiado, aproximación que es ampliamente utilizada desde una perspectiva demográfica. Esto resulta más complejo si se considera la localización de esta población. En efecto, la heterogeneidad de un campo aleatorio geográfico involucra dos elementos: la varianza global de la población y la estructura espacial de esta varianza, debiendo considerarse ambos para un correcto diseño muestral (Griffith, 2005). En este caso, como ya argumentamos con anterioridad, el supuesto de que la población estudiada es independiente e idénticamente distribuida es falso, por lo que resulta inadecuado utilizar diseños de muestreo tradicionales (Wang, Stein, Gao, & Ge, 2012).

El acercamiento tradicional a la estratificación consiste en subdividir la muestra, en muestras más homogéneas, esto permite que la aleatoriedad funcione sobre un marco de menor varianza y por lo tanto mejora los niveles de error, pero a costo de un *n* más elevado. Por lo tanto, a máxima

varianza se deben considerar 348 casos[11] por estrato lo que eleva de manera significativa el tamaño y costo de las encuestas.

Una alternativa es incorporar el espacio en la estratificación, por medio de los métodos de clusterización, regionalización o interpolación (Brus et al., 2019), es decir hacer que por medio de la captura del fenómeno de autocorrelación espacial disminuya la varianza dentro de zonas homogéneas, estratificando de esta manera la muestra y permitiendo reducir casos redundantes.

Para abordar la estratificación espacial existen dos grandes enfoques: las estrategias basadas en modelos, propias de la geoestadística y las basadas en diseño de la estadística espacial.

Estrategias basadas en modelos

El muestreo y estimación basados en modelos permiten realizar interpolaciones bastante precisas con un mínimo de observaciones (Heaton & Gelfand, 2012; Pettitt & McBratney, 1993; Phillips et al., 2006; Stein & Ettema, 2003; Vallejos & Osorio, 2014). Esta estrategia ha sido ampliamente utilizada en el ámbito de la geografía física y del estudio de ecosistemas, pero su utilización es discutible en muchos casos. En particular, el muestreo basado en modelos requiere de supuestos fuertes, implica la definición de una función objetivo y limita la pertinencia de los resultados a las variables definidas en ésta (Brus y de Gruijter, 1997; Gruijter y Braak, 1990; Wang et al, 2012). En consecuencia, desde el ámbito de las ciencias sociales, es preferible utilizar técnicas de muestreo con un mínimo de supuestos, más flexibles y aptas para el análisis simultáneo de múltiples variables, en desmedro de estimaciones basadas en modelos que son más adecuadas para la predicción a partir de interpolación (Wang et al., 2010).

Muestreo Basado en Diseño y Muestreo Espacializado

Técnicas tradicionales de muestreo basado en diseño, como son la selección aleatoria simple, sistemática, con diseño estratificado y/o bi-etápico, pueden ser extendidas y combinadas para elaborar metodologías de muestreo aptas para información que presenta una estructura espacial (Griffith, 2005; Wang *et al*., 2013; Wang *et al*., 2012). Esta estrategia implica menos suposiciones que el muestreo basado en modelos, también permite evitar los errores asociados a la autocorrelación y heterogeneidad espacial y es altamente recomendable en casos en que no existe suficiente información previa acerca de la distribución de las variables de interés (Brus & De Gruijter, 1997; de Gruijter & ter Braak, 1990). En este caso, debe considerarse dos factores fundamentales para el diseño muestral: la inclusión de índices de autocorrelación espacial en la estimación de los parámetros de interés y del error muestral, y la generación de un diseño que permita una cobertura espacial adecuada (Griffith, 2005).

Diversos diseños de muestreo espacial estratificado permiten cumplir esta última condición, utilizando una combinación de métodos de selección aleatoria simple y sistemática. Esta estrategia es usualmente implementada en dos etapas. En primer lugar, se genera una partición o jerarquización exhaustiva del espacio y luego se selecciona aleatoriamente a uno o más individuos en

[11] Con máxima varianza, población finita y un 95% de confianza el *n* de una encuesta tiende a estabilizarse en 348 para poblaciones mayores a 35.000 habitantes, por lo tanto, para más de 6 millones se deja como número fijo, lo mismo para la muestra estratificada que es un tercio de 6 millones.

cada una de las áreas así definidas. La etapa de selección sistemática permite generar una cobertura espacial adecuada para evitar sesgos y omisiones derivados de la heterogeneidad espacial, ya sea mediante control de distancias promedio a vecinos más cercanos, grillas regulares o definición de zonas homogéneas (Brus & De Gruijter, 1997; Griffith, 2005; Wang et al., 2010). Existe evidencia teórica y empírica que muestra una mayor precisión del muestreo con grillas regulares y selección aleatoria, en comparación con métodos de selección aleatoria simple o basada en modelos (Griffith, 2005). Metodologías más recientes utilizan una partición espacial en zonas homogéneas, lo que permite además controlar los sesgos y errores derivados de la autocorrelación espacial, reduciendo los errores de estimación e incrementando el tamaño muestral efectivo con un mismo número de observaciones (Wang *et al*., 2013; Wang *et al*., 2012).

Métodos de Regionalización

Los algoritmos de regionalización se desarrollaron con profusión cuando la geografía da un giro hacia lo cuantitativo, durante los años 70, en consonancia con los avances de la computación. Puntualmente alcanza su máxima profusión al final de dicha década, a partir de los numerosos trabajos de Openshaw (Openshaw, 1977, 1978; Openshaw & Baxter, 1977).

En términos generales estos métodos giran en torno al concepto geográfico de zona o región (Montello, 2003) y cómo delimitarlo en función de sus características. Cuando esto se combina a partir de métodos computacionales, el objeto se transforma en la definición de criterios de optimización a partir de las posibilidades derivadas de la combinatoria definida por la contigüidad, funciones objetivos y restricciones, las que además deben considerar los efectos espaciales del Problema de la Unidad Espacial Modificable (MAUP, por sus siglas en inglés), la autocorrelación espacial y la multicolinealidad (Garreton & Sánchez, 2016).

En tal sentido la regionalización, como método tiene numerosos puntos en común con los algoritmos de clusterización, cuando su diseño está pensado para dos dimensiones (Folch & Spielman, 2014). Sin embargo, en el caso de los métodos de regionalización, la topología, entendida en su dimensión de contigüidad, debe ser una condición fundamental (distancia física además de la función objetivo de optimización). En función de estos preceptos aparecen diferentes aproximaciones algorítmicas para intentar resolver de manera consistente y plausible[12].

