uc3m | Universidad Carlos III de Madrid

TESIS DOCTORAL

Marco de Referencia para el Análisis Comparativo de Métodos de Generación de Poblaciones Sintéticas

Autor

Guillermo Castilla Alcalá

Directores

Alfonso Durán Heras

Isabel García Gutiérrez

Departamento Ingeniería Mecánica

Leganés, Julio 2018

Agradecimientos

Antes de todo, quiero agradecer a mi esposa por su cariño y estar a mi lado. También a mis hijos por el apoyo moral que me han prestado.

No puede faltar mi más amplio agradecimiento a mis directores de tesis, Dr. Alfonso Duran Heras y Dra. Isabel García Gutiérrez, por permitirme adentrarme en el mundo de la investigación, apoyarme incondicionalmente en todo el proceso de realización de esta tesis y guiarme por el mundo de las publicaciones científicas. He aprendido mucho de ellos, y gracias a sus numerosas y valiosas ideas he podido llevar a cabo este trabajo de investigación.

De la misma manera, quiero dejar constancia de mi gratitud a Javier Rodríguez, Juan Antonio Baratas y Raúl Hernández por su desinteresado apoyo y colaboración en la confección de algunos de los programas Java y R, necesarios para el desarrollo de esta tesis.

Por último, y no por ello menos importante, agradecer a mis amigos y compañeros del área de Organización del Departamento de Ingeniería Mecánica de la Universidad Carlos III de Madrid por su apoyo y palabras de ánimo.

Gracias a todos.

Resumen

La generación de poblaciones y datos sintéticos es un proceso básico en múltiples áreas tales como la estimación de *small area,* microsimulación, simulación basada en agentes, simulación basada en actividad, anonimización de datos o validación y pruebas de aplicaciones. La principal finalidad es obtener una población de agentes, o microdatos sintéticos, con distribuciones de atributos correlacionados que se asemeje a una población o datos reales. Se han desarrollado múltiples métodos orientados a tal fin, sin que existan claras recomendaciones para la elección entre ellos.

Dentro del ámbito de los métodos de generación de poblaciones, no existe un marco de referencia de general aceptación para realizar de forma efectiva la evaluación comparativa de la eficacia y precisión de estos métodos, por lo que el objetivo general de esta investigación es definir dicho marco de referencia en el que posicionar los distintos métodos de generación de poblaciones sintéticas respecto a los escenarios de uso (necesidades y datos disponibles), y establecer una clara metodología para llevar a cabo un análisis cuantitativo del rendimiento de estos métodos.

Una vez establecido el marco de referencia se plantean tres estudios cuantitativos experimentales conforme al marco propuesto, con los que se obtienen conclusiones sobre el rendimiento relativo de los métodos. En los dos primeros se plantean escenarios de generación de poblaciones en los que se diseñan los experimentos con los que se efectúan los correspondientes análisis comparativos concretos. A partir de necesidades y datos de partida concretos, se determinan los métodos más adecuados para obtener los microdatos sintéticos. Se desarrolla un esquema de validación comparativa entre métodos que permite comparar las poblaciones sintetizadas con la población real, utilizando una comparativa estadística con contraste de hipótesis.

El tercer estudio se apoya en la metodología del análisis cuantitativo para comparar distintas estrategias que se utilizan para abordar el problema de muestras con algún valor marginal nulo. Este problema se presenta cuando se extraen muestras pequeñas de poblaciones con categorías poco frecuentes, por lo que es habitual que en las muestras obtenidas falten representantes de dichas categorías.

Para los experimentos se utiliza una población real de 60 "*small area*s" y múltiples muestras de distinto tamaño. Entre los métodos de generación de poblaciones que se comparan se encuentran métodos de amplia difusión como: *Iterative Proportional Fitting* con redondeo y con muestreo de Monte Carlo, *Simulated Annealing, Generalized Raking* e *Iterative Proportional Updating*.

En esta tesis se describen y clasifican los principales métodos de generación de poblaciones. Se construye un mapa con los métodos más adecuados según distintos escenarios de generación. También se analizan diversos sistemas de medida de similitud de poblaciones con objeto de facilitar la selección de métricas de la calidad de la población generada. Con respecto a la utilización de diferentes métricas, se constata que clasifican y ordenan los métodos comparados de forma equivalente, por lo que la recomendación de uso se basa en su simplicidad y su fácil comprensión.

El marco de referencia permite determinar si hay diferencias estadísticamente significativas entre los rendimientos de los métodos; por tanto, permite posicionar los métodos, y analizar la influencia de distintos factores en el rendimiento de los mismos.

A raíz de los resultados obtenidos con los experimentos planteados pueden extraerse distintas conclusiones e implicaciones, tales como la importancia del redondeo o del tamaño de la muestra y la superioridad de ciertos métodos sobre otros.

Entre los factores utilizados en la generación de poblaciones se han considerado el número de atributos y categorías que describen la población, el tamaño de la muestra y el tamaño de la "*small area*".

El conjunto de esta tesis aporta claridad en el campo de los métodos y técnicas de generación de poblaciones sintéticas, facilitando el análisis y la toma de decisiones ante la necesidad de disponer de una población de agentes, y proporcionando una guía para llevar a cabo un análisis comparativo entre métodos.

Esta tesis contribuye a ayudar al investigador a determinar, en cada situación, los métodos que generan las poblaciones que representan al mundo real con mayor precisión, permitiéndole mejorar la calidad de los resultados de las simulaciones y obtener resultados más verídicos. También permite probar algoritmos en condiciones más próximas a la realidad utilizando los datos generados por estos métodos, con la posibilidad de difundir y compartir dichos datos y los resultados de los algoritmos sin problemas de confidencialidad, como datos abiertos.

Palabras clave: generación de poblaciones; población sintética; métodos de generación de poblaciones; métricas de comparación de poblaciones; área pequeña.

Abstract

Synthetic population generation is a basic process in multiple areas such as microsimulation, agent-based simulation, activity-based simulation, validation and application testing or data anonymization. The main purpose is to obtain a population of agents, or synthetic microdata, with distributions of correlated attributes that resemble a population or real data. Multiple methods have been developed for this purpose, with no recommendations regarding their comparative use.

In the academic world of population generation methods, there is no generally accepted reference framework to effectively perform the comparative assessment of the effectiveness and precision of these methods, so this study has the objective of defining such reference framework in which to position the different methods of population generation methods with respect to the scenarios (needs and available data), and establish a clear methodology to carry out a quantitative analysis of the performance of these methods.

Three experimental quantitative research studies are accomplished according to the proposed framework and conclusions are obtained on the relative performance of the methods. Two scenarios of population generation are suggested in the two first researches where experiments are designed and comparative analyzes are carried out according to the proposed reference framework.

Based on specific data and input data, the most appropriate methods are determined for obtaining the best synthetic microdata. A comparative validation scheme between methods is developed to compare the synthesized populations with the real population, based on a statistical comparison with hypothesis testing.

The quantitative analysis methodology is also used to compare the strategies used to address the zero-marginal problem of the sample in the third research. This problem occurs when there are rare categories in the population, from which a small sample is extracted in which the representation of these rare categories are missing.

For the research experiments, a true population of 60 "*small area*s" and multiple samples of different sizes are used. The population generation methods that are compared are widely diffused methods such as *Iterative Proportional Fitting* with rounding and with *Monte Carlo sampling, Simulated Annealing*, *Generalized Raking* and *Iterative Proportional Update*.

The main methods of population generation are described and classified. A map is constructed with the most appropriate methods according to the generation scenario. Metrics for evaluating the goodness of fit between estimated and observed sets of population microdata are also analyzed in order to facilitate the selection of measurements for the

accuracy of the generated population. Metrics are found that it classifies and rank the compared methods in an equivalent way, so a goodness-of-fit evaluation metric for synthetic data is recommended to use based on its simplicity and its easy interpretation of results.

The reference framework allows to determine if there are statistically significant differences between methods performance; therefore, it allows methods positioning, and analyze the influence of multiple factors on the performance.

Following the results obtained with the experiments, final conclusions and implications are extracted, such as the importance of rounding or sample size, and the superiority of some methods.

The number of categories and attributes that describe the population, the sample size, and the "*small area*" size have been considered among the factors that were used in the population generation process.

This thesis as a whole provides clarity in the field of methods and techniques for synthetic population generation, facilitates the analysis and decision making in the posed scenarios when a population of agents is needed and provides a guide to performing a comparative analysis between methods.

This thesis helps the researcher to establish the methods that generate populations that best depicts the real world, and it allows to improve the quality of the simulations results and to obtain the most truthful trending values. It also allows them to test algorithms in conditions more similar to reality using the data generated by these methods, with the possibility of spreading and sharing these data and algorithms results without confidentiality problems, as open data.

Keywords: population generation; synthetic population; population generation methods; population comparison metrics; small area.

Índice

Índice de Tablas

Índice de Figuras

1 Introducción

1.1 Motivación

Muchos estudios de ingeniería y ciencias sociales se apoyan en técnicas de simulación que requieren el uso de poblaciones de agentes que representan las entidades objeto de investigación. En la era de los datos, no siempre se dispone fácilmente de los datos que describen estas poblaciones, bien porque su coste sea excesivo, por la falta de los mismos o por la confidencialidad que requieren.

Igualmente, en otras ocasiones se precisa disponer de datos de ensayo para pruebas que se asemejen a entornos reales, con objeto de evaluar y testear el funcionamiento de determinados sistemas de información.

A falta de estos datos, se han desarrollado una serie de técnicas y algoritmos para obtenerlos de forma artificial, a los que se denomina métodos de generación de datos o poblaciones sintéticas, de uso frecuente en los entornos de simulación y pruebas.

En esta tesis nos referiremos a población sintética como un conjunto de registros de agentes (personas, hogares, empresas, etc.) caracterizados por una serie de atributos, la cual se asemeja a una población real determinada.

En los últimos 20 años se han desarrollado distintos métodos de generación de poblaciones sintéticas sin que exista un enfoque de general aceptación para analizar de forma práctica la efectividad de los mismos. Aunque se han llevado a cabo algunas comparaciones entre los métodos, dichas comparaciones no permiten posicionar con rigor los métodos y establecer claramente las ventajas e inconvenientes de cada uno, ni tampoco los escenarios de uso más apropiados para cada método. Este hecho es atribuible a una serie de carencias metodológicas, como pueden ser, falta de consenso en las medidas de error, deficiencias en el análisis estadístico o débil caracterización de las condiciones de utilización de los métodos.

Esta situación ha motivado el desarrollo de esta tesis, donde se establece un marco de referencia que permite posicionar los distintos métodos de generación de poblaciones sintéticas, y en el que se define una metodología para realizar un análisis comparativo del rendimiento. De acuerdo con el marco planteado se presentan varios estudios comparativos dentro del ámbito de la generación de poblaciones sintéticas.

Esta tesis se enmarca en la línea de investigación sobre plataformas flexibles de simulación, desarrolladas por los dos directores en el área de Ingeniería de Organización de la universidad Carlos III desde hace más de una década, y ha dado lugar a la publicación del artículo

Comparison of Iterative Proportional Fitting and Simulated Annealing as synthetic population generation techniques: Importance of the rounding method en la revista Computers, Environment and Urban Systems, en diciembre del año 2017.

1.2 Objetivos

Como se ha expuesto en la motivación, no existen directrices claras en relación a comparativas entre métodos. Los distintos estudios comparativos publicados son, en la mayoría de los casos, comparaciones entre dos métodos que utilizan sus propias medidas de error, distintas unas de otras, y suelen basarse en un solo caso de aplicación.

Por tanto, la finalidad de la presente investigación es aportar claridad en el campo de los métodos y algoritmos de generación de poblaciones sintéticas, mediante el desarrollo de un marco de referencia que facilite el análisis de los escenarios que se plantean ante la necesidad de disponer de una población de agentes y que permita llevar a cabo un análisis comparativo entre métodos. Para lograr tal fin se establecen los siguientes cuatro objetivos concretos:

1. Hacer una revisión y clasificación de los principales métodos de generación de poblaciones sintéticas que aparecen en la literatura, así como de las principales aplicaciones de los datos sintéticos.
2. Analizar las distintas problemáticas que se plantean en el proceso de selección de métodos de construcción de poblaciones sintéticas y justificar la propuesta de marco de referencia que ayude en dicho proceso.
3. Definir un marco de referencia para analizar comparativamente los métodos, tanto con respecto a recomendaciones de utilización de métodos ante las distintas necesidades y disponibilidades de datos, como en el ámbito de la metodología de análisis comparativo.
4. Realizar un mínimo de 3 estudios cuantitativos experimentales conforme al marco propuesto, con los que se obtengan conclusiones sobre el rendimiento relativo de los métodos y se posibiliten análisis de verificación del impacto del uso de nuevas técnicas y estrategias, análisis del comportamiento de distintos métodos de generación y determinación del rendimiento relativo de métodos.

1.3 Organización y esquema de la Tesis

En el capítulo 2 de esta tesis se introduce el concepto de datos y poblaciones sintéticas, el ámbito de los métodos de generación de datos y poblaciones sintéticas, así como las distintas áreas donde tienen utilidad.

En el capítulo 3 se clasifican y describen los distintos métodos de generación de datos sintéticos.

En los capítulos 4 y 5 se definen los componentes del Marco de Referencia. En el primero se definen los potenciales escenarios y se posicionan los métodos más adecuados para los distintos escenarios. Y en el segundo se establece la metodología para el análisis comparativo de métodos, incluyendo las métricas y pruebas estadísticas que se proponen.

Los siguientes tres capítulos corresponden a tres estudios experimentales con aplicación del marco de referencia definido. Los dos primeros, capítulos 6 y 7, presentan comparativas de métodos en distintos escenarios: una población de agentes sin jerarquía y otra con jerarquía multinivel. En ambos capítulos se analiza la importancia que tiene el redondeo en el rendimiento de los métodos.

El capítulo 8 describe un análisis comparativo de distintas estrategias que se utilizan para abordar el problema de muestras con algún valor marginal nulo, que se presenta cuando se tienen categorías poco frecuentes en la población y muestras de pequeño tamaño que no incluyen representantes con dichas categorías.

Estos últimos tres capítulos, correspondientes a los estudios experimentales, se han redactado de forma independiente, por lo que no es necesaria una lectura secuencial.

Se finaliza la tesis con un capítulo de conclusiones, limitaciones y futuros estudios.

2 Datos y Poblaciones Sintéticas

2.1 Introducción

Antes de entrar en la descripción de los métodos de construcción de datos sintéticos, y delimitar el entorno en el que se plantea el marco de referencia que se propone, se introducirá el concepto de "estimación de *small area*" que permitirá encuadrar los métodos de generación de datos sintéticos, y en las siguientes secciones del capítulo se analizarán las distintas áreas de aplicación de los datos sintéticos, entre las que se incluyen la microsimulación espacial, la simulación basada en agentes y la basada en actividad (*activity-based modelling*), los ensayos y pruebas de aplicaciones, y la anonimización para proteger la confidencialidad de los datos.

Previo a la descripción de los conceptos y áreas de aplicación mencionadas se definen algunos términos y notación que se utilizarán a lo largo de la tesis.

El término población de agentes se refiere a un conjunto de registros, uno para cada agente, los cuales contienen valores de variables que describen sus características. Asumiremos que dichas características o atributos solo pueden tomar valores discretos, que denominaremos categorías, dado que si se trata de variables de naturaleza continua, siempre se podrán discretizar mediante el uso de intervalos y tratarse como discretas. Esta forma de estructurar los datos de la población se conoce con el nombre de microdatos o información desagregada. La información de los microdatos puede presentarse en forma condensada mediante una tabla multidimensional, donde cada celda contendrá el número de agentes que tienen los mismos valores de los atributos. De este modo se facilita el almacenamiento y la recuperación de los datos pudiéndose manejar información de miles de agentes fácilmente. En estas tablas se pueden eliminar atributos, colapsando sus categorías, con lo que se puede obtener información agregada de los mismos, que llamamos distribuciones marginales de los atributos. Estas tablas tendrán tantas dimensiones como atributos se consideren. Una tabla cuyas celdas corresponden a categorías de un único atributo se denominará tabla unidimensional, y nos referiremos a ella como la distribución de marginales unidimensionales de dicho atributo.

En la Figura 1 se muestra un ejemplo que ilustra las distintas formas en la que pueden agregarse los microdatos para un caso con tres atributos. El primer nivel corresponde a la tabla tridimensional. Las columnas de estas tablas siempre corresponden a las categorías del último atributo. En función del atributo que se elimine se obtendrán distintas tablas 2-dimensionales y, eliminado otro atributo, las 1-dimensionales.

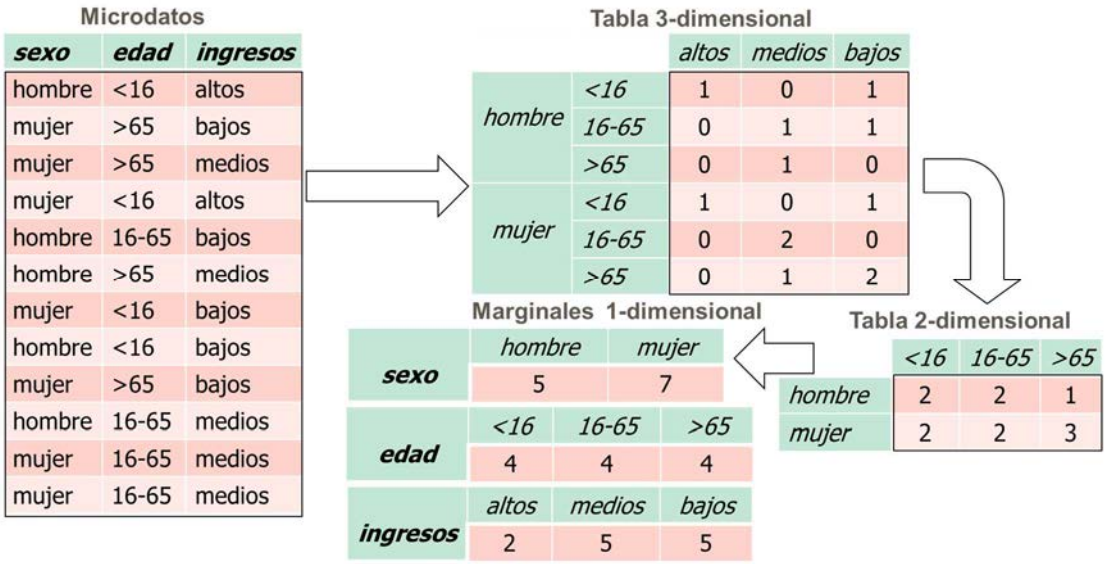

Figura 1 Microdatos, tabla 3-dimensional, 2-dimensional y marginales unidimensionales.

En el caso general de una población caracterizada por M atributos categóricos $X = (x_1, x_2, \dots x_{\mathrm{M}})$, cada celda de la tabla M-dimensional contendrá el número de agentes de la población que tienen una determinada combinación de valores de los atributos. Si $(\mathbf{C_1}, \mathbf{C_2}, \dots \mathbf{C_M})$ representan el número de categorías de cada atributo, el número total de celdas de la tabla multidimensional será el producto de estos números de categorías, $\mathbf{C_1} * \mathbf{C_2} * \dots * \mathbf{C_M}$. Si los atributos son socio-económico-demográficos, cada celda se corresponderá a un grupo socio-económico-demográfico distinto.

Si $t_{c_1 c_2 \dots c_{\mathrm{M}}}$ representa el valor de la celda $c_1 c_2 \dots c_{\mathrm{M}}$, y $t_{c_i + \dots +}$ es el valor del marginal unidimensional de la categoría c_i del primer atributo, la suma de los marginales unidimensionales de cada atributo será el número total de individuos de la población, N. Esto es,

$$\sum_{c_1=1}^{\mathbf{C_1}} t_{c_1 + \dots +} = \sum_{c_2=1}^{\mathbf{C_2}} t_{+ c_2 + \dots +} = \dots = \sum_{c_{\mathrm{M}}=1}^{\mathbf{C_M}} t_{+ \dots + c_{\mathrm{M}}} = \mathrm{N}$$

Para caracterizar y comparar las distribuciones de las tablas multidimensionales, es conveniente introducir dos tipos de medidas que serán de utilidad cuando se describan los métodos de generación de poblaciones. El primero es la entropía y el segundo los *odd ratio*[1] condicionales (COR, del inglés *Conditional Odd Ratio*).

[1] El término en español es "razones de odds" o "razón de razones", ya que el odd es un cociente que expresa cuantas veces es mayor la probabilidad de que ocurra un suceso que la de que no ocurra, aunque en este documento se ha preferido mantener el nombre en inglés.

La entropía se usa para determinar el grado de dispersión o variabilidad de una distribución de probabilidad (Agresti, 2007). Si se consideran los valores de la tabla multidimensional como una distribución de probabilidad, esto es, dividiendo el valor de cada celda entre el total de la población, puede calcularse la entropía de la distribución de la tabla, que para el caso de tres atributos es:

$$H = -\sum_{c_1}\sum_{c_2}\sum_{c_3} \dot{t}_{c_1c_2c_3} \log(\dot{t}_{c_1c_2c_3})$$

El punto sobre la $\dot{t}$ indica que se trata de probabilidades, valores de la tabla divididos por el total N. Y la entropía relativa de dos tablas tridimensionales $t_{c_1c_2c_3}$ y $s_{c_1c_2c_3}$ es:

$$H(t|s) = \sum_{c_1}\sum_{c_2}\sum_{c_3} \dot{t}_{c_1c_2c_3} \log(\dot{t}_{c_1c_2c_3}/\dot{s}_{c_1c_2c_3})$$

Los *odd ratio* condicionales (CORs) de una tabla multidimensional miden la asociación entre los atributos de la tabla. Para el caso de 3 atributos (Figura 2), el COR para las categorías i_1 e i_2 del atributo x_1 y las categorías j_1 y j_2 del atributo x_2, condicionado a las categorías k_1, k_2 del atributo x_3, puede visualizarse como la razón del producto cruzado (*cross product*) entre 4 celdas de la tabla 3-dimensional.

		x_2		
x_3	x_1	j_1	j_2	...
k_1	i_1	$\pi_{i_1j_1\|k_1}$	$\pi_{i_1j_2\|k_1}$	
	i_2			
	...			
k_2	i_1			
	i_2	$\pi_{i_2j_1\|k_2}$	$\pi_{i_2j_2\|k_2}$	
	...			
...	...			

$$COR^{x_1x_2|x_3}_{i_1i_2j_1j_2|k_1k_2} = \frac{\pi_{i_1j_1|k_1}\,\pi_{i_2j_2|k_2}}{\pi_{i_2j_1|k_2}\,\pi_{i_1j_2|k_1}}$$

Figura 2 Odd Ratio Condicional en una tabla 3-dimensional.

Los CORs de una tabla estarán condicionados a una cantidad mayor de categorías de atributos a media que aumenta la dimensionalidad de la tabla, ya que para determinar una celda de la tabla habrá que especificar más categorías de atributos.

En el caso de una tabla con dos atributos, al no existir un tercer atributo que condiciona, solo existen los *odd ratios* (cociente de productos cruzados), los cuales son una medida de la asociación entre los dos atributos. Si todos los *odd ratios* son 1, no hay correlación entre los atributos, por lo que los atributos son independientes. Los *odd ratios* pueden tomar valores entre 0 e infinito, y cuanto más alejados de la unidad (considerando distancia logarítmica) mayor será la asociación entre las categorías de los atributos. El sentido de la asociación cambia si el odd ratio es mayor o menor que 1.

Por último, para finalizar con la terminología, se define el concepto de población multinivel. Puede ocurrir que sea necesario agrupar los agentes de la población en grupos que constituyen entidades de otro nivel. Es el caso en que los distintos individuos de una población se agrupan en familias constituidas por 2 o más individuos. Estos nuevos colectivos constituyen una población de segundo nivel, que a su vez pueden volver a agruparse y constituir una población de tercer nivel, como sería la agrupación de familias e individuos en hogares asociados a una vivienda. Cada entidad de cada nivel (individuo, familia y hogar) tendrá atributos específicos propios. Se denomina población multinivel al conjunto de poblaciones que representan los distintos niveles. El uso de poblaciones multinivel es un requisito fundamental para algunas aplicaciones, como se verá más adelante.

2.2 Estimación de "small area": SAE

Las encuestas realizadas a una muestra de una población se utilizan para estimar variables mediante estimadores estadísticos directos.

El término estimación de área pequeña (SAE, del inglés *Small Area Estimation*) se refiere a la situación en la que se desea estimar el valor de al menos una variable de interés de una subpoblación, mediante una encuesta realizada a una muestra de la población superior, y no es posible estimar directamente, con suficiente fiabilidad, el valor de la variable de interés en la subpoblación.

La razón para no poder estimar fiablemente la distribución de valores de la variable de interés mediante estimadores estadísticos directos puede deberse a dos motivos.

Puede ocurrir que la encuesta se haya realizado en una población mayor, que incluye a la subpoblación, y no se disponga de una variable que identifique los elementos de la muestra

que pertenecen a dicha subpoblación, con lo que no es posible aplicar métodos de estimación directa.

Pero también puede ocurrir que la encuesta disponga de dicha variable, pero el número de elementos de la muestra pertenecientes a la subpoblación sea demasiado pequeño, con lo que la estimación estadística con métodos directos produce estimadores excesivamente poco precisos.

En ambos casos, a esta subpoblación que representa un dominio donde no puede estimarse con suficiente precisión el valor de la variable de interés con estimadores estadísticos directos se denomina "*small area*".

Es por esto que "*small area*" se refiere al "dominio" (*small domain*), "subpoblación", o grupos socio-demográficos dentro de un área, en los que el tamaño de muestra es pequeño, con independencia del tamaño del área o de la población.

Ejemplos de estos dominios que cubren distintos tamaños geográficos son: nación, región, provincia, comarca, municipio, barrio, distrito, área de servicio de salud, etc., y un ejemplo de dominio, en su acepción de "grupo socio-demográfico", es un grupo de personas de una raza y edad determinadas dentro de un área geográfica extensa. Todos son "*small areas*".

A lo largo de este documento se utilizará el término en inglés de "*small area*", para evitar que se asocie con un área de pequeñas dimensiones, puesto que este nombre va ligado al tamaño y características de la muestra más que al tamaño del área o de la población, e indica una partición de una población de acuerdo a criterios geográficos o a características socio-demográficas o económicas, etc.

Los métodos de estimación de "*small area*" (*SAE methods*) tratan el problema de obtener estimadores fiables de las variables de interés en dominios en los que no se dispone de una muestra que incluya la variable interés, o que aun incluyéndola, no sea posible la estimación directa debido al pequeño tamaño de la muestra.

Estos métodos se denominan estimadores indirectos, ya que se apoyan en información de muestras de dominios superiores, que pueden englobar a la *"small area".* Utilizan los datos de una encuesta diseñada para generar estimaciones fiables en un gran dominio, para construir estimaciones fiables en una subpoblación de dicho domino, sin tener que aumentar el tamaño de la muestra, lo cual exigiría costes adicionales.

Cada vez hay mayor demanda de información estadística de "*small area*s", siendo preciso estimar variables de interés en estas "*small area*s" sin tener que recurrir a realizar encuestas con el tamaño adecuado para proporcionar estimadores directos. Por ejemplo, es el caso de

la estimación de la tasa de desempleo de un municipio a partir de los datos de la Encuesta de Población Activa (EPA), que se hace por comunidades autónomas y donde la mayoría de los municipios no están representados. Otros casos típicos de información de "*small area*" son la estimación de la tasa de pobreza en comunidades de minoría étnica, la tasa de madres solteras en paro, la proporción de jubilados que precisan cuidados específicos en los suburbios de una ciudad, etc.

Para llevar a cabo estas estimaciones existen múltiples métodos indirectos. Esta tesis se centra principalmente en métodos con un enfoque estadístico-geográfico que comprenden la generación de microdatos sintéticos con los que realizar las estimaciones. Estos métodos combinan datos de encuestas que incluyen la variable de interés, pero no están desagregados a nivel de "*small area*", con datos de otras fuentes, principalmente datos censales y de las administraciones públicas, donde ocurre lo contrario (desagregados por "*small area*", pero no incluyen la variable de interés, Figura 3).

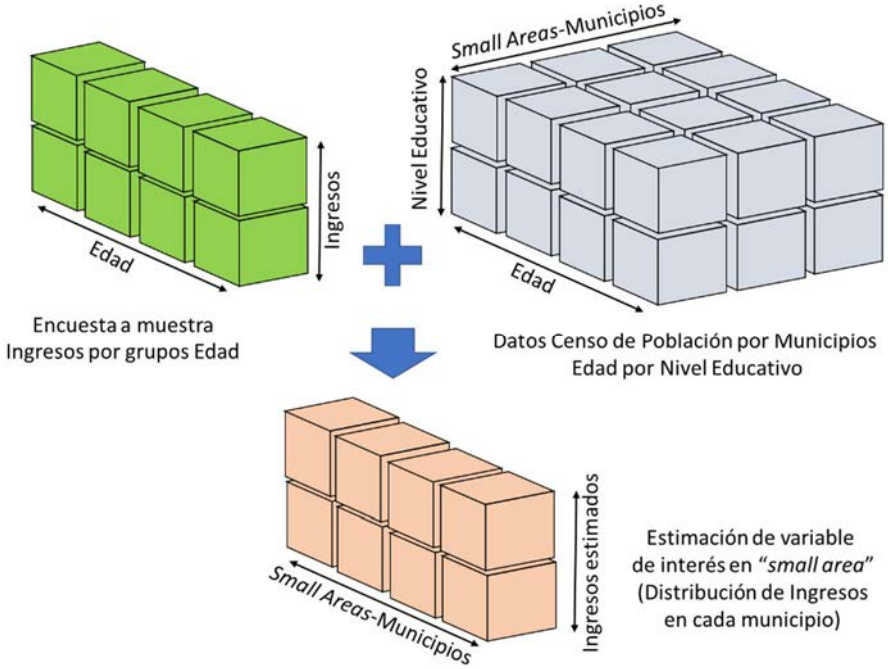

Figura 3 Estimación de "small area". Adaptado de (Dobre & Caragea, 2015).

El resto de métodos indirectos son métodos con un enfoque probabilístico, basados en modelos estadísticos predictivos de la variable de interés en función de otras variables auxiliares, que utilizan los principios de estadística bayesiana y otras técnicas estadísticas (Rao & Molina, 2015).

Aunque esta tesis pone especial foco en el primer grupo de métodos, con enfoque geográfico y de creación de microdatos (métodos de reponderación de muestras), se describirán

también algunos métodos de generación de poblaciones del tipo probabilístico, con objeto de presentar una visión panorámica de los métodos de generación de poblaciones sintéticas.

2.3 Modelos de Microsimulación Espacial

Los modelos de microsimulación (MSM, del inglés *Microsimulation Model*) son representaciones simplificadas de una población donde se simulan políticas, acciones o comportamientos basados principalmente en probabilidades y estadística que modifican los valores de los atributos de los individuos de la población, y donde cada ocurrencia de un suceso se basa en su probabilidad.

Con estos modelos es posible analizar el impacto de determinadas políticas en la población y hacer análisis comparativos del impacto probable que tendrán dichas políticas. Un ejemplo de este tipo de sistemas es el Euromod, un modelo de microsimulación fiscal de libre acceso (Institute for Social and Economic Research, 2018) diseñado para evaluar el impacto de distintas políticas fiscales en los países de la unión europea, teniendo en cuenta las características de la población de los mismos.

Los modelos de microsimulación espacial (SMSM, del inglés *Spatial Microsimulation Model*) son modelos MSM que generan la población que se desea modelar incorporando atributos de localización espacial de "*small area*", en su acepción de partición de una población con criterios geográficos. Los SMSM añaden la componente de localización a las variables y atributos de dicha población, con lo que se posibilita el análisis del impacto de políticas y actuaciones específicas (sociales, económicas, sanitarias, laborales, demográficas, de seguridad, etc.) entre los integrantes de dicha población (principalmente hogares e individuos) a nivel geográfico, esto es, en las distintas áreas espaciales.

Estos modelos combinan datos de encuestas (Figura 3), por ejemplo la Encuesta de Presupuestos Familiares (EPF), con datos del Censo, o con totales poblacionales proporcionados por distintas organizaciones, con objeto de determinar la distribución de las variables de interés (ej. Ingresos de los hogares) en las distintas áreas, y de este modo poder estimar, por ejemplo, la distribución de la tasa pobreza en los distintos barrios de un municipio a partir de determinadas políticas fiscales o de bienestar (Fenton, 2016).

Un representante de este tipo de modelos es SMILE, del inglés *Simulation Model for the Irish Local Economy*, (Ballas, Clarke, & Wiemers, 2006) que modela la distribución regional del trabajo e ingresos de los hogares de Irlanda, permitiendo elaborar un análisis económico de las distintas zonas rurales de este país.

Los modelos SMSM pueden incluir una componente dinámica que describa la evolución temporal de la población (conjunto de probabilidades de cambio de estado civil, de estado de salud, nacimiento, muerte, cambio de residencia, cambio de vehículo, etc.). De este modo, permiten proyectar en el tiempo la población de cada "*small area*" con objeto de poder estudiar las características de la misma dentro del contexto local, con lo que se tienen predicciones a escala de "*small area*" que son importantes a la hora de planificar o hacer análisis de políticas en una dimensión espacial. A estos modelos se los conoce con el nombre de SMSM dinámicos. En ocasiones, a partir de estas microsimulaciones dinámicas se construyen modelos de interacción espacial, donde se analiza el flujo de individuos entre las "*small areas*", en función de las características de los individuos y del atractivo de las mismas.

Un ejemplo de SMSM dinámico es SimBritain que simula toda la población de Gran Bretaña con un tamaño de "*small area*" correspondiente a distrito electoral (5.500 individuos aprox.) hasta 2021 (Ballas, Clarke, et al., 2005).

Todos estos modelos precisan de poblaciones descritas por los atributos de interés para el modelo, las cuales no siempre están disponibles, por lo que han de construirse sintéticamente.

2.4 Modelos de Simulación basados en Agentes (Agent-based modelling)

Los modelos de simulación basados en agentes (SBA), también llamados ABM, del inglés *Agent-based models*, son unas herramientas muy utilizadas en la mayoría de las ramas de las ciencias sociales. Se trata de modelos computacionales que permiten crear, analizar y experimentar con sistemas compuestos por agentes que interactúan dentro de un entorno (Gilbert, 2008).

Los modelos de SBA modelan el comportamiento independiente de los agentes. Cada agente sigue unas reglas intrínsecas, tomando sus propias decisiones y ejecutando acciones, conforme a estas reglas y al conocimiento que va adquiriendo, mediante la interacción con otros agentes y con el entorno en el que se desenvuelve. Muchas veces, el entorno se describe mediante localizaciones y espacios geográficos donde el agente puede situarse.

Mediante interacciones simples y predecibles, se consiguen patrones de conducta global que no podrían haberse predicho con solo entender a cada agente en particular.

A diferencia de los MSM de la sección anterior, los modelos de SBA se apoyan en reglas de comportamiento de los agentes. Estas reglas pueden incorporar probabilidades, pero no es un requisito (Birkin, 2008). Las reglas de comportamiento hacen que se modifiquen los

valores de los atributos de los agentes. Los experimentos de simulación permiten analizar el efecto de distintas reglas.

Existen múltiples herramientas para desarrollar las SBA, entre las que podemos destacar herramientas de código abierto tales como NetLogo o Repast Symphony. *("ABM Software Comparison,"* 2018).

Para construir la simulación, es preciso disponer de un modelo de comportamiento de los agentes, y simular las operaciones simultáneas de múltiples agentes en un intento de recrear y predecir el comportamiento global del sistema que se simula.

Se trata de un proceso ascendente *bottom-up*, desde el nivel más elemental (micro) al más elevado (macro). Es decir, los modelos SBA se basan en un enfoque que se centra en los participantes individuales del sistema, y obtiene conclusiones acerca del comportamiento del sistema macroscópico.

Así por ejemplo, en el caso de un modelo de SBA en el entorno de mercados financieros de bolsa de valores, se simularía el comportamiento de dicho mercado a partir del comportamiento de la población de operadores de bolsa participantes en el mismo, en contraste al enfoque tradicional de análisis de mercado de capitales.

Estos modelos permiten no asumir que los participantes tomen decisiones de inversión racionales basadas en la información de datos macroeconómicos, como la tasa de interés o de cambio de divisas, ni que el mercado responda a un modelo descendente *top-down* según ecuaciones matemáticas.

En su lugar, consideran que las decisiones de los operadores de bolsa y el comportamiento del mercado son el resultado de las acciones de los agentes que participan en el mismo, los cuales reaccionan, según ciertos parámetros, ante los precios de las acciones y los cambios que se producen en los mismos.

De esta manera, mediante la SBA puede llegarse a entender la relación entre los parámetros y atributos que definen el comportamiento de los participantes en el mercado de valores y los distintos patrones de precio de acciones a nivel macro creados mediante las decisiones de compra/venta de los agentes individuales. En definitiva, este tipo de modelos de simulación pretenden ser útiles para entender mejor el mercado.

Generalmente los agentes individuales actúan según lo que perciben como sus intereses propios, tales como beneficio económico, status social, o reproducción y su conocimiento es limitado. Los agentes pueden llegar a experimentar aprendizaje, adaptación y/o reproducción.

Tal como indican Macal & North (2008), un agente es:

- Un individuo identificable, discreto o modular con un conjunto de características, y reglas que rigen sus comportamientos y capacidad de toma de decisiones.
- Autónomo y auto dirigido.
- Social, por lo que interacciona con otros agentes.
- Reside dentro de un entorno externo con el que el agente interactúa, además de interactuar con otros agentes del entorno.
- Diseñado para conseguir unas metas y objetivos, lo que permite comparar los logros de su comportamiento respecto a dichos objetivos.
- Flexible y con capacidad de aprender y adaptar sus comportamientos conforme a su experiencia.

En la Figura 4 se resumen los componentes de un agente.

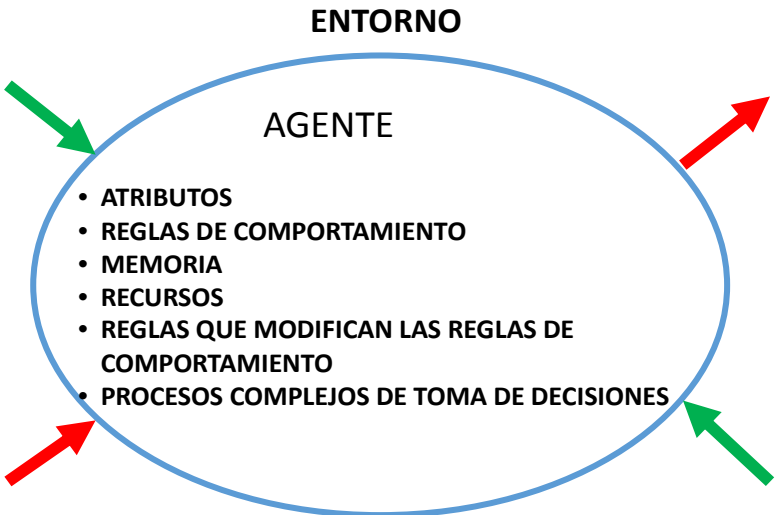

Figura 4 Agente. Elaboración propia basada en (Macal & North, 2008).

Las reglas de comportamiento de cada agente son desiguales, pueden variar en función de múltiples factores, como la cantidad de información disponible por el agente (carga cognitiva), los modelos internos del agente sobre el mundo exterior, las reacciones o comportamientos de otros agentes, o la cantidad de recuerdos de acontecimientos pasados que se utilicen.

Los agentes se distinguen unos de otros según sus atributos y recursos acumulados. Es muy importante para muchas de las simulaciones disponer de una población de agentes con los atributos, recursos y capacidades, lo más real posible.

La importancia de este tipo de simulaciones va en aumento, y prueba de ello es el gran número de áreas en las que se utilizan. En la Tabla 1, se muestran algunas de estas áreas. Las

aplicaciones de la SBA cubren áreas muy diversas, tales como la indicada de análisis del comportamiento del mercado de valores, u otros temas como el análisis de cadenas de suministro, la predicción de la propagación de epidemias o de la amenaza de guerra biológica, modelado del crecimiento y declive de las civilizaciones antiguas, modelado del complejo sistema inmune del ser humano, la desregulación del mercado de la energía, o estudio de dilemas sociales, solo por nombrar algunos.

Área	Aplicación	Ejemplos Poblacionales
Organización de Empresas	• Operaciones de Fabricación • Cadena de Suministro • Mercados de Consumo • Seguros	Pedidos/órdenes de fabricación Consumidores/clientes Empresas Proveedores Asegurados
Economía	• Mercados Financieros • Redes comerciales • Comportamientos del consumidor	Propietarios de Acciones Clientes y Socios comerciales Redes de distribución (Órdenes de compra/venta)
Infraestructuras	• Tráfico/Transporte • Mercado eléctrico • Infraestructuras carga gas Hidrógeno/carga eléctrica	Viajeros (modelos basados en Actividad) Consumidores Consumidores y Proveedores de puntos de carga para vehículos a gas.
Multitudes	• Movimiento peatonal • Modelos de evacuación	Peatones Individuos en un Centro
Sociedad y Cultura	• Civilizaciones antiguas • Desobediencia civil • Determinantes sociales del terrorismo • Dilemas Sociales Organizacionales	Población de sociedades / transmisión de información Población de un entorno cultural
Medicina	• Salud Pública • Planificación de respuesta ante epidemias. Propagación enfermedades • Desarrollo de fármacos	Enfermos Poblaciones vulnerables Pacientes
Defensa	• Mando y Control • Enfrentamientos	Ejército Comando de Guerrillas
Biología	• Dinámica de poblaciones • Redes Ecológicas • Comportamientos de grupos de animales • Comportamiento de células y procesos subcelulares	Población de células Ecosistemas Población de animales

Tabla 1 Ejemplos de áreas de aplicación de la Simulación Basada en Agentes. Elaboración propia a partir de (Macal & North, 2008)

Las aplicaciones de SBA varían desde pequeños modelos que solo tienen en cuenta las principales características de los sistemas a simular, hasta modelos a gran escala de sistemas de apoyo a la toma de decisiones. Los primeros suelen ser modelos exploratorios que a modo de laboratorio de ensayos permiten simular múltiples supuestos. Los modelos de simulación para los sistemas de apoyo a la toma de decisiones, se diseñan para responder a una amplia gama de preguntas sobre el comportamiento del mundo real. Estos modelos se distinguen por incluir datos reales y por tener que someterse a algún grado de pruebas de validación para establecer la credibilidad en sus resultados.

Los datos reales que incluyen estos modelos, aparte de las reglas de comportamiento de los agentes que se desean validar, son atributos de la población que se trate, bien sean clientes, animales, viajeros o células. El modelo basa sus resultados en la interacción de estos agentes, bien entre ellos o con el entorno. Por lo que, aunque se disponga de un modelo de comportamiento de los agentes muy elaborado, si no se dispone de una población real, no se obtendrán conclusiones aplicables al mundo real, o no podrán validarse con observaciones reales.

Por tanto, estas aplicaciones de SBA requieren un conjunto de agentes que representen de la manera más fiel posible a la población que se quiere simular, con una distribución de atributos lo más realista posible.

Así por ejemplo, se usa SBA para construir modelos urbanos que den explicación a la segregación racial que se observa en las ciudades, y dichos modelos precisan de datos de la población de la ciudad que se asemejen a la realidad que se quiere modelar.

Los modelos de análisis de la dinámica de las opiniones, tratan de entender el desarrollo de opiniones políticas, por ejemplo, explicar el desarrollo de opiniones extremas dentro de una población, y necesitan partir de una población lo más real posible, donde se van a transmitir las opiniones, con arreglo a unos supuestos.

En medicina, la SBA ha demostrado ser gran valor para analizar las políticas de salud pública destinadas a la preparación o respuesta a epidemias, utilizando poblaciones sintéticas vulnerables. También se usan SBA para simular ensayos clínicos y mejorar el desarrollo de fármacos, tal como describe Tannenbaum, Holford, Lee, Peck, & Mould (2006). En este caso se simula el desarrollo de ensayos clínicos de fármacos en una población virtual de pacientes, considerando el cumplimiento de los horarios de dosificación y abandonos del ensayo por parte de algunos pacientes, e incluyendo modelos que describen el progreso de la enfermedad a lo largo del ensayo, y la farmacocinética y farmacodinamia del fármaco que se prueba. La población de pacientes se describe con atributos con determinadas distribuciones correlacionadas. Entre estos atributos se incluyen factores específicos del paciente que

pueden explicar las diferencias farmacocinéticas y de farmacodinamia observadas a nivel individual, teniendo en cuenta también las características demográficas de los pacientes.

Estudios recientes sobre simulación de transporte y modelado de enfermedades han resaltado el valor de las poblaciones sintéticas (Guo & Bhat, 2007a), especialmente cuando las heterogeneidades en la mezcla de poblaciones juegan un papel crucial, como en la transmisión de enfermedades (Geard et al., 2015), o cuando la incidencia de la enfermedad o los riesgos de infección varían significativamente entre subgrupos, como los grupos de edad.

Cuando se desarrolla una simulación social a escala real, es todo un desafío la preparación de una población con agentes en su entorno a escala real (Gilbert, 2008). Cada agente debe tener sexo, edad, y una familia como mínimo.

En resumen, en muchos estudios de SMSM y SBA se precisa disponer de una población caracterizada por múltiples atributos socio-económico-demográficos y espaciales, que sea el fiel reflejo de una población real (Pei-jun Ye, Wang, Chen, Lin, & Wang, 2016). Dada la falta de fuentes que proporcionen el detalle de la población completa, y el coste que puede suponer el conseguirla, los investigadores acuden a generar dichas poblaciones de forma sintética (Tanton & Edwards, 2013).

2.5 Modelos basados en Actividad

Las simulaciones basadas en agentes que se han explicado permiten simular el tráfico en una región. Los modelos consideran el transporte como fruto de la planificación de actividades en el tiempo, y tienen en cuenta los condicionantes espaciales y personales de cada hogar y de los individuos que lo constituyen. Por tanto, los desplazamientos de los individuos se establecen en función de las actividades que realizan, *activity-based* (ir al trabajo, ir a la compra, ir a recoger a los hijos al colegio, etc.).

Para determinar estos desplazamientos se construyen los llamados Modelos basados en Actividad (ABM, del inglés *Activity Based Modeling*), también conocidos como modelos de generación de demanda de desplazamientos basados en actividad (*Activity-Based Demand Generation*, ABDG) para no confundirlos con las siglas ABM de las simulaciones basadas en agentes. Con estos modelos se generan cadenas de actividad que son utilizadas para producir los planes de desplazamiento de los agentes que participarán posteriormente en la simulación de tráfico basada en agentes.

Los ABM son modelos de generación de demanda de tráfico dentro de una región. A partir de los desplazamientos de una muestra de la población se obtienen los desplazamientos de

los distintos grupos de individuos con atributos socio-económico-demográficos similares, cada uno de los cuales se supone tienen un comportamiento homogéneo, de esta forma se predice el patrón de desplazamientos de cada individuo. Con este tipo de modelos se determinan los modos de transporte, tiempos, localizaciones y patrones de actividad de cada uno de los individuos de la región donde se desea analizar el tráfico.

A cada uno de estos grupos de individuos se asigna un grupo de cadenas de actividad. Cada cadena de actividad contiene información sobre un determinado conjunto de actividades, el lugar donde se desarrollan y la frecuencia de ocurrencia (diaria, en hora punta, matutina o vespertina, días concretos, etc.). Las actividades se describen por el tipo, ubicación y modos de transporte disponibles para desplazarse a esa ubicación (coche, tren, a pie, etc.). La localización de la vivienda del hogar es el punto de inicio y fin de cada secuencia de actividad. La información sobre ubicaciones suele estar referida a zonas de análisis de tráfico (TAZ, del inglés *Traffic Analysis Zone*) dentro de la región, que representan zonas que son origen o destino de corrientes de tráfico. Las cadenas de actividad no incluyen información sobre el instante de tiempo en que se llevan a cabo las actividades, ni el modo de transporte utilizado. Para transformar las secuencias de actividad en planes de desplazamiento adaptados en el tiempo de cada individuo se precisa conocer la información desagregada de las características del individuo. Entre las características de los individuos que influyen en la generación del plan de desplazamientos, se encuentran atributos como: estructura y tamaño del hogar, nivel de ingresos, disponibilidad de vehículo, disponibilidad de transporte público, actividad, edad y localización del individuo respecto a la localización de la actividad.

Existen dos tipos principales de enfoques para determinar los planes de viajes de los individuos: los enfoques econométricos basados en la maximización de la utilidad y los basados en reglas. En los primeros, los modelos de actividad están basados en un conjunto predefinido de secuencias de actividad (Casa-Trabajo-Restaurante-Trabajo-Casa; Casa-Trabajo-Compras-Ocio-Casa; Casa-Escuela-Almuerzo-Visita a Familiares-Casa; etc.), que utilizan modelos que maximizan la utilidad para construir los planes de desplazamiento, y se construyen con modelos de elección discreta (probit, logit multinomial, etc.) con objeto de explicar la decisión que reporta más utilidad al individuo entre las alternativas de viajes que tiene disponibles (Ben-Akiva, Bowman, & Gopinath, 1996). Los segundos crean planes de desplazamiento basados en reglas heurísticas, como es el caso de ALBATROSS (Arentze & Timmermans, 2004) o FEATHERS (Bellemans et al., 2010). Aunque también hay implementaciones que combinan ambos enfoques (Pendyala, Kitamura, Kikuchi, Yamamoto, & Fujji, 2005), o que incorporan enfoques estocásticos utilizando modelos aleatorios para asignar las cadenas de actividad a cada agente (Barthelemy & Toint, 2015).

Las alternativas de viaje no tienen utilidad por sí solas, sino que la utilidad se deriva de la combinación de las características de las alternativas y de los individuos. Como se ha indicado, para transformar las cadenas de actividad en planes de desplazamiento de los individuos, bien en función de utilidad o reglas, hay que disponer de información desagregada sobre la ubicación y los restantes atributos de interés de cada individuo.

Por consiguiente, es fundamental que estos modelos dispongan de datos desagregados relativos a la población de la región cuyo comportamiento se desea modelar. Por tanto, para cada TAZ de la región donde se simulan los desplazamientos es necesario conocer la distribución de las características/atributos de los hogares y de todos sus miembros.

En la mayoría de los casos no están disponibles los datos de los atributos desagregados de la población completa de la región en la que se desea modelar la generación de la demanda de transporte, por lo que se hace necesario generar poblaciones sintéticas que se ajusten a la distribución de atributos de las personas y de los hogares de las TAZ con distintos métodos de generación de poblaciones sintéticas.

2.6 Otros Usos: Protección de datos y Datos para Ensayos y Pruebas.

En muchas ocasiones se dispone de datos originales (completos o incompletos), y se desea obtener unos datos equivalentes para poder difundirlos protegiendo la confidencialidad de los mismos, permitiéndose la difusión de los datos de forma controlada. Las agencias estadísticas, como el Instituto Nacional de Estadística (INE), divulgan sus datos de forma anónima, protegiendo los datos confidenciales con distintos métodos de anonimización.

Existen distintos métodos de anonimización de los datos, como son los que se basan en perturbar los datos tratando de mantener sus propiedades. Entre los métodos de perturbación de datos más habituales están la agregación, recodificación, supresión de valores sensibles, adición de ruido o métodos de redondeo utilizados para publicar datos en forma tabular y proteger la información sensible.

Adicionalmente a estos, hay otros métodos de protección de datos confidenciales que usan técnicas de imputación de datos basadas en modelos estadísticos, los cuales están considerados como métodos de generación de datos sintéticos a partir de los datos reales.

Estos métodos permiten generar múltiples poblaciones sintéticas a partir de una población original, manteniendo las propiedades estadísticas de la misma, pero protegiendo la información sensible.

Con objeto de ofrecer un mapa completo de métodos de generación de datos sintéticos, en el siguiente capítulo se revisan distintos tipos de métodos de generación de poblaciones, entre los que se incluyen métodos estadísticos basados en modelos con imputación múltiple, utilizados para generar datos sintéticos para anonimización.

Otra típica utilización de datos sintéticos es la de ensayos y pruebas, donde se exige que dichos datos exhiban determinadas propiedades, tales como, niveles determinados de auto-correlación, ciertos grados de disparidad o de proximidad, etc.

Entre los principales motivos por los que se usan datos sintéticos para ensayos y pruebas están su menor coste comparado con la obtención de datos reales y la protección de la privacidad que suponen frente al uso de datos reales. La producción de datos sintéticos a través de un modelo de generación es significativamente más rentable y eficiente que la recopilación de datos del mundo real.

Es el caso de los sistemas de información dedicados a la minería de datos, que se diseñan para identificar correlaciones múltiples entre los datos e identificar sucesos relevantes. Estos sistemas se simulan y validan con datos de prueba que requieren datos con determinadas correlaciones. La obtención de datos de prueba puede ser compleja debido a problemas de privacidad, de tiempo y el coste que pueden suponer.

Otras veces es la confidencialidad de los datos lo que motiva que se recurra a datos sintéticos. Para estudiar las limitaciones del software de las redes sociales se utilizan datos sintéticos, dada la naturaleza privada de los datos que en ellas se almacenan. De este modo, han surgido aplicaciones de código abierto como SONETOR, un generador de tráfico sintético para redes sociales que simula el comportamiento de los usuarios en la red, la publicación y consumo de los contenidos, como cuando estos usuarios comparten y comentan la información de sus amigos (Bernardini, Silverston, & Festor, 2014).

En algunas ocasiones es la falta de datos reales la razón por la que se utilizan datos sintéticos. Los sistemas de detección de intrusiones son un ejemplo de sistemas donde se trabaja con este tipo de datos, tal como indican Lundin, Kvarnstrom, & Jonsson (2002), en la mayoría de los proyectos de sistemas de detección de intrusión se utilizan datos sintéticos debido a la falta de datos reales o la falta de datos con las propiedades deseadas, debido a la escasez de datos de registro de ataques reales a los sistemas.

Barse et al. (2003) muestran un caso de generación de datos sintéticos de prueba para entrenar y probar un sistema de detección de fraude en un servicio de *video-on-demand* basado en IP, donde se generan datos de registro (*log*) sintéticos con las mismas propiedades estadísticas que un pequeño conjunto de datos de l*og* reales.

En un reciente estudio de Schatsky & Chauhan (2017), de la consultora Deloitte, se indica que uno los factores que disminuirá la barrera de acceso a las aplicaciones con aprendizaje automático es la reducción de la cantidad de datos de entrenamiento que dichas aplicaciones requieren, la cual puede conseguirse con la generación de datos sintéticos. La consultora realizó un experimento desarrollando una aplicación de aprendizaje automático con datos de entrenamiento reales, posteriormente utilizó una muestra del 20% de dichos datos para generar datos de entrenamiento sintéticos. Con los nuevos datos sintéticos repitió el mismo trabajo y obtuvo el mismo resultado. Con esto evidenció un caso en el que se reducían los datos de entrenamiento reales en un 80%.

Jeske, Gokhale, & Ye (2006) aborda el problema de cómo integrar toda la información parcial en un esquema no-paramétrico de generación de datos sintéticos para pruebas de minería de datos. Estos investigadores describen un esquema que incorpora toda la información que puede encontrarse sobre asociación entre atributos, pero sin forzar una estructura adicional dentro del esquema. Dicho esquema se basa en uno de los métodos más comunes de generación de datos sintéticos, como es el *Iterative Proportional Fitting* (IPF), que se describirá en detalle en el siguiente capítulo.

Con este método reconstruyen una tabla multidimensional de datos sintéticos a partir de ciertas distribuciones bidimensionales que representan la asociación entre atributos.

Estos investigadores llegan a construir tablas de datos para pruebas con más de 16 millones de celdas, usando 23 distribuciones marginales bidimensionales de 9 atributos.

En conclusión, y como resumen de este capítulo, podemos afirmar que los datos y poblaciones sintéticas son una necesidad real en la era de los datos, son objeto de estudio y se utilizan cada vez con mayor frecuencia en múltiples campos.

3 Métodos de generación de poblaciones sintéticas

3.1 Introducción

En esta sección se hace una revisión de los principales métodos de generación de datos sintéticos, con especial foco en datos de poblaciones. Antes de proceder a clasificar y describir los tipos de métodos, hacemos la distinción entre método y técnica.

En este estudio, el concepto de método se refiere a un modo de hacer que se traduce en una serie de acciones para conseguir un determinado objetivo, realizándose dichas actuaciones con diversas técnicas. Mientras que el concepto de técnica apunta a uno o varios procedimientos, algoritmos, instrucciones o reglas para llevar a cabo una actividad de forma efectiva. En este caso, los métodos de generación de datos sintéticos tienen como objetivo la obtención de una población que pueda reemplazar a una población real, para lo cual utilizan técnicas específicas como la técnica del IPF, técnicas de búsquedas meta-heurística, técnicas de muestreo y de redondeo, técnicas de programación lineal, técnicas o modelos probabilísticos, o cualquier otra técnica estadística. Por tanto, encontraremos métodos que utilizan una misma técnica, aunque integrada en un proceso distinto.

En relación a la clasificación de estos métodos, fue en primer lugar Rahman (2009) y posteriormente Hermes & Poulsen (2012) quienes efectuaron una revisión de los tipos de métodos de generación de poblaciones sintéticas. Estos investigadores distinguieron dos grandes familias: los métodos de re-ponderación de la muestra (*Sample Reweighting*[2]) y los de Reconstrucción Sintética (*Synthetic Reconstruction*). Los primeros requieren disponer de una muestra desagregada con todos los atributos, cuyos individuos se replican en las proporciones adecuadas para producir la población sintética, por lo que en la población sintetizada solo aparecen elementos que están representados en la muestra. Los segundos no requieren una muestra con todos los atributos de la población, sino que están orientados a integrar información de distintas fuentes.

La clasificación anterior desatiende un gran conjunto de métodos con enfoque probabilístico, algunos de los cuales han sido desarrollados en los últimos años, por lo que en esta tesis se ha preferido reclasificar los métodos agrupando por un lado los de *reweighting*, denominados por algunos autores como métodos de ajuste (*fitting*) basados en muestra (Saadi et al., 2016), ya que ajustan a marginales, y por otro todos los demás métodos que no pueden considerarse de *reweighting,* a los que denominaremos "probabilísticos". Dentro de estos últimos se

[2] En adelante se utilizará el término en inglés de *reweighting*

incluyen métodos que, con un enfoque probabilístico, no utilizan información desagregada de una muestra como los anteriores, aunque pueden usar una tabla agregada de la muestra para obtener los valores de las probabilidades. Los métodos de Reconstrucción Sintética forman parte de estos métodos probabilísticos. Este tipo de métodos no dirigen su estrategia hacia el ajuste de los marginales de la población, aunque también utilizan técnicas para tratar de ajustarlos, centrando su estrategia alrededor de modelos probabilísticos de los datos. Dado que también existen métodos que son combinación de métodos de *reweighting* y probabilísticos, se incluye un tercer grupo con estos métodos híbridos.

Los métodos de *reweighting* son los más utilizados actualmente dada la gran disponibilidad de datos de encuestas existentes hoy en día porque requieren menos pre-procesado de los datos (Barthelemy & Toint, 2013; Lenormand & Deffuant, 2013). Dentro de estos se encuentran los métodos basados en IPF y los de Optimización Combinatoria. Estos dos tipos de métodos son los de uso más frecuente por parte de muchos investigadores (Cho et al., 2014; Harland, Heppenstall, Smith, & Birkin, 2012; Hermes & Poulsen, 2012a; Robin Lovelace, Ballas, & Watson, 2014; Williamson, 2013; Zhu & Ferreira, 2014).

