UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

Métodos estadísticos para la estimación de la abundancia

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

Actuario

PRESENTA :

Carlos René Flores Mendive

TUTORA:

Mat. Margarita Elvira Chávez Cano

Ciudad Universitaria, Cd. Mx., 2019

Universidad Nacional Autónoma de México
UNAM
Biblioteca Central
Dirección General de Bibliotecas de la UNAM

UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Índice general

1. Muestreo por parcelas **6**

1.1. Introducción 6

1.2. Estimación por máxima verosimilitud 7

1.3. Intervalos de confianza 9

1.4. Consecuencias cuando se violan los supuestos 16

2. Muestreo por distancias **17**

2.1. Introducción 17

2.1.1. Función de detección 19

2.1.2. Truncamiento de datos 20

2.1.3. Poblaciones agrupadas 21

2.1.4. Constantes conocidas y parámetros 23

2.1.5. Supuestos 23

2.2. Transectos lineales 25

2.2.1. Introducción 25

2.2.2. Estimación por máxima verosimilitud 26

2.2.3. Estimación de la función de detección 30

2.2.4. Elección del modelo 37

2.2.5. Estimación por intervalos 41

2.3. Transectos puntuales 45

2.3.1. Introducción 45

2.3.2. Estimación por máxima verosimilitud 46

2.3.3. Estimación la función de detección 48

2.3.4. Elección del modelo 53

2.3.5. Estimación por intervalos 55

3. Vecino más cercano **57**
3.1. Introducción 57
3.2. Estimación por máxima verosimilitud 58
3.3. Estimación por intervalos 61

4. Aplicaciones **64**
4.1. Introducción 64
4.2. Primera aplicación 65
4.2.1. Muestreo por parcelas 65
4.2.2. Muestreo por distancias 68
4.2.2.1. Muestreo por transectos lineales 68
4.2.2.2. Muestreo por transectos puntuales 72
4.2.3. Vecino más cercano 75
4.2.4. Conclusiones 77
4.3. Segunda aplicación 79
4.3.1. Muestreo por parcelas 80
4.3.2. Muestreo por distancias 82
4.3.2.1. Muestreo por transectos lineales 82
4.3.2.2. Muestreo por transectos puntuales 85
4.3.3. Vecino más cercano 89
4.3.4. Conclusiones 91

A. Códigos en R **93**
A.1. Primera aplicación 93
A.2. Segunda aplicación 101

Introducción

La Ecología es el estudio de las interacciones entre los organismos y su entorno. Los diferentes organismos crecen, se multiplican, ocupan diferentes regiones, compiten con otras especies por recursos, etc. Muchos cuestionamientos interesantes se pueden realizar acerca de este amplio conjunto de objetos que hallamos en la naturaleza y al intentar encontrar respuesta a ellos nos encontraremos constantemente proponiendo modelos de la forma en que el mundo funciona basados en nuestras observaciones, intuición, y el conocimiento con el que contamos. Estos modelos pueden ser gráficos, verbales o representaciones matemáticas de un proceso.

Los modelos matemáticos son de gran utilidad en cualquier área de investigación y en particular en la Ecología, ya que son proposiciones formales que enfoncan nuestro razonamiento y nos obligan a ser explícitos con los supuestos que utilizamos y la manera en que visualizamos la relación que presentan las variables de nuestro objeto de estudio. Cuando modelamos un objeto ecológico como una población o un ecosistema, tenemos siempre que comenzar con una analogía entre este objeto y un objeto matemático, y es al hacer esta abstracción en donde usando el rigor matemático podemos trabajar a través de nuestro razonamiento lógico y obtener una idea del funcionamiento de los distintos objetos ecológicos.

Por lo anterior, al modelar matemáticamente los diferentes elementos ecológicos no tratamos directamente con objetos naturales, sino que se trabaja con objetos matemáticos como variables, operaciones o ecuaciones como análogos de elementos de la naturaleza, lo que resulta en que al construir un modelo matemático, este no contendrá toda la información que podamos conocer acerca de la naturaleza, sino solo la pertinente para el problema de nuestro interés. En esta modelación matemática, hemos abstraido la naturaleza en una estructura más simple de forma que podamos entenderla, para lo cual podremos usar la rama de las matemáticas que más convenga en un estudio en particular. En este

trabajo nos concentraremos en la modelación desde el punto de vista estadístico.

