Muestreo espacialmente balanceado. Aplicaciones en R

TRABAJO FIN DE GRADO

Doble Grado en Matemáticas+Estadística

Autor/a: Cintia María Ojeda Silva
Tutor/a: Ana María Muñoz Reyes

Sevilla, Junio de 2023

Índice general

Resumen

Las técnicas de muestreo tradicionales no suelen contemplar las características propias de los datos distribuidos espacialmente, como la dependencia espacial. Vamos a estudiar procedimientos a partir del muestreo aleatorio simple sin reemplazamiento, del muestreo exponencial y del método pivotal, con probabilidades de inclusión iguales o distintas, en los que integraremos la dependencia espacial. El enfoque adoptado nos servirá para corroborar argumentos intuitivos sobre la necesidad de integrar la dependencia en el muestreo.

Podemos trabajar en cualquier dimensión, pero en el caso de datos con tendencia espacial, si seleccionamos una muestra bien distribuida (balanceada), conseguiremos mejores estimadores.

Abstract

Traditional sampling techniques often do not consider the specific characteristics of spatially distributed data, such as spatial dependence. We will study procedures based on simple random sampling without replacement, the exponential sampling and the pivotal method, with equal or different inclusion probabilities, in which we will integrate spatial dependence. The approach adopted will help us corroborate intuitive arguments about the need to integrate dependence in sampling.

We can work in any dimension, but in the case of data with spatial trends, if we select a well-distributed (balanced) sample, we will achieve better estimators.

Índice de figuras

Índice de tablas

Introducción

Los datos georreferenciados poseen características especiales que han influenciado fuertemente el desarrollo de modelos utilizados para el análisis de datos espaciales (Haining, 2003). Sin embargo, hay mucho menos esfuerzo de investigación dirigido a integrar información espacial en los diseños de muestra y la metodología de recolección de datos.

Las estimaciones de totales, medias y proporciones de algunas variables objetivo son típicamente el principal resultado de una encuesta por muestreo. Desafortunadamente, se suele aplicar métodos de poblaciones infinitas y no se tiene en cuenta cómo se obtuvieron los datos de la muestra. En muchos casos, el uso de esquemas de muestreo complejos implica que se deben utilizar pesos y que las varianzas de los estimadores de la encuesta deben calcularse de una manera que refleje la complejidad del diseño de la muestra. Esto viene recogido ampliamente en Chambers et al. (2012) y Benedetti et al. (2015).

Cuando se seleccionan al azar unidades espaciales de una población finita, su principal característica identificativa es su información de georreferenciación.

Definición: la **georreferenciación** es el proceso de referenciar datos contra un sistema de coordenadas geoespacial conocido, ajustándose a puntos conocidos en el sistema de coordenadas, de manera que los datos puedan ser visualizados, procesados, consultados y analizados junto con otros datos geográficos.

En consecuencia, es evidente que esta distribución espacial debe utilizarse como información estratégica al diseñar el procedimiento de selección de muestra (Vallée et al., 2015; Dickson y Tillé, 2016; Benedetti et al., 2016).

Basándose en la noción intuitiva de que las unidades poblacionales que están más cerca proporcionan menos información sobre un parámetro de la población que las unidades más alejadas, Benedetti & Palma (1995), Arbia & Lafratta (2002), Rogerson & Delmelle (2004) y Bohorquez et al. (2016) utilizaron información espacial sobre las unidades poblacionales para diseñar muestras bien distribuidas, con las que se trabajará en este documento. Una muestra está geográficamente bien distribuida si el número de unidades seleccionadas está cerca de lo que se espera en promedio en cada parte de la región de estudio (Grafström y Lundström, 2013). Estos tipos de diseños de muestreo evitan la selección de unidades geográficas vecinas. Al distribuir espacialmente la muestra en la población objetivo, estos diseños de muestra apuntan a una propiedad específica, que generalmente se conoce como equilibrio espacial (Steven y Olsen, 2004).

La justificación para un enfoque espacialmente balanceado en la selección de muestras es principalmente intuitiva. Dado que se han introducido varios algoritmos de muestreo que buscan lograr un equilibrio espacial (Christman, 2000; Wang et al., 2012), vale la pena abordar esta brecha en la literatura de muestreo.

Este documento revisa el estado actual del estudio del muestreo espacialmente equilibrado. En particular, se describen las principales técnicas para seleccionar muestras espacialmente equilibradas que se han introducido recientemente y se evalúan sus ventajas y desventajas, sus estimadores de la varianza y se desarrollará su implementación más reciente en un paquete de R-Studio (Pantalone, F., Benedetti, R., & Piersimoni, F. (2022). Spbsampling: An R Package for Spatially Balanced Sampling. Journal of Statistical Software, 103, 1-22).

En este trabajo se adopta la perspectiva de encuesta por muestreo de una población finita. Siguiendo este enfoque, también es posible abordar algunos problemas en el monitoreo de recursos naturales y ambientales, aunque estos problemas a menudo se exploran utilizando el marco de poblaciones infinitas (Thompson, 2013).

Se estudiarán distintas técnicas para conseguir muestras balanceadas espacialmente, en particular, veremos en profundidad el método pivotal simple, introducido por Deville y Tillé (1998), para seleccionar muestras con un alto grado de equilibrio espacial. En este procedimiento se incluye la dependencia espacial mediante las distancias entre las unidades. La mayoría de las aplicaciones espaciales se refieren a poblaciones distribuidas en una, dos o tres dimensiones. Sin embargo, el método se puede utilizar para cualquier número de dimensiones porque todo lo que se necesita es una medida de la distancia entre unidades.

Como otra opción de método que utiliza la distancia entre los elementos de la población, se comentará también el método Poisson Condicional (Tillé, Y. (2006), Capítulo 5), que no es más que un diseño exponencial que se define con un soporte concreto.

Por último, se estudiará en profundidad el método GTRS. Stevens y Olsen (2004) generalizaron la idea del muestreo πps (del inglés *Probability Proportional to Size*, muestreo en el que la probabilidad de selección de cada elemento es proporcional a su tamaño dentro de la población) para dos dimensiones, y dio lugar al método GRTS (muestreo estratificado con teselado aleatorio generalizado, del inglés *generalised random tessellation stratified*).

Definición: un **teselado** es una partición del plano mediante polígonos idénticos, o a un polígono o grupo de polígonos idénticos que convenientemente agrupados recubren enteramente el plano.

El GRTS (Benedetti, R., Piersimoni, F., & Postiglione, P. (2017).) usa una correspondencia de la localización en dos dimensiones en una dimensión, conservando cierto orden espacial, para luego poder aplicar el muestreo πps y asegurar que obtenemos una muestra bien distribuida.

Se introduce la siguiente notación: sea $U = (1, 2, .., N)$ una población finita, donde cada unidad i tiene una probabilidad de inclusión prescrita $0 < \pi_i \leq 1$ y una ubicación conocida. Sea $C = (c_1, c_2, ..., c_g, ..., c_h)$ un conjunto de h coordenadas obtenidas mediante la geo-rreferenciación de cada unidad poblacional. Sea $c_g = (c_{1g}, c_{2g}, ...c_{ig}, ...c_{Ng})$ la g-ésima coordenada. Dada C, se calcula la matriz de distancias $D_U = (d_{kl} : k = 1, .., N; l = 1, .., N)$, que especifica la distancia entre dos unidades cualesquiera de la población.

Además, pongamos que el tamaño de muestra esperado es $n = \sum_{i \in U} \pi_i$. El objetivo es estimar el total $Y = \sum_{i \in U} y_i$ de alguna variable interesante con valor yi para la unidad i. Cuando se selecciona una muestra, el total Y se puede estimar mediante el estimador insesgado de Horvitz-Thompson (HT) (Horvitz y Thompson, 1952).

Capítulo 1

Métodos basados en distancias

1.1. Método Pivotal

1.1.1. Descripción

El método pivotal, introducido por Deville y Tillé (1998), es un método de muestreo que permite considerar probabilidades de inclusión desiguales. En cada paso del método pivotal, se actualizan las probabilidades de inclusión para dos unidades de tal manera que el resultado del muestreo se decide, al menos, para una de las dos unidades. Por lo tanto, se obtiene una muestra en un máximo de N pasos. Para describir la regla de actualización, introducimos la siguiente notación:

Se denotan las probabilidades de inclusión posiblemente actualizadas con π_i^*, y se dirá que una unidad i está terminada si $\pi_i^* = 0$ o $\pi_i^* = 1$. Una vez que una unidad está terminada, no se puede elegir de nuevo.

En primer lugar, se deben elegir dos unidades de la población, que serán i y j, con probabilidades de inclusión π_i y π_j respectivamente. Deville y Tillé sugieren que estas unidades se elijan al azar. El vector reducido (π_i,π_j) se actualiza según las siguientes reglas:

$$\cdot \text{ Si } \pi_i + \pi_j < 1, \text{ entonces, } (\pi_i^*, \pi_j^*) = \begin{cases} (0, \pi_i + \pi_j) & \text{con probabilidad } \frac{\pi_j}{\pi_i + \pi_j} \\ (\pi_i + \pi_j, 0) & \text{con probabilidad } \frac{\pi_i}{\pi_i + \pi_j} \end{cases}$$

$$\cdot \text{ Si } \pi_i + \pi_j \geq 1, \text{ entonces, } (\pi_i^*, \pi_j^*) = \begin{cases} (1, \pi_i + \pi_j - 1) & \text{con probabilidad } \frac{1-\pi_j}{2-\pi_i-\pi_j} \\ (\pi_i + \pi_j - 1, 1) & \text{con probabilidad } \frac{1-\pi_i}{2-\pi_i-\pi_j} \end{cases}$$

Las probabilidades se actualizan hasta que todas las unidades hayan terminado, y obtendremos un vector centrado en las n coordenadas de la muestra a partir del vector $(\pi_1, \pi_2, ..., \pi_N)$.

Si ambas unidades se eligen al azar en cada paso, entonces el método pivotal tiene una entropía alta (Grafström, 2010), es decir, la incertidumbre del procedimiento es alta. Para mejorar el equilibrio espacial en el método pivotal, queremos mantener constantes (localmente) la suma de las probabilidades de inclusión actualizadas. Para ello, presentamos los métodos pivotales locales.

1.1.2. Métodos Pivotales Locales

Los métodos pivotales locales (Grafström, A., Lundström, N. L., & Schelin, L. (2012)) actualizan las probabilidades de inclusión según la regla de actualización mencionada anteriormente para dos unidades cercanas en cada paso. A continuación, se muestran dos formas diferentes de elegir las dos unidades cercanas i y j en cada paso.

Estos dos métodos se denominan método pivotal local 1 (LPM 1) y método pivotal local 2 (LPM 2).

El método pivotal local 1.

1. Elija al azar una unidad i.
2. Elija la unidad j, un vecino más cercano a i. Si dos o más unidades tienen la misma distancia a i, entonces elija al azar entre ellas con igual probabilidad.
3. Si j tiene a i como su vecino más cercano, actualice las probabilidades de inclusión según la regla de actualización de la sección anterior. En caso contrario, vaya al paso 1.
4. Si todas las unidades han terminado, el proceso ha terminado. De lo contrario, vuelva al paso 1.

El método pivotal local 2.

Los pasos 1, 2, y 4 son como para LPM 1. En lugar de aplicar el paso 3, actualizamos directamente las probabilidades de inclusión para las unidades i y j según la regla de actualización mencionada en el punto anterior.

Una vez que la unidad ha terminado, se elimina de la población y no se puede elegir de nuevo, y deja de ser considerada vecina de ninguna otra unidad.

El número esperado de cálculos necesarios para seleccionar una muestra LPM 1 es, en el peor de los casos, proporcional a N^3, y en el mejor de los casos proporcional a N^2.

Para LPM 2, el número de cálculos necesarios para seleccionar una muestra es proporcional a N^2.

LPM 1 es más equilibrado, y LPM 2 es más simple y rápido, ya que en el paso 3 aplica la regla de actualización sin tomar otras consideraciones, como sí hace LPM 1.

1.1.3. Equilibrio espacial

A continuación se exponen resultados que indican que los métodos pivotales locales están balanceados espacialmente. En términos generales, esperamos que $\sum_{i\in U_1} I_i \approx \sum_{i\in U_1} \pi_i$ si U_1 son todas las unidades de una subregión de U.

Para el método LPM 1, tenemos el siguiente teorema:

TEOREMA 1. *Sea U una población y sea $U1 \subset U$ con la propiedad de que LPM 1 debe terminar todas las unidades excepto una unidad en U1 antes de posiblemente cambiar $\sum_{i\in U_1} \pi_i^* *$. Entonces, la suma de los indicadores de inclusión satisface,*

$$\lfloor n_1 \rfloor \leq \sum_{i\in U_1} I_i \leq \lceil n_1 \rceil$$

donde $n_1 = \sum_{i\in U_1} \pi_i$. Por tanto, el tamaño de la muestra de U_1 tiene mínima varianza.