A continuación se describen los métodos más relevantes que se incluyen en cada grupo, especialmente dentro del grupo de *reweighting* que constituye la parte fundamental de esta tesis, ya que en estos métodos se centran los tres estudios experimentales de análisis comparativo de los últimos capítulos.

3.2 Métodos de "Sample Reweighting"

En esta sección se incluyen los métodos que se caracterizan por utilizar marginales objetivos y una muestra de la población como datos de partida para generar la población sintética ajustada a los marginales, mediante réplicas de los elementos de la muestra. Muchos de estos métodos incluyen un proceso de asignación de pesos a los distintos registros de la muestra con objeto de ajustar (*fitting*) la muestra y obtener la distribución ajustada de probabilidades conjuntas, a partir de la cual utilizan otras técnicas para construir la población. En primer lugar se describe el algoritmo del IPF y los métodos que hacen uso de esta técnica de ajuste, incluyendo el *Deterministic Reweighting, que* es la forma de *reweighting* equivalente al IPF.

3.2.1 Iterative Proportional Fitting (IPF) y Deterministic Reweighting (DR)

Los métodos basados en esta técnica permiten generar la población en dos pasos: un primer paso de "ajuste", donde mediante el uso del algoritmo *Iterative Proportional Fitting* (IPF) propuesto por Demings & Stephan (1940) se construye una tabla multidimensional con la distribución deseada de marginales de los atributos, y un segundo paso de "selección de

agentes", también denominado de "conversión a enteros", donde se construye la población a partir de la tabla obtenida en el primer paso.

En esta sección nos centramos en el primer paso de "ajuste", y en la siguiente 0 se tratará el segundo paso. Utilizamos la notación y conceptos introducidos al comienzo del capítulo anterior, para explicar el algoritmo IPF.

A partir de una muestra de la población, muestra S de tamaño *n* con M atributos categóricos $(x_1, x_2, \dots x_{\mathrm{M}})$, se construye la tabla M-dimensional cuyas celdas se representan por $S_{c_1 c_2 \dots c_{\mathrm{M}}}$. El algoritmo IPF permite ajustar los valores de las celdas de esta tabla, preservando los valores de los *odd ratio* condicionales (CORs) de la misma, y obtener unos nuevos valores de celda $E_{c_1 c_2 \dots c_{\mathrm{M}}}$ de modo que la nueva tabla tenga unas distribuciones marginales especificadas (marginales objetivo). Estos marginales pueden ser 1-dimensionales o ser de mayor dimensionalidad (tablas bidimensionales, etc.).

La tabla multidimensional de la muestra inicial no ajustada se denomina "semilla", y a las distribuciones marginales especificadas se las denomina marginales objetivo o restricciones marginales.

Para el caso de una población con 3 atributos x_1, x_2, x_3 siendo los marginales unidimensionales objetivo (*target*) $t_{c_1++}, t_{+c_2+}, t_{++c_3}$ y siendo N el número total de individuos o agentes a generar, se tiene que:

$$\sum_{c_1=1}^{\mathbf{C_1}} t_{c_1++} = \sum_{c_2=1}^{\mathbf{C_2}} t_{+c_2+} = \sum_{c_3=1}^{\mathbf{C_3}} t_{++c_3} = \mathrm{N}$$

El algoritmo comienza inicializando los valores que se buscan con los valores de la semilla: $E^{(0)}_{c_1 c_2 c_3} = S_{c_1 c_2 c_3}$ para todas las posibles combinaciones de categorías $\forall c_1 = 1..\mathbf{C_1};\ c_2 = 1..\mathbf{C_2};\ c_3 = 1..\mathbf{C_3}$.

Posteriormente se repiten los tres siguientes pasos que hacen converger los marginales de la tabla ($E_{c_1++}, E_{+c_2+}, E_{++c_3}$) con los marginales objetivo, preservando la estructura de *odd ratio* condicionales.

Paso 1: $$E^{(1)}_{c_1 c_2 c_3} = E^{(0)}_{c_1 c_2 c_3} \frac{t_{c_1++}}{E^{(0)}_{c_1++}} \qquad \forall c_1 = 1..\mathbf{C_1};\ c_2 = 1..\mathbf{C_2};\ c_3 = 1..\mathbf{C_3}$$

Paso 2: $$E^{(2)}_{c_1 c_2 c_3} = E^{(1)}_{c_1 c_2 c_3} \frac{t_{+c_2+}}{E^{(1)}_{+c_2+}} \qquad \forall c_1 = 1..\mathbf{C_1};\ c_2 = 1..\mathbf{C_2};\ c_3 = 1..\mathbf{C_3}$$

Paso 3: $$E^{(3)}_{c_1 c_2 c_3} = E^{(2)}_{c_1 c_2 c_3} \frac{t_{++c_3}}{E^{(2)}_{++c_3}} \qquad \forall c_1 = 1..\mathbf{C_1};\ c_2 = 1..\mathbf{C_2};\ c_3 = 1..\mathbf{C_3}$$

Estos tres pasos se repiten hasta que la diferencia entre los valores estimados $E_{c_1c_2c_3}^{(it+1)}, E_{c_1c_2c_3}^{(it+2)}, y\ E_{c_1c_2c_3}^{(it+3)}\ \forall c_1, c_2, c_3$ sean menores que la precisión deseada, lo cual conduce a tener un error total absoluto de 0, entre los marginales de la tabla y los objetivos en la iteración it (TAE, del inglés *Total Absolute Error*):

$$TAE^{(it)} = \sum_i |\ E_{c_1++}^{(it)} - t_{c_1++}| + \sum_j |\ E_{+c_2+}^{(\mathrm{it})} - t_{+c_2+}| + \sum_k |\ E_{++c_3}^{(it)} - t_{++c_3}| = 0$$

Por tanto, como condición de finalización del algoritmo, también se usa que $TAE^{(\mathrm{it})}$ sea menor que la precisión deseada.

El algoritmo IPF converge cuando todas las celdas de la tabla inicial "semilla" son mayores que cero $S_{c_1c_2c_3} > 0$ y los marginales objetivo unidimensionales son consistentes (Fienberg, 1970), pero en caso de que haya ceros en las celdas de la semilla inicial, no hay garantía de convergencia. Como las muestras utilizadas para generar poblaciones sintéticas son tablas inherentemente con muchas celdas cero, cuyo número aumenta exponencialmente con el número de atributos y categorías, no es extraño que se produzcan situaciones de no convergencia en algunos casos. Esta posibilidad de no convergencia de IPF es el principal problema de este algoritmo (R Lovelace, Birkin, Ballas, & van Leeuwen, 2015).

Recientemente Pukelsheim (2014) demostró, para el caso de tablas bidimensionales con celdas cero, que el algoritmo IPF converge si la secuencia de $TAE^{(it)}$de las sucesivas iteraciones tiende a cero.

$$\lim_{it\to\infty} TAE^{(it)} = 0$$

En caso de no convergencia del IPF, es posible reemplazar las celdas cero de la tabla con un valor infinitesimal arbitrariamente pequeño, para garantizar la convergencia del algoritmo. Todos los infinitesimales han de tener el mismo valor, con independencia del valor que se utilice. Estos valores infinitesimales permiten que las celdas con cero puedan tomar un valor distinto de cero a lo largo de las iteraciones.

En la Figura 5 se muestra el pseudocódigo de este algoritmo iterativo, que a partir de la tabla multidimensional de la muestra S incorpora un bucle iterativo para cada uno de los marginales impuestos de cada atributo, donde las celdas de la tabla se multiplican por el correspondiente “factor”. Posteriormente se determina el desajuste entre los marginales de la tabla y los marginales impuestos como objetivo, esto es, el Error Total Absoluto unidimensional. Este bucle se repite hasta que dicho desajuste sea inferior a la precisión deseada o se supere el número máximo de iteraciones.

```
#------------------ IPF--------------------------------
#---------------Inputs-------------------------------------
M ← number_of_attributes;
C(1)…C(M) ← Number categories for each attribute;
max_iter ← maximum_number_of_iterations;
precision ← máximum_precision; iter ← 0;
current_table ← sample_table(S_x);
for k =1 to M
        for  m= 1 to C(k)
        marginal_const(k,m) ← target marginal value for categ. m of attr.k
        end for
end for
#----------------Iteration loop -------------------------
Do
    for k =1 to M
        for  m= 1 to C(k)
             factor(k,m)← marginal (k,m,current_table)/marginal_const(k,m)
             current_table(k,m)←current_table(k,m)*factor(k,m)
        endfor
    endfor
    TAE ← 0 ; iter ← iter+1;
    for k=1 to M
        for m=1 to C(k)
           TAE ← TAE + Abs[ marginal(k,m,current_table)-marginal_const(k,m)]
        endfor
    endfor
While (TAE > precision) or (iter<max_iter);
#-----------------Output---------------------------------
return current_table;
```

Figura 5 Pseudocódigo del Algoritmo IPF para generación de poblaciones.

La aplicación de este algoritmo produce una nueva tabla multidimensional que verifica los marginales impuestos, a la vez que preserva la estructura de asociación interna de la muestra en términos de los *odd ratio* condicionales (COR).

La tabla estimada con IPF tiene la propiedad de que hace mínima la entropía relativa entre dicha tabla y la tabla de la muestra, asumiendo que log 0=-∞,log a/0=+∞,0·∞=0 (Csiszar, 1975) (ver definición de entropía en el capítulo 2). Para el caso de una tabla $E_{c_1c_2c_3}$ de 3 dimensiones (con 3 atributos) y una muestra $S_{c_1c_2c_3}$, la entropía relativa viene dada, como se indicó en el capítulo 2, por:

$$Min \sum_{c_1}\sum_{c_2}\sum_{c_3} E_{c_1c_2c_3} \log(E_{c_1c_2c_3}/S_{c_1c_2c_3})$$

El *Deterministic Reweighting* es una forma equivalente de ejecutar el algoritmo IPF. El resultado es el mismo, aunque el algoritmo es del tipo de *reweighting* de la muestra. El algoritmo calcula pesos para los distintos registros de la muestra para construir la población final, de forma que la suma de los pesos será el número de agentes de la población final, que tendrá los marginales ajustados a los objetivos. Por tanto, el resultado del algoritmo son los pesos de cada registro de la muestra, tal como se muestra en el pseudocódigo de la Figura 6.

Ballas, Rossiter, Thomas, Clarke, & Dorling (2005) fueron los primeros que utilizaron esta forma de IPF aplicada a la generación de poblaciones.

```
#------------ Deterministic Reweighting----------
#---------------Inputs---------------------------------
M ← number_of_attributes;
N ← number_of_indivuals_of_sample_table(S_x);
C(1)…C(M) ← Number of categories for each attribute;
max_iter ← maximum_number_of_iterations;
precision ← máximum_precision; iter ← 0;
W(1)…W(N) ← 1 ; weights initialization
sample ← sample_list(S_x);
for k=1 to M
    for each m in C(k)
        marginal_const(k,m) ← target marginal value for cat. m att. k;
    endfor
endfor
#---------------function definition----------------------
delta(j,k,m)← if {( categoria (j,k) = m ) then 1 else 0 }
           ; 1 if the k attribute of j element in the sample is m
#-----------------Iterations Loop------------------------
Do
    for k =1 to M
       for  m= 1 to C(k)
              factor(k,m)←0
              for  j= 1 to N
                 factor(k,m) ← factor(k,m) + W(j)* delta(j,k,m);
              endfor
              for  j= 1 to N
                  W(j) ← W(j)*marginal_const(k,m)/factor(k,m);
              endfor
       endfor
    endfor
    TAE ← 0 ; iter ← iter+1;
    for k=1 to M
       for m=1 to C(k)
              TAE ← TAE + Abs[ factor(k,m)-marginal_const(k,m)]
       endfor
    endfor
While (TAE > precisión) or (iter<max_iter);
#-----------------Output ---------------------------------
return W(1)…W(N);
```

Figura 6 Pseudocódigo del Algoritmo Deterministic Reweighting para generación de poblaciones.

Se trata de un algoritmo muy popular, del que existen múltiples implementaciones en los paquetes de software estadístico (Matlab, Stata, SAS, etc.), incluyendo implementaciones en R, como son las de las librerías cat (2012), ipfp (2016) o mipfp (2018).

Existen algunas variantes de los métodos basados en IPF. Cabe destacar el utilizado por Beckman, Baggerly, & McKay (1996) pioneros de la construcción sintética, que sintetizaron las poblaciones de todas las "*small area*s" incluidas en una región de mayor tamaño aplicando el IPF en dos fases. Observaron que haciendo IPF para cada una de las "*small area*s", usando una muestra de la región y los marginales de cada "*small area*", obtenían una tabla de población para cada "*small area*", pero la tabla resultante de la suma de todas las tablas, la cual corresponde a la región completa, tenía distintos CORs (asociaciones) que la muestra de la región.

Para conseguir que la tabla resultante de la región completa tenga la misma asociación que la muestra, hicieron un primer IPF utilizando los marginales de toda la región (suma de los

marginales de todas las "*small areas*") y usando la muestra como semilla. En una segunda fase, añaden una nueva dimensión a la tabla, que es la dimensión que representa las distintas "*small areas*" que constituyen la región. A partir de los totales de población de cada "*small area*" obtienen los marginales de esta nueva dimensión, y vuelven a hacer un segundo IPF utilizando una semilla que solo contiene valores 1 en todas las celdas e imponiendo en este nuevo IPF la tabla obtenida en la primera fase y los marginales de la nueva dimensión (los totales de cada "*small area*"). Como resultado, obtienen una nueva tabla en la que cada columna representa la población de cada "*small area*".

Las diferencias entre las tablas obtenidas con el IPF en una fase y el IPF en dos fases, son muy pequeñas, tal como ya indicaron Beckman et al. (1996). Dentro del marco de esta tesis, se han realizado experimentos de comprobación mediante generación de tablas aplicando el método IPF Beckman de 2 fases y el método de IPF secuencial para cada "*small area*". Se ha comprobado que los resultados que se obtienen son prácticamente equivalentes y no hay diferencia estadística significativa, es decir, no puede determinarse, con un nivel de confianza del 95%, la tabla que tiene mayor error respecto de la referencia. No se incluyen los detalles de estos experimentos, pero sí una breve descripción de los mismos.

El escenario con el que se han efectuado los experimentos comparativos ha sido una población de referencia distribuida en 60 "*small areas*", descrita con 6 atributos nominales, constituyendo una tabla multidimensional con 2.700 celdas, de la que se han extraído 5 muestras aleatorias del 5% de la población de referencia. Se han calculado los marginales de los atributos en cada una de las 60 "*small areas*". En el capítulo 6 se describe en detalle esta población dentro del caso de Andalucía (ver 6.2.1). Con estos datos se han ejecutado los dos algoritmos, el IPF tradicional y el IPF de Beckman en dos fases. Se ha determinado la similitud entre las 300 poblaciones sintetizadas con cada método (60 *small areas* x 5 muestras) y la población de referencia, utilizando distintas métricas de error para determinar la discrepancia entre tablas (en la sección 4.2 se incluye una relación de este tipo de métricas). Tras hacer una prueba de contraste de hipótesis no se ha encontrado una diferencia estadísticamente significativa entre las métricas (con un nivel de significación $\alpha = 0{,}05$).

Por último, antes de finalizar con los métodos basados en la técnica del IPF, destacamos el método planteado por Arentze, Timmermans, & Hofman (2007) el cual fue diseñado con el objetivo de superar la limitación del IPF tradicional para ajustar los marginales a nivel hogar e individuo simultáneamente. Este método utiliza unas "matrices de relación entre individuos" que permite pasar de distribuciones de individuos a distribuciones de hogares. Cada dimensión de estas matrices representa individuos con distintos valores de los atributos, y los elementos de las matrices representan hogares constituidos por dichos individuos. Con un primer IPF ajusta los grupos de los distintos atributos a nivel individuo, con

los que crean distintos tipos de hogares y con un segundo IPF obtienen la distribución de hogares. El método consigue poblaciones ajustadas a ciertas distribuciones de atributos demográficos de hogares e individuos, pero no ajusta otros atributos exclusivos de hogares como por ejemplo, presencia de hijos, disponibilidad de modos de transportes o nivel socio-económico.

3.2.1.1 Conversión a enteros

Una vez que se ha ejecutado el algoritmo IPF se dispone de una tabla multidimensional con la distribución de marginales impuesta, a continuación se procede con el segundo paso de "selección de agentes".

La técnica convencional que se utiliza para este segundo paso es la utilizada por Beckman et al. (1996) la cual consiste en utilizar la tabla multidimensional obtenida con el IPF como una distribución de probabilidades, con la que hacer un muestreo de Monte Carlo, con o sin repetición, para seleccionar los elementos de la muestra. Mediante el muestreo se genera el número de agentes que se deseen, que se aproximarán a los marginales objetivos impuestos en el IPF. Uno de los problemas surge cuando la población es multinivel, de hogares e individuos (se remite al comienzo del capítulo 2 donde se introdujo el concepto de población multinivel), ya que con esta técnica pueden conseguirse poblaciones de hogares ajustadas al total de hogares y en cierto grado a los marginales objetivos impuestos en el IPF, pero quedan sin ajustar el total de la población de individuos y los marginales de estos.

Entre las primeras variantes desarrolladas para superar esta limitación del IPF, aparte de la ya mencionada de Arentze et al. (2007), que produce buenos resultados con determinados tipos de atributos, se encuentra la ideada por Guo & Bhat (2007). Estos investigadores plantearon hacer IPF para cada nivel de la población de forma independiente. Con la muestra de hogares y los marginales de hogares generan la distribución multidimensional de hogares, y con la tabla de la muestra de individuos (miembros de los hogares de la muestra) y los marginales de individuos obtienen la distribución multidimensional de individuos.

A partir de la distribución de hogares proponen un muestreo de Monte Carlo. Seleccionan hogares de la muestra con los que ir completando las celdas de la tabla de población de hogares, pero haciendo un seguimiento del número de hogares y personas que van entrando en cada celda (con los mismos valores de atributos) y tras cada selección modifican la distribución de probabilidades de selección de los hogares, con objeto de que la probabilidad de selección de un hogar de una celda de la tabla de la muestra disminuya a medida que se eligen hogares de dicha celda para la población sintética. Los hogares seleccionados que supongan superar un determinado porcentaje de los tamaños de marginales objetivos, ya sean de hogares o de individuos, son descartados.

Pritchard & Miller (2012) también propusieron modificar el proceso de selección con un muestreo de Monte Carlo condicional de individuos según los hogares seleccionados.

Zhu & Ferreira (2014) plantearon varias etapas de IPF y selección de Monte Carlo para mejorar la precisión del ajuste. J. Auld & A. Mohammadian (2010) también plantearon la modificación de las probabilidades de selección en función del número de individuos que queden por crear tras cada selección.

Todos estos métodos que parten de las tablas obtenidas con IPF consiguen buenos resultados, pero no llegan a conseguir el ajuste perfecto de todos los marginales de forma sistemática. Para conseguirlo es preciso utilizar otro tipo de estrategias.

Otra estrategia alternativa para la "selección de agentes" es la "conversión a enteros", es decir, el redondeo de la tabla obtenida con IPF. Es por esto, que a este segundo paso de "selección de agentes", también se le denomina de "conversión a enteros".

Estos procesos de redondeo desajustan los valores de los marginales, al igual que el de selección de agentes estocástica, pero recientemente se ha desarrollado una técnica de redondeo que permite construir una población perfectamente ajustada a los valores marginales de los atributos. Este técnica utiliza programación lineal binaria (BLP, del inglés *Binary Linear Programming*) para determinar el valor redondeado de las celdas de la tabla multidimensional, logrando minimizar la diferencia absoluta entre la tabla redondeada y la tabla sin redondear generada por el IPF (Choupani & Mamdoohi, 2015).

Antes de entrar en mayor detalle sobre esta técnica, se enumeran los distintos tipos de redondeo que se utilizan en el campo de la generación de poblaciones:

- Redondeo aritmético: son los redondeos tradicionales, derivados de los utilizados en computación (Santoro, Bewick, & Horowitz, 1989) para el redondeo entero. Existen distintas modalidades: hacia arriba (*round-up*), hacia abajo (*round-down*) y el clásico hacia el entero más cercano (*round-toward-nearest*). Y según se redondee el valor central entre dos enteros consecutivos, se tiene el (*half-up, half-down, half-even, half-odd, half-alternate, half-random*). Este redondeo se ha aplicado considerando el valor central como un valor "umbral", que se fija de forma que el resultado del redondeo aplicado a todo el conjunto de valores produzca una cantidad determinada.
- Redondeo *bucket*: se redondean los números de forma secuencial (con un redondeo aritmético hacia abajo), y el residuo de cada operación de redondeo se conserva para aplicarlo al siguiente redondeo. ("*PopGen software*," 2017). Por ejemplo, la secuencia 1,3; 2,6; 3,2 y 4,9 aplicando este redondeo resulta 1; 2; 4 y 5.

- Redondeo probabilístico: en este tipo de redondeo, la decisión entre redondeo hacia arriba o hacia abajo se realiza en función de una distribución de probabilidades dada. En el caso de que la distribución de probabilidad sea uniforme, se tiene el redondeo aleatorio, el cual evita el sesgo del redondeo de forma muy efectiva cuando se trata de grandes cantidades de números.
- *Truncate-Replicate-Sampling*: redondeo propuesto por R Lovelace & Ballas (2013) para aplicar a los pesos que se obtienen con *Deterministic Reweighting*, donde el redondeo hacia arriba se hace en función de una función de probabilidad determinada por los propios pesos. Estos autores compararon este tipo de redondeo con otros tipos (redondeo hacia arriba *half-up*, redondeos según umbral y otro redondeo probabilístico). Después de analizar distintas métricas de ajuste de marginales de las poblaciones obtenidas, determinaron la superioridad de este método, aunque no consigue el ajuste perfecto de todos los marginales.
- Redondeo con programación lineal binaria (BLP): planteado por Choupani & Mamdoohi (2015). Lo compararon con otros tipos de redondeo (aritméticos y probabilísticos) y tras analizar los ajustes de los marginales de las poblaciones resultantes, verificaron el mejor ajuste de las poblaciones obtenidas con este redondeo, dado que con este redondeo se mantienen los marginales ajustados.

Todo lo expuesto sobre "conversión a enteros", no solo aplica a los métodos que usan el IPF, sino que también puede utilizarse con todos los métodos que asignan pesos no enteros a los representantes de la muestra (métodos de *reweighting*), que se describen en las siguientes secciones.

Antes de pasar a la descripción de estos métodos, se ha incluido una sección para explicar con mayor detalle la técnica BLP aplicada a una tabla ajustada con IPF (población de un nivel). Cuando se describa el primer método de *reweighting* con ajuste simultáneo de marginales de hogares y de individuos, se incluirá otra sección para explicar la adaptación de esta técnica BLP al caso de este tipo de poblaciones multinivel.

3.2.1.2 Redondeo BLP para un nivel

El algoritmo IPF aplicado a una muestra de una población de un solo nivel con M atributos produce una tabla con unos valores de celda $E_{c_1 \dots c_M}$. Esta tabla, al ser el resultado del IPF, verifica las condiciones de los marginales objetivo t_{c_m+}, siendo c_m el índice de categoría para cada atributo *m*, $c_1 = 1 \dots \mathrm{C}_1; \dots; \ c_M = 1 \dots \mathrm{C}_\mathrm{M}$, es decir:

$$E_{c_m+} = \sum_{c_1=1}^{\mathbf{C_1}} \dots \sum_{c_{m-1}=1}^{\mathbf{C_{m-1}}} \sum_{c_{m+1}=1}^{\mathbf{C_{m+1}}} \dots \sum_{c_M=1}^{\mathbf{C_M}} E_{c_1 \dots c_m \dots c_M} = t_{c_m+} \qquad (1)$$

Recordemos que $\mathbf{C_1}, \mathbf{C_2}, \dots \mathbf{C_M}$ representan el número de categorías de cada uno de los M atributos.

Cada valor de una celda de la tabla $E_{c_1 \dots\dots c_M}$tendrá una parte entera $I_{c_1 \dots\dots c_M}$y una parte fraccionaria $f_{c_1 \dots\dots c_M}$.

Llamando $F_{c_1 \dots\dots c_M}$ al valor redondeado de la parte fraccionaria, el cual será 0 o 1, el método de redondeo BLP determina los valores redondeados $F_{c_1 \dots\dots c_M}$ que cumplan:

$$Min \sum_{c_1=1 \dots}^{\mathbf{C_1}} \dots \sum_{c_M=1}^{\mathbf{C_M}} -f_{c_1 \dots.. c_M} * F_{c_1 \dots\dots c_M} \qquad (2)$$

Sujeto a los M conjuntos de condiciones $(m = 1 \dots \mathrm{M})$:

$$\sum_{c_1=1}^{\mathbf{C_1}} \dots \sum_{c_{m-1}=1}^{\mathbf{C_{m-1}}} \sum_{c_{m+1}=1}^{\mathbf{C_{m+1}}} \dots \sum_{c_M=1}^{\mathbf{C_M}} F_{c_1 \dots c_m \dots c_M} = \sum_{c_1=1}^{\mathbf{C_1}} \dots \sum_{c_{m-1}=1}^{\mathbf{C_{m-1}}} \sum_{c_{m+1}=1}^{\mathbf{C_{m+1}}} \dots \sum_{c_M=1}^{\mathbf{C_M}} f_{c_1 \dots c_m \dots c_M}$$

$$\forall c_m = 1 \dots \mathbf{C_m} \qquad (3)$$

Cada conjunto de condiciones tendrá por tanto $\mathbf{C_m}$ ecuaciones, y en total habrá $(\mathbf{C_1} + \mathbf{C_2} + \dots + \mathbf{C_M})$ condiciones que garantizan que el proceso de redondeo no modifica los marginales de la tabla. Por tanto, la solución de redondeo mantendrá el ajuste conseguido por el IPF, ya que este problema de programación lineal binaria siempre tiene solución si los marginales E_{c_m+} son enteros.

3.2.2 Métodos de ajuste complementarios al IPF (ML, MCHI2, LSQ)

Dentro de este grupo se incluyen tres métodos con enfoque de ajuste (*fitting)* de marginales, los cuales son métodos clásicos de estimación de probabilidades de celdas de tablas de contingencia, con valores marginales fijados y una muestra, al igual que el IPF, y que bajo el supuesto de una muestra totalmente aleatoria y representativa de la población, se ha demostrado que son asintóticamente equivalentes al IPF.

Estos métodos han sido objeto de análisis conjunto con el IPF en varias ocasiones por distintos matemáticos. Todos los métodos ajustan los marginales de las tablas, pero a diferencia del IPF, no mantienen la asociación entre los atributos de la tabla de la muestra, en la tabla resultante final.

Se trata de los métodos: máxima verosimilitud bajo supuesto de muestra aleatoria (ML, del inglés *Maximum Likelihood*), minimización del chi-cuadrado χ^2 (MCHI2), y el de mínimos cuadrados (LSQ, del inglés *Least Squares*).

Sin pérdida de generalidad, y siguiendo con el ejemplo de la población de tamaño N descrita por 3 atributos x_1, x_2, x_3 con un número de categorías $\mathbf{C_1}, \mathbf{C_2}, \mathbf{C_3}$ y marginales objetivos $t_{c_1++}, t_{+c_2+}, t_{++c_3}$, estos métodos, al igual que el IPF permiten estimar los valores $E_{c_1c_2c_3}$ de la tabla tridimensional que cumple con las condiciones marginales:

$$\sum_{c_1=1}^{\mathbf{C_1}} E_{c_1++} = t_{c_1++} \;\; ; \;\; \sum_{c_2=1}^{\mathbf{C_2}} E_{+c_2+} = t_{+c_2+} \;\; ; \;\; \sum_{c_3=1}^{\mathbf{C_3}} E_{++c_3} = t_{++c_3}$$

$$c_1 = 1, \dots \mathbf{C_1} \quad ; \quad c_2 = 1, \dots \mathbf{C_2} \quad ; \quad c_3 = 1, \dots \mathbf{C_3}$$

La tabla *E* que se obtiene con IPF, tal como se indicó al explicar este algoritmo, es la que a partir de la tabla de la muestra *S* de tamaño *n*, minimiza la entropía relativa:

$$Min \sum_{c_1}\sum_{c_2}\sum_{c_3} E_{c_1c_2c_3} \log(E_{c_1c_2c_3}/S_{c_1c_2c_3})$$

La tabla de máxima verosimilitud bajo la suposición de una muestra aleatoria es la que maximiza:

$$Max \sum_{c_1}\sum_{c_2}\sum_{c_3} S_{c_1c_2c_3} \log(E_{c_1c_2c_3})$$

La que se obtiene con el método de los mínimos cuadrados, es la que hace:

$$Min \sum_{c_1}\sum_{c_2}\sum_{c_3} \frac{(S_{c_1c_2c_3} - \frac{n}{N} E_{c_1c_2c_3})^2}{S_{c_1c_2c_3}}$$

Y la que se tiene con el método de mínimo chi-cuadrado χ^2:

$$Min \sum_{c_1}\sum_{c_2}\sum_{c_3} \frac{(S_{c_1c_2c_3} - \frac{n}{N} E_{c_1c_2c_3})^2}{E_{c_1c_2c_3}}$$

Existen distintos modos para resolver estas ecuaciones y encontrar los valores $E_{c_1c_2c_3}$ que las satisfacen, como el basado en los multiplicadores de Lagrange. Hoy en día existen librerías del lenguaje estadístico R, como cmm (Bergsma & van der Ark, 2018) que proporciona una función para encontrar la solución ML para los valores de $E_{c_1c_2c_3}$, o la librería mipfp (Barthelemy, Suesse, & Namazi-Rad, 2018) que implementa los cuatro métodos.

Casey (1983) fue uno de los primeros matemáticos que comparó estos métodos, aunque solo comparó tablas 2x2, trató de establecer el método con el que se obtenía la tabla con menor error absoluto esperado y menor error cuadrático medio esperado para la primera celda de

la tabla, bajo el supuesto de un muestreo aleatorio simple. Ordenó los errores obtenidos en cada población y aunque las diferencias entre ellas eran pequeñas, concluyó que el MCHI2 era el mejor método y el LSQ el peor.

Posteriormente, Little & Wu (1991) llevaron a cabo una comparación más rigurosa y sistemática, también para el caso bidimensional, usando múltiples muestras de distintos tamaños y concluyeron que el IPF y el ML producían los mejores resultados.

Recientemente, Suesse, Namazi-Rad, Mokhtarian, & Barthélemy (2017) han publicado un estudio de simulación con tablas multidimensionales, en el que han verificado que los cuatro métodos (IPF, ML, MCHI2 y LSQ) funcionan de manera similar, pero ML es generalmente el mejor en términos de eficiencia y presentando el mejor rendimiento con muestras de menor tamaño, siempre y cuando se trate de muestras representativas con un muestreo aleatorio simple, lo cual no es fácil de encontrar en muchas ocasiones. En su estudio, tratan el problema de convergencia que presentan estos métodos debido a las celdas cero de la tabla de la muestra y, como solución a dicho problema, proponen añadir a todas las celdas de la tabla de la muestra un mismo valor de ajuste próximo a la unidad. Después de analizar los errores obtenidos con estos métodos aplicados a 10.000 muestras de poblaciones de distinto tamaño (tablas tridimensionales de 40 celdas), concluyen que mediante ajustes de 0,5 y 1 en todas las celdas se consigue generar tablas con menos error.

3.2.3 Iterative Proportional Updating (IPU)

Este algoritmo fue introducido por X. Ye, Konduri, Pendyala, Sana, & Waddell (2009), y fue diseñado para poder ajustar todos los marginales de una población multinivel (hogares e individuos), ya que el IPF y los métodos complementarios de la sección anterior no contemplan esta situación.

El algoritmo opera iterativamente sobre una estructura de datos matricial construida con la muestra multinivel de la población real, con el objetivo de encontrar unos pesos para cada hogar que produzcan una distribución conjunta de la población ajustada simultáneamente a los marginales objetivos de hogares e individuos (*target marginals* t_{c_m+}). Dicha estructura matricial tiene una fila por cada hogar k de la muestra y una columna por cada categoría c_m de atributo con marginal objetivo. Para cada categoría de atributo a nivel hogar, la estructura matricial contiene el valor 1 o 0, correspondiendo a si dicho hogar tiene dicha categoría o no, y para cada categoría de atributo de individuo contiene la cantidad de personas del hogar con dicha categoría.

Si x_{km} representa el valor de esta estructura matricial para el hogar de la fila k, y de la columna de la categoría c_m, el proceso iterativo del IPU permite determinar los pesos w_k para

cada hogar k de la muestra *S,* tales que ajusten la muestra a los marginales objetivos especificados $t_{c_{m+}}$ minimizando la expresión:

$$Min \sum_{c_m} \left[\frac{(\sum_{k \in S} w_k x_{kc_m} - t_{c_{m+}})}{t_{c_{m+}}} \right]^2 \text{ siendo los marginales } t_{c_{m+}} \geq 0 \qquad (4)$$

Donde el valor máximo de c_m será el número total de categorías de todos los atributos (de hogar e individuo), es decir, $(\mathbf{C_1} + \mathbf{C_2} + \cdots + \mathbf{C_M})$.

En la Tabla 2 se reproduce dicha estructura para el mismo caso del ejemplo explicado en el artículo de Ye et al. (2009) con una muestra de 23 personas agrupadas en 8 hogares (zona sombreada). Cada hogar se describe con un atributo de 2 categorías y cada persona se describe con otro atributo de 3 categorías. La última columna corresponde a los pesos calculados con este algoritmo, por lo que la suma total de los pesos se corresponde con el número total de hogares que se desea obtener.

Hogar (k)	**Pesos iniciales**	**Hogar tipo 1**	**Hogar tipo 2**	**Persona tipo 1**	**Persona tipo 2**	**Persona tipo 3**	**Pesos Finales (w_k)**
1	1	1	0	1	1	1	1,36
2	1	1	0	1	0	1	25,66
3	1	1	0	2	1	0	7,98
4	1	0	1	1	0	2	27,79
5	1	0	1	0	2	1	18,45
6	1	0	1	1	1	0	8,64
7	1	0	1	2	1	2	1,47
8	1	0	1	1	1	0	8,64
Suma ponderada con pesos iniciales		3	5	9	7	7	
Objetivos ($t_{c_{m+}}$)		35	65	91	65	104	100
Suma ponderada con pesos finales $\sum_k w_k x_{kc_m}$		35	64,99	90,99	64,99	103,99	

Tabla 2 Estructura de datos con la muestra para el algoritmo IPU.

Con algunas muestras, tal como el caso del ejemplo de la tabla anterior, se obtienen pesos que hacen el numerador de (4) sea 0, es decir,

$$\sum_{k \in S} w_k x_{kc_m} = t_{c_{m+}} \qquad \forall\, c_m = 1,2, \ldots (\mathbf{C_1} + \mathbf{C_2} + \cdots + \mathbf{C_M}) \qquad (5)$$

Esta ecuación es la que nos indica que los pesos obtenidos ajustan la muestra a los marginales objetivos. Pero en ciertos casos, dependiendo de la muestra y sobre todo a medida que aumenta el número de valores marginales objetivo, el algoritmo IPU converge a un valor mínimo de la ecuación (4), pero los pesos w_k que se obtienen no consiguen cumplir exactamente la ecuación (5).

Con independencia del cumplimiento de la ecuación (5) los pesos obtenidos pueden redondearse y convertirse en enteros. Como ya se indicó anteriormente, este proceso puede realizarse mediante distintas técnicas, una de las cuales es la del redondeo BLP que para este caso multinivel se explicará en detalle en la siguiente subsección.

Existen distintas implementaciones de código abierto para el algoritmo IPU, por lo que es un algoritmo con bastante popularidad utilizado en muchos trabajos de investigación. Una de las implementaciones más conocidas es la de PopGen en Python, desarrollada por el SimTRAVEL Research Initiative de la Arizona State University, disponible online ("PopGen software," 2017). Recientemente han aparecido implementaciones en el lenguaje R, tales como son los paquetes simPop (Templ, Meindl, Kowarik, & Dupriez, 2017) y MultiLevelIPF (Müller, 2017b).

3.2.3.1 Redondeo BLP para poblaciones multinivel

Al igual que se explicó el redondeo BLP aplicado a tablas de población obtenidas con IPF (sección 3.2.1.2), vamos a explicar la adaptación de esta técnica BLP al caso de tablas de población del tipo multinivel. En aquella ocasión se explicó la idea original de Choupani & Mamdoohi (2015) aplicada a tablas multidimensionales ajustadas a marginales enteros, ahora se trata de una adaptación de dicha idea al caso de los pesos asignados a la muestra, en el que no siempre ajustan exactamente la población a los valores marginales (enteros) de todas las categorías, y mediante la adaptación que se propone se consigue mejorar el ajuste a los marginales. Esta adaptación no ha sido planteada en ningún estudio con anterioridad.

Los métodos de generación de poblaciones multinivel, como el IPU, generan pesos w_k para cada hogar de la muestra *S.* Estos pesos verifican la ecuación de calibración:

$$\sum_{k \in S} w_k x_{k c_m} = t_{c_m +} \qquad \forall c_m = 1 \dots (\mathbf{C_1} + \mathbf{C_2} + \cdots + \mathbf{C_M}) \tag{6}$$

Siendo $x_{k c_m}$ los valores de categorías para cada elemento k de la muestra, que a nivel hogar son valores 0/1 que indican si el hogar tiene dicha categoría, y a nivel individuo contiene la cantidad de personas del hogar con dicha categoría. $t_{c_m +}$ representa los marginales objetivo, marginales de la población a sintetizar, y c_m el índice de categoría para cada atributo *m*. La ecuación (6) es la equivalente a la ecuación (1) para el caso de población multinivel.

Cada peso de la muestra w_k tiene una parte entera $I_k \geq 0$ y una parte fraccionaria f_k. Llamando F_k al valor redondeado de la parte fraccionaria, cuyo valor será 0 o 1, la ecuación (6) con los pesos redondeados puede escribirse como:

$$\sum_{k \in S} (I_k + f_k) x_{kc_m} = \sum_{k \in S} (I_k + F_k) x_{kc_m} = t_{c_m+} \quad \forall c_m = 1 \dots (\mathbf{C_1} + \mathbf{C_2} + \cdots + \mathbf{C_M}) \qquad (7)$$

En este caso, la ecuación a minimizar, equivalente a la (2), es:

$$Min \sum_{k \in S} -f_k * F_k \qquad (8)$$

Sujeto a las $(\mathbf{C_1} + \mathbf{C_2} + \cdots + \mathbf{C_M})$ condiciones:

$$\sum_{k \in S} F_k \, x_{kc_m} = \sum_{k \in S} f_k \, x_{kc_m} \qquad \forall c_m = 1 \dots (\mathbf{C_1} + \mathbf{C_2} + \cdots + \mathbf{C_M}) \qquad (9)$$

En los casos en que los pesos asignados inicialmente no verifiquen exactamente la ecuación (6) para todas las categorías y se tenga para alguna categoría c_m una desviación $|w_k x_{kc_m} - t_{c_m+}| > 0{,}5$ no hay garantía de solución factible al problema de programación lineal binaria.

Para estos casos, se ha comprobado que es más probable encontrar una solución factible de pesos redondeados, si en lugar de imponer las restricciones (7) a los redondeos F_k, se impone para cada categoría c_m el valor marginal correspondiente t_{c_m+}:

$$\sum_{k \in S} F_k \, x_{kc_m} = \sum_{k \in S} |I_k x_{kc_m} - t_{c_m+}| \qquad \forall c_m = 1 \dots (\mathbf{C_1} + \mathbf{C_2} + \cdots + \mathbf{C_M})$$

De este modo el redondeo contribuye a mejorar el ajuste de los marginales, consiguiéndose pesos enteros que ajustan perfectamente la muestra a los marginales objetivos, para los casos en que los pesos (previos al redondeo) no ajustan exactamente.

Las mencionadas comprobaciones se han realizado en 6.000 casos de pesos (poblaciones distintas) que mayoritariamente producían desviaciones de los marginales superiores al 0,5. En la sección 7.3 se detallarán los escenarios para los que se obtuvieron los pesos.

Aplicando esta técnica de redondeo BLP al caso del ejemplo explicado con el IPU de la sección 3.2.3, se obtienen los pesos de la última columna de la Tabla 3. Se observa que dado que hay dos hogares iguales en la muestra, con los mismos tipos de personas, previamente al redondeo han de agruparse los registros de hogares de la muestra que sean iguales, sumando los pesos correspondientes, y son estos nuevos pesos lo que se redondean.

Hogar (k)	Pesos iniciales	Hogar tipo 1	Hogar tipo 2	Persona tipo 1	Persona tipo 2	Persona tipo 3	Pesos Finales (w_k)	Pesos redondeados BLP (w_k)
1	1	1	0	1	1	1	1,36	2
2	1	1	0	1	0	1	25,66	26
3	1	1	0	2	1	0	7,98	7
4	1	0	1	1	0	2	27,79	27
5	1	0	1	0	2	1	18,45	18
6	2	0	1	1	1	0	17,28	18
7	1	0	1	2	1	2	1,47	2
Suma ponderad. pesos iniciales		3	5	9	7	7		
Objetivos (t_{c_m+})		35	65	91	65	104	100	100
Suma ponderad. pesos redond. $\sum_k w_k x_{kc_m}$		35	65	91	65	104		

Tabla 3 Pesos del algoritmo IPU redondeados.

3.2.4 *Generalized Raking* (GR)

Los métodos de *Generalized Raking* son conocidos desde los primeros años de los 90. Fueron desarrollados como técnicas de calibración de encuestas. El Australian Bureau of Statistics (ABS) desarrolló un programa en SAS que implementaba uno de modelos de *Generalized Raking (*GREGWT, del inglés *Generalized Regression Weighting*) que fue utilizado principalmente por investigadores del National Centre for Social and Economic Modelling (NATSEM). GREGWT se utilizó para generar poblaciones sintéticas, y desarrollar microsimulaciones espaciales para llevar a cabo distintas investigaciones, como la distribución espacial del problema de acceso a la vivienda (Taylor, Harding, Lloyd, & Blake, 2004), o la distribución de la pobreza en distintas zonas de Australia (Tanton, 2011). Las referencias de utilización de este método para generación de poblaciones sintéticas se restringen a dicho entorno australiano.

Recientemente se ha puesto a disposición de los investigadores una implementación en R del GREGWT (Muñoz, Vidyattama, & Tanton, 2015), disponible en (Muñoz, 2016). Este método ha sido comparado con *Simulated Annealing* (Tanton, Williamson, & Harding, 2014) (ver sección 3.2.8) y con IPF (Muñoz et al., 2015), aunque IPF solo ajusta un único nivel y este método permite ajustar poblaciones multinivel.

Las técnicas de calibración de encuestas asignan distintos pesos a las respuestas de una encuesta realizada a una muestra de la población, de modo que dicha muestra refleje las mismas proporciones de la población. Es decir, si en la población hay un 30% de hombres y 70% de mujeres, se asignarán pesos a las respuestas para que se tengan dichos porcentajes

en la muestra usada para la encuesta. Por lo tanto, es necesario conocer los marginales de las categorías hombre/mujer para la asignación de dichos pesos.

El IPF (descrito en la Sección 3.2.1), también conocido por el nombre inglés de *raking*, puede utilizarse para ajustar o calibrar los pesos de las respuestas de la encuesta a datos marginales externos, que representan los totales de la población. Este fue el motivo por el que Deville, Sarndal, & Sautory (1993) que desarrollaron unos métodos para estimar pesos de muestras para encuestas, con el fin de que la encuesta refleje unos marginales determinados, los calificaron con el nombre de *Generalized Raking,* ya que el *raking* de Demings & Stephan (1940) puede considerarse como un caso particular de sus métodos. *Generalized Raking* utiliza por tanto información auxiliar en forma de datos marginales de múltiples atributos para estimar los pesos de las respuestas de la encuesta realizada a la muestra.

Deville et al. (1993) desarrollaron diferentes métodos utilizando distintos algoritmos, que se diferenciaban en la función a optimizar. Para entender estas funciones introducimos algunos conceptos y nomenclatura relativa a la calibración de encuestas.

Con las encuestas a muestras es frecuente estimar un parámetro o característica descriptiva Y de una población U de tamaño N. Si y_k es el valor de la característica de interés y de la unidad de población k, se estima el parámetro poblacional $Y = \sum_{k \in U} y_k$ utilizando una muestra S de la población U, con tamaño *n*, obtenida con un método donde se conoce la probabilidad de selección de cada elemento de la población, siendo la "probabilidad de inclusión" π_k la probabilidad de que un individuo k esté en la muestra de tamaño n, $\pi_k = P(k \epsilon S)$.

Si se trata de un muestreo aleatorio simple, cada elemento de la población tiene la misma probabilidad de selección. El estimador Horvitz-Thompson (HT) del parámetro poblacional *Y* se construye con los valores inversos de las probabilidades de selección.

$$\hat{Y}_{HT} = \sum_{k \in S} \frac{y_k}{\pi_k} = \sum_{k \in S} y_k d_k$$

Donde d_k son los inversos de las probabilidades de selección, también llamados "pesos de diseño". Esta suma ponderada de las características y_k de los elementos de la muestra es un estimador insesgado del total poblacional, es decir, la diferencia entre el valor esperado (esperanza) del estimador y el verdadero valor del parámetro que se estima es nula.

Si mediante fuentes externas distintas a la encuesta se conocen otros parámetros $\boldsymbol{t}_{\boldsymbol{c_m}+}$ de la población U, correspondientes a los atributos $X = \{x_1, x_2, \ldots x_{\mathrm{M}}\}$, con categorías $\mathbf{C_1}, \mathbf{C_2}, \ldots \mathbf{C_M}$, estos parámetros serán:

$$t_{c_m+} = \sum_{k \in U} x_{kc_m} \qquad \forall c_m = 1 \ldots (\mathbf{C_1} + \mathbf{C_2} + \cdots + \mathbf{C_M})$$

Y para cada elemento de la muestra se conocen los valores de todos los atributos $(y_k, \boldsymbol{X_k})$, donde $\boldsymbol{X_k} = \{x_{1_k}, x_{2_k}, \ldots x_{M_k}\}$, entonces puede crearse un conjunto de pesos w_k tales que la suma ponderada (con estos w_k) de los valores $\boldsymbol{X_k}$ de la muestra sea exactamente igual a los parámetros poblacionales totales t_{c_m+} $(\sum_{k \in S} w_k x_{kc_m} = \sum_{k \in U} x_{kc_m})$, imponiendo que dichos pesos estén lo más cerca posible de los "pesos de diseño" d_k. Esta condición se conoce como ecuación de calibración.

$$\sum_{k \in S} w_k x_{kc_m} = t_{c_m+} \qquad \forall c_m \tag{10}$$

La cercanía se define en términos de una función distancia $\boldsymbol{G}(w_k/d_k)$ entre los pesos w_k y los pesos de diseño d_k.

Estos nuevos pesos se determinan encontrando los valores w_k que minimizan la distancia total para todos los pesos de la muestra:

$$Min \sum_{k \in S} d_k \boldsymbol{G}(w_k/d_k) \tag{11}$$

Sujeta a la ecuación de calibración $\sum_{k \in S} w_k x_{kc_m} = \sum_U x_{kc_m} \qquad \forall c_m$

Deville et al. (1993) estudiaron distintos métodos alternativos para el *raking* con distintos algoritmos. La diferencia entre los métodos estaba en la función distancia a optimizar $\boldsymbol{G}(w_k/d_k)$:

- Método lineal. Utiliza la función distancia $G(z) = (1/2)\,(z-1)^2$ siendo $z = w_k/d_k$.
- Método lineal truncado (Inf, Sup). Es el método lineal en el que se especifican los valores inferior y superior que pueden tomar los pesos. De este modo se evita que puedan aparecer valores de pesos no deseados.
- Método multiplicativo (o método *raking ratio*). En este caso la función distancia es $G(z) = z \log z - z + 1$ la cual produce el mismo resultado que el IPF de Deming y Stefan o el método de Optimización de la Entropía de Bar-Gera (ver sección 3.2.6).
- Método logit (Inf, Sup). Similar al anterior, pero estableciendo el valor inferior (Inf) y superior (Sup) con lo que se tienen pesos con un menor número de valores extremos.

Para cada función resolvieron el problema de optimización con el método de los multiplicadores de Lagrange.

Como en todos los métodos tipo *reweighting* de la muestra, a partir de los pesos obtenidos con cualquiera de los métodos, se procede al proceso de conversión a enteros.

El paquete de R, MultiLevelIPF, implementa los algoritmos indicados.

Durante mucho tiempo se ha ignorado en la literatura la relación entre GREGWT y *Generalized Raking* lineal, cuando en realidad son el mismo método, ya que ambos minimizan la misma función: $Min \sum_{k \in S} \frac{(w_k - d_k)^2}{d_k}$. Dentro del marco de esta tesis, se ha comprobado la igualdad de resultados de ambos métodos, utilizado una muestra con 1.000 hogares descritos con 1 atributo de hogar (tamaño de hogar) y 2 atributos para sus individuos (edad y estado civil). Los pesos obtenidos con la implementación del *Generalized Raking* lineal del paquete MultiLevelIPF son iguales a los obtenidos con la implementación de GREGWT de (Muñoz et al., 2015), utilizando la misma muestra y los mismos marginales objetivo.

3.2.5 IPF Jerárquico (Hierarchical IPF - HIPF)

Al igual que el IPU, este algoritmo de *reweighting* de la muestra fue diseñado específicamente para crear poblaciones multinivel ajustadas a los marginales de cada nivel. El método establece pesos para los elementos de la muestra (similares al método anterior) llamados factores de expansión de los hogares f_h, pero además plantea otros pesos para los individuos de los hogares, llamados factores de expansión de las personas f_{h_p}.

A partir de estos factores de expansión, que inicialmente son 1, mediante un proceso iterativo genera unos nuevos factores de expansión f_h' que han de verificar:

$$\sum_h f_h' = \mathrm{N} \tag{12}$$

$$\sum_{h \in S} p_h \cdot f_h' = \nu \tag{13}$$

Siendo p_h el número de individuos que constituyen el hogar h, N el número de hogares objetivo que se desea generar, ν el número de personas objetivo a generar y h un hogar de la muestra S. No entramos en el detalle de los procedimientos y solo se describe el proceso en su conjunto (Müller & Axhausen, 2011).

Si las categorías de los atributos de hogar son $a, b, c, \ldots$ etc, se define la suma de factores de expansión correspondientes a H_a (hogares con un valor determinado de un atributo *a)*, como $F_a = \sum_{h \in H_a} f_h$; y para los hogares con el valor *b* como $F_b = \sum_{h \in H_b} f_h$, etc.

E igualmente, denominando con letras griegas a las categorías de los atributos de personas $\alpha, \beta, \gamma, \ldots etc$, se define la suma de factores de expansión de las personas con un valor determinado de un atributo α (P_α),

$$F_\alpha = \sum_{h \in P_\alpha} f_{h_p}$$

Una vez se establecen las restricciones de los valores de los marginales de los atributos de los hogares $t_a, t_b, t_c, \ldots$ y los de atributos de personas $t_\alpha, t_\beta, t_\gamma, \ldots$ se ejecutan los procedimientos iterativos de ajuste de estas sumas de factores de expansión a los marginales de hogares, es decir, $F_x = t_x$ para todas las categorías x de los atributos de hogares, y $F_\xi = t_\xi$ para todas las categorías ξ de los atributos de personas.

Se continua alternando entre factores de expansión a nivel hogar y a nivel persona ($f_h = \frac{1}{p_h}\sum_{p \in P_h} f_p$) buscando nuevos f_h' a partir de los f_h que minimicen la expresión:

$$Min \sum_{h \in S} f_h \ln\left(\frac{f_h'}{f_h}\right)$$

Sujetos a las condiciones (12) y (13). Los procedimientos iterativos finalizan cuando las sucesivas iteraciones convergen y no se modifica el valor de los factores de expansión.

Como en todos los métodos tipo *reweighting*, a partir de los pesos obtenidos se procede al proceso de generación de la población a partir de la distribución de pesos.

Los autores de este método, Müller & Axhausen (2011), generaron poblaciones sintéticas multinivel para los cantones suizos aplicando muestreo de Monte Carlo, y verificaron que la precisión de las poblaciones generadas era superior a la de otros métodos, como el IPU.

3.2.6 Optimización de la Entropía (Entropy Optimization - EO)

Fue planteado por Bar-Gera, Konduri, Sana, Ye, & Pendyala (2009), y como los anteriores, es del tipo *reweighting* de la muestra. El método comienza a partir de los pesos iniciales de la muestra multinivel w_k^0. Estos pesos se obtienen agrupando los elementos iguales de la muestra (hogares con los mismos valores de atributos de hogar y de personas), tal como aparece en la segunda columna de la Tabla 3 para el caso de la muestra de 8 hogares del ejemplo del IPU de la Tabla 2.

Este método utiliza la misma estructura matricial de la muestra que el IPU. El objetivo del algoritmo es encontrar unos nuevos pesos para cada hogar de la muestra que produzcan una distribución de marginales de la población ajustados a los marginales objetivos de hogares e individuos ($t_{c_{m+}}$).

Estos pesos, o ponderaciones de la muestra w_k, se calculan de forma que minimicen la función:

$$Min \sum_{k \in S} w_k \left[\log \left(\frac{w_k}{w_k^0} \right) - 1 \right] \qquad (14)$$

Sujeto a las restricciones: $\sum_{k \in S} w_k x_{k c_m} = t_{c_m +} \qquad \forall c_m$

Los pesos w_k que hacen mínima la función (14) estarán próximos a los pesos iniciales w_k^0.

La función (14) equivale a minimizar la entropía relativa entre la tabla de la población y de la muestra (ver definición de entropía y entropía relativa en el capítulo anterior).

Para el caso de muestras aleatorias simples, en las que todos los elementos de la población tienen la misma probabilidad de pertenecer a la muestra, la función a minimizar (14) es la misma que la del *Generalized Raking* multiplicativo, por lo que ambos métodos producen los mismos resultados. Existen distintos métodos numéricos para resolver el problema de minimización, algunos iterativos como el planteado por los propios autores.

Como en todos los métodos del tipo *reweighting*, a partir de los pesos obtenidos se procede al proceso de conversión a enteros con alguna de las técnicas anteriormente indicadas.

Existe una variante de este método que se utiliza con poblaciones de un nivel, en la que, asumiendo pesos iniciales 1, se determinan los pesos que minimizan $Min \sum_{k \in S} w_k \log(w_k) + \sum_{c_m} error_{c_m}$ (sujetos a las restricciones de marginales), donde se tiene en cuenta el posible error o incertidumbre de los valores de los marginales $t_{c_m +}$ mediante la inclusión de un factor de penalización $error_{c_m}$ en la ecuación de optimización de la entropía (Nagle, Buttenfield, Leyk, & Spielman, 2013).

Este método se ha utilizado recientemente para estimar los índices de mortalidad, entre los años 2000 y 2003, de más de los 500 distritos censales (entre 1.300 y 8.000 individuos) de 6 de los 64 condados del estado de Colorado (Estados Unidos). Esta estimación se ha realizado generando la población de los distritos censales a partir de los microdatos individuales disponibles únicamente a nivel condado, datos de mortalidad de los años 2000 a 2003 proporcionados por el National Center for Health Statistics para cada condado y de los marginales de los distritos censales, incluyendo la diferencia de población de cada distrito censal entre el año 2000 y 2003 (Ruther, Leyk, & Buttenfield, 2017).

3.2.7 Método Heurístico Pop-H

Es un algoritmo desarrollado por Zhuge, Li, Ku, Gao, & Zhang (2017) que también parte de los pesos iniciales de la muestra multinivel w_k^0. Estos pesos se van modificando gradualmente en pequeñas cantidades de forma iterativa, tratando de ajustar la muestra ponderada a los marginales objetivos, pero teniendo en cuenta la desviación típica de los desajustes de las

categorías. No es lo mismo tener una discrepancia absoluta relativa $|t_{c_{m+}} - \sum_k w_k x_{kc_m}|/t_{c_{m+}}$ de 0,01 en 9 categorías, y en la décima tener una discrepancia de 0,1, que tener 5 categorías con discrepancias relativa de 0,025 y otras 5 con valores de 0,015. En el primer caso la discrepancia total es de 0,19 y la desviación de estas 0,027, ya que la última categoría tiene 10 veces más discrepancia que las restantes. Por lo que es preferible tener una discrepancia absoluta relativa media mayor, de 0,02 (5 categorías con discrepancias relativa de 0,025 y otras 5 con valores de 0,015) con una desviación de 0,005, sobre todo si la aplicación que usa la población sintética utiliza el atributo con el mayor error como variable determinante para sus resultados.

El método utiliza la misma estructura matricial de la muestra descrita con el IPU. Dados unos valores de pesos, la discrepancia con cada marginal objetivo de cada categoría t_{c_i} se mide mediante el Error Absoluto Relativo μ_{c_j} que para cada categoría se define como:

$$\mu_{c_m} = \frac{\left|t_{c_{m+}} - \sum_k w_k x_{kc_m}\right|}{t_{c_{m+}}}$$

Este método utiliza la función objetivo:

$$f(\mu, \sigma) = (1 - \lambda)R(\mu) + \lambda R(\sigma) \qquad (15)$$

Siendo λ el factor objetivo ($0 < \lambda < 1$), μ la media de todos los desajustes (MAPE de la sección 4.2) y σ la desviación típica:

$$\mu = \frac{1}{(\mathbf{C_1} + \cdots + \mathbf{C_M})} \sum_{c_m=1}^{\mathbf{C_1}} \cdots \sum_{c_m=\mathrm{C}_1+..+\mathrm{C}_{\mathrm{M}-1}+1}^{\mathbf{C_1}+..+\mathbf{C_{M-1}}+\mathbf{C_M}} \mu_{c_m} \ ;$$

$$\sigma = \sqrt{\frac{1}{(\mathbf{C_1} + \cdots + \mathbf{C_M})} \sum_{c_m=1}^{\mathbf{C_1}} \sum_{c_m=\mathbf{C_1}+1}^{\mathbf{C_1}+\mathbf{C_2}} \cdots \sum_{c_m=\mathbf{C_1}+..+\mathbf{C_{M-1}}+1}^{\mathbf{C_1}+\cdots+\mathbf{C_{M-1}}+\mathbf{C_M}} (\mu_{c_m} - \mu)^2}$$

La función $R(x)$ clasifica en modo ascendente un conjunto de valores y devuelve como resultado el número de orden que corresponda en la clasificación. Se utiliza para ordenar los valores de μ y los de σ. Por ejemplo, si hay 10 valores de μ, el resultado del μ mayor será $R(\mu) = 10$.

Las iteraciones del algoritmo seleccionan los ajustes de los pesos a partir de los que minimicen la función objetivo (15).

A diferencia del IPU, que solo tiene en cuenta minimizar el error absoluto relativo,

$$Min \sum_{c_m} \frac{|\sum_k w_k x_{kc_m} - t_{c_m+}|}{t_{c_m+}}$$

Este algoritmo incorpora en el proceso de minimización la desviación típica de los errores absolutos de los marginales objetivos, según el valor que se establezca para el factor objetivo λ.

3.2.8 Optimización Combinatoria: Simulated Annealing (CO/SA)

Otro de los enfoques basados en muestra más referenciado por los investigadores, es el que considera el problema de generación de la población sintética como un problema de optimización combinatoria, donde el espacio de soluciones está formado por las tablas multidimensionales cuyas distribuciones marginales son las de la población que se quiere crear, imponiendo que solo contengan agentes contenidos en la muestra. En este caso, mediante un enfoque de búsqueda metaheurística tal como el de escala simple (*Hill Climbing*), o de *Simulated Annealing*[3] se trata de hallar una de las soluciones viables que hace mínima la función de bondad de ajuste (GoF) que se establezca.