Estimacíon de la abundancia

De entre los interminables aspectos que pudieran estudiarse de una población biológica, uno que es de gran interés en la mayoría de estudios, y el cual será el objeto principal de este texto, es el de estimar la densidad de población de una comunidad o el tamaño absoluto de ésta. Un ingeniero forestal quiere hacer un inventario de la flora en un área. Un biólogo conservacionista necesita ver el efecto de un programa de protección a una especie en peligro de extinción. Un manejador de vida silvestre desea observar si existe un balance razonable entre animales carnívoros y herbívoros. En cualquier caso optar por un enfoque estadístico para obtener una estimación de la abundancia será de gran beneficio ya que nos permitirá obtener estimadores confiables sin la necesidad de hacer un censo completo.

La técnica con que afrontemos este problema tendrá que ver directamente con las características del organismo en cuestión. Objetos estáticos como árboles, nidos, hormigueros, pueden ser cuantificados de una manera que no sea adecuada para organismos con más movilidad como aves, peces, insectos, mamíferos.

Como un primer acercamiento se presentará en el capítulo 1 el muestreo por parcelas, que si bien tiene un rango de aplicación limitada por contar con condiciones restrictivas, funcionará como base para los métodos explicados en los capítulos siguientes.

Una metodología menos restrictiva es el muestreo por distancias, mismo que se estudiará con más detalle en el capítulo 2 y que recibe su nombre del hecho de que la información utilizada es la de distancias obtenidas a los objetos de interés, medidas a partir de un punto o línea (transecto) de observación. Esta metodología incorpora la estimación de la probabilidad de detección en la cual se basa la mayor parte de la modelación y es lo que hace del muestreo por distancias un método eficiente y aplicable a un gran número de situaciones.

En el capítulo 3 se mostrará el método del vecino más cercano, en el cual se modelará a partir de la variable aleatoria representada por la distancia de un punto seleccionado aleatoriamente sobre el área de estudio al individuo más cercano.

Finalmente en el capítulo 4 se mostrarán aplicaciones prácticas a poblaciones simuladas lo que permitirá hacer comparaciones entre cada uno de los métodos

y analizar en qué situaciones es conveniente usar cada uno.

Capítulo 1

Muestreo por parcelas

1.1. Introducción

El muestreo por parcelas es un método para la estimación de la abundancia, denotada por N, en el cual se seleccionan aleatoriamente áreas ('parcelas') dentro de una región de estudio (con área A) y se cuentan **totalmente** los individuos dentro de las parcelas. Al total del área de las parcelas se le llamará **región cubierta**. La otra característica esencial de este método es la de suponer que los individuos de la población de interés se distribuyen uniformemente e independientemente sobre la región de estudio (figura 1.1).

Este enfoque puede resultar adecuado para estimar la densidad de plantas pero puede presentar problemas para animales que se encuentran en constante movimiento. Para un muestreo eficiente de poblaciones animales, las parcelas a utilizar deben ser generalmente de gran tamaño, y en este caso podría ser complicado asegurar que todos los animales dentro de la parcela sean contados. La dificultad recae también en que los animales podrían entrar y salir de la parcela durante el conteo.

Si bien el muestreo por parcelas es raramente utilizado debido a que generalmente los supuestos necesarios no son compatibles con poblaciones reales, estos métodos proveen una base para métodos más complejos y ampliamente utilizados de los cuales se hablará más adelante.

El muestreo por parcelas suele ser nombrado de diferentes maneras dependiendo la forma de las parcelas, las posibles formas de las parcelas no tienen importancia alguna si es posible medir su área.

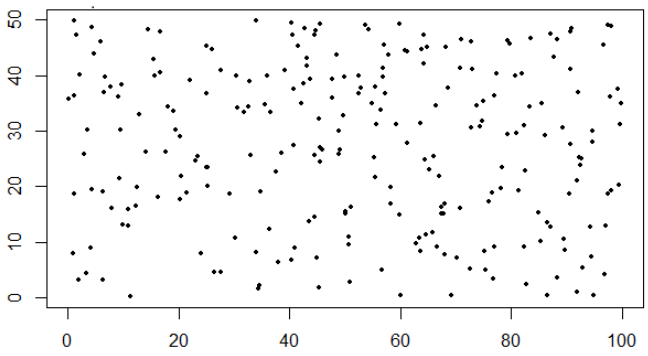

Figura 1.1: Población simulada con un total de $N = 250$ individuos, en un área total de tamaño $A = 5000Km^2$. La distribución de los individuos es uniforme en la región de estudio.