Los subconjuntos que satisfacen la hipótesis del Teorema 1 son bastante generales. Sea d(i,j) la distancia entre las unidades i y j. Por ejemplo, si $U1 \subset U$ satisface

$$\max_{i,j\in U_1} d(i,j) < D1 \quad donde \quad D_1 = \min_{i\in U1, j\notin U1} d(i,j)$$

entonces el Teorema 1 se aplica a U1. Una población de este tipo se muestra en la Figura 1a. Mediante el uso repetido del Teorema 1, también obtenemos límites para las uniones de subconjuntos que satisfacen estas condiciones.

Ahora enunciamos una suposición más fuerte sobre la población, lo que da la posibilidad de probar límites para los indicadores de inclusión en el método LPM 2.

TEOREMA 2. *Sea U una población y sea $U_1, U_2, ..., U_m$ una partición de U tal que para k = 1, 2, . . . , m, tenemos*

$$\max_{i,j\in k} d(i,j) < D_k \quad donde \quad D_k = \min_{i\in U_k, j\in U_l, k\neq l} d(i,j)$$

Si $n_k = \sum_{i\in U_k} \pi_i$, entonces la suma de los indicadores de probabilidad verifican que

$$\lfloor n \rfloor - \sum_{l\neq k} \lceil n_l \rceil \leq \sum_{i\in U_k} I_i \leq \lceil n \rceil - \sum_{l\neq k} \lfloor n_l \rfloor$$

para k=1,2,. . . ,m.

Los Teoremas 1 y 2 muestran que los métodos pivotales locales inducen estratificación laxa sin bordes estrictos. En algunos casos, la estratificación con tamaño de muestra fijo dentro de los estratos aparece de manera natural.

Prueba del teorema 2

Asumimos sin pérdida de generalidad que k=1. Comenzamos probando el límite inferior, que se puede encontrar considerando el caso en que todos los conjuntos $U_l, l \neq 1$ eliminan la masa máxima de U1. En particular, para encontrar el "peor caso" para U1, asumimos lo siguiente:

(ı) Todos los conjuntos $U_l, l \neq 1$ eliminan masa de probabilidad de U1 antes de que todas las unidades en U1 terminen.

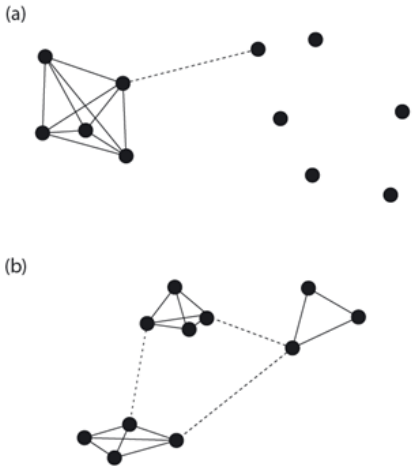


Figura 1.1: *Figura-.(a) Una población con un subconjunto U1, que consiste en las 5 unidades conectadas con una línea, que cumple el Teorema 1. La distancia máxima dentro de U_1 es menor que la distancia mínima D_1 a las unidades fuera de U_1 (línea punteada). (b) Tres subconjuntos, que consisten en las unidades conectadas con una línea continua, que cumplen el Teorema 2. Para cada subconjunto, la distancia máxima dentro del conjunto es menor que la distancia mínima a las unidades en otros conjuntos (líneas punteadas).*

(II) Todos los conjuntos $U_l, l \neq 1$ eliminan masa de U_1 antes de que la masa de probabilidad se mueva entre los conjuntos $U_l, l \neq 1$. Para justificar la suposición (ii), considera que se eligen dos unidades i y j. Antes de actualizar las probabilidades de inclusión, estas unidades pueden eliminar como máximo la masa $2 - (\pi_i^* + \pi_j^*)$ de U_1. Después de la actualización de las probabilidades de inclusión, estas unidades pueden eliminar como máximo

$$1 - (\pi_i^* + \pi_j^*), \quad si \quad \pi_i^* + \pi_j^* < 1$$

$$2 - (\pi_i^* + \pi_j^*), \quad si \quad \pi_i^* + \pi_j^* \geq 1$$

Si asumimos las condiciones (i) y (ii) mencionadas anteriormente, vemos que la masa máxima $U_l, l \neq 1$ que podemos eliminar de U_1 es

$$\lceil n_l \rceil - n_l$$

Por lo tanto, la masa de probabilidad que se puede eliminar de U_1 de otros conjuntos no excede

$$\sum_{l \neq 1} (\lceil n_l \rceil - n_l)$$

Y como se verifica que $\sum_{l \neq 1} n_l = n - n1$, una cota inferior para la suma de los indicadores de inclusión del conjunto U_1 es

$$\lfloor n_1 \rfloor - \sum_{l \neq 1} (\lceil n_l \rceil - n_l) \rfloor = \lfloor n \rfloor - \sum_{l \neq 1} \lceil n_l \rceil \leq \sum_{i \in U_1} I_i$$

Análogamente, obtenemos una cota superior en el caso en el que U_1 recibe masa del resto de conjuntos $U_l, l \neq 1$: la máxima probabilidad de masa que se puede añadir a U_1 es $\sum_{l \neq 1}(n_l - \lfloor n_l \rfloor)$, por tanto,

$$\sum_{i \in U_1} I_i \leq \lceil n_1 \rceil - \sum_{l \neq 1} (n_l - \lfloor n_l \rfloor) \rceil = \lceil n \rceil - \sum_{l \neq 1} \lfloor n_l \rfloor$$

y con esto, queda probada la desigualdad del teorema 2.

■

1.2. Poisson Condicional

1.2.1. Diseño

Es un diseño exponencial que se define con el soporte $S_n = \{s \in S \mid \sum_{k\in U} s_k = n\}$, donde $S = \{0,1\}^N$, llamado soporte simétrico sin reemplazamiento con tamaño muestral fijo. Su implementación es más compleja que la del diseño exponencial habitual. Este diseño se llama Muestreo Poisson Condicional (CP, de las siglas en inglés *Conditional Poisson*) porque se puede obtener seleccionando muestras mediante un muestreo Poisson sin reemplazamiento hasta obtener un tamaño de muestra dado. Varios autores se refieren a él como "diseño exponencial sin reemplazamiento" o "diseño de entropía máxima" porque se puede obtener maximizando la medida de entropía

$$I(p) = -\sum_{s\in S_n} p(s) log p(s)$$

sujeto a probabilidades de inclusión dadas.

El principal problema con la implementación de este diseño es que la función característica no se puede simplificar y parece imposible calcular el vector de probabilidades de inclusión sin enumerar todas las posibles muestras. Sin embargo, Chen et al. (1994) proporcionaron un resultado muy importante: propusieron un algoritmo que permite derivar las probabilidades de inclusión a partir del parámetro y viceversa.

Deville (2000) mejoró este algoritmo. Chen et al. (1994), Chen and Liu (1997), Chen (1998, 2000) y Deville (2000) señalaron que un cálculo rápido del parámetro permite una implementación rápida de este diseño de muestreo.

Como el CP es un diseño exponencial definido en el soporte S_n, se verifica lo siguiente

$$p_{CP}(s,\lambda,n) = p_{EXP}(s,\lambda,S_n) = e^{\lambda s - \alpha(\lambda,S_n)}$$

Para toda $s \in S_n$ y con $\lambda \in \mathbb{R}^n$. Hasta el estudio de Chen et al. (1994), como $\alpha(\lambda, S_n)$ no se podía simplificar, para seleccionar una muestra, había que enumerar todas las muestras posibles de S_n

El vector de probabilidades de inclusión es

$$\pi(\lambda, S_n) = \sum_{s\in S_n} s p_{EXP}(s,\lambda,S_n)$$

Como $Inv(S_n) = \{x \in \mathbb{R}^n | x = a1, \forall a \in \mathbb{R}\}$ el vector λ se puede redefinir para que sea único. Esto se consigue buscando que $\sum_{k\in U} \lambda_k = 0$.

Tenemos un resultado que enuncia que dado un diseño exponencial con $p_{EXP}(s,\lambda,Q)$ sobre un soporte Q, entonces, $\mu(\lambda) = \sum_{s\in Q} s p_{EXP}(s,\lambda,Q)$ es un homeomorfismo entre $\overrightarrow{Q}$ (dirección de Q) y $\overset{\circ}{\mathrm{Q}}$ (interior de Q). Este resultado, aplicado a nuestro caso y teniendo en cuesta que se verifica que $\sum_{k\in U} \pi_k = n$, nos lleva a la conclusión de que la aplicación $\pi(\lambda, S_n)$ es una biyección de

$$\overrightarrow{S_n} = \{\lambda \in \mathbb{R}^n | \sum_{k\in U} \lambda_k = 0\}$$

a

$$\overset{\circ}{\mathrm{S}}_n = \{\pi \in]0,1[^N \mid \sum_{k \in U} \pi_k = n\}$$

Luego, se tiene que

$$\pi(\lambda, S_n) = \sum_{s \in S_n} sp_{EXP}(s, \lambda, S_n) = \frac{\sum_{s \in S_n} s e^{\lambda' s}}{\sum_{s \in S_n} e^{\lambda' s}}$$

Lo cual es imposible de computar cuando U es muy grande. Chen et al. (1994) y posteriormente completado por Deville (2000) buscaron relaciones recursivas entre $\pi(\lambda, S_{n-1})$ y $\pi(\lambda, S_n)$, lo cual permite derivar $\pi(\lambda, S_n)$ con respecto a λ sin enumerar todas las muestras posibles de S.

Aunque hay varias expresiones, un ejemplo es

$$\pi(\lambda, S_n) = (e^{\lambda_k})[1 - \pi_k(\lambda, S_{n-1})] \frac{e^{\alpha(\lambda, S_{n-1})}}{e^{\alpha(\lambda, S_n)}}$$

Capítulo 2

Método GRTS.

2.1. Origen

Hedayat et al. (1988b) sugirieron que se puede obtener más información sobre la población si la muestra no considera pares de unidades contiguas, y propusieron el uso de un diseño de muestreo en el cual las probabilidades de inclusión de segundo orden π_{kl} son no decrecientes en la distancia entre las unidades k y l.

Además, introdujeron un diseño básico denominado diseño de muestreo balanceado excluyendo unidades contiguas (BSEC, por sus siglas en inglés *balanced sampling design excluding contiguous units*). El BSEC es un diseño de tamaño fijo n con la restricción $\pi_{kl} = 0$ si las unidades k y l son contiguas, mientras que todas las otras π_{kl} se establecen como iguales a una constante apropiada. Sin embargo, la desventaja más relevante del BSEC es la suposición de un ordenamiento lineal, luego la idea de excluir las unidades contiguas parece ser demasiado simple si se aplica a datos distribuidos geográficamente.

Es por esto por lo que este último enfoque se ha discutido ampliamente. Stufken (1993) definió los planes de muestreo balanceados excluyendo unidades adyacentes. Este diseño generalizó el BSEC al excluir todos aquellos pares de unidades cuya distancia es menor o igual a un umbral determinado m. Stufken et al. (1999) introdujeron diseños poligonales y mostraron que los diseños poligonales son equivalentes a los planes de muestreo balanceados que excluyen unidades adyacentes. En los trabajos de Hedayat & Stufken (1998), Mandal et al. (2008) y Wright & Stufken (2008) se puede ampliar la información sobre estos métodos.

Para obtener muestras bien distribuidas, una opción común es utilizar la estratificación de las unidades en función de sus posiciones geográficas. Desafortunadamente, esta estrategia no tiene un impacto significativo en las probabilidades de inclusión de segundo orden. Por lo tanto, no está claro cómo obtener una muestra que tenga unidades que no estén cerca entre sí. Estas prácticas inspiraron el diseño GRTS (Stevens y Olsen, 2004). El GRTS selecciona sistemáticamente las unidades y mapea la población espacial bidimensional en una dimensión mientras intenta preservar parte del orden multidimensional. Las primeras contribuciones en el área de los planes basados en teselación se deben a Olea (1984) y Overton y Stehman (1993), quienes analizaron el diseño de estratificación de teselación aleatoria (RTS, por sus siglas en inglés *random tessellation stratified*). El diseño RTS selecciona aleatoriamente puntos a través de un procedimiento de dos pasos.

1. Se ubica aleatoriamente una teselación coherente con una cuadrícula regular sobre el dominio.

2. Se selecciona un punto aleatorio dentro de cada celda de la teselación aleatoria.

Stevens (1997) introdujo el diseño de estratificación de teselación aleatoria anidada de densidad múltiple (MDNRTS, del inglés *multiple-density nested random tessellation stratified*). Esta metodología agrega puntos a una cuadrícula regular de manera que produce una cuadrícula regular más fina con celdas de teselación de forma similar pero más pequeñas. El MDNRTS utiliza datos geográficos en múltiples niveles de detalle espacial y, por lo tanto, puede considerarse un método espacial a escala múltiple.

Stevens y Olsen (2004) aplicaron este último enfoque al diseño GRTS.

2.2. Definición del modelo GRTS

El diseño GRTS extiende la metodología del diseño MDNRTS para producir una serie infinita de cuadrículas anidadas y coherentes. Este proceso genera una función $f(\cdot)$ que transforma un espacio bidimensional en un espacio unidimensional, preservando cierto orden espacial y algunas relaciones de proximidad.

La técnica GRTS se puede resumir en los siguientes pasos:

1. Supongamos que el marco de muestra consta de N unidades distribuidas geográficamente. En este paso, colocamos una cuadrícula cuadrada sobre el marco de la región de estudio.