El método de *Simulated Annealing*, basado en el intercambio aleatorio de agentes, es uno de los métodos que mejores resultados produce. Para llevar a cabo dicho método, en primer lugar, se establecen los valores de los parámetros iniciales, como son, el número máximo de iteraciones y la temperatura. Este último parámetro controla la probabilidad de aceptar un intercambio si no se mejora la solución y su valor irá disminuyendo gradualmente a lo largo del tiempo.

En la Figura 7 se muestra el pseudocódigo de este algoritmo de optimización iterativo. El algoritmo incorpora dos bucles, uno externo que regula el descenso de temperatura, y otro interno donde se exploran caminos de poblaciones alternativas con un límite máximo de iteraciones y donde es posible seleccionar alternativas que empeoren el ajuste (GoF, del inglés *Goodness-of-Fit*) en función del grado de empeoramiento y de la temperatura.

El algoritmo comienza con la construcción de una población inicial, seleccionando agentes de la muestra de forma aleatoria, y se evalúa el ajuste de los marginales de dicha población con una función bondad de ajuste GoF (habitualmente la suma de los valores absolutos de las diferencias de marginales, TAE). Posteriormente se selecciona un nuevo agente de la muestra y se intercambia de forma aleatoria por otro de la población inicial.

[3] El término en español es “recocido simulado”, aunque en este documento se ha preferido mantener el nombre en inglés.

```
#------------------Simulated Annealing------------------
#------------------Inputs-------------------------------
M ← number_of_attributes;
N ← number_of_indivuals_of_sample_table(Sx);
C(1)…C(M) ← Number of categories for each attribute;
sample<- sample_table(Sx)
for k =1 to M
       for  m= 1 to C(k)
       marginal_const(k,m) ← target marginal value for categ. m of attr.k
       end for
end for
Temp ← Initial_temperature;
max_iter ← maximum_number_of_iterations;
Decrement ← temperature_reduction;
current_table ← initial_random_population_table(sample);
#---------------function GoF definition-----------------
GoF(table) ← calculate TAE adding errors for each marginal_const()
#--------------------Loop-------------------------------
Do
       iter ← 0;
       Do
              next_table ← random_swapping(current_table);
              Δ GoF ← GoF(next_table)-GoF(current_table);
              If (Δ GoF < 0) then
                     current_table←next_table;
              else
                     q ← Min{1,exp(- ΔE/Temp};
                     If random(0,1)<q then current_table←next_table;
              endif;
              iter ← iter+1;
       While (iter < max_iter);
       Temp ← Temp*decrement;
While Temp>0,0001;
#-----------------Ouput-------------------------------
return current_table;
```

Figura 7 Pseudocódigo del Algoritmo Simulated Annealing.

La población con intercambio aleatorio (*random swapping*) se convierte en la siguiente "población" si supone un GoF menor (se está minimizando el ajuste de los marginales y, por tanto, es mejor que su antecesor), pero con una cierta probabilidad también puede ser la siguiente aunque suponga un incremento del GoF (es peor que su antecesor). La probabilidad de que esto último suceda disminuye exponencialmente con el valor de empeoramiento relativo, y a medida que disminuye la temperatura.

Si la muestra es multinivel (hogares con los miembros que los constituyen) este método permite encontrar ajustes de ambos niveles, buscando ajustar simultáneamente los marginales de hogares e individuos, para lo cual la función bondad de ajuste GoF ha de tener en cuenta los marginales de ambos niveles. Es habitual utilizar como GoF la suma de errores absolutos de todos los marginales.

La primera aplicación del *Simulated Annealing* a la generación de poblaciones sintéticas se debe a Williamson, Birkin, & Rees (1998), que utilizaron como función de bondad de ajuste la suma de Z-score cuadrado modificada (ver sección 4.2). Actualmente existen otras implementaciones disponibles en R como el paquete simpop (Templ et al., 2017). Este paquete incluye la función calibPop que asigna pesos enteros a los registros de la muestra

para encontrar la combinación óptima que hace mínima la suma de los valores absolutos de las diferencias de marginales (TAE).

Dado que la población final está constituida por los elementos de la muestra, cada uno en diferentes cantidades o pesos, estos métodos se encuadran dentro del tipo *reweighting* de optimización combinatoria, tal como fueron clasificados por Hermes & Poulsen (2012).

3.2.9 Síntesis Mediante Ajuste (Fitness-Based Synthesis - FBS)

Al igual que el método anterior de *Simulated Annealing*, este método se encuadra dentro del tipo *reweighting,* en este caso heurístico, pues la población final se construye mediante clonación de los elementos de la muestra, cada uno en diferentes cantidades, determinadas mediante reglas simples.

Este método fue desarrollado por L. Ma (2011), y al igual que el método anterior, se trata de un método que genera una población multinivel directamente, sin tener que construir la distribución conjunta multidimensional de los atributos.

El método puede iniciarse con una población aleatoria de hogares de la muestra, como en el caso anterior, pero también puede comenzar con una población vacía sin hogares e ir añadiéndolos progresivamente.

Es un método iterativo, que en cada iteración agrega o elimina un hogar de la muestra a/de la población sintética (hogar con todos sus miembros), para lo cual, para cada hogar de la muestra define dos valores de ajuste, el primero es el error cuadrático de los marginales en el caso de que se añada el hogar a la población sintética, y el segundo el error cuadrático de los marginales de la población en el caso de que se elimine el hogar de la población.

Si en la iteración $i-1$ se tiene una tabla de población cuyos valores de celda son $E^{i-1}_{c_1 \dots c_M}$, y cuyos valores marginales son $E^{i-1}_{c_m+} = \sum_{c_1=1}^{\mathbf{C_1}} \dots \sum_{c_{m-1}=1}^{\mathbf{C_{m-1}}} \sum_{c_{m+1}=1}^{\mathbf{C_{m+1}}} \dots \sum_{c_M=1}^{\mathbf{C_M}} E^{i-1}_{c_1 \dots c_m \dots c_M}$ siendo c_m el índice de categoría para cada atributo *m*, $c_1 = 1 \dots \mathbf{C_1}; \dots; c_M = 1 \dots \mathbf{C_M}$, los 2 valores de ajuste mencionados son:

$$F_{\mathrm{I}}^{ki} = \sum_{c_1=1 \dots}^{\mathbf{C_1}} \dots \sum_{c_M=1}^{\mathbf{C_M}} \left[\left(R^{i-1}_{c_m+} \right)^2 - \left(R^{i-1}_{c_m+} - HT^{k}_{c_m+} \right)^2 \right]$$

$$F_{\mathrm{II}}^{ki} = \sum_{c_1=1 \dots}^{\mathbf{C_1}} \dots \sum_{c_M=1}^{\mathbf{C_M}} \left[\left(R^{i-1}_{c_m+} \right)^2 - \left(R^{i-1}_{c_m+} + HT^{k}_{c_m+} \right)^2 \right]$$

Siendo t_{c_m+} los valores marginales objetivos, y:

$R_{c_{m+}}^{i-1} = t_{c_{m+}} - E_{c_{m+}}^{i-1}$ número de hogares / personas que faltan para alcanzar el marginal objetivo $O_{c_{m+}}$ después de la iteración *i-1*.

$HT_{c_{m+}}^{k}$ contribución del *k*-ésimo hogar de la muestra al marginal c_m.

$R_{c_{m+}}^{i-1} - HT_{c_{m+}}^{k}$número de hogares que se precisan para alcanzar el objetivo $t_{c_{m+}}$ si se añade el hogar k, después de la iteración *i-1*.

$R_{c_{m+}}^{i-1} + HT_{c_{m+}}^{k}$número de hogares que se precisan para alcanzar el objetivo $t_{c_{m+}}$ si se elimina el hogar k, después de la iteración *i-1*.

Por tanto, en cada iteración se calculan los dos valores de ajuste para cada hogar de la muestra. Los hogares candidatos a ser añadidos tendrán el valor de ajuste I positivo y los hogares de la muestra que ya estén en la población sintética que tengan el valor de ajuste II positivo serán candidatos a ser eliminados de la población sintética. Entre todos los candidatos, se selecciona uno al azar y tras añadirlo o retirarlo se vuelven a actualizar todos los valores de ajuste y se ejecuta una nueva iteración. El proceso finaliza cuando ningún hogar de la muestra tenga valores de ajuste I o II positivos, habitualmente tras un número de iteraciones aproximadamente equivalente al doble del tamaño de la muestra.

Se han desarrollado variantes de este método, tal como la implementada en Matlab por Hafezi & Habib (2014) usando técnicas de matriz dispersa, matriz cuyos elementos mayoritariamente son cero, técnicas especiales para sacar ventaja del gran número de elementos ceros, con la que se ha construido la población sintética, con 9 atributos (5 de individuo y 4 de hogar), de las 4 provincias atlánticas de Canadá.

3.3 Métodos Probabilísticos

En esta sección se describen los métodos que utilizan modelos probabilísticos para generar una distribución de probabilidades con la que generar la población. Estos métodos utilizan valores de probabilidades, que en algunos casos pueden obtenerse de tablas de muestras parciales (con parte de los atributos), y en otros, se obtienen de información de terceros. Si las probabilidades no se obtienen de una muestra, se conocen como métodos *Sample-Free* (SF). Los primeros métodos de este tipo son los de “Reconstrucción Sintética”, que utilizan principalmente datos agregados de las agencias estadísticas y muestras parciales. Aunque este enfoque tiene cierta antigüedad, se han desarrollado algunas variantes en los últimos años que con un mínimo de datos de partida obtienen relativamente buenos resultados, con alto grado de ajuste a los marginales de hogares e individuos, tal como indican Lenormand & Deffuant,(2013).

3.3.1 Reconstrucción Sintética (Synthetic Reconstruction - SR)

Wilson & Pownal (1976) fueron los primeros que propusieron este método para generar poblaciones sintéticas a partir de datos agregados proporcionados por las agencias estadísticas, siendo hasta hace pocos años el método más utilizado cuando solo se disponía de datos agregados del censo. A partir de las tablas del censo calcula tablas de probabilidad condicionada de los atributos, con las que genera los agentes mediante muestreo de Monte Carlo (Huang & Williamson, 2001). Este método trata de reconstruir la población original, de forma que se repliquen las tablas del censo.

El orden de uso de las probabilidades condicionadas se construye según las dependencias entre los atributos y las diferentes escalas de cada uno de los datos (datos a nivel distritos, municipios, provincias, etc.). Así, por ejemplo, el atributo sexo y edad suele ser independiente del resto, pero el atributo estado civil estará en función de los anteriores. Por tanto, a la hora de generar individuos de un municipio, en primer lugar se generarán hombres y mujeres con determinada edad, tantos como aparezcan en las tablas del censo de sexo por edad, y una vez se tienen estos individuos, se asigna el estado civil, según las probabilidades condicionadas de los distintos valores de estado civil dada la edad y sexo del individuo. Así mismo, a la hora de generar individuos de los distritos del municipio, primero se generan los del municipio, y según las probabilidades de pertenencia a cada distrito se les asigna el distrito.

En la Figura 8 se ilustra el proceso simplificado de reconstrucción sintética. En este ejemplo, se parte de la tabla del censo correspondiente a edad por sexo de un municipio, y se crean tantos individuos con una determinada edad y sexo como aparezcan en la tabla del censo de dicho municipio. A partir de una tabla provincial de edad *x* sexo *x* situación civil, se determinan las probabilidades condicionadas de los distintos valores de estado civil dado la edad y el sexo del individuo. Una vez se transforman en probabilidades acumuladas, se hace un muestreo de Monte Carlo, generando números aleatorios para ir asignando la situación civil a cada uno de los individuos creados.

El proceso continúa con la asignación de otros atributos, como situación laboral, esposa/o e hijos según su estado civil, etc., y a partir de las tablas de población de los distritos de cada municipio, se asigna el distrito a los individuos de cada municipio, etc.; todo a partir de las tablas del censo.

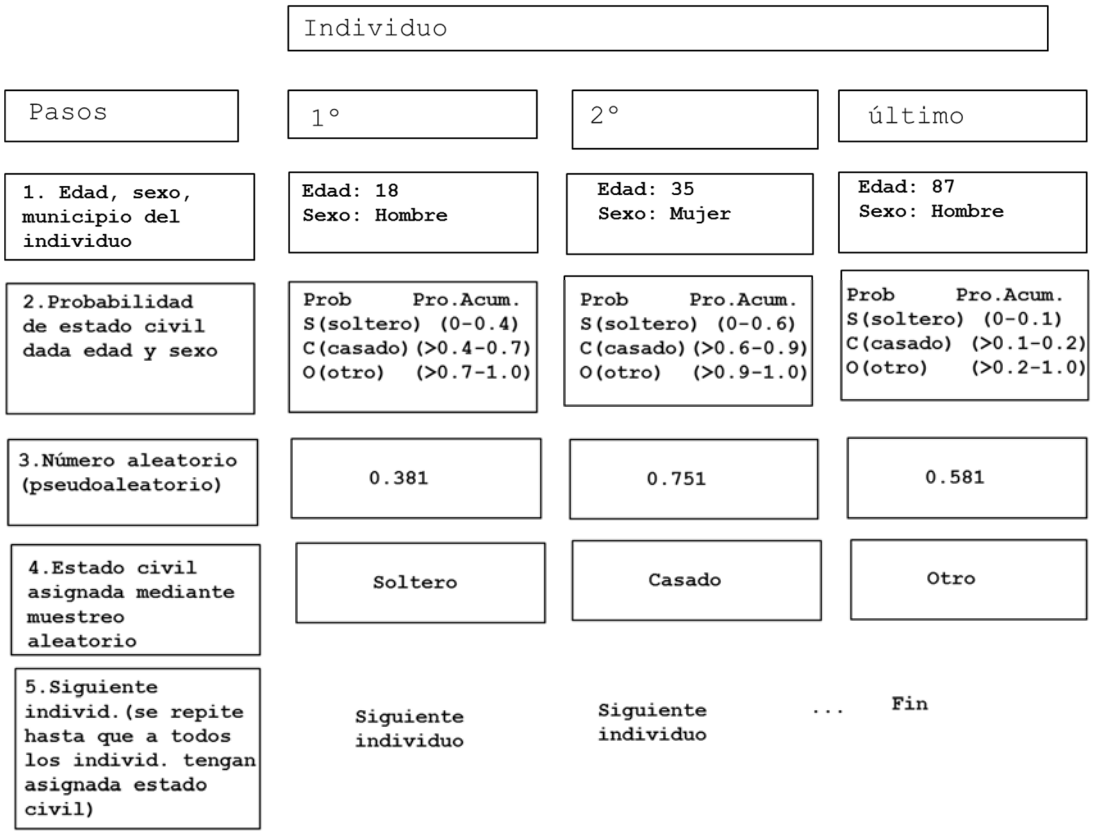

Figura 8 Procedimiento simplificado de reconstrucción sintética. Adaptado de (Williamson, 2013).

Además de utilizar los datos de las agencias estadísticas, también se apoyan en la "Fusión de Datos" de distintas fuentes (*Data Fusion*) para construir tablas de probabilidad condicionada. También utilizan muestras parciales (con un número reducido de atributos), que junto con valores marginales y mediante la técnica del IPF (descrita en sección 3.2.1), se transforman en tablas de distribución de probabilidad para generar agentes o asignar valores de atributos a agentes previamente generados.

Existen variantes de la "Reconstrucción Sintética" que, en lugar de utilizar la técnica de generación de números aleatorios para la asignación de los atributos de los agentes, utilizan otras técnicas (Murata, Harada, & Masui, 2017), a la vez que utilizan algoritmos específicos para mantener la población ajustada a los marginales objetivo.

Gargiulo, Ternes, Huet, & Deffuant (2010) aplicaron este método para construir una población multinivel de hogares e individuos. Utilizaron como datos de entrada distribuciones marginales de atributos de hogares (tamaños y tipos) y distribuciones marginales de atributos de individuos (edad, edad del cabeza de familia, diferencias de edad entre cónyuges, edad de los hijos respecto a la de la madre).

A partir de estas distribuciones crean una lista de individuos ajustada a la distribución de edades. Los hogares se seleccionan conforme a los datos de distribuciones de probabilidad

disponibles (tipos de hogar) y se van incorporando los miembros del hogar de forma secuencial, generándolos conforme a su distribución de probabilidad, por ejemplo, cada hogar tendrá un cabeza de familia con una edad cuya distribución de edades es dato (distribución de edad del cabeza de familia en función del tipo de hogar). En función del tipo de hogar, se genera una pareja de acuerdo a la distribución de probabilidades de las diferencias de edades entre las parejas y unos hijos conforme a la distribución de la diferencia de edades con la madre. Antes de incorporar un individuo al hogar, se comprueba que existe disponibilidad del mismo con dicha edad en la lista de individuos, y si no existe, se desecha el hogar y se vuelve a seleccionar un nuevo hogar. Una vez se han completado todos los individuos del hogar, se procede a eliminar dichos individuos de la lista de individuos.

El algoritmo finaliza cuando ya no pueden construirse más hogares con los individuos que quedan disponibles. Considerando que los restantes hogares son hogares atípicos y no responden a la estructura de habitual, siendo, por ejemplo, un grupo de personas que simplemente comparten piso. Al no tener información sobre estas estructuras de hogar, se asignan aleatoriamente las personas restantes a los hogares que quedan por generar, hasta completar los marginales objetivos. Con esta estrategia consiguen ajustar los marginales de individuos y hogares.

En el artículo de Xu et al. (2017) se utiliza este método para generar una población sintética en la Samoa Estadounidense con objeto de modelar una población para estudiar la transmisión de una enfermedad infecciosa en dicha población.

3.3.2 Reconstrucción Sintética con Optimización

Barthelemy & Toint (2013) también se plantearon generar una población multinivel, con hogares e individuos con el método descrito de “Reconstrucción Sintética”, y desarrollaron una variante basada en varios procedimientos de optimización continua y discreta, con el que sintetizaron la población de hogares e individuos de Bélgica.

El método se basa en utilizar los datos disponibles al máximo nivel de desagregación posible para definir las distribuciones conjuntas de hogares e individuos con las que hacer el muestreo de Monte Carlo, permitiendo cierta flexibilidad en relación a los datos disponibles de partida.

El procedimiento se representa en la Figura 9 e incluye tres pasos de ajuste jerárquico. En primer lugar se genera una lista de individuos a partir de las distribuciones de atributos disponibles, ajustando los atributos de los individuos para que la lista total concuerde con los marginales deseados. En segundo lugar se genera la distribución conjunta de atributos de hogares a partir de las distribuciones y datos disponibles, pero imponiendo que dicha

distribución tenga máxima entropía (ver definición de entropía en el capítulo anterior) y mejorando dicho máximo con un algoritmo meta-heurístico utilizado en optimización combinatoria, como la búsqueda tabú. Finalmente, se construye la población sintética extrayendo aleatoriamente individuos de la lista y asignándolos a hogares, a la vez que imponiendo que se mantenga la distribución de atributos de los hogares obtenida, y se minimice la distancia entre los marginales de la población generada y los objetivo, para lo cual hace uso de la métrica Chi-cuadrado de Pearson como medida de la distancia entre los marginales.

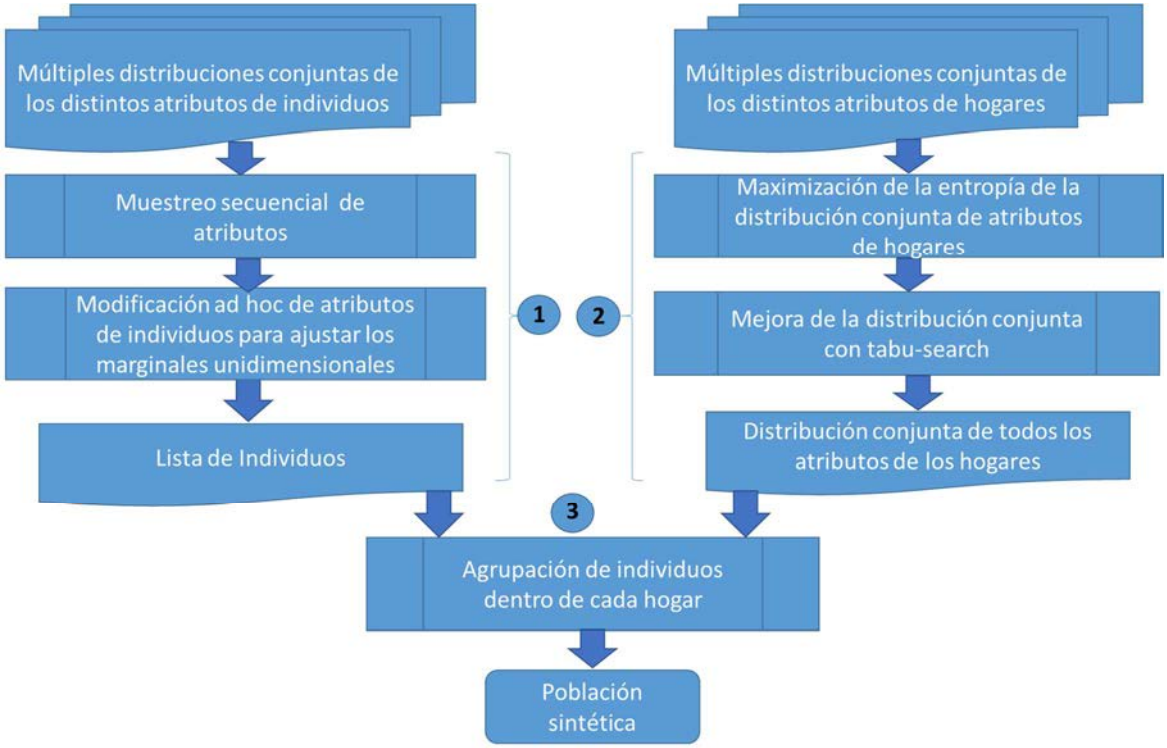

Figura 9 Generador de Población Sintética. Adaptado de Barthelemy & Toint (2013).

Como indican Farooq, Bierlaire, Hurtubia, & Flötteröd (2013) este enfoque no garantiza que la población obtenida quede ajustada simultáneamente a los marginales objetivo de hogares e individuos y además, falta disponer de un modelo general de este método, que generalice y facilite su utilización en otros casos.

3.3.3 Probabilidad Condicional (Conditional Probabilitiy - CP)

Se trata de otra variante de la *reconstrucción sintética*, planteado para utilizar una tabla de una muestra a partir de la que determinar la distribución de probabilidades condicionadas de las distintas categorías de los atributos. Con dichas probabilidades se asignan los valores de los atributos a los agentes de la población. Por tanto, en este caso, donde las probabilidades se obtienen de la muestra, no pueden aparecer agentes en la población final que no estén inicialmente en la muestra.

Al igual que la *reconstrucción sintética*, opera de forma secuencial con los atributos, comenzando con los que se consideran más independientes del resto, por ejemplo, si se comienza por el sexo, se crean tantos individuos hombres y mujeres como se especifiquen en los marginales objetivos. Posteriormente a partir de la muestra se determina la distribución de probabilidades del segundo atributo, por ejemplo edad condicionadas al primero, *P*(edad|sexo), se hace un muestreo de Monte Carlo y se asigna la edad a los individuos. Luego se procede a comparar el número de individuos generados de cada categoría de edad con los valores marginales objetivos, y si se han creado demasiados de alguna categoría se pondera a la baja la probabilidad de tener dicha categoría, o se re-pondera al alza en caso de haber obtenido menos de los deseados, y se procede iterativamente a otro muestreo hasta que se consigan los marginales objetivos de dicho atributo. Este proceso suele converger de forma rápida y una vez se han ajustado los marginales se continúa con el siguiente atributo hasta el último (Harland et al., 2012).

3.3.4 Cadena de Markov Monte Carlo (Markov Chain Monte Carlo - MCMC)

Este nombre es la combinación de dos enfoques: Monte Carlo y Cadena de Markov. El enfoque de Monte Carlo corresponde al de métodos que utilizan números aleatorios para resolver una variedad de problemas matemáticos, desde estimar la media de una distribución hasta determinar el valor de la integral o área de una función.

El enfoque de la cadena de Markov es el de una secuencia de estados tales que la probabilidad de que ocurra cada estado solo depende del estado inmediatamente anterior.

Por tanto, el enfoque MCMC corresponde a la idea de que el muestreo aleatorio se genere mediante un proceso secuencial en el que cada muestra aleatoria se usa para generar la siguiente muestra aleatoria (de ahí la cadena) y aunque cada nueva muestra depende de la anterior, las nuevas muestras no dependen de ninguna muestra anterior a la anterior.

Farooq et al. (2013) propusieron este enfoque para ir generando conjuntos de atributos que representan a los agentes, utilizando las distribuciones conjuntas de los atributos de los agentes.

En concreto, el método se basa en el muestreo de Gibbs (*Gibbs sampling*) el cual es un método tipo *Markov Chain Monte Carlo* (MCMC) que utiliza las distribuciones condicionadas de los atributos para la generación de individuos. Utilizando la notación introducida en la sección 2, si la población se describe con M atributos $X = (x_1, x_2, \dots x_{\mathrm{M}})$, siendo $(\mathbf{C_1}, \mathbf{C_2}, \dots \mathbf{C_M})$ el número de categorías de cada atributo, podemos expresar las probabilidades condicionadas de las que se parten, para el atributo x_i, como:

$p(x_i|x_1, x_2, \dots x_{i-1}, x_{i+1} \dots . x_{\mathrm{M}})= p(x_i|x_{-i})$, donde el subíndice negativo representa a todos los atributos excepto el propio negativo.

Este método produce poblaciones de calidad si se utilizan buenas distribuciones condicionadas de los atributos, lo cual no siempre es posible, ya que en muchas ocasiones no existen datos disponibles para todos los atributos de la población. Si no es posible disponer de las distribuciones condicionadas completas con todos los atributos $p(x_i|x_{-\mathrm{i}})$, para $i = 1 \dots M$, y solo está disponible la distribución condicional parcial con parte de los atributos $p(x_k|x_1, x_2, \dots x_{k-1})$, se asume la independencia condicional del atributo x_k respecto a los restantes atributos $x_{k+1} \dots . x_{\mathrm{M}}$.

Las probabilidades condicionadas pueden construirse a partir de la muestra, o a partir de probabilidades parciales, o de los marginales en el caso extremo, o incluso pueden derivarse de asunciones o conocimiento previo. Por ejemplo, la probabilidad de que un individuo <16 tenga una situación laboral de jubilado será 0.

A continuación se indican los pasos del proceso de generación secuencial de la población con el muestreo de Gibbs:

- A partir de las probabilidades condicionadas iniciales se calculan las probabilidades condicionadas acumuladas para cada categoría de cada atributo *i*, $p(x_i|x_{-\mathrm{i}}), i = 1 \dots M$, por tanto, para la última categoría de cada atributo x_i la probabilidad será 1, ya que son acumuladas.
- Se genera un agente inicial de forma aleatoria con valores de atributos $(c_1, c_2, \dots c_{\mathrm{M}})$, siendo los c_{i} el valor de una categoría del atributo *i*, es decir, c_1 podrá tomar cualquier valor entre $1 \dots \mathbf{C_1}$,… y c_M podrá valer entre $1 \dots \mathbf{C_M}$; este agente pasa a ser el primer agente generado.
- A partir del agente anterior se genera un nuevo agente utilizando las probabilidades condicionadas acumuladas, de cada atributo, es decir, con la distribución de probabilidad condicional acumulada del primer atributo $p(x_1|x_{-1})$, se determina el valor del primer atributo. Los valores de los atributos x_{-1} son los del primer agente. Se genera un número aleatorio entre 0 y 1, y con las probabilidades acumuladas se determina la nueva categoría del primer atributo. Una vez fijado el valor del primer atributo, se procede a generar otro número aleatorio para el segundo atributo usando las $p(x_2|x_{-2})$ y así hasta el último atributo con el que se usará la $p(x_n|x_{-\mathrm{n}})$, siendo los valores de $x_{-\mathrm{n}}$ los valores previamente generados de los atributos.
- Se procede de igual modo, generando nuevos agentes, partiendo del último generado.
- Se genera un número suficiente grande de agentes (por ejemplo, 20.000) hasta que se considere el sistema está en régimen estacionario. A esta fase se denomina “calentamiento”.

- Una vez en régimen estacionario, este proceso converge con el muestreo de Monte Carlo de la distribución conjunta $p(X)$, esto es, la generación de nuevos agentes es como si se obtuvieran de la distribución conjunta $p(X)$ (Train, 2002). Se continúan generando agentes y seleccionando uno de cada cierta cantidad (por ejemplo, cada 10) para la población sintética, para evitar posibles correlaciones entre generaciones consecutivas de agentes, hasta que se tenga el número de agentes deseado.

Este método no requiere que se conozcan los marginales de la población a sintetizar, solo el número total de agentes a generar. Existe una versión posterior de este método denominada "MCMC Jerárquico" (Casati, Müller, Fourie, Erath, & Axhausen, 2015), que permite generar poblaciones multinivel combinando atributos de individuos y hogares al mismo tiempo. Aunque en ningún caso se consigue sintetizar una población con unas distribuciones marginales determinadas.

3.3.5 Modelo Oculto de Markov (Hidden Markov Model - HMM)

Este es otro método basado en procesos de Markov, en este caso en un modelo de Markov oculto, el cual es un modelo probabilístico de Markov donde los estados no son observables explícitamente y solo se observa una secuencia de símbolos que genera el modelo (Saadi et al., 2016).

De hecho, cada estado tiene una distribución de probabilidad sobre los posibles símbolos de salida. Por lo tanto, la secuencia de símbolos generada por un HMM proporciona cierta información sobre la secuencia de estados. El HMM en cada instante t está en un estado q_{t} y en el instante siguiente pasa a un nuevo estado $q_{\mathrm{t+1}}$, siendo la secuencia de observaciones $O = \{O_t, O_{t+1} \dots\}$ la cual se corresponde con los valores de los atributos de cada agente.

El resultado se produce de acuerdo a las funciones de probabilidad de las observaciones. La observación de cada estado se selecciona entre el conjunto de las *m* categorías posibles. Entonces, el modelo puede transitar de un estado a otro de acuerdo a la probabilidad de transición entre estados.

Se trata de un modelo discreto de M estados, el número de atributos de la población, con $\mathrm{C_m}$ salidas posibles cada uno. Formalmente, el HMM se define con los siguientes 5 parámetros $(X, \mathrm{C_m}, T, \pi, E) = \lambda$.

1. Conjunto finito de M estados ocultos, correspondientes a los M atributos $X = (x_1, x_2, \dots x_{\mathrm{M}})$. Los atributos se ordenan de forma descendente (de izquierda a derecha) según el número de categorías, poniendo primero el que tenga más categorías y el último el que menos.

2. Conjunto finito C_m de símbolos de observación por estado, es decir, la salida física observable, que corresponde a los valores de las categorías de cada atributo $C_{\mathrm{m}} = (1,2, \dots \mathbf{C_m}),\ m = 1 \dots M$.
3. La matriz de transición T, que representa la probabilidad de pasar del estado x_i al estado x_j; $T = \{t_{ij}; i,j = 1 \dots \mathrm{M}\}$, es decir, entre la categoría *i* del atributo de la izquierda y la categoría j del siguiente atributo:

$$t_{\mathrm{ij}} = P[q_t = x_{\mathrm{i}}\ y\ q_{t+1} = x_{\mathrm{j}}] \quad i,j = 1, \dots \mathrm{M}\ ;\ \ t_{\mathrm{ij}} \geq 0\ ;\ \sum_{j=1}^{\mathrm{M}} t_{ij} = 1$$

$t_{ij} = \frac{p_{ij}}{\sum_{k=1}^{N} p_{ik}}\ \forall\ i,j = 1, \dots \varepsilon$; siendo p_{ij} el número de transiciones que hay en la muestra entre el estado *i* de un atributo y el estado *j* del siguiente atributo, y N el número total de transiciones posibles que se inician desde el estado *i.*

Y, $\varepsilon = \sum_{i=1}^{M-1} Cat(x_i) * Cat(x_{i+1}) =$ número de probabilidades de transición, es decir, la suma de los M-1 productos del número de categorías del atributo *i* por el número de categoría del siguiente atributo *i+1*.
4. El vector de probabilidad de estado inicial π, que representa probabilidades de estados iniciales: $\pi = \{\pi_i; i = 1 \dots \mathrm{M}\}$; $\pi_{\mathrm{i}} = P[q_1 = x_{\mathrm{i}}]$; $\pi_{\mathrm{i}} \geq 0$; $\sum_{i=1}^{\mathrm{M}} \pi_i$=1; si es el primer atributo, la distribución inicial puede obtenerse de la distribución marginal objetivo.
5. La matriz de probabilidades de observación de los símbolos E, que representa la probabilidad de producir la observación c_m en el instante t en el estado x_i; $E = \{e_{jk}(O_t)\}\ O_t = k$

$$e_{jk}(O_t) = P[c_k\ en\ t |\ q_t = x_{\mathrm{j}}]\ , j = 1, \dots \varepsilon\ ;\ k = 1, \dots m$$

En este modelo se considera que cada estado emite solo un símbolo observable en un estado dado, que corresponde a una categoría, por lo que la matriz de probabilidades de observación se construye con probabilidades que son 1, y las filas de esta matriz por tanto suman 1:

$$\sum_{k=1}^{\mathrm{C_M^*}} e_{jk} = 1\ \ \forall j = 1 \dots \varepsilon$$

Siendo ε el número total de estados y $\mathrm{C_M^*}$ el número de categorías del atributo con más categorías (mayor número posible de símbolos emitido por uno de los estados).

Dada una secuencia de símbolos observada $O = O_1, O_2, O_3 \dots O_t$ (que se tiene en los agentes de muestra) y una secuencia de estados $Q = q_1, q_2, q_3 \dots q_t$ (el conjunto de atributos), la $P(O, Q|\lambda)$ se obtiene:

$$(O, Q|\lambda) = P(O|Q, \lambda)P(Q|\lambda) = \pi_{q_1} e_{q_1}(O_1) t_{q_1 q_2} e_{q_2}(O_2) t_{q_2 q_3} \dots t_{q_{t-1} q_t} e_{q_t}(O_t)$$

Ya que:

$$P(O|Q,\lambda) = e_{q_1}(O_1)e_{q_2}(O_2)\dots e_{q_t}(O_t)$$
$$P(Q|\lambda) = \pi_{q_1} t_{q_1 q_2} t_{q_2 q_3} \dots t_{q_{t-1} q_t}$$

Con los agentes de la muestra se calibra el modelo y se determinan los 5 conjuntos de parámetros de la HMM. Una vez se tienen los parámetros del modelo, se genera la secuencia de atributos de cada agente de la población sintética, lo cual puede realizarse con la función simHMM del paquete HMM (Himmelmann, 2010) de R, o con la función hmmgenerate de MATLAB.

La principal ventaja de este método es que no exige disponer de una muestra completa con todos los atributos, sino que permite fusionar información proporcionada por múltiples fuentes, por ejemplo sería el caso de dos muestras parciales de la población, que tuvieran dos atributos en común y el resto distintos, a partir de las que se determinaría la matriz de probabilidades de transición.

3.3.6 Inferencia de la Distribución Conjunta (Joint Distribution Inference - JDI)

Este método infiere la distribución de probabilidades conjunta de la población a partir de distribuciones parciales y de distribuciones de los marginales objetivos totales de la población. Sus autores (P Ye, Hu, Yuan, & Wang, 2017) defienden que el método crea poblaciones sintéticas que se ajustan de forma más precisa a distribuciones parciales de atributos que las poblaciones obtenidas con métodos de reconstrucción sintética, MCMC o de tipo *reweighting*.

El proceso comienza verificando que no existen atributos, o conjuntos de atributos, independientes del resto, ya que si fuera así, bastaría conocer la distribución de dichos atributos, o grupos de atributos independientes, y la distribución conjunta se obtendría a partir de las distribuciones parciales de estos grupos. Por ejemplo, si hay M atributos $(x_1, x_2, \dots x_{\mathrm{M}})$, y los tres primeros son un conjunto independiente del resto:

$$f(x_1, x_2, \dots x_{\mathrm{M}}) = f(x_1, x_2, x_3) f(x_4, x_5, \dots x_{\mathrm{M}})$$

Esta comprobación se realiza mediante la prueba de la chi cuadrado χ^2 de las distribuciones bidimensionales de atributos que haya disponibles. Por ejemplo, si solo se dispone las distribuciones bidimensionales del primer atributo con los demás, se realiza la prueba con cada una de estas distribuciones para verificar si el primer atributo es independiente del resto.

Este test es una prueba de contraste clásica que permite determinar, mediante el estadístico chi cuadrado χ^2, si dos atributos de los individuos de la población están relacionados.

Seguidamente, para cada conjunto de atributos no independientes se infiere la distribución conjunta a partir de sus distribuciones marginales unidimensionales objetivo y distribuciones conjuntas parciales (bidimensionales, etc.), si están disponibles.

Por ejemplo, para el caso de inferir la distribución conjunta de 3 atributos $f(x,y,z)$ conociendo solo las distribuciones parciales $f(x,y)$ y $f(x,z)$, se usa el producto de las distribuciones condicionadas:

$$f(x,y,z) = f(x)f(y|x)f(z|x)$$

En el caso de que se conociera la distribución $f(y,z)$ y las distribuciones marginales unidimensionales, entonces se calcula como:

$$f(x,y,z) = f(x)f'(y,z)$$

Siendo $f'(y,z)$ la distribución obtenida haciendo IPF (ver siguiente sección) con las distribuciones marginales $f(y)$ y $f(z)$ y usando como semilla $f(y,z)$.

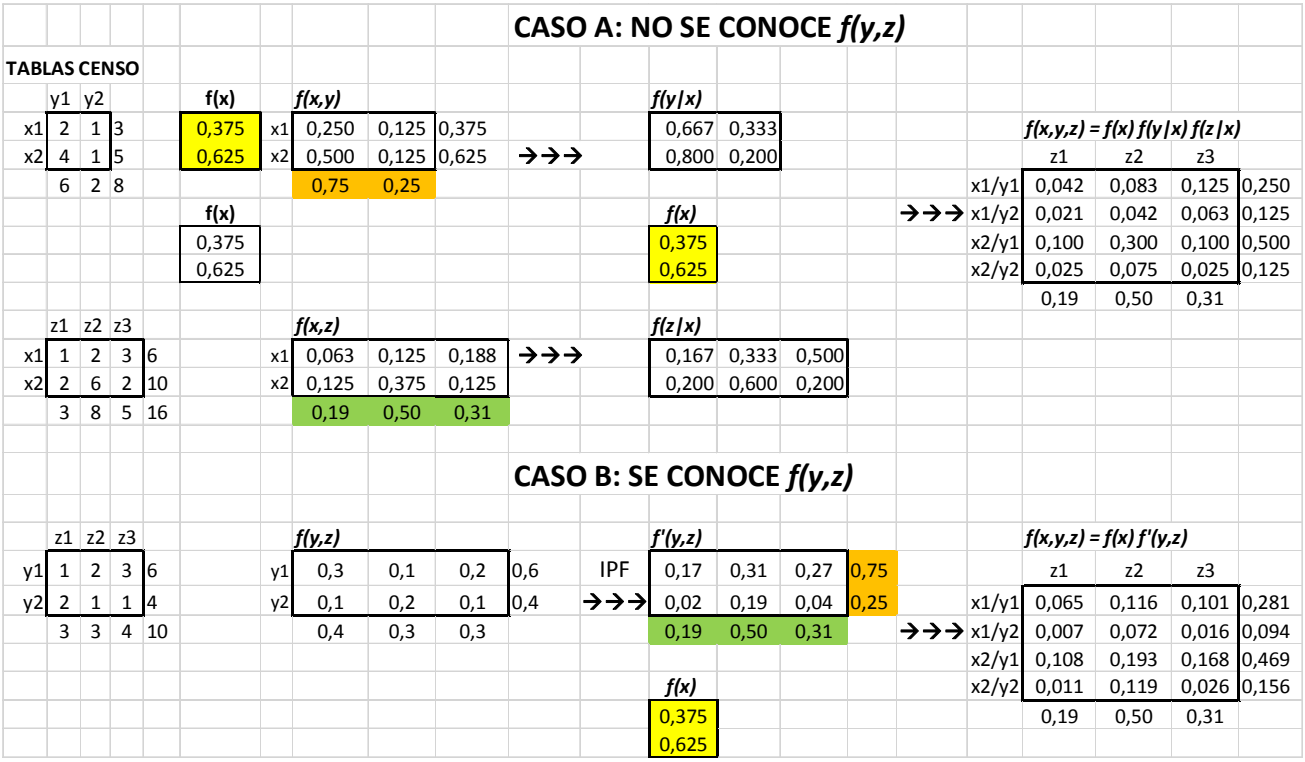

Figura 10 Ejemplos de inferencia de distribución f(x,y,z).

Posteriormente multiplica las distribuciones de cada grupo para obtener la distribución conjunta de la población total. Y una vez se tiene la distribución conjunta (*joint distribution*) completa se procede con un muestreo de Monte Carlo para obtener la población sintética.

En la Figura 10 se muestra un ejemplo de ambos casos.

*3.3.7 Redes Bayesianas (*Bayesian Networks - *BN)*

Una red Bayesiana G representa las relaciones probabilísticas entre un conjunto de M variables, que en este caso son los atributos de la población $X = (x_1, x_2, \dots x_M)$. Esencialmente consta de 2 componentes $G = \{(G_S, G_P)\}$. La primera es la estructura de la red, un grafo dirigido acíclico (GDA), cuyos nodos son los atributos y las aristas van dirigidas de unos nodos a otros, indicando la dependencia entre los atributos. Se expresa como $G_S = \{(x_i, \pi_i)\}$, siendo π_i el conjunto de padres del nodo del atributo x_i. La segunda componente son los parámetros, constituidos por un conjunto de tablas de probabilidad condicional que representan el grado de interdependencia entre los atributos, $G_P = \{P(x_i|\pi_i)\}$. Según el principio de independencia condicional, la distribución de probabilidades conjuntas de X puede obtenerse mediante la expresión:

$$P(X) = \prod_{i=1}^{M} P(x_i|\pi_i)$$

A partir de una red bayesiana G puede determinarse la distribución $P(X)$ con la que proceder a un muestreo de Monte Carlo para generar los agentes que constituyen la población sintética. Por tanto, este método consiste en determinar la red bayesiana de la población a partir de los datos disponibles, que habitualmente es una muestra aleatoria junto con los marginales, aunque en este método los marginales no son requisito necesario, y basta con disponer de las distribuciones de probabilidades de las categorías de cada atributo (probabilidades a priori $P(x_i)$).

Así mismo, a partir de la red Bayesiana también sería posible determinar las distribuciones condicionadas completas con todos los atributos $p(x_i|x_{-\mathrm{i}})$ utilizadas en el método MCMC de la sección anterior, basándose en la cobertura (*Markov Blanket*) de cada nodo, la cual comprende los padres del nodo, los hijos y los padres de los hijos.

$$P(x_i|x_{-i}) \propto P(x_i|\pi_i) \prod_{k \in hijos(i)} P(x_k|\pi_k)$$

Por lo que este método también podría aplicarse combinadamente con el anterior MCMC.

Este método de Redes Bayesianas define el proceso de selección de la mejor red a partir de unas observaciones dadas (muestra *S*). Este proceso consiste en determinar la mejor estructura de la red G_S en función de la que posea el mejor BIC (Criterio de Información Bayesiano) o AIC (Criterio de Información de Akaike), dos criterios de uso frecuente para la

selección de modelos bayesianos, que miden la verosimilitud del grafo para explicar la muestra observada[4].

Dado que el número de posibles grafos GDA, entre los que seleccionar el mejor, se incrementa exponencialmente con el número de nodos (con 6 nodos hay casi 3 millones de grafos posibles y con 7 nodos más de mil millones), el método utiliza las funciones del paquete de R bnlearn (Scutari, 2010) que implementan algoritmos como la búsqueda tabú o de la escalada simple (*Hill Climbing*), para encontrar la estructura de red bayesiana que maximice los criterios de selección de modelos.

Una vez se dispone de la mejor estructura de la red, se utilizan las probabilidades condicionadas calculadas a partir de la muestra y mediante muestreo de la red bayesiana inferida se generan la población de agentes. La población generada con este método no se ajusta con precisión a los marginales objetivos.

Este método fue ideado por Sun & Erath (2015), que lo aplicaron para generar la población de Singapur, concluyendo que es un método muy potente para determinar la distribución conjunta de atributos de la población cuando se tienen muestras pequeñas (<20%) y al no estar basado en clonar los elementos de la muestra produce poblaciones de gran heterogeneidad.

3.3.8 Kernel Cruce "K-Vecinos más cercanos" (K-nearest neighbors Crossover Kernel - KNN)

Se basa en estimar la función densidad de la población a sintetizar a partir de la muestra. Posteriormente, mediante un muestreo probabilístico de Monte Carlo usando la distribución de la función densidad se generan los agentes sintéticos (Hamada, Homma, Higuchi, & Kikuchi, 2015).

La función densidad (de probabilidad) de una variable aleatoria discreta es la función definida mediante: $f(x) = P(X = x)$ es decir, se corresponde con la frecuencia relativa. En el caso de poblaciones con un conjunto de atributos $x = (x_1, x_2, \dots x_m)$ la función densidad se corresponde con los valores de las celdas de la tabla multidimensional de la población.

Este método consiste en construir una estimación no paramétrica de la función densidad mediante el método *Kernel*, también conocido como estimadores tipo núcleo, el cual estima

[4] BIC y AIC miden la calidad relativa de un modelo estadístico con k parámetros, para un conjunto dado de datos: $BIC = -2\ln\hat{L} + k\ln n$ y $AIC = 2k - 2\ln\hat{L}$, siendo $\hat{L}$ el máximo valor de la verosimilitud del modelo para los datos dados, k el número de parámetros del modelo y *n* número de datos.

la función en torno a los valores de los elementos de la muestra de tamaño n, $x^i = (x_1^i, x_2^i, \dots x_m^i)$ *(i=1..n).* La idea en la que se inspira el método *Kernel* es similar al concepto de media móvil, que puede visualizarse como una ventana que se desliza a lo largo de una serie, y aquí la ventana tendría la forma definida en la función *kernel* que se establezca. El método estima la densidad como:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{i=n} K\left(\frac{x - x^i}{h}\right)$$

Siendo h lo que se denomina "anchura de la ventana", que representa el área de influencia que se pretende dar a cada elemento de la muestra $x^1, \dots x^n$, y K la función *kernel.* Se ha utilizado el símbolo de acento circunflejo para indicar que se trata de un estimador de la función densidad *f(x).*

En lugar de utilizar como *kernel* las funciones habituales (normal, binomial, exponencial, etc.) este método utiliza un tipo de función *kernel*, llamada "*kernel* cruce" (*crossover kernel*), que utiliza el operador "cruce", utilizado en los algoritmos genéticos, como una función de *kernel*. Este operador crea un individuo de nueva generación a partir de otros individuos que son sus padres (generación previa).

Este "*kernel* cruce" no precisa parámetro de anchura de ventana para determinar la forma de la función, en su lugar utiliza un subconjunto constituido por "*p"* elementos de la muestra, que serán los padres para el "cruce", llamado "conjunto de construcción del *kernel*" (x^* KCS, del inglés *Kernel Construction Set*). Los elementos del KCS se eligen seleccionándolos entre los "k-vecinos más cercanos" (K-NN) a cada elemento de la muestra y con esto se consigue que el método de estimación de la densidad tenga mayor exactitud que si se utilizan otros *kernels* convencionales de tipo gaussiano.

A partir de los "*n*" KCS se generan nuevos conjuntos, generando combinaciones a partir de cada uno, tomando al azar elementos de los KCS de manera uniforme y con reemplazo (por tanto, algunos elementos del conjunto original pueden aparecer repetidos en los nuevos conjuntos generados), y se determina el valor esperado de todas las posibles elecciones de KCS:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{i=n} E_{x^{*i}}[K(x|x^{*i})]$$

Siendo $E_{x^{*i}}[\cdot]$ el valor esperado sobre todas las posibles elecciones del KCS x^{*i} de tamaño "*p*" de los "k -vecinos más cercanos" de x^i. Para ejecutar el algoritmo hay que fijar por tanto el valor de "*p*" y de "*k*".

Una vez se ha determinado la función densidad, se genera la población sintética mediante un muestreo de Monte Carlo. Para cuadrar la población con marginales objetivo, hay que proceder a seleccionar individuos en la fase final de la generación, lo cual distorsiona la población que se obtiene.

3.3.9 Síntesis con Función Cópula

Se trata de un método estadístico planteado por Kao, Kim, Liu, Cui, & Bhaduri (2012), que preserva la distribución marginal de los atributos y mantiene la estructura de correlación de rangos entre los mismos.

Si por ejemplo, se ordenan los hogares de la muestra de menor a mayor valor de sus ingresos y se determina el número de orden de cada uno, y por otro lado se ordenan de menor a mayor según el número de vehículos del hogar y se vuelve a determinar el número de orden, la correlación que existe entre los dos números de orden (medida con el coeficiente de correlación de Spearman[5]) es la que este método preserva en la población sintética (correlación de rangos).

El método se basa en construir lo que en estadística se llama "cópula", una función de distribución de probabilidad conjunta con distribuciones marginales uniformes. El nombre de "cópula" proviene de que se trata de funciones que "unen" funciones de distribución marginales unidimensionales uniformes en el intervalo [0, 1]. Existen distintas familias de cópulas, siendo la cópula gaussiana una de las más habituales. En la Figura 11 se muestra un ejemplo de la densidad de una cópula bivariada con marginales normales estándar.

Requiere que los valores de los atributos sean ordenables, por lo que atributos como sexo, estado civil, o nacionalidad, no son directamente manejables por esta técnica.

A partir de la muestra se determinan los coeficientes de correlación de Spearman entre cada dos atributos y se calcula la matriz de correlación con la que se construye la función cópula (normalmente una función cópula gaussiana) conjuntamente con las distribuciones de los marginales objetivo (si no se conocen las distribuciones de marginales objetivos se utilizan las distribuciones de marginales de la muestra), y se procede a un muestreo de Monte Carlo para generar la población sintética.

[5] Este coeficiente viene dado por la expresión $\rho = 1 - 6\sum D^2/(\mathrm{N}(\mathrm{N}^2 - 1))$, siendo D la diferencia entre los números de orden de las dos ordenaciones de cada hogar de la muestra, y N el número de hogares de la muestra. Al igual que el coeficiente de correlación de Pearson su valor está entre 1 y -1.

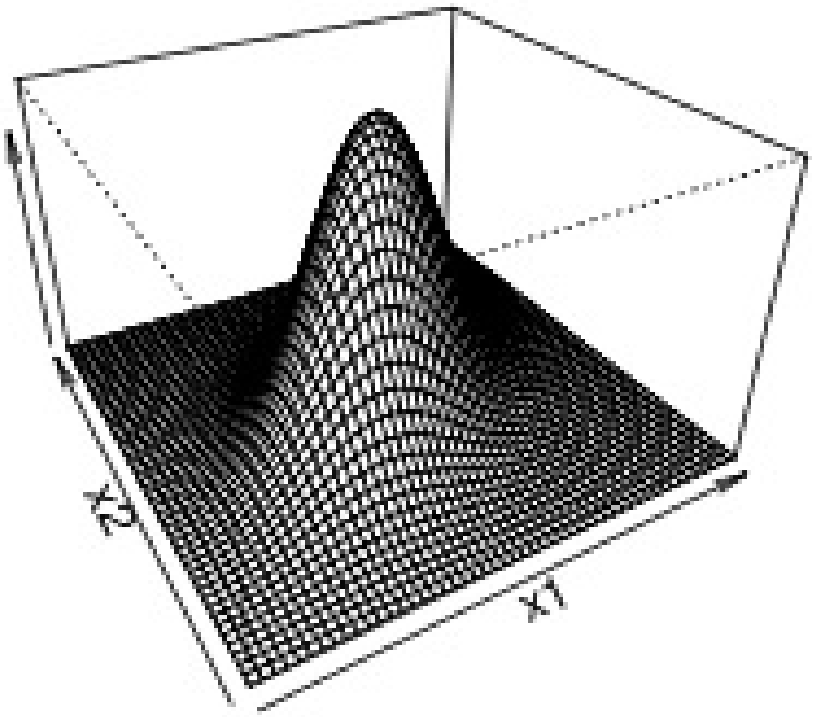

Figura 11 Perspectiva de la función densidad de una cópula bivariada con marginales normales. Fuente: https://en.wikipedia.org/wiki/File:Gaussian_copula_gaussian_marginals.png Licencia: Creative Commons Attribution-Share Alike 3.0 (CC BY-SA 3.0) Autor: Matteo Zandi.

Los autores de este método compararon el coeficiente de correlación de Spearman de las parejas de atributos de una muestra con 6 atributos y comprobaron que las correlaciones de la población sintetizada con este método se parecían más a las de la muestra que las correlaciones de las poblaciones obtenidas con IPF.

Un método similar basado en una función cópula no gaussiana, que también preserva la estructura de dependencia de los atributos de la muestra puede encontrarse en el estudio de Jeong, Lee, Kim, & Shin (2016).

3.3.10 Síntesis con modelos de correlación o regresión

Estos métodos se utilizan cuando se dispone de información sobre las distribuciones de los datos a sintetizar y de las correlaciones entre los atributos. Se trata de métodos de gran sencillez, que permiten generar secuencias de valores aleatorios correlacionados con las distribuciones predeterminadas (media y desviación), sin buscar el ajuste de los marginales de la población. Están diseñados principalmente para atributos con valores continuos, aunque pueden adaptarse para incluir atributos discretos.

El método fue propuesto por Nissen & Saft (2014) para construir una población de agentes con atributos continuos correlacionados, con el objetivo de ser utilizados en simulaciones de

agentes. En la descripción del método se plantean generar una población de pintores caracterizados por su edad, la calidad de sus cuadros, el precio de los mismos y la reputación del artista como creador de escuela. Cuatro atributos relacionados para los que se fijan unos valores medios, desviaciones y su matriz de correlación.

Con este método se generan valores aleatorios correlacionados de los 4 atributos x_1, x_2, x_3, x_4 a partir de un conjunto de valores aleatorios y_1, y_2, y_3, y_4 no correlacionados, conociendo los valores medios $\mu_1, \mu_2, \mu_3, \mu_4$ de los atributos, las desviaciones estándar de cada uno $\sigma_1, \sigma_2, \sigma_3, \sigma_4$ y la matriz (simétrica) de coeficientes de correlación ρ_{ij} entre los atributos.

$$C = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{21} & 1 & \rho_{23} & \rho_{24} \\ \rho_{31} & \rho_{32} & 1 & \rho_{34} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 \end{pmatrix}$$

El enfoque matemático propuesto en este método consiste en encontrar la descomposición de Cholesky de la matriz de correlaciones C, esto es, encontrar una matriz triangular inferior L tal que el producto de dicha matriz por su traspuesta sea la matriz $C = L * L'$ (esta matriz puede obtenerse fácilmente con la función Cholesky del paquete estándar de R Matrix). Y ejecutar los siguientes pasos:

a) Generar el vector de valores aleatorios $Y' = (y_1, y_2, y_3, y_4)$ donde los y_i tienen distribución normal *N(0,1).*
b) Calcular el vector $Z = L * Y$
c) Obtener el vector $X = (x_1, x_2, x_3, x_4)$, mediante: $x_i = \mu_i + \sigma_i * z_i$

Repitiendo N veces estos tres pasos, se obtienen N agentes caracterizados por los 4 atributos, correlacionados dos a dos según las correlaciones de la matriz de correlación, y con la media y desviación estándar establecidas. Este mismo método se aplica también para crear agentes que, además de tener definidas las correlaciones entre sus atributos, tienen correlación entre ellos mismos. De este modo generan la población de pintores pertenecientes a distintas escuelas pictóricas, de forma que los que pertenecen a la misma escuela tienen determinada correlación entre el precio y calidad de sus cuadros. La matriz de correlación de los atributos de estos agentes pintores, para el caso de que el segundo y tercer atributo (precio y calidad) del primer agente tenga unas determinadas correlaciones ρ' con el segundo y tercer atributo del agente n es:

$$C = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} & & 0 & 0 & 0 & 0 \\ \rho_{21} & 1 & \rho_{23} & \rho_{24} & \cdots & 0 & \rho'_{22} & \rho'_{23} & 0 \\ \rho_{31} & \rho_{32} & 1 & \rho_{34} & & 0 & \rho'_{32} & \rho'_{33} & 0 \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 & & 0 & 0 & 0 & 0 \\ \vdots & & & & \ddots & & & \vdots & \\ 0 & 0 & 0 & 0 & & 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ 0 & \rho'_{22} & \rho'_{23} & 0 & \cdots & \rho_{21} & 1 & \rho_{23} & \rho_{24} \\ 0 & \rho'_{32} & \rho'_{33} & 0 & & \rho_{31} & \rho_{32} & 1 & \rho_{34} \\ 0 & 0 & 0 & 0 & & \rho_{41} & \rho_{42} & \rho_{43} & 1 \end{pmatrix}$$

Si en lugar de disponer de las correlaciones entre parejas de atributos se dispone de un modelo de regresión multivariada, es posible aplicar la misma técnica de generación de valores aleatorios para los atributos independientes y obtener así los valores de los atributos dependientes del modelo.

Ejemplos de construcción de poblaciones sintéticas haciendo uso de regresiones pueden encontrarse en (Frazier & Alfons, 2012) y (Chen, Evans, Frisby, Izquierdo, & Plale, 2016). En el primero de estos estudios se genera la población de Ghana mediante una regresión logística multinomial a partir de los datos de la encuesta de condiciones de vida (*Ghana Living Standard Survey*). La calidad de la población sintética está en función de la calidad del modelo que se obtiene con los datos de la encuesta.

En el segundo estudio, ante la falta de dos tercios de los valores de un determinado atributo (tamaño del hogar) de una población de 53.491 hogares de una zona agraria de Zambia, se utilizó el tercio de valores conocido para construir un modelo lineal generalizado (estos modelos son una extensión de los modelos de regresión tradicionales) con el que predecir los valores desconocidos del atributo bajo la asunción de que existe una correlación con otro atributo con valor conocido (atributo dependiente: tamaño del hogar; atributo independiente: superficie de cultivo). A partir de este modelo se completan los valores del atributo tamaño del hogar de toda la población de 53.491 hogares de agricultores.

3.3.11 Síntesis con Algoritmos Genéticos

Los algoritmos genéticos (GA, del inglés *Genetic Algorithm*s) son una alternativa para asignar valores de atributos en los casos en que la información que se disponga consista en determinados valores marginales y los rangos de los posibles valores de los atributos, no disponiéndose de muestra de la población.

Los algoritmos genéticos son algoritmos de optimización inspirados en los procesos de evolución natural y de evolución genética. Las posibles soluciones del problema de optimización se representan por cromosomas compuestos por genes que en nuestro caso se corresponden con los atributos que se desean determinar. En cada iteración se crea un

conjunto de cromosomas que mediante operaciones de selección, cruce, mutación de los distintos genes de los cromosomas de la generación anterior van evolucionando y creando nuevas generaciones de cromosomas. En cada generación se van descartando los peores cromosomas y se vuelven a aplicar las operaciones de cruce y mutación hasta obtener los mejores cromosomas que satisfagan la condición de ajuste de marginales que se imponga.

En el último estudio citado en la sección anterior (sección 3.3.10), correspondiente a la población de hogares de agricultores de una zona de Zambia, además de utilizar un modelo de regresión para determinar los valores desconocidos del tamaño de los hogares a partir de la superficie cultivable de la parcela, se utilizaron algoritmos genéticos para asignar los valores de otros atributos desconocidos para los que no existían encuestas.