1.2. Estimación por máxima verosimilitud

Al suponer que es igualmente probable que un individuo se encuentre en cualquier parte de la región de estudio, se sigue que la probabilidad de que un individuo esté en la región cubierta con área a es $\pi_c = a/A$. Si además las localizaciones de los individuos son independientes y definimos el evento en que un individuo se encuentre en el área cubierta como un 'éxito', entonces n, el número de individuos contados en la región cubierta, es una variable aleatoria binomial con parámetros N y π_c, y función de verosimilitud:

$$L(N) = \binom{N}{n} \pi_c^n \left(1 - \pi_c\right)^{N-n} \tag{1.2.1}$$

Podemos maximizar esta función de verosimilitud al encontrar valores de N para los cuales la función de verosimilitud es creciente con respecto a N, i.e.:

$$\frac{L(N)}{L(N-1)} > 1$$

$$\Rightarrow \frac{\frac{N!}{(N-n)!n!}\pi_c^n(1-\pi_c)^{N-n}}{\frac{N-1!}{(N-1-n)!n!}\pi_c^n(1-\pi_c)^{N-1-n}} > 1$$

$$\Rightarrow \frac{N}{N-n}(1-\pi_c) > 1$$

$$\Rightarrow 1 - \pi_c > 1 - \frac{N}{n}$$

$$\Rightarrow N < \frac{n}{\pi_c}$$

Por lo tanto el máximo de la función de verosimilitud se alcanza en el mayor$\hat{N}$ tal que $\hat{N} < \frac{n}{\pi_c}$. Al sólo permitir soluciones enteras en la práctica, entonces $\hat{N}$ será la parte entera de $\frac{n}{\pi_c}$ i.e., $\hat{N} = \left[\frac{n}{\pi_c}\right]$.

Consideremos un ejemplo de una población simulada con N =250 individuos mostrada en la figura 1.2. La región de estudio tendrá un área total A=5000km^2 y cada una de las 20 parcelas elegidas aleatoriamente (en este caso rectangulares) tendrá un área $a_i = 50Km^2$ para $i = 1, \ldots, 20$, dentro de los cuales se contaron un total de $n = 46$ individuos. El área total de la región cubierta es $a = \sum_{i=1}^{20} a_i = 1000\, Km^2$, entonces el estimador $\hat{N}$ para el tamaño de población real N es:

$$\hat{N} = \frac{n}{\pi_c} = n\,\frac{A}{a} = 46\frac{5000\, Km^2}{1000\, Km^2} = 230\ individuos$$

lo cual resulta relativamente similar a $N = 250$ que es el tamaño real de población.

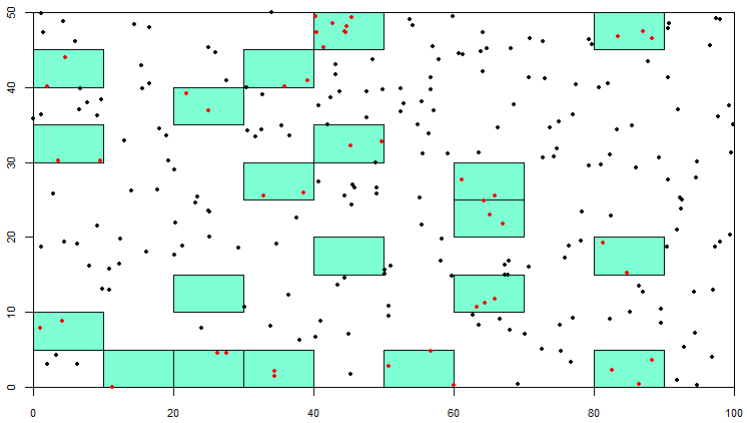

Figura 1.2: Población simulada con 250 individuos. Los puntos dentro de cada rectángulo representan los individuos observados (46). En este caso el área cubierta a representa el 20 % de la región de estudio total.