2. Dividimos la región geográfica en cuatro subregiones y asignamos números al azar a las subregiones. Dividimos de nuevo las subregiones en otras cuatro subregiones y asignamos números al azar de manera independiente a cada nueva subregión, creando direcciones jerárquicas. Continuamos subdividiendo hasta que solo haya una unidad por celda. Este proceso es conocido como método recursivo de cuadrantes. Este proceso preserva las relaciones espaciales de las unidades de muestra.

3. Identificamos cada unidad con una dirección jerárquica de las obtenidas en el paso anterior. Luego, las unidades de muestreo se disponen en línea siguiendo el orden numérico de las direcciones jerárquicas. La línea tiene N longitudes. Este proceso asigna un espacio bidimensional a un espacio unidimensional.

4. La línea se divide en un número de segmentos de igual longitud según el tamaño de muestra requerido. Seleccionamos al azar una unidad de cada segmento.

El GRTS proporciona un diseño de igual probabilidad que se distribuye bien en el área de estudio. Para definir muestras de probabilidad desigual, es necesario asignar una longitud proporcional a su probabilidad de inclusión a cada unidad. Veamos una aplicación de ejemplo del método GRTS:

Los polígonos Voronoi se pueden utilizar para definir una medida estadística que proporcione información sobre la distribución espacial de las muestras. Este elemento es muy

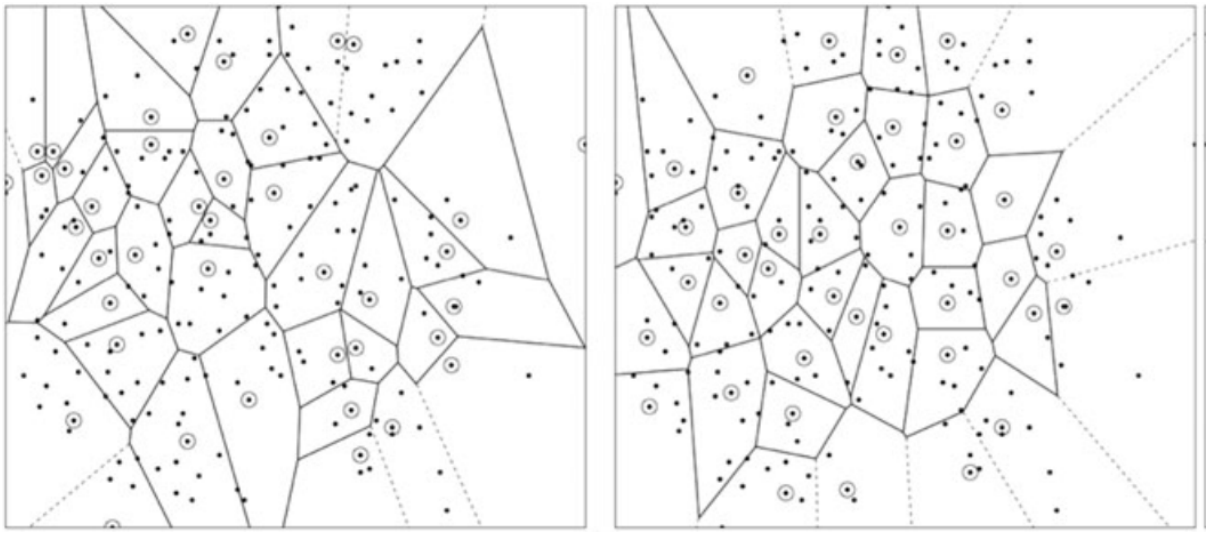


Figura 2.1: *Muestras seleccionadas con polígonos de Voronoi, usando una población de n=40, a la izquierda con MAS y a la derecha con GRTS, donde los ejes son las coordenadas geográficas (x,y)*

útil para comparar la capacidad de un algoritmo de producir un conjunto de puntos bien distribuido sobre la región de estudio.

Ahora definimos $v_k = \sum_{i \in VP(k)} \pi_i$, que es la suma de las probabilidades de inclusión de primer orden de las unidades de la población en el k-ésimo polígono de Voronoi. Para cualquier unidad de muestra, tendremos $E(v_k) = 1$, y para una muestra equilibrada espacialmente, todos los v_k deberían estar cerca de 1 (Stevens y Olsen, 2004).

La varianza $Var(v_k)$ se puede utilizar como una medida de equilibrio espacial para una muestra: cuanto menor sea esta varianza, mayor equilibrio. El problema es que el índice $Var(v_k)$ involucra las π_k, probabilidades de inclusión de primer orden, que son fijas y no se pueden modificar. Por lo tanto, en la práctica existen dificultades para adoptar este índice para diseñar muestras directamente. Por ello, puede ser más apropiado utilizar reglas de selección basadas en la distancia entre las unidades muestreadas y la evaluación a posteriori de $Var(v_k)$ obtenida.

Capítulo 3

Estimación de la varianza.

3.1. Métodos pivotales y Poisson Condicional

En primer lugar, vamos a estudiar distintos estimadores de la varianza para los métodos pivotales y Poisson condicional (CP).

La varianza con el estimador de HT para un tamaño de muestra fijo es

$$V(\hat{Y}) = -\frac{1}{2}\sum_{i,j\in U}(\pi_{ij} - \pi_i\pi_j)\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

donde π_{ij} es la probabilidad de que tanto la unidad i como la j sean incluidas en la muestra (probabilidad de 2º orden).

Si el tamaño muestral es fijo, podemos utilizar el estimador de Sen-Yates-Grundy como estimador de la varianza:

$$\hat{V}(\hat{Y}) = -\frac{1}{2}\sum_{i,j\in s}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

Si todas las probabilidades de inclusión de 2º orden son estrictamente positivas, el estimador es insesgado.

En los métodos pivotales locales, las probabilidades de inclusión de 2º orden dependen de la estructura espacial de la población. Hay casos en los que el estimador de varianza de Sen-Yates-Grundy es siempre cero, aunque la varianza real puede ser grande. Esta situación se da, por ejemplo, cuando las unidades están agrupadas en grupos pequeños y las probabilidades de inclusión suman uno dentro de cada grupo. En tal caso, se selecciona una unidad de cada grupo y las unidades de diferentes grupos se seleccionan de manera independiente entre sí, es decir, para todas las unidades seleccionadas tenemos $\pi_{ij} = \pi_i\pi_j$.

Para la mayoría de las poblaciones, muchas probabilidades de inclusión de segundo orden serán cero al usar los métodos pivotales locales. Por lo tanto, no es posible crear un estimador insesgado basado en diseño de la varianza. Si fuera factible calcular las probabilidades de inclusión de segundo orden, el sesgo de Sen-Yates-Grundy podría ser importante. Es un problema común en el muestreo espacialmente balanceado que las

probabilidades de inclusión de segundo orden puedan ser cero o muy cercanas a cero para unidades (o puntos) que están cerca en distancia. Sin embargo, es importante presentar una medida de variabilidad junto con el estimador.

Una posibilidad es "fingir" que la muestra fue seleccionada con otro diseño. Si se obtiene una muestra mediante muestreo aleatorio independiente (IRS), la varianza podría estimarse por:

$$\hat{V}_{IRS}(\hat{Y}) = \frac{n}{n-1}\sum_{i\in s}\left(\frac{y_i}{\pi_i} - \frac{1}{n}\sum_{j\in s}\frac{y_j}{\pi_j}\right)^2$$

donde s es la muestra. Ver, por ejemplo, Stevens y Olsen (2003). Es habitual que este estimador sobreestime la varianza, independientemente del tamaño de la muestra.

Otra opción es utilizar un estimador de varianza para el diseño CP. Puede ser conservador y exagerar la varianza, pero se espera que funcione mejor que $\hat{V}_{IRS}$. El diseño CP tiene entropía máxima y un estimador aproximado de varianza simple, utilizando solo probabilidades de inclusión de primer orden, que es el estimador de Hájek-Rosén:

$$\hat{V}_{HR}(\hat{Y}) = \frac{n}{n-1}\sum_{i\in s}(1-\pi_i)\left(\frac{y_i}{\pi_i} - \frac{\sum_{j\in s} y_j(1-\pi_j)/\pi_j}{\sum_{j\in s}(1-\pi_j)}\right)^2$$

donde s es la muestra, ver Hájek (1981) y Rosén (1997a, 1997b).

3.2. Método GRTS

En segundo lugar, veamos los estimadores de la varianza usados para el método GRTS.

El diseño GRTS produce una muestra que tiene π_k fijos. En este caso, se puede aplicar el estimador de HT para obtener estimaciones de las características de la población, como hacíamos con los métodos pivotales.

Stevens (1997) proporcionó expresiones exactas para las π_{kl} en un caso específico del GRTS, pero estas expresiones impiden el uso del estimador de varianza basado en HT o en los estimadores de Sen-Yates-Grundy (Särndal et al., 1992) porque tienden a ser inestables debido a la presencia de varios π_{kl} que son muy cercanos a cero (Stevens y Olsen, 2004).

La solución propuesta por Stevens y Olsen (2003) fue un estimador de varianza basado en el contraste para el diseño GRTS. Este enfoque se puede considerar similar al estimador suavizado (Overton y Stehman, 1993). Dado que tanto GRTS como los métodos pivotales locales producen muestras πps bien distribuidas en la población, se espera que el estimador de media-varianza local también funcione bien para los métodos pivotales locales.El estimador de varianza propuesto se define como:

$$\hat{V}_{NBH}(\hat{t}_y) = \sum_{k\in s}\sum_{l\in NB_k} wd_{kl}\left(\frac{y_k}{\pi_k} - \sum_{i\in NB_k} wd_{kt}\frac{y_t}{\pi_t}\right)^2$$

donde NB_k es un vecindario local de la unidad k, y los pesos wd_{kl} disminuyen a medida que aumenta la distancia entre la unidad k y l, con $\sum_k wd_{kl} = \sum_l wd_{kl} = 1$.

La eficiencia de una muestra estratificada espacialmente, como GRTS, mejora a medida que aumenta el número de estratos y disminuye el tamaño de muestra por estrato. La máxima eficiencia se obtiene para un diseño de una unidad por estrato.

3.3. Comparación

Vamos a comparar el equilibrio espacial de los métodos pivotales locales frente al método GRTS. También se incluye el diseño de Poisson condicional (CP) (Hájek, 1981; Tillé, 2006, Capítulo 5) como referencia para comparar con los otros métodos.

Cuando las probabilidades de inclusión son iguales, el diseño CP corresponde a SRS (Muestreo Aleatorio Simple, del inglés *simple random sampling*). El diseño CP ignora por completo el aspecto espacial porque el diseño no se ve afectado por una reubicación de unidades dentro de la población.

El equilibrio espacial se puede medir de diferentes maneras. Utilizaremos el enfoque de polígonos de Voronoi sugerido por Stevens y Olsen (2004). Suponemos que $n = \sum_{i \in U} \pi_i$ es un número entero positivo. Para una muestra s, el polígono de Voronoi para la unidad de muestra $i \in s$ incluye todas las unidades de población más cercanas a i que a cualquier otra unidad muestral.

Sea v_i la suma de las probabilidades de inclusión de todas las unidades en el i-ésimo polígono de Voronoi. Si una unidad de población tiene igual distancia a dos o más unidades de muestra, entonces se incluye en más de un polígono. La probabilidad de inclusión de esa unidad se divide por igual entre cada polígono en el que se incluye.

Para una unidad elegida al azar $i \in s$, tenemos $E(v_i) = 1$ porque hay n unidades en la muestra y $\sum_{i \in s} v_i = \sum_{j \in U} \pi_j = n$. Para una muestra equilibrada espacialmente, todos los valores v_i deberían estar cerca de 1. Por lo tanto, la varianza

$$\frac{1}{n} \sum_{i \in s} (v_i - 1)^2$$

puede ser utilizada como medida de equilibrio espacial para una muestra. Para comparar diferentes diseños de muestreo, necesitamos comparar la media de esta varianza sobre muchas muestras repetidas.

Los siguientes ejemplos y sus correspondientes resultados se extraen del documento de Grafström, A., Lundström, N. L., & Schelin, L. (2012).

EJEMPLO 1. En este ejemplo, comparamos el equilibrio espacial utilizando una población de 267 árboles. Se utilizan probabilidades de inclusión iguales y desiguales.

Cuando son desiguales, las probabilidades se eligen en proporción al área basal de los árboles. Se utilizan dos tamaños de muestra, n = 20 y n = 50. Con probabilidades desiguales y un tamaño de muestra de n = 50, las probabilidades varían entre 0.005 y 0.6. Para ambos tamaños de muestra, se han generado un total de 1000 muestras por cada diseño. El resultado se presenta en una tabla.

Tabla 3.1: Resultados para el Ejemplo 1. El grado de equilibrio espacial comparado con distintos diseños muestrales.

	n=20		n=50	
Diseño	Unequal	Equal	Unequal.	Equal.
LPM 1	0.108	0.126	0.139	0.132
LPM 2	0.113	0.129	0.143	0.141
GRTS	0.163	0.170	0.173	0.173
CP	0.340	0.362	0.322	0.366

La tabla muestra la media de la varianza para 1000 muestras. Un valor bajo indica un nivel alto de equilibrio espacial.