En ese caso se trataba de asignar a cada hogar de agricultores las características de su parcela de cultivo: tipo de suelo con 15 valores posibles, porcentaje de producción de maíz local frente a maíz híbrido (entre 0% y 100%) y fecha de siembra (8 valores posibles). Para cada parcela se conocía la superficie de cultivo. A nivel país se disponían de los datos de una encuesta con la distribución del rendimiento agrícola de la producción de maíz local (PHS, del inglés *Post Harvest Survey*). Esta encuesta proporcionaba el número medio de parcelas con cada nivel de producción de maíz, en incrementos de 1.000kg/hectárea, para la cosecha de 2011-2012.

La asignación de los valores de los atributos se realizó con algoritmos genéticos que encontraron los cromosomas que asignados a los distintos hogares de la población de la zona producían la distribución del rendimiento agrario, de las parcelas de la zona estudiada, más próxima al rendimiento observado en la PHS a nivel país.

Este método facilita la creación de poblaciones con el nivel de heterogeneidad que se determine.

3.3.12 Síntesis con técnicas de Imputación múltiple

Para finalizar la descripción de los métodos de generación de datos sintéticos tipo probabilístico terminamos esta sección con la descripción de un tipo de métodos que se utilizan para realizar imputación estadística, y que permiten generar datos sintéticos a partir de los datos originales, por lo que en primer lugar se introduce el concepto de imputación.

La imputación estadística, o imputación de datos, es el proceso de sustitución de valores faltantes dentro de un conjunto de datos, por otros sustitutivos. Entre los métodos más conocidos de imputación de datos se encuentran métodos basados en la distancia, como "k-vecinos más cercanos" (k-*nearest neighbors* kNN), imputación por media condicional o los métodos basados en modelos, como el de regresión secuencial multivariante o imputación

múltiple por ecuaciones encadenadas (*Multivariate imputation by chained equations* MICE), también conocido con el nombre en inglés de *sequential regression multiple imputation*.

Estos últimos se basan en definir modelos para los datos o atributos faltantes. Los modelos han de predecir los datos faltantes en función de los valores de los otros atributos no faltantes, creando imputaciones por medio de una secuencia de ecuaciones encadenadas. Para construir los modelos se utilizan técnicas de regresión paramétrica que imputan el valor de cada variable (atributo) x_i a partir de la regresión de x_i en función de las otras variables $(x_1, \dots x_{i-1}, x_{i+1} \dots x_m)$. Cada modelo de regresión se diseña según el tipo de variable que se imputa, por ejemplo, una regresión logística cuando x_i es binaria, logit multinomial si la variable fuera categórica, regresión lineal si fuera continua, Poisson para variables numéricas discretas tipo contador, etc.

Los métodos de imputación múltiple consisten en sustituir cada valor faltante por un conjunto de valores, con lo que se tienen varios conjuntos completos de datos, a partir de los que se obtiene estimaciones que se combinan para determinar un valor final para cada dato faltante.

Hoy en día se utilizan técnicas como el remuestreo (*bootstrap*) o *Markov Chain Monte Carlo* (MCMC) para la imputación múltiple. Estas técnicas generan varias copias de los datos imputados, reemplazando los valores de las variables por nuevos valores generados a partir de un modelo. Las distintas copias se integran en una población sintética final, mediante reglas simples, como puede ser la media de todas, de forma que está diseñada para que el análisis de la población sintética final produzca estimadores consistentes, como si se obtuvieran de los datos verdaderos.

En la 12 se muestra un esquema de un ejemplo de imputación múltiple en una población de N agentes con 3 atributos. En este ejemplo solo hay un atributo faltante que es la edad (en determinados registros) y se utiliza un modelo de relación entre el atributo faltante y el resto de atributos tipo regresión lineal:

$$edad = \beta_0 + \beta_1 * ingresos + \beta_2 * sexo + \varepsilon(\sigma) \qquad (16)$$

En este modelo, los errores ε tienen una distribución normal con valor esperado $E(\varepsilon) = 0$ y $Var(\varepsilon) = \sigma^2$. La variable $sexo$ solo podrá tomar valores binarios 0/1.

El proceso se desarrolla del siguiente modo:

1. Selección de varias muestras de tamaño *n* con un determinado porcentaje de valores faltantes repartidos aleatoriamente.

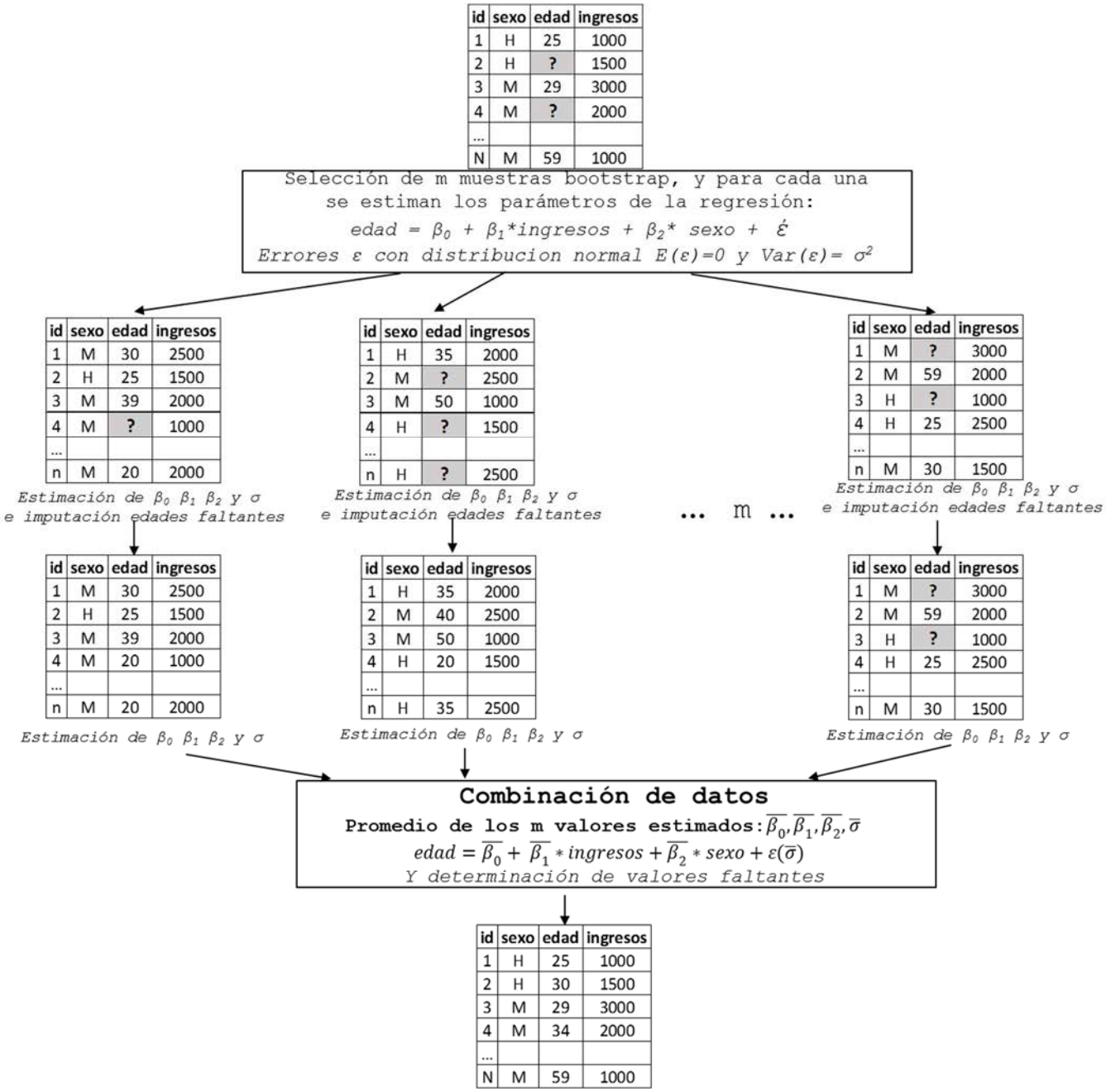

Figura 12 Esquema de imputación múltiple con regresión lineal.

2. Selección de varias muestras de tamaño *n* con un determinado porcentaje de valores faltantes repartidos aleatoriamente.
3. Para cada muestra se estiman los parámetros de la regresión (16) con los valores no faltantes y se imputan dichos valores estimados por la regresión a los valores faltantes de edad.
4. Para cada muestra se vuelven a estimar los parámetros de la regresión (16) utilizando todos los registros de la muestra.

5. Una vez se han obtenido los múltiples conjuntos de datos imputados y los múltiples parámetros de regresión, se combinan los parámetros para determinar unos parámetros promedio del modelo de regresión.
6. A partir del nuevo modelo de regresión se imputan los valores faltantes en la población inicial.

La imputación de datos faltantes puede considerarse como un modo de generación de datos sintéticos, pudiéndose aplicar estas técnicas de imputación al caso generación de datos sintéticos en situaciones en las que se disponga del 100% de la población y se desee construir un conjunto de datos equivalente.

Esto es lo que tradicionalmente hacen las agencias estadísticas para preservar la confidencialidad de ciertos datos, la anonimización de los datos, que es reemplazar los valores originales sensibles con nuevos datos causando una mínima distorsión de la información estadística contenida en el conjunto de datos. Este tipo de anominización puede considerarse como un método de creación de datos sintéticos.

Hay que aclarar que existen otras técnicas de anonimización que no tienen relación con los datos sintéticos, como son los métodos clásicos de perturbación (agregación, recodificación, intercambio de registros, supresión de valores, adición de ruido, redondeo, etc.).

Rubin (1993) fue el primero que propuso, con objeto de preservar la confidencialidad de los datos sensibles, que las organizaciones no difundieran los microdatos reales, sino que publicaran datos sintéticos construidos mediante modelos aplicando las técnicas de imputación múltiple para que pudieran ser analizados utilizando las técnicas estadísticas estándar como si fueran los microdatos reales.

La generación de datos sintéticos con imputación múltiple basada en modelos es un enfoque muy válido y flexible que se utiliza habitualmente para producir datos sintéticos en general. Como se ha explicado, consiste en derivar un modelo predictivo de los atributos de la población a partir de los microdatos disponibles (o muestra) y posteriormente predecir múltiples conjunto de datos sintéticos, a partir de los que crear el nuevo conjunto de datos.

Si los microdatos disponibles son todos los originales y el modelo reproduce las relaciones entre las variables que ajustan el mismo, la generación de datos será completa y efectiva. Si los microdatos disponibles son solo una muestra y el modelo no considera las interacciones importantes entre variables, los datos sintéticos no serán tan perfectos y tendrán propiedades estadísticas diferentes a los originales.

Por tanto, la idea detrás de los datos sintéticos, obtenidos de este modo, es disponer de un modelo a partir de los propios datos originales y luego extraer muestras de este modelo, que constituirán los datos sintéticos.

Se ha utilizado la imputación múltiple para la creación de datos sintéticos en distintas situaciones (datos parcialmente sintéticos para valores raros o poco comunes, imputación múltiple selectiva, etc.).

El método propuesto por Rubin de imputación múltiple utiliza modelos paramétricos para crear los múltiples conjuntos de datos sintéticos, por lo que pueden producirse errores derivados de que el modelo no se ajuste bien a los datos o de asumir que los datos se ajustan a una determinada distribución, cuando realmente no sea así. Con objeto de evitar estos errores, se utilizan modelos que no precisan estas asunciones, es decir, pueden utilizarse modelos no-paramétricos como árboles de clasificación y regresión (CART, del inglés *Classification And Regression Tree*), *random forest* o máquinas de vectores soporte (SVM, del inglés *Support Vector Machine*), entre otros.

Los CART son arboles de clasificación (para variables categóricas) o de regresión (para variables continuas) que en este caso se construyen con los microdatos existentes (variables explicativas) para caracterizar las distribuciones condicionadas de las variables a imputar (sintetizar), a partir de las que se muestrean los agentes (Reiter, 2005).

Los *random forest* (Caiola & Reiter, 2010), en lugar de utilizar CART para el modelo de datos, utilizan una combinación de árboles de clasificación. Y las máquinas de soporte vectorial son algoritmos de aprendizaje supervisado, para resolver problemas de clasificación y regresión, que se basan en encontrar el mejor “hiperplano” que permita separar y clasificar los datos.

Todos estos modelos son utilizados como técnicas de imputación de datos para generar datos sintéticos a partir de los datos originales.

Este tipo de métodos de generación de datos sintéticos, tal como se desprende la descripción de los mismos, no tienen en cuenta los marginales y no persiguen el ajuste de los mismos.

Existen paquetes software como synthpop de R (Nowok, Raab, & Dibben, 2016) que dispone de una función (syn) para generar datos sintéticos a partir de los datos originales. Esta función permite especificar el modelo para cada variable de los datos. Entre los posibles modelos paramétricos están todos los tipos de regresión (lineal normal, logística, multinomial, etc.) y entre los no paramétricos, los ya citados de CART y *random forest*. La función también permite crear un único conjunto de datos sintéticos o múltiples síntesis que pueden combinarse posteriormente.

Este paquete de generación de datos sintéticos incluye una función para comparar los marginales de los datos sintéticos con los de los datos originales. Asimismo, permite ajustar modelos lineales y modelos lineales generalizados (*Generalized Linear Model* GLM) a los datos sintéticos y comparar los resultados estimados por estos modelos con los datos originales.

Recientemente se han desarrollado nuevos métodos de generación de datos sintéticos a partir de datos originales que utilizando técnicas de programación lineal entera, tratan de preservar ciertas propiedades estadísticas de los datos originales, tales como los momentos[6] (marginales y combinados) de distinto orden. Estos nuevos métodos suponen una alternativa a los métodos basados en modelos de imputación múltiple con modelos paramétricos y no paramétricos. Según un estudio realizado por los desarrolladores de estos métodos (Bogle & Mehrotra, 2016), la calidad de los modelos de regresión que se obtienen a partir de datos sintéticos que se generan mediante ajuste de momentos de cuarto orden de los datos originales, es superior a la de los modelos de regresión obtenidos con datos sintéticos generados con imputación múltiple.

3.4 Otros métodos

Por último, en este grupo se incluyen métodos que podrían ser clasificados dentro de cualquiera de las dos familias de métodos planteadas debido a que son combinación de ambas, por lo que se ha considerado apropiado incluirlos en un grupo separado.

Siempre es posible combinar las distintas técnicas utilizadas en los métodos anteriormente descritos y plantear métodos híbridos de generación de poblaciones. Es el caso del método planteado por Casati, Müller, Fourie, Erath, & Axhausen (2015) , un modelo probabilístico con ajuste de marginales que combina dos de los métodos explicados anteriormente. En una primera fase, utilizan las técnicas del método probabilístico "MCMC Jerárquico" explicado en el punto 3.3.4, para generar una población multinivel, la cual es utilizada en una segunda fase, como una gran muestra a la que aplican la técnica de *Generalized Raking* multiplicativo. En esta segunda fase se asignan pesos a cada uno de los hogares de la población previamente obtenida, de forma que con estos pesos se ajusta la población a los marginales objetivos de los distintos niveles, hogares e individuos.

Los pesos resultantes no son números enteros, por lo que se utilizan como una nueva distribución de probabilidades con la que efectuar el muestreo de Monte Carlo para obtener

[6] Los momentos de un conjunto de datos x_i son medidas de dispersión sobre un determinado valor. El momento de orden *r* respecto del valor *c* de N valores x_i es $\frac{1}{\mathrm{N}} \sum_{\mathrm{i=1\ ...}}^{\mathrm{N}} (x_i - c)^r$

la nueva población, la cual tendrá la heterogeneidad que aporta el método probabilístico de la primera fase, ya que pueden obtenerse hogares no contenidos en la muestra inicial (asumiendo que las probabilidades condicionadas se obtienen de diversas muestras parciales).

Sun & Erath (2015) también combinaron su método de Redes Bayesianas con el de *Generalized Raking*. De este modo consiguen asignar pesos a los agentes y ajustar los marginales de la población de agentes obtenida mediante muestro de la red bayesiana inferida.

4 Marco de Referencia: Posicionamiento de Métodos

4.1 Introducción

Cada vez que se plantea un nuevo método de generación de poblaciones es normal proceder a compararlo con métodos de referencia como el IPF o IPU. Tales comparaciones, en ocasiones, se confeccionan estableciendo criterios heterogéneos y sin un análisis estadístico riguroso, lo cual conduce a la existencia de múltiples comparativas de métodos basadas en criterios dispares de las que difícilmente pueden extraerse conclusiones generales. Con estas afirmaciones no se pretende descalificar estos trabajos, pero sí realizar una llamada de atención sobre esta situación.

Dentro del ámbito de los métodos de generación de poblaciones no existe un marco de referencia, de general aceptación, para realizar de forma efectiva la evaluación comparativa de la exactitud de las poblaciones generadas. Por tal motivo, uno de los objetivos que se han planteado en esta investigación es la definición de un marco de referencia donde posicionar los distintos métodos de generación de poblaciones respecto a los escenarios de uso (necesidades y datos disponibles) y donde establecer una clara metodología para ejecutar análisis comparativos del rendimiento de los métodos.

El marco de referencia que se propone se define basándose en 3 componentes principales:

- Definición de escenarios con distintas problemáticas y necesidades.
- Mapa de Escenarios-Métodos.
- Metodología para llevar a cabo el análisis comparativo de métodos.

Antes de entrar en la descripción de estos componentes se revisan las métricas que se utilizan en la comparación de poblaciones (sección 4.2), así como los principales estudios publicados en relación a la comparación de métodos (sección 4.3), con objeto de verificar la situación descrita.

4.2 Métricas de comparación de poblaciones

No existe una métrica estandarizada para evaluar la bondad del ajuste entre dos poblaciones. Esto se debe a que existe cierta controversia en la comunidad científica al determinar qué métricas son las más apropiadas para la comparación de poblaciones, porque es necesario tener en cuenta posibles variaciones en las distribuciones de las discrepancias (las distribuciones de los errores), y no hay consenso para encontrar el punto de equilibrio para

medir tanto las discrepancias como su distribución (Chai & Draxler, 2014; R Lovelace et al., 2015).

Por este motivo, muchos investigadores utilizan varias métricas para confirmar los resultados de sus estudios de rendimiento.

En esta sección se describen distintas métricas que se utilizan en la literatura. No se trata de una lista exhaustiva, pero sí incluye los principales tipos o familias de métricas de uso más frecuente. Para la descripción de las métricas se va a utilizar la siguiente nomenclatura con objeto de referirnos a dos tablas M-dimensionales de dos poblaciones, una que denominamos "Observada" y otra sintetizada a la que llamaremos "Estimada":

$c_m =$ índice de categoría para el atributo m, $c_1 = 1 \dots \mathbf{C_1}; \dots; \ c_M = 1 \dots \mathbf{C_M}$.

$O_{c_1 \dots c_M} =$ número de agentes en la celda $c_1 \dots c_M$ de la tabla Observada.

$E_{c_1 \dots c_M} =$ número de agentes en la celda $c_1 \dots c_M$ de la tabla Estimada (sintetizada).

$$\mathrm{N} = Núm.\, total\ de\ agentes = N_O = \sum_{c_1=1}^{\mathbf{C_1}} \dots \sum_{c_M=1}^{\mathbf{C_M}} O_{c_1 \dots c_M} = \sum_{c_1=1}^{\mathbf{C_1}} \dots \sum_{c_M=1}^{\mathbf{C_M}} E_{c_1 \dots c_M} = N_E$$

A continuación se describen las distintas métricas:

1. **Error Absoluto Total y Error Absoluto Estandarizado** (TAE y SAE, del inglés *Total Absolute Error* y *Standarized Absolute Error*). Se trata de métricas muy intuitivas. El TAE es la suma las discrepancias absolutas entre las celdas de las dos tablas multidimensionales. Si se normaliza con el total de la población se habla del Error estandarizado (SAE), una medida relativa que a veces también se conoce como Error Absoluto Medio (MAE, del inglés *Mean Absolute Error*):

$$TAE = \sum_{c_1=1}^{\mathbf{C_1}} \dots \sum_{c_M=1}^{\mathbf{C_M}} \left| O_{c_1 \dots c_M} - \ E_{c_1 \dots c_M} \right|$$

$$SAE = TAE/N$$

Si en lugar de comparar discrepancias absolutas se comparan proporciones, lo cual es útil para el caso de poblaciones con distinto tamaño, se tiene el Error Porcentual Absoluto Medio (MAPE, del inglés *Mean Absolute Percentage Error*), a veces llamado *Total Absolute Proportional Error* (TAPE):

$$MAPE = \frac{1}{(\mathbf{C_1} + \mathbf{C_2} + \dots + \mathbf{C_M})} \sum_{c_1=1}^{\mathbf{C_1}} \dots \sum_{c_M=1}^{\mathbf{C_M}} \frac{\left| O_{c_1 \dots c_M} - \ E_{c_1 \dots c_M} \right|}{O_{c_1 \dots c_M}}$$

En muchos estudios se desconocen los valores observados, es decir, las celdas $O_{c_1 \dots c_M}$, y solo se conocen valores agregados de menor dimensionalidad, por lo que el TAE se calcula en relación a estos valores, por ejemplo, si en el caso de 3 atributos, en lugar de conocerse $O_{c_1 c_2 c_3}$ solo se conocen los marginales unidimensionales del primer atributo O_{c_1}y los bidimensionales del segundo y tercer atributo: $O_{c_2 c_3}$, el TAE se calcula como:

$$TAE = \sum_{c_1=1}^{\mathbf{C_1}} \left| O_{c_1} - E_{c_1} \right| + \sum_{c_2=1}^{\mathbf{C_2}} \sum_{c_3=1}^{\mathbf{C_3}} \left| O_{c_2 c_3} - E_{c_2 c_3} \right|$$

En los casos de generación de poblaciones en los que se imponen marginales unidimensionales, es habitual utilizar el TAE unidimensional para determinar cómo se ajusta la población a las restricciones marginales, esto es, medir la diferencia absoluta entre los marginales objetivos de la "*small area*" (datos de entrada) y los marginales de la población sintetizada para dicha área.

En el caso del ejemplo con 3 atributos, si los marginales unidimensionales objetivos son $O_{c_1}, O_{c_2}, O_{c_3}$ $(c_1 = 1 \dots \mathrm{C}_1; c_2 = 1 \dots \mathrm{C}_2;\ c_3 = 1 \dots \mathrm{C}_3$ el TAE y el MAPE se calculan como:

$$TAE = \sum_{i=1}^{3} \left(\sum_{c_i=1}^{\mathbf{C_i}} \left| O_{c_i} - E_{c_i} \right| \right) \ ; \ \ MAPE = \frac{1}{(\mathbf{C_1}+\mathbf{C_2}+\mathbf{C_3})} \sum_{i=1}^{3} \left(\sum_{c_i=i}^{\mathbf{C_i}} \frac{\left| O_{c_i} - E_{c_i} \right|}{O_{c_i}} \right)$$

2. **Error Cuadrático Medio**. (RMSE y SRMSE, del inglés *Root Mean Square Error* y *Standardized Root Mean Square Error*): RMSE es una de las métricas más utilizada en los estudios de ciencias sociales y en generación de poblaciones. Esta métrica divide la suma de las discrepancias al cuadrado entre el número total de agentes (N) o el número total de celdas. Se calcula como:

$$RMSE = \sqrt{\frac{1}{N} \sum_{c_1=1 \ \dots \ c_M=1}^{\mathbf{C_1}} \sum^{\mathbf{C_M}} (E_{c_1 \dots c_M} - O_{c_1 \dots c_M})^2}$$

En otras ocasiones se normaliza con el número de celdas $(\mathbf{C_1} * \mathbf{C_2} * \dots * \mathbf{C_M})$ en lugar de con el total de agentes. Algunos investigadores (Müller & Axhausen, 2011) (Pritchard & Miller, 2012) utilizan este valor normalizado con el número de celdas, y estandarizado en relación al número total de agentes entre el número de celdas $(\mathbf{C_1} * \mathbf{C_2} * \dots * \mathbf{C_M})$:

$$SRMSE = \frac{\sqrt{1/(\mathbf{C_1} * \mathbf{C_2} * \dots * \mathbf{C_M}) \ \sum_{c_1=1 \ \dots}^{C_1} \sum_{c_M=1}^{\mathrm{C_M}} (E_{c_1 \dots c_M} - O_{c_1 \dots c_M})^2}}{N/(\mathbf{C_1} * \mathbf{C}_2 * \dots * \mathbf{C_M})}$$

Otros investigadores prefieren utilizar el MSE, sin normalizar y eliminando la raíz cuadrada (L. Ma, 2011).

3. **Estadísticos Z**. Existe el Z-score y algunas variantes. El Z-score es una medida de distancia entre puntos expresada en unidades de desviación estándar. El Z-score para un punto de una tabla se define como:

$$Z = \frac{\dot{E}_{c_1 \dots c_M} - \dot{O}_{c_1 \dots c_M}}{\sqrt{\frac{\dot{O}_{c_1 \dots c_M}(1 - \dot{O}_{c_1 \dots c_M})}{N}}}$$

Y la suma de Z-score cuadrado (SSZ, del inglés *Sum of Square Z scores*) se define como:

$$SSZ = \sum_{c_1=1}^{\mathbf{C_1}} \dots \sum_{c_M=1}^{\mathbf{C_M}} \frac{(\dot{E}_{c_1 \dots c_M} - \dot{O}_{c_1 \dots c_M})^2}{\left(\frac{\dot{O}_{c_1 \dots c_M}(1 - \dot{O}_{c_1 \dots c_M})}{N}\right)}$$

El punto sobre la O y el E, indican que son probabilidades, es decir: $\dot{E}_{c_1 \dots c_M} = E_{c_1 \dots c_M}/N$ y $\dot{O}_{c_1 \dots c_M} = O_{c_1 \dots c_M}/N$

Para el caso en que la población observada tengan distintos tamaño que la estimada $N_O \neq N_E$, se propuso el SSZ^* modificado (Voas & Williamson, 2001), el cual coincide con el anterior SSZ si el número de agentes de las dos poblaciones es el mismo:

$$SSZ^* = \sum_{c_1=1}^{\mathbf{C_1}} \dots \sum_{c_M=1}^{\mathbf{C_M}} \frac{\left(E_{c_1 \dots c_M}/N_E - O_{c_1 \dots c_M}/N_O\right)^2}{\left(\frac{O_{c_1 \dots c_M}(1 - O_{c_1 \dots c_M}/N_O)}{N_O N_E}\right)}$$

El algoritmo de *Simulated Annealing* de Williamson, utiliza esta métrica como alternativa al TAE para la función de bondad de ajuste (GoF).

4. **Medidas del desajuste**. Se trata de métricas muy simples, básicamente son dos: el porcentaje de celdas que son distintas a lo esperado (*Non Fitting Cells* NFC) y la proporción de celdas con una diferencia respecto a lo esperado superior a un determinado porcentaje (1% por ejemplo, E>1%).

   En algunas ocasiones, el NFC se determina calculando el estadístico Z-score de la celda, y si es mayor que un valor crítico (1,96 para un nivel de confianza del 95%, α=0,05), se concluye la celda no es conforme al valor esperado. De igual modo, también se determina si la tabla completa no está ajustada (*Non Fitting Table*, NFT), si la suma de todos los Z^2 de todas las celdas de la tabla $\sum Z^2$es mayor que un valor crítico de la distribución $\boldsymbol{\chi^2}$.

   La métrica del E>5%, junto con la RMSE, son dos de las métricas más utilizadas (R Lovelace & Ballas, 2013) (L. Ma, 2011). La E>5% representa la proporción de valores que están alejados

más del 5%, de lo observado. Esta métrica favorece a las poblaciones donde haya unas pocas celdas con valores grandes muy desviados de lo observado frente a otras poblaciones que tienen muchas celdas con pequeñas desviaciones (pero mayores del 5%).

5. **Pendiente de la línea de ajuste y coeficiente de correlación R^2.** Un modo sencillo de comparar las dos poblaciones es representar el valor de cada celda observada $O_{c_1 \dots c_M}$ en el eje de abscisas y el valor de cada celda equivalente estimada $E_{c_1 \dots c_M}$ en ordenadas. De este modo, por cada celda se tendrá un punto. Con la nube de puntos correspondiente a todas las celdas se construye una recta de regresión, y = m x. Si las tablas son iguales los puntos se ajustarán a una línea y = x, en caso contrario, la línea de ajuste tendrá una pendiente distinta de 1, por lo que la pendiente es una medida de la proximidad entre las poblaciones. Como medida de la bondad de ajuste de esta línea de regresión a los puntos de la nube, se calcula el coeficiente de determinación lineal R^2, que es el cociente entre la dispersión de las ordenadas respecto a su media y la dispersión de las abscisas respecto a su media, es decir:

$$R^2 = \frac{\sum_{c_1=1 \dots}^{\mathbf{C_1}} \sum_{c_M=1}^{\mathbf{C_M}} (E_{c_1 \dots c_M} - \bar{O}_{c_1 \dots c_M})^2}{\sum_{c_1=1 \dots}^{\mathbf{C_1}} \sum_{c_M=1}^{\mathbf{C_M}} (O_{c_1 \dots c_M} - \bar{O}_{c_1 \dots c_M})^2}$$

   El coeficiente varía entre 0 y 1. Un valor de 1 indica que ambas poblaciones son iguales, y si es próximo a cero indica que la población estimada $E_{c_1 \dots c_M}$no es capaz de reproducir las desviaciones de la población observada $O_{c_1 \dots c_M}$.

   El valor de R^2 (de este modelo de regresión lineal) es el cuadrado del **coeficiente de correlación de Pearson *r*** que mide el grado de asociación lineal entre las dos variables, en este caso las celdas de E y las de O, el cual varía entre -1 y 1.

6. **Error de clasificación (%CE)**. Se trata de una medida relativa de uso generalizado (Harland et al., 2012) (Voas & Williamson, 2001) fácil de calcular y de interpretar. Representa el porcentaje de agentes de la población que no están clasificados correctamente en la celda correcta de la tabla, sin tener en cuenta la distancia entre cada agente de la tabla observada, es decir, con independencia del número de categorías asignadas de forma incorrecta.

$$\%CE = \frac{\sum_{c_1=1 \dots}^{\mathbf{C_1}} \sum_{c_M=1}^{\mathbf{C_M}} \left| O_{c_1 \dots c_M} - E_{c_1 \dots c_M} \right|}{2\,N} 100 = \frac{TAE}{2\,N} 100$$

   Al igual que el TAE, en función de la dimensionalidad de la tabla, se pueden calcular %CE de distinta dimensionalidad. El anterior corresponde al M-dimensional y el más básico es el 1-dimensional, que indica el porcentaje de ajuste de los marginales unidimensionales. Algunos investigadores usan el PGP (*Proportion of Good Prediction*) que es $PGP = 1 - CE$ (Lenormand & Deffuant, 2013).

7. **Chi-cuadrado / χ^2.** La Chi-cuadrado de Pearson mide la discrepancia entre dos distribuciones:

$$\chi^2 = \sum_{c_1=1}^{\mathbf{C_1}} \dots \sum_{c_M=1}^{\mathbf{C_M}} \frac{(O_{c_1 \dots c_M} - E_{c_1 \dots c_M})^2}{E_{c_1 \dots c_M}}$$

   El principal inconveniente se tiene cuando los valores $E_{c_1 \dots c_M}$ son cero o muy pequeños, con lo que esta medida no se recomienda para tablas con ceros o que contengan valores extremadamente pequeños, por lo cual se han propuesto alternativas, como sustituir el denominador por los valores promedios entre los observados y estimados, o 1 en caso de que ambos sean 0 (estadístico de Neyman).

8. **Métricas basadas en la Entropía**. La entropía se usa para determinar el grado de variabilidad de una distribución de probabilidad, especialmente cuando las variables son categóricas, ya que en estos casos no se puede calcular la media ni la desviación estándar, y en su lugar se mide la entropía. La expresión más extendida de la entropía para casos discretos es la de Shannon:

$$E = -\sum_{c_1=1}^{\mathbf{C_1}} \dots \sum_{c_M=1}^{\mathbf{C_M}} O_{c_1 \dots c_M} \log O_{c_1 \dots c_M}$$

   Basado en este concepto se define la divergencia de Kullback-Leibler, también llamada entropía relativa, que nos mide la similitud que existe entre dos funciones de distribución de probabilidad, que aplicado a nuestro caso de poblaciones es:

$$D_{KL} = \frac{1}{N} \sum_{c_1=1}^{\mathbf{C_1}} \dots \sum_{c_M=1}^{\mathbf{C_M}} O_{c_1 \dots c_M} \log \frac{O_{c_1 \dots c_M}}{E_{c_1 \dots c_M}}$$

9. **Razón de Verosimilitud $\mathbf{G^2}$ (*Likelihood ratio*).** Este estadístico es muy habitual como medida de la bondad de ajuste entre tablas de contingencia:

$$G^2 = 2 \sum_{c_1=1}^{\mathbf{C_1}} \dots \sum_{c_M=1}^{\mathbf{C_M}} O_{c_1 \dots c_M} \log(\frac{O_{c_1 \dots c_M}}{E_{c_1 \dots c_M}})$$

   Tiene el mismo inconveniente que el χ^2(problemas con los valores 0).

   Dividiendo esta métrica por $2N$, resulta ser equivalente a la entropía relativa entre las dos tablas (Voas & Williamson, 2001).

10. **Freeman–Tukey**. Métrica que se utiliza para llevar acabo test estadísticos de verificación de la independencia entre tablas de datos.

$$T^2 = 4 \sum_{c_1=1}^{\mathbf{C_1}} \dots \sum_{c_M=1}^{\mathbf{C_M}} (\sqrt{E_{c_1 \dots c_M}} - \sqrt{O_{c_1 \dots c_M}})^2$$

Una variante de esta métrica, muy usada para comparar distribuciones de probabilidad, es la distancia de Hellinger, utilizada por Ma (2011):

$$H^2 = \frac{1}{2} \sum_{c_1=1}^{\mathbf{C_1}} \dots \sum_{c_M=1}^{\mathbf{C_M}} (\sqrt{E_{c_1 \dots c_M}/N_E} - \sqrt{O_{c_1 \dots c_M}/N_O})^2$$

Estas métricas no tienen el problema de estabilidad con los valores 0.

11. **Error de Clasificación de atributos.** Una métrica propuesta por Otani, Sugiki, & Miyamoto (2012) específicamente para comparar poblaciones de agentes. La idea que subyace en esta métrica es la necesidad de tener en cuenta cuan bien caracterizados están los agentes de la población, considerando el tipo y número de diferentes valores de atributo cuando se clasifican en una tabla. En primer lugar determinan las distancias entre cada agente de la población estimada y cada uno de la población de referencia, basándose en la diferencia de los valores de cada uno sus atributos. La métrica total se obtiene determinando la mínima suma de distancias para todos los agentes de la población.

    A diferencia del %CE (Error de Clasificación) que solo tiene en cuenta si un agente está mal clasificado, con independencia del número de categorías erróneas, esta medida si tiene en cuenta para cada agente el número de categorías incorrectamente asignadas.

    El problema con esta métrica es que el número de cálculos aumenta en proporción al factorial de la cantidad de agentes de la población (N!). Por lo tanto, a pesar de la precisión que aporta esta métrica, se trata de una métrica poco práctica, de difícil utilización y que requiere de algoritmos especialmente complejos (algoritmos genéticos evolutivos) para su determinación.

12. **Índice de Similitud de Bray-Curtis.** Se trata de una métrica para comparación de poblaciones con distinto tamaño, esto es, $N_O \neq N_E$.

    Esta métrica se calcula como:

$$BCI = \frac{\sum_{c_1=1 \dots}^{\mathbf{C_1}} \sum_{c_M=1}^{\mathbf{C_M}} |O_{c_1 \dots c_M} - E_{c_1 \dots c_M}|}{N_O + N_E} 100$$

    En el caso de que las poblaciones tengan igual tamaño coincide con el error de clasificación *%CE*. Y al igual que este, es una métrica muy intuitiva y de fácil interpretación. Un valor 0

corresponde a dos poblaciones con tablas de contingencia idénticas, y un valor 1 a tablas totalmente disjuntas.

4.3 Estudios comparativos entre métodos

Tal como indicaron Hermes & Poulsen (2012) la falta de conocimiento sobre la fiabilidad de las poblaciones generadas obstaculiza muchas potenciales aplicaciones de los microdatos espaciales sintéticos. Es preciso abordar nuevos estudios de investigaciones que establezcan los métodos más adecuados para cada aplicación. (Clarke y Harding, 2013).

En relación a los estudios de investigación llevados a cabo con tal fin, la mayoría presentan resultados de aplicación de las poblaciones sintéticas y pocos se centran en un verdadero análisis comparativo entre métodos.

Entre los estudios comparativos cabe distinguir los de carácter cualitativo o metodológico, y los cuantitativos. Los primeros se limitan a enumerar propiedades de los métodos, así como ventajas y desventajas, y son poco frecuentes. Un ejemplo de este tipo es el de (Rahman, 2009) donde se describe el funcionamiento de los métodos GREGWT y SA, centrándose en la comparativa de los aspectos metodológicos de los mismos.

A nivel cuantitativo, a medida que han ido apareciendo nuevos métodos se han realizado estudios comparativos, cuyo objetivo ha sido la comparación del rendimiento entre el nuevo método y algunos de los métodos de referencia descritos en el capítulo anterior.

En la Tabla 4 se muestran los principales estudios publicados, de carácter cuantitativo, junto con algunas de sus características. Entre las características consideradas se encuentra: el tipo de población utilizada en los experimentos (multinivel o no), el conjunto de métricas utilizadas, el número de celdas de la tabla de población que se compara, la dimensionalidad con la que se hace la comparación, el número de atributos de la tabla utilizada en la comparación, el número de muestras utilizadas y el número de "*small area*s" que se comparan (número de ensayos). Se remite al lector a la anterior sección 4.2 en la que se describe las métricas y la notación que aparece en esta tabla.

Cada estudio enfrenta diferentes métodos utilizando distintos criterios de comparación, tanto a nivel de métrica como de dimensionalidad de la misma. Así mismo, en cada estudio se comparan poblaciones de número variable de "*small area*s", algunos con una única "*small area*" y otros con más de 100.

En pocos estudios se utilizan múltiples muestras, y cuando se usan varias, se trata de muestras de distinto tamaño para analizar el comportamiento del método con los distintos tamaños de muestra (Farooq et al., 2013; Ryan, Maoh, & Kanaroglou, 2009; Saadi et al., 2016;

Sun & Erath, 2015). Solo Ryan et al. (2009) tras hacer una primera comparativa entre métodos usando 2 muestras aleatorias de cada tamaño (4 tamaños diferentes) efectúa una comparativa exhaustiva, usando 50 muestras aleatorias del mismo tamaño (5%) para hacer un análisis comparativo fiable, aunque siempre utilizando la misma "*small area*".

ESTUDIO COMPARATIVO	Multi-nivel	Métrica	Max Núm. Celdas	Núm. atrib.	Max. Dimen. validac.	Muestras/ áreas(ensayos)
IPF-Beckman vs IPF-CR Auld et al (2008)	No	SAE	-	5	1-dim	1/1
IPF vs SA (Ryan et al., 2009)	No	FT^2	51.816	3	3-dim	8/1 (2) ; y 50/1
FBS vs IPF vs IPU (L. Ma, 2011)	Si	MAE,MSE,H^2,E >5	8.529	Hog:3 Ind:3	6-dim	1/22
HIPF vs ENT vs IPU (Müller & Axhausen, 2011)	Si	SRMSE,G^2	16.016	Hog:7 Ind:7	3-dim	1/23
Cond. Prob. vs DR vs SA (Harland et al., 2012)	No	%CE	42	6	1,2 & 3-dim	1/300 y 1/2.439
SR vs SA (Williamson, 2013)	Si	TAE, RSSZ*, NFT,NFC	180	Hog:4 Ind:4	3-dim	1/86 (100)
MCMC vs IPF (Farooq et al., 2013)	No	RMSE,R^2, pendiente	384	4	4-dim	5/1
SR vs IPU (Lenormand & Deffuant, 2013)	Si	CE/PGP	147 HH+I 17+130	Hog:2 Ind:3	3-dim 2-dim	1/1.310 (10)
GREGWT vs SA (Tanton et al., 2014)	Si	TAE,TAPE, SSZ*	32	Hog:8 Ind:4	3-dim	1/307
IPF vs FBS (L. Ma & Srinivasan, 2015)	No	MSE,MAE, Hellinger, E>5	84	Hog:3 Ind:3	3-dim	1/12
KNN vs IPU (Hamada et al., 2015)	No	Hellinger	-	6	6-dim	1/10
MCMC vs BN vs IPF (Sun & Erath, 2015)	No	SRMSE	32.256	7	7-dim	20/1
IPF vs GREGWT (Muñoz et al., 2015)	No	TAE,SAE, χ^2		3	3-dim	1/4
IPF vs SA (Kim & Lee, 2016)	No	RMSE,SAE,r, χ^2	24	2	2-dim	1/1
HMMC vs IPF (Saadi et al., 2016)	No	SRME, pendiente, R^2	19.264	6	3-4-5-6-dim	2/589
IPFSR vs SA vs SFF vs MCMC vs JDI (P Ye et al., 2017)	No	MAE, RMSE	58	6	1-dim	1/1
GR vs HIPF vs EO vs IPU (Müller, 2017a)	Si	G^2, SSE (vs pesos unidad)	16	Hog:2 Ind:2	4-dim	7/7
Pop-H vs IPU (Zhuge et al., 2017)	Si	MAPE	4	Hog:2 Ind:2	1-dim	1/1

Tabla 4 Estudios Comparativos.

Lo más habitual es que en estos estudios se analicen casos específicos donde solo se utiliza una muestra de la población, por lo que dichos estudios, más que estudios comparativos, pueden considerarse comparaciones de casos puntuales que no prestan atención a la incertidumbre que se deriva del uso de una muestra.

Algunos estudios comparan métodos que utilizan diferentes datos de entrada, por lo que al no compararse en igualdad de condiciones, los resultados no son concluyentes. Es el caso de la comparativa IPU vs Reconstrucción Sintética sin muestra de Lenormand & Deffuant (2013), utilizando una población descrita por 2 atributos de hogar y 3 de individuo. El IPU utiliza una muestra del 25% de la población y Reconstrucción Sintética utiliza probabilidades calculadas a partir de 6 distribuciones de atributos de la población (1 unidimensional, 2 bidimensionales y 3 tridimensionales).

Otro caso de este tipo de comparativa es el planteado en el estudio de Farooq et al. (2013) MCMC vs IPF, en el que se generan poblaciones con MCMC utilizando las distribuciones de probabilidad condicionadas de la población, mientras que el IPF se usa con muestras de distinto tamaño. Este estudio presenta unos resultados en los que el error cuadrático medio SRMSE de la población generada mediante IPF con una muestra del 20% es peor que el SRMSE de cualquiera de las poblaciones obtenidas con MCMC usando distribuciones de probabilidades condicionadas parciales (para este caso con 4 atributos: usando 3 distribuciones de probabilidad condicionada con 2 atributos, y 1 distribución condicionada con 3 atributos). Lo cual podría calificarse de “esperado”, dada la diferencia entre la información contenida en los conjuntos de datos de entrada utilizados por cada método (distribuciones de probabilidad exactas y muestra del 20%).

En otras ocasiones se comparan poblaciones obtenidas con métodos que utilizan poblaciones con estructuras de distintos niveles, como son, IPF vs FBS. El IPF solo opera con la estructura del nivel de hogar (L. Ma & Srinivasan, 2015) por lo que solo impone marginales a nivel hogar, mientras que el FBS impone marginales tanto a nivel de hogar como de individuo.

Otro de los aspectos a tener en cuenta en los estudios es la aleatoriedad de los resultados. Casi todos los métodos utilizan técnicas estocásticas. Estas técnicas no solo son utilizadas por los métodos probabilísticos como el MCMC o la Reconstrucción Sintética, sino también por el SA y todos los métodos de *reweighting* que utilizan muestreo de Montecarlo. Con el fin de reducir el impacto de la aleatoriedad de los mismos, los experimentos de generación de poblaciones deben repetirse varias veces en función de la variabilidad (varianza), y posteriormente calcular el valor promedio para obtener los resultados finales. La mayoría de los estudios no indican explícitamente si realmente llevan a cabo la repetición de experimentos. En los casos en que si se indica, aparece este valor entre paréntesis en la cuarta columna de la Tabla 4.

Las poblaciones utilizadas en los estudios de la tabla están descritas con distinto número de atributos y categorías, lo cual dificulta la comparación entre los resultados de dichos estudios. Además, las escalas de los mismos presentan grandes diferencias. Alguna de las comparativas se hacen usando poblaciones de "*small area*" a muy gran escala, como es el caso de (P Ye et al., 2017) donde se genera la población sintética de China. Sin embargo, en otras de las comparativas se utilizan poblaciones muy pequeñas, como es el caso de Müller con poblaciones de menos de 400 hogares.

En varios de los estudios, la comparativa se reduce a determinar el método que mejor replica los marginales objetivos en la población generada, es decir, a una comprobación de que la población generada se ajusta a las restricciones marginales impuestas. Es el caso de los estudios cuya máxima dimensión de validación es 1-dim.

También se observa que son minoría los estudios comparativos con métodos de poblaciones multinivel, siendo más frecuentes las comparativas entre métodos con poblaciones de un único nivel: IPF vs GR Linear, (Muñoz et al., 2015); SA vs SR, (Williamson, 2013); "k-vecinos más cercanos" vs IPU, (Hamada et al., 2015); (Ryan et al., 2009), (Harland et al., 2012).

A la luz de los estudios comparativos publicados se concluye la necesidad de un esquema común que facilite las comparativas globales. Este esquema debe tener en cuenta a todos los tipos de métodos, utilizar un conjunto de datos homogéneo y emplear un mismo esquema de comparación de poblaciones basado en una misma métrica. Un esquema unificado aportaría un conjunto claro de directrices en relación al uso comparado de los métodos. Es por ello que se plantea este marco de referencia dentro del que encuadrar dicha comparativa global.

4.4 Escenarios

Ante la necesidad de disponer de una población de agentes pueden presentarse distintos escenarios. Estos escenarios se caracterizan en relación al uso de dicha población y a los datos disponibles. A continuación se analizan los factores que intervienen en estos escenarios.

4.4.1 Población con un nivel o multinivel

El uso de poblaciones con una jerarquía de niveles, hogares e individuos, mejora la calidad de los resultados de las aplicaciones que las utilizan. Es el caso de la comparativa entre los resultados de dos microsimulaciones espaciales llevadas a cabo por Fenton (2016). En la primera utilizó IPF para calcular la distribución de ingresos de los hogares de 33 barrios (*local authority areas*) de Londres en 2001 y 2011, modelando una población compuesta por hogares descritos con 5 atributos. En la segunda utilizó HIPF considerando, además de los

atributos indicados, 4 atributos adicionales para los individuos que componían los mencionados hogares. En ambos casos utilizó los datos de las encuestas gubernamentales FRS (*Family Resources Survey*) y HBAI (*Households Below Average Incomes*) para determinar la distribución de la pobreza en los distintos barrios de Londres. Con dicho estudio observó que la segunda microsimulación mejoraba la estimación de los índices de pobreza, acercándose más a los datos publicados por otras fuentes. El estudio concluyó que la microsimulación espacial con un solo nivel subestimaba la pobreza comparándola con los datos de HBAI y de otras fuentes de datos, mientras que la microsimulación multinivel producía mejores estimadores.

No solo en casos de microsimulación, sino también en los modelos de SBA, se requiere disponer de una población multinivel en función de los objetivos de la propia simulación. En los modelos SBA, además de los agentes y del propio entorno, hay ocasiones en las que han de tenerse en cuenta entidades de otro nivel, denominadas “colectivos” (Grimm et al., 2010), que pueden interactuar con otros agentes o grupos “colectivos”, o verse afectados por factores del entorno externo. Estos “colectivos” representan grupos de agentes con un comportamiento y unas reglas propias del grupo; por ejemplo, aparte de hogares de individuos, podrían tratarse de grupos sociales de agentes, empleados de una empresa, individuos de una organización o red social, etc., que llevan a cabo actividades de grupo como hacer un paro laboral, un boicot, ejercer una presión de grupo determinada sobre otros agentes o ejecutar acciones que solo se realizan en grupo. En general, estos “colectivos” generalmente se caracterizan por la lista de sus agentes, por atributos específicos del colectivo y acciones específicas que solo son realizadas por el colectivo.

Por otro lado, el comportamiento del agente puede estar condicionado por el colectivo al que pertenece. En los modelos de demanda de viajes basados en actividad ya se indicó la importancia que tienen los atributos del hogar al que pertenece el individuo (tamaño, ingresos, número de vehículos, etc.) para determinar el plan de viajes de dicho individuo.

El uso de poblaciones multinivel no solo mejora la calidad de los estimadores de las aplicaciones, sino que amplía el número de atributos que pueden utilizarse en las simulaciones y microsimulaciones (en los casos de que estén disponibles valores marginales de estos atributos). Por otro lado, produce una población más próxima a la "realidad", lo que permite elaborar mejores pruebas de simulación de distintas políticas de actuación sobre la población.

En general, la necesidad de poblaciones multinivel es uno de los escenarios básicos que se presenta en un alto porcentaje de situaciones de generación de poblaciones sintéticas.

4.4.2 Problema de Población de "small area" y Población Completa

Uno de los factores a tener en cuenta a la hora de generar una población mediante una muestra es el tipo de aplicación para el que se precisa la población. Básicamente pueden distinguirse dos tipos:

- Generación de una población de "*small area*" basándose en una muestra de la población real de una región que incluye la "*small area*" e información sobre la distribución marginal de los atributos en la "*small area*". En el caso extremo, puede tratarse de generar la población de todas las "*small area*s" contenidas en la región. En este tipo de casos, el tamaño de la muestra (*n*) puede ser menor que el tamaño de la población (*N*) a generar *n≤N*, situación *A* de la Figura 13, o puede ocurrir que *n>N,* reflejada en la situación *B.*
- Generación de una población completa de un área determinada, basándose en una muestra de la población real de dicha área y la información de distribución de marginales de los atributos en dicha área. En esta clase de problemas el tamaño de la muestra siempre será menor (o igual) que el tamaño de la población a generar *n≤N* (situación *C* de la Figura 13).

Estos dos tipos de aplicación corresponden a dos tipos de problema diferentes cuya principal diferencia es el tamaño relativo de la muestra y la población sintetizada, ya que el tamaño de la muestra n>N solo ocurre en el primer tipo, lo cual afecta en gran medida al resultado, con independencia del método que se utilice para generar la población. El resultado se verá afectado en mayor grado si el método que se utiliza está diseñado para ser utilizado con una muestra aleatoria, que en el caso de *n>N*, al tener la muestra mayor tamaño que la población, es un requisito difícil encajar.

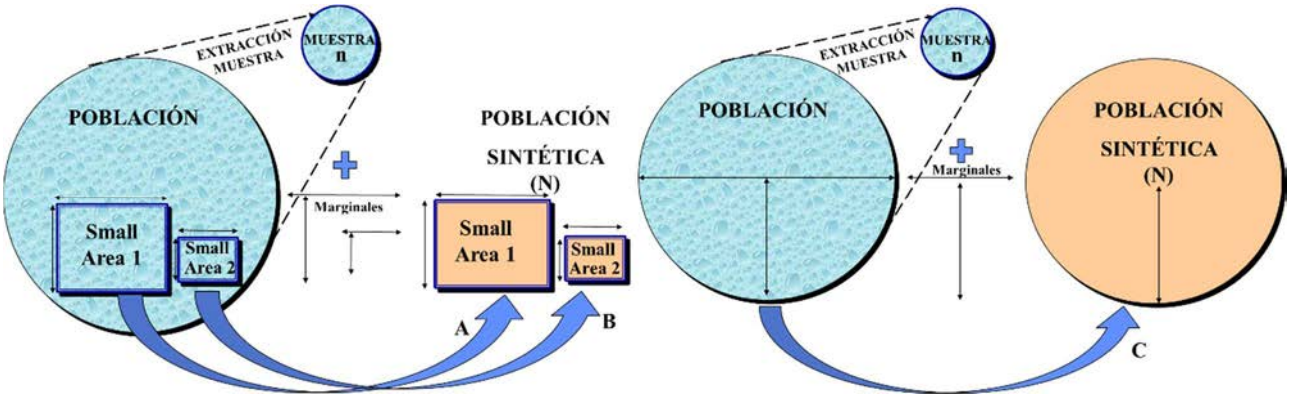

Figura 13 Casos de generación de poblaciones sintéticas de "small area" y completa.

Si se utilizan métodos de generación tipo *reweighting,* que replican los agentes de la muestra para sintetizar la población, y se trata de un problema de "*small area*" (*A* o *B*), las poblaciones que se sintetizan presentan una clase de errores que no suceden si el problema es tipo

población completa (*C*). Nos referimos a que puede ocurrir que se generen agentes sintéticos con combinaciones de atributos que realmente no existen en la población de la "*small area*". Este tipo de error no ocurre en el caso C donde todos los tipos de agentes sintéticos que se crean existen en la población completa, ya que provienen de la muestra de dicha población.

En general, los errores de las poblaciones sintetizadas con métodos de *reweighting* a partir de una muestra de la población provienen de 3 fuentes:

- La muestra no incluye agentes que representan a todos los tipos de agentes de la población real. En este caso, en la población sintetizada faltarán agentes con las combinaciones de atributos que no están representados en la muestra.
- La muestra incluye agentes que no existen en la población real de la "*small area*". En este caso, en la población sintetizada podrá contener agentes con combinaciones de atributos que no deberían existir.
- El método no tiene la capacidad de generar el número correcto de cada tipo agentes (cada combinación de atributos) por lo que crea un número de agentes diferente al real.

Lo habitual es encontrarse con errores derivados de la primera y tercera fuente. La segunda fuente solo se presenta con problemas tipo "*small area*" (*A* y *B*), siendo más probable que se presenten en los tipo *B*. Por tanto, a igualdad de método de generación y tamaño de muestra, hay más posibles fuentes de error de las poblaciones sintetizadas en problemas tipo "*small area*" que en problemas tipo población completa.

Si no nos limitamos a métodos de *reweighting* y consideramos métodos probabilísticos aparece una nueva posible fuente de error, similar a la tercera, que proviene del propio método probabilístico, ya que estos métodos pueden generar agentes imposibles, que no existen realmente en la población real.

Por todo lo dicho, es importante utilizar el mismo tipo de problema para comparar métodos, en caso de comparar problemas distinto tipo pueden presentarse situaciones incoherentes. Así por ejemplo, puede darse que un método X sea superior a otro Z (con problemas del mismo tipo) por lo que las poblaciones generadas con X serán más similares a la real, sin embargo, si el método X se aplica a un problema de "*small area*" puede obtenerse una población menos precisa que la población generada con el método Y aplicado a un problema de población completa, usando la misma muestra.

Una correcta comparativa entre métodos debe de medir la precisión de las poblaciones generadas en las mismas circunstancias, por lo que ha de utilizarse el mismo tipo de problema y la misma información de partida para llevarla a cabo. Si se comparan resultados de

problemas de "*small area*" con resultados de problemas de población completa, aunque la muestra sea la misma en ambos, la comparativa será irrelevante.

Por otro lado, si dos métodos utilizan distinta información de partida, por ejemplo, uno utiliza distribuciones de probabilidad y otro una muestra, la comparativa de los resultados estará en función de la calidad de la información de partida utilizada por cada método, y la comparativa no será concluyente.

Es importante que la comparación de métodos se realice utilizando los mismos datos en los métodos que se comparan. Esta recomendación no siempre se tiene en cuenta, tal como el caso, ya comentado, de la comparativa entre MCMC e IPF de Farooq et al. (2013) para obtener poblaciones completas de la parte occidental de la ciudad suiza de Lausana, con probabilidades condicionadas y con muestra del 20% de dicha población.

Si la comparativa se hubiera realizado con MCMC utilizando distribuciones condicionadas con los atributos de una región más amplia (toda la zona metropolitana de la cuidad o del cantón), posiblemente el resultado habría sido diferente.

En las comparativas entre métodos que utilizan muestras, si las muestras tienen distintas características (por ejemplo, distinto tamaño), los resultados de la comparativa tampoco serán concluyentes. Lo adecuado es comparar poblaciones generadas con múltiples muestras de iguales características.

Por otro lado, si la comparativa entre métodos se plantea con el mismo problema tipo "*small areas*" y ambos métodos utilizan la misma muestra, es también importante utilizar la misma "*small area*", ya que la muestra puede ser representativa para una "*small area*" determinada, pero no para otra. A medida que el tamaño de la "*small area*" disminuye, la representatividad de la muestra para la población de la "*small area*" es menor. Lo recomendable es comparar poblaciones generadas para múltiples "*small areas*" con distinto tamaño y posteriormente combinar los resultados de forma estadística.

4.4.3 Disponibilidad y Calidad de los datos de partida

Los posibles escenarios, en cuanto a disponibilidad de información para generar datos sintéticos, son muy variados, ya que en cada situación los datos disponibles y la calidad de los mismos son diferentes. Las distintas disponibilidades son el resultado de la combinación de las características de la muestra, de los datos agregados y de los atributos. Entre estas características están:

- Muestra: No hay muestra; muestra de baja calidad (probabilística: aleatoria, estratificada, etc., o no probabilística); muestra con pocos atributos; múltiples

muestras parciales (una o varias; de igual o distinto tamaño; con atributos comunes o no); muestra completa con todos los atributos; tamaño de la muestra; todos los datos originales completos.

- Datos estadísticos agregados: se distingue entre distribuciones de probabilidad y distribuciones de marginales.
  - Distribuciones de probabilidad multidimensional: Probabilidades condicionadas completas; probabilidades condicionadas parciales; probabilidades marginales.
  - Distribuciones de marginales: Sin marginales; marginales incompletos; marginales unidimensionales, marginales de distinta dimensionalidad.
- Atributos: Pocos (baja dimensionalidad); muchos (alta dimensionalidad); con pocas categorías; con muchas categorías; atributos no discretos; relaciones no lineales entre atributos; mix de atributos discretos y continuos; propiedades de los atributos (correlaciones, modas, medias, medianas, varianzas, valores máximos y mínimos, modelos de regresión, momentos, etc.).

A continuación se citan algunas de las situaciones más representativas que pueden presentarse en relación a la disponibilidad y calidad de los de los datos. Estas situaciones son el resultado de combinación de las anteriores características de los datos para la generación de poblaciones sintéticas:

- No se dispone de muestra con todos los atributos de la población. Solo existen datos estadísticos agregados de los atributos, a nivel unidimensional (con un solo nivel o multiniveles jerárquicos de la población).
- No se dispone de una muestra con todos los atributos de la población. Pero existen distribuciones parciales de baja dimensionalidad (bidimensionales o tridimensionales) de atributos de la población. Puede estar disponible información de distribución de marginales o no.
- Se dispone de múltiples muestras parciales con parte de los atributos de la población. También se conocen los marginales de los atributos (con un solo nivel o multiniveles jerárquicos de la población).
- Se dispone de datos de probabilidades condicionadas (completas o incompletas), y valores marginales de algunos atributos (con un solo nivel o multiniveles jerárquicos de la población).
- Se dispone de distribuciones marginales de algunos o todos los atributos con distinta dimensionalidad (unidimensionales, bidimensionales, tridimensionales, etc.).
- Se dispone de una muestra en la que faltan agentes de alguna categoría y se conocen los marginales.

- Existe una muestra con todos los atributos de la población (con un nivel o multiniveles jerárquicos) y se conocen los valores marginales (o solo se conocen algunos o ninguno). Los microdatos de las encuestas proporcionados por los organismos oficiales son habitualmente muestras aleatorias (caso del censo) o el 100% de los datos de la encuesta, anonimizados en cualquier caso.
- Se dispone de las correlaciones entre atributos de la población, o modelos de regresión de los atributos, sin información de marginales o de otro tipo como datos de distribuciones de atributos.