Frecuentemente al estimar la abundancia podemos centrar nuestra atención en la densidad D, ya que ésta y la abundancia están relacionadas por N=$D \times$

A. Por lo tanto podemos también enfocarnos en obtener un estimador de la densidad. Usando los resultados anteriores tenemos:

$$\begin{aligned} \hat{N} &= \hat{D} \cdot A \\ \Rightarrow \hat{D} &= \hat{N}/A \\ \Rightarrow \hat{D} &= \frac{n}{\pi_c} \cdot \frac{1}{A} = \frac{nA}{a} \cdot \frac{1}{A} = \frac{n}{a} \end{aligned}$$

En el ejemplo anterior el estimador máximo verosímil para la densidad es $\hat{D}$=0.046 $\frac{Individuos}{Km^2}$.

1.3. Intervalos de confianza

Intervalos de confianza exactos

Bajo los supuestos acerca de la manera de realizar el muestreo por parcelas, tenemos que la probabilidad de contar exactamente n individuos está dada por una variable aleatoria binomial con parámetros N y π_c, (i.e $n \sim Bin(N, \pi_c)$). Por lo que para cualquier N podemos encontrar la probabilidad de que n sea menor o igual que el número de individuos contados, en el caso del ejemplo, 46. De igual forma podríamos encontrar la probabilidad de que n sea mayor o igual que 46.

Siguiendo con la población ejemplo, si la abundancia real fuera N=179 tendríamos que $\mathbb{P}[n \geq 46] = 0.025$, si la abundancia real fuera N=299 entonces $\mathbb{P}[n \leq 46] = 0.025$. Equivalentemente si N=179 entonces $\mathbb{P}[\hat{N} \geq 230] = 0.025$ y si N=299 entonces $\mathbb{P}[\hat{N} \leq 230] = 0.025$. De donde se tiene que dado que contamos 46 individuos el intervalo del 95 % de confianza es (179, 299).

Pensando este procedimiento como una prueba de hipótesis, para cualquier $N < 179$ o para $N > 299$, nuestra observación de $n = 46$ individuos estaría en la región de rechazo con un nivel de significancia del 5 % de la prueba cuya hipótesis nula es N=230 (el valor estimado en el ejemplo), contra la hipótesis alternativa de dos colas en que $N \neq 230$.

Supongamos entonces que se observan $n = n_0$ individuos en la región cubierta, entonces para obtener el intervalo del (1-α) $\times$ 100 % de confianza para N, debemos encontrar N_1 y N_2 tales que si $n \sim Bin(N_1, \pi_c)$ entonces $\mathbb{P}[n \geq n_0] = \alpha/2$, y si $n \sim Bin(N_2, \pi_c)$ entonces $\mathbb{P}[n \leq n_0] = \alpha/2$, de tal manera obtenemos el intervalo (N_1, N_2).

Normalidad asintótica

Una manera sencilla de obtener un intervalo de confianza para la abundancia N, es aproximando la distribución del estimador máximo verosímil $\hat{N}$ con una distribución normal.

En la sección 1.2 mencionamos que la distribución del número de individuos observados n, sigue una distribución binomial con parámetros N y $p = \pi_c$, (i.e. $n \sim Bin(N, p)$) por lo que se tiene lo siguiente:

$$\mathbb{E}\left[\hat{N}\right] = \mathbb{E}\left[\frac{n}{p}\right] = \frac{1}{p}\mathbb{E}\left[n\right] = \frac{Np}{p} = N \tag{1.3.1}$$

y

$$\widehat{Var}\left[\hat{N}\right] = \frac{\hat{N}p(1-p)}{p^2} \tag{1.3.2}$$

Las ecuaciones (1.3.1) y (1.3.2) y la normalidad asintótica de los estimadores máximo verosímiles proporcionan la justificación teórica para suponer la normalidad de $\hat{N}$, con lo cual al construir el intervalo de confianza con un nivel de confianza $(1-\alpha)\times 100\,\%$ para la abundancia N obtenemos el intervalo:

$$\left(\hat{N} - Z_{1-\frac{\alpha}{2}}\sqrt{\widehat{Var}\left[\hat{N}\right]},\, \hat{N} + Z_{1-\frac{\alpha}{2}}\sqrt{\widehat{Var}\left[\hat{N}\right]}\right) =$$
$$\left(\hat{N} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{N}p(1-p)}{p^2}}\,,\, \hat{N} + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{N}p(1-p)}{p^2}}\right) \tag{1.3.3}$$

donde $Z_{1-\frac{\alpha}{2}}$ es el cuantil $1-\frac{\alpha}{2}$ de una la distribución normal estándar.