Tipologías de Regionalización

Las tipologías de regionalización se abordan de forma diferente dependiendo del enfoque del estudio, pero básicamente se pueden establecer dos categorías principales: revisiones basadas en el objetivo del modelo de regionalización y las basadas en la descripción algorítmica de este. Dada la importancia que tiene la elección del método, es que se establecerá una clasificación orientada a la descripción algorítmica, ya que las basadas en objetivos son menos consistentes y tienen exploraciones transversales en torno a su aplicación metodológica. (Moreno et al., 2011).

Se propone entonces, un esquema de caracterización metodológica en función de las especificidades del algoritmo, lo anterior con el objeto de establecer cuáles son las ventajas y des-

[12] La resolución de un problema de combinatoria espacial se puede resolver por medio de "fuerza bruta" computacionalmente hablando, pero es inabordable a escala temporal.

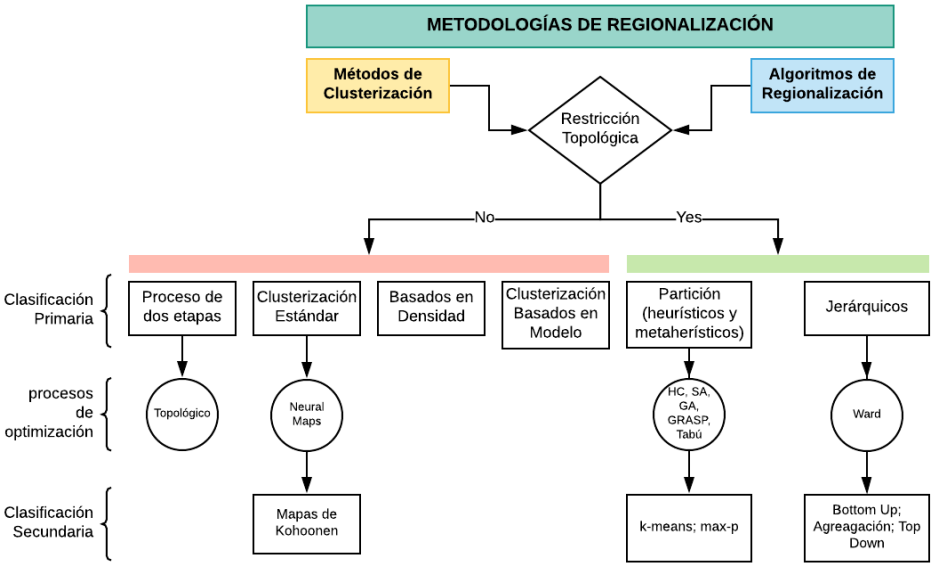
ventajas de los métodos y respaldar la elección del método de regionalización por medio de su vinculación con la teoría.

Para esto se utilizó como referencias las clasificaciones realizadas por Moreno et al. (2011), Duque et al. (2007) y Sáenz (2016), pero sobre todo la planteada por Sánchez (2015) en su tesis de doctorado.

Como se observa en la Figura N°3, la primera línea de metodologías derivan de los análisis de clusterización tradicionales, por lo tanto en un primer acercamiento no poseen una restricción topológica[13] formal en el proceso de regionalización (Sánchez, 2015). Dentro de estos procesos, además se destacan algunas aproximaciones novedosas, como la utilización de contextos multiescalares, clusterización basadas en modelos o utilización de centroides (Horn, 1995; Spielman & Logan, 2013), que permiten corregir problemas de fragmentación, no obstante estos no garantizan, necesariamente, la consistencia geográfica de la zonificación obtenida (Sáenz Vela, 2016; Sánchez, 2015).

En contrapartida, los algoritmos de regionalización formales consideran como parte de su concepción la restricción espacial o topológica la que debe primar, incluso sobre la función de optimización o el criterio de agregación que se defina (Openshaw, 1977). Dentro de éstos podemos encontrar dos clasificaciones predominantes, los de jerarquización y los de partición.

Figura N°3.
Clasificación de Metodologías de Regionalización[14]



Fuente: Elaboración propia en base a Duque, Ramos, & Suriñach, 2007; Moreno *et al*., 2011; Sáenz Vela, 2016; Sánchez, 2015

[13] Se refiere a la necesidad de que los resultados de la regionalización deriven en zonas contiguas, que respeten los criterios de vecindad de Von Neumann (vecindad de Torre) o Moore (vecindad de Reina)

[14] HC = Hill climbing; SA= Simulated Annealing; GA= Genetic Algorithms; GRASP= greedy randomized adaptive search procedure

Los de jerarquización trabajan a partir de una cadena anidada de clústeres con contigüidad espacial en donde se computar la similitud entre dos clústeres, similar a la función de Ward, o siguiendo criterios de optimización local basándose en procesos de agregación que minimicen la heterogeneidad (Guo, 2008). Estas aproximaciones metodológicas son espacialmente útiles cuando parte del objeto de estudio es resolver cuál es la escala de análisis más adecuada (Garreton & Sánchez, 2016), ya que produce soluciones anidadas a diferentes niveles, las que son posibles de analizar contra el efecto aleatorio de agregación (Sánchez, 2015). También han demostrado su eficacia para evitar sesgos espaciales de agregación estadística en el cálculo de indicadores de segregación urbana (Garreton et al., 2020). No obstante, estos algoritmos presentan algunos problemas de consistencia cartográfica al momento de generar los clústeres.

Por su parte, los algoritmos de regionalización por partición tienen diversos enfoques; la mayoría de ellos con una segunda etapa de optimización local, generalmente heurística o metaheurística, a partir de la aplicación de Hill Climbing, Simulated Annealing, búsqueda Tabú, Greedy e incluso la aplicación de algoritmos genéticos (Moreno *et al*., 2011; Sánchez, 2015). Estos procesos de optimización local además son los responsables en buena parte del mejoramiento y evolución de la efectividad computacional del método, dando ventajas sobre otras aproximaciones metodológicas.

La segunda clasificación gira en torno a las restricciones del método propiamente tal, principalmente las derivadas de los métodos de "organización exacta" (Openshaw, 1977; Sáenz Vela, 2016) que son aquellos que requieren de un número de particiones a partir de la cual se define la solución del mismo (Duque, Anselin, & Rey, 2012; Wei, Rey, & Knaap, 2020). Esto obviamente contraviene el sentido de definición de escala óptima de la regionalización, razón por la cual en ocasiones se decanta por la utilización de metodologías jerárquicas (Garreton & Sánchez, 2016). No obstante, Duque et al. (2012) plantean alternativas basadas en pisos máximos de población - límite de habitantes - que permiten una aproximación acorde a la restricción de escala a la que se quiera optar, la que no necesariamente se traduce en la óptima.