En referencia a la calidad de los datos disponibles, puede afirmarse que la calidad de las poblaciones sintéticas está en función de la calidad de los datos que se utilicen para generarlas. Así quedó de manifiesto en uno de los estudios de microsimulación espacial realizado por Hermes & Poulsen (2012b). En dicho estudio se estimó por duplicado la tasa de fumadores de los barrios del área del Gran Londres (33 *local authorities* con una población de más de 8 millones) generando con *Simulated Annealing* la población de cada barrio por duplicado, en base a dos encuestas de diferente calidad y los marginales de cada barrio.

La primera encuesta era la más completa, *General Household Survey* (GHS) correspondiente a los años 2000-01 y 2001-02, y la segunda, *British Household Panel Survey* (BHPS) del 2001, procedía de un estudio longitudinal de años anteriores, lo que la hacía menos representativa y donde las minorías étnicas no estaban representadas con exactitud. Ambas encuestas contenían datos socio- demográficos, incluyendo información sobre el consumo de tabaco, y fueron reducidas a los hogares localizado en el área del Gran Londres, la primera tenía un tamaño de muestra de 4.527 hogares y la segunda de 1.232.

Con las poblaciones generadas con ambas encuestas detectaron que muchos barrios con un alto porcentaje de fumadores tenían a su vez altos porcentajes de individuos nacidos fuera del Reino Unido, individuos jóvenes, individuos con bajo nivel educativo y eran barrios con altas tasas de desempleo. Tras analizar las diferencias entre la población de los barrios, que con la primera encuesta eran calificados de mayoría de fumadores, pero con la segunda encuesta eran calificados de minorías de fumadores, y viceversa, comprobaron que los barrios en los que había disparidad de resultados se debían a la baja representatividad de las minorías étnicas de la segunda encuesta.

4.5 Mapa Escenarios-Métodos

Como parte del marco de referencia se incluye un mapa (Figura 14) que posiciona los distintos métodos de generación de poblaciones respecto al tipo de población (multinivel o no), en la dimensión horizontal, y respecto a la información básica necesaria para utilizar el método, en

la dimensión vertical. La información básica cubre un amplio espectro: la población con todos los atributos, una muestra de la población con todos los atributos (muestra completa), varias muestras con parte de los atributos (muestras parciales), distribuciones marginales (desde unidimensionales a M-1 dimensionales, probabilidades condicionadas y datos o modelos de correlación entre los atributos.

En la tabla de abreviaturas, al final de la tesis, puede encontrarse la equivalencia de los distintos acrónimos indicados en el mapa.

La información básica representa la información necesaria para poder utilizar el método. Adicional a esta información básica, se ha marcado un rectángulo más claro con los métodos que imponen distribuciones marginales objetivo, la cual es parte de la información básica que utiliza el método.

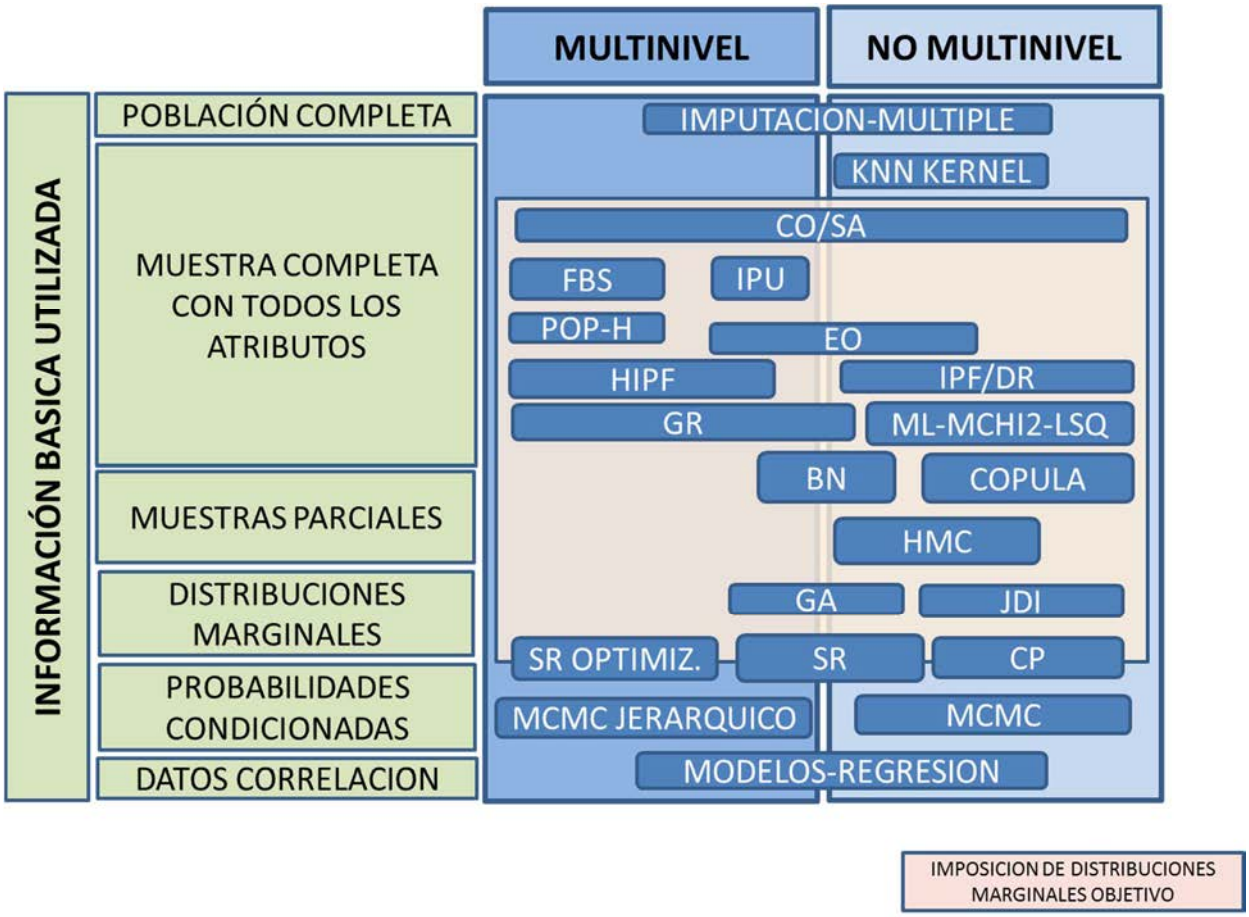

Figura 14Mapa de Métodos.

En la parte superior del mapa se posicionan los métodos de imputación múltiple, utilizados para anonimización de poblaciones, que requieren como información básica los datos de la población completa a anonimizar. A continuación comienzan los métodos que utilizan la información de una muestra completa. El primero es el método *K-nearest neighbors*

Crossover Kernel (KNN KERNEL), no requiere utilizar distribuciones marginales objetivo, aunque puede adaptarse para incorporarlas y producir poblaciones que tiendan a ajustarse a los marginales objetivo.

Seguidamente se posicionan los métodos de *reweighting* que utilizan la información de una muestra completa junto con información de distribuciones marginales de distinta dimensionalidad.

En la zona media baja del mapa se posicionan mayoritariamente los métodos probabilísticos. *Bayesian network* (BN), Síntesis con Cópula y *Hidden Markov Model* (HMM) operan con múltiples muestras parciales (con atributos comunes), aunque con un muestras completas los métodos son más eficientes. Siguen los métodos que solo requieren información de distribuciones marginales; algoritmos genéticos (GA) y *Joint Distribution Inference* (JDI). En la parte inferior del mapa se posicionan los métodos probabilísticos que utilizan probabilidades condicionadas, y datos sobre correlación o modelos de regresión de los atributos (coeficientes de correlación, modelos lineales generalizados, etc.).

A partir de este mapa se ha construido la Tabla 5 con distintos escenarios representativos y los métodos recomendados en cada caso. Los escenarios vienen definidos en relación a la combinación de necesidades y datos disponibles para generar la población sintética.

En la tabla se han incluido como necesidades representativas:

- Población con un único nivel o Multinivel.
- Ajuste de Marginales o del total.
- Máxima heterogeneidad de los atributos de la población.
- Preservar confidencialidad.

En cuanto a disponibilidad de datos para generar la población, ya se han indicado en la sección anterior las posibles situaciones representativas.

Escenario	Método
Sin muestra, pero con información auxiliar (Distribuciones de atributos de diferente dimensionalidad de la población con un único nivel).	SR(Data Fusion), CP, JDI, GA
Sin muestra, pero con información auxiliar (Distribuciones de atributos de diferente dimensionalidad de una población multinivel).	SR con optimización(Data Fusion), GA
Muestra con todos los atributos de interés de hogares e individuos (multinivel) y datos marginales.	GR, SA, IPU, HIPF, EO, Pop-H, FBS
No se desea limitar la heterogeneidad de los atributos individuales y de hogares a los de la muestra. No se requiere ajuste de marginales.	Métodos Probabilísticos
Muestra con único nivel y datos marginales completos o incompletos.	IPF, ML, MCHI2, LSQ, Función Cópula, *Kernel* Cruce K-NN, BN
Probabilidades condicionadas dados M-1 atributos, o M-2 atributos, etc. (parciales o incompletas).	MCMC; HMC; CP; BN
Múltiples muestras de distinto tamaño, conteniendo distintos atributos (algunos comunes), y marginales de atributos.	HMC; CP
Muestra con menos atributos que la población y datos estadísticos agregados.	SR(Data Fusion); SA
Múltiples muestras parciales sin marginales (o con marginales).	SR(Data Fusion); HMC
Información parcial (distribuciones parciales), para simulación y pruebas (con o sin imposición de marginales).	IPF, Métodos Probabilísticos, JDI, Técnicas de Imputación
Modelo de datos (correlaciones entre atributos, o regresiones…), sin marginales (también con información de marginales).	Imputación basada en modelos
Muestra completa representativa de la que se pueden obtener modelos de los datos, correlaciones y distribuciones de los valores de los atributos. Sin necesidad de ajustar la población a marginales.	Técnicas basadas en modelos de correlación o regresión
Datos originales completos: anonimización para preservar la confidencialidad de los datos (datos que puedan usarse para inferencia en lugar de los datos reales originales).	Técnicas de imputación múltiple y basadas en modelos

Tabla 5 Escenarios y Métodos.

5 Marco de Referencia: Análisis Comparativo

Parte del marco de referencia que se describe en este capítulo ha sido publicado en la revista *Computers, Environment and Urban Systems*, factor de impacto 2.650 (Durán-Heras, García-Gutiérrez, & Castilla-Alcalá, 2017).

5.1 Introducción: Esquema General

Con objeto de poder llevar a cabo experimentos de generación de poblaciones con distintos métodos, a partir de los que puedan realizarse análisis estadísticos de los resultados y obtener así conclusiones generales válidas, se propone una metodología de análisis comparativo, dentro del marco de referencia, basada en cinco puntos:

- Población de Referencia.
- Métrica de comparación.
- Validación interna y externa.
- Prueba de Hipótesis.
- Análisis de sensibilidad.

Antes de entrar en detalle en cada uno de estos puntos, se describe el proceso general de experimentación y análisis comparativo entre dos métodos de generación de poblaciones. En la Figura 15 se muestra un esquema del proceso.

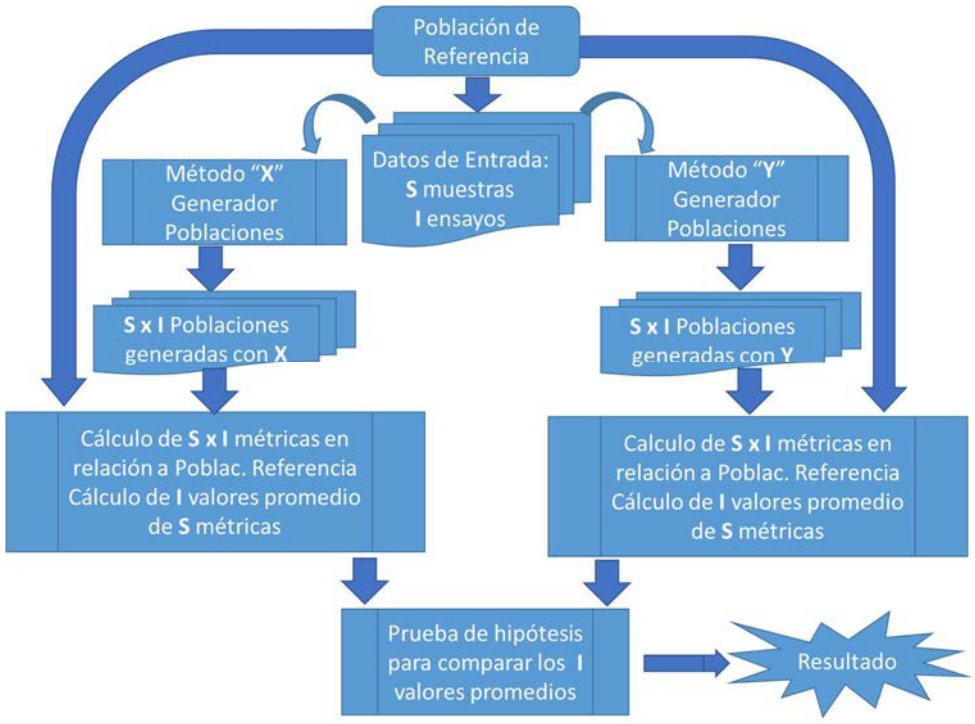

Figura 15 Esquema del proceso de análisis comparativo.

El proceso comienza con los datos de una población real que puede estar en formato de microdatos o en formato de tabla multidimensional. Esta población será utilizada como población de referencia y es la que se va a replicar artificialmente con distintos métodos. Esta población vendrá descrita mediante un conjunto de atributos, algunos de los cuales corresponderán a atributos espaciales que indicarán la "*small area*" de localización (o atributos que definen el "dominio"). A partir de esta población se calculan los datos de entrada que utilizarán los métodos a comparar. Si los métodos operan con poblaciones multinivel, la población de referencia estará constituida por la correspondiente jerarquía de niveles.

Para explicar el proceso nos centramos en la situación de comparación de dos métodos tipo *reweighting*, en la que se conocen los marginales de los atributos de la población a generar y una muestra de la misma.

Para este caso, a partir de los datos de la población de referencia se determinan los marginales de los atributos para cada una de las "*small area*" que constituyen lo que llamaremos el número de ensayos "**I**", y se extraen un conjunto "**S**" de muestras aleatorias de igual tamaño de la población de referencia. Por tanto se tendrán **S*I** conjuntos de datos de entrada.

Con cada uno de los métodos a comparar y con cada conjunto de datos de entrada, muestra y distribución de marginales de cada "*small area*", se genera una población por "*small area*" ajustada a sus marginales. Por tanto se tendrán **S*I** poblaciones generadas con cada método.

Una vez se han sintetizado las poblaciones, se selecciona una métrica con la que determinar cuán distantes están de la población de referencia.

Ya que con cada método se han utilizado **S** muestras para cada "*small area*", se habrán generado **S** poblaciones para cada una, por lo que se tendrán **S** valores de la métrica. Promediando estos **S** valores se tendrán **I** valores medios por cada método (uno por "*small area*"). A estos valores medios de la métrica los llamaremos errores medios.

En este momento se dispone 2 conjuntos de **I** errores medios correspondientes a **I** "*small areas*". Cada uno de estos conjuntos de valores corresponde a un método de generación. A continuación se procede a contrastar las medias de estos dos conjuntos para determinar el método que produce las poblaciones con menores errores medios.

A partir de estos dos conjuntos de **I** valores se lleva a cabo una prueba de contraste de hipótesis. Con esta prueba se determina la media de la diferencia entre los errores medios y el intervalo de confianza de la misma.

Si en lugar de métodos tipo *reweighting* se compararan métodos que no utilizan muestra, y en su lugar utilizan otra información de la población de referencia, como distribuciones de probabilidad condicional, se determinará dicha información con los datos de la población de referencia una vez (**S**=1) y se generará una población para cada "*small area*" con cada método.

Tal como se indicó en el capítulo anterior, para que puedan extraerse resultados concluyentes de la comparativa, los métodos a comparar han de utilizar la misma información de partida. Si dos métodos utilizan diferentes tipos de datos, generarán poblaciones cuya comparación será controvertida. En estos casos, las diferencias entre las poblaciones generadas no solo tendrán su origen en el uso de métodos diferentes, sino también en la diferente información de partida que utilizan, no siendo posible determinar qué parte de las diferencias se deben a los datos de partida y qué parte a la algorítmica del método.

Si a pesar de todo se plantea una comparación de métodos que utilizan distintos tipos de datos de entrada hay que tener en cuenta este hecho cuando se interpreten los resultados.

Los resultados de la comparativa pueden validarse seleccionando una métrica diferente y repitiendo el proceso para comprobar que el resultado de la prueba de contraste de hipótesis para la diferencia de medias es el mismo.

Si la comparativa se desea efectuar con múltiples tamaños de muestra se extraerá un conjunto "S" de muestras por cada uno de los tamaños y se volverá a repetir el proceso para cada conjunto.

5.2 Población de Referencia

Evaluar la precisión de las poblaciones obtenidas mediante un método de síntesis no es una tarea fácil porque la población de referencia perfecta para la población sintética sería la población real. Sin embargo, generalmente no están disponibles los datos completos de la población real, debido por ejemplo, a cuestiones de privacidad. De hecho, esta falta de disponibilidad es la principal razón para sintetizar poblaciones.

Es debido a esta ausencia de datos reales, por lo que los distintos estudios de comparación de métodos mostrados en la Tabla 4 adoptan diferentes estrategias. En la quinta columna de dicha tabla se observa que en la mayoría de los estudios la máxima dimensión utilizada en el proceso de comparación es inferior al número de atributos de la población. Es decir, no se compara con la distribución M-dimensional, sino que se hacen comparativas de menor dimensionalidad, ya que frecuentemente estos son los datos disponibles (por ejemplo, tablas del censo).

Así por ejemplo, en el estudio de Harland et al. (2012) se validan las poblaciones sintetizadas utilizando como referencia información marginal complementaria de la población real, en concreto, utiliza 10 tablas bidimensionales (tablas de contingencia de 2 atributos) y 1 tabla con 3 atributos. Este tipo de comparativas no es totalmente concluyente, pues unos buenos resultados de baja dimensionalidad no presuponen que sean buenos a mayor dimensionalidad.

Hay algunos estudios, como el de Ryan et al. (2009) en el que tuvieron acceso a datos completos sobre una población real de 11.499 empresas con tres atributos, sintetizaron poblaciones con dos métodos diferentes y midieron la precisión de dichas poblaciones utilizando la población "real" como población de referencia.

El enfoque que se propone aquí es similar al de Ryan et al. (2009) en el sentido de que se propone utilizar poblaciones con información de validación completa. Pero, a diferencia de Ryan et al, es una población repartida en múltiples "*small areas*" para las que se generarán las correspondientes poblaciones. Para el marco de referencia se propone una población completa (con un único nivel jerárquico y/o con multinivel, según el caso) repartida en múltiples "*small area*s", contra la que comparar con la máxima dimensionalidad posible. De este modo la comparativa será concluyente.

Dado que disponer de una población completa no es siempre posible, en los análisis de los siguientes capítulos se utiliza una población obtenida de los ficheros de microdatos públicos del censo, que corresponden a un pequeño porcentaje de la población real (5%), tratándolo como si fuera el 100% del censo. Se tomaran muestras aleatorias de distinto tamaño de esta población, que serán utilizadas para sintetizar las poblaciones de las distintas "*small area*s".

Las agencias estadísticas gubernamentales ponen a la libre disposición de los ciudadanos muchos otros ficheros de microdatos completos, correspondientes a encuestas en distintos ámbitos (sociedad, condiciones de vida, mercado laboral, servicios, etc.) los cuales pueden utilizarse como población de referencia.

Es fundamental disponer de una población de referencia, ya que sin ella difícilmente se podrá preparar el conjunto de muestras necesario para lleva a cabo el estudio comparativo. Así mismo, tampoco podría procederse al cálculo de métricas con la máxima dimensionalidad de la población, por lo que habría que hacerlo con una dimensionalidad inferior, lo cual hace que los resultados del análisis no sean precisos.

5.3 Métricas para poblaciones con un único nivel y multinivel

Cualquiera de las métricas descritas en la sección 4.2 podría utilizarse para estudiar la similitud entre tablas multidimensionales de poblaciones con un único nivel. Cuanto mayor

sea el número de atributos *M* y el número de categorías con las que se describen los agentes de la población C_m, mayor será el número de posibles tipos de agentes de la población, es decir, el número de celdas a comparar $\mathrm{C}_1 * \mathrm{C}_2 * \ldots * \mathrm{C}_\mathrm{M}$. Este valor corresponde al valor máximo. El número real de tipos posibles normalmente será inferior, ya que muchas celdas serán ceros estructurales por tratarse de combinaciones de categorías de atributos imposibles, como por ejemplo, un menor de 14 años, jubilado o divorciado.

Para el caso de poblaciones multinivel de hogares e individuos también pueden usarse las mismas métricas, aunque en este caso pueden plantearse dos enfoques de comparación diferentes.

El primero sería considerar cada tipo de hogar de acuerdo a los valores de los atributos de nivel hogar y de los atributos del nivel de cada individuo miembro del hogar, es decir, dos hogares que tengan los mismos atributos a nivel hogar, y cuyos miembros también tengan los mismos atributos a nivel persona, serán considerados idénticos, y por tanto, estarán en la misma celda de la tabla multidimensional. Este enfoque conduce a tablas con un número de celdas enorme, pues aunque solo se tengan en cuenta hogares de 1.000 tipos e individuos de otros 1.000, y se consideren hogares con 2 miembros, se tendrán 10^9 tipos posibles de hogares.

Dentro de este enfoque, en ocasiones se opta por limitar el número de tipos (número de celdas a comparar), con lo que podría reducirse el número de tipos/celdas, y tener un número razonable de celdas a comparar.

El otro enfoque es construir una tabla multidimensional para cada nivel, con los atributos y categorías de cada nivel, y comparar las dos tablas de cada nivel con las dos tablas de los niveles correspondientes de la segunda población. De este modo se comparan la tabla de hogares por un lado y la de individuos por otro, con lo que se tendrán dos métricas distintas, una para cada nivel. En el caso del ejemplo anterior habría que comparar dos tablas de 1.000 celdas con las dos de la población de referencia.

A partir de estas dos métricas puede construirse una métrica combinada que pondere ambas métricas.

5.4 Métrica de Comparación

Voas & Williamson (2001) elaboraron un análisis sobre las métricas de comparación de poblaciones agrupándolas en tres familias separadas: métricas tipo Chi-cuadrado, métricas Z y métricas asociadas a la teoría de la información. Concluyeron que todas las alternativas tienen ventajas e inconvenientes específicos y que no existe una métrica generalmente

preferida para todos los casos. Lovelace et al. (2015) también trataron este problema, incluyendo una comparación cuantitativa entre seis métricas aplicadas a un ejemplo de síntesis de población: coeficiente de correlación de Pearson, Error Absoluto Total (TAE), Error Absoluto Estandarizado (SAE), Error Cuadrático Medio (RMSE), Z-Score y el error E>5% del valor observado. Encontraron un alto acuerdo entre ellas, concluyendo que aunque el valor absoluto de las métricas depende de cada una, la posición relativa de los diferentes escenarios permanece sin cambios. Cuando los escenarios son dispares en términos de tamaño de la población total, las métricas relativas tienen la ventaja de proporcionar una mejor comparabilidad cruzada que las métricas absolutas (por ejemplo, SAE sería preferible a TAE).

En esta tesis también se ha realizado un estudio similar al indicado, utilizando cuatro de las métricas más comunes: error de clasificación (%CE), error cuadrático medio (RMSE), Chi cuadrado de Pearson y razón de verosimilitud G^2. No se han incluido los detalles de este estudio con objeto de simplificar las descripciones dentro del marco de referencia comparativo y facilitar su lectura, pero se resume brevemente y se destacan las conclusiones del mismo.

Se han utilizado poblaciones generadas con los métodos de *Simulated Annealing* (SA) e IPF-BLP, para realizar la comparación entre las métricas con el objetivo de verificar si la métrica de error puede influir en la preferencia entre uno u otro método.

Los experimentos se han realizado en un escenario de población con 6 atributos (2700 celdas); muestra del 5% de la población; y 60 municipios andaluces ("*small areas*"). Se ha medido el valor de las 4 métricas para cada una de las poblaciones generadas. Se han llevado a cabo pruebas de hipótesis para determinar que método genera las poblaciones con menores valores de las métricas. La conclusión a las que se ha llegado es que las cuatro métricas de error (%CE, RMSE, Chi-cuadrado y G^2) conducen a la misma clasificación de preferencia entre IPF y SA, siempre mostrando una superioridad significativa del mismo método, IPF-BLP sobre SA, en términos de cada medida particular de error.

A la vista de estas conclusiones, en línea con la de los estudios citados, se propone utilizar una única métrica para el proceso de comparación. En concreto, se propone una métrica "relativa" fácil de interpretar y calcular, el Error de Clasificación Porcentual (%CE), métrica de uso generalizado (Harland et al., 2012; Voas & Williamson, 2001), que mide el porcentaje de individuos que hay que cambiar de celda (grupo socio-económico-demográfico) para obtener la población de referencia.

Dado que esta métrica puede ser requerida para comparar tablas con distinta dimensionalidad, se utilizará con un superíndice para indicar la dimensionalidad. Si se usa para comparar tablas M-dimensionales completas, se tiene:

$$\%CE^{M-dim} = \frac{\sum_{c_1=1}^{\mathbf{C_1}} \dots \sum_{c_M=1}^{\mathbf{C_M}} \left| O_{c_1 \dots c_M} - E_{c_1 \dots c_M} \right|}{2\,N} 100 \qquad (17)$$

Y en el caso de comparar solo los valores marginales unidimensionales del atributo m de dicha tabla:

$$\%CE^{1-dim}(m) = \frac{\sum_{c_m=1}^{\mathbf{C_m}} \left| O_{c_{\mathrm{m}+}} - E_{c_{\mathrm{m}+}} \right|}{2\,N} 100 \qquad (18)$$

Donde O_{c_m+} son los valores marginales de la tabla de referencia observada, y E_{c_m+} los de la tabla estimada:

$$E_{c_m+} = \sum_{c_1=1}^{\mathbf{C_1}} \dots \sum_{c_{m-1}=1}^{\mathbf{C_{m-1}}} \sum_{c_{m+1}=1}^{\mathbf{C_{m+1}}} \dots \sum_{c_M=1}^{\mathbf{C_M}} E_{c_1 \dots c_m \dots c_M}$$

Esta métrica también permite ser utilizada con una parte de la población, así por ejemplo, puede obtenerse el error de clasificación M-dimensional de un determinado conjunto de individuos $N_{\boldsymbol{c}^*}$ que tienen un valor específico de un atributo $\boldsymbol{c}^*$, en cuyo caso los sumatorios de la ecuación (17) no incluyen el sumatorio sobre dicho atributo, y solo se suman las celdas que tiene dicho valor $\boldsymbol{c}^*$**:**

$$\%CE^{M-dim}(\boldsymbol{c}^*) = \frac{\sum_{c_1=1}^{\mathbf{C_1}} \dots \sum_{c_M=1}^{\mathbf{C_M}} \left| O_{c_1 \dots \boldsymbol{c}^* \dots c_M} - E_{c_1 \dots \boldsymbol{c}^* \dots c_M} \right|}{2\,N_{\boldsymbol{c}^*}} 100 \qquad (19)$$

El error de clasificación es mayor a medida que se incrementa la dimensionalidad de la tabla. Es fácil demostrar que para una tabla M-dimensional el error %CE se hace mayor a medida que aumenta la dimensionalidad de la comparativa:

$$\max_m \%CE^{1-dim}(m) \leq \max \%CE^{2-dim} \leq \cdots \leq \max \%CE^{(M-1)-dim} \leq \%CE^{M-dim}$$

Para los casos en que se comparen poblaciones multinivel, tal como se ha explicado en la sección anterior, se determinará el valor de la métrica para cada nivel jerárquico de la población y se obtendrán tantos valores como niveles. Estos valores se pueden condensar en uno calculando el promedio ponderado con el número de atributos de cada nivel. De este modo se consigue disponer de una única métrica consolidada para la población multinivel.

5.5 Validación Interna y Externa

Haciendo uso de distinción que hace Lovelace et al. (2015) entre la validación interna y externa en la evaluación de los métodos de síntesis, se propone incluir estos dos tipos de validación dentro del proceso de comparación. A través de la validación externa, se pretende evaluar si la población sintetizada es una representación precisa de la población de referencia, mientras que la validación interna aborda qué tan bien cumple la población sintetizada con las restricciones de entrada impuestas durante el proceso de síntesis.

La validación externa es en realidad la más relevante de las dos, porque mide qué tan similar es la población sintetizada respecto a la población de referencia. Sin embargo, esta comparación no suele ser factible en aplicaciones del mundo real, ya que generalmente no está disponible la información completa sobre la población de referencia, por lo tanto, con cierta frecuencia los estudios comparativos se limitan a la validación interna. Ante la falta de datos de la población de referencia, en ocasiones se recurre a utilizar otros esquemas de validación externa alternativos, tales como: utilizar un modelo de regresión de los atributos procedente de la muestra o de otros datos disponibles, para generar la distribución probable y comparar los resultados; la agregación de datos sintéticos a una escala superior para la que se dispongan de datos reales (distrito, municipio, provincia, etc.); o el uso de determinados datos de referencia externos, procedentes de otras fuentes, con los que comparar los datos sintetizados.

La característica clave del marco de referencia es que plantea producir poblaciones sintéticas que representan poblaciones conocidas; por lo tanto, puede medirse perfectamente la precisión de los resultados a través de la comparación celda a celda de las tablas multidimensionales que contienen las poblaciones sintetizadas y la de referencia.

Complementariamente, también se plantea medir la validez interna para analizar algunos aspectos específicos con respecto al rendimiento de los métodos.

Los procesos de validación interna, aparte de permitir identificar errores de implementación de los métodos, permiten determinar el error que introduce el propio método. En los métodos con componente estocástica, que toman como dato de entrada la distribución de marginales objetivo de la población a sintetizar, la validación interna determina el error en dichos marginales en la población generada. Estos errores son de un orden inferior a los que se tienen en la validación externa (Auld & Mohammadian, 2010).

Es habitual utilizar la métrica del *Total Absolute Error* (TAE), con la misma dimensionalidad que los marginales objetivo, para la validación interna de los métodos de generación de

poblaciones, ya que mide la diferencia absoluta entre los marginales objetivos de las "*small areas*" (datos de entrada) y los marginales de la población sintetizada para dichas áreas.

La mayoría de los métodos de generación de poblaciones se han sometido a procesos de validación interna, pero no todos se han sometido a validaciones externas con precisión que establezcan la concordancia entre la población generada y la población observada de referencia.

Por ejemplo, en el estudio comparativo entre GREGWT y SA (Tanton et al., 2014) de la Tabla 4, los marginales objetivo que se imponen son una tabla 3-dimensional de atributos de individuos (sexo, edad, situación laboral), 6 tablas bidimensionales de atributos de hogar (tipo de hogar, régimen de tenencia, ingresos, renta de alquiler pagada, importe hipoteca pagada, composición del hogar), y una tabla unidimensional (tamaño del hogar). Las métricas de las poblaciones generadas con estos métodos se calculan tomando como referencia este conjunto de marginales impuestos. Por tanto, la comparativa de las métricas determina el método que produce la población con los valores más próximos a dichas distribuciones marginales.

A esta validación interna sigue, en dicho estudio, una validación externa "parcial" ad hoc, con uno de los esquemas alternativos anteriormente indicados, diferente al que aquí se plantea. Este tipo de validación compara determinadas celdas de la tabla 3-dimensional de la población de hogares generada con cada método, con datos externos procedentes de otras fuentes, que representan el valor dichas celdas. En concreto, a partir de la población sintetizada de hogares, se determina el porcentaje de hogares con un coste de vivienda (ya sea en alquiler o hipoteca) igual o superior al 30% de los ingresos del hogar a partir de la tabla de ingresos x renta de alquiler x importe hipoteca pagada, y se compara el porcentaje obtenido con datos publicados por un organismo independiente, de este modo determina el método con el mejor nivel de coincidencia.

5.6 Prueba de Hipótesis

Una vez se dispone de los 2 conjuntos de valores de la métrica (errores medios) correspondientes a las distintas "*small areas*", hay que determinar el método con el que se han obtenido las poblaciones de "*small area*" con menores errores medios, esto es, las poblaciones más coincidentes con las de referencia. Para lo cual, asumiendo que las poblaciones de las "*small areas*" son independientes, se plantea una prueba de hipótesis con la que determinar si se puede aceptar o rechazar la afirmación (hipótesis nula H_o) de que no hay diferencia entre los errores medios de las poblaciones generadas con cada método para cada "*small area*".

El resultado de esta prueba estará en función del valor-p que se tenga en dicha prueba. Si el valor es menor que el nivel de significación establecido α, entonces se podrá rechazar la afirmación de la hipótesis nula. Este valor-p nos aporta la probabilidad de obtener, siendo H_o verdadera, las diferencias de errores para las distintas "*small areas*" iguales o superiores a las obtenidas.

Se plantean dos posibles pruebas, una paramétrica y otra no paramétricas, de comparación de muestras pareadas.

5.6.1 Prueba t de Student

Esta prueba con muestras pareadas, también conocida como t-test, requiere normalidad de las diferencias de errores medios, por lo que previamente ha de ejecutarse una prueba que confirme la hipótesis de normalidad de las diferencias para que el resultado del t-test sea fiable. Para casos en los que se tenga un número no muy grande de diferencias se recomienda la prueba de normalidad de Shapiro-Wilk, una de las pruebas de normalidad más consolidadas y con mayor potencia estadística (Yap & Sim, 2011).

La prueba t-test pareada no requiere suponer que ambas muestras tengan la misma varianza (homocedasticidad); esa suposición solo se precisa en el t-test no apareado utilizado para comparar dos muestras independientes (no apareadas) (McDonald, 2014). Para esos casos del t-test no apareado, con dos muestras independientes con varianzas diferentes, se utiliza la variante del t-test de Welch.

En esta prueba se evalúa la hipótesis nula de que la media de las diferencias de errores es igual a un valor especificado μ_0, en nuestro caso $\mu_0 = 0$, haciendo uso del estadístico:

$$t = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{I}}$$

Donde $\bar{x}$ es la media muestral (media de los errores medios), σ es la desviación estándar muestral e I es el tamaño de la muestra (número de "*small areas*") Los grados de libertad utilizados en esta prueba se corresponden al valor (I -1).

5.6.2 Prueba de los rangos con signo de Wilcoxon

Esta prueba de hipótesis (*Wilcoxon signed Rank*) es una alternativa a la prueba t de Student cuando no se cumple el requisito de normalidad de los datos.

Esta prueba requiere que las diferencias de errores tengan una distribución continua y simétrica respecto a la mediana. Si los datos de diferencia tienen una distribución continua simétrica, la mediana y la media coinciden.

La hipótesis nula en este caso es que la mediana de las diferencias de los errores medios es 0, por lo que se está indicando que los errores medios son iguales.

Para determinar si hay una diferencia estadísticamente significativa entre los errores de las poblaciones producidas por dos métodos, al igual que con el t-test, se emparejan para cada "*small area*" *i*, los valores de $\overline{\%CE_i}\ i = 1 \ldots I$ para los dos métodos (Navidi, 2015). Por ejemplo, para comparar el método X con el método Y, y determinar si la diferencia de medias no es nula (mayor error en la población generada con el método X), se realiza la prueba calculando los valores para la variable de diferencia $\overline{\%CE_i}(x) - \overline{\%CE_i}(y)$. Posteriormente se calcula la suma de los rangos positivos S_+. Como la hipótesis nula es de la forma H_o: $\mu = 0$, esta es una prueba de dos colas, los valores altos de S_+ proporcionan evidencia contra H_o. Un valor-p < α permite rechazar la hipótesis nula con un nivel de significación de α, es decir, un método produce consistentemente más errores que el otro (Navidi, 2015).

Suponiendo la simetría y la continuidad de la distribución de la variable de diferencia, se calcula un intervalo de confianza no paramétrico y un estimador para la pseudomediana de la variable diferencia utilizando estos *I* puntos. Esta prueba no paramétrica funciona razonablemente bien bajo violaciones moderadas de las suposiciones de simetría (Voraprateep, 2013), lo cual puede comprobarse realizando la prueba del signo.

La prueba del signo es una prueba estadística de comparación de medias que no precisa imponer este requisito de simetría, la cual puede realizarse en los casos en que se detecten diferencias importantes entre la media y la mediana, y por tanto que no se cumpla el requisito de simetría. De este modo se puede verificar que se obtienen las mismas conclusiones con ambas pruebas.

5.7 Análisis de sensibilidad.

Por último, el marco de referencia planteado estaría incompleto si no tuviera la capacidad de detectar diferencias de rendimiento de los métodos en relación a los parámetros de experimentación, para lo cual en el marco se incorpora la noción de análisis de sensibilidad.

Con el objetivo de determinar si los resultados observados con las pruebas de hipótesis dependen de los valores específicos asignados a los diversos parámetros (es decir, si los resultados y las conclusiones son "sensibles" a la valores específicos elegidos), el marco de referencia contempla la repetición del análisis en múltiples "escenarios". Cada escenario está definido por una combinación dada de valores de parámetros, para que de este modo puedan probarse todas las combinaciones de interés. Entre los parámetros a incluir en el análisis de sensibilidad (es decir, parámetros para los que han de probarse varios valores alternativos)

se considerarán al menos: el tamaño de muestra, el número de celdas (que es el producto del número de categorías de cada atributo de la población) y el tamaño de las "*small area*s".

Por otro lado están los parámetros de los propios métodos, por ejemplo, en el caso de los métodos metaheurísticos, los resultados pueden ser sensibles a la configuración de los parámetros del método; así por ejemplo, el método *Simulated Annealing* puede ser sensible al número máximo de iteraciones que se impongan, a la "temperatura inicial" o a la métrica que se utilice como medida de la bondad del ajuste.

Si se añadieran estos últimos parámetros a la configuración de escenarios, se incrementaría exponencialmente el número de escenarios. Con el fin de no incrementar en exceso el número de escenarios de experimentación con la inclusión de estos parámetros del propio método, se propone fijar estos parámetros con los valores recomendados. De este modo, aunque no se consigan los mejores resultados del algoritmo, no estarán muy alejados de los óptimos. En algunos casos, como el ya comentado de *Simulated Annealing*, los propios implementadores del método recomiendan valores de parámetros con los que se consiguen los resultados más eficientes.

6 Comparación de métodos caso un nivel: IPF vs SA

Este capítulo ha sido publicado en la revista *Computers, Environment and Urban Systems*, factor de impacto 2.650 (Durán-Heras et al., 2017).

6.1 Introducción

En este capítulo se presenta el análisis comparativo de métodos basados en dos enfoques del tipo *reweighting* de uso muy frecuentemente en la generación de poblaciones, como son los métodos basados en IPF y métodos de optimización combinatoria (CO).

Antes de proceder a la descripción de los experimentos con los que se ha realizado el análisis comparativo, se resumen los resultados de 3 estudios previos, incluidos en la Tabla 4, que han confrontado métodos con estos dos enfoques.

En primer lugar, Ryan et al. (2009) compararon los resultados del *Simulated Annealing* (SA) e IPF con muestreo de Monte Carlo, al que se refieren como IPFSR (IPF *Synthetic Reconstruction*). Con el IPF obtuvieron la tabla multidimensional con la distribución conjunta de atributos de la población, que posteriormente utilizaron para muestrear y generar la población. Una vez sintetizaron las poblaciones con ambas técnicas, compararon las poblaciones obtenidas con la población de referencia. Evaluaron el rendimiento de las técnicas con muestras de distinto tamaño (entre el 1% y el 100% de la población total), y utilizaron distintos conjuntos de distribuciones marginales objetivo, imponiendo como restricción de marginales una distribución unidimensional, una bidimensional y una tridimensional. En sus experimentos utilizaron una población de 11.499 empresas descritas con 3 atributos, uno de los cuales era un atributo de localización espacial dentro de la "*small area*". Estos autores concluyeron que, aunque ambos métodos permiten sintetizar la población de empresas con bastante semejanza a la población de referencia, no obstante, SA generalmente produce poblaciones más precisas que el método IPFSR.

El estudio de Harland et al. (2012) compara el rendimiento de los métodos *Deterministic Reweighting*, Probabilidad Condicional y SA, generando poblaciones de "*small area*" en el área metropolitana de la ciudad de Leeds (Reino Unido). El método *Deterministic Reweighting* que utilizan se basa en el IPF para ajustar la muestra a las distribuciones marginales objetivo, pero no proporcionan detalles sobre el método utilizado para la conversión a valores enteros de la población. Evalúan el impacto de diferentes escalas espaciales en el proceso de generación de poblaciones, con tres niveles de agregación espacial. Y llegan a la conclusión de que SA supera a los otros dos métodos en todas las métricas y escenarios que analizan.

Kim & Lee (2016) estudian la generación de poblaciones utilizando SA, e intentan establecer el conjunto de parámetros óptimo para este algoritmo. Logran los mejores resultados aumentando la "temperatura inicial" y el número máximo de iteraciones. También llevan a cabo una comparación de este método con un único caso de generación de población generada con IPF, concluyendo que no hay un resultado claramente superior.

No se encuentra ninguna otra comparación cuantitativa entre métodos basados en IPF y métodos de Optimización Combinatoria (CO). Williamson (2013) analiza un estudio previo, publicado en 2001, en el que compara SA con Reconstrucción Sintética aplicándolos a un caso de generación de población multinivel de hogares e individuos con ajuste simultáneo de sus marginales. Basándose en las comparaciones entre las poblaciones sintetizadas con ambos métodos para este caso, concluye que el enfoque de CO es superior al de Reconstrucción Sintética para la generación de microdatos de "*small area*", ya que, aunque en promedio producen poblaciones de características similares, el primero tiene menos variabilidad de rendimiento entre ejecuciones. Pueden encontrarse todos los detalles de este estudio en (Huang & Williamson, 2001).

Varios autores se refieren a estos estudios cuantitativos cuando comparan métodos del tipo CO e IPF en sus investigaciones, concluyendo en muchos casos, que el primero parece superior al segundo (Abraham, Stephan, & Hunt, 2012; Cho et al., 2014; Farooq et al., 2013; Hermes & Poulsen, 2012a; Kurban, Gallagher, & Persky, 2012; Levy, Fabian, & Peters, 2014; L. Ma & Srinivasan, 2015; Sun & Erath, 2015; Pei-jun Ye et al., 2016). Sin embargo, aunque reconociendo el valor de dichos estudios, sus conclusiones han de interpretarse teniendo en cuenta que se trata de casos específicos de aplicación, difícilmente generalizables. Dichos estudios no evalúan la significación estadística de sus resultados, la cual podría obtenerse con la realización de pruebas de hipótesis con datos de múltiples casos que analizaran la variabilidad inherente de los resultados.

Además de estas comparaciones experimentales, algunos autores han analizado las diferencias estructurales de los algoritmos en los que se basan estos tipos de métodos. La principal desventaja atribuida a los métodos basados en IPF es la necesidad de un segundo paso para obtener los valores enteros que representan la población (Choupani & Mamdoohi, 2016; R Lovelace & Ballas, 2013; Williamson et al., 1998). Este proceso de conversión a enteros puede provocar el desajuste de los marginales y de la estructura de asociación entre atributos recogida en la muestra. En cuanto a los inconvenientes más citados de los métodos con algoritmos de optimización combinatoria, se cita que no preserva la asociación de atributos de la muestra, computacionalmente son muy intensivos, y requieren gran capacidad de almacenamiento (Pritchard & Miller, 2012).

En este capítulo se complementan los estudios comparativos existentes con un análisis estructurado conforme al marco de referencia expuesto anteriormente.

En relación a los métodos específicos a utilizar en la comparativa, se ha de seleccionar un representante para cada enfoque, los cuales han de tener relevancia significativa justificada por su amplia difusión o por sus acreditadas características de buen rendimiento.

En el caso de CO, Williamson et al. (1998) comparó varias alternativas, concluyendo que SA proporcionaba los mejores resultados para la síntesis de poblaciones. Posteriormente propuso algunos refinamientos al algoritmo (Williamson, 2013). La mayoría de los autores se refieren a SA cuando hablan del métodos CO, y muchos utilizan la implementación en lenguaje FORTRAN de Williamson (2017) cuando lo aplican en sus investigaciones (Hanaoka & Clarke, 2007; Harland et al., 2012; Hermes & Poulsen, 2012a; Morrissey, Clarke, Ballas, Hynes, & O'Donoghue, 2008; Ryan et al., 2009). Por tanto, como representante del enfoque CO se selecciona dicha implementación de SA.

También existen varias alternativas disponibles entre los métodos basados en IPF. Teniendo en cuenta los dos pasos conceptuales de que consta el proceso de generación con las técnica del IPF, ya explicados en la sección 3.2.1, para el primer paso pueden plantearse dos alternativas: aplicar el algoritmo clásico de Demings & Stephan (1940) a la tabla multidimensional de la muestra imponiéndole los marginales o ejecutar el algoritmo IPF en dos fases propuesto por Beckman et al. (1996). Como ya se indicó cuando se explicó el IPF de Beckman (sección 3.2.1), los dos métodos producen resultados muy similares, por lo que dada la similitud de los mismos se selecciona el algoritmo clásico, ya que es más utilizado (Guo & Bhat, 2007a; Harland et al., 2012; Robin Lovelace et al., 2014; Ryan et al., 2009; Zhu & Ferreira, 2014), y desde el punto de vista computacional es menos exigente (Pritchard & Miller, 2012).

En cuanto al segundo paso, existen distintas alternativas para convertir a enteros los valores de la tabla de distribución de probabilidades conjunta de los atributos de la población obtenida con el IPF. Autores como Lovelace & Ballas (2013) y Choupani & Mamdoohi (2015) han demostrado que la técnica utilizada para obtener valores enteros de la población puede tener una influencia significativa en el error de la población sintética. Por lo tanto, se decide utilizar dos técnicas distintas de conversión a enteros para llevar a cabo los experimentos de IPF. El primero es uno de los métodos más utilizados para obtener enteros, el muestreo de Monte Carlo (Beckman et al., 1996; Frick & Axhausen, 2004; Pritchard & Miller, 2012; X. Ye et al., 2009). Nos referiremos a este método como IPF-MCS (*Monte Carlo Sampling*). El segundo es el método propuesto por Choupani & Mamdoohi (2015), que produce un ajuste perfecto con respecto a las distribuciones marginales objetivo de los atributos. Tal como se explicó en la sección 3.2.1.2, el procedimiento se basa en un modelo de programación lineal binaria para

redondear los valores de la tabla, en el que las distribuciones marginales de los atributos se tratan como restricciones obligatorias (logrando así el ajuste de marginales perfecto), y minimizando la desviación respecto la tabla multidimensional original no entera. Los autores argumentan que siempre puede encontrarse una solución óptima si los objetivos marginales son valores enteros. Como resultado, cada celda con valor no entero de la tabla original se redondea hacia arriba o hacia abajo al valor entero más cercano, de tal manera que se logra un ajuste perfecto de todos los marginales objetivos. Nos referimos a este método como IPF-BLP (*Binary Linear Programming*).

Por tanto, se diseñan experimentos para comparar los tres métodos indicados: IPF-MCS, IPF-BLP y SA.

6.2 Diseño de los Experimentos

En esta sección se describen los datos de la población de referencia y los experimentos que se ejecutarán.

Para incrementar la generalidad de los resultados, se ha repetido el análisis comparativo utilizando datos de dos zonas geográficas diferentes. Por tanto, la comparativa se realiza con dos poblaciones de referencia distintas, una de 60 municipios andaluces, y otra de 25 cantones suizos, por lo que los resultados se describen separadamente. Además, el análisis de Andalucía incluye resultados con mayor detalle para un escenario básico y un análisis de sensibilidad de tres parámetros.

6.2.1 Población de Referencia de Andalucía

Se utilizan los microdatos del censo la comunidad autónoma española más poblada, Andalucía, con 8,4 millones de habitantes. Los microdatos del censo comprenden información detallada de distintos atributos, incluyendo el código del municipio, para aproximadamente el 5% de la población.

La población de referencia se extrae de los microdatos públicos del Censo de 2001 (“Instituto Nacional de Estadística,” 2017) de los 60 municipios con más de 20.000 habitantes (no se han considerado municipios más pequeños porque en los microdatos vienen agrupados bajo un código genérico). Cada municipio, corresponde a una división administrativa urbana definida por un territorio con límites fijos establecidos.

En la Figura 16 se muestra un mapa de Andalucía; los 60 municipios con más de 20.000 habitantes que se utilizan en la población de referencia vienen marcados en color más oscuro.

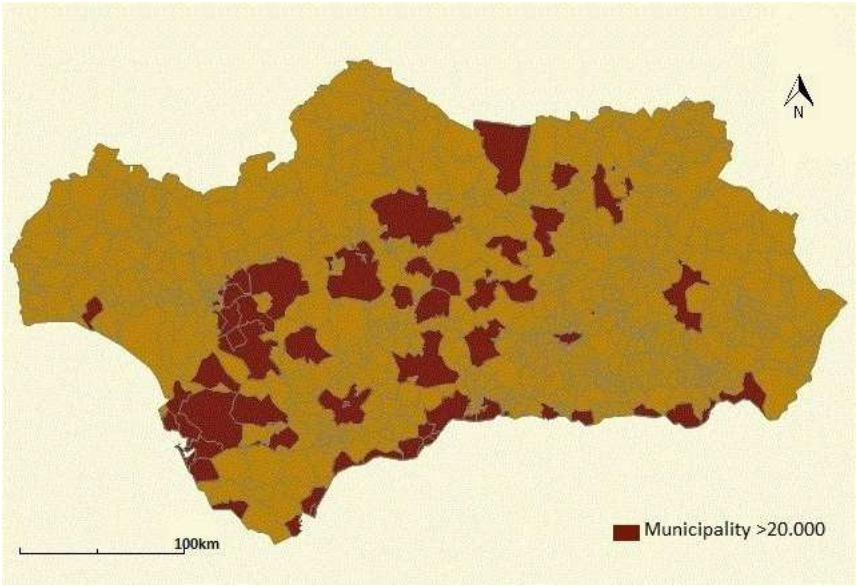

Figura 16 Municipios de Andalucía con más de 20.000 habitantes. Fuente: Censo de Población y Viviendas 2001del INE.

Por tanto, para ejecutar los experimentos, se construye la población de referencia extrayendo de los microdatos del censo la información parcial correspondiente a los atributos y categorías de los individuos residentes en viviendas familiares que se muestran en la Tabla 6 (se excluyen individuos en viviendas colectivas). La población total "real" queda constituida por 228.212 individuos distribuidos en los 60 municipios.

Atributo	**Categorías (número)**
sexo	hombre ; mujer (2)
edad	<15; 15-19 ; 20-24 ; 25-29 ; 30-34 ; 35-39 ; 40-44 ; 45-49 ; 50-54 ; 55-59 ; 60-64 ; 65-69 ; 70-74 ; >75 (14)
nacionalidad	español ; extranjero (2)
estado civil	soltero; casado; separado; viudo; divorciado (5)
situación laboral preferente	estudiante; ocupado; parado buscando primer empleo; parado con trabajo anterior; pensionista de invalidez; pensionista de viudedad u orfandad; pensionista de jubilación; Realizando o compartiendo las tareas del hogar; Otra situación (menores sin escolarizar, rentistas...) (9)
nivel educativo (grados)	analfabeto; sin estudios; primer grado; segundo grado (ESO, EGB, Bachillerato; FP); tercer grado (Diplomatura, Licenciatura, Doctorado) (5)

Tabla 6 Atributos y Categorías del escenario Andalucía.

A partir de esta población "real" se extrae la información de marginales de cada municipio y las correspondientes muestras, que serán utilizadas como datos de entrada para los procesos de síntesis. Finamente se calculará el error entre lo sintetizado y la población "real", a la que nos referimos como la población de referencia. Se utilizará la métrica propuesta del Error de Clasificación con máxima dimensionalidad.

También se van a realizar experimentos agrupando las categorías del atributo edad de distintos modos, con lo que obtendrán tablas de población con diferente número de celdas.

6.2.2 Población de Referencia de Suiza

En este caso, la población de referencia se obtiene de los microdatos de la PUS (*Public Use Sample*) del Censo del 2000 de la Swiss Federal Statistical Office (2017) con el 5% de la población de los cantones suizos.

Inicialmente el PUS del año 2000 contiene 364.401 individuos, pero se excluyen los que viven en hogares colectivos (*ménages collectifs*), resultando un total de 348.238 individuos.

Aunque en Suiza hay 26 cantones, solo se consideran 25 "*small area*s", ya que en el PUS los individuos de los 2 semi-cantones, Appenzell Rodas Exteriores e Interiores (AR y AI), vienen codificados con el mismo valor.

Se seleccionan solo aquellos atributos y categorías indicados en la Tabla 7, y se eliminan aquellos individuos de los que falte información de alguno de los atributos (hay un 0,1% aproximadamente con valor de -1 en atributos distintos al máximo nivel educativo).

Atributo	**Categorías (número)**
sexo	hombre ; mujer (2)
edad	<15; 15-24 ; 25-34 ; 35-44 ; 45-54 ; 55-64 ; 65-74 ; >74 (8)
nacionalidad	suizo ; extranjero (2)
estado civil	soltero; casado; viudo; divorciado (4)
situación laboral	a tiempo parcial; a tiempo completo; desempleado; activo con otros ingresos; trabajo en casa; estudiante (escuela, instituto); pre-escolar; pensionista; otro sin ingresos (9)
máximo nivel educativo	sin indicación; sin estudios; educación obligatoria; aprendizaje profesional; madurez; escuela superior; escuela secundaria, universidad (7)

Tabla 7 Atributos y Categorías del escenario Suiza.

En el censo suizo, el valor del atributo de máximo nivel educativo para las personas de 14 años o menos es “-1”, por lo que se incluyen dentro de la categoría "sin indicación”.

También se van a realizar experimentos considerando el atributo edad con diferente número de categorías. Además de las 8 categorías indicadas en la tabla, se agruparán las edades en 3 categorías ($\leq$14; 15-64, y $\geq$65), por lo que se tendrán tablas de población cantonales con 3.024 y 8.064 celdas. La Figura 17 muestra un mapa de Suiza con los cantones considerados.

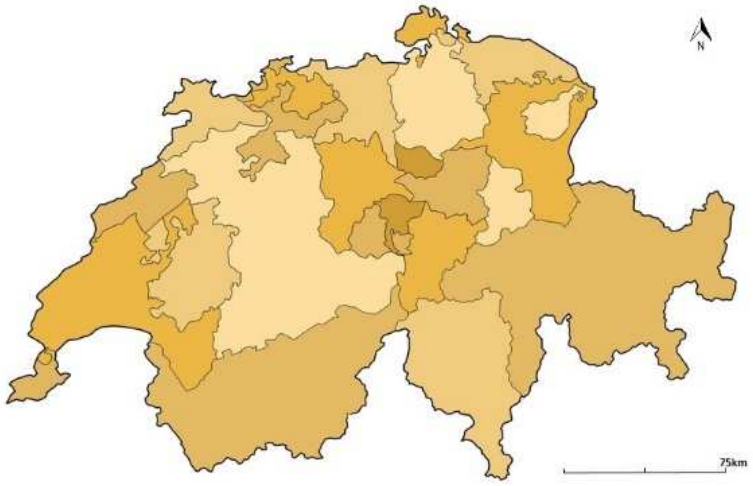

Figura 17 Mapa de Suiza con los 25 cantones codificados. Fuente: Dreamstime.com Copyright: Darknightsky.

6.3 Comparación para Andalucía

Cada combinación de valores de los tres parámetros de configuración explorados (tamaño de muestra, número de celdas y tamaño de las “*small areas*”) configura un escenario de experimentación. La discusión del escenario básico en la sección 6.3.1 contiene una comparación detallada, en términos de validación externa e interna, de los tres métodos seleccionados para el análisis comparativo: IPF-BLP, IPF-MCS y SA. La sección 6.3.2 resume los resultados obtenidos para los 17 escenarios adicionales, obtenidos mediante la combinación de dos o tres valores alternativos de cada uno de los tres parámetros de configuración.

6.3.1 Comparación en el escenario básico de Andalucía

Se ha seleccionado un escenario básico con el objetivo de describir con mayor detalle el experimento para una configuración típica, en la que se han fijado los tres parámetros con valores comúnmente utilizados en la práctica, dejando otros valores más extremos o inusuales para el análisis de sensibilidad:

- Tamaño de la muestra: 5% de la población de referencia, que es un valor habitual para muestras de uso público en el censo nacional.
- Número de celdas: las categorías descritas en la Tabla 6 conducen a una tabla multidimensional de 12.600 celdas (correspondiente a 2 x 14 x 2 x 5 x 9 x 5 posibles combinaciones de categorías de los atributos).
- Tamaño de la "*small area*": se ha elegido el nivel de municipio para "*small area*", generando poblaciones para las 60 "*small areas*" (el número de municipios considerados de la zona geográfica andaluza).

Siguiendo el procedimiento descrito en el capítulo anterior, se generan 5.400 poblaciones sintéticas para este escenario, que corresponden a las 60 "*small areas*" usando 30 muestras aleatorias, y con los tres métodos utilizados (SA, IPF-BLP e IPF-MCS).

6.3.1.1 Validación Externa

En la Tabla 8 se muestran los valores medios del error de clasificación 6-dimensional para el escenario básico con los tres métodos (% $CE^{6\text{-dim}}$). De acuerdo con estos resultados, las poblaciones de "*small areas*" sintetizadas con IPF-BLP tienen en promedio un 23,39% de individuos mal clasificados respecto a la población de referencia, mientras que en las poblaciones generadas con SA e IPF-MCS los errores promedio son más altos: 28,53% y 28,89%. Estos resultados ponen de relieve la crítica influencia del segundo paso en la precisión de la población generada con IPF y podrían ayudar a explicar los hallazgos de Ryan et al. (2009) quienes, en su comparación cuantitativa entre IPF y SA, encontraron que SA producía poblaciones con menos error que IPF. Esto podría deberse al hecho de que realizaban el segundo paso, posterior al IPF, con muestreo de Monte Carlo.

	IPF-BLP	**SA**	**IPF-MCS**
$\overline{\%CE^{6-dim}}$	23,39%	28,53%	28,89%

Tabla 8 Valor medio del Error de Clasificación 6-dim en el escenario básico de Andalucía.

La Tabla 9 muestra los resultados de la prueba t-test que compara SA con IPF-BLP, que confirman la conclusión sugerida en la comparación directa de los valores de $\overline{\%CE^{6-dim}}$. La media de la variable diferencia $\overline{\%CE_i^{6\text{-dim}}}$(SA) - $\overline{\%CE_i^{6\text{-dim}}}$(IPF-BLP) está comprendida entre 4,58% y 5,70% (α = 0,01). Como ambos extremos del intervalo son positivos, puede concluirse que en este escenario SA genera poblaciones con un error significativamente mayor que IPF-BLP. La significación es muy elevada (valor-p <0,0001). También se ha realizado la prueba de Shapiro-Wilk para contrastar la hipótesis de normalidad, obteniendo un p-valor >0,05, por lo que no se rechaza dicha hipótesis.