Utilizando los datos de la población ejemplo de la figura 1.2, obtuvimos $\hat{N}$=230, $p = 0.2$ y calculamos $\widehat{Var}\left[\hat{N}\right]$ =920, de donde si α=0.05 se sigue de (1.3.3) que el intervalo de confianza al 95 % para N , suponiendo la normalidad de $\hat{N}$, es $(170, 290)$ que en este caso contiene al verdadero valor de $N = 250$ y es similar al intervalo de confianza exacto obtenido en la sección anterior de (179,299).

De lo anterior podemos ver que la única información que usamos para construir este intervalo de confianza es el estimador puntual $\hat{N}$ y la estimación de su varianza, por lo que al utilizar la normalidad asintótica no tomamos en cuenta información acerca de la verdadera forma de la distribución de n (el número de individuos observados), lo que ocasiona que al tener un tamaño de muestra pequeño, este intervalo aproximado podría proporcionar una estimación muy

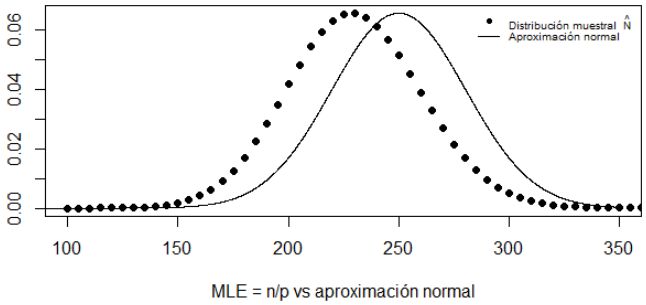

Figura 1.3: Distribución del EMV cuando $n \sim Bin(230, 0.2)$ (línea punteada) y densidad normal aproximada con media igual a la abundancia real $N = 250$, y con varianza igual a la varianza estimada de la ecuación (1.3.2) , en este caso como $\hat{N}$=230 la varianza estimada es 920 (línea continua).

pobre. Las figuras 1.3, 1.4 y 1.5 muestran las posibles diferencias entre la distribución de n y la aproximación normal, utilizando los datos del ejemplo de la sección 1.2.

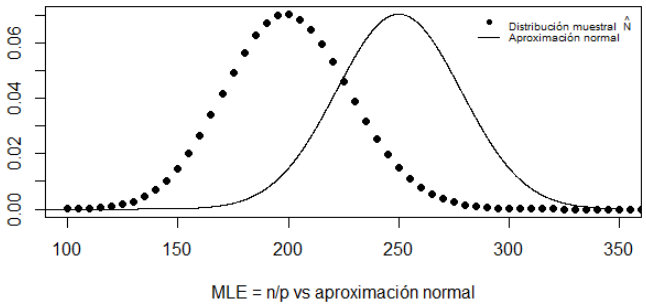

Figura 1.4: Distribución del EMV cuando $n \sim Bin(200, 0.2)$ (línea punteada) y densidad normal aproximada con media igual a la abundancia real $N = 250$, y en este caso como $\hat{N}$=200 la varianza estimada es 800 (línea continua).

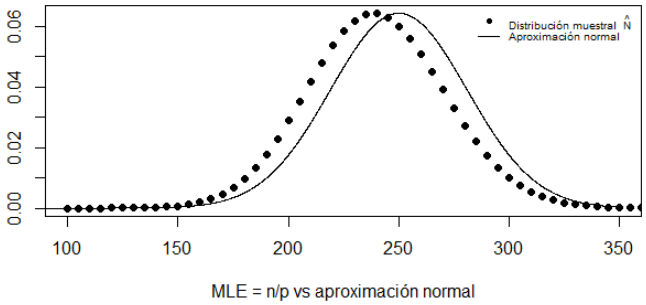

Figura 1.5: Distribución del EMV cuando $n \sim Bin(240, 0.2)$ (línea punteada) y densidad normal aproximada con media igual a la abundancia real $N = 250$, y en este caso como $\hat{N}$=240 la varianza estimada es 38.4 (línea continua).