Los métodos pivotales locales produjeron las muestras más equilibradas para esta población. Los dos LPM son similares en términos de equilibrio espacial, pero se espera que el LPM 1 sea ligeramente mejor.

El método GRTS produce muestras mucho más equilibradas que el diseño CP, pero no muestras tan equilibradas como los métodos pivotales locales.

EJEMPLO 2. Para la población del ejemplo 1, también conocemos el volumen estimado para cada árbol. Usamos estos valores como nuestra variable objetivo. Nuestro objetivo es estimar el volumen total del árbol para la población. Usaremos 3 tamaños de muestra diferentes: 20, 50 y 70. Las probabilidades de inclusión se eligen para ser proporcionales al área basal de los árboles. Veámoslo en la siguiente tabla:

Tabla 3.2: Resultados para el Ejemplo 2

Diseño	Vsim	Mean.VIRS.	Mean.VHR.	Mean.VNBH.
n=20				
LPM 1	59.79	68.54	65.09	48.46
LPM 2	58.81	68.84	65.4	48.51
GRTS	61.32	69.12	65.67	48.49
CP	64.73	68.3	64.86	47.56
SRS	2870.71			
n=50				
LPM 1	21.39	27.62	24.25	19.1
LPM 2	20.83	27.33	23.99	18.9
GRTS	22.24	27.64	24.26	19.01
CP	24.41	27.48	24.12	18.61
SRS	1008.82			
n=70				
LPM 1	13.84	19.58	16.18	13.3
LPM 2	14.22	19.47	16.08	13.2
GRTS	14.46	19.7	16.3	13.27
CP	16.05	19.54	16.15	13.05
SRS	654.17			

En la tabla se muestra V_{Sim} para 10,000 muestras. Para el resto de estimadores de la varianza, la tabla contiene la media de 10,000 estimaciones. Se muestran los valores

exactos para SRS, sin utilizar probabilidades de inclusión desiguales.

El muestreo balanceado espacialmente con probabilidades de inclusión desiguales es eficiente cuando hay tendencias espaciales en las razones y_i/π_i. Sin embargo, para esta población, no hay una tendencia suave en las razones. Hay algunos grupos pequeños con razones pequeñas, por lo que las razones tienen algunos saltos grandes en lugar de cambiar suavemente en el espacio. Esto puede explicar por qué el estimador de varianza media local ($\hat{V}_{NBH}$) produce subestimaciones de la varianza.

También explica por qué la ganancia de usar un diseño de muestreo equilibrado espacial es bastante pequeña. Sin embargo, ambos métodos pivotales locales muestran un comportamiento similar al método GRTS.

El estimador de varianza para el método GRTS da varianzas estimadas similares para los métodos pivotales locales y para el método GRTS. Lo mismo ocurre para los otros estimadores de varianza.

Capítulo 4

Aplicaciones en R

4.1. Métodos basados en distancias

En esta sección nos centramos en dos diseños de muestreo introducidos por Benedetti y Piersimoni (2017b) e implementados en el paquete Spbsampling. Estos métodos seleccionan una muestra con un tamaño fijo n y se basan en un procedimiento iterativo que utiliza una matriz de distancias y un índice resumen. La matriz de distancias $D_U = \{d_{ij}; i = 1, ..., N; j = 1, ..., N\}$ contiene las distancias para todos los pares de unidades de la población.

Dichas distancias se pueden calcular con cualquier métrica y en cualquier dimensión. Las más habituales son la distancia euclídea y la distancia de Manhattan.

Para medir el equilibrio espacial de una muestra, recurrimos al enfoque de Stevens y Olsen (2004) y utilizamos un índice de equilibrio espacial (SBI, siglas del inglés *Spatial Balance Index*) basado en polígonos de Voronoi. Dada una muestra s, podemos construir un polígono de Voronoi para cada unidad i de la muestra de manera que incluya todas las unidades de la población más cercanas a i que a cualquier otro elemento j. Luego, el SBI toma la siguiente forma:

$$SBI(s) = \frac{1}{n}\sum_{i\in s}(v_i - 1)^2$$

donde v_i es la suma de las probabilidades de inclusión de todas las unidades en el polígono de Voronoi correspondiente a la i-ésima unidad. Dado que la unión de los polígonos de Voronoi constituye una partición de la población y el tamaño de la muestra es igual a n, es decir, $\sum_{i\in s} v_i = \sum_{j=1}^{N} \pi_j = n$, se verifica que, para cualquier unidad de muestra i, tenemos $E(v_i) = 1$.

Para una muestra espacialmente balanceada, $vi \approx 1$ para cada i, de modo que los valores de SBI cercanos a cero corresponden a muestras más dispersas, por tanto, cuanto menor sea el SBI, mayor será la dispersión.

Sea $M_*(D_s)$ un índice de la matriz de distancias de una muestra dada s que resume las distancias de las unidades de la muestra. Esperamos que cuanto mayor sea $M_*(D_s)$, menor sea el SBI.

Consideramos

$$M_0(D_s) = \prod_{i \in s} \prod_{j \in s, j \neq i} d_{ij}$$

y

$$M_1(D_s) = \sum_{i \in s} \sum_{j \in s, j \neq i} d_{ij}$$

El proceso iterativo que define los métodos es el siguiente: denotemos con $s^{(t)} = \{l_1, ..., l_n\}$ la configuración en la iteración t, que es un vector de etiquetas de unidades. Una configuración es, de hecho, una muestra. El algoritmo comienza con una muestra aleatoria simple de dimensión n, $s^{(0)}$. Luego, cada iteración t se compone de los siguientes pasos:

1. Seleccionar dos unidades al azar: una dentro y otra fuera de la configuración actual $s^{(t)}$.
2. Definir una nueva configuración $s_e^{(t)}$,tal que las unidades i y j seleccionadas en el paso anterior se intercambien. Por lo tanto, $s^{(t)}$ y $s_e^{(t)}$ difieren solo en un elemento.
3. Actualizar a la nueva configuración $s_e^{(t)}$ o conservar la configuración actual $s^{(t)}$, siguiendo alguna regla de aceptación.

En particular, actualizar a la nueva configuración $s_e^{(t)}$ con probabilidad:

$$p = \text{mín}\left[1, \left(\frac{M(D_{s_e^{(t)}})}{M(D_{s^{(t)}})}\right)^{\beta}\right]$$

donde β es una constante conocida. Cabe destacar que, cuanto mayor sea el índice $M(D_{s^{(t)}})$, mayor será la distancia general de la nueva configuración y mayor será la probabilidad de actualizar la configuración al nuevo estado. Por lo tanto, la regla de aceptación brinda más oportunidades para permanecer en configuraciones que están relativamente más dispersas.

Además, el parámetro $\beta \in \mathbb{R}$ regula el grado de dispersión: cuanto mayor sea su valor, más dispersa será la muestra. Por lo tanto, este parámetro nos permite elegir de antemano el nivel de dispersión deseado, si corresponde.

4. Repetimos N veces los pasos 1,2,3 mencionados anteriormente.

En teoría podemos usar cualquier índice $M_*(D_s)$, pero los implementados en el paquete SpbSampling corresponden a los diseños muestrales PWD (siglas en inglés de *product within distance*, es decir, producto dentro de una distancia) y SWD (suma dentro de una distancia, del inglés *sum within distance)*), usando los índices $M_0(D_s)$ y $M_1(D_s)$, respectivamente.

Los dos procedimientos descritos hasta ahora tienen un costo computacional de orden, al menos, N^2. Cuando la población objetivo es muy grande o estamos interesados en seleccionar un gran número de muestras, estos métodos podrían ser muy complejos en términos de cálculos.

En esta situación, el uso del diseño de muestreo HPWD (heurística de producto dentro de la distancia, de las siglas en inglés *heuristic product within distance*) (Benedetti y Piersimoni 2017a) puede abordar el problema, ya que es un algoritmo dibujado paso a paso que en un máximo de n pasos selecciona una muestra espacialmente balanceada de dimensión fija n a través de una secuencia de selecciones aleatorias de tamaño 1 con probabilidades de selección que varían, las cuales se actualizarán en cada paso dependiendo de la unidad seleccionada en el paso anterior.

El algoritmo elige en cada paso el óptimo local (Cormen, Leiserson, Rivest y Stein 2009). El procedimiento a seguir es el siguiente:

1. Selección aleatoria de una unidad i con igual probabilidad.
2. En cada paso $t \leq n$, las probabilidades de selección $\pi_j^{(t)}$ para el resto de elementos j de la población se actualizan según la regla

$$\pi_j^{(t)} = \frac{\pi_j^{(t-1)} \bar{d}_{ij}}{\sum_{j=1;j\neq i} \pi_j^{(t-1)} \bar{d}_{ij}} \quad \forall j = 1, ...N; j \neq i$$

donde i es la unidad seleccionada aleatoriamente en el paso t-1, y $\bar{d}_{ij} = \phi(d_{ij}^{\beta})$, donde ϕ es una estandarización aplicada a D_U con el fin de tener productos conocidos y fijos por fila (y columna), y β es una constante conocida y tiene el mismo papel que se ha mencionado anteriormente para los diseños SWD y PWD.

Cabe destacar que $d_{ii} = 0$ por definición, por lo que las unidades ya seleccionadas en la muestra tienen $\pi_i = 0$, por lo tanto, la selección aleatoria es sin reemplazamiento.

4.2. Paquete Spbsampling

El paquete Spbsampling (Pantalone, F., Benedetti, R., & Piersimoni, F. (2022)) implementa diseños de muestreo balanceado espacialmente. En particular, los diseños implementados son PWD, SWD y HPWD, que hemos ilustrado en la sección anterior, y se implementan con `` `pwd()` ``, `` `swd()` `` y `` `hpwd()` ``, respectivamente.

Las funciones `` `pwd()` `` y `` `swd()` `` requieren los siguientes argumentos:

- dis: matriz de distancia de la población objetivo (de dimensión N × N);
- n: tamaño de la muestra;
- beta: parámetro β para el algoritmo, por defecto establecido en 10;
- nrepl: número de muestras a extraer, por defecto establecido en 1;
- niter: número máximo de iteraciones para el algoritmo, establecido en 10 por defecto.

Ambas funciones devuelven una lista con el objeto s, que es una matriz de dimensión nrepl × n que contiene las nrepl muestras seleccionadas, cada una de ellas almacenada en una fila (en particular, la fila b contiene todas las etiquetas de las unidades seleccionadas en la muestra b), y el objeto iteraciones, que es un vector de longitud nrepl donde el elemento b-ésimo contiene el número de iteraciones realizadas por el algoritmo para seleccionar la muestra b.

Como el análisis y las conclusiones son análogas, vamos a dejar el valor por defecto de nrepl, es decir, vamos a extraer una sola muestra en cada uno de los ejemplos (b=1).

La función `` `hpwd()` `` utiliza los mismos argumentos que `` `pwd()` `` y `` `swd()` ``, excepto niter. La salida es una matriz de muestras seleccionadas, como se describe para el objeto s de la salida de `` `pwd()` `` y `` `swd()` ``.

Necesitamos otras funciones que realicen la estandarización de la matriz de distancia y calculen el índice de equilibrio espacial, ya que estos elementos tienen grandes implicaciones en nuestros diseños de muestreo.

Las funciones utilizadas para realizar esta tarea son `` `stprod()` ``, diseñada para el diseño PWD y HPWD, y `` `stsum()` ``, diseñada para el diseño SWD. Tienen los mismos argumentos, y son los siguientes:

- mat: matriz de distancia que se va a estandarizar (de dimensión N × N);
- con: restricciones, de longitud N (porque tenemos una restricción para cada fila/columna de la matriz de distancia);
- differ: diferencia máxima aceptada entre las sumas de las filas/columnas de la matriz de distancia y las restricciones requeridas; por defecto establecido en 1e-15;
- niter: número máximo de iteraciones; por defecto establecido en 1000.

La salida es una lista con los objetos mat, iterations y conv, que son la matriz estandarizada, el número de iteraciones realizadas por el algoritmo y la convergencia alcanzada, respectivamente. Es necesario señalar que la estandarización de la matriz de distancia a través del uso de la función `` `stprod()` `` podría tomar algunos minutos, especialmente cuando el tamaño de la población N es grande.

Finalmente, la función `` `sbi()` `` permite calcular el índice de equilibrio espacial dado por

$$SBI(s) = \frac{1}{n}\sum_{i\in s}(v_i - 1)^2$$

Los datos de entrada son:

- dis: matriz de distancia;
- pi: vector de las probabilidades de inclusión de primer orden de las unidades en la población;
- s: vector de etiquetas de unidades muestreadas.

4.2.1. Ejemplos

4.2.1.1. LUCAS

Usaremos el conjunto de datos de lucas_abruzzo de la librería Spbsampling, que contiene datos de las tierras de la región italiana de Abruzzo, proporcionados por el progarama europeo en el campo de encuestas (LUCAS, land use/cover area frame statistical survey), creado y ejecutado por Eurostat.

Tenemos una población de tamaño 2699.

En primer lugar, vamos a ver el procedimiento general para realizar el muestreo, y luego haremos un breve análisis sobre el efecto del parámetro β en las muestras seleccionadas.