	$\overline{\%CE_i^{6\text{-dim}}}$(SA) - $\overline{\%CE_i^{6\text{-dim}}}$(IPF-BLP)
Media	5,14%
Intervalo de confianza para la media (α=0,01)	(4,58; 5,70)%
valor-p	<0,0001

Tabla 9 Resultados del t-test pareado de las diferencias del $\overline{\%CE^{6-dim}}$ entre SA e IPF-BLP, en el escenario básico de Andalucía.

En resumen, en el escenario básico, el método IPF con redondeo basado en programación lineal binaria es el que proporciona poblaciones sintéticas más precisas en relación a las poblaciones de referencia. Este método proporciona un porcentaje de error de clasificación 6-dimensional, cuya media está entre 4,58% y 5,70% por debajo del error que se obtiene con SA (α = 0,01).

6.3.1.2 Validación Interna

Se presentan aquí los resultados de la validación interna para el escenario básico utilizando la métrica (18) definida en la sección 5.4, es decir, el error de clasificación 1-dimensional. Los valores de la Tabla 10 muestran que, tanto IPF-BLP como SA cumplen perfectamente con las distribuciones de marginales objetivo de los seis atributos (primeras dos columnas de la tabla), mientras que el IPF-MCS produce desviaciones en todos los atributos, con un valor medio para todos los atributos de 1,54%. Estas desviaciones deben atribuirse exclusivamente al segundo paso del método, ya que el primer paso logra una coincidencia perfecta con respecto a esta métrica. Estos resultados son coherentes con la naturaleza de la técnica: IPF-BLP y SA siempre cumplirán con las distribuciones marginales objetivo si la muestra es lo suficientemente variada, mientras que la variabilidad asociada con la técnica de muestreo de Monte Carlo siempre conducirá a una cierta falta de coincidencia.

		IPF-BLP	**SA**	**IPF-MCS**
$\overline{\%CE_m^{1\text{-dim}}}$	sexo	0%	0%	0,87%
	edad	0%	0%	3,07%
	nacionalidad	0%	0%	0,23%
	estado civil	0%	0%	1,22%
	situación laboral	0%	0%	2,09%
	educación	0%	0%	1,64%
	valor medio de los 6 atributos	**0%**	**0%**	**1,54%**

Tabla 10 Resultados del Error de Clasificación 1-dim para los 6 atributos en el escenario básico de Andalucía.

6.3.2 Análisis de sensibilidad para Andalucía

Para completar el análisis realizado con los datos de Andalucía, se ha ejecutado un conjunto de experimentos adicionales destinados a explorar si los resultados se ven afectados por los valores de los parámetros: tamaño de muestra, número de celdas y tamaño del "*small area*" (ver sección 5.7). En este análisis de sensibilidad se utilizan los valores del escenario básico y se añaden dos nuevos valores para el tamaño de muestra y número de celdas, y un nuevo valor para tamaño de "*small area*":

- Tamaño de la muestra. En el escenario básico se usó una muestra del 5% de la población de referencia; aquí agregamos una muestra de menor tamaño, 1% y otra de mayor, 20%.
- Número de celdas. En el escenario básico se utilizaron 14 categorías para el atributo de edad. Aquí exploramos dos valores adicionales colapsando este atributo a 7 y 3 categorías. Esto conduce a tres valores alternativos para el número total de celdas en la tabla multidimensional: 2.700, 7.200 y 12.600 celdas.
- Tamaño de "*small area*". En el escenario básico, utilizamos 60 "*small areas*", con un tamaño medio de 3.800 individuos. Ahora se añade un nuevo escenario colapsando los 60 municipios en las 8 provincias andaluzas, que tienen un tamaño medio de 28.500 individuos. La base de datos utilizada solo contiene los atributos espaciales municipio y provincia.

En la gráfica de la Figura 18 se muestran los valores medios del error de clasificación 6-dimensional (% $CE^{6\text{-dim}}$) en los 18 escenarios resultantes de todas las combinaciones de valores de parámetros. Los resultados correspondientes a las técnicas de IPF se muestran en azul (azul oscuro para IPF-BLP, y azul claro para IPF-MCS), y los correspondientes a SA se muestran en naranja. Los escenarios se dividen en dos gráficos, el de la izquierda muestra los resultados de los escenarios con 60 "*small areas*", y el de la derecha muestra los obtenidos para 8 "*small areas*". El eje horizontal de ambos gráficos corresponde a los distintos tamaños de muestra, y el eje vertical al %$CE^{6\text{-dim}}$. Para cada técnica se ha dibujado una línea que conecta los resultados obtenidos en cada escenario con distinto tamaño de muestra. Por ejemplo, en el gráfico del lado izquierdo, la línea más baja (línea más gruesa azul oscuro) muestra los resultados obtenidos para IPF-BLP con tamaños de muestra de 1%, 5% y 20%, 60 "*small areas*" y 2,700 celdas. Las líneas naranja y azul claro superior muestran los resultados en los mismos escenarios para SA e IPF-MCS, respectivamente.

Los resultados para los escenarios con 7.200 y 12.600 celdas, también se muestran en estos gráficos con diferentes estilos de línea. Por tanto, los resultados para cada uno de los métodos IPF-BLP, IPF-MCS y SA, que corresponden a un mismo escenario, están representados por puntos con la misma abscisa, pertenecientes a líneas con el mismo estilo,

pero con diferentes colores. Por ejemplo, los resultados del escenario básico (que se muestran en la Tabla 8) están resaltados con un rectángulo en el gráfico de la izquierda.

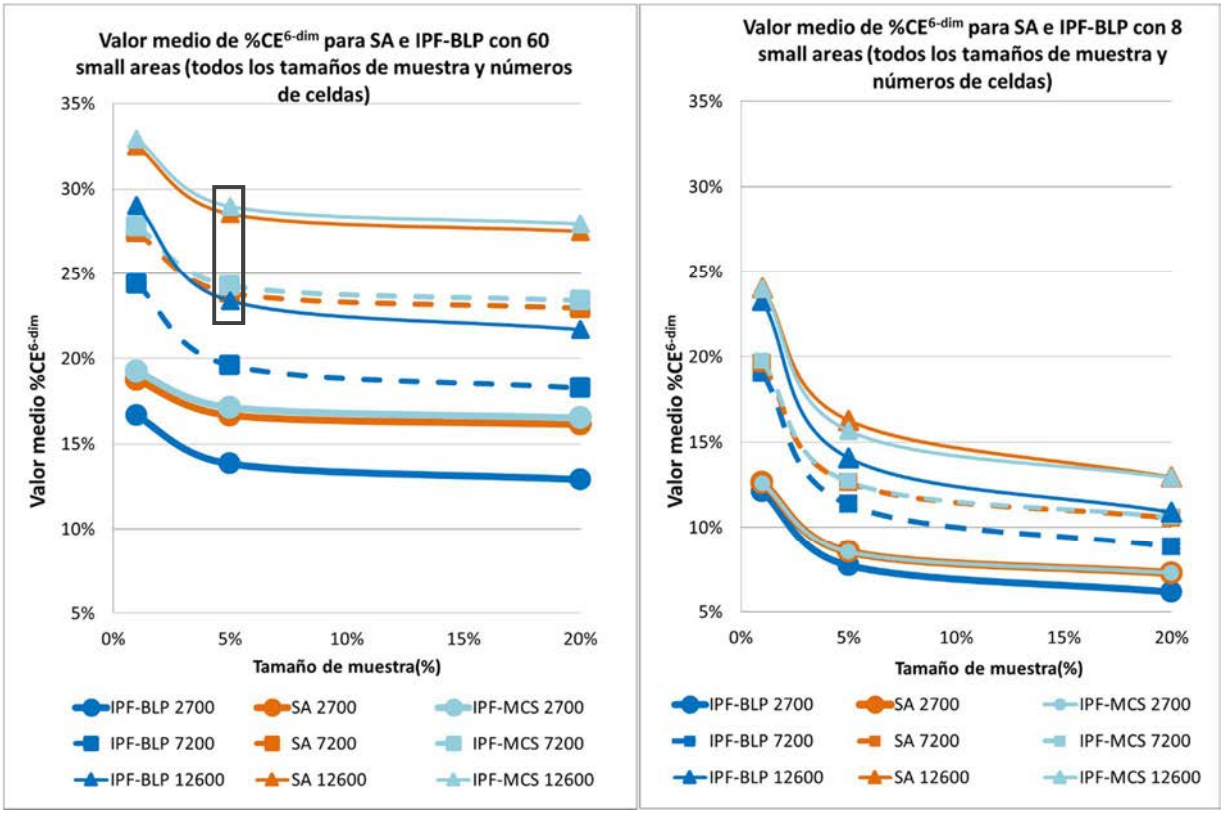

Figura 18 Valores medios de $\%CE^{6\text{-}dim}$ para IPF-BLP, IPF-MCS y SA con variación en el tamaño de la muestra, número de celdas y tamaño de la "small area" para el conjunto de datos de Andalucía.

	Intervalo de confianza (α=0,01) para la media de $\overline{\%CE_i^{6\text{-}dim}}$(SA) - $\overline{\%CE_i^{6\text{-}dim}}$(IPF-BLP)	
Número de celdas-Tamaño Muestra	**60 small areas**	**8 small areas**
2.700-1%	(1,80%; 2,41%)	(0,14%; 0,88%)
2.700-5%	(2,47%; 3,22%)	(0,32%; 1,34%)
2.700-20%	(2,84%; 3,60%)	(0,60%; 1,63%)
7.200-1%	(2,62%; 3,50%)	(-0,01%; 1,12%)
7.200-5%	(3,75%; 4,77%)	(0,56%; 2,05%)
7.200-20%	(4,21%; 5,15%)	(0,96%; 2,73%)
12.600-1%	(3,00%; 3,97%)	(0,11%; 1,41%)
12.600-5%	(4,58%; 5,70%)	(1,32%; 3,16%)
12.600-20%	(5,26%; 6,36%)	(1,43%; 3,20%)

Tabla 11 Intervalos de confianza de los t-test pareados, $\overline{\%CE_i^{6\text{-}dim}}$(SA) - $\overline{\%CE_i^{6\text{-}dim}}$(IPF-BLP) para todos los escenarios del análisis de sensibilidad de Andalucía.

Estos resultados confirman la conclusión principal del escenario básico: IPF-BLP es la técnica que produce las poblaciones más precisas.

Esto sucede en todas las situaciones evaluadas, para todos los tamaños de muestra, número de celdas de la tabla multidimensional y tamaños de "*small area*".

En la Tabla 11 se muestran los intervalos de confianza para las diferencias de % $CE^{6\text{-dim}}$ entre SA e IPF-BLP. Se observa que las diferencias de error entre esos métodos son en todos los casos estadísticamente significativas, excepto en un caso, el escenario con 7.200 celdas, 1% de tamaño de muestra y 8 "*small area*s". En este único caso, el intervalo de confianza contiene el valor cero, y por lo tanto, la diferencia entre SA e IPF-BLP no es estadísticamente significativa. Sin embargo, esta falta de significatividad podría estar relacionada con la utilización de solo 8 puntos (cada punto es el promedio de 5 muestras) para construir el intervalo de confianza.

Observando la Figura 18, pueden apreciarse ciertos efectos similares en las tres técnicas cuando varían los parámetros de configuración explorados en el análisis de sensibilidad. Con respecto al tamaño de la muestra, se observa para los tres métodos una disminución significativa en el error a medida que el tamaño de la muestra aumenta de 1% a 5%. Incrementos adicionales del tamaño de la muestra hasta el 20% producen menos mejora del error. Esto coincide con lo observado en el estudio de Ryan et al. (2009). En cuanto al número de celdas de la tabla multidimensional, el incremento de las mismas produce un aumento del error %$CE^{6\text{-dim}}$ para los tres métodos. Finalmente, hay una reducción generalizada del error en los escenarios en los que las poblaciones se colapsan en las 8 provincias (gráfico a la derecha), en comparación con el escenario más desagregado con 60 poblaciones (gráfico de la izquierda).

6.4 Comparación para Suiza

Para probar la sensibilidad de los resultados a la zona geográfica, se ha utilizado otro conjunto de datos de otro país y se han realizado de nuevo los experimentos, utilizando la misma metodología. En este caso, se han definido 6 escenarios combinando los siguientes valores de los parámetros de configuración:

- Tamaño de la muestra. Se han utilizado los mismos tres valores que en el caso de Andalucía, 1%, 5% y 20%.
- Número de celdas. Colapsando las ocho categorías de edad en tres, obteniéndose tablas con distinto número de celdas: 3.024 celdas (edad con 3 categorías) y 8.064 celdas (edad con 8 categorías).

- Tamaño del "*small area*". En el caso suizo, el mayor nivel de detalle espacial en el que se agrega la población de referencia en la PUS (Muestra de Uso Público) es el cantón. Por lo tanto, solo se ha considerado un valor para el tamaño de la "*small area*" que es el tamaño de las 25 "*small areas*" correspondientes a los cantones, con un tamaño medio de 13.900 individuos.

	Intervalo de confianza (α=0,01) para la media de $\overline{\%CE_i^{6-dim}}(SA) - \overline{\%CE_i^{6-dim}}(IPF-BLP)$
Número de celdas-Tamaño Muestra	**25 small areas**
3.024-1%	(0,96%; 2,10%)
3.024-5%	(1,56%; 2,77%)
3.024-20%	(1,81%; 3,14%)
8.064-1%	(1,34%; 3,05%)
8.064-5%	(2,41%; 4,23%)
8.064-20%	(3,09%; 4,99%)

Tabla 12 . Intervalos de confianza de los t-test pareados $\overline{\%CE_i^{6-dim}}(SA)$ - $\overline{\%CE_i^{6-dim}}(IPF-BLP)$ para todos los escenarios del análisis de sensibilidad de Suiza.

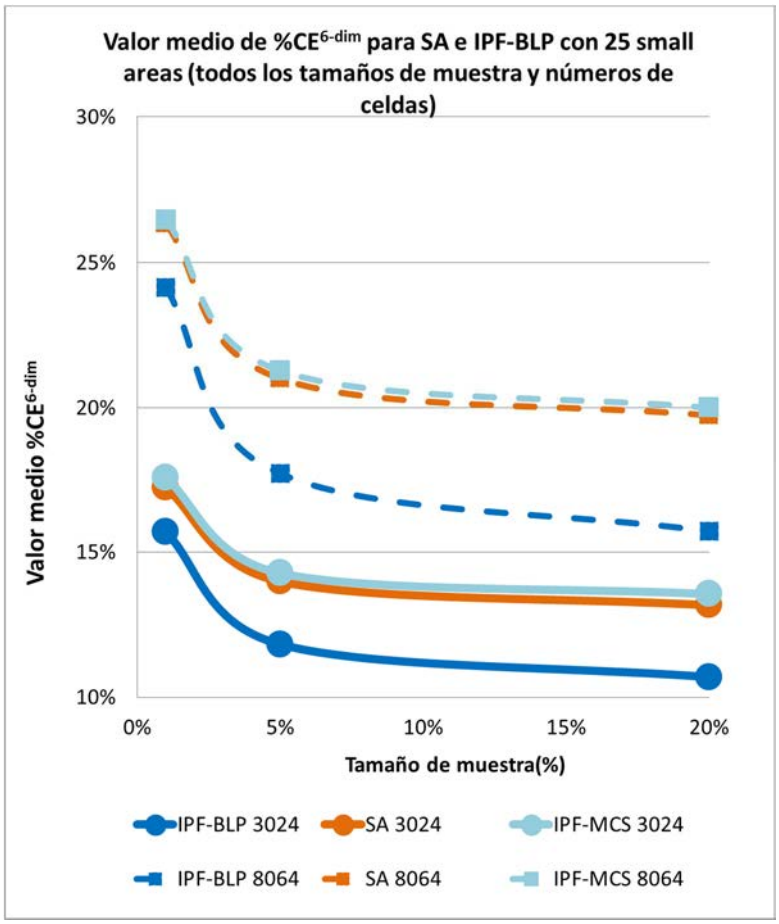

Figura 19 Valor medio de $\%CE^{6-dim}$ para IPF-BLP, IPF-MCS y SA con variación del tamaño de muestra, número de celdas y tamaño de "small area" para los datos de Suiza.

Los resultados, en términos de $\%CE^{6\text{-dim}}$ (Figura 19) e intervalos de confianza para las diferencias de errores (Tabla 12), conducen a las mismas conclusiones que en el caso de Andalucía. En todos los escenarios evaluados, IPF-BLP es la técnica que produce los valores de error de clasificación $\% CE^{6\text{-dim}}$ más pequeños. La Tabla 12 muestra que las diferencias en los errores son estadísticamente significativas en todos los escenarios.

La comparación de los resultados obtenidos para los distintos escenarios de Andalucía (Figura 18) y de Suiza (Figura 19) evidencia un comportamiento relativo equivalente de los tres métodos en las dos zonas geográficas, tanto en términos de rendimiento, como de impacto de los valores de los parámetros en el error. Esto refuerza las principales conclusiones del análisis, a saber, que el error producido por IPF-BLP es menor que el error producido por SA en todos los escenarios estudiados; y que SA proporciona unos resultados similares al IPF-MCS, aunque en general ligeramente mejor.

6.5 Conclusiones

Los métodos basados en IPF y SA son dos enfoques alternativos para generar microdatos espaciales sintéticos utilizados en un número creciente de campos. Solo se han publicado un número limitado de estudios comparativos de casos específicos, y no incluyen análisis de significación estadística. Varios autores han enfatizado la necesidad de una mayor investigación en esta área (Clarke & Harding, 2013).

En este capítulo, se ha seguido la metodología propuesta en el marco de referencia propuesto en esta tesis, para llevar a cabo una comparación sistemática entre los métodos. La piedra angular de esta metodología es la población de referencia, con la que las poblaciones sintetizadas se comparan, la cual es la población *real* a la que se supone representan. Por lo tanto, la divergencia entre las dos poblaciones, la sintetizada y la *real*, se calcula comparando cada par de celdas a través de las matrices multidimensionales de ambas poblaciones. La prueba de hipótesis se ha utilizado para establecer la significatividad estadística de los resultados obtenidos en cada escenario.

Se ha aplicado esta metodología a uno de los problemas más frecuentes de la literatura. Se han comparado dos técnicas, IPF y SA, identificadas como representantes relevantes dentro del enfoque de métodos de generación tipo *reweighting* de la muestra. Para el IPF, se han probado dos procedimientos alternativos de conversión a enteros, BLP y MCS.

Se han estudiado dos zonas geográficas usando microdatos censales: Andalucía (España) y Suiza. Dentro de cada zona se han analizado varios escenarios (18 y 6, respectivamente) para efectuar un análisis de sensibilidad.

Los resultados obtenidos en los 24 escenarios han sido bastante homogéneos. Una de las principales conclusiones que se obtienen del estudio es que, en el caso del IPF, el procedimiento utilizado para el paso de conversión a valores enteros afecta críticamente la precisión de los resultados. Si los números enteros se calculan utilizando MCS, los resultados suelen ser ligeramente peor que los obtenidos con SA, además de no conseguir ajustar las distribuciones de los marginales objetivos de la población. Sin embargo, cuando se usa el redondeo BLP propuesto por Choupani & Mamdoohi (2015), las poblaciones que se obtienen con IPF son más próximas a la población de referencia que las conseguidas con SA en los 24 escenarios, siendo la diferencia estadísticamente significativa en 23 de los escenarios. Por último, merece la pena resaltar que algunos de estos escenarios corresponden a tablas multidimensionales con gran cantidad de celdas cero, para las que en un principio podría pensarse que IPF no se considera la técnica adecuada (Wong, 1992). Aunque este análisis está circunscrito a un determinado problema (incluso si es altamente prevalente), y a dos geografías concretas, las conclusiones pueden ser útiles para los profesionales, particularmente teniendo en cuenta que el redondeo BLP fue propuesto después de que se hubieran publicado análisis comparativos donde distintos autores obtienen mejores resultados para SA que para IPF.

7 Comparación de técnicas de conversión a enteros en métodos de generación de poblaciones multinivel tipo *reweighting*

7.1 Introducción

En la mayoría de los modelos de demanda de transporte basados en actividad, y en muchos modelos de microsimulación, es necesario tener en cuenta la influencia de los atributos a nivel de persona y a nivel de hogar, para conseguir una representación realista del efecto de las circunstancias del viaje sobre la actividad y la elección del tipo de transporte. Así, los modelos basados en actividad requieren una representación de las poblaciones de individuos dentro de los hogares, para predecir cómo, cuándo y qué actividades llevarán a cabo dichos individuos.

Los requerimientos de disponer de estructuras de agentes multinivel en los modelos de simulación no solo se restringen a hogares e individuos, sino que también son aplicables a otros tipos de jerarquías, como familias dentro de los hogares, y a otros modelos de simulación que los precisen: empleados en empresas, vehículos disponibles en hogares, huéspedes en hoteles, etc. A lo largo de este capítulo nos referiremos a la jerarquía de hogares e individuos para hacer referencia en general a las estructuras de agentes multinivel.

El algoritmo IPF es la técnica más habitual para crear estas poblaciones de agentes con información sobre atributos socio-económico-demográficos y geográficos, proporcionada por las agencias estadísticas regionales y/o nacionales.

Con el algoritmo IPF es posible crear una población sintética de individuos o de hogares ajustada a los valores marginales objetivos establecidos para los atributos que describen la población. Pero este algoritmo no permite ajustar simultáneamente los marginales de los individuos y los hogares, esto es lo que se denomina el problema del ajuste multinivel.

En el caso de poblaciones con un solo nivel, el algoritmo IPF parte de una tabla multidimensional de agentes, correspondiente a una muestra de la población, y consigue generar una nueva tabla multidimensional ajustada a los marginales objetivos establecidos manteniendo la asociación entre los atributos de la muestra. Cada celda de la tabla representa la cantidad de agentes con una combinación determinada de atributos socio-económico-demográficos. Esta tabla se normaliza posteriormente, dividiendo los valores de las celdas por el total de agentes, y se obtienen valores fraccionarios que representan las probabilidades conjuntas. A partir de esta distribución de probabilidades conjunta de

atributos de los agentes de la población, se procede a la conversión a enteros, o generación de la población, mediante selección de agentes de la muestra según las probabilidades calculadas.

El método más usual para este proceso de conversión a enteros es el muestreo de Monte Carlo (MCS), con o sin reemplazo (X. Ye et al., 2009) donde la probabilidad de selección es proporcional a la probabilidad calculada. Existen otros métodos como el *Truncate-Replicate-Sample* de Lovelace & Ballas (2013), pero todos estos métodos de conversión a enteros producen poblaciones con valores marginales desajustados. En la sección 3.2.1.2 se ha descrito un método de conversión a enteros que, mediante programación lineal binaria, asegura que los valores marginales de las distintas categorías no se modifican (BLP).

En este capítulo se va a analizar el impacto que tienen las distintas técnicas de conversión a enteros en métodos de generación de poblaciones multinivel, en concreto, se centrará en comparar la estrategia de redondeo frente a la de muestreo de Monte Carlo.

Dado que la estrategia de redondeo solo se aplica en los métodos tipo *reweighting* que producen pesos no enteros, el análisis se centra en este tipo de métodos. Recordemos que hay métodos de *reweighting* heurístico y combinatorio tales como *Simulated Annealing* o Síntesis mediante Ajuste (FBS) que producen pesos con valores enteros.

En el capítulo 3 se han descrito los principales métodos de *reweighting*, que ofrecen una alternativa a los métodos basados en IPF para crear poblaciones con estructuras multinivel (hogares e individuos) a partir de una muestra multinivel.

Estos métodos de *reweighting* asignan pesos no enteros a cada uno de los hogares de la muestra (con sus individuos constitutivos), a partir de los que se procede al muestreo Monte Carlo, o al redondeo, para obtener la población sintética. A este tipo de métodos corresponden los métodos: IPF jerárquico (HIPF), IPU (X. Ye et al., 2009), Optimización de la Entropía (EO) (Bar-Gera et al., 2009), *Generalized Raking* (GR) (Deville et al., 1993) (lineal, multiplicativo, logit (L, U), truncado lineal también conocido como GREGWT), y también el heurístico Pop-H (Zhuge et al., 2017).

Algunos de estos métodos producen pesos iguales a los que se obtienen con el IPF (*Deterministic Reweighting*) cuando se utilizan con estructuras no jerárquicas (solo individuos o solo hogares), como es el caso del GR multiplicativo (también llamado *raking ratio* por Deville, Sarndal, & Sautory (1993)) y el de Optimización de la Entropía.

Estos dos últimos métodos son conceptualmente equivalentes, como puede comprobarse comparando las funciones de minimización (11) y (14) en las que se basan.

El estudio más representativo de comparación entre métodos de *reweighting* no entero ha sido elaborado por los desarrolladores del método HIPF (Müller & Axhausen, 2011). Estos investigadores compararon su método con IPU y Optimización de la Entropía (EO), utilizando en todos los métodos muestreo de Monte Carlo. Con cada método generaron poblaciones sintéticas multinivel para los cantones suizos con 7 atributos para cada nivel. Utilizaron una muestra aleatoria del 5% de los hogares y compararon las distribuciones conjuntas tridimensionales de las poblaciones resultantes con las mismas distribuciones conjuntas tridimensionales (de cada combinación de atributos) proporcionadas por el censo suizo. Concluyendo que su método HIPF producía mejores resultados que IPU y EO.

Posteriormente, Müller (2017a) realizó una nueva comparativa de estos métodos, usando pequeñas poblaciones de prueba con 2 y 3 atributos binarios, añadiendo a la comparativa los tres métodos de GR (lineal, multiplicativo y logit). Este autor se centra en el análisis cualitativo y de funcionamiento de los métodos, examinando en qué medida los pesos obtenidos se asemejan a los de la muestra inicial. El estudio concluye que el GR multiplicativo es el algoritmo que mejor grado de convergencia presenta y produce los pesos que más se asemejan a los iniciales de la muestra.

Se han verificado las conclusiones de este último estudio y se ha comprobado la superior convergencia del GR multiplicativo frente a los otros métodos, por lo que, dado que produce los mismos resultados que OE, pero con menos iteraciones, se ha seleccionado este método para analizar el impacto que tienen las distintas técnicas de conversión a enteros de los pesos asignados a la muestra. Con GR nos referiremos al GR multiplicativo.

También se han realizado las mismas comparaciones considerando los métodos de *reweighting* HIPF e IPU. Este último es el método de ajuste multinivel más popular en la literatura de generación de poblaciones sintéticas, con el que prácticamente todos los métodos se han comparado de uno u otro modo.

Tras comparar las dos estrategias de conversión a enteros con el método HIPF, se ha comprobado que los resultados obtenidos no tienen diferencia estadísticamente significativa con los resultados obtenidos con GR multiplicativo (α=0,01) por lo que, en este estudio solo se describe el análisis comparativo de las estrategias de conversión a enteros con los métodos GR e IPU.

A partir de los pesos calculados con cada método se procede a generar dos poblaciones sintéticas aplicando cada una de las estrategias mencionadas. Por un lado, se utilizará el tradicional muestreo de Monte Carlo con reemplazo, y por otro la técnica BLP de redondeo de los pesos. Aunque no existen referencias en la literatura acerca de la utilización de esta técnica de redondeo BLP al caso multinivel, se aplicará esta técnica adaptada a este caso,

siempre que se encuentre una solución factible, tal como se explicó en la sección 3.2.3.1. Por tanto, se realizan experimentos de generación de poblaciones con las dos estrategias aplicadas a los dos métodos.

En la siguiente sección se describen los datos de la población de referencia y los experimentos para el estudio comparativo. A continuación, en la sección 7.3 se presenta una comparación de la calidad de los pesos no enteros, previos al redondeo, obtenidos con IPU y GR. Esta comparativa permite entender mejor los resultados de los experimentos. En la sección 7.4 se describen y analizan los resultados obtenidos en los distintos experimentos de generación de poblaciones con las dos estrategias de conversión a enteros aplicadas a los dos métodos GR e IPU. Se finaliza el capítulo con las conclusiones que se derivan del análisis de los resultados.

7.2 Diseño de los Experimentos

En esta sección se describen los datos que constituyen la población de referencia y los experimentos realizados conforme a la metodología explicada anteriormente.

Al igual que en el capítulo anterior, la población de referencia que se utiliza en este segundo estudio comparativo se obtiene de los microdatos del censo de 2001 de Andalucía facilitados por el INE, los cuales representan el 5% de la población total.

Esta población multinivel "real" (aunque solo sea el 5% de la verdadera) está constituida por 74.142 hogares con 228.212 individuos de los 60 municipios andaluces con mayor número de residentes (Figura 20).

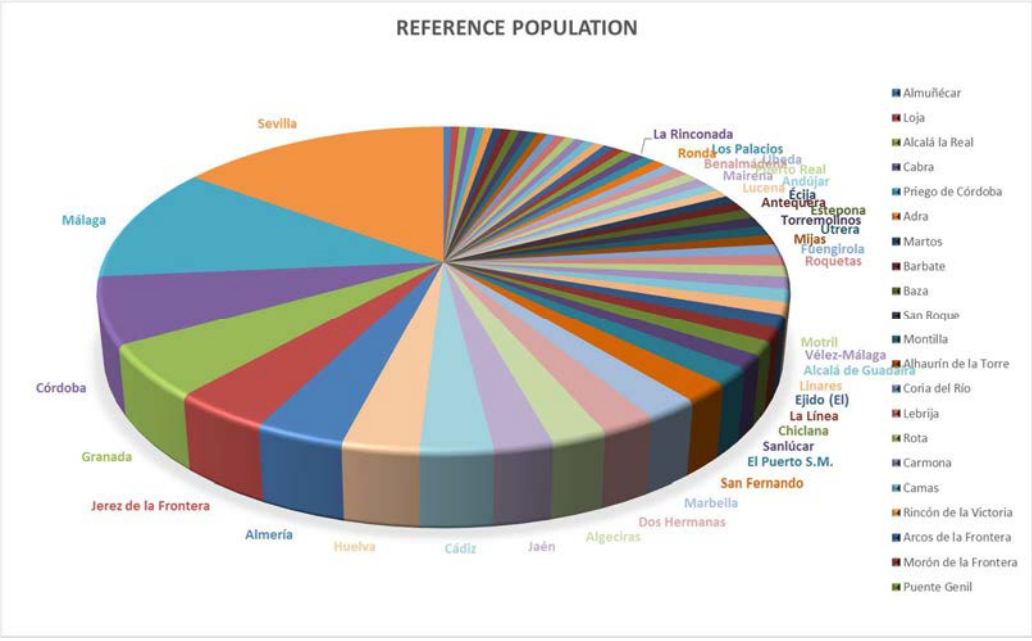

Figura 20 Los 60 municipios de Andalucía mayor población. Fuente INE, Censo 2001.

En la Tabla 13 se muestran los atributos y categorías que describen el nivel hogar, y en la Tabla 6 del capítulo anterior los correspondientes al nivel de individuo.

Atributo	Categorías (número)
Régimen de tenencia	En propiedad por compra, totalmente pagada; En propiedad por compra, con pagos pendientes (hipotecas…); En propiedad por herencia o donación; En alquiler; Cedida gratis o a bajo precio por otro hogar, la empresa..; Otras formas(6)
Número de coches	1;2;>=3; 0; (4)
Estructura del Hogar	Unipersonal; Un adulto y algún menor <16; Dos adultos con/sin menores<16; 3 adultos con/sin menores<16; 4 adultos con/sin menores<16; 5 o más adultos con/sin menores<16 (6)
Número de personas en el hogar	1;2;3;4;5;6;>=7(7)
Municipio	1 a 60 (60)

Tabla 13 Atributos y categorías del nivel Hogar.

Con estos atributos se obtiene, por cada uno de los 60 municipios, una tabla multidimensional para la población de hogares con 1.008 celdas y otra tabla con 12.600 celdas para la población de individuos.

A partir de la población de referencia se determinan los marginales de los atributos en cada municipio y se extraen 20 muestras aleatorias de distintos tamaños 1%, 3%, 5%, 10% y 20% de la población (total 20x5 muestras). Con cada una de estas 100 muestras se procede a la generación de las poblaciones sintéticas para los 60 municipios. Esto se realiza con cada método y cada técnica de conversión a enteros. Por tanto, en total se obtienen 24.000 poblaciones sintéticas multinivel.

La comparativa entre los procesos de conversión a enteros se realiza midiendo la similitud, celda a celda, entre las poblaciones generadas y la de referencia, para cada municipio, determinando cual se asemeja más a la población de referencia multinivel.

La semejanza se mide utilizando el índice de similitud Bray-Curtis (BCI), equivalente al error de clasificación %CE, y que tal como se indicó, permite la comparación de poblaciones de distinto tamaño, pues aunque el número de hogares sintetizados con el muestreo de Monte Carlo N^{hh} coincide con el objetivo impuesto en la población de referencia, esto no ocurre cuando se suman los individuos que componen dichos hogares. El número total de individuos de los hogares generados no coincide con el número de individuos objetivo de la población de referencia.

Como la población tiene dos niveles, hogares con M^{hh} atributos e individuos con M^{ind} atributos, para cada municipio i se tendrán dos tablas multidimensionales de referencia, una para hogares $O^{hh}_{ic_1 \dots c_{M^{hh}}}$ y otra para individuos $O^{ind}_{ic_1 \dots c_{M^{ind}}}$, siendo c_m el índice de categoría para el atributo "*m*". Por tanto, habrá un índice de similitud BCI para cada tabla (que para el caso de los hogares equivale al porcentaje de hogares incorrectamente clasificados respecto a la población de referencia (%CE)).

Estos dos índices, de hogares e individuos del municipio *"i"* obtenidos con la muestra *"s"* (1…20) son:

$$BCI^{hh}_{is} = \frac{\sum_{c_1=1 \dots}^{\mathbf{C_1}} \sum_{c_M=1}^{\mathbf{C_M}} \left| O^{hh}_{ic_1 \dots c_{M^{hh}}} - E^{hh}_{isc_1 \dots c_{M^{hh}}} \right|}{2\, N^{hh}_i} \cdot 100 \qquad (20)$$

$$BCI^{ind}_{is} = \frac{\sum_{c_1=1 \dots}^{\mathbf{C_1}} \sum_{c_n=1}^{\mathbf{C_M}} \left| O^{ind}_{ic_1 \dots c_{M^{ind}}} - E^{ind}_{isc_1 \dots c_{M^{ind}}} \right|}{N^{ind}_i + N^{ind}_{is}} \cdot 100 \qquad (21)$$

Siendo:

$O^{hh}_{ic_1 \dots c_{M^{hh}}}$ y $E^{hh}_{isc_1 \dots c_{M^{hh}}}$ = número de hogares de la celda $c_1 \dots c_{M^{hh}}$ de la tabla multidimensional de referencia del municipio i, y de la tabla calculada para el municipio *i* con la muestra *s*.

N^{hh}_i = número total de hogares de la tabla de referencia del municipio *i.*

$O^{ind}_{ic_1 \dots c_{M^{ind}}}$ y $E^{ind}_{isc_1 \dots c_{M^{ind}}}$= núm. de individuos de la celda $c_1 \dots c_{M^{ind}}$ de la tabla multidimensional de referencia del municipio *i*, y de la tabla calculada para el municipio *i* con la muestra *s*.

N^{ind}_i y N^{ind}_{is} =número total de individuos de la tabla de referencia del municipio *i* y de la tabla sintetizada para el municipio *i* con la muestra *s.*

Como ya se indicó, el índice BCI es una métrica de fácil de entender, ya que un valor de 0 indica que las poblaciones son exactamente iguales (las poblaciones son idénticas: $O^{ind}_{ic_1 \dots c_M} = E^{ind}_{isc_1 \dots c_M}$ para todos las combinaciones $c_1 \dots c_M$) mientras que un valor de 100 indica la máxima diferencia que puede observarse entre las poblaciones (poblaciones completamente disjuntas). Este índice, también conocido como índice de disimilitud, se utiliza en ciencias ambientales, especialmente para comparar poblaciones en estadística aplicada a la ecología (Bray & Curtis, 1957). Cuanta más disimilitud existe entre las poblaciones, mayor es el índice.

Con objeto de construir una única métrica que permita evaluar la semejanza de toda la población multinivel, se considera la media ponderada de los índices (20) y (21) de cada nivel,

ponderando con el número de atributos utilizados para describir cada nivel de la población, con lo que la métrica a utilizar es el índice:

$$BCI_{is} = \frac{na_{hh}\ \mathrm{BCI}_{is}^{hh} + \ na_{ind}\ \mathrm{BCI}_{is}^{ind}}{(na_{hh} + na_{ind})} \qquad (22)$$

Siendo: na_{hh} y na_{ind}el número de atributos que describen la población de hogares y de individuos, respectivamente. De esta forma se otorga una mayor relevancia a la población con mayor dificultad de ajuste.

Para comparar los índices BCI de las poblaciones generadas con cada método (GR-*Rounding*, GR-MCS, IPU-*Rounding* e IPU-MCS), se calcula el valor medio del índice BCI de las 20 poblaciones obtenidas con cada una de las 20 muestras (de un determinado tamaño):

$$\overline{BCI_{\imath}} = \frac{\sum_{s=1}^{20} BCI_{si}}{20}$$

Y como indicador global para cada método, se obtendrá el valor medio de los 60 valores $\overline{BCI_i}$ correspondientes a los 60 municipios:

$$\overline{BCI} = \frac{\sum_{i=1}^{60} \overline{BCI_i}}{60}$$

Seguidamente, tal como se explicó en la sección 5.6.2, se utiliza la prueba de rangos con signo de Wilcoxon para determinar si hay diferencia estadísticamente significativa entre los índices ponderados BCI's de las poblaciones generadas con cada una de las estrategias de conversión a enteros.

Esta prueba de contraste se realiza con las 60 parejas de valores $\overline{BCI_i}$ para cada municipio *i*, obtenidas con las dos estrategias (Navidi, 2015). Si $\overline{BCI_i}(_{GR\text{-}MCS})$ son los índices ponderados medios de la población del municipio *i*, obtenida con el muestreo de Monte Carlo aplicado a los pesos de GR, y $\overline{BCI_i}(_{GR\text{-}Rounding})$ los índices ponderados medios de la población del mismo municipio, obtenida aplicando el redondeo a dichos pesos, se determinan las 60 diferencias $\overline{BCI_i}(_{GR\text{-}MCS}) - \overline{BCI_i}(_{GR\text{-}Rounding})$ y las sumas de los rangos positivos S_+. Ya que la hipótesis nula es de la forma H_o: $\mu = 0$, se trata de una prueba de dos colas, en la que valores altos o bajos de S_+ proporcionan evidencia contra H_o. Un valor-p<α permite rechazar la hipótesis nula con un nivel de significación α (es decir, el primer método produce consistentemente poblaciones con mayor o menor disimilitud que el otro).

Esta prueba permite obtener un estimador e intervalo de confianza de la pseudomediana de la distribución de la variable diferencia asumiendo la simetría y continuidad de dicha

distribución. La proximidad entre los valores de la media y la mediana confirma la simetría de la distribución.

Se ha seleccionado una prueba no paramétrica que se comporta razonablemente bien ante violaciones moderadas de dicha asunción de simetría, ya que pruebas paramétricas como la t de Student requieren asunción de que las diferencias estén normalmente distribuidas (lo cual no ocurre con el IPU). Los resultados obtenidos con esta prueba son totalmente coherentes con los que se obtienen con otras pruebas con menos requisitos, como la prueba del signo.

Por último, a semejanza de la expresión (18) se define el índice el índice $BCI^{1\text{-dim}}(m)$ para cada atributo m (de hogar o individuo) con categorías $c_m = 1..\mathbf{C_m}$:

$$BCI^{1-dim}(m) = \frac{\sum_{c_m=1}^{\mathbf{C_m}} \left| O_{c_{m+}} - E_{c_{m+}} \right|}{N_O + N_E} 100 \qquad (23)$$

Donde O_{c_m+} son los valores marginales objetivos de la tabla de referencia y E_{c_m+} los de la tabla de población generada:

$$E_{c_m+} = \sum_{c_1=1}^{\mathbf{C_1}} \dots \sum_{c_{m-1}=1}^{\mathbf{C_{m-1}}} \sum_{c_{m+1}=1}^{\mathbf{C_{m+1}}} \dots \sum_{c_M=1}^{\mathbf{C_M}} E_{c_1 \dots c_m \dots c_M}$$

N_O , N_E = número de agentes de la tabla de observada (de referencia) y el de la tabla estimada.

Este índice se utilizará para llevar a cabo la validación interna de los métodos, ya que mide la diferencia entre los marginales unidimensionales de los atributos de las tablas de población generadas y los marginales unidimensionales objetivo.

7.3 Comparación de pesos obtenidos con GR e IPU

Antes de detallar los resultados de los experimentos de generación de poblaciones con las dos técnicas de conversión a enteros de los pesos, se estudia el nivel de ajuste que presentan los pesos que se obtienen con los métodos que se están analizando, pues la precisión del ajuste a los marginales impuestos es un factor clave para aplicar con éxito el proceso de redondeo.

Se han utilizado las implementaciones de GR e IPU del paquete R MultiLevelIPF (Müller, 2017b) para determinar los pesos (funciones *ml_fit_dss* y *ml_fit_ipu*). Con cada una de las 100 muestras aleatorias se ha generado una población para cada municipio, con cada método. Como hay 60 municipios, se han obtenido 6.000 conjuntos de pesos con cada

método. Cada conjunto de pesos ajusta, con distinta precisión, los hogares de una muestra a las distribuciones marginales de los atributos (de hogares e individuos) de un municipio.

La ejecución de estas funciones se da por concluida cuando se alcanza la convergencia o se alcanza un máximo de 10.000 iteraciones. La convergencia ocurre cuando la discrepancia relativa para todas las categorías c_m de todos los atributos es inferior a 0,001, esto es (siguiendo la notación introducida en las secciones 3.2.3 y 3.2.4 donde se describieron los métodos GR e IPU):

$$\left(\frac{|\sum_{k\in S} w_k x_{kc_m} - t_{c_m+}|}{t_{c_m+}}\right) < 0{,}001 \qquad \forall c_m = 1 \dots \mathbf{C_m} \quad \forall m = 1 \dots (M^{hh} + M^{ind})$$

Para determinar la precisión del ajuste (a los marginales objetivo) que se consigue con los pesos w_k de una muestra, que corresponden a la población de un municipio, se mide el Error Absoluto del marginal de cada categoría c_m, $AE_{c_m} = |\sum_{k\in S} w_k x_{kc_m} - t_{c_m+}|$ y se determina el valor máximo de este error entre todas las categorías, es decir, $Max\ AE_{c_m}$ es el error absoluto de la categoría con la que se tiene la mayor desviación respecto al valor marginal objetivo. Si este valor es cero $Max\ AE_{c_m} = 0$ habrá un ajuste perfecto para todas las categorías, por tanto:

$$\sum_{k\in S} w_k x_{kc_m} = t_{c_m+} \qquad \forall c_m = 1 \dots \mathbf{C_m},\ \forall m = 1 \dots (M^{hh} + M^{ind})$$

En general, el problema del ajuste multinivel puede tener múltiples soluciones, pero también es posible que no tenga solución, debido a que exista una inconsistencia o un marginal objetivo que entre en conflicto con la muestra, por ejemplo, se tenga una muestra que contenga exclusivamente hogares sin niños, pero los marginales objetivos impongan la existencia de niños. También podría deberse a que, aunque no existan inconsistencias, sea imposible ajustar todos los marginales objetivos simultáneamente, porque la muestra no sea suficientemente representativa y existan pocas posibilidades de combinación de hogares para conseguir los marginales objetivos.

Las muestras que se utilizan en los experimentos aquí planteados son consistentes con los marginales objetivos, y tienen la suficiente representatividad y tamaño (≥1% de la población; 741 hogares en el caso de muestras del 1%) de modo que siempre existen posibilidades de combinar hogares y obtener los marginales objetivos, tal como puede deducirse de los resultados obtenidos con el GR, que se describen a continuación.

En la Figura 21 se muestra la distribución de los valores $Max\ AE_{c_m}$ de los 6.000 conjuntos de pesos, correspondientes a las 6.000 poblaciones obtenidas con cada método, GR e IPU. El eje horizontal es una escala pseudologarítmica, donde cada uno de los siete puntos representa

un intervalo de valores, por ejemplo, el punto 5-49,99 representa el porcentaje de las 6.000 poblaciones en las que la máxima discrepancia de una categoría está comprendida entre 5 y 49,99, que para el caso de GR es 0, y para el caso de IPU representa casi el 40% de las poblaciones.

En casi el 90% de las poblaciones, los pesos que se obtienen con GR verifican la ecuación de calibración$\sum_{k\in S} w_k x_{kc_m} = t_{c_m+}$(10) con una precisión inferior a 0,5 para todas las categorías, y la probabilidad de obtener unos pesos con los que $Max\ AE_{c_m}$sea mayor que 1 es prácticamente nula (<0,001).

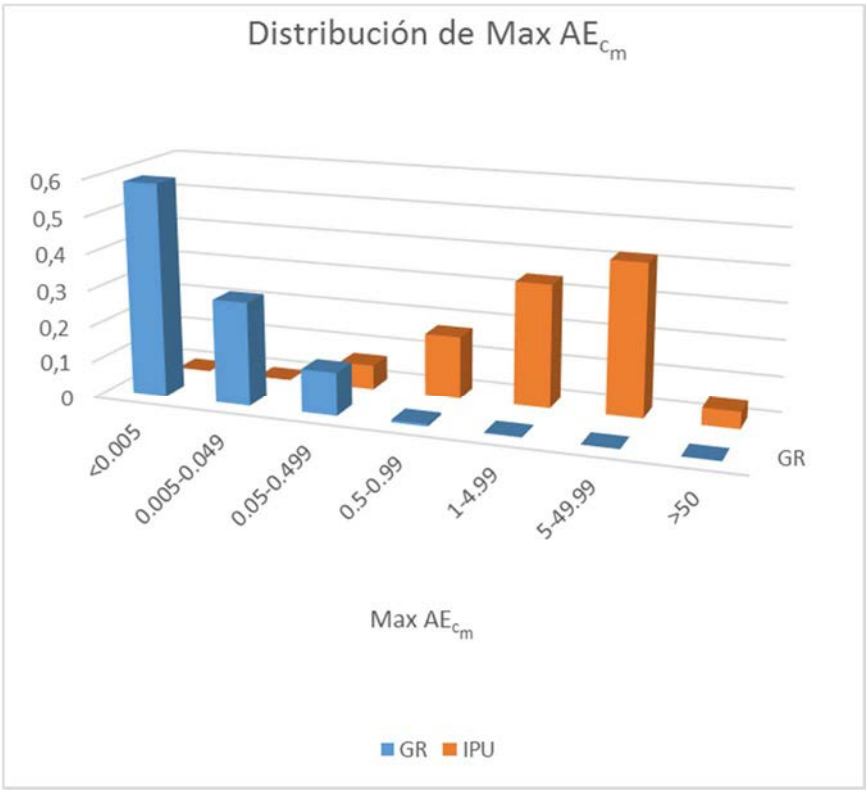

Figura 21 Distribución del Error Absoluto Máximo para los dos métodos: GR e IPU.

Sin embargo, la gráfica muestra que la convergencia de los pesos del IPU (a los marginales objetivo) no es tan satisfactoria, los pesos que se obtienen con este método corresponden a poblaciones que mayoritariamente tienen errores absolutos AE_{c_m}por encima de la unidad. Hay muy pocos casos en que el ajuste de marginales tenga el nivel de precisión similar al obtenido con los pesos de GR, $Max\ AE_{c_m} < 0{,}5$. Con el IPU existe una probabilidad superior al 90% de que el desajuste $Max\ AE_{c_m}$sea > 1, y aproximadamente el 40% de que $Max\ AE_{c_m} > 5$.

Estas diferencias entre la calidad del ajuste de los pesos de ambos métodos ocasionan que el redondeo de los mismos con la técnica BLP tenga distinto efecto en cada uno.

La técnica de redondeo BLP aplicada a los pesos obtenidos con GR, impone como restricción obligatoria el conservar el ajuste de los marginales, y consigue con los pesos redondeados que todos los AE_{c_m} de todas las categorías sean 0, tal como se describió en la sección 3.2.3.1.

Sin embargo, en el caso de los pesos del IPU (con $Max\ AE_{c_m} > 5$), cuando puede aplicarse la técnica de redondeo BLP, se mejora el ajuste y se consigue que los pesos redondeados logren que $AE_{c_m} = 0$.

Si fuera posible aplicar la técnica BLP en todos los casos de pesos del IPU, se tendrían poblaciones con el mismo nivel de ajuste de marginales que con los pesos de GR redondeados, pero no siempre es posible aplicar este redondeo.

Como ya se indicó, para que exista solución factible entera del redondeo BLP, se requiere que el máximo error sea $Max\ AE_{c_m} < 0{,}5$, asumiendo exista suficiente cantidad de representantes en la muestra; a medida que el $Max\ AE_{c_m}$ aumenta es menos probable encontrar una solución factible al redondeo BLP; por otro lado, el número de pesos > 0 a redondear es otro factor que influye para encontrar solución factible del redondeo, y a medida que disminuye este número es más difícil encontrar solución factible al problema de redondeo. Puede ocurrir que el algoritmo de *reweighting* asigne muchos pesos de la muestra con valor 0, y haya un número tan pequeño de pesos > 0, que no sea posible la solución de redondeo ajustada.

Con las muestras utilizadas en los experimentos (las del 1% tienen 741 hogares) los pesos generados con IPU pueden redondearse con la técnica BLP en más del 60% de los casos, en el resto de ocasiones nos vemos obligados a aplicar una técnica de redondeo diferente, tal como el *bucket*.

El redondeo tipo *bucket*, es una técnica de redondeo alternativa, que aunque modifica los marginales (previamente desajustados, en el caso del IPU) conserva el total de la población generada. Esta técnica es una opción incluida en la versión de código abierto de la implementación “PopGen”. Se trata de un redondeo secuencial en el que el “residuo” de cada operación de redondeo se retiene para aplicárselo a la siguiente operación, y así sucesivamente (“PopGen software,” 2017). Esta técnica de redondeo es la que se utilizará en los experimentos de generación de poblaciones planteados, en los casos en que no exista una solución factible de redondeo BLP.

Por tanto, en aproximadamente el 40% de las poblaciones obtenidas con IPU más redondeo tipo *bucket*, no se conseguirá el ajuste a los marginales objetivos impuestos, al igual que en las poblaciones que se obtienen aplicando la estrategia de muestreo de Monte Carlo, usando los pesos como una distribución conjunta de probabilidades.

En la siguiente sección se presenta la comparativa entre las precisiones de las poblaciones obtenidas con los dos métodos de GR e IPU, usando ambas estrategias. También se comparan los dos métodos, GR e IPU. Para encontrar la solución del modelo de programación lineal BLP se utiliza el paquete R glpkAPI (Fritzemeier, Dietrich, & Luangkesorn, 2015), que proporciona un interfaz de alto nivel del software GLPK (*GNU Linear Programming Kit*), un conjunto de rutinas de código abierto para resolver problemas de programación lineal y de programación entera mixta (*Mixed Integer Programming* MIP).

7.4 Resultados

7.4.1 Validación Interna

En primer lugar se presentan los resultados de la validación interna de las distintas combinaciones de método y estrategia de conversión a entero. En la Tabla 14 se muestran los valores medios de los errores de las 6.000 poblaciones obtenidas con cada uno de los métodos. La métrica de error utilizada es el índice de disimilitud $BCI^{1\text{-}dim}$ definido en la expresión (23). Este índice mide la discrepancia entre los valores marginales unidimensionales de los atributos de las tablas de población sintetizadas y los valores marginales unidimensionales objetivo que se imponen para sintetizar la población.

Dado que el número de hogares objetivo y estimado es el mismo en todos los casos, esta métrica es igual que el error de clasificación $\%CE^{1\text{-}dim}$ para los atributos de hogar. No ocurre así con los individuos sintetizados con MCS, ya que el total de individuos generados y el total de individuos objetivo no coinciden exactamente, por lo que el índice $BCI^{1\text{-}dim}$ es diferente al error de clasificación $\%CE^{1\text{-}dim}$. Los atributos con mayor número de categorías, como la edad o tamaño del hogar, son los que tienen mayor error.

Los valores de la Tabla 14 muestran que el redondeo produce poblaciones con un buen ajuste a las distribuciones de marginales objetivo, tanto de los cuatro atributos de hogares, como de los seis atributos de individuos (primeras dos columnas de la tabla).

En el caso del GR, el redondeo logra crear una población ajustada a unos valores marginales con gran precisión, con un error de clasificación de 2 agentes (hogares o individuos) por cada 10.000. Sin embargo, en el caso del IPU, este error, aunque pequeño, se multiplica por más de 10, por causa de la imposibilidad de aplicar la técnica de redondo BLP en todos los casos y tener que aplicar el redondeo *bucket* en un 40% de casos, con el que no se consiguen ajustar los marginales con precisión.

En el caso del muestreo de Monte Carlo, ambos métodos, GR-MCS y IPU-MCS, producen desviaciones en todos los atributos con un valor medio superior al 2%. Estas desviaciones

deben atribuirse exclusivamente a la variabilidad asociada a la técnica de muestreo de Monte Carlo.

		GR-*Rounding*	IPU-*Rounding*	GR-MCS	IPU-MCS
$\overline{BCI_{(m)}^{1\text{-}dim}}$	tenencia	0,03%	0,26%	3,02%	3,01%
	número de coches	0,02%	0,22%	2,73%	2,72%
	estructura del hogar	0,02%	0,19%	2,27%	2,25%
	tamaño de hogar	0,03%	0,29%	3,43%	3,45%
	sexo	0,01%	0,21%	0,98%	1,13%
	edad	0,03%	0,37%	3,45%	3,56%
	nacionalidad	0,01%	0,21%	0,96	1,01%
	estado civil	0,01%	0,24%	1,44%	1,69%
	situación laboral	0,02%	0,29%	2,32%	2,43%
	educación	0,02%	0,27%	2,07%	2,13%
	Valor Medio de los 10 atributos	**0,02%**	**0,25%**	**2,27%**	**2,34%**

Tabla 14 Valores medios del $BCI^{1\text{-}dim}$ para 4 atributos de hogar y 6 de individuo, de las poblaciones sintetizadas con cada método.

7.4.2 Validación externa

Tal como se explicó en la sección 7.2, con cada método y estrategia se ha calculado el valor medio del índice BCI ponderado (hogares e individuos) de las 1.200 poblaciones sintéticas de hogares e individuos obtenidas para los 60 municipios con las 20 muestras aleatorias de un tamaño dado. Recordemos que se han utilizado 20 muestras de cada tamaño (1%, 3%, 5%, 10% y 20%) para cada uno de los 60 municipios (20x60), por tanto, se tendrán 5 valores medios del índice BCI, ponderado según la expresión (22) con los números de atributos $na_{hh} = 4$ y $na_{ind} = 6$.

En la Figura 22 se muestran estos 5 valores medios del BCI. En el eje horizontal de la gráfica se representa el tamaño de muestra y en el eje vertical el valor del $\overline{BCI}$. Cada punto corresponde con un valor medio de 1.200 poblaciones para un tamaño de muestra, y las líneas discontinuas a lo largo del eje horizontal conectan los resultados obtenidos para distintos tamaños de muestra por cada método.

La Tabla 15 muestra los resultados de la prueba de rangos con signo de Wilcoxon para las 60 diferencias $\overline{BCI_i}(GR\text{-MCS})$ -$\overline{BCI_i}(GR\text{-}Rounding)$. En la misma se observa que, para tamaños de muestra inferior o igual al 5%, el valor medio del índice de disimilitud de las poblaciones

obtenidas con MCS es mayor que el de las poblaciones obtenidas con redondeo, sin embargo, esto no ocurre con muestras de mayor tamaño.

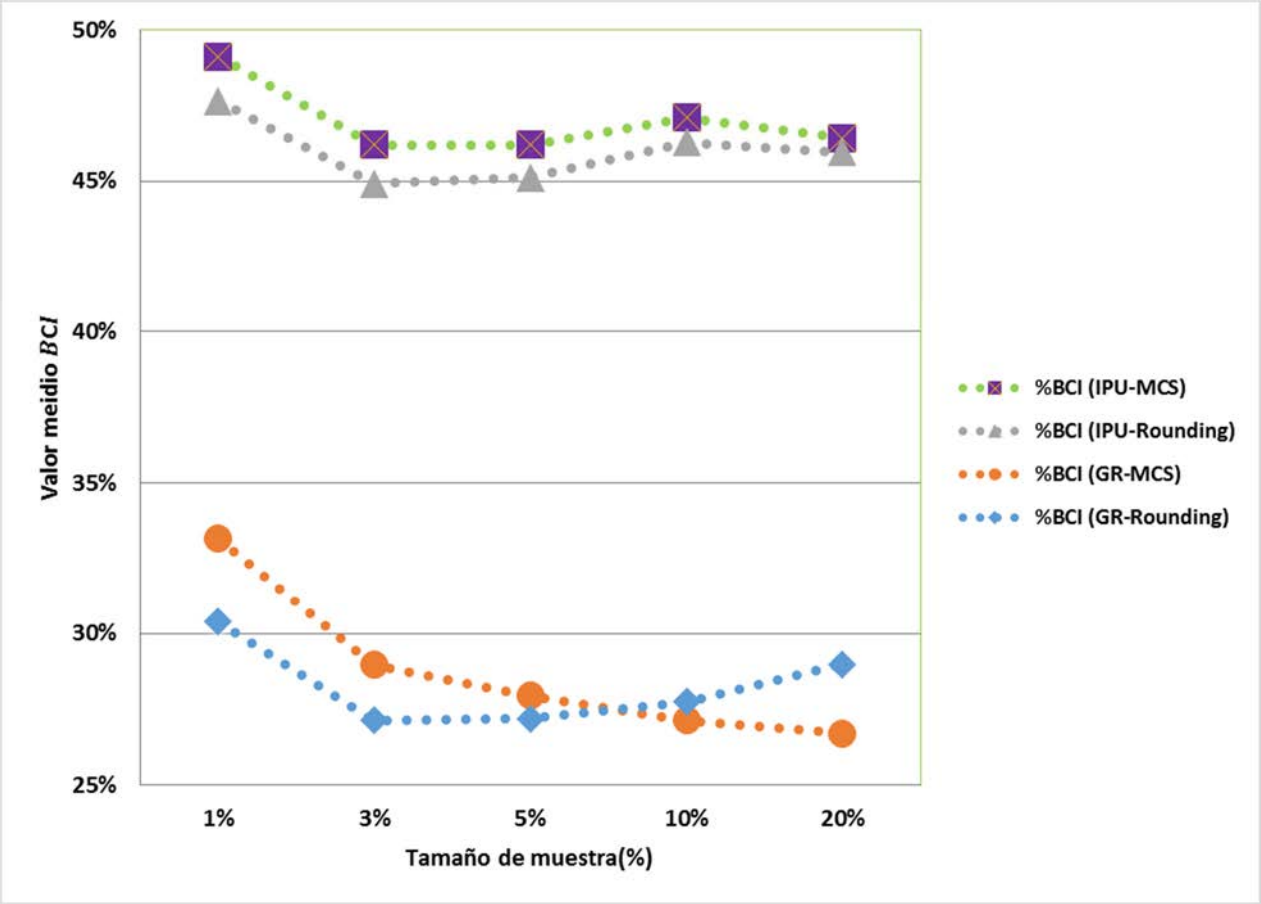

Figura 22 Valor medio del índice de disimilitud de Bray-Curtis para diferentes métodos y técnicas de conversión a enteros.