Podemos observar que a medida que el EMV $\hat{N}$ es más parecido a la abundancia real N, esta aproximación será mejor y los intervalos serán más cercanos a los exactos.

A continuación veremos otros métodos más robustos para la estimación por intervalos.

Bootstrap

El bootstrap es un método de simulación que se utiliza para aproximar la distribución muestral de un estimador y a partir de esta distribución poder estimar diferentes medidas de variación o en nuestro caso, intervalos de confianza. Una de la principales ventajas del bootstrap es la simplicidad, es un método muy directo para obtener estimadores de desviaciones estándar e intervalos de confianza, y a pesar de que en muchos casos es prácticamente imposible econtrar los intervalos de confianza exactos, los intervalos obtenidos mediante bootstrap resultan asintóticamente más precisos que los obtenidos suponiendo normalidad asintótica.

En nuestro contexto, el bootstrap consistirá en construir la distribución muestral del estimador $\hat{N}$ realizando un gran número de simulaciones por computadora y contando la proporción de veces que obtenemos cada valor de $\hat{N}$. Existen varias maneras de realizar esto pero nos enfocaremos en dos métodos:

- El **bootstrap paramétrico** en donde supondremos un modelo estadístico para nuestro estimador de interés $\hat{N}$ y así simularemos la distribución propuesta, la cual tendrá un parámetro desconocido que estimaremos a partir de la muestra.

- El **bootstrap no-parámetrico** en el cual se utilizará muestreo con reemplazo para obtener la distribución de $\hat{N}$ sin suponer ninguna distribución.

Bootstrap paramétrico

Para este método, obtendremos un gran número de observaciones simuladas usando el modelo estadístico que elegimos para estimar N, en nuestro caso generaremos $n_1, \ldots, n_j$ observaciones de una variable aleatoria binomial con parámetros $\hat{N}$ y π_c, continuando con los datos del ejemplo de las secciones anteriores tomaremos $\hat{N} = 230$ y π_c=0.2, una vez obtenidas las simulaciones solo resta dividir cada una entre π_c y así obtendremos la muestra $\hat{N}_1 = \frac{n_1}{\pi_c}, \ldots, \hat{N}_j = \frac{n_j}{\pi_c}$.

Una vez obtenida la muestra utilizando bootstrap, existen distintos métodos para obtener los intervalos de confianza, en este texto usaremos el 'método de los percentiles', el cual es un método simple y a diferencia de otros métodos, este no requiere la simetría alrededor del valor estimado para N. Para construir el intervalo del (1-α)×100 % de confianza para N por el método de percentiles solo basta encontrar los percentiles del $\frac{\alpha}{2} \times 100\,\%$ y $(1 - \frac{\alpha}{2}) \times 100\,\%$ de la muestra obtenida por bootstrap. Tomando los datos de nuestro ejemplo y α=0.05, el intervalo del (1-α)×100 % de confianza obtenido por bootstrap con 100 simulaciones es (177, 282), con 300 simulaciones obtenemos el intervalo (170, 285) y finalmente con 10,000 simulaciones obtenemos el intervalo (170, 290) que en este ejemplo coincide con el obtenido al suponer normalidad asintótica; si bien en este caso el intervalo obtenido por bootstrap no resulta mejor que el de la aproximación normal, el método de percentiles para obtener intervalos de confianza a partir de la muestra bootstrap es generalmente más robusto en casos donde el tamaño de muestra es pequeño o la distribución es sesgada.

Notemos también que a partir de la muestra bootstrap podemos calcular fácilmente un estimador para la varianza de $\hat{N}$ simplemente calculando la varianza de la muestra $\hat{N}_1, \ldots, \hat{N}_j$.

En las figuras 1.6, 1.7 y 1.8 se ilustra la construcción de la distribución $\hat{N} = \frac{n}{p}$, al generar observaciones simuladas para n.