En tal sentido una aproximación funcional basada en la población aparece como un aporte importante al problema de investigación vinculado con la optimización de un marco muestral, ya que le da un cariz funcional a la aplicación empírica del mismo y vinculado con las necesidades desde instituciones gestoras de información censal y ejecutoras de encuestas, como por ejemplo el Instituto Nacional de Estadística de Chile (INE).

Metodología: Muestro espacializado para el área Metropolitana de Santiago con verificación censal

Insumos y Métodos

En la problematización del presente documento se estableció la importancia de generar evidencia empírica de la efectividad de la estratificación espacial a través de un método de regionalización, situación que ha sido probada en otras investigaciones (Duque et al., 2012; Folch & Spielman, 2014), pero no con información efectiva y de carácter exhaustivo. De esta forma se propone testear 3 métodos de regionalización y establecer una comparación con un muestro

aleatorio simple y un muestreo aleatorio en dos etapas. Para todos los casos se utilizarán los siguientes insumos:

- Como variable para estratificar se usará un indicador sintético de estratificación sociodemográfico, el Índice Socio-Material Territorial (ISMT), que resume el nivel socio material en base a 4 variables censales, estas son: escolaridad del jefe de hogar (79% varianza), el nivel de hacinamiento del hogar (7% de varianza), allegamiento del hogar (3% de varianza) y la materialidad de la vivienda (11% varianza)[15]. Cabe destacar que se usa el ISMT por ser el único indicador socioeconómico actualizado en base al Censo de 2017, dada la problemática de representatividad territorial de la CAASEN 2017.

- Se ocupará como base las zonas censales para 2017, como unidades mínimas geográficas censales.

- Para todos los efectos se usa el Área Urbana Consolidada de Santiago[16], dada la continuidad topológica necesaria para general la matriz de pesos espaciales.

- Para testear la validez de los resultados se ocupa el Censo 2017 a nivel de hogares, con un universo efectivo de alrededor de 1.8MM de casos.

Los métodos de regionalización ocupados para el muestreo espacializado son tres:

- Max-P: algoritmo de partición propuesto por Duque en (Duque *et al*., 2012). Trabaja con una matriz de pesos espaciales, y puede segmentar por un piso de población, lo que da la posibilidad de dar un peso regular a las zonas. Posee métodos heurísticos (greedy y búsqueda tabú) para optimizar los procesos de combinatoria y evitar caer en óptimos locales.

- Grouping Analysis (GA): algoritmo de clusterización multivariables, que puede utilizarse con o sin constricción espacial, ocupa el método k-Means y tiene una componente heurística, aunque no de optimización local (ESRI, 2018).

- REDCAP: algoritmo de partición jerárquica, con seis aproximaciones de aglomeración de clústeres. Posee una componente heurística, para optimización local (Guo, 2008).

La efectividad de estos tres métodos es a su vez medida bajo distintas parcelaciones territoriales medidas en k = número de zonas (cuadro N°1).

[15] Para más detalles ver (Observatorio de Ciudades PUC, 2018).

[16] Corresponde a las áreas urbanas consolidadas generadas en la mesa de trabajo conjunta entre el MINVU e INE con fecha de publicación al 2019.

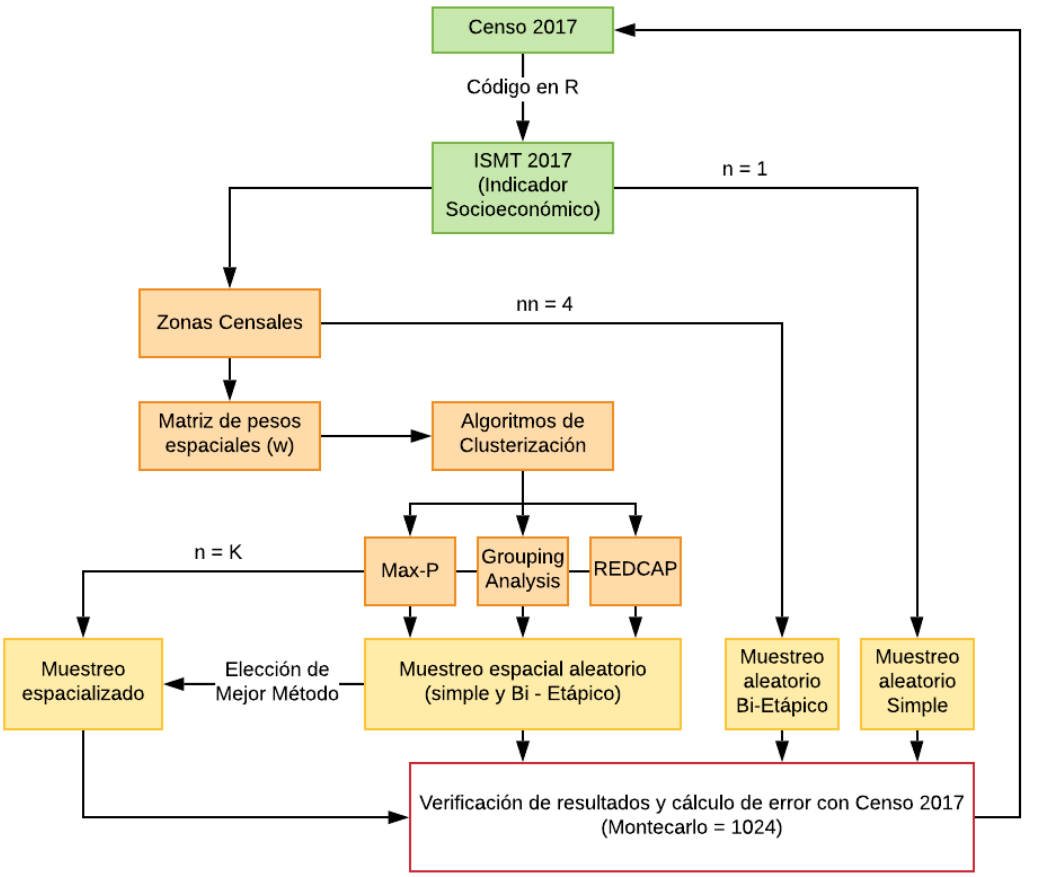
Cuadro N°1.