	$\overline{BCI_i}$(GR-MCS) -$\overline{BCI_i}$(GR-Rounding)				
Sample Size	1%	3%	5%	10%	20%
Media	2,77%	1,83%	0,76%	-0,60%	-2,29%
(pseudo)mediana	2,88%	1,88%	0,78%	-0,62%	-2,39%
Intervalo de Confianza (α=0,01)	2,49% ; 3,18%	1,66% ; 2,09%	0,49% ; 1,08%	-1,00% ; -0,28%	-2,77% ; -2,00%
S_+	1.830	1.767	1.643	341	18
valor-p	< 0,0001	< 0,0001	< 0,0001	< 0,0001	< 0,0001

Tabla 15 Resultados de la prueba de rangos con signo de Wilcoxon para las diferencias entre los índices de disimilitud de Bray-Curtis para GR con Redondeo y MCS.

En cambio, en la Tabla 16, donde se muestran los mismos resultados para el IPU, la diferencia es siempre positiva, por tanto, puede afirmarse que con IPU el redondeo produce menos disimilitud que el muestreo de Monte Carlo.

	$\overline{\%BCI_i}$(**IPU-MCS**) -$\overline{\%BCI_i}$(***IPU-Rounding***)				
Sample Size	1%	3%	5%	10%	20%
Media	1,47%	1,28%	1,08%	0,82%	0,47%
(pseudo)mediana	1,43%	1,28%	1,07%	0,82%	0,58%
Intervalo de Confianza (α=0,01)	1,19% ; 1,70%	1,10% ; 1,44%	0,92% ; 1,22%	0,68%; 0,95%	0,21% ; 0,90%
S_+	1.830	1.830	1.830	1.830	1.421
valor-p	< 0,0001	<0,0001	<0,0001	<0,0001	0,0002

Tabla 16 Resultados de la prueba de rangos con signo de Wilcoxon para las diferencias entre los índices de disimilitud Bray-Curtis para IPU Redondeo y MCS.

No se han podido determinar las causas de este distinto efecto del redondeo en ambos métodos, pero pueden señalarse algunos factores que pueden contribuir a este distinto comportamiento del redondeo de los pesos del IPU y del GR.

En primer lugar, y tal como se ha descrito en la sección 7.3, los pesos asignados por el IPU no ajustan la muestra a los marginales objetivo con la misma precisión que los pesos del GR. En más de la mitad de los casos (>60%), la técnica de redondeo BLP mejora el ajuste de marginales conseguido por los pesos del IPU sin redondear.

Sin embargo, para el caso del GR el redondeo BLP no modifica el ajuste de los marginales de la población, pues ya estaban ajustados con los pesos no enteros, por lo que no aporta ninguna mejora (en cuanto a marginales) respecto a la distribución de pesos no redondeados.

Un segundo factor que podría explicar el diferente efecto del redondeo, es el de la fragmentación de pesos del GR. Con muestras de tamaño superior al 5%, se produce un efecto de fragmentación de los pesos asignados por el método GR, lo cual reduce su efectividad. Cuando se usan muestras con un tamaño superior al 5%, GR asigna pocos pesos con valor >1, y además, dichos pesos son valores pequeños (en promedio no suman el 15% del total de hogares objetivo). La mayoría de los pesos asignados tienen valor < 1, estando muy fragmentados y concentrados por debajo de 0,1 (en promedio suman el 85% del total de hogares objetivo). A modo de ejemplo, en la gráfica de la Figura 23 se ha dibujado la densidad de los pesos obtenidos con una de las muestras del 20% para el primer municipio. Se observa una gran concentración (densidad) de pesos con valores muy pequeños, lo cual no ocurre con los pesos que se generan con IPU (trazo azul discontinuo).

Para comprobar las diferencias en las distribuciones de los pesos generados por ambos métodos, se han analizado los pesos de las 1.200 poblaciones generadas con 20 muestras aleatorias del 20% de la población (14.828 hogares) para los 60 municipios. En la Figura 24 se comparan las distribuciones de las sumas de pesos > 1 de estas poblaciones y las sumas de los pesos < 1. Se observa que con GR, los pesos inferiores a 0,1 representan una parte importante de la población sintética, por lo que el redondeo de los mismos es de vital

importancia para la construcción de la población, mientras que en el caso del IPU, con la parte entera de los pesos >1 ya se dispone de más de la mitad de la población, con lo que el redondeo determinará una pequeña proporción de hogares de la población.

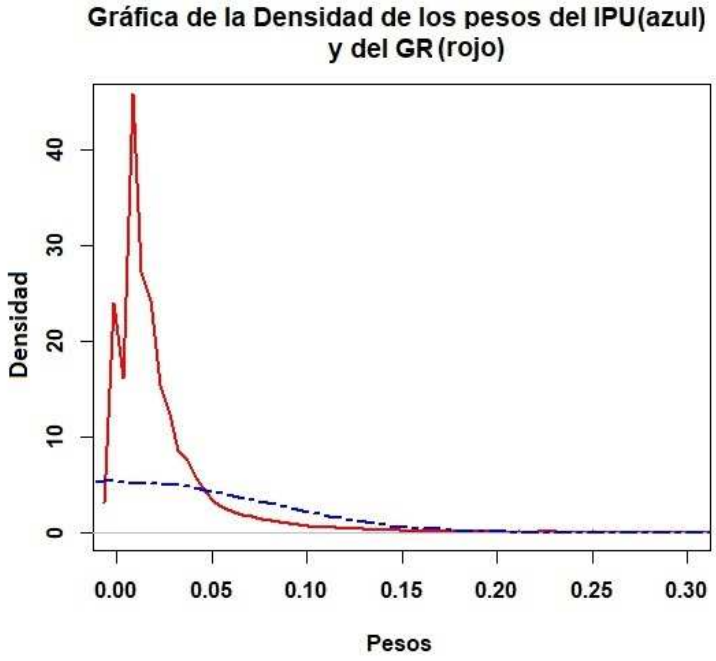

Figura 23 Densidad de Pesos obtenidos con GR e IPU, usando una muestra de 14.828 hogares.

Teniendo en cuenta esta distinta distribución de los pesos de ambos métodos, se deriva que el redondeo de los pesos de GR modifica fuertemente la distribución de pesos generada (alejándola de la condición óptima de minimización impuesta por el GR), mientras que en el caso de los pesos del IPU, el redondeo de los pesos modifica en menor grado (levemente) la distribución inicial de pesos generada por el IPU. Este hecho puede traducirse en un mejor comportamiento del muestro de Monte Carlo para el GR (con pesos no tan modificados) que utiliza los pesos iniciales sin redondear para el muestreo.

Por último, otra posible causa del distinto efecto del redondeo en GR, podría deberse a la propia técnica utilizada para resolver el problema de programación lineal binaria del redondeo. Esta técnica utiliza el *solver* GLPK MIP basado en el método *branch-and-cut* para encontrar una solución binaria factible, que no necesariamente es la óptima, ya que el algoritmo se interrumpe cuando se encuentra una solución factible a una distancia máxima del 10% del valor óptimo, con objeto de limitar el tiempo de computo. Dado que con muestras de tamaño superior al 5% el GR genera gran cantidad de pesos pequeños, el algoritmo de búsqueda de la solución de redondeo puede que produzca soluciones de peor calidad.

Considerando lo expuesto y los resultados de la Tabla 15 y Tabla 16, podría concluirse la conveniencia de utilizar el redondeo de los pesos obtenidos con GR y muestras de pequeño

tamaño (≤5%), sin embargo, con GR y grandes muestras la recomendación sería utilizar el muestreo de Monte Carlo.

Con respecto al método IPU, los resultados muestran que la técnica de redondeo BLP es una estrategia complementaria aconsejable, ya que mejora el ajuste de la población a los valores marginales objetivos y propicia poblaciones más precisas que las que se obtienen muestreando los pesos.

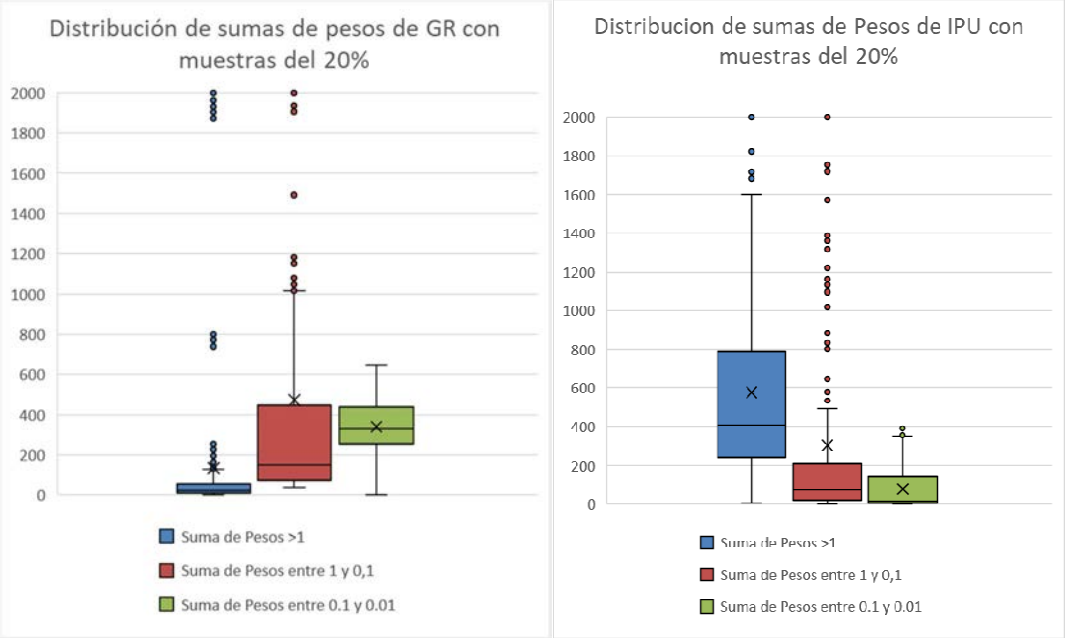

Figura 24 Diagramas de Cajas y Bigotes de las distribuciones de las sumas de los pesos >1 y <1 de las poblaciones de los 60 municipios sintetizadas con 20 muestras del 20%.

Influencia del tamaño de la "*small area*".

Además de estudiar la influencia del tamaño de la muestra en el rendimiento de los métodos y estrategias, los experimentos realizados permiten analizar el rendimiento de los métodos en relación con el tamaño del municipio, es decir, el tamaño de la "*small area*". En la Figura 25 se muestra la influencia del tamaño de la población de "*small area*" para el caso de muestras del 5%. La tendencia mostrada es la misma para cualquier tamaño de muestra, y para cualquier método. Hay una reducción en el índice de disimilitud (menor error) a medida que se sintetizan poblaciones de mayor tamaño.

La línea de tendencia potencial dibujada con trazo continuo $(y = \alpha * x^{\beta})$ muestra claramente la disminución del error con el tamaño de la población. El valor del coeficiente de determinación R^2 evidencia el buen ajuste de la línea de tendencia a los datos, sobre todo para el caso del GR.

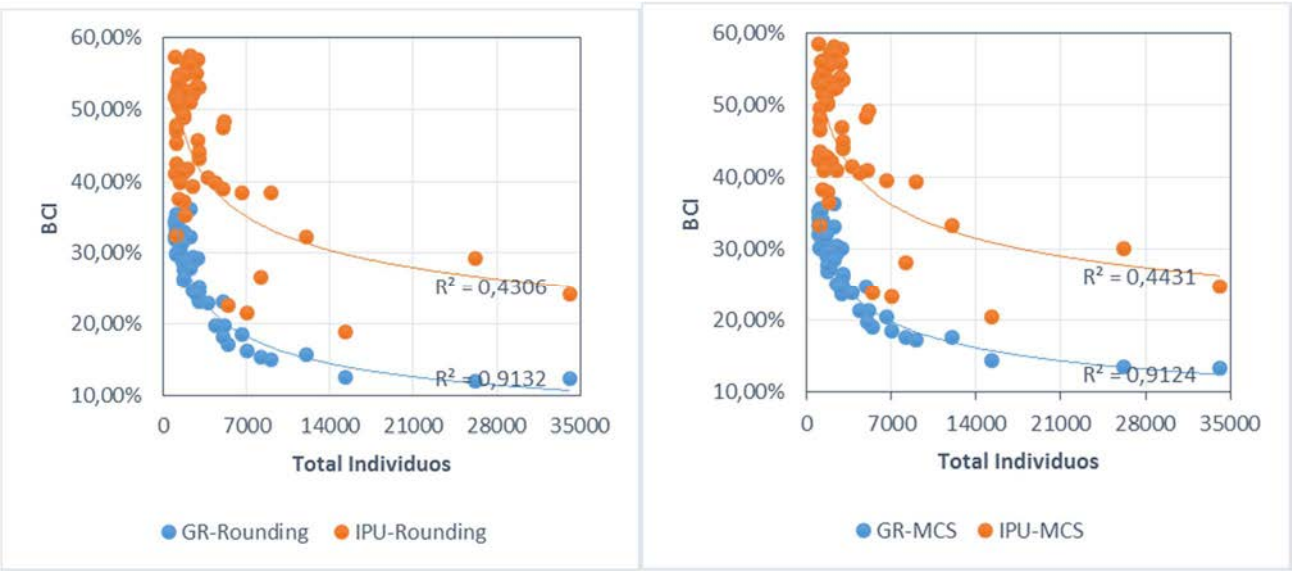

Figura 25 Valor medio del índice de disimilitud de Bray-Curtis para 60 municipios de distinto tamaño (tamaño de muestra del 5%).

Comparación entre métodos de *reweighting*.

La comparativa de los experimentos realizados también permite concluir que GR produce poblaciones claramente con menor error, o disimilitud, que el IPU, con independencia de la técnica de conversión a enteros utilizada, y del tamaño de muestra.

Tal como se indicó al comienzo del capítulo, también se han realizado experimentos comparando GR con IPF Jerárquico (HIPF), utilizando las funciones del paquete MultiLevelIPF.

Se han realizado los mismos experimentos con HIPF, aplicado las dos técnicas de conversión a enteros, MCS y redondeo, y se han comparado los resultados, tanto de $\overline{BCI_i}(GR\text{-}Rounding)$ - $\overline{BCI_i}(HIPF\text{-}Rounding)$, como $\overline{BCI_i}(GR\text{-MCS})$ -$\overline{BCI_i}(HIPF\text{-MCS})$, y en ninguna de las dos comparativas se han obtenido resultados cuyas diferencias sean estadísticamente significativas (α=0,05). Incluso, los pesos del HIPF adolecen del mismo efecto de fragmentación que los pesos del GR cuando se utilizan muestras de gran tamaño. En todos los casos, los intervalos de confianza de las diferencias de error contienen el valor 0. Por lo tanto, teniendo en cuenta estos resultados no puede afirmarse que un método sea superior a otro si ambos utilizan la misma técnica de conversión a enteros, tal como concluyeron los desarrolladores del método HIPF en su estudio. Las diferencias entre estos métodos se encuentran en la velocidad de la convergencia y el consumo de recursos computacionales.

En relación al método de Optimización de la Entropía, la semejanza de resultados con el GR multiplicativo es aún mayor, ya que como se ha explicado, se trata de dos métodos conceptualmente equivalentes, y las pequeñas diferencias provienen exclusivamente de la propia implementación de la resolución del problema de minimización.

7.5 Conclusiones

Los métodos de *reweighting* de la muestra, con pesos no enteros, son métodos habitualmente utilizados para generar poblaciones sintéticas multinivel, que se usan en campos como el de los modelos de demanda de transporte basados en actividad, con los que se determinan los patrones de desplazamiento de la población a partir de los valores de los atributos de las personas y de sus hogares.

Estos métodos asignan pesos a los hogares de la muestra, hogares con individuos, generando la población mediante un proceso de selección de hogares de la muestra utilizando los pesos como distribución de probabilidades para un muestreo de Monte Carlo (MCS). Así mismo, también existe la opción de convertir los pesos a enteros mediante técnicas de redondeo que utilizan programación lineal binaria (BLP), para generar de este modo la población directamente. Esta segunda opción no había sido evaluada en poblaciones multinivel, por lo que no hay estudios comparativos publicados sobre la conveniencia de la misma.

En este capítulo se ha aplicado la técnica del muestreo de Monte Carlo y la de conversión a enteros por redondeo, a los pesos de una muestra multinivel obtenidos con distintos métodos. Se han realizado comparaciones de las poblaciones resultantes, dentro del marco de referencia planteado en esta tesis, utilizando como medida del error el índice de disimilitud Bray-Curtis ponderado, con lo que se ha determinado la técnica y método que produce poblaciones con menor error, esto es, mayor similitud con la población de la que se extrae la muestra.

Se han comparado poblaciones generadas con métodos de *reweighting*: *Generalized Raking* multiplicativo (GR), *Iterative Proportional Updating* (IPU), IPF Jerárquico (HIPF) y Optimización de la Entropía (EO) usando el redondeo BLP y MCS. Se han considerado 5 tamaños de muestra diferentes, 20 muestras de cada tamaño, y 60 áreas geográficas.

Una conclusión importante es que la conversión de los pesos a enteros mejora sistemáticamente el rendimiento del método IPU, y, si la muestra es pequeña (<10%), también del GR multiplicativo.

Adicionalmente, no se han encontrado diferencias significativas de precisión entre los métodos GR multiplicativo, HIPF y EO. Sin embargo, con GR multiplicativo se han obtenido poblaciones más precisas que con IPU, con independencia del tamaño de la muestra y de la técnica de conversión a enteros utilizada. Con ambos métodos, GR e IPU, la precisión de la población sintetizada es mayor (la disimilitud con la población de referencia se reduce) cuanto mayor es el tamaño de la misma.

El análisis comparativo realizado aporta claridad en el campo de los métodos y algoritmos de generación de poblaciones sintéticas multinivel, facilitando a los investigadores nuevas valoraciones para la selección de métodos y técnicas de generación de poblaciones.

8 Estrategias para abordar el problema del marginal-cero

8.1 Introducción

Los métodos de generación de poblaciones sintéticas, IPF y *Simulated Annealing* (SA), tienen buen comportamiento con muestras no excesivamente pequeñas, como quedó de manifiesto en el capítulo 6. Sin embargo, no siempre es posible utilizar muestras de gran tamaño y a medida que el tamaño de la muestra disminuye, se incrementa el error de las poblaciones generadas, llegando a aparecer el problema del marginal-cero[7], esto es, en la muestra faltan representantes de alguna categoría de las presentes en los marginales objetivo (categoría faltante) y los métodos presentan problemas: IPF produce una división por 0, aparte de problemas de convergencia, y SA no puede alcanzar el objetivo de generación individuos de la categoría faltante.

Existen algunas estrategias para abordar estas situaciones, como son la de "reducción de categorías" o "redefinición de categorías" y el "*tweaking*[8]" de la muestra con pequeños valores arbitrarios. Sin embargo, estas estrategias tienen inconvenientes, lo cual hace que en ciertas ocasiones sean poco efectivas o no recomendables.

Por otro lado, en muchas ocasiones existe la posibilidad de obtener información adicional sobre la categoría faltante, proveniente de diversas fuentes, la cual puede combinarse con la muestra original. Las técnicas de "Fusión de Datos" permiten integrar dicha información con la muestra. Por lo que en estos casos en los que se dispone de información auxiliar sobre las "categorías faltantes", se puede utilizar para mejorar la generación de los individuos con dicha categoría dentro de la población sintética.

En este capítulo se analiza el problema del marginal-cero en la muestra, y se propone una estrategia aplicable a cualquier método de generación que utilice los datos de una muestra. Esta estrategia utiliza el enfoque de "Fusión de Datos" para modificar la muestra, pero solo en las categorías faltantes, apoyándose en el uso de información auxiliar sobre dicha categoría. Se analizará el efecto que tiene el enfoque propuesto sobre el rendimiento de los dos métodos citados, utilizando el marco de referencia planteado en esta tesis, y se verificará

[7] En inglés se utiliza el término *zero-marginal problem*, para referirse al problema de muestras con algún valor marginal nulo. Aquí se utilizará el término problema del marginal-cero.

[8] "Tweaking", literalmente "retoque", consiste en añadir un pequeño valor arbitrario a las celdas. Se utilizará el término en inglés introducido por Beckman et al. (1996).

que el nuevo enfoque mejora la exactitud de la población sintetizada, produciendo poblaciones con menor error de clasificación, y sobre todo, generando individuos de la "categoría faltante" con mayor precisión, lo cual en ciertos casos puede ser un factor diferencial de la calidad de la población sintética.

El capítulo se estructura del siguiente modo: en la siguiente sección 8.2 se describe el problema que plantea el marginal-cero y las alternativas de solución que se ofrecen en la literatura. Seguidamente, en la sección 8.3 se introduce el concepto de "Fusión de Datos" y se presenta la nueva estrategia para abordar el mencionado problema de marginal-cero, basada en la modificación de la muestra, con la que se consigue mejorar la fiabilidad de la población sintética. Posteriormente, en la sección 8.4 se describe el diseño de los experimentos realizados, para hacer un análisis comparativo entre las estrategias dentro del marco de referencia propuesto. Y finalmente se presentan los resultados (sección 8.5) y conclusiones (sección 8.6) de dicho análisis.

8.2 Celdas-Cero y Marginal-Cero

En el capítulo 2 de esta tesis se describieron distintos métodos y técnicas de generación de poblaciones sintéticas tipo *reweighting* de una muestra de la población que se desea "replicar". El IPF es una de las técnicas más utilizadas para la síntesis de poblaciones, que, tal como se ha explicado, permite estimar una tabla de distribución de probabilidades conjuntas de los atributos de la población, ajustadas a unas distribuciones de marginales de los atributos y preservando la asociación existente entre los atributos de una muestra. A partir de esta distribución de probabilidades se genera la población mediante selección de individuos de la muestra de acuerdo a estas probabilidades o mediante un proceso de conversión a enteros por redondeo de la tabla generada.

Otro método muy utilizado es el *Simulated Annealing*, que incluye un procedimiento de muestreo aleatorio para selección de los individuos de la muestra.

Ambos métodos han sido utilizados en múltiples estudios con muestras distinto tamaño (Beckman et al., 1996; Hanaoka & Clarke, 2007; Kim & Lee, 2016; Levy et al., 2014; J. Ma, Heppenstall, Harland, & Mitchell, 2014; Pritchard & Miller, 2012; Simpson & Tranmer, 2005).

En todas las tablas de estas muestras existen un gran número de celdas con valor cero, celdas-cero, llegándose a hablar del "problema de las celdas-cero", ya que esto puede llegar a causar la no convergencia del IPF. Algunas de estas celdas son "verdaderas celdas-cero", porque realmente no existe ese tipo de individuos en la población, pero otras celdas-cero son "falsas celdas-cero" porque al tratarse de una muestra es normal tener "observaciones faltantes", lo cual produce que algunos individuos de la población no estén representados en la muestra.

Por otro lado, cuando la población está caracterizada por atributos con categorías poco frecuentes, o la muestra es excesivamente pequeña, se incrementa el número de celdas-cero (falsas y verdaderas). Puede ocurrir que se tengan celdas cero en todas las celdas correspondientes a una de las categorías infrecuentes, con lo que en este caso se habla de "categoría faltante" en la muestra, o muestra con marginal cero, pudiendo ser un fenómeno que afecte a más de una categoría. Por tanto, se trata de dos problemas simultáneos: "problema de las celdas-cero", y el "problema de marginal-cero", este último es el que trataremos de abordar.

Por tanto, en función de la calidad y tamaño de la muestra puede presentarse esta situación de "categorías faltantes" en la muestra, ante la necesidad de generar una población de la cual se conoce a priori que ha de incluir individuos con dichas categorías, ya que se impone como restricción marginal de generación de la población sintética. Este hecho produce una división por cero en el caso del algoritmo IPF, aparte de los problemas de convergencia que presenta el propio algoritmo a medida que se incrementa el número de celdas-cero en la muestra, ya que uno de los efectos de las celdas-cero es el rebajar las posibilidades de convergencia del algoritmo IPF. Por tal motivo se habla del problema marginal-cero.

Con los métodos tipo *Simulated Annealing* (SA) no hay tal problema de convergencia propiamente dicho, sino que el marginal-cero en una categoría de la muestra produce un pérdida de efectividad del método, ya que aunque la función de bondad de ajuste (GoF) sea mínima, no es lo suficientemente buena, y al haber una categoría faltante en la muestra no será posible generar individuos con dicha categoría, por lo que no se podrá conseguir una población con los marginales objetivo para dicha categoría.

Para evitar los inconvenientes del problema del marginal-cero se han propuesto distintas alternativas en la literatura: redefinición de categorías, reducción de categorías y *tweaking*.

La redefinición de categorías fue estudiada por primera vez por Wong (1992) el cuál fue uno de los primeros investigadores que analizó el error de poblaciones obtenidas con IPF usando muestras de distinto tamaño. Este investigador planteó un esquema para establecer las categorías de forma que haya aproximadamente la misma cantidad de individuos en cada una (esquema de categorización de "igual tamaño") con lo que no se producen tantas celdas-cero en la tabla de la muestra, y se consiguen evitar los problemas de convergencia del IPF y del marginal-cero. Esto lo pudo hacer porque operaba únicamente con tablas bidimensionales y atributos numéricos de fácil redefinición, tales como edad e ingresos. Pero cuando las poblaciones son pequeñas y hay muchos atributos con categorías nominales representando propiedades cualitativas como nacionalidad o estado civil, cuyos valores no son numéricos, no es factible hacer una redefinición de categorías con este esquema de "igual tamaño", apareciendo una alta proporción de celdas-cero en la tabla multidimensional, lo

cual conduce a la falta de convergencia del IPF y a no conseguir evitar el problema del marginal-cero.

Una variante de la redefinición de categorías es la propuesta de Guo & Bhat (2007) basada en la reducción de categorías, suprimiendo las categorías poco frecuentes y combinándolas con otras para reducir la probabilidad de que se produzcan ceros en la tabla. Esta práctica conlleva el coste de pérdida de resolución de una variable, la cual se supone importante para el objetivo de la población, por ejemplo, la toma de decisiones del agente podría estar en función de alguna de estas categorías, por lo que si se eliminan, no sería posible basar la toma de decisión en el valor de dicha variable.

Beckman et al. (1996) abordó el problema de las celdas-cero, observaciones faltantes en la muestra, sin analizar el caso específico de toda una "categoría faltante" completa (el problema del marginal-cero es un problema de orden superior al problema de celdas-cero). Su propuesta contra el problema de celdas-cero fue el *tweaking* de las observaciones faltantes con pequeños valores arbitrarios, es decir, añadir a todas las celdas de la tabla de la muestra que tienen valor cero, un mismo valor arbitrario de ajuste mucho menor que 1 (por ejemplo, 0,001) de este modo solventaron el problema de la convergencia del IPF, pero no lograron mejorar la calidad del resultado.

También, Auld et al. (2009) describen y analizan el problema de las celdas-cero, señalando la correlación directa entre el número de "falsas celdas-cero" y los problemas de convergencia del IPF. Estos investigadores comparan las poblaciones obtenidas con IPF, utilizando la estrategia de "Reducción de Categorías", con las que se obtiene con IPF sin reducir de categorías. Esta comparación la realizan con referencia a totales reales conocidos, marginales de atributos no utilizados en el IPF como marginales objetivos, concluyendo a favor de la superioridad de esta estrategia, ya que produce poblaciones más ajustadas a los mencionados valores de marginales conocidos.

Recientemente, Suesse et al. (2017) han realizado un estudio de simulación del rendimiento del IPF y los métodos de ajuste de marginales descritos en la sección 3.2.2 (Máxima verosimilitud, Mínimo chi-cuadrado y Mínimos cuadrados) en tablas multidimensionales. Estos investigadores evitan los problemas de convergencia aplicando un tipo de *tweaking* diferente al propuesto por Beckman et al., que implica el ajuste de todas las celdas de la tabla, celdas-cero y distintas de cero. Este tipo de *tweaking* consiste en incrementar los valores de todas las celdas en una cantidad determinada "+α". Mediante simulación de múltiples muestras han investigado el rendimiento de los métodos con distintos valores de α: +0,5; +1; o +2, llegando a la conclusión que unos valores de *tweaking* de +0,5 y +1 son los más adecuados para el IPF, aunque esto implica un incremento del error relativo de las tablas estimadas.

En general, es común encontrarse en la literatura y libros de texto las recomendaciones de utilización de las estrategias indicadas: “Redefinición de Categorías” o “Reducción de Categorías” (CR), y el “*Tweaking*” (TW), para abordar el problema del marginal-cero (Janssens, Yasar, & Knapen, 2014).

Dado que estas recomendaciones no ofrecen una solución muy convincente, en algunas investigaciones se opta por evitar el problema del marginal-cero, seleccionando exclusivamente muestras que no adolezcan de dicho problema, con lo que evitan tener que aplicar alguna de las alternativas descritas para tratar la situación.

8.3 Fusión de Datos y Estrategia Propuesta contra el marginal-cero

En esta sección se explica la estrategia que se propone para abordar el problema del marginal-cero. Dado que esta estrategia se basa en el concepto de “Fusión de Datos” (“*Data Fusion*”), se introduce este concepto.

La Fusión de Datos (también llamada *statistical matching*) es una metodología de integración de datos que proporciona los medios para combinar información proveniente de distintas fuentes en un único conjunto de datos.

La Fusión de Datos utiliza múltiples técnicas (imputación, regresión, *hot deck*, *log linear*, técnicas paramétricas, etc.) para combinar información de dos o más fuentes distintas, las cuales contienen algunas variables comunes y otras no comunes (sobre un conjunto de individuos que pertenecen a la población de interés). El objetivo es conseguir estudiar la relación existente entre variables no observadas conjuntamente (no comunes), como sería el caso de conseguir la distribución conjunta de todas las variables (comunes y no comunes), o conseguir un único conjunto de datos con todas las variables (Palaumbo, Montanari, & Vichi, 2017).

Ambos resultados pueden conseguirse mediante dos posibles enfoques, uno macro y otro micro, tal como los define (D’Orazio, Di Zio, & Scanu, 2006). El enfoque macro tiene como objetivo el transformar los conjuntos de datos en resultados agregados, identificando estructuras que describan la relación entre las variables de las dos fuentes que no han sido observadas de manera conjunta (por ejemplo, distribuciones conjuntas). El objetivo del enfoque micro es construir un conjunto de datos sintético completo, transformando los distintos conjuntos de datos en uno solo integrado, cuyos registros se refieren a la misma unidad de análisis (los individuos) y donde todas las variables de interés están presentes.

El uso de la “Fusión de Datos” está creciendo cada día, principalmente debido al incremento de las fuentes de datos. Una de sus principales aplicaciones es en los estudios de investigación

de mercado (Roser, Aluja-Banet, & Nonell, 1999) donde se utiliza para combinar datos provenientes de bases de datos con información de encuestas de mercado.

En esta área se presenta el caso típico de fusión de datos, cuando una compañía tiene un fichero de clientes con variables del histórico de compra (*A*) e información socio-económica (B_1) de los clientes, y desea identificar los clientes con alta probabilidad de compra de un nuevo producto. En estos casos, la compañía decide realizar una encuesta anónima a una pequeña muestra de clientes, preguntando de forma anónima sobre el histórico de compra del cliente y la intención de compra del nuevo producto (B_2). Posteriormente se procede a combinar la información de la encuesta con el fichero de clientes para conseguir identificar los clientes que comprarán su nuevo producto.

En la Figura 26.a se representa la combinación del fichero de clientes con la encuesta de mercado. Esta combinación se hace apoyándose en las variables comunes ($A=\{x_m , m = 1..M\}$) de cada uno de los conjuntos de datos, con el fin de obtener el fichero de datos de clientes extendido con todas las variables $\{A, B_1, B_2\}$. Al fichero de datos de clientes que se desea extender se denomina "receptor" y a la encuesta de mercado de la cual se determina la intención de compra, "donante".

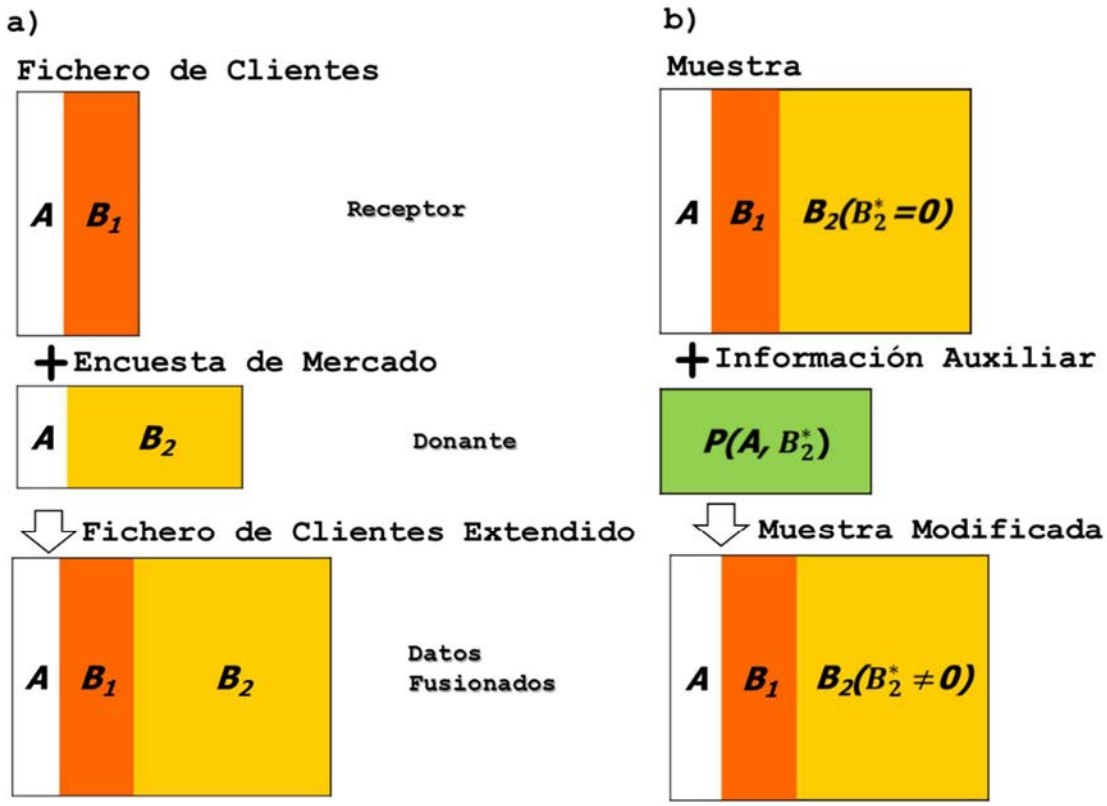

Figura 26 Dos escenarios de Fusión de Datos.

Una de las técnicas más fácil de implementar para conseguir la combinación de los datos es la de "distancia *hot deck*" (Andridge & Little, 2010) donde a cada registro del "receptor" se

asigna el registro más cercano del fichero "donante". Para medir la distancia entre registros existen distintas posibles definiciones, las cuales se calculan utilizando las variables comunes de ambos ficheros *A={x_m , m = 1..M}*, como por ejemplo, la distancia euclidiana, $d(i,j) = \sqrt{\sum_{\mathrm{m=1\ ...}}^{\mathrm{M}}(x_{\mathrm{im}} - x_{\mathrm{jm}})^2}$, la máxima desviación $d(i,j) = \max_m |x_{\mathrm{im}} - x_{\mathrm{jm}}|$ o una distancia ponderada con distinto peso a las variables.

Con objeto de conseguir que esta técnica produzca un buen resultado, las variables que se utilizan en la definición de distancia han de tener poder explicativo, es decir, ser buenos predictores de B_2, y los valores de B_2 han de ser independientes de los valores de B_1 para cualquier valor dado de *A*, esto es lo que se conoce como la asunción de independencia condicional (*Conditional Independence Assumption* CIA). En nuestro ejemplo, esto se traduce en que los datos históricos de compra del cliente sean un buen predictor (puedan explicar) de la propensión de compra y la propensión de compra sea independiente de las características socio-económicas del cliente dado el histórico de compra.

En el caso en el que todas las combinaciones de valores de las variables comunes (*A*) presentes en el fichero "receptor" estuvieran asimismo presentes en el fichero "donante", la distribución condicional conjunta de B_1 y B_2 dado *A* es:

$$p(B_1, B_2|A) = p(B_1|A)p(B_2|A)$$

Por tanto, la distribución conjunta de todas las variables $p(A, B_1, B_2)$ puede determinarse estimando $p(B_1|A)$ y la distribución marginal $p(A)$ del "receptor", y estimando $p(B_2|A)$ del "donante".

$$p(A, B_1, B_2) = p(B_1|A)p(B_2|A)p(A)$$

La distribución condicional $p(B_1|A)$ puede obtenerse a partir de la tabla multidimensional de datos agregados del "receptor" $Recep(A, B_1)$:

$$p(B_1|A) = Recep(A, B_1)/Recep(A) = Recep(A, B_1) \Big/ \sum_{\forall B_1} Recep(A, B_1)$$

Donde $Recep(A, B_1)$ representa las celdas de la tabla "receptor" con dimensiones *A* y B_1, y $Recep(A)$ las celdas de la tabla unidimensional A, es decir, los valores marginales de *A*.

Para el caso del problema marginal-cero se propone aplicar un esquema lógico análogo para completar las celdas de la muestra correspondientes a las "categorías faltantes". La estrategia que se propone consiste en utilizar información auxiliar específica de baja dimensionalidad sobre dichas categorías, e incorporarla en la muestra en forma de nuevos individuos que tienen la "categoría faltante", con objeto de evitar el problema del marginal-cero. La muestra jugará el papel de "receptor", y la información auxiliar de "donante".

En la literatura de generación de poblaciones sintéticas existen precedentes sobre el uso de información auxiliar (X. Ye et al., 2009). En ese caso se disponía de una muestra de una población de un área determinada, la cual tenía el problema de celdas-cero, es decir, "individuos faltantes" en algunas celdas, pero se disponía de otra muestra (Información auxiliar) de una área superior más amplia con las mismas variables, la cual no tenía valores cero en las mismas celdas. La estrategia consistió en reemplazar la muestra inicial con una distribución de probabilidad, tomando como valor de las celdas-cero el de la muestra del área más amplia, y modificando el resto de probabilidades para que la distribución completa sume 1 (multiplicando cada valor por 1 menos la suma de las probabilidades extraídas de la muestra del área de mayor tamaño).

En la Figura 26.b aparece como "receptor" la tabla de la muestra $S(A, B_1, B_2)$ con el problema del marginal-cero, donde (A, B_1, B_2) representa el conjunto de atributos que caracteriza la población de la muestra. A y B_1 son conjuntos de atributos disjuntos, y B_2 el atributo con la "categoría faltante" la cual se representa por $\mathbf{B_2^*}$. El "donante" aparece como información auxiliar agregada que describe la relación entre la categoría $\mathbf{B_2^*}$, y A, por ejemplo, valores porcentuales, es decir, la distribución de probabilidad condicional $p(\mathbf{B_2^*}|A)$, siendo A uno o varios atributos usados en la muestra, que no incluye el B_2.

El objetivo es establecer un valor para las celdas de la muestra correspondientes a la categoría faltante $S(A, B_1, B_2 = \mathbf{B_2^*})$ (que inicialmente son 0), dada la información p($\mathbf{B_2^*}|A$), sin alterar el resto de celdas de la muestra $S(A, B_1, B_2 \neq \mathbf{B_2^*})$, y conseguir así una muestra modificada $S^{mod}(A, B_1, B_2)$, que no tenga valores 0 en las celdas $S^{mod}(A, B_1, B_2 = \mathbf{B_2^*})$. Por lo tanto, tras incorporar la información auxiliar, en la muestra modificada se cumplirá:

$$p(\mathbf{B_2^*}|A) = \frac{S^{mod}(A, B_2 = \mathbf{B_2^*})}{\sum_{\forall B_2} S^{mod}(A, B_2)}$$

$S^{mod}(A, B_2)$ representa la misma tabla colapsada, sin los atributos B_1. El sumatorio del denominador es sobre todas las posibles categorías del atributo B_2.

Si se acepta la asunción de independencia de los atributos B_1 y B_2 dado *A*, esto implica que:

$$S^{mod}(A, B_1, B_2) = S^{mod}(A, B_2) \cdot p(B_1|A)$$

Por lo que también es cierto que:

$$p(\mathbf{B_2^*}|A) = \frac{S^{mod}(A, B_1, B_2 = \mathbf{B_2^*})}{\sum_{\forall B_2} S^{mod}(A, B_1, B_2)}$$

Como hemos dicho que la tabla $S^{mod}(A, B_1, B_2)$ es igual que la tabla de la muestra $S(A, B_1, B_2)$, salvo en las celdas correspondientes a la categoría faltante $B_2 = \mathbf{B_2^*}$, la anterior expresión puede escribirse como:

$$p(\mathbf{B_2^*}|A) = \frac{S^{mod}(A, B_1, B_2 = \mathbf{B_2^*})}{\sum_{\forall B_2 \neq \mathbf{B_2^*}} S(A, B_1, B_2) + S^{mod}(A, B_1, B_2 = \mathbf{B_2^*})}$$

De donde se tiene que:

$$S^{mod}(A, B_1, B_2 = \mathbf{B_2^*}) = \frac{p(\mathbf{B_2^*}|A) \sum_{\forall B_2 \neq \mathbf{B_2^*}} S(A, B_1, B_2)}{(1 - p(\mathbf{B_2^*}|A))}$$

El sumatorio del numerador se corresponde con las celdas de la tabla de la muestra con el atributo B_2 colapsado, $S(A, B_1)$ (ya que las celdas que se excluye del sumatorio son nulas en la muestra $S(A, B_1, B_2)$). Por tanto las celdas de la muestra con la categoría faltante pueden estimarse con la expresión:

$$S^{mod}(A, B_1, B_2 = \mathbf{B_2^*}) = \frac{p(B_2^*|A)\, S(A, B_1)}{(1 - p(\mathbf{B_2^*}|A))} \qquad (24)$$

Con este número de agentes, de la categoría $\mathbf{B_2^*}$, la condición aportada por la información auxiliar se cumple en la muestra modificada. Se han añadido lo individuos necesarios, con la categoría "$\mathbf{B_2^*}$ ", para se cumplan los porcentajes dados en la información auxiliar.

La información auxiliar a combinar con la muestra ha de ser información agregada sobre la "categoría faltante" la cual puede presentarse en diferentes formatos, por ejemplo, podrían ser ceros estructurales o valores porcentuales de otras categorías, pudiendo involucrar a uno o a varios atributos.

Por ejemplo, si un atributo de la población es el estado civil (con las 5 categorías indicadas en la Tabla 6) podría haber una muestra que no contenga individuos de la categoría divorciado, en cuyo caso, divorciado sería la "categoría faltante".

Si otro atributo de la población es la edad, con tres categorías (<16, 16-65, >65), se puede disponer de información auxiliar en forma de "ceros estructurales", ya que por motivos legales no puede haber individuos divorciados con edad inferior a 16 años. Y también podría haber información porcentual sobre los divorciados, combinada con otros atributos, procedente de fuentes alternativas.

En el caso de "ceros estructurales", la celda correspondiente a (divorciado, <16) de la tabla bidimensional estado civil x edad tendrá un valor nulo, $p(\text{divorciado}| < 16) = 0$, al igual que todas las celdas de la tabla multidimensional de la muestra que correspondan a estas dos categorías $S(ecivil = \text{divorciado}, edad = <16, \dots) = 0$. Los "ceros estructurales" pueden venir definidos por los valores de múltiples categorías.

Si se dispone de información auxiliar de la categoría faltante combinada con otros atributos, por ejemplo, se conoce información porcentual de divorciados por edad dadas ciertas categorías de un tercer atributo de nacionalidad con 2 categorías: español y extranjero, con

lo que se conocería el porcentaje de divorciados de nacionalidad española con edad comprendida entre 16-65 años, $p(\text{divorciado}|\text{español}, 16-65) = 5\%$, y $p(\text{divorciado}|\text{español}, 16-65) = 1\%$.

	edad		
nacionalidad	<16	16-65	>65
español	0%	5%	1%

Tabla 17 Porcentaje de divorciados entre los españoles según las distintas categorías de edad.

Aunque lo más habitual es disponer de información de porcentaje de divorciados dado un atributo, por ejemplo, $p(\text{divorciado}|16-65) = 4\%$, $p(\text{divorciado}|>65) = 2\%$.

edad		
<16	16-65	>65
0%	4%	2%

Tabla 18 Porcentaje de divorciados para las distintas categorías de edad. Fuente: Censo nacional 2011 INE.

En este caso, el 4% de las personas con edad entre 16-65 tienen estado civil de divorciado, comparado con el restante 96% de personas de esta edad que tendrán otro estado. El 0% corresponde al "cero estructural" de no existencia de individuos divorciados menores de 16 años.

No es preciso que la información auxiliar sea exacta, basta con que sea aproximada, con lo que podría obtenerse de una fuente estadística desactualizada (censo global de una fecha distinta o de otro país) o simplemente a través de un conocimiento heurístico.

En la Figura 27 se muestra la aplicación de la expresión (24) para un caso ejemplo con una muestra 4-dimensional constituida por 200 individuos. Los individuos se describen con 4 atributos de interés, sexo, edad, nacionalidad y estado civil, y la muestra no incluye ninguna persona con estado civil divorciado, pero se dispone de la información auxiliar $p(\text{divorciado}|edad)$ de la Tabla 18, por lo que se aplica la ecuación (24) para completar las celdas de la muestra correspondientes a divorciados. De este modo se construye una nueva muestra modificada para utilizar con los métodos de generación de poblaciones, en la que se cumple:

$$p(\text{divorciado}|edad < 16) = 0$$

$$p(\text{divorciado}|edad = 16-65) = \frac{(0{,}42 + 1{,}17 + 0{,}54 + 1{,}5)}{(10{,}42 + 29{,}17 + 13{,}54 + 37{,}50)} = 0{,}04$$

$$p(\text{divorciado}|edad > 65) = \frac{(0{,}29 + 0{,}47 + 0{,}27 + 0{,}47)}{(14{,}29 + 23{,}47 + 13{,}27 + 23{,}47)} = 0{,}02$$

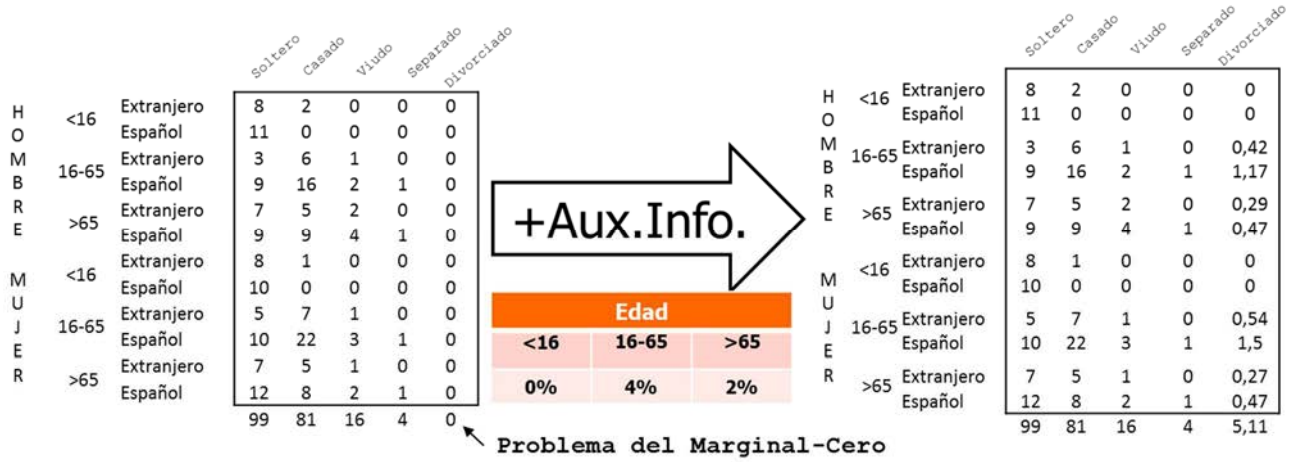

Figura 27 Ejemplo de modificación de la muestra con la información auxiliar de la categoría faltante.

Aunque se han añadido 5,11 nuevos individuos a la muestra, no se han modificado los *odd ratio* condicionales (CORs), salvo los que involucran a $\mathbf{B}_2^*$, y como se indicó en la sección 3.2.1 el algoritmo IPF no utiliza los individuos de la muestra, lo que utiliza son los CORs de la muestra para preservar la asociación entre los atributos, y trasladarla a la población sintética.

Por tanto, los nuevos valores aportados a la muestra, permiten construir una estimación de los CORs que involucran a las celdas con las categorías faltantes, sin modificar los CORs previamente existentes de la muestra.

8.4 Diseño de los Experimentos

El objetivo de los experimentos que se plantean en este capítulo es el de comparar la estrategia propuesta y las estrategias existentes, en casos de síntesis de poblaciones mediante dos métodos que utilizan muestras con el problema de marginal-cero. Los métodos de síntesis son el IPF con redondeo y el *Simulated Annealing*.

Los experimentos se llevarán a cabo dentro del marco de referencia propuesto en esta tesis, por lo que aportarán información estadística relevante para permitir comparar la eficacia de las estrategias indicadas.

Para el caso del IPF se compararán las tres estrategias:

- IPF con *tweaking* exclusivo de las celdas-cero de las "categorías faltantes" de la muestra con un pequeño valor arbitrario (IPF-TW).
- IPF con "reducción de categoría", combinando las categorías del marginal-cero con otras categorías (IPF-CR); y posterior expansión de las categorías consolidadas de forma proporcional a los marginales objetivos.
- IPF con muestra modificada con nuevos individuos con la categoría del marginal-cero, basándose en información auxiliar (IPF-AI).

Como ya hemos mencionado, a medida que el número de celdas cero aumenta en la muestra, la probabilidad de no convergencia del IPF se incrementa. Por tanto, cuando se detecte no convergencia en algún experimento, con independencia de la estrategia que se utilice, se aplicará un *tweaking* generalizado a todas las celdas-cero de la muestra, de esta forma se evita la no convergencia del algoritmo, esto implica que en el caso del IPF-TW no solo se aplica el *tweaking* a las celdas con la "categoría faltante", sino que se extiende a todas las celdas-cero de la tabla, asumiendo el riesgo que esto supone de creación de individuos imposibles.

Para el caso del SA, dado que no existe la alternativa de *tweaking*, solo se compararán:

- SA con "reducción de categorías", combinando las categorías de marginal-cero con otras categorías (SA-CR), y posterior expansión de las categorías consolidadas de forma proporcional a los marginales objetivos.
- SA con la estrategia propuesta para evitar el marginal-cero, añadiendo nuevos individuos a la muestra con la categoría del marginal-cero usando la información auxiliar (SA-AI).

Los resultados de estos experimentos también nos permitirán comparar el efecto de la nueva estrategia propuesta en ambos métodos (IPF-AI versus SA-AI).

La población de referencia que se utilizará en los experimentos de generación de poblaciones es la misma que se ha utilizado en los capítulos anteriores, obtenida de los microdatos del Censo 2001 de 60 municipios andaluces, pero en este caso se limita a individuos con 5 de los atributos de la Tabla 6. Se excluye el atributo situación laboral preferente, y se utilizan 3 categorías de edad (<16; 16-65; >65), con lo que el número de celdas de la tabla se reduce a 300 (2x3x2x5x5).

La cantidad total de individuos de la población de referencia es de 228.212, de los que casi el 1%, 2.220, tienen situación civil de divorciado, y 6.209 son extranjeros.

Los experimentos consistirán en sintetizar una población para cada uno de los 60 municipios, con los 5 atributos indicados, utilizando como datos las distribuciones marginales de los atributos para cada municipio y la correspondiente muestra de la región (la región completa es la agregación de los 60 municipios).

Estos experimentos se repetirán con 10 muestras aleatorias diferentes en las que no aparezcan divorciados, es decir, al menos exista la categoría faltante divorciado del atributo estado civil.

Para obtener las muestras con este problema marginal-cero, se extraerán muestras de pequeño tamaño, 0,1% de la población, dado que con este tamaño hay una probabilidad significativa de obtener muestras que no contengan individuos de la categoría divorciado. El proceso de obtención de las muestras consiste en extraer muestras aleatorias del 0,1% de la

población de referencia y seleccionar las primeras 10 muestras con al menos dicha categoría faltante, descartando las muestras que no tenga "categoría faltante", ya que estamos interesados en el comportamiento de los métodos bajo esta circunstancia.

Como los microdatos del censo con los que se construye la población de referencia están constituidos por hogares con sus individuos componentes, las muestras se construyen extrayendo aleatoriamente un porcentaje del 0,1% de los hogares del total de la población de referencia, y posteriormente se agregan las personas correspondientes a dichos hogares. Esto causa ligeras variaciones del número total de individuos entre las muestras.

De este modo, se consiguen 10 muestras aleatorias con el problema de cero marginal. Dos de las cuales tienen una segunda categoría faltante, como se detalla en la Tabla 19.

Para la estrategia de "reducción de categorías" se procederá a combinar dos categorías poco frecuentes del atributo estado civil, la de separado y divorciado. En España existe la situación de separación de pareja, en la que las partes permanecen casadas entre sí y, por tanto, no son libres de volverse a casar, pero están libres de las obligaciones derivadas del matrimonio. Con el divorcio, el matrimonio se finaliza oficialmente y es posible contraer matrimonio de nuevo. El divorcio fue autorizado por la legislación española en 1981, 20 años después, en 2001, había solo un 1,1% de la población divorciada y un 1,8% separada (10 años después, en 2011 estas cifras eran 3,2% y 1,5 %, respectivamente).

Muestra	Tamaño	Separados	Nº.celdas ≠ 0	Categorías Faltantes
1	221	2	38	extranjero y divorciado
2	238	7	49	divorciado
3	224	2	44	divorciado
4	227	5	52	divorciado
5	208	10	45	divorciado
6	241	2	40	divorciado
7	226	1	42	divorciado
8	231	5	40	extranjero y divorciado
9	223	5	44	divorciado
10	235	4	51	divorciado

Tabla 19 Muestras aleatorias con categorías faltantes.

En la tercera columna de la Tabla 19 se muestra el número de individuos separados de cada muestra aleatoria que pasarán a la nueva categoría consolidada de "separado y divorciado" con la estrategia de "reducción de categorías". En este caso, la tabla de la muestra tendrá 240 celdas en lugar de 300 celdas, ya que solo hay 4 categorías en el atributo de estado civil, en lugar de las 5 iniciales.

La información auxiliar (AI) utilizada para modificar la muestra es una información aproximada, en este caso, de un año censal diferente y de una región diferente, como es el censo nacional de 2011 para todo el país. La información auxiliar que se utiliza en los experimentos es la indicada en la Tabla 18, la cual es diferente de la que puede extraerse de la población de referencia (0%; 1,3%; 0,5%).

Para la segunda categoría faltante, extranjero, de la 1ª y 8ª muestra, la información auxiliar que se utiliza es la de la Tabla 20, también obtenida del censo nacional de 2011.

edad		
<16	16-65	>65
12%	13%	4%

Tabla 20 Porcentaje de extranjeros para las distintas categorías de edad. Fuente: Censo nacional INE 2011.

El IPF que se aplica en los experimentos es el correspondiente al algoritmo clásico de un paso, ajustando la tabla 5-dimensional de la muestra a las 5 restricciones marginales unidimensionales correspondientes, y posteriormente se aplica el redondeo BLP para la convertir los valores de la tabla preservando los marginales.

El valor que se utiliza para el *tweaking* es 0,001. Podría pensarse que este valor de *tweaking* seleccionado puede afectar a la calidad de la población sintetizada. Para comprobarlo, se ha aplicado el *tweaking* con un rango de valores comprendido entre 10^2 y 10^{-5} (100; 10; 1; 0,5; 0,1; 0,01; 0,001; 0,0001; 0,00001) y se han comparado las poblaciones obtenidas verificándose que la significatividad de los resultados no cambia en ningún caso.

Este valor también se usa para el *tweaking* masivo de todas las celdas cero ante una situación de no convergencia del algoritmo IPF. Los experimentos de generación de poblaciones se realizan con tablas de 300 celdas que tienen entre 80% y 85% de celdas-cero. Con este porcentaje de celdas-cero, un 5% de las ejecuciones del IPF no convergen; este porcentaje de no convergencia se incrementaría a media que las tablas tengan más celdas, o mayor porcentaje de celdas-cero. Se han hecho ensayos aumentando el número de atributos y categorías, es decir, utilizando tablas multidimensionales con mayor número de celdas y las conclusiones son las mismas que las que se obtienen con las tablas de 300 celdas, a pesar del mayor porcentaje de celdas cero (*sparsity*), lo que provoca que aumente el número de casos de no convergencia del IPF, lo que obliga a recurrir con mayor frecuencia al *tweaking* completo de todas las celdas cero de la tabla para evitar la no convergencia.

Para la técnica del *Simulated Annealing* se utiliza la implementación de Williamson (2007). Este algoritmo no precisa de conversión a enteros, ya que el resultado es entero.

Por último, la métrica que se utiliza para medir la diferencia entre las poblaciones sintetizadas y la de referencia es el error de clasificación (% CE) propuesto en 5.4, cuya expresión (17) en este caso será 5-dimensional. También se medirá el mismo error de clasificación, pero referido a las los individuos con estado civil divorciado, como es el $\%CE^{5\text{-dim}}$(divorciado) de la expresión (18).

Para evaluar estas métricas en las poblaciones de 240 celdas, sintetizadas con la estrategia de reducción de categorías, y poder determinar el %CE referido a la población de referencia de 300 celdas, es necesario repartir de forma proporcional los valores de las celdas correspondientes a la categoría consolidada entre las categorías de separados y divorciados. El reparto de los valores de las celdas se realizará en proporción a los valores de los marginales de estos individuos en cada municipio, lo cual produce valores fraccionarios, por lo que en estos casos se utiliza el mismo método de redondeo BLP para convertirlos a enteros manteniendo así los marginales de todos los atributos de la tabla, sin producir ningún desajuste de los mismos.

8.5 Resultados

En esta sección se presentan los resultados de los experimentos realizados con cada uno de los métodos y estrategias.

Con cada método-estrategia se han generado 600 poblaciones (60 municipios x 10 muestras diferentes), se han determinado los errores de clasificación de cada una ($\%CE^{5\text{-dim}}$) y se ha calculado el valor medio de los 600 valores de la métrica. También se han calculado los valores medios de los porcentajes de individuos divorciados y separados de cada municipio, que están incorrectamente clasificados en las poblaciones sintetizadas, $\%CE^{5\text{-dim}}$(divorciados) y $\%CE^{5\text{-dim}}$(separados).

En la Tabla 21 se muestran estos valores medios para cada una de las combinaciones de método- estrategia frente al problema cero marginal. En la Figura 28, se representan gráficamente los valores de las dos primeras filas de dicha tabla.

Dado que en dos de las 10 muestras hay una segunda categoría faltante, la de extranjero, en la última fila de la Tabla 21 se incluye el valor medio del porcentaje de individuos extranjeros mal clasificados en las poblaciones generadas con estas dos muestras (120 poblaciones).

Estos resultados muestran que la estrategia de modificar la muestra a partir de información auxiliar conduce a poblaciones más precisas, tanto con IPF como con SA. Recordemos que la información auxiliar referente a individuos divorciados solo afecta a 2.200 individuos (1% de la población), de los que un 85% quedan incorrectamente clasificados con el IPF-TW, por lo

que en el mejor de los casos, si con IPF-AI se hubiera logrado clasificar a todos los divorciados correctamente se tendría una disminución del $\%CE^{5\text{-dim}}$ del 0,8% respecto al IPF-TW. Aunque la mejora en el error global $\%CE^{5\text{-dim}}$ producido por IPF-AI es pequeña, dicha mejora es apreciable en relación al error de clasificación de los individuos divorciados. Con IPF-AI se consigue clasificar correctamente a casi el 60% (1.308 individuos divorciados). Igualmente, en el caso de la categoría de extranjero, hay una reducción del error de más del 35% en relación con la estrategia de *tweaking*.