Uso del paquete.

· El procedimiento clásico lleva a cabo estos pasos:

1. Matriz de distancias D_U
2. Estandarizar la matriz según el método de muestreo utilizado (`stprod `o `stsum`)
3. Selección de la muestra usando la función correspondiente.

Podemos calcular la matriz de distancias con el comando `dist()`, que calcula por defecto la distancia Euclídea, aunque podemos indicar la distancia deseada mediante el argumento **method**. También podemos usar `st_distance` del paquete **sf**.

Veamos cómo se calcularía la matriz de distancias:

```
library(Spbsampling)
data("lucas_abruzzo", package = "Spbsampling")
dis_la <- as.matrix(dist(cbind(lucas_abruzzo$x, lucas_abruzzo$y)))
```

O con el paquete sf, cuya ventaja es que la matriz de distancias se expresa en metros, aunque las coordenadas se den en latitud y longitud.

```
library(sf)
lucas_abruzzo_sf <- st_as_sf(lucas_abruzzo, coords = c("x", "y"))
dis_la_sf <- st_distance(lucas_abruzzo_sf)
```

PWD

Asumimos que las probabilidades de inclusión son iguales, n/N. El primer paso es estandarizar la matriz de distancias. Cuando usamos el modelo PWD, necesitamos restringir la suma de las filas (o la columnas) de la matriz con la transformación logarítmica a una constante conocida (por ejemplo, a 0). La función `stprod` está diseñada específicamente para esto, y la realizamos con restricciones iguales a 0, a través del siguiente código:

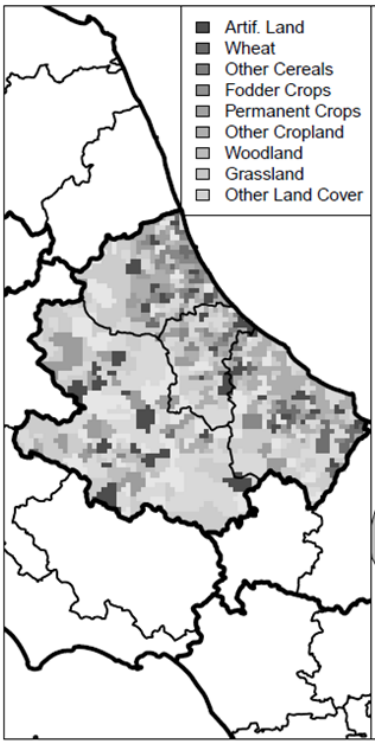


Figura 4.1: *Población lucas_abruzzo*

```
con <- rep(0, nrow(dis_la))
stand_dist_la_pwd <- stprod(mat = dis_la, con = con)$mat
```

Una vez que tenemos estandarizada la matriz de distancias, seleccionamos una muestra de tamaño 10, según el método PWD:

```
set.seed(12345)
s_pwd_la <- pwd(dis = stand_dist_la_pwd, n = 10)$s
s_pwd_la
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,] 1381  746   43 2262 2297  484 1144  634 2338  2071
```

Aquí tenemos las etiquetas de las unidades seleccionadas. Si queremos las coordenadas de los puntos seleccionados, podemos obtenerlas con el siguiente código:

```
lucas_abruzzo[s_pwd_la[1, ], c("x", "y")]
```

```
##             x       y
## 42995 1928000 4680000
## 41462 1852000 4708000
## 38722 1904000 4748000
## 44882 1858000 4654000
```

```
## 45014 1922000 4654000
## 40711 1866000 4720000
## 42467 1884000 4690000
## 41170 1934000 4710000
## 45166 1896000 4652000
## 44408 1952000 4664000
```

Por último, calculamos el SBI con la función `sbi()`. Para ello, necesitamos las probabilidades de inclusión, que son todas iguales a n/N ya que hemos estandarizado la matriz, y las etiquetas de las unidades seleccionadas:

```
pi <- rep(10 / nrow(lucas_abruzzo), nrow(lucas_abruzzo))
sbi(dis = dis_la, pi = pi, s = s_pwd_la[1, ])
```

```
## [1] 0.0617412
```

Recordemos que a menor SBI, mejor distribuidos estarán los datos.

SWD

Seleccionamos una muestra de tamaño 10 a partir de la matriz de distancias calculada anteriormente. Como utilizamos un modelo distinto, estandarizamos la matriz y restringimos las sumas de las filas (o de las columnas) de la matriz a una constante conocida (por ejemplo, 1), usando la función `stsum()` y la restricción:

```
con <- rep(1, nrow(dis_la))
stand_dist_la_swd <- stsum(mat = dis_la, con = con)$mat
```

Seleccionamos la muestra con el método, mediante la función `swd()`

```
set.seed(12345)
s_swd_la <- swd(dis = stand_dist_la_swd, n = 10)$s
s_swd_la
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,] 1858 2091 2665   12 1334  664  852 1968 2012  2266
```

Al igual que antes, esto nos da las etiquetas de los datos seleccionados. Veamos el SBI:

```
pi <- rep(10 / nrow(lucas_abruzzo), nrow(lucas_abruzzo))
sbi(dis = dis_la, pi = pi, s = s_swd_la[1, ])
```

```
## [1] 0.4297499
```

HPWD

De nuevo, necesitamos estandarizar la matriz de distancias y tomar una muestra. Usaremos para ello la matriz `stand_dist_la_pwd` y la función `hpwd()`:

```
set.seed(12345)
s_hpwd_la <- hpwd(dis = stand_dist_la_pwd, n = 10)
s_hpwd_la
```

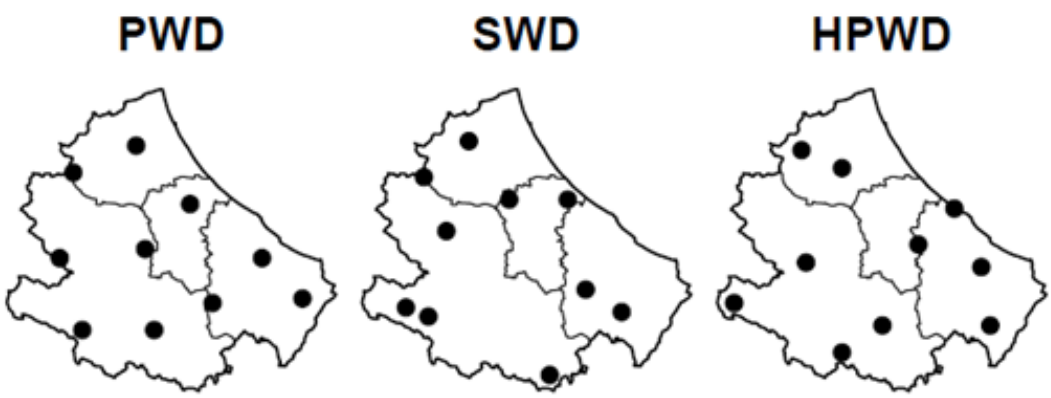
```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,] 1946 1199  542 2592  142 1287 1496 2023 1068  2311
```

Y calculamos el índice SBI:

```
pi <- rep(10 / nrow(lucas_abruzzo), nrow(lucas_abruzzo))
sbi(dis = dis_la, pi = pi, s = s_hpwd_la[1, ])
```

```
## [1] 0.05943573
```

Vamos a ver cómo se distribuyen las muestras que hemos obtenido por los 3 métodos:



4.2.1.2. Meuse.grid

El conjunto de datos meuse.grid, del paquete sp, contiene datos de la llanura del río Meuse, cerca de la población de Stein (Países Bajos), y se refiere a las concentraciones de metales pesados en la capa superior del suelo (junto con una serie de variables del suelo y del paisaje en las ubicaciones de la encuesta).

En particular, dada la cuadrícula construida sobre la zona de estudio, en este análisis se seleccionarán muestras de la propia rejilla, en las que posteriormente se tomarán mediciones de las distintas variables.

La población tiene 3103 elementos.Con este conjunto de datos vamos a estudiar el efecto del parámetro β en las muestras seleccionadas.

El papel de β es crucial en estos diseños muestrales, ya que controla que la muestra esté bien distribuida: a mayor β, mejor distribuida se espera que esté. Además, vale la pena señalar que para valores negativos de β, estos diseños de muestreo proporcionan muestras por conglomerados (no estamos interesados en este caso, así que no contemplaremos tales valores de β).

Además, para $\beta = 0$ observamos que estos diseños de muestreo se comportan aproximadamente como el método de muestreo aleatorio simple.

En primer lugar, calculamos la matriz de distancia para esta población.

```
library("sp")
data("meuse.grid", package = "sp")
dis_m <- as.matrix(dist(cbind(meuse.grid$x, meuse.grid$y)))

head(meuse.grid)
```

```
##        x      y part.a part.b      dist soil ffreq
## 1 181180 333740      1      0 0.0000000    1     1
## 2 181140 333700      1      0 0.0000000    1     1
## 3 181180 333700      1      0 0.0122243    1     1
## 4 181220 333700      1      0 0.0434678    1     1
## 5 181100 333660      1      0 0.0000000    1     1
## 6 181140 333660      1      0 0.0122243    1     1
```

Vamos a obtener las muestras mediante los distintos métodos vistos: fijamos n=20, y usaremos distintos valores de $\beta = 1, 5, 10$

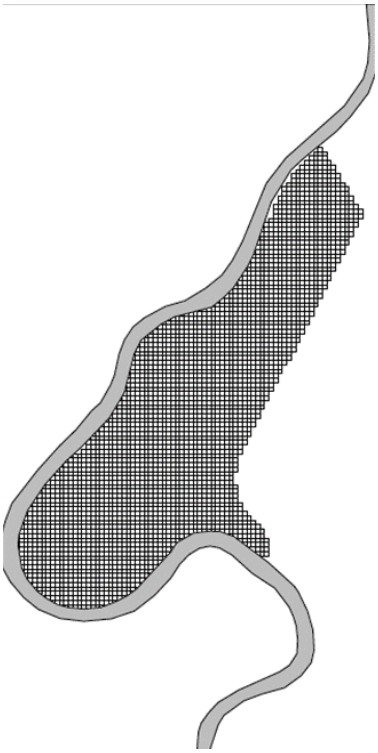


Figura 4.2: *Población de los datos Meuse*

PWD

En primer lugar, estandarizamos la matriz de distancias:

```
con <- rep(0, nrow(meuse.grid))
stand_dist_m_pwd <- stprod(mat = dis_m, con = con)$mat
```

Obtenemos las muestras para los distintos β y calculamos el índice SBI:

$\beta = 1$

```
set.seed(12345)
s_pwd1 <- pwd(dis = stand_dist_m_pwd, n = 20, beta = 1)$s
```

```
pi <- rep(20 / nrow(meuse.grid), nrow(meuse.grid))
sbi(dis = dis_m, pi = pi, s = s_pwd1[1, ])
```

```
## [1] 0.07351108
```

$\beta = 5$

```
set.seed(12345)
s_pwd2 <- pwd(dis = stand_dist_m_pwd, n = 20, beta = 5)$s
```

```
pi <- rep(20 / nrow(meuse.grid), nrow(meuse.grid))
sbi(dis = dis_m, pi = pi, s = s_pwd2[1, ])
```

```
## [1] 0.06826356
```

$\beta = 10$

```
set.seed(12345)
s_pwd3 <- pwd(dis = stand_dist_m_pwd, n = 20, beta = 10)$s
```

```
pi <- rep(20 / nrow(meuse.grid), nrow(meuse.grid))
sbi(dis = dis_m, pi = pi, s = s_pwd3[1, ])
```

```
## [1] 0.04786995
```

Observamos que, a mayor valor de β, menor es el SBI, lo cual indica que la muestra está más balanceada.

HPWD

Una vez estandarizada la matriz, ya obtenida en el método anterior, vamos a obtener las muestras y el índice SBI:

$\beta = 1$

```
set.seed(12345)
s_hpwd1 <- hpwd(dis = stand_dist_m_pwd, n = 20, beta = 1)
```

```
pi <- rep(20 / nrow(meuse.grid), nrow(meuse.grid))
sbi(dis = dis_m, pi = pi, s = s_hpwd1[1, ])
```

```
## [1] 0.105215
```

$\beta = 5$

```
set.seed(12345)
s_hpwd2 <- hpwd(dis = stand_dist_m_pwd, n = 20, beta = 5)
```

```
pi <- rep(20 / nrow(meuse.grid), nrow(meuse.grid))
sbi(dis = dis_m, pi = pi, s = s_hpwd2[1, ])
```

```
## [1] 0.08956204
```

$\beta = 10$

```
set.seed(12345)
s_hpwd3 <- hpwd(dis = stand_dist_m_pwd, n = 20, beta = 10)
```

```
pi <- rep(20 / nrow(meuse.grid), nrow(meuse.grid))
sbi(dis = dis_m, pi = pi, s = s_hpwd3[1, ])
```

```
## [1] 0.06676638
```

ESe observa que el índice SBI va disminuyendo a media que aumenta el coeficiente β.