Especificaciones metodológicas por tipo de muestro

Nombre	Tipo de muestro	n (número de observaciones)	k (número de zonas)	Progresión del muestreo
Aleatorio	se escoge aleatoriamente un hogar sobre base de 1,8MM	350	no aplica	10; 25; 50; 100; 150; 200; 250; 300; 350
Zona Censal 2E-1	aleatorio en dos etapas; se escoge aleatoriamente 350 zonas y luego un hogar por zona	350	1.665	
Zona Censal 2E-4	aleatorio en dos etapas; se escoge aleatoriamente 87 zonas y luego cuatro hogares por zona	87*4	1.665	
Max-P 350	aleatorio espacializado. Se parcela el territorio en 350 zonas y se escoge un hogar por zona	87*4	350	
Max-P 87	aleatorio espacializado. Se parcela el territorio en 87 zonas y se escogen cuatro hogares por zona	87*4	87	
Ga-350	aleatorio espacializado. Se parcela el territorio en 350 zonas y se escoge un hogar por zona	350	350	
Ga-87	aleatorio espacializado. Se parcela el territorio en 87 zonas y se escogen cuatro hogares por zona	87*4	87	
Rc-350	aleatorio espacializado. Se parcela el territorio en 350 zonas y se escoge un hogar por zona	350	350	
Rc-87	aleatorio espacializado. Se parcela el territorio en 87 zonas y se escogen cuatro hogares por zona	87*4	87	
E-Max-P 350	muestro espacializado. Se parcela el territorio en n= k y se escoge un hogar por persona.	350	n = k (350)	

Fuente: Elaboración propia

Figura N°4.
Modelación de métodos de muestreo



Fuente: Elaboración propia.

Secuencia Metodológica

Como se observa en la Figura N°4 la aplicación de la metodología se resume en una serie de pasos, secuenciales, que luego se iteran en los diferentes algoritmos de regionalización (Cuadro N°1). Estos procesos fueron escogidos de acuerdo con su pertinencia, en función de la discusión teórica planteada con anterioridad.

De esta manera tomando como variable el ISMT a nivel de zona censal, se construye una matriz de pesos espaciales basada en contigüidad (*w*), específicamente probando la vecindad de Von Neumann y de Moore (ambas pruebas construidas en el software GeoDa).

Posteriormente se aplica el proceso de clusterización que considera el uso de GeoDa para Max-P y REDCAP y ArcGIS para GA. En el caso de MAX-P no sólo se optimiza por la función objetivo (disminuir la varianza de la variable ISMT), sino que además se define un piso de población plausible tal que sea posible construir las zonas (desde 10 a 350 zonas). Con las zonas construidas, se procedió a realizar los procesos de muestreo.

a) Muestreos aleatorios tradicionales sin espacio:

Se realizan tres tipos de muestreos: (1) aleatorio simple, en donde los casos son directamente obtenidos de la base Censal (2) muestro aleatorio en dos etapas, en el cual se escoge aleatoriamente una zona censal y posteriormente se determina la extracción de 1 o 4 casos[17] por zona.

b) Muestreos espacializados:

Para los muestreos derivados de los procesos de regionalización se siguen dos estrategias diferentes: la primera es simular un proceso aleatorio en dos etapas con n = 1 y n = 4. Esto se hace con el objeto de comparar los procesos de clusterización con las zonas censales, unidades mínimas del Censo 2017. De esta manera tendremos seis (6) pruebas de métodos espaciales, 2 por cada método de regionalización (ver Cuadro 1).

El séptimo (7) tipo de muestreo corresponde a la metodología espacializada como tal, ocupando una estrategia multiescalar, en donde el número de zona se igual al número de casos, es decir un k flexible (n=k)[18]. De esta manera se construye para cada k una zonificación independiente, sin anidación, maximizando la función objetivo para cada caso.

Para todos los experimentos antes realizados se estableció una selección bi-etápica, en donde se escogió una zona censal o una zona homogénea (proceso de regionalización) y sobre sobre esto se sampleo un caso aleatorio de un 1 hogar por zona elegida (k=350) o bien 4 (k=87). Para dar cuenta de un proceso en donde el espacio tenga mayor preponderancia en el muestro, se realizó una zonificación homogénea multiescalar (Garreton & Sánchez, 2016), no para buscar la escala más efectiva, sino que para priorizar la configuración de zona para cada n, más ajustada a las condiciones de la distribución de la variable de estratificación.

De la misma forma. para todas las metodologías propuestas se establece el nivel de error como diferencia porcentual contra la distribución censal total de la variable socioeconómica (ISMT). Este error es sometido a un procedimiento Montecarlo, es decir se testean 1024 iteraciones[19] en donde se extraen, aleatoriamente, casos del Censo 2017 para cada zona; luego se revisa el promedio del error. Con esto se asegura que los resultados nos son producto del azar y son estadísticamente significativos.

[17] Se extraen cuatro casos para alcanzar un n de 348 observaciones aproximadamente, n usada comúnmente en los nuestros representativos aleatorios con máxima varianza, con un 95% de confianza y población finita.

[18] Se define como n al número de casos obtenido del sampleo sobre el universo; mientras que k es el número de zonas obtenidas de los procesos de regionalización. Se equiparán según la progresión del muestro (Cuadro N°1).