	IPF-TW	IPF-CR	IPF-AI	SA	SA-CR	SA-AI
$\overline{\%CE}$	18,75%	18,89%	18,46%	19,61%	19,71%	19,38%
$\overline{\%CE^{divorced}}$	85,03%	66,44%	40,55%	100%*	67,41%	49,13%
$\overline{\%CE^{separated}}$	62,81%	62,94%	63,07%	65,52%	64,11%	65,15%
$\overline{\%CE^{non-spaniard}}$	90,26%	87,52%	52,63%	100%*	100%*	54,60%

Tabla 21 Error de Clasificación para distintos métodos y estrategias para abordar el marginal-cero.

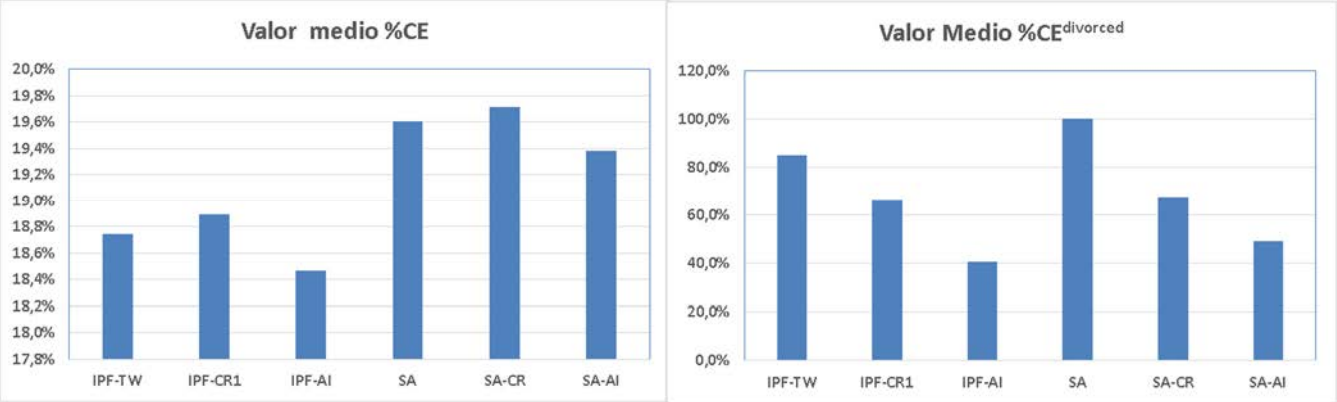

Figura 28 Promedio de Error de Clasificación para diferentes métodos y estrategias para abordar el marginal-cero.

El incremento de precisión en la determinación de divorciados o separados puede ser determinante según el fin que se persiga con dicha población. Por ejemplo, si se tratara de generar una población sintética con individuos divorciados y separados con el objetivo de estimar el número de nuevos matrimonios o de adquisiciones de vivienda en propiedad en caso de no tenerla, ya que dicho colectivo tiene mayor probabilidad de contraer matrimonio o de adquirir una nueva vivienda, frente al resto de individuos de su misma edad. Por lo tanto, es de vital importancia que queden correctamente clasificados tales individuos en la población sintética para que las estimaciones de matrimonio o de compra de vivienda sean más realistas.

En los resultados expuestos, también se observa que la consolidación de categorías (CR) es la estrategia con la que se obtienen las poblaciones con mayor error en ambos métodos. Así mismo, el IPF con redondeo produce poblaciones con menor error de clasificación que el

Simulated Annealing, con independencia de la estrategia utilizada. Resultado que coincide con las conclusiones de la comparativa del capítulo 6 con muestras de mayor tamaño.

El "*" que aparece en los valores de la tabla para el SA, indica que el valor utilizado corresponde al índice Bray-Curtis (explicado en la sección 4.2), ya que en estos casos, al no haber divorciados en la muestra, SA no genera individuos divorciados, y por lo tanto, las poblaciones que se comparan tienen distinto tamaño. La métrica %CE no es válida en estos casos, por lo que se utiliza el BCI. El valor de este índice es 100%, ya que todas las personas divorciadas están mal clasificadas al no haber ningún divorciado en las poblaciones generadas con SA. Tampoco hay extranjeros en las poblaciones generadas con SA y las muestras sin extranjeros.

8.5.1 Análisis estadístico de las diferencias de error

A continuación se procede a comparar mediante pruebas de contraste de hipótesis las medias de los errores de las poblaciones generadas con cada método-estrategia.

A partir de las 600 observaciones de error, de las poblaciones sintetizadas con cada método-estrategia, se determinan los valores promedios para cada municipio. Por tanto, se obtienen 60 valores medios para cada método-estrategia. Con estos 60 valores se plantean las pruebas de comparación de medias de las distintas combinaciones método-estrategia.

Dado que las diferencias entre los errores de dos distintos métodos-estrategias no responden a una distribución normal, según pone de manifiesto la prueba de normalidad de Shapiro-Wilk, no es posible realizar una prueba de hipótesis paramétrica como el t-test. Por tanto se realiza la prueba de hipótesis no paramétrica de rangos con signo de Wilcoxon (ver descripción de este test en sección 5.6.2).

La Tabla 22 muestra los resultados de dicha prueba. Estos resultados confirman las conclusiones sugeridas en la comparación directa de valores de $\overline{\%CE_i}$. Además, coinciden con los de la prueba del signo (no incluidas en la tabla), ya que la mediana y la media son muy similares, por lo que es asumible la simetría de las distribuciones de las diferencias.

Las pruebas de hipótesis que aparecen en las tres primeras columnas de la tabla, correspondientes a las diferencias de errores de las poblaciones generadas con IPF-AI, IPF-TW, IPF-CR y SA-AI muestran, para los tres casos, que la mediana de las diferencias entre estos errores no es cero y es significativamente positiva. Por tanto, puede afirmarse que IPF-AI produce poblaciones con menos error de clasificación *(%CE*) que IPF-TW, IPF-CR y SA-AI.

La prueba de la cuarta columna de la tabla no es tan clara. La columna corresponde a las diferencias $\overline{\%CE_i}_{(\text{SA})}$ -$\overline{\%CE_i}_{(\text{SA-AI}),}$ y aunque la mediana de la variable diferencia no es cero, el

intervalo de confianza está entre -0,05% y 0,39% (para un nivel de confianza del 99% y significación α = 0,01). Por lo que, dado que ambos extremos del intervalo son de diferente signo, no puede concluirse, con este nivel de significación, que SA con la estrategia AI produzca poblaciones con un error de clasificación estadísticamente significativo menor que sin ella. Tampoco puede concluirse nada en relación a las diferencias $\overline{\%CE_i}$(SA) -$\overline{\%CE_i}$(SA-CR), no incluidas en la tabla para simplificar. Sin embargo en la comparación entre SA-CR y SA-AI de la última columna de la tabla, sí puede afirmarse que SA-AI produce poblaciones con valores de error ligeramente inferiores, ya que la estrategia CR empeora la precisión de la población sintetizada, comparándola con usar SA con la muestra con marginal-cero.

	$\overline{\%CE_i}$(**IPF-TW**) - $\overline{\%CE_i}$(**IPF-AI**)	$\overline{\%CE_i}$(**IPF-CR**) - $\overline{\%CE_i}$(**IPF-AI**)	$\overline{\%CE_i}$(**SA-AI**) - $\overline{\%CE_i}$(**IPF-AI**)	$\overline{\%CE_i}$(**SA**) - $\overline{\%CE_i}$(**SA-AI**)	$\overline{\%CE_i}$(**SA-CR**) - $\overline{\%CE_i}$(**SA-AI**)
(pseudo)mediana	0,27%	0,30%	0,88%	0,15%	0,27%
Media	0,28%	0,43%	0,92%	0,23%	0,33%
Intervalo de confianza de la mediana (α=0,01)	(0,18%; 0,34)	(0,21%; 0,43)	(0,68%; 1,08%)	(-0,05%; 0,39%)	(0,05%; 0,47%)
S_+	1.825	1.711	1.830	1.182	1.350
Valor-p	<0,0001	<0,0001	<0,0001	0,0493	0,0013

Tabla 22 Resultados de la prueba de los rangos con signo de Wilcoxon de las diferencias entre Errores de Clasificación Porcentual.

Tampoco se ha incluido en la tabla la comparativa de la diferencia entre el %CE de IPF-CR y de IPF-TW, ya que tampoco se ha encontrado una diferencia estadísticamente significativa entre estos errores.

Si se repiten las pruebas de hipótesis, pero esta vez con los datos de los errores de clasificación de los individuos divorciados exclusivamente, los resultados son de otra proporción. Las primeras dos columnas de la Tabla 23 contienen las diferencias entre los errores de clasificación de las poblaciones obtenidas con IPF con las tres estrategias. Este error de clasificación se refiere exclusivamente a los individuos divorciados, esto es, corresponde al error de clasificación 5-dimensional de la expresión (19). El error obtenido con *tweaking* es el doble del obtenido con "información auxiliar", siendo la diferencia del 45,15%. La última columna de la tabla muestra que la diferencia entre el error de clasificación de divorciados de la población obtenida con la estrategia de información auxiliar aplicando IPF es un 9% inferior al de las poblaciones obtenidas con SA.

Los intervalos de confianza de las diferencias de $\overline{\%CE^{divorced}}$ de la Tabla 23 muestran que las diferencias de error entre las poblaciones obtenidas con IPF-AI y las obtenidas con IPF-TW, IPF-CR y SA-AI son en todos los casos estadísticamente significativas.

En resumen, los resultados confirman que el uso de información auxiliar, para compensar la incertidumbre que genera una muestra con el problema de marginal-cero en ciertas categorías, permite obtener poblaciones sintéticas con individuos de las categorías faltantes mejor clasificados que si se utilizara otras estrategias tradicionales.

	$\overline{\%CE_i^{\text{divorciado}}}$(IPF-TW) - $\overline{\%CE_i^{\text{divorciado}}}$(IPF-AI)	$\overline{\%CE_i^{\text{divorciado}}}$(IPF-CR) - $\overline{\%CE_i^{\text{divorciado}}}$(IPF-AI)	$\overline{\%CE_i^{\text{divorciado}}}(SA\text{-}AI)$ - $\overline{\%CE_i^{\text{divorciado}}}$(IPF-AI)
Pseudo(median)	45,15%	27,54%	9,00%
Mean	43,74%	25,46%	8,44%
Confidence Interval (α=0,01)	(40,76% ; 49,34%)	(23,36% ; 30,87%)	(6,66% ; 11,25%)
S_+	1.770	1.752	1.734
p-value	<0,0001	<0,0001	<0,0001

Tabla 23 Resultados de la prueba de rangos con signo de Wilcoxon con las diferencias del Error de Clasificación de individuos divorciados.

Después de completar todos los experimentos, se han vuelto a repetir parte de los mismos utilizando información auxiliar diferente, información obtenida del censo 2001, en lugar de 2011, con lo que la nueva información auxiliar se ajusta mejor a la realidad de la población de referencia. Con los nuevos experimentos se ha comprobado que los errores de las poblaciones generadas con IPF-AI son menores al de los experimentos descritos. Dada la pequeña disminución del error, no es posible verificar estadísticamente si la disminución del error está directamente relacionada con la calidad de la información auxiliar.

Pero es lógico que al utilizar una información auxiliar más exacta se obtengan mejores estimaciones de los *odd* ratio condicionales de la categoría faltante de la muestra, lo cual es aprovechado por el IPF para generar poblaciones más exactas.

En resumen, la utilización de información auxiliar para completar las muestras con el problema de marginal-cero mejora la precisión de las poblaciones que se obtienen con el método IPF, en comparación con las poblaciones que se obtienen aplicando las estrategias tradicionales para abordar dicho problema. Esta mejora proviene de una importante reducción del error de clasificación de los individuos con la categoría faltante.

8.6 Conclusiones

En este capítulo se ha analizado el problema de las muestras con marginal-cero, el cual afecta especialmente a los métodos de generación de poblaciones que utilizan una muestra como dato de entrada. Se ha examinado el rendimiento de dos de estos métodos con muestras con

marginal-cero: uno basado en IPF y otro en *Simulated Annealing*, utilizando con ambos, distintas estrategias para abordar dicho problema.

Se han planteado las diferentes estrategias para tratar el problema del marginal-cero en la muestra, aproximaciones tradicionales para remediar el problema, y un nuevo enfoque basado en "fusión de datos", combinación de la muestra con información auxiliar agregada sobre la categoría faltante.

Se ha realizado una evaluación del rendimiento de los métodos y estrategias planteadas conforme al marco de referencia descrito en esta tesis, estableciendo una población de referencia, y determinando la semejanza entre las poblaciones generadas y la población de referencia.

La evaluación se ha basado en la comparación 5-dimensional de las poblaciones sintetizadas con la población de referencia, la cual se supone que representa. La divergencia entre las dos poblaciones se obtiene comparando cada pareja de celdas de la tabla 5-dimensional de ambas poblaciones. Se ha realizado una prueba de hipótesis para verificar la significación de los resultados. Se ha comprobado la importancia de utilizar la información auxiliar sobre la categoría faltante para afrontar con éxito el problema.

Tras realizar la comparación estadística entre los resultados obtenidos con las distintas combinaciones de método-estrategia, la principal conclusión es que el rendimiento del IPF mejora significativamente si se modifica la muestra mediante el uso de información auxiliar sobre la categoría faltante. Sin embargo para el caso del SA esta mejora no ha sido estadísticamente significativa en este estudio, aunque ha mejorado la generación de los individuos con la categoría faltante.

Así mismo, tras comparar el impacto de la estrategia de "fusión de datos" con el de otras estrategias convencionales para abordar el problema marginal-cero de la muestra, se ha comprobado que la "fusión de datos" supone un mayor incremento en la mejora el rendimiento del IPF y reduce el problema de falta de convergencia.

9 Conclusiones y futuras líneas de investigación

Aunque se han ido señalando las aportaciones y conclusiones propias de cada uno de los tres estudios planteados, se exponen aquí las conclusiones generales y el cumplimiento con los objetivos propuestos. También se incluye una sección dedicada a futuras líneas de trabajo.

En esta tesis se han analizado distintos métodos de generación de poblaciones sintéticas, principalmente los métodos de *reweighting* sobre los que se ha centrado la investigación. Así mismo, se han analizado diversas problemáticas que pueden plantearse ante la necesidad de disponer de una población sintética. Este análisis ha incluido la adaptación de la técnica de redondeo BLP al caso de poblaciones multinivel con objeto de mejorar la eficiencia de los métodos de *reweighting* con pesos no enteros. Tras dicho análisis, se ha propuesto un marco de referencia metodológico en el que se ha definido un esquema de validación y comparación cuantitativa de los distintos métodos de generación de poblaciones sintéticas.

El marco de referencia plantea un banco de pruebas para los métodos de generación de poblaciones que posibilita estudiar el rendimiento comparado de métodos con múltiples posibilidades de elección de los datos de entrada, muestras aleatorias, distribuciones de marginales, distribuciones de probabilidades condicionadas, marginales con múltiple dimensionalidad, atributos de distintos tipos, etc.

Esta propuesta es aplicable a múltiples problemas y métodos de síntesis, por lo que puede ser de ayuda a investigadores que deseen explorar los entornos y circunstancias en los que unos métodos producen poblaciones más precisas que otros.

También puede ser útil para los investigadores que desarrollan nuevos enfoques de síntesis de poblaciones, con el fin de disponer de un esquema de comparación sistemático entre los nuevos enfoques y los actuales.

Finalmente, se ha utilizado este marco de referencia en tres estudios de investigación cuantitativa sobre métodos relevantes en el campo de la generación de poblaciones sintéticas, analizando los métodos más utilizados por los investigadores, y comparando las mejoras de precisión de las poblaciones sintetizadas al aplicar técnicas de redondeo.

9.1 Conclusiones

A continuación se revisa el cumplimiento de los objetivos que se plantearon al comienzo de esta tesis, junto con las aportaciones que se han realizado en línea con cada uno de los mismos.

Se presentan las aportaciones bajo cada uno de los objetivos planteados.

1. Hacer una revisión y clasificación de los principales métodos de generación de poblaciones sintéticas que aparecen en la literatura, así como de las principales aplicaciones de los datos sintéticos.
    - Se han analizado las principales aplicaciones de los datos de poblaciones sintéticas, así como la importancia de los mismos, verificando que la generación de poblaciones para "*small area*" es un paso esencial para el modelado espacial y la simulación.
    - Se han clasificado y revisado los múltiples enfoques y métodos que se utilizan para generar poblaciones sintéticas, de un modo que ayuda a otros investigadores a encontrar el mejor enfoque para generar sus microdatos sintéticos.
    - Se han contrastado distintas formas de aplicación de la técnica del IPF, como la del IPF tradicional y el IPF de Beckman.
    - Se ha comprobado de forma práctica la equivalencia entre métodos, como *Generalized Regression Weighting* (GREGWT) y *Generalized Raking* lineal, y entre el método de Optimización de la Entropía y *Generalized Raking* multiplicativo.
2. Analizar las distintas problemáticas que se plantean en el proceso de selección de métodos de construcción de poblaciones sintéticas y justificar la propuesta de marco metodológico que ayude en dicho proceso.
    - Se ha analizado el requerimiento de la población multinivel, los distintos tipos de problema de generación de poblaciones ("*small area*" y población completa), la problemática de la variedad de métricas y disponibilidad de datos de entrada.
    - Se han analizado los distintos estudios publicados sobre comparación de métodos, examinando las variables de los esquemas de comparación de dichos estudios.
    - Se ha confirmado la situación que justifica la propuesta de un marco metodológico.
    - Se ha adaptado el algoritmo de redondeo BLP para distribuciones multidimensionales al caso de poblaciones multinivel.
3. Definir un marco de referencia para analizar comparativamente los métodos, tanto con respecto a recomendaciones de utilización de métodos ante las distintas necesidades y disponibilidades de datos, como en el ámbito de la metodología de análisis comparativo.

- Se ha propuesto un marco de referencia para el análisis comparativo de métodos, que se ha estructurado en 3 componentes: definición de escenarios de necesidades, mapa de escenarios-métodos y metodología para llevar a cabo el análisis comparativo de métodos.
- Se ha definido una metodología con análisis estadístico para comparar distintos enfoques de generación de población, articulada a partir de 5 pilares: población de referencia, métrica de comparación, esquema de validación interna y externa, prueba de hipótesis y análisis de sensibilidad.

4. Realizar un mínimo de 3 estudios cuantitativos experimentales conforme al marco propuesto, con los que se obtengan conclusiones sobre el rendimiento relativo de los métodos y se posibiliten análisis de verificación del impacto del uso de nuevas técnicas y estrategias, análisis del comportamiento de distintos métodos de generación y determinación del rendimiento relativo de métodos.
    - Se han elaborado 3 estudios comparativos de métodos con análisis de significación estadística, siguiendo indicaciones de varios autores que han enfatizado la necesidad de una mayor investigación en esta área. En el primero de los estudios:
        - Se ha examinado el rendimiento de métodos de generación de poblaciones basados en dos enfoques del tipo *reweighting* de uso muy frecuentemente (IPF y optimización combinatoria).
        - Se ha utilizado el marco de referencia para verificar el impacto del uso del redondeo en los métodos. Se ha concluido que el proceso de redondeo de la tabla obtenida con IPF, manteniendo el ajuste de marginales, es un factor clave, que puede hacer más eficiente ciertos métodos, hasta tal punto que métodos como IPF con redondeo BLP presentan un rendimiento superior al SA, lo cual no sucede con otros métodos de redondeo comúnmente utilizados.
        - Contrariamente a lo que varios estudios han publicado, en este estudio, IPF (con el redondeo apropiado) supera al *Simulated Annealing* (SA).
        - Se ha analizado el comportamiento del IPF y SA ante variaciones de parámetros como son el tamaño de muestra, tamaño de la "*small area*" y número total de categorías.
    - En el segundo de los estudios:
        - Se ha comparado en distintos métodos de *reweighting* de generación de poblaciones multinivel, el muestreo de Monte Carlo y la adaptación propuesta del procedimiento de redondeo BLP de distribuciones multidimensionales multinivel.

  - Se ha verificado que el IPU con redondeo BLP es superior al IPU con muestreo de Monte Carlo. Igualmente, el GR multiplicativo con redondeo BLP es más preciso que con muestreo de Monte Carlo cuando las muestras son de pequeño tamaño (<5%).
  - Se ha verificado, de forma estadística, el superior rendimiento del método GR multiplicativo frente a IPU, y la equivalencia de las poblaciones obtenidas con métodos como GR multiplicativo, HIPF y OE, sin poder afirmar que ninguno de estos 3 métodos tiene un rendimiento superior al resto, en contra de lo que argumentan otros investigadores.
  - Se ha estudiado la influencia del tamaño de la muestra y del tamaño de la "*small area*" en el rendimiento de los métodos de *reweighting* multinivel. El error de la población sintetizada disminuye a medida que se sintetizan poblaciones de mayor tamaño y se usan muestras de mayor tamaño.
- En el tercer estudio se ha aplicado el marco de referencia para estudiar el problema de muestras con marginal-cero.
  - Se ha examinado el rendimiento de dos métodos de generación de poblaciones utilizando muestras con marginal-cero, aplicando las estrategias convencionales referenciadas en la literatura para abordar este problema.
  - Se ha propuesto una estrategia novedosa para abordar dicho problema.
  - Se ha estudiado el efecto de la estrategia propuesta en dos métodos (IPF-BLP y SA), usando muestras con marginal-cero, y se ha comparado con el efecto de las estrategias convencionales aplicadas a dichos métodos.
  - Se ha verificado la mejora que supone la utilización de la estrategia propuesta, basada en utilizar información auxiliar, con el método IPF frente a otros métodos y estrategias. Se ha comprobado que se trata de una mejora estadísticamente significativa.
  - En este estudio no se ha observado una mejora estadísticamente significativa de la precisión de las poblaciones generadas con SA, utilizando muestras modificadas mediante información auxiliar sobre la categoría faltante, aunque si se ha observado mejora en la generación de los individuos con la categoría faltante.

9.2 Futuras líneas de trabajo

El desarrollo del marco de referencia propuesto abre nuevas posibilidades para replantear los estudios comparativos entre métodos, así como plantear nuevas comparativas entre métodos dentro de este nuevo marco. Las nuevas comparativas permitirán construir una guía que confronte los múltiples métodos disponibles, y en la que se tengan en cuenta los distintos factores que afectan al rendimiento de los métodos.

Entre los caminos no explorados se proponen tres líneas de trabajo de cara al futuro: mejora de la eficiencia del redondeo, la implementación de una plataforma software para el marco de referencia y desarrollo de nuevos métodos híbridos de generación de poblaciones.

En relación a la primera línea, se propone investigar nuevas formas de mejora de la eficiencia de los procesos de redondeo, dado que el redondeo BLP no siempre ofrece una solución factible. El empleo de la técnica de programación por metas aplicada a pesos de poblaciones multinivel podría mejorar los resultados obtenidos.

Los métodos de *reweighting* de la muestra asignan pesos a los elementos de la muestra que optimizan distintas funciones objetivos, sujetas a un conjunto de restricciones como el ajuste de los marginales de la población. El redondeo BLP posibilita encontrar una solución redondeada de estos pesos, modificando ligeramente el valor óptimo encontrado, pero satisfaciendo las restricciones de los marginales.

Tal como se ha descrito, esto no es posible en todas las situaciones, por lo que con técnicas de programación por metas se podría establecer una jerarquía de prioridades de ajuste que en lugar de obligar a ajustar todos los marginales, encuentre el mejor compromiso de ajuste entre ellos. La programación por metas permitiría que en los casos en los que no se encuentran soluciones satisfactorias que verifiquen todas las restricciones marginales, se transformen algunas restricciones en objetivos, estableciendo un orden de prioridad entre los mismos, que subordine unos objetivos a otros. Por ejemplo, tendría baja prioridad ajustar las categorías poco frecuentes, frente a otras categorías importantes, y se incorporarían nuevas funciones objetivo como es minimizar los errores absolutos de dichas categorías poco frecuentes.

Otra posible alternativa de investigación para mejorar el redondeo se basa en la de exploración de métodos heurísticos. Este tipo de métodos representa una buena alternativa para encontrar soluciones eficientes en los casos de falta de solución factible del redondeo BLP.

En relación a la segunda línea de trabajo propuesta, se plantea la utilización del software desarrollado para esta tesis (principalmente en Java y R) para implementar una plataforma

software flexible que facilite las tareas de generación y comparación de datos y poblaciones sintéticas. La plataforma permitiría establecer fácilmente una población de referencia con atributos de distintos tipos, y de forma automatizada obtener diferentes datos, como marginales de distinta dimensionalidad para cada una de las "*small area*s", muestras con distintos atributos y tamaños, y/o distribuciones de probabilidad condicionadas.

La plataforma permitiría seleccionar distintos métodos de generación de poblaciones y sintetizar las poblaciones ajustadas a las restricciones impuestas, o importar las poblaciones obtenidas con otros métodos. Incluiría el establecimiento de un interfaz de programación de aplicaciones (*Application Programming Interface* API) para conectar con implementaciones de métodos de generación.

Mediante un proceso de selección de métricas y de la dimensionalidad a utilizar, se definirían los esquemas de los procesos de validación interna e externa. La plataforma calcularía dichas métricas para cada una de las poblaciones generadas. Con los valores de las métricas, se realizaría la comparativa de validaciones de forma automática conforme al esquema seleccionado.

Con un software dotado de esta flexibilidad se facilitaría la realización de comparaciones y validación de los métodos. Se podría construir un mapa coherente, con fundamento, basado en datos de rendimiento de los múltiples métodos disponibles. Así mismo, los desarrollares de nuevos métodos de generación de poblaciones podrían utilizarla a modo de plataforma de benchmarking.

Sería de interés para poder disponer fácilmente de nuevas comparativas, tales como entre los nuevos métodos tipo probabilístico aparecidos recientemente y métodos clásicos como el IPF, y determinar con que tamaño de muestra se obtienen rendimientos equivalentes.

Una alternativa de construcción de esta plataforma software sería plantearla como una extensión de otras plataformas software de uso extendido. Esto contribuiría a la popularización del uso de poblaciones y datos sintéticos. Recientemente se han popularizado plataformas software de código abierto para minería de datos y aprendizaje automático (análisis predictivo), tales como WEKA, RapidMiner, KNIME, etc.,(basadas en Java) las cuales disponen de un interfaz gráfico que permiten realizar de forma muy amigable tareas de pre-procesamiento, imputación, *clustering*, clasificación y selección de datos entre otras.

Este tipo de aplicaciones están construidas con arquitecturas modulares que permiten incorporar fácilmente paquetes o *plugings* con nueva funcionalidad. Sería de utilidad el desarrollo de nuevos paquetes, integrados en estas plataformas, que incorporaran distintos métodos de generación de poblaciones y el propio marco de referencia planteado, lo cual popularizaría su uso.

Dado que en los últimos años están apareciendo múltiples librerías de R que implementan los métodos de generación de poblaciones, la integración con estas librerías se puede plantear mediante el uso de los *plugins* de R que existen en estas plataformas, como Rplugin de WEKA, o R Scripting extensión de RapidMiner por ejemplo. De este modo se facilitaría y simplificaría la tarea de creación y utilización de poblaciones sintéticas. Esto es un factor esencial para conseguir popularizar el uso de los datos y poblaciones sintéticas, con el que seguramente aparecerían nuevas aplicaciones.

La última línea de investigación planteada es la de desarrollar nuevos métodos híbridos de generación de poblaciones sintéticas. Estos métodos se construirían combinando técnicas probabilísticas, utilizadas en los problemas de estimación de "*small area*" (SAE), con métodos de *reweighting*. Es decir, métodos que se encuadrarían en el tercer grupo de la clasificación de métodos aquí presentada.

Abreviaturas

ABS	*Australian Bureau of Statistics*
AE	*Absolute Error*
ABDG	*Activity-Based Demand Generation*
ABM	*Activity-based modelling*
ABM	*Agent Based Modelling*
AI	*Auxiliary Information*
BCI	*Bray-Curtis index*
BLP	*Binary Linear Programming*
BN	*Bayesian Networks*
CART	*Classification and Regression Tree*
CE	*Classification Error*
CO	*Combinatorial Optimization*
COR	*Conditional Odd Ratio*
CP	*Conditional Probability*
CR	*Category Reduction*
DR	*Deterministic Reweighting*
EO	*Entropy Optimization*
EPA	Encuesta de Población Activa
EPF	Encuesta de Presupuestos Familiares
FBS	*Fitness-Based Synthesis*
FRS	*Family Resources Survey*
GA	*Genetic Algorithm*
GDA	Gráfico Dirigido Acíclico
GLM	*Generalized Linear Model*
GoF	*Goodness of Fit*
GR	*Generalized Raking*
GREGWT	*Generalized Regression Weighting*
HBAI	*Households Below Average Incomes*
HIPF	*Hierarchical IPF*
HMM	*Hidden Markov Model*
IPF	*Iterative Proportional Fitting*
IPFSR	*IPF Synthetic Reconstruction*
IPU	*Iterative Proportional Updating*
INE	Instituto Nacional de Estadística
JDI	*Joint Distribution Inference*
KCS	*Kernel Construction Set*
KNN	*K-Nearest neighbors*
LSQ	*Least Squares*
MAE	*Mean Absolute Error*

MAPE	*Mean Absolute Percentage Error*
MCHI2	*Minimum Chi-Square*
MCMC	*Markov Chain Monte Carlo*
MCS	*Monte Carlo Sampling*
MICE	*Multivariate imputation by Chained Equations*
ML	*Maximum Likelihood*
MSE	*Mean Square Error*
MSM	*Microsimulation Modelling*
NATSEM	National Centre for Social and Economic Modelling
NFC	*Non Fitting Cells*
NFT	*Non Fitting Table*
PGP	*Proportion of Good Prediction*
PHS	*Post Harvest Survey*
RMSE	*Root Mean Square Error*
SA	*Simulated Annealing*
SAE	*Small area Estimation*
SAE	*Standarized Absolute Error*
SBA	Simulación Basada en Agentes
SFF	*Sample Free Fitting*
SMILE	*Simulation Model for the Irish Local Economy*
SMSM	*Spatial Microsimulation Model*
SR	*Sinthetic Reconstruction*
SRMSE	*Standardized Root Mean Square Error*
SSZ	*Sum of Square Z scores*
TAE	*Total Absolute Error*
TAZ	*Traffic Analysis Zone*

Bibliografía

ABM Software Comparison. (2018). Retrieved from http://en.wikipedia.org/wiki/ABM_Software_Comparison

Abraham, J. E., Stephan, K. J., & Hunt, J. D. (2012). Population Synthesis Using Combinatorial Optimization at Multiple Levels. In *91st Annual Meeting of the Transportation Research Board*. Washington, DC.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. *Statistics*. http://doi.org/10.1002/0471249688

Andridge, R. R., & Little, R. J. A. (2010). A Review of Hot Deck Imputation for Survey Non-response. *Int Stat Rev.*, (78(1)), 40–64. http://doi.org/10.1111/j.1751-5823.2010.00103.x

Arentze, T., & Timmermans, H. (2004). Albatross: A Learning Based Transportation Oriented Simulation System. *Transportation Research Part B Methodological*, *38(7)*, 613–633. http://doi.org/10.1016/j.trb.2002.10.001

Arentze, T., Timmermans, H., & Hofman, F. (2007). Creating synthetic household populations. *Trannsportation Research Record: Journal of Transportation Research Board*, (2014), 85–91. http://doi.org/10.3141/2014-11

Auld, J., & Mohammadian, A. (2010). Efficient Methodology for Generaty Sinthetic Populations with Multiple Control Levels. *Trannsportation Research Record: Journal of Transportation Research Board*, (2175), 138–147. http://doi.org/10.3141/2175-16

Auld, J., Mohammadian, A., & Wies, K. (2009). Population Synthesis with Subregion-Level Control Variable Aggregation. *Journal of Transportation Engineering*, *135*(9), 632–639. http://doi.org/10.1061/(ASCE)TE.1943-5436.0000040

Ballas, D., Clarke, G., Dorling, D., Eyre, H., Thomas, B., & Rossiter, D. (2005). SimBritain: a spatial microsimulation approach to population dynamics. *Population, Space and Place*, *11*(1), 13–34. http://doi.org/10.1002/psp.351

Ballas, D., Clarke, G. P., & Wiemers, E. (2006). Spatial microsim- ulation model for rural policy analysis in Ireland: the impli- cations of the CAP reforms for the national spatial strategy. *Journal of Rural Studies*, *22*, 367–78.

Ballas, D., Rossiter, D., Thomas, B., Clarke, G., & Dorling, D. (2005). *Geography Matters: Simulating the local impacts of national social policies*. *Joseph Rowntree Foundation contemporary research issues* (Vol. 64). Joseph Rowntree Foundation, York. http://doi.org/10.2307/3650139

Bar-Gera, H., Konduri, K. C., Sana, B., Ye, X., & Pendyala, R. M. (2009). Estimating Survey Weights with Multiple Constraints Using Entropy Optimization Methods. In *Papers Presented at the 88th AnnualMeeting of Transportation Research Board*. Washington,

D.C.

Barse, E. L., Kvarnstrom, H., & Johnson, E. (2003). Synthesizing test data for fraud detection systems. *19th Annual Computer Security Applications Conference, 2003. Proceedings.*, (Acsac), 384–394. http://doi.org/10.1109/CSAC.2003.1254343

Barthelemy, J., Suesse, T., & Namazi-Rad, M. (2018). Package "mipfp." Retrieved from https://cran.r-project.org/package=mipfp

Barthelemy, J., & Toint, P. L. (2013). Synthetic Population Generation Without a Sample. *Transportation Science*, *47*(2), 226–279. http://doi.org/10.1287/trsc.1120.0408

Barthelemy, J., & Toint, P. L. (2015). A Stochastic and Flexible Activity Based Model for Large Population . Application to Belgium. *Journal of Artificial Societies and Social Simulation*, *18*, 1–20.

Beckman, R. J., Baggerly, K. a., & McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, *30*(6), 415–429. http://doi.org/10.1016/0965-8564(96)00004-3

Bellemans, T., Kochan, B., Janssens, D., Wets, G., Arentze, T., & Timmermans, H. (2010). Implementation Framework and Development Trajectory of FEATHERS Activity-Based Simulation Platform. *Transportation Research Record: Journal of the Transportation Research Board*, *2175*, 111–119. http://doi.org/10.3141/2175-13

Ben-Akiva, M., Bowman, J. L., & Gopinath, D. (1996). Travel demand model system for the information era. *Transportation*, *23*(3), 241–266. http://doi.org/10.1007/BF00165704

Bergsma, W. P., & van der Ark, L. A. (2018). Package "cmm." Retrieved from https://cran.r-project.org/package=cmm

Bernardini, C., Silverston, T., & Festor, O. (2014). SONETOR: A social network traffic generator. *2014 IEEE International Conference on Communications, ICC 2014*, 3734–3739. http://doi.org/10.1109/ICC.2014.6883902

Birkin, M. (2008). Hybrid Geographical Models of Urban Spatial Structure and Behaviour. In S. Sergio Albeverio, D. Andrey, P. Giordano, & A. Vancheri (Eds.), *The Dynamics of Complex Urban Systems. An Interdisciplinary Approach* (pp. 95–109). Heidelberg: Physica-Verlag.

Bogle, B. M., & Mehrotra, S. (2016). A Moment Matching Approach for Generating Synthetic Data.pdf. *Big Data*, *4*. http://doi.org/10.1089/big.2016.0015

Bray, J. R., & Curtis, J. T. (1957). An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs*, *27*(4), 326–349.

Caiola, G., & Reiter, J. P. (2010). Random forests for generating partially synthetic, categorical data. *Transactions on Data Privacy*, *3*(1), 27–42.

Casati, D., Müller, K., Fourie, P. J., Erath, A., & Axhausen, K. W. (2015). Synthetic Population Generation by Combining a Hierarchical, Simulation-Based Approach with Reweighting by Generalized Raking. *Transportation Research Record: Journal of the Transportation Research Board*, *2493*, 107–116. http://doi.org/10.3141/2493-12

Casey, B. D. (1983). Estimation of proportions for multinomial contingency tables subject to marginal constraints. *Communications in Statistics - Theory and Methods*, *12*(22). http://doi.org/10.1080/03610928308828624

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. http://doi.org/10.5194/gmd-7-1247-2014

Chen, P., Evans, T., Frisby, M., Izquierdo, E., & Plale, B. (2016). A hybrid approach to population construction for agricultural agent-based simulation. *Proceedings of the 2016 IEEE 12th International Conference on E-Science, e-Science 2016*, 313–322. http://doi.org/10.1109/eScience.2016.7870914

Cho, S., Bellemans, T., Creeemers, L., Knapen, L., Janssens, D., & Wets, G. (2014). Synthetic Population Techniques in Activity-Based Research. In D. Janssens, A.-U.-H. Yasar, & L. Knapen (Eds.), *Data Science and Simulation in Transportation Research* (pp. 48–70). Hershey PA: IGI Global.

Choupani, A. A., & Mamdoohi, A. R. (2015). Tabular Rounding in Iterative Proportional Fitting for Population Synthesis in Activity-Based Models. *TRB 94th Annual Meeting Compendium of Papers*. http://doi.org/10.3141/2493-01

Choupani, A. A., & Mamdoohi, A. R. (2016). Population synthesis using Iterative Proportional Fitting (IPF): A Review and Future Research. *Transportation Research Procedia*, *17*(December 2014), 223–233. http://doi.org/10.1016/j.trpro.2016.11.078

Clarke, G., & Harding, A. (2013). Conclusions and Future Research Directions. In R. Tanton & K. L. Edwards (Eds.), *Spatial Microsimulation: A Reference Guide for Users* (pp. 259–273). Berlin: Springer.

Csiszar, I. (1975). I-Divergence Geometry of Probability Distributions and Minimization Problems. *The Annals of Probability*, *3*(1), 146–158. http://doi.org/10.1214/aop/1176996454

D'Orazio, M., Di Zio, M., & Scanu, M. (2006). *Statistical Matching. Theory and practice.* (John Wiley & Sons, Ed.). Chichester: John Wiley & Sons. http://doi.org/10.1002/0470023554.scard

Demings, W. E., & Stephan, F. F. (1940). On a Least Square Adjustment of a Sampled Frequency Table with the Expected Marginal Totals are Known. *J. of Am Statistics Assoc.*, *11*(4), 427–444.

Deville, J., Sarndal, C., & Sautory, O. (1993). Generalized Raking Procedures in Survey

Sampling Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association*, *88*(423), 1013–1020.

Dobre, A. M., & Caragea, N. (2015). Producing small area estimation using R in the Romanian official statistics. *Romanian Journal of Economics*, *40*, 115–125.

Durán-Heras, A., García-Gutiérrez, I., & Castilla-Alcalá, G. (2017). Comparison of Iterative Proportional Fitting and Simulated Annealing as synthetic population generation techniques: Importance of the rounding method. *Computers, Environment and Urban Systems*, *68*(December 2017), 78–88. http://doi.org/10.1016/j.compenvurbsys.2017.11.001

Farooq, B., Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B: Methodological*, *58*, 243–263. http://doi.org/10.1016/j.trb.2013.09.012

Fenton, A. (2016). Spatial microsimulation estimates of household income distributions in London boroughs, 2001 and 2011. *Centre for Analysis of Social Exclusion. London School of Economics*. London.

Fienberg, S. E. (1970). An Iterative Procedure for Estimation in Contingency Tables. *The Annals of Mathematical Statistics*, *41*(3), 907–917. http://doi.org/10.1214/aoms/1177696968

Frazier, T., & Alfons, A. (2012). Generating a close-to-reality synthetic population of Ghana. *Social Science Research Network (SSRN) 208634*.

Frick, M., & Axhausen, K. W. (2004). Generating Synthetic Populations using IPF and Monte Carlo Techniques: Some New Results. *4th Swiss Transport Research Conference*. Retrieved from http://matsim.org/uploads/ab225.pdf

Fritzemeier, C. J., Dietrich, G. G., & Luangkesorn, L. (2015). Package “glpkAPI” R Interface to C API of GLPK. Retrieved from https://cran.r-project.org/package=glpkAPI

Gargiulo, F., Ternes, S., Huet, S., & Deffuant, G. (2010). An Iterative Approach for Generating Statistically Realistic Populations of Households. *PLoS ONE*, *5*(1), e8828. http://doi.org/10.1371/journal.pone.0008828

Geard, N., Glass, K., McCaw, J. M., McBryde, E. S., Korb, K. B., Keeling, M. J., & McVernon, J. (2015). The effects of demographic change on disease transmission and vaccine impact in a household structured population. *Epidemics*, *13*, 56–64. http://doi.org/10.1016/j.epidem.2015.08.002

Gilbert, N. (2008). *Agent-Based Models*. Sage Publications, Inc.

Grimm, V., Berger, U., DeAngelis, D. L., Polhill, J. G., Giske, J., & Railsback, S. F. (2010). The ODD protocol: A review and first update. *Ecological Modelling*, *221*(23), 2760–2768. http://doi.org/10.1016/j.ecolmodel.2010.08.019

Guo, J. Y., & Bhat, C. R. (2007a). Population Synthesis for Microsimulating Travel Behavior. *Transportation Research Record: Journal of the Transportation Research Board*, 92–101. http://doi.org/10.3141/2014-12

Guo, J. Y., & Bhat, C. R. (2007b). Population Synthesis for Microsimulating Travel Behavior. *Transportation Research Record*, *2014*, 92–101.

Hafezi, M. H., & Habib, M. A. (2014). Synthesizing Population for Microsimulation-based Integrated Transport Models Using Atlantic Canada Micro-data. *Procedia Computer Science*, *37*, 410–415. http://doi.org/10.1016/j.procs.2014.08.061

Hamada, N., Homma, K., Higuchi, H., & Kikuchi, H. (2015). Population Synthesis via k -Nearest Neighbor Crossover Kernel. *IEEE International Conference on Data Mining*, 763–768. http://doi.org/10.1109/ICDM.2015.65

Hanaoka, K., & Clarke, G. P. (2007). Spatial microsimulation modelling for retail market analysis at the small-area level. *Computers, Environment and Urban Systems*, *31*(2), 162–187. http://doi.org/10.1016/j.compenvurbsys.2006.06.003

Harland, K., Heppenstall, A., Smith, D., & Birkin, M. (2012). Creating Realistic Synthetic Populations at Varying Spatial Scales: A Comparative Critique of Population Synthesis Techniques. *Journal of Artificial Societies and Social Simulation*, *15*(1). http://doi.org/10.18564/jasss.1909

Hermes, K., & Poulsen, M. (2012a). A review of current methods to generate synthetic spatial microdata using reweighting and future directions. *Computers, Environment and Urban Systems*, *36*(4), 281–290. http://doi.org/10.1016/j.compenvurbsys.2012.03.005

Hermes, K., & Poulsen, M. (2012b). Small area estimates of smoking prevalence in London. Testing the effect of input data. *Health and Place*, *18*(3), 630–638. http://doi.org/10.1016/j.healthplace.2011.12.010

Himmelmann, L. (2010). Package "HMM." Retrieved from https://www.rdocumentation.org/packages/HMM

Huang, Z., & Williamson, P. (2001). *A Comparison of Synthetic Reconstruction and Combinatorial Optimisation Approaches to the Creation of Small-Area Microdata*. *Department of Geography University of Liverpool* (Vol. workingpap).

Institute for Social and Economic Research. (2018). Euromod. Tax-benefit microsimulation model for the European Union. Retrieved from https://www.euromod.ac.uk/

Instituto Nacional de Estadística. (2017). Retrieved from http://www.ine.es/en/censos2011_datos/cen11_datos_microdatos_en.htm

Janssens, D., Yasar, A.-U.-H., & Knapen, L. (2014). *Data Science and Simulation in Transportation Research*. Hershey PA: IGI Global.

Jeong, B., Lee, W., Kim, D. S., & Shin, H. (2016). Copula-Based approach to synthetic population generation. *PLoS ONE*, *11*(8), 1–28. http://doi.org/10.1371/journal.pone.0159496

Jeske, D. R., Gokhale, D. V., & Ye, L. (2006). Generating synthetic data from marginal fitting for testing the efficacy of data-mining tools. *International Journal of Production Research*, *44*(14), 2711–2730. http://doi.org/10.1080/00207540600622514

Kao, S.-C., Kim, H. K., Liu, C., Cui, X., & Bhaduri, B. L. (2012). Dependence Preserving Approach to Synthesizing HouseHold Characteristics. *Transportation Research Record: Journal of the Transportation Research Board*, (2302), 192–200. http://doi.org/10.3141/2302-21

Kim, J., & Lee, S. (2016). A Simulated Annealing Algorithm for the Creation of Synthetic Population in Activity-Based Travel Demand Model. *KSCE Journal of Civil Engineering*, *20*(6), 2513–2523. http://doi.org/10.1007/s12205-015-0691-7

Kurban, H., Gallagher, R. M., & Persky, J. J. (2012). Estimating Local Redistribution through Property-Tax-Funded Public School Systems. *National Tax Journal*, *65*(3), 629–651. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=ecn&AN=1328291&site=ehost-live&scope=site

Lenormand, M., & Deffuant, G. (2013). Generating a synthetic population of individuals in households: Sample-free vs sample-based methods. *Journal of Artificial Societies and Social Simulation*, *16*(4), 1–16. Retrieved from http://arxiv.org/abs/1208.6403

Levy, J. I., Fabian, M. P., & Peters, J. L. (2014). Community-wide health risk assessment using geographically resolved demographic data: A synthetic population approach. *PLoS ONE*, *9*(1). http://doi.org/10.1371/journal.pone.0087144

Little, R. J. A., & Wu, M.-M. (1991). Models for Contingency Tables With Known Margins When Target and Sampled Populations Differ. *Journal of the American Statistical Association*, *86*(413), 87–95. Retrieved from http://www.jstor.org/stable/2289718

Lovelace, R., & Ballas, D. (2013). 'Truncate, replicate, sample': A method for creating integer weights for spatial microsimulation. *Computers, Environment and Urban Systems*, *41*, 1–11. http://doi.org/10.1016/j.compenvurbsys.2013.03.004

Lovelace, R., Ballas, D., & Watson, M. (2014). A spatial microsimulation approach for the analysis of commuter patterns: from individual to regional levels. *Journal of Transport Geography*, *34*, 282–296. http://doi.org/10.1016/j.jtrangeo.2013.07.008

Lovelace, R., Birkin, M., Ballas, D., & van Leeuwen, E. (2015). Evaluating the Performance of Iterative Proportional Fitting for Spatial Microsimulation: New Tests for an Established Technique. *Journal of Artificial Societies and Social Simulation*, *18*(2), 1–15.

Lundin, E., Kvarnstrom, H., & Jonsson, E. (2002). A Synthetic Fraud Data Generation Methodology. In R. Deng, S. Qing, F. Bao, & J. Zhou (Eds.), *4th. International*

Conference,Information and Comunications Security (pp. 265–277). Singapore.

Ma, J., Heppenstall, A., Harland, K., & Mitchell, G. (2014). Synthesising carbon emission for mega-cities: A static spatial microsimulation of transport CO2 from urban travel in Beijing. *Computers, Environment and Urban Systems*, *45*, 78–88. http://doi.org/10.1016/j.compenvurbsys.2014.02.006

Ma, L. (2011). *Generating Disaggregate Population Characteristics for Input to Travel-Demand Models*. Florida.

Ma, L., & Srinivasan, S. (2015). Synthetic Population Generation with Multilevel Controls: A Fitness-Based Synthesis Approach and Validations. *Computer-Aided Civil and Infrastructure Engineering*, *30*(2), 135–150. http://doi.org/10.1111/mice.12085

Macal, C. M., & North, M. . (2008). Agent Based Modeling and Simulation: ABMS Examples. In S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, & J. W. Fowler (Eds.), *Proceedings of the 2008 Winter Simulation Conference*.

McDonald, J. H. (2014). *Handbook of Biological Statistics* (3rd ed.). Baltimore, Maryland: Sparky House Publishing. Retrieved from http://www.biostathandbook.com/HandbookBioStatThird.pdf

Morrissey, K., Clarke, G., Ballas, D., Hynes, S., & O'Donoghue, C. (2008). Examining access to GP services in rural Ireland using microsimulation analysis. *Area*, *40*(3), 354–364. http://doi.org/10.1111/j.1475-4762.2008.00844.x

Müller, K. (2017a). *A Generalized Approach to Population Synthesis*. ETH Zurich.

Müller, K. (2017b). MultiLevelIPF. Retrieved from https://github.com/krlmlr/MultiLevelIPF

Müller, K., & Axhausen, K. W. (2011). Hierarchical IPF: Generating a synthetic population for Switzerland. In *51st Congress of the European Regional Science Association, Barcelona*. Retrieved from http://ideas.repec.org/p/wiw/wiwrsa/ersa11p305.html

Muñoz, E. (2016). A GREGWT implementation on R. Retrieved from https://github.com/emunozH/GREGWT

Muñoz, E., Vidyattama, Y., & Tanton, R. (2015). A Comparison of the GREGWT and IPF Methods for the Reweighting of surveys. In *Conference: 5th World Congress of the International Microsimulation Association (IMA)*.

Murata, T., Harada, T., & Masui, D. (2017). Comparing Transition Procedures in Modified Simulated-Annealing-Based Synthetic Reconstruction Method without Samples. *SICE Journal of Control, Measurement and System Integration*, *10*(6), 513–519.

Nagle, N. N., Buttenfield, B. P., Leyk, S., & Spielman, S. E. (2013). Dasymetric Modeling and Uncertainty Dasymetric Modeling and Uncertainty. *Annals of the Association of American Geographers*, *10*(May 2013), 1–15.

Navidi, W. C. (2015). *Statistics for engineers and scientists* (4th Editio). New York: McGraw-Hill.

Nissen, V., & Saft, D. (2014). A Practical Guide for the Creation of Random Number Sequences from Aggregated Correlation Data for Multi-Agent Simulations. *Journal of Artificial Societies and Social Simulation*, *17*(2014), 2–9. Retrieved from http://jasss.soc.surrey.ac.uk/17/4/7.html

Nowok, B., Raab, G. M., & Dibben, C. (2016). synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software*, *74(11)*, 1–26. http://doi.org/10.18637/jss.v074.i11

Otani, N., Sugiki, N., & Miyamoto, K. (2012). Goodness-of-Fit Evaluation Method for Agent-Based Household Microdata Sets Composed of Generalized Attributes. *Transportation Research Record: Journal of the Transportation Research Board*, *2254*(1), 97–103. http://doi.org/10.3141/2254-10

Palaumbo, F., Montanari, A., & Vichi, M. (2017). *Data Science: Innovative Developments in Data Analysis and Clustering*. Springer. http://doi.org/10.1007/978-3-319-55723-6

Pendyala, R. M., Kitamura, R., Kikuchi, A., Yamamoto, T., & Fujji, S. (2005). FAMOS: Florida activity mobility simulator. In *Proceedings of the 84th Annual Meeting of the Transportation Research Board*. Washington, DC.

PopGen software. (2017). Retrieved from https://www.mobilityanalytics.org/popgen.html

Pritchard, D. R., & Miller, E. J. (2012). Advances in population synthesis: Fitting many attributes per agent and fitting to household and person margins simultaneously. *Transportation*, *39*(3), 685–704. http://doi.org/10.1007/s11116-011-9367-4

Pukelsheim, F. (2014). Biproportional scaling of matrices and the iterative proportional fitting procedure. *Annals of Operations Research*, *215*(1), 269–283. http://doi.org/10.1007/s10479-013-1468-3

Rahman, A. (2009). Small Area Estimation Through Spatial Microsimulation Models: Some Methodological Issues. In *2 nd International Microsimulation Association Conference , Ottawa* (pp. 8–10).

Rao, J. N. K., & Molina, I. (2015). *Small area estimation*. John Wiley & Sons. http://doi.org/10.1002/9781118735855

Reiter, J. P. (2005). Using CART to Generate Partially Synthetic, Public Use Microdata. *Journal of Official Statistics*, *21*.

Roser, R., Aluja-Banet, T., & Nonell, R. (1999). File Grafting In Market Research. *Applied Stochastic Models in Business and Industry*, *15*(May), 451–460.

Rubin, D. B. (1993). Discussion on statistical disclosure limitation. *Journal of Official Statistics*, *9*(2), 461–468.

Ruther, M., Leyk, S., & Buttenfield, B. (2017). Deriving Small Area Mortality Estimates Using a Probabilistic Reweighting Method. *Annals of the American Association of Geographers*, *107*(6), 1299–1314. http://doi.org/10.1080/24694452.2017.1320213

Ryan, J., Maoh, H., & Kanaroglou, P. (2009). Population synthesis: Comparing the major techniques using a small, complete population of firms. *Geographical Analysis*, *41*(2), 181–203. http://doi.org/10.1111/j.1538-4632.2009.00750.x

Saadi, I., Mustafa, A., Teller, J., Farooq, B., & Cools, M. (2016). Hidden Markov Model-based population synthesis. *Transportation Research Part B: Methodological*, *90*, 1–21. http://doi.org/10.1016/j.trb.2016.04.007

Santoro, M. R., Bewick, G., & Horowitz, M. a. (1989). Rounding algorithms for IEEE multipliers. *Proceedings of 9th Symposium on Computer Arithmetic*, 176–183. http://doi.org/10.1109/ARITH.1989.72824

Schatsky, D., & Chauhan, R. (2017). Machine learning and the five vectors of progress. Retrieved from https://www2.deloitte.com/insights/us/en/focus/signals-for-strategists/machine-learning-technology-five-vectors-of-progress.html

Scutari, M. (2010). Learning Bayesian Networks with the bnlearn R Package. *Journal of Statistical Software*, *35*(3), 1–22. http://doi.org/10.18637/jss.v035.i03

Simpson, L., & Tranmer, M. (2005). Combining Sample and Census Data in Small Area Estimates: Iterative Proportional Fitting with Standard Software. *The Professional Geographer*, *57*(2), 222–234. http://doi.org/10.1111/j.0033-0124.2005.00474.x

Suesse, T., Namazi-Rad, M., Mokhtarian, P., & Barthelemy, J. (2017). Estimating Cross-Classified Population Counts of Multidimensional Tables: An Application to Regional Australia to Obtain Pseudo-Census Counts. *Journal of Official Statistics*, *33*(4), 1021–1050. http://doi.org/10.1515/jos-2017-0048

Sun, L., & Erath, A. (2015). A Bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies*, *61*, 49–62. http://doi.org/10.1016/j.trc.2015.10.010

Swiss Federal Statistical Office. (2017). Retrieved from http://www.portal-stat.admin.ch/pus/files/data_e.html

Tannenbaum, S. J., Holford, N. H. G., Lee, H., Peck, C. C., & Mould, D. R. (2006). Simulation of correlated continuous and categorical variables using a single multivariate distribution. *Journal of Pharmacokinetics and Pharmacodynamics*, *33*(6), 773–794. http://doi.org/10.1007/s10928-006-9033-1

Tanton, R. (2011). Spatial microsimulation as a method for estimating different poverty rates in Australia. *Population, Space and Place*, *17*(3), 222–235. http://doi.org/10.1002/psp.601

Tanton, R., & Edwards, K. L. (Eds.). (2013). *Spatial Microsimulation: A Reference Guide for Users. Understanding population trends and processes 6*. Springer. http://doi.org/10.1007/978-94-007-4623-7

Tanton, R., Williamson, P., & Harding, a. (2014). Comparing two methods of reweighting a survey file to small area data. *International Journal of Microsimulation*, *7*(1), 76–99. http://doi.org/10.3153/jfscom.2008021

Taylor, E., Harding, A., Lloyd, R., & Blake, M. (2004). Housing unaffordability at the statistical local area level: new estimates using spatial microsimulation. *Australasian Journal of Regional Studies*, *10*(3), 279–300. Retrieved from http://search.informit.com.au/documentSummary;dn=096898118346795;res=E-LIBRARY

Templ, M., Meindl, B., Kowarik, A., & Dupriez, O. (2017). Simulation of Synthetic Complex Data: The *R* Package **simPop**. *Journal of Statistical Software*, *79*(10). http://doi.org/10.18637/jss.v079.i10

Train, K. (2002). Discrete Choice Methods with Simulation. *Cambridge University Press*, 1–388. http://doi.org/10.1017/CBO9780511753930

Voas, D., & Williamson, P. (2001). Evaluating Goodness-of-Fit Measures for Synthetic Microdata. *Geographical and Environmental Modelling*, *5*(2), 177–200. http://doi.org/10.1080/13615930120086078

Voraprateep, J. (2013). *Robustness of Wilcoxon Signed-Rank Test Against the Assumption of Symmetry*. Birmingham.

Williamson, P. (2007). Paul Williamson's Home Page. Retrieved October 15, 2015, from http://pcwww.liv.ac.uk/~william/microdata/CO 070615/CO_software.html

Williamson, P. (2013). An evaluation of two synthetic small-area microdata mimulation methodologies: Synthetic Reconstruction and Combinatorial Optimization. In R. Tanton & K. Edwards (Eds.), *Spatial Microsimulation: A Reference Guide for Users* (pp. 19–47). Berlin: Springer.

Williamson, P. (2017). Williamson's Home Page - University of Liverpool. Retrieved from http://pcwww.liv.ac.uk/~william/microdata/CO 070615/CO_software.html

Williamson, P., Birkin, M., & Rees, P. H. (1998). *The Estimation of Population Microdata by using data from Small Area Statistics and Samples of Anonymised records. Environment & planning A* (Vol. 30). http://doi.org/10.1068/a300785

Wilson, A. G., & Pownal, C. E. (1976). A new representation of the urban system for modeling and for the study of micro-level interdependence. *Area*, *8*(4), 246–254.

Wong, D. W. S. (1992). The Reliability of Using the Iterative Proportional Fitting. *The Professional Geographer*, *44*, 340–348.

Xu, Z., Glass, K., Lau, C. L., Geard, N., Graves, P., & Clements, A. (2017). A Synthetic Population for Modelling the Dynamics of Infectious Disease Transmission in American Samoa. *Scientific Reports*, *7*(1), 1–9. http://doi.org/10.1038/s41598-017-17093-8

Yap, B. W., & Sim, C. H. (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, *81*(12), 2141–2155. http://doi.org/10.1080/00949655.2010.520163

Ye, P., Hu, X., Yuan, Y., & Wang, F. (2017). Population Synthesis Based on Joint Distribution Inference Without Disaggregate Samples. *Journal of Artificial Societies and Social Simulation*, *20*(4), 16. http://doi.org/10.18564/jasss.3533

Ye, P., Wang, X., Chen, C., Lin, Y., & Wang, F. (2016). Hybrid Agent Modeling in Population Simulation: Current Approaches and Future Directions. *Journal of Artificial Societies and Social Simulation*, *19*(1). http://doi.org/10.18564/jasss.2849

Ye, X., Konduri, K. C., Pendyala, R. M., Sana, B., & Waddell, P. (2009). A Methodology to Match Distributions of Both Household and Person Attributes in Generation of Synthetic Populations. *88th Annual Meeting of the Transportation Research Board, Washington, D.C.*

Zhu, Y., & Ferreira, J. (2014). Synthetic Population Generation at Disaggregated Spatial Scales for Land Use and Transportation Microsimulation. *Transportation Research Record: Journal of the Transportation Research Board*, *2429*. http://doi.org/10.3141/2429-18

Zhuge, C., Li, X., Ku, C. A., Gao, J., & Zhang, H. (2017). A heuristic-based population synthesis method for micro-simulation in transportation. *KSCE Journal of Civil Engineering*, *21*(6), 2373–2383. http://doi.org/10.1007/s12205-016-0704-1