SWD

Seguimos el procedimiento general: el primer paso es estandarizar la matriz.

```
con <- rep(1, nrow(meuse.grid))
stand_dist_m_swd <- stsum(mat = dis_m, con = con)$mat
```

Obtenemos las muestras y calculamos el SBI:

$\beta = 1$

```
set.seed(12345)
s_swd1 <- swd(dis = stand_dist_m_swd, n = 20, beta = 1)$s
```

```
pi <- rep(20 / nrow(meuse.grid), nrow(meuse.grid))
sbi(dis = dis_m, pi = pi, s = s_swd1[1, ])
```

```
## [1] 0.564162
```

$\beta = 5$

```
set.seed(12345)
s_swd2 <- swd(dis = stand_dist_m_swd, n = 20, beta = 5)$s
```

```
pi <- rep(20 / nrow(meuse.grid), nrow(meuse.grid))
sbi(dis = dis_m, pi = pi, s = s_swd2[1, ])
```

```
## [1] 0.3696178
```

$\beta = 10$

```
set.seed(12345)
s_swd3 <- swd(dis = stand_dist_m_swd, n = 20, beta = 10)$s
```

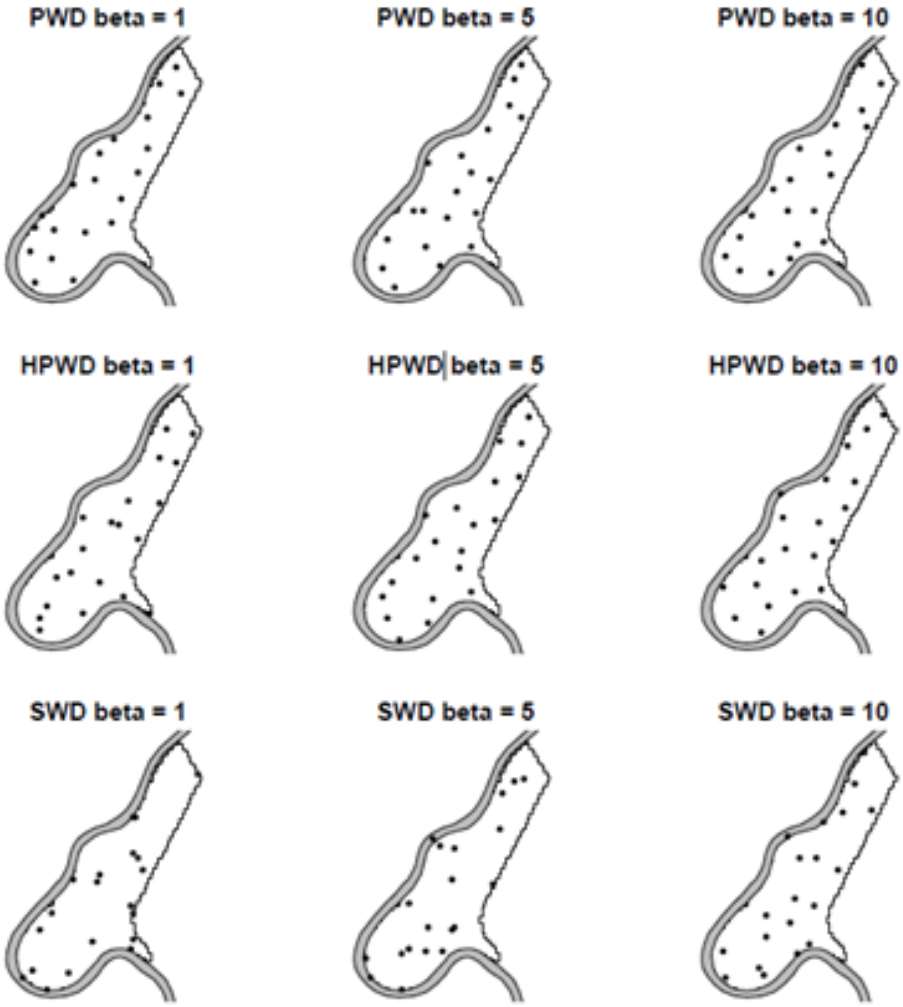


Figura 4.3: *Muestras de la población Meuse obtenidas mediante los 3 métodos estudiados y para distintos valores del parámetro*

```
pi <- rep(20 / nrow(meuse.grid), nrow(meuse.grid))
sbi(dis = dis_m, pi = pi, s = s_swd3[1, ])
```

```
## [1] 0.3745429
```

Llegamos a la misma conclusión: como era de esperar, el índice SBI va disminuyendo a media que aumenta el coeficiente β.

4.2.1.3. Conclusiones

Benedetti y Piersimoni (2017a) y Benedetti y Piersimoni (2017b) demostraron que los métodos implementados en el paquete suelen tener un mejor desempeño en términos de equilibrio espacial que la mayoría de los otros diseños de muestreo espacialmente balanceados implementados en R. En particular, mostraron que el diseño PWD con un valor de $\beta = 10$ tiene un mejor desempeño que los diseños GRTS (Stevens y Olsen 2004), CP (Grafström 2012) y LPM (Grafström et al. 2012). Una posible desventaja es la necesidad de una matriz de distancias estandarizada como entrada, lo cual podría llevar un tiempo considerable de cálculo en caso de una población grande.

4.3. Paquete spsurvey (método GRTS)

En Dumelle et al.(2023) se desarrolla en profundidad el paquete spsurvey, en el que se implementa el método GRTS.

Utilizamos los datos NE_Lakes del paquete spsurvey. Los datos NE_Lakes son un objeto 'sf' de 195 lagos en el noreste de Estados Unidos.

```
data("NE_Lakes", package = "spsurvey")
NE_Lakes <- sp_frame(NE_Lakes)
head(NE_Lakes)
```

```
## Simple feature collection with 6 features and 4 fields
## Geometry type: POINT
## Dimension:     XY
## Bounding box:  xmin: 1849399 ymin: 2313865 xmax: 2017323 ymax: 2417191
## Projected CRS: NAD83 / Conus Albers
##        AREA AREA_CAT   ELEV ELEV_CAT                geometry
## 1 10.648825    large 264.69     high POINT (1930929 2417191)
## 2  2.504606    small 557.63     high POINT (1849399 2375085)
## 3  3.979199    small  28.79      low POINT (2017323 2393723)
## 4  1.645657    small 212.60     high POINT (1874135 2313865)
## 5  7.489052    small 239.67     high POINT (1922712 2392868)
## 6 86.533725    large 195.37     high POINT (1977163 2350744)
```

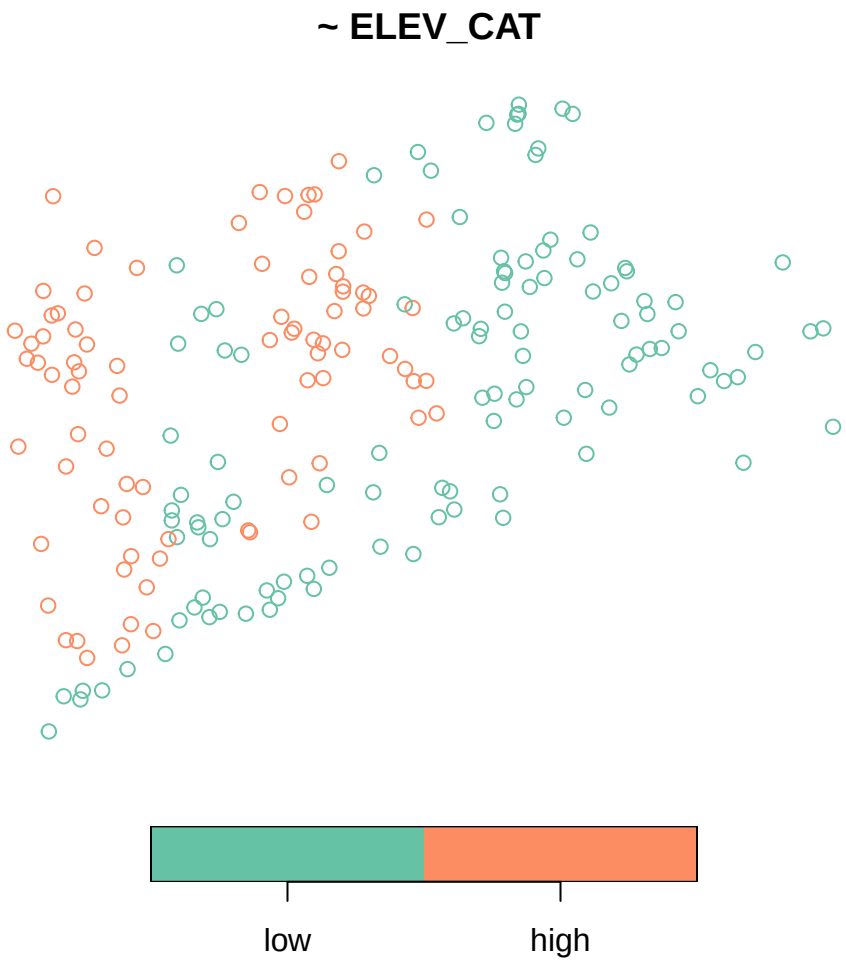
Hay cinco variables: AREA, una variable continua que representa el área del lago (en hectáreas); AREA_CAT, una variable categórica que representa niveles de área del lago pequeña (1 a 10 hectáreas) y grande (más de 10 hectáreas); ELEV, una variable continua que representa la elevación del lago (en metros); y ELEV_CAT, una variable categórica que representa niveles de elevación del lago baja (0 a 100 metros) y alta (más de 100 metros), y geometry (dato de tipo punto).

Se puede mostrar resumen y la representación gráfica de las variables con los siguientes comandos (análogo para cualquier variable):

```
summary(NE_Lakes, formula = ~ ELEV_CAT)
```

```
##    total     ELEV_CAT
##  total:195   low :112
##              high: 83
```

```
plot(NE_Lakes, formula = ~ ELEV_CAT)
```

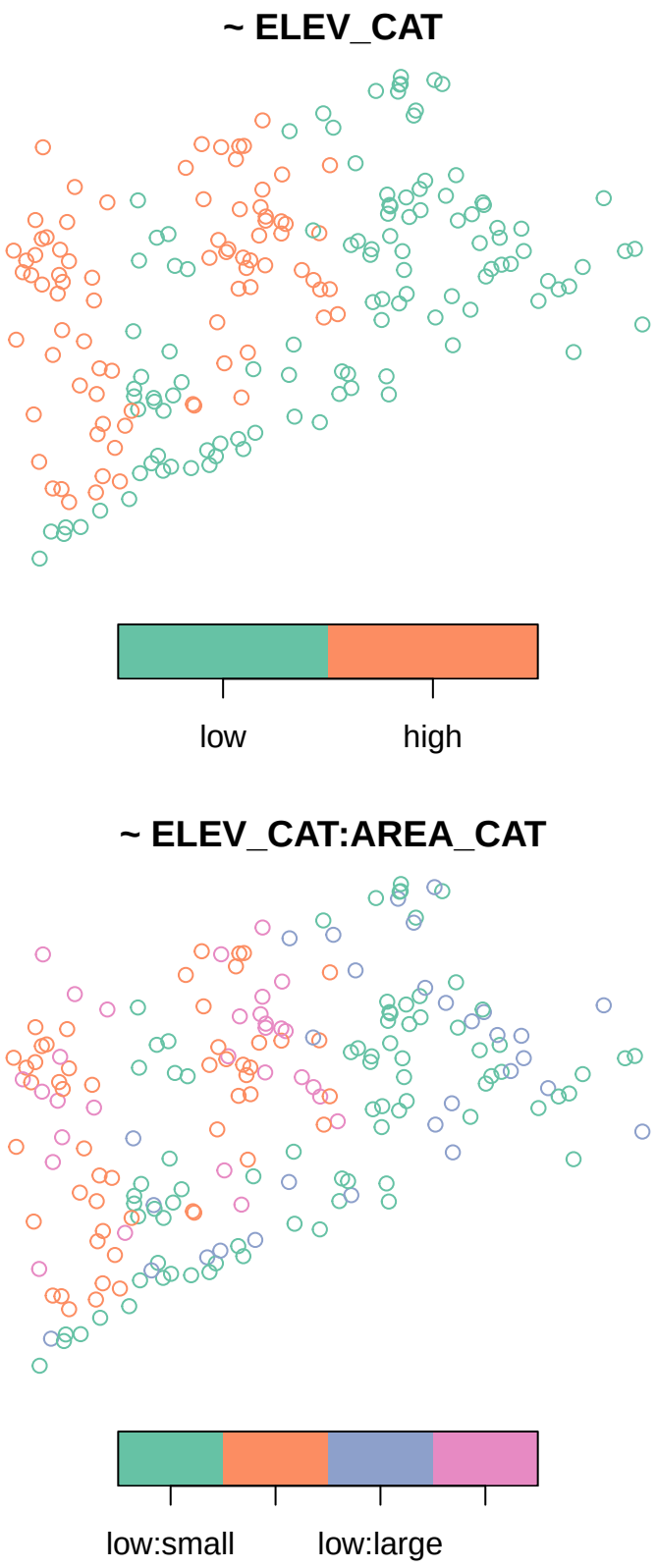


También se puede estudiar la interacción entre dos variables:

```
summary(NE_Lakes, formula = ~ ELEV_CAT + ELEV_CAT:AREA_CAT)
```

```
##    total      ELEV_CAT     ELEV_CAT:AREA_CAT
##  total:195   low :112   low:small :82
##              high: 83   high:small:53
##                         low:large :30
##                         high:large:30
```

```
plot(NE_Lakes, formula = ~ ELEV_CAT + ELEV_CAT:AREA_CAT)
```



A continuación, se estudia la implementación del método GRTS.