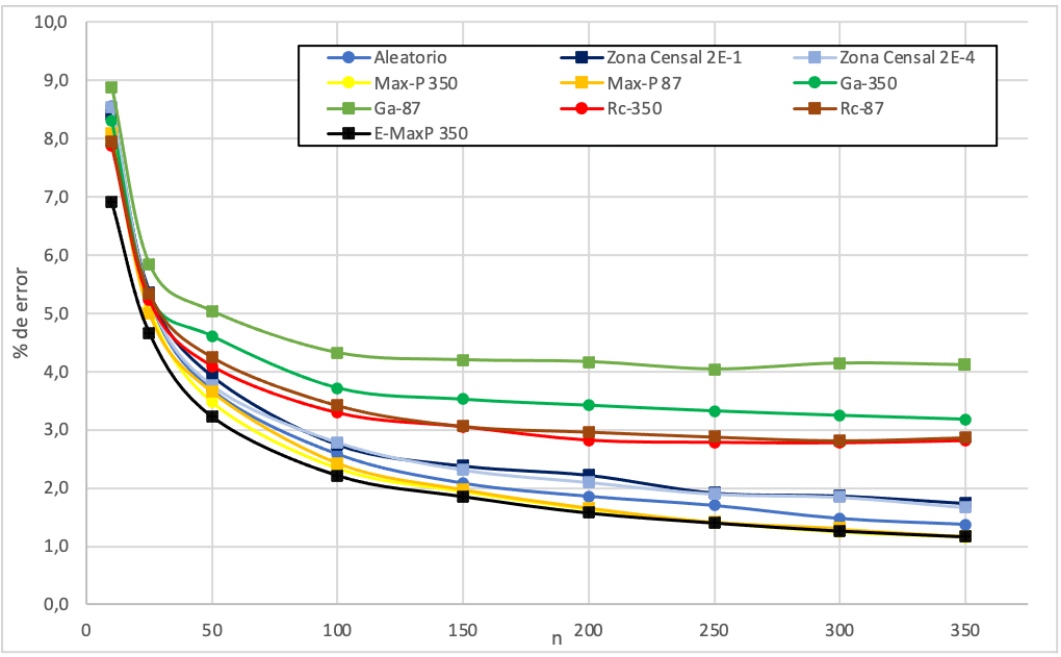
[19] El n fue obtenido empíricamente mediante un análisis de la estabilización de los resultados con la desviación estándar total del Censo, tomando como referencia en que iteración se estabiliza el error muestral de un conjunto de datos teórico (simulado en R).

Resultados y Discusión: Muestreo aleatorio simple con y sin espacio

El cálculo del error muestral (*Em*) para una distribución normal con población finita, con un 95% de confianza, se establece por medio de la formula $Em = \frac{\alpha}{2} * n * (\frac{\sigma}{\sqrt{n}})$; donde $\frac{\alpha}{2}$ corresponde a 1,96 (95% de confianza) y σ a la desviación estándar. Esto aplicado a todos los modelos de forma teórica, arroja errores similares, ya que el único factor efectivo en variar el resultado es la desviación estándar de la muestra. Sin embargo, al aplicar el sampleo efectivo, los resultados son muy diferentes, dando cuenta del rendimiento de cada uno de los métodos de regionalización y su comparación correspondiente con las metodologías tradicionales.

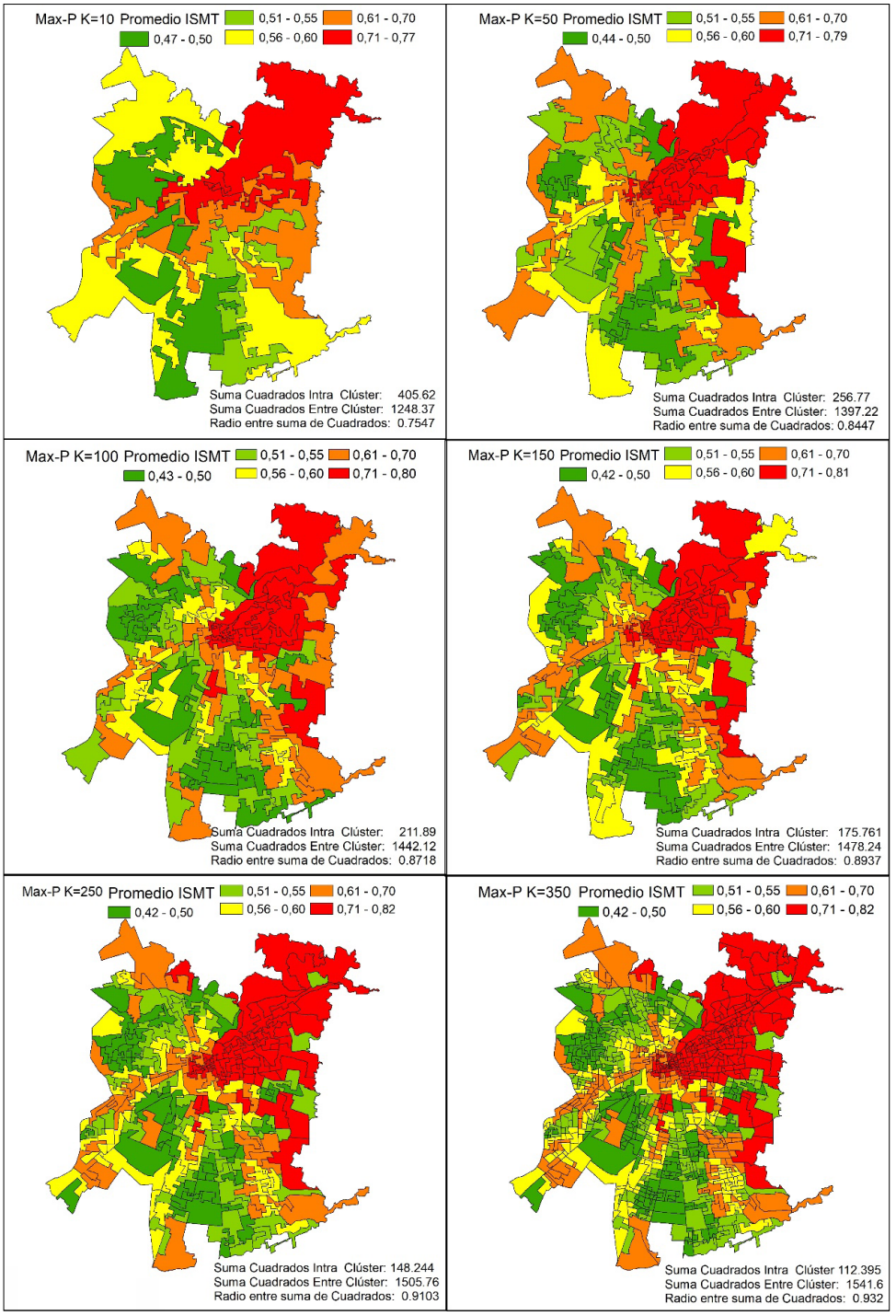
Como se observa en la Figura N°5 y el Cuadro N°2, los modelos REDCAP y el análisis de grupos no son competitivos contra las técnicas de muestreo tradicionales (aleatorio y en dos etapas), estabilizando su error después de los 100 o 150 casos entre 3 a 5%. No obstante, el método Max-P tanto en su versión simple como bi-etápica mejora en todos los N el rendimiento, sobre todo el simple (Max-P 350) el que, para la zonificación con 250 casos, tiene un comportamiento equivalente al n=350, es decir es posible ahorrar 100 personas de la muestra total, con el mismo nivel de error.

Figura N°5.
Gráfico con porcentaje de error de muestreos aleatorios con y sin espacio



Fuente: Elaboración propia.