El paquete spsurvey implementa el algoritmo GRTS utilizando la función grts(). Hay dos argumentos obligatorios para `grts()`: el marco de muestreo y un tamaño de muestra base.

• El primer argumento requerido es el marco de muestreo, que debe ser un objeto 'sf'. Para recursos puntuales, las geometrías sf deben ser todas de tipo POINT o MULTIPOINT; para recursos lineales, las geometrías sf deben ser todas LINESTRING o MULTILINESTRING

• El segundo argumento es el tamaño de muestra deseado para la muestra base, n_base.

De forma opcional, admite otros argumentos, como legacy_sites, n_over y n_near.

• legacy_sites: a menudo se desea garantizar que algunas ubicaciones seleccionados en una muestra antigua sean seleccionadas en una nueva muestra. Foster et al. (2017) discute dos tipos de sitios que se pueden utilizar para lograr este objetivo: sitios heredados y sitios icónicos. Los sitios heredados fueron seleccionados al azar en la muestra antigua, están en el marco de muestreo actual y deben estar en la muestra actual.

• n_over, n_near: a veces se selecciona una ubicación en la muestra en la que no se pueden recopilar datos. Esto ocurre normalmente cuando el propietario de la tierra no da permiso o cuando es muy costoso, entre otras razones. Cuando esto ocurre, es útil tener un conjunto de sitios de reemplazo para que se pueda alcanzar el tamaño de muestra deseado. La función `grts()` proporciona dos opciones para los sitios de reemplazo: el orden jerárquico inverso y el vecino más cercano.

La salida de la función `grts()` es una lista con cinco componentes: sites_legacy, sites_base, sites_over, sites_near y design. Las 4 primeras componentes son objetos 'sf'.

• sites_legacy: muestra los sitios heredados seleccionados al azar en la muestra antigua que deben estar en la muestra actual.

• sites_base: sitios base (excepto aquellos que ya están incluidos en sites_legacy).

• sites_over: corresponde a los sitios de reemplazo utilizando un orden jerárquico inverso.

• sites_near: los sitios de reemplazo utilizando el vecino más cercano.

En conjunto, esta colección de objetos de sitios se llama sitios del diseño. Cada objeto de sitios contiene todas las columnas originales del marco de muestreo y algunas columnas adicionales relacionadas con el diseño de muestreo. La última componente de la salida de la función es una lista llamada diseño, que contiene detalles sobre el diseño de muestreo.

Como se ha visto ampliamente en el desarrollo teórico de la sección 2, Stevens y Olsen (2004) propusieron medir el equilibrio espacial utilizando polígonos de Voronoi (es decir, teselaciones de Dirichlet). Stevens y Olsen (2004) definen v_i como la suma de las probabilidades de inclusión de todos los sitios en el marco de muestreo contenidos en el i-ésimo polígono de Voronoi. Ellos demostraron que el valor $E(v_i) = 1$ para todos los i.

Este marco motiva el uso de métricas de pérdida basadas en polígonos de Voronoi para medir el equilibrio espacial. Una métrica de pérdida es el índice de uniformidad de Pielou (PEI; Shannon 1948; Pielou 1966), que se define como

$$PEI = 1 + \sum_{i=1}^{n} \frac{v_i}{n} ln(v_i/n)/ln(n)$$

donde n es el tamaño de la muestra. El PEI está limitado entre cero y uno. Un PEI de cero indica un equilibrio espacial perfecto. A medida que el PEI aumenta, el equilibrio espacial empeora.

La función `` `sp_balance()` `` en spsurvey mide el equilibrio espacial, y requiere tres argumentos:

- Un conjunto de sitios de diseño.
- El marco de muestreo.
- Un vector de métricas de pérdida. La métrica de pérdida por defecto es "pielou" para el PEI, aunque también hay disponibles varias otras métricas.

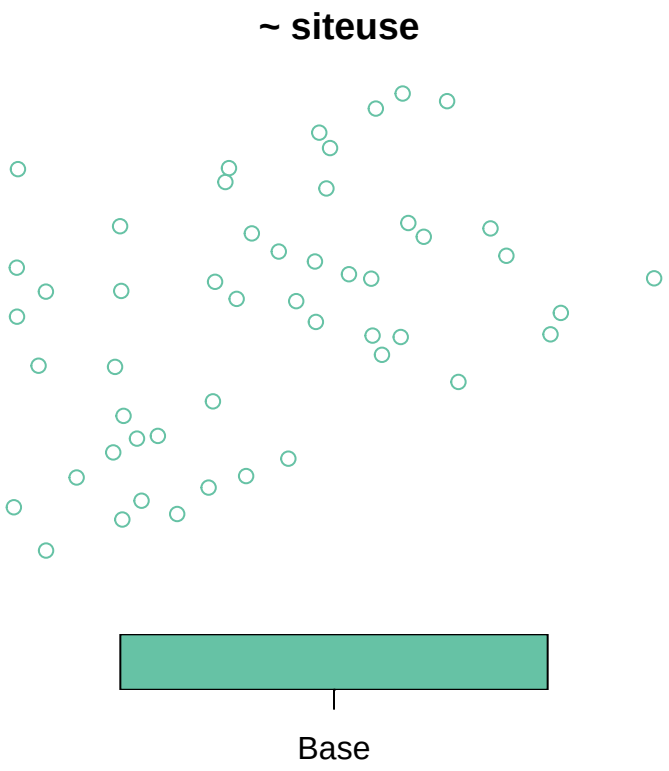
A continuación, se ilustra el funcionamiento con un ejemplo, en el que vamos a obtener una muestra de tamaño 50 a través del método GRTS con probabilidades de inclusión iguales.

En primer lugar, se proporciona únicamente los argumentos obligatorios:

```
grts <- grts(NE_Lakes, n_base = 50)
grts
```

```
## Summary of Site Counts:
##
##    total    siteuse
##  total:50   Base:50
```

```
plot(grts)
```



Veamos el índice de equilibrio espacial PEI:

```
sp_balance(grts$sites_base,NE_Lakes)$value
```

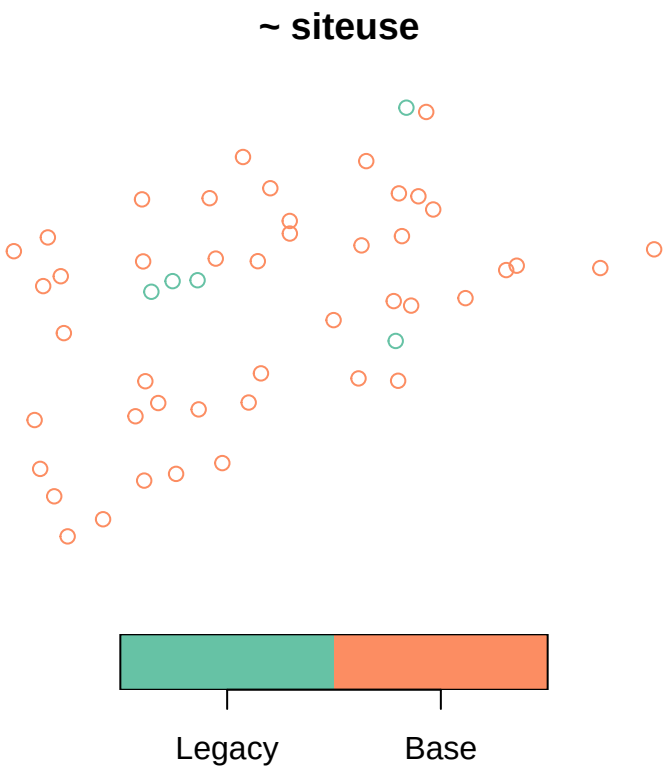
```
## [1] 0.03180812
```

Ahora, le añadimos ubicaciones de herencia, en este caso, 5 sitios que vienen almacenados en el conjunto de datos NE_Lakes_Legacy:

```
data("NE_Lakes_Legacy")
grts_legacy <- grts(NE_Lakes, n_base = 50,
                    legacy_sites = NE_Lakes_Legacy)
grts_legacy
```

```
## Summary of Site Counts:
##
##    total      siteuse
##  total:50   Legacy: 5
##             Base  :45
```

```
plot(grts_legacy)
```



Veamos el índice de equilibrio espacial PEI:

```
sp_balance(grts_legacy$sites_base,NE_Lakes)$value
```

```
## [1] 0.02294874
```

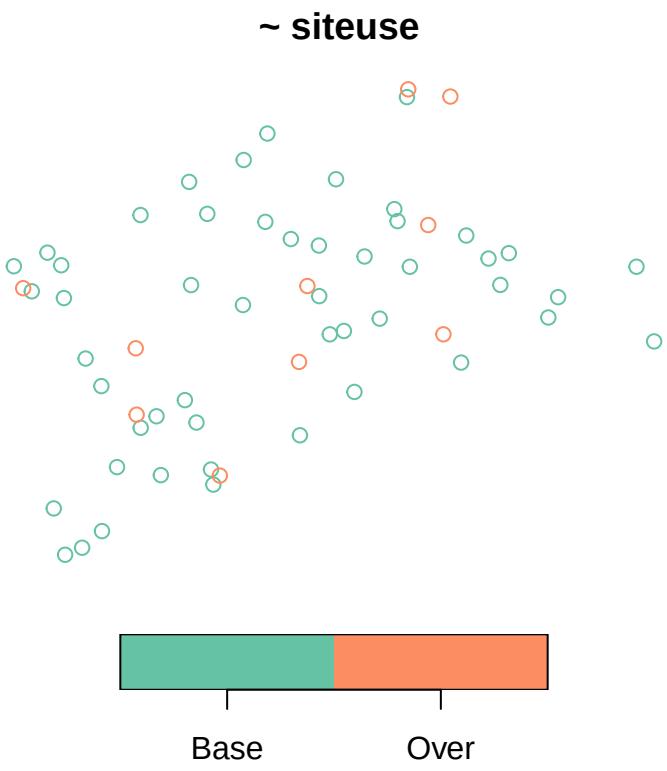
Por último, se introducen los argumentos de las ubicaciones de reemplazo:

Con n_over:

```
grts_over <- grts(NE_Lakes, n_base = 50, n_over = 10)
grts_over
```

```
## Summary of Site Counts:
##
##    total    siteuse
##  total:60   Base:50
##             Over:10
```

```
plot(grts_over)
```



Veamos el índice de equilibrio espacial PEI:

```
sp_balance(grts_over$sites_base,NE_Lakes)$value
```

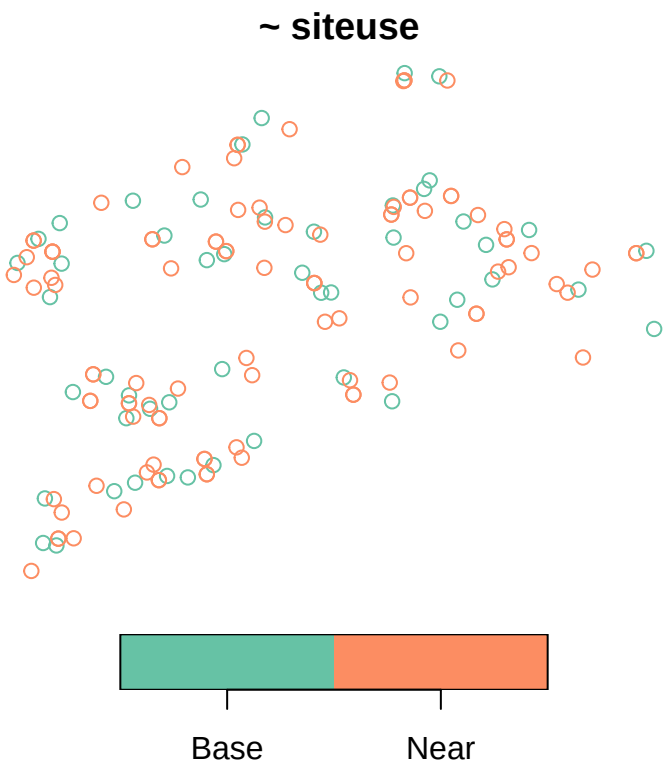
```
## [1] 0.03313944
```

Con n_near:

```
grts_near <- grts(NE_Lakes, n_base =50, n_near = 2)
grts_near
```

```
## Summary of Site Counts:
##
##    total     siteuse
##  total:150   Base: 50
##              Near:100
```

```
plot(grts_near)
```



Veamos el índice de equilibrio espacial PEI:

```
sp_balance(grts_near$sites_base,NE_Lakes)$value
```

```
## [1] 0.02604607
```

Capítulo 5

Conclusiones

La definición de diseños de muestreo espacial apropiados representa un enorme desafío para los estadísticos e investigadores que trabajan con datos geográficamente distribuidos. Las características específicas de las poblaciones georreferenciadas deben ser consideradas al diseñar una muestra (Grafström et al., 2014; Benedetti et al., 2015; Benedetti et al., 2016). Muchas poblaciones en estudios ambientales, agrícolas y forestales se distribuyenen el espacio. Es claro que las unidades espaciales no pueden ser muestreadas como si hubieran sido generadas bajo el modelo clásico de urna independiente.

Este argumento se debe principalmente a algunos de los efectos que muestran los datos espaciales, como el agrupamiento de las coordenadas, la homogeneidad, las tendencias espaciales y la homogeneidad local.

El principal desafío para los investigadores es cómo incluir estos efectos en los diseños de muestreo para reducir la varianza de los estimadores. Los métodos comúnmente utilizados de muestreo sistemático espacial y muestreo estratificado espacial sólo utilizan parcialmente estos efectos espaciales. Por estas razones, en las últimas décadas se han introducido muchos diseños de muestreo que consideran explícitamente estas características espaciales, como los que se han visto a lo largo de este trabajo.

El principal problema se refiere a la capacidad de una muestra para estar bien distribuida y aprovechar la presencia de cualquier estructura espacial que esté presente en el análisis de las poblaciones geocodificadas.

Comparamos los métodos utilizando dos poblaciones reales. Los principales resultados muestran que, si estas características espaciales de los datos existen y el método las considera, entonces puede haber una reducción notable en el error de muestreo en comparación con el muestreo aleatorio simple (como se ha visto en la sección de *Estimadores de la varianza*).

5.1. Ventajas y desventajas

5.1.1. Métodos basados en distancias

Los métodos pivotales locales y el método CP son simples y pueden ser utilizados para cualquier número de dimensiones. Las posiciones de las unidades pueden ser distintas a la

espacial. Además, los métodos pivotales locales pueden manejar fácilmente probabilidades de inclusión desiguales.

Por otra parte, hemos observado que proporcionan un alto grado de equilibrio espacial, lo que implica una pequeña varianza para el estimador HT cuando existen tendencias espaciales en $\frac{y_i}{\pi_i}$. En todos los ejemplos presentados en la comparación de los modelos, hemos encontrado que $var_{LPM} < var_{GRTS} < var_{SRS}$. Esto indica que los métodos pivotales locales son eficientes en poblaciones que tienen una clara tendencia en $\frac{y_i}{\pi_i}$. La razón principal por la que los métodos pivotales locales equilibran el tamaño de la muestra localmente es que estos métodos producen interacción solo para unidades cercanas.

Esto implica que, en cada paso, estos métodos mueven la masa de probabilidad solo una corta distancia. Debido a que LPM 1 mueve la masa de probabilidad solo entre pares en los que cada unidad es vecina más cercana de la otra, este método está ligeramente más equilibrado que LPM 2.

Una desventaja de los métodos pivotales locales es que no es posible producir un estimador de varianza sin sesgo de diseño para el estimador HT. Sin embargo, se puede utilizar el estimador de varianza media local. En nuestros ejemplos, este estimador de varianza parece funcionar tan bien para los métodos pivotes locales como para GRTS. En base a estos resultados, se recomienda utilizar el estimador de varianza media local $\hat{V}_{NBH}$

5.1.2. GRTS

El método GRTS es actualmente el método más utilizado para diseñar muestras equilibradas espacialmente, ya que tiene varias ventajas. En primer lugar, es una técnica de muestreo basada en probabilidad que garantiza un buen grado de equilibrio espacial. Además, se puede utilizar para seleccionar muestras con probabilidades de selección desiguales.

El diseño de GRTS puede abordar fácilmente problemas que ocurren en el muestreo de poblaciones, como la información de marco deficiente, la inaccesibilidad, la probabilidad variable, los patrones espaciales irregulares, los datos faltantes y las estructuras de panel.

Sin embargo, aunque no hay resultados teóricos ni suficiente evidencia empírica sobre la ganancia en eficiencia que surge del uso de GRTS al abordar poblaciones finitas, su aplicación en el muestreo continuo de superficies es muy beneficioso, ya que proporciona estimadores muy precisos y normalmente distribuidos para muestras grandes, con una tasa de convergencia de varianza de orden n^{-y}, con $1 < y < 3$ (Barabesi y Franceschi,2011; Barabesi y Marcheselli, 2008).

El método GRTS también tiene alguna desventaja. El principal problema es que el mapeo que se utiliza no es siempre eficiente, ya que las unidades que están cerca en el espacio bidimensional podrían estar siendo llevadas lejos en el espacio unidimensional.

Bibliografía

Arbia, G. & Lafratta, G. (2002). *Anisotropic spatial sampling designs for urban pollution*. J. Roy. Stat. Soc. Ser. C, 51, 223–234.

Barabesi, L., & Franceschi, S. (2011). *Sampling properties of spatial total estimators under tessellation stratified designs*. Environmetrics, 22(3), 271-278.

Barabesi, L., & Marcheselli, M. (2008). *Improved strategies for coverage estimation by using replicated line-intercept sampling*. Environmental and Ecological Statistics, 15, 215-239.

Benedetti R, Piersimoni F (2017a). *Fast Selection of Spatially Balanced Samples*. arXiv:1710.09116 [stat.ME], https://arxiv.org/abs/1710.09116.

Benedetti R, Piersimoni F, Postiglione P (2017b). *Spatially Balanced Sampling: A Review and a Reappraisal*. International Statistical Review, 85(3), 439–454. doi:10.1111/insr.12216.

Benedetti, R. & Palma, D. (1995). *Optimal sampling designs for dependent spatial units*. Envi-ronmetrics, 6, 101–114.

Benedetti, R., Piersimoni, F. & Postiglione, P. (2015). *Sampling Spatial Units for Agricultural Surveys, Advances in Spatial Science*. Series Berlin Heidelberg: Springer.

Benedetti, R., Piersimoni, F., & Postiglione, P. (2017). *Spatially balanced sampling: a review and a reappraisal*. International Statistical Review, 85(3), 439-454.

Bohorquez, M., Giraldo, R. & Mateu, J. (2016). *Optimal sampling for spatial prediction of functional data*. Stat. Methods Appl., 25, 39. doi:10.1007/s10260-015-0340-9.

Calvo, P. L. L. *Escribir un Trabajo Fin de Estudios con R Markdown*. http://destio.us.es/calvo/memoriatfe/MemoriaTFE_PedroLuque_2017Nov_librodigital.pdf.

Chambers, R.L., Steel, D.G., Wang, S. & Welsh, A. (2012). *Maximum Likelihood Estimation for Sample Surveys*. Boca Raton, USA: Chapman & Hall/CRC.

Chen, S.X. (1998). *Weighted polynomial models and weighted sampling schemes for finite population*. Annals of Statistics, 26, 1894–1915.

Chen, S.X. (2000). *General properties and estimation of conditional Bernoulli models*. Journal of Multivariate Analysis, 74, 67–87.

Chen, S.X. and Liu, J.S. (1997). *Statistical applications of the Poissonbinomial and conditional Bernoulli distributions*. Statistica Sinica, 7, 875–892.

Chen, S.X., Dempster, A.P. and Liu, J.S. (1994). *Weighted finite population sampling to maximize entropy*. Biometrika, 81, 457–469.

Christman, M.C. (2000). *A review of quadrat-based sampling of rare, geographically clustered populations*. J. Agric. Biol. Environ. Stat., 5, 168–201.

Cormen TH, Leiserson CE, Rivest RL, Stein C (2009). *Introduction to Algorithms*. MIT Press, Cambridge.

Deville, J. C., & Särndal, C. E. (1992). *Calibration estimators in survey sampling*. Journal of the American statistical Association, 87(418), 376-382.

Deville, J. C., & Tille, Y. (1998). *Unequal probability sampling without replacement through a splitting method*. Biometrika, 85(1), 89-101.

Deville, J. C., & Tillé, Y. (2000). *Selection of several unequal probability samples from the same population*. Journal of Statistical Planning and Inference, 86(1), 215-227.

Deville, J. C. (2000). *Note sur l'algorithme de Chen, Dempster et Liu*. Tech. rept. CREST-ENSAI, Rennes.

Dickson, M.M. & Tillé, Y. (2016). *Ordered spatial sampling by means of the traveling salesman problem*. Comput.Stat., 31, 1359–1372.

Dumelle, M., Kincaid, T., Olsen, A. R., & Weber, M. (2023). *spsurvey: Spatial Sampling Design and Analysis in R*. Journal of Statistical Software, 105, 1-29.

Grafström, A. (2010). *Entropy of unequal probability sampling designs*. Statistical Methodology, 7(2), 84-97.

Grafström, A., & Lundström, N. L. (2013). *Why well spread probability samples are balanced*. Open Journal of Statistics, 3(1), 36-41.

Grafström, A., Lundström, N. L., & Schelin, L. (2012). *Spatially balanced sampling through the pivotal method*. Biometrics, 68(2), 514-520.

Haining, R.P. (2003). *Spatial Data Analysis: Theory and Practice*. Cambridge: Cambridge University Press.

Hájek, J. (1981). *Sampling from a Finite Population*. New York: Marcel Dekker.

Hedayat, A. & Stufken, J. (1998). *Sampling designs to control selection probabilities of contiguous units*. J. Statist. Plann. Inference, 72, 333–345.

Hedayat, A., Rao, C.R. & Stufken, J. (1988b). *Designs for survey sampling avoiding contiguous units*. In Handbook of Statistics, Vol. 6: Sampling, Eds. P.R. Krishnaiah & C.R. Rao, pp. 575–583. The Netherlands: Elsevier.

Hedayat, A.S., Bing-Ying, L. and Stufken, J. (1989). *The construction of* ΠPS *sampling designs through a method of emptying boxes*. Annals of Statistics, 4, 1886–1905.

Horvitz, D. G. and Thompson, D. J. (1952). *A generalization of sampling without replacement from a finite universe*. Journal of the American Statistical Association 47, 663–685. Journal of Theoretical Biology, 13, 131–144. doi:10.1016/0022-5193(66)90013-0.

Mandal, B., Parsad, R. & Gupta, V. (2008). ΠPS *sampling plans excluding adjacent units*. Commun. Stat. – Theory, 3, 2532–2550.

Olea, R. A. (1984). *Sampling design optimization for spatial functions*. Journal of the interna-tional Association for Mathematical Geology, 16, 369-392.

Overton, W.S. & Stehman, S.V. (1993). *Properties of designs for sampling continuous spatial resources from a triangular grid*. Commun. Stat. Theory, 22, 2641–2660.

Pantalone, F., Benedetti, R., & Piersimoni, F. (2022). *Reference Manual for the R-Package Spbsampling*. Retrieved October 12, 2022, from https://cran.r-project.org/web/packages/Spbsampling/Spbsampling.pdf.

Pantalone, F., Benedetti, R., & Piersimoni, F. (2022). *Spbsampling: An R Package for Spatially Balanced Sampling*. Journal of Statistical Software, 103, 1-22.

Pielou EC (1966). *The Measurement of Diversity in Different Types of Biological Collections*.

Rogerson, P. & Delmelle, E. (2004). *Optimal sampling design for variables with varying spatial importance*. Geog. Anal., 36, 177–194.

Rosén, B. (1997a). *Asymptotic theory for order sampling*. Journal of Statistical Planning and Inference 62, 135–158.

Rosén, B. (1997b). *On sampling with probability proportional to size*. Journal of Statistical Planning and Inference 62, 159–191.

Scott Overton, W., & Stehman, S. V. (1993). *Properties of designs for sampling continuous spatial resources from a triangular grid*. Communications in Statistics–Theory and Met-hods, 22(9), 251-264.

Shannon CE (1948). *A Mathematical Theory of Communication*. The Bell System Technical Journal, 27(3), 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x.

Stevens Jr, D. L., & Olsen, A. R. (2003). *Variance estimation for spatially balanced samples of environmental resources*. Environmetrics, 14(6), 593-610.

Stevens Jr, D. L., & Olsen, A. R. (2004). *Spatially balanced sampling of natural resources*. Journal of the American statistical Association, 99(465), 262-278.

Stevens, D.L. Jr. (1997). *Variable density grid-based sampling designs for continuous spatial population*. Environmetrics, 8, 167–195.

Stufken, J. (1993). *Combinatorial and statistical aspects of sampling plans to avoid the selection of adjacent units*. J. Combin. Inform. System Sci., 18, 81–92.

Stufken, J., Song, S.Y., See, K. & Driessel, K.R. (1999). *Polygonal designs: some existence and non-existence results*. J. Statist. Plann. Inference, 77, 155–166.

Thompson, S.K. (2013). *Sampling,* 3^{rd} *edition*. Hoboken, New Jersey: John Wiley and Sons Inc.

Tillé, Y. (2006). *Sampling algorithms (pp. 31-39)*. Springer New York.

Vallée, AA, Ferland-Raymond, B, Rivest, LP & Tillé, Y. (2015). *Incorporating spatial and operational constraints in the sampling designs for forest inventories*. Environmetrics, 26, 557–570. doi: 10.1002/env.2366.

Wang, J.F., Stein, A., Gao, B.B. & Ge, Y. (2012). *A review of spatial sampling*. Spat. Stat., 2, 1-14.

Wright, J. & Stufken, J. (2008). *New balanced sampling plans excluding adjacent units*. J. Statist. Plann. Inference,138, 3326–3335.