Figura N°6.
Cartografía muestreo espacializado con k flexible



Fuente: Elaboración propia.

Cuadro N°2.
Porcentajes de error de muestreos aleatorios con y sin espacio

	Aleatorio	Zona Censal 2E- 1	Zona Censal 2E - 4	Max-P 350	Max-P 87	Ga-350	Ga-87	Rc-350	Rc-87	E-Max-P 350
10	8,56	8,42	8,54	7,96	8,09	8,30	8,88	7,87	7,94	**6,90**
25	5,26	5,36	5,28	5,08	4,99	5,30	5,83	5,22	5,34	**4,65**
50	3,68	3,91	3,76	3,47	3,65	4,61	5,03	4,09	4,25	**3,22**
100	2,58	2,74	2,78	2,33	2,42	3,72	4,31	3,29	3,42	**2,21**
150	2,08	2,37	2,31	1,93	1,97	3,52	4,19	3,05	3,06	**1,85**
200	1,85	2,21	2,09	1,63	1,66	3,42	4,16	2,82	2,96	**1,57**
250	1,70	1,91	1,89	1,40	1,41	3,32	4,03	2,78	2,88	**1,40**
300	1,47	1,86	1,84	**1,24**	1,29	3,25	4,14	2,77	2,81	1,26
350	1,37	1,73	1,66	**1,16**	1,17	3,18	4,11	2,81	2,86	1,16
$\bar{x}$	**3,17**	**3,39**	**3,35**	**2,91**	**2,96**	**4,29**	**4,97**	**3,85**	**3,95**	**2,69**
σ	**2,37**	**2,22**	**2,26**	**2,28**	**2,30**	**1,66**	**1,58**	**1,71**	**1,71**	**1,93**

Fuente: Elaboración propia.

Esto aparece significativo, ya que permite reducir el error, en el tramo entre 250 y 350 observaciones, entre un 15 y 17%, y en promedio para todos los n, en un 8%. La explicación se debe, en parte, a la optimización heurística, clave sobre todo en las particiones con mayor número de zonas, por la evidente mayor probabilidad de combinatorias presentes para el algoritmo.

Los resultados dan cuenta que los procesos de regionalización basados en la minimización de la varianza efectivamente funcionan para una variable sociodemográfica como el ISMT (y cualquier otra variable con esta tipología). Esto se justifica en función de la distribución en el Área Metropolitana de Santiago altamente clusterizada y segregada. Lo anterior es efectivo además considerando, la aplicación de métodos efectivos de optimización local.

Esto es altamente significativo considerando que se aplica a partir de zonas censales que ya tienen de base una heterogeneidad relevante a diferencia de las manzanas que hubiesen sido mucho mejor prospecto como unidad espacial (no disponibles debido all proceso de indeterminación censal)

Muestreo espacializado y consideraciones para su aplicación

En la Figura N°5 y el Cuadro N°2, se puede evidenciar que el muestro espacializado (E-Max-p 350) tiene mejores rendimientos contra todos los métodos de muestro tradicionales con un promedio de 14% de mejora sobre el método aleatorio y un 20% sobre el bi-etápico basado en zonas censales.

En base a este análisis empírico, se da cuenta que la estrategia multiescalar es mucho más flexible y coherente que la utilización de una zonificación fija. Esto se explica porque el algoritmo busca la obtención del óptimo global para cada una de las escalas, sobre todo en presencia de procesos heurísticos, lo que permite mejorar los resultados contra una zonificación fija.

El nivel de error del muestreo espacializado (E-Max-p 350) con 250 casos es similar al de 350 casos del muestreo tradicional aleatorio, esto supone una mejora relevante y un ahorro significativo en el n de la encuesta.

A modo de consideraciones prácticas, se discuten algunos puntos necesarios para consolidar las técnicas de muestreo espacializado de forma de asegurar una aplicación efectiva en el contexto del diseño de un marco muestral, estas son:

- Desde el punto de vista de la cobertura es ideal considerar una zonificación con continuidad topológica. Las vecindades de Von Neumann o Moore funcionan mejor que otras aproximaciones como triangulación de Delaunay o generación de matriz de pesos espaciales con distancia.

- Es necesario trabajar con zonas continuas, como la definición del área urbana consolidada de una ciudad, evitando islas. Esto puede ser complejo en ciudades con configuraciones más complejas, como el área metropolitana de Concepción, por lo que es necesario mejorar estrategias mixtas en la creación de matrices de pesos espaciales. (ver Figura N°6).

- El rendimiento de los resultados es altamente dependiente de la presencia de procesos de optimización local, específicamente, técnicas heurísticas, que además sean relativamente simples para trabajar con un set de datos de gran tamaño[20].

- Se destaca la flexibilidad de los métodos que permiten tener un piso mínimo de población y restricción topológica como Max-P. Si bien esto puede ser conflictivo para la función de optimización central (minimizar la varianza de la variable de estratificación), es altamente deseable para las estrategias territoriales de muestreo, contribuyendo la organización efectiva y división del trabajo en terreno para los encuestadores.

- Se debe profundizar en temas como la consistencia topológica de los resultados, ya que independiente de la calidad estadísticas de los mismos, es necesario modular con la efectividad para que las zonas sirvan como unidades mínimas estadísticas para su aplicabilidad en el contexto de marcos muestrales (ver Figura N°6).

Conclusiones: estratificación espacial y disminución del error muestral

Los resultados presentados, en términos teóricos dan cuenta que el proceso de regionalización, efectivamente es capaz de disminuir datos redundantes, permitiendo mantener un nivel de representación efectivo con menos observaciones (Griffith, 2005).

[20] Las zonas censales para el AMS tienen 1.665 filas y las metodologías funcionan bien con hasta 7.000 casos, por lo que trabajar con manzanas para Santiago es inviable (50.000 manzanas).

De esta forma, la efectividad de la estratificación queda supeditada a un diseño espacialmente continuo y a la correspondencia entre el n de la zonificación y el tamaño muestral. Esto queda en evidencia al analizar los resultados de estrategias bi-etápicas con más de un hogar extraído por zona, las cuales aparecen menos precisas debido a la redundancia de observaciones.

En términos cuantitativos la estratificación espacial con mejor rendimiento (E-max-p350) permite disminuir el error muestral de manera significativa obteniendo mejoras entre 15 y 20%, o en su defecto, disminuyendo el número de observaciones de forma relevante (350 a 250 con un 1,4% de error muestral).

Se destaca además que, para todos los efectos, E-max-p350 obtiene mejores resultados que las zonas censales, lo que aparece como relevante dado que es la unidad base desagregada de origen de la clusterización; es decir a pesar de la agregación se logra disminuir varianza obteniendo zonas más representativas de la realidad socioeconómica del Área Metropolitana de Santiago.

Tomando en consideración los porcentajes mencionados, el impacto global en encuestas de gran tamaño, como la CASEN 2017, podría traducirse en disminuciones de 10.000[21] observaciones. Esto requeriría un importante trabajo previo de sistematización geográfica, incluyendo la generación de continuidad topológica en ciudades y una diferenciación robusta entre el ámbito rural y urbano. En suma, es un trabajo analítico previo complejo que compensaría con creces la reducción de costo o mejora de calidad estadística de encuestas socioeconómicas.

Respecto a los resultados específicos se puede concluir que es necesario profundizar la forma de aplicación para corregir la compacidad cartográfica, ya que su aplicación baja la calidad de los procesos de regionalización, pero por sobre todo genera problemas de equilibrio de población de los pisos de población, situación que se debe mantener para la aplicación de encuestas y su respectiva consistencia en función de los muestreos.

Referencias

Arretx, C. (1989). La conciliacion censal. *CELADE*, *Santiago C*, 23. http://repositorio.cepal.org/handle/11362/32637

Borchsenius, L. (2001). From a Conventional To a Register-Based Census of Population. *Census Seminar*, 20–21. http://www.demography-lab.prd.uth.gr/european-census/Files/general-data/Insee-Eurostat/borchsenius.pdf

Bravo, D., Larrañaga, O., Millán, I., Ruiz, M., & Zamorano, F. (2013). Informe final Comisión externa revisora del CENSO 2012. *Resumenes I Congreso Iberoamericano de Gestión Integrada de Áreas Litorales.*, 23–30. http://www.censo.cl/documentos/informe_final-comision-nacional.pdf

[21] La CASEN tiene un marco muestral que varía en torno a las 69.000 observaciones (Ministerio de Desarrollo Social, 2018).

Brus, D. J., & De Gruijter, J. J. (1997). Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with Discussion). In *Geoderma* (Vol. 80, Issues 1–2, pp. 1–44). Elsevier. https://doi.org/10.1016/S0016-7061(97)00072-4

Brus, D. J., Yang, L., & Zhu, A. X. (2019). Accounting for differences in costs among sampling locations in optimal stratification. *European Journal of Soil Science*, *70*(1), 200–212. https://doi.org/10.1111/ejss.12731

Cochran W.G. (1977). *Sampling Techniques*. New York, N.Y. (USA) Wiley. http://agris.fao.org/agris-search/search.do?recordID=XF2015028634

Cohen, B. (2006). Urbanization in developing countries: Current trends, future projections, and key challenges for sustainability. *Technology in Society*, *28*(1–2), 63–80. https://doi.org/10.1016/j.techsoc.2005.10.005

Cook, L. (2004). The quality and qualities of population statistics, and the place of the census. *Area*, *36*(2), 111–123. https://doi.org/10.1111/j.0004-0894.2004.00208.x

Cressie, N. A. C. (1993). 01 Statistics for Spatial Data. In *Statistics for Spatial Data* (pp. 1–26). https://doi.org/10.1002/9781119115151

de Gruijter, J. J., & ter Braak, C. J. F. (1990). Model-free estimation from spatial samples: A reappraisal of classical sampling theory. *Mathematical Geology*, *22*(4), 407–415. https://doi.org/10.1007/BF00890327

Duque, J., Anselin, L., & Rey, S. (2012). The max-p-regions problem. *Journal of Regional Science*, *52*(3), 397–419. https://doi.org/10.1111/j.1467-9787.2011.00743.x

Duque, J., Ramos, R., & Suriñach, J. (2007). Supervised regionalization methods: A survey. *International Regional Science Review*, *30*(3), 195–220. https://doi.org/10.1177/0160017607301605

ESRI. (2018). *Análisis de agrupamiento—Ayuda | ArcGIS Desktop*. https://desktop.arcgis.com/es/arcmap/10.3/tools/spatial-statistics-toolbox/grouping-analysis.htm

Folch, D. C., & Spielman, S. E. (2014). Identifying regions based on flexible user-defined constraints. *International Journal of Geographical Information Science*, *28*(1), 164–184. https://doi.org/10.1080/13658816.2013.848986

Garreton, M., Basauri, A., & Valenzuela, L. (2020). Exploring the correlation between city size and residential segregation: comparing Chilean cities with spatially unbiased indexes. *Environment and Urbanization*, 095624782091898. https://doi.org/10.1177/0956247820918983

Garreton, M., & Sánchez, R. (2016). Identifying an optimal analysis level in multiscalar regionalization: A study case of social distress in Greater Santiago. *Computers, Environment and Urban Systems*, *56*, 14–24. https://doi.org/10.1016/j.compenvurbsys.2015.10.007

Griffith, D. A. (2005). Effective Geographic Sample Size in the Presence of Spatial Autocorrelation. *Annals of the Association of American Geographers*, *95*(4), 740–760. https://doi.org/10.1111/j.1467-8306.2005.00484.x

Guo, D. (2008). Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science*, *22*(7), 801–823. https://doi.org/10.1080/13658810701674970

Heaton, M. J., & Gelfand, A. E. (2012). Kernel averaged predictors for spatio-temporal regression models. *Spatial Statistics*, *2*, 15–32. https://doi.org/10.1016/J.SPASTA.2012.05.001

Horn, M. E. T. (1995). Solution Techniques for Large Regional Partitioning Problems. *Geographical Analysis*, *27*(3), 230–248. https://doi.org/10.1111/j.1538-4632.1995.tb00907.x

Jin, X., Wah, B. W., Cheng, X., & Wang, Y. (2015). Significance and Challenges of Big Data Research. *Big Data Research*, *2*(2), 59–64. https://doi.org/10.1016/J.BDR.2015.01.006

Lefebvre, H. (1991). *The production of space*. Blackwell.

Legendre, P. (1993). Spatial autocorrelation: trouble or new paradigm? *Ecology*, *74*(6), 1659–1673. https://doi.org/10.2307/1939924

Lin, L., & Wang, F. (2019). Geographical proximity vs network tie: innovation of equipment manufacturing firms in Shanghai, China. *Erdkunde*, 185–198. https://doi.org/10.3112/erdkunde.2019.03.03

Lindley, D. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, *27*, 986–1005. http://www.jstor.org/stable/2237191?casa_token=CB0UZDZivncAAAAA:-6VZV_tAaxiAQbOXVGJGf3VXwodidklVhnUmtZbdjavY2Hk9LOMxJZrRCWbc6IQMV9wnBQAUz5JYV_I_GIoNQaXQti7q_tvHlXGyiAZ78MqdFGyB448

Miller, H. J. (2004). Tobler's first law and spatial analysis. *Annals of the Association of American Geographers*, *94*(2), 284–289. https://doi.org/10.1111/j.1467-8306.2004.09402005.x

Ministerio de Desarrollo Social. (2018). *Metodología de Diseño Muestral*. http://observatorio.ministeriodesarrollosocial.gob.cl/casen-multidimensional/casen/docs/Diseno_Muestral_Casen_2017_MDS.pdf

Montello, D. R. (2003). Regions in geography: Process and content. In M. Duckham, M. F. Goodchild, & M. F. Worboys (Eds.), *Foundations of geographic information science* (Taylor & F, pp. 173–189). https://doi.org/doi:10.1201/9780203009543.ch9

Moreno, P., García, J., & Lacalle, L. D. E. (2011). Estado del Arte en procesos de zonificación. *Geofocus*, *11*, 155–181. www.geo-focus.org

Observatorio de Ciudades PUC. (2018). *ISMT | Infraestructura de Datos Espaciales OCUC*. IDE OCUC. https://ideocuc-ocuc.hub.arcgis.com/datasets/97ae30fe071349e89d9d5ebd5dfa2aec_0

Openshaw, S. (1977). A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling. *Transactions of the Institute of British Geographers*, *2*(4), 459–472. https://doi.org/10.2307/622300

Openshaw, S. (1978). *An optimal zoning approach to the study of spatially aggregated data* (pp. 95–113). Springer, Boston, MA. https://doi.org/10.1007/978-1-4613-4067-6_5

Openshaw, S., & Baxter, R. S. (1977). Algorithm 3; a procedure to generate pseudo random aggregations of N zones into M zones where M is less than N. *Environment and Planning A*, *9*(12), 1423–1428. https://doi.org/10.1068/a091423

Pettitt, A. N., & McBratney, A. B. (1993). Sampling Designs for Estimating Spatial Variance Components. *Applied Statistics*, *42*(1), 185. https://doi.org/10.2307/2347420

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*(3–4), 231–259. https://doi.org/10.1016/J.ECOLMODEL.2005.03.026

Rodríguez-iglesias, G., & Teresa, M. (2010). La importancia de la especificidad territorial en la construccion de indicadores locales. *Ciencia Ergo Sum*, *18*(m), 145–152. http://www.redalyc.org/articulo.oa?id=10418753005

Sáenz Vela, H. M. (2016). Revisando los métodos de agregación de unidades espaciales: MAUP, algoritmos y un breve ejemplo / Reviewing spatial unit aggregation methods: MAUP, algorithms and a brief example. In *Estudios Demográficos y Urbanos* (Vol. 31, Issue 2). El Colegio de México. https://doi.org/10.24201/edu.v31i2.1592

Sánchez, R. (2015). *Spatial self-organization in Santiago. Methods and Applications*. Universidad Adolfo Ibáñez.

Shewry, M. C., & Wynn, H. P. (1987). Maximum entropy sampling. *Journal of Applied Statistics*, *14*(2), 165–170. https://doi.org/10.1080/02664768700000020

Spielman, S. E., & Logan, J. R. (2013). Using High-Resolution Population Data to Identify Neighborhoods and Establish Their Boundaries. *Annals of the Association of American Geographers*, *103*(1), 67–84. https://doi.org/10.1080/00045608.2012.685049

Stein, A., & Ettema, C. (2003). An overview of spatial sampling procedures and experimental design of spatial studies for ecosystem comparisons. *Agriculture, Ecosystems & Environment*, *94*(1), 31–47. https://doi.org/10.1016/S0167-8809(02)00013-0

Stock, B. C., Ward, E. J., Thorson, J. T., Jannot, J. E., & Semmens, B. X. (2019). The utility of spatial model-based estimators of unobserved bycatch. *ICES Journal of Marine Science*, *76*(1), 255–267. https://doi.org/10.1093/icesjms/fsy153

Tobler, W. R. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, *72*(5), 234. https://doi.org/10.1037/h0028106

Vallejos, R., & Osorio, F. (2014). Effective sample size of spatial process models. *Spatial Statistics*, *9*(C), 66–92. https://doi.org/10.1016/j.spasta.2014.03.003

Wallgren, A., & Wallgren, B. (2007). Register-based Statistics: Administrative Data for Statistical Purposes. In *Register-based Statistics: Administrative Data for Statistical Purposes*. John Wiley & Sons. https://doi.org/10.1002/9780470061350

Wallgren, A., & Wallgren, B. (2016). Frames and Populations in a Register-based National Statistical system. *Journal of Mathematics and Statistical Science*, *2016*, 208–216. http://www.ss-pub.org/wp-content/uploads/2016/04/JMSS15121601.pdf

Wang, J.-F., Jiang, C.-S., Hu, M.-G., Cao, Z.-D., Guo, Y.-S., Li, L.-F., Liu, T.-J., & Meng, B. (2013). Design-based spatial sampling: Theory and implementation. *Environmental Modelling & Software*, *40*, 280–288. https://doi.org/10.1016/j.envsoft.2012.09.015

A review of spatial sampling, 2 Spatial Statistics 1 (2012). https://doi.org/10.1016/j.spasta.2012.08.001

Wang, J., Haining, R., & Cao, Z. (2010). Sample surveying to estimate the mean of a heterogeneous surface: Reducing the error variance through zoning. *International Journal of Geographical Information Science*, *24*(4), 523–543. https://doi.org/10.1080/13658810902873512

Wei, R., Rey, S., & Knaap, E. (2020). Efficient regionalization for spatially explicit neighborhood delineation. *International Journal of Geographical Information Science*. https://doi.org/10.1080/13658816.2020.1759806

Williamson, I., Rajabifard, A., & Binns, A. (2006). Challenges and Issues for SDI Development. *International Journal of Spatial Data Infrastructures Research*, *1*(1), 24–35. https://doi.org/10.2902/

Yates, F. (1946). A Review of Recent Statistical Developments in Sampling and Sampling Surveys. *Journal of the Royal Statistical Society*, *109*(1), 12–43. https://doi.org/10.2307/2981390