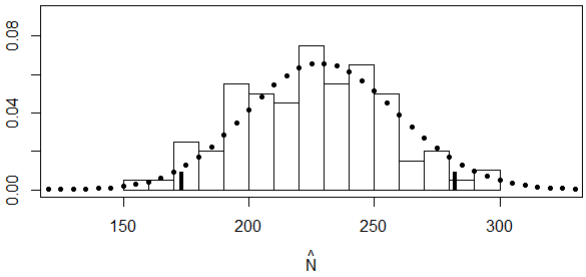

Figura 1.6: Histograma de la muestra bootstrap de $\hat{N} = \frac{n}{p}$ cuando $n \sim Bin(230, 0.2)$, con 100 observaciones simuladas. La línea punteada es la función de masa de probabilidad de una variable aleatoria $Bin(230, 0.2)$. Las líneas cortas verticales representan los límites del intervalo de confianza estimado usando el método de percentiles.

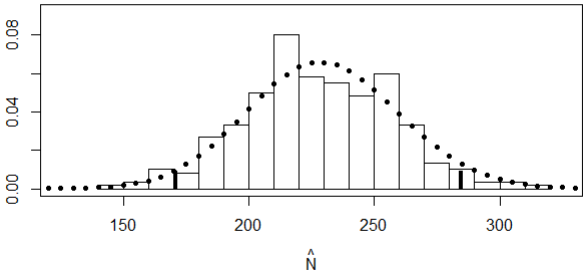

Figura 1.7: Igual que la figura 1.6 pero con 300 observaciones simuladas.

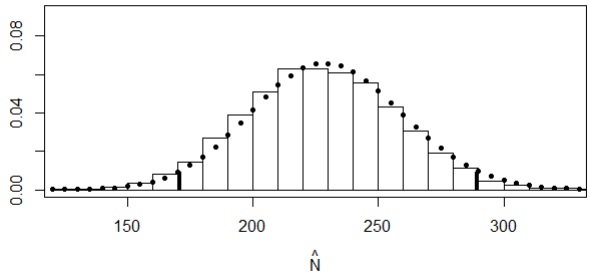

Figura 1.8: Igual que la figura 1.7 pero con 10,000 observaciones simuladas.

Bootstrap no-paramétrico

El bootstrap no-parámetrico funciona de manera similar al bootstrap paramétrico con la diferencia que al simular observaciones de $\hat{N}$ no haremos supuestos distribucionales, en cambio realizaremos muestreos con reemplazo a partir de la muestra obtenida. Continuando con el ejemplo, para el bootstrap no-paramétrico cada observación de la variable aleatoria $\hat{N}$ la generaremos al obtener una muestra con reemplazo de tamaño igual al valor estimado para N (en este caso de tamaño 230), de una población con $n = 46$ individuos observados y $\hat{N} - n = 230 - 46 = 184$ individuos no observados.

Veremos cómo funciona lo anterior. Generaremos primero una población de tamaño 230 formada por 46 unos y 184 ceros, donde los unos serán individuos observados y los ceros serán individuos no observados, obtendremos después una muestra con reemplazo de esta población con tamaño 230, cada una de esas muestras con reemplazo serán las muestras bootstrap para las cuales sumaremos los unos obtenidos y dividiremos el resultado entre π_c=0.2. Repitiendo este procedimiento j veces obtendremos la muestra $\hat{N}_1 = \frac{n_1}{\pi_c}, \ldots, \hat{N}_j = \frac{n_j}{\pi_c}$ donde cada n_i es la suma de unos obtenidos en cada muestra bootstrap.

En el caso particular de muestreo por parcelas podemos observar que el bootstrap paramétrico y el no-paramétrico son equivalentes ya que al realizar el procedimiento de bootstrap no-paramétrico, el número de unos que hay en cada muestra bootstrap es la suma de 230 variables aleatorias Bernoulli con parámetro $\frac{n}{\hat{N}} = \frac{n}{\frac{n}{\pi_c}} = \pi_c = 0.2$, es decir son observaciones de una Binomial con parámetros 230 y 0.2 como en el caso paramétrico. En general los dos métodos son distintos..

1.4. Consecuencias cuando se violan los supuestos

- **Todos los individuos en la región cubierta son detectados**

  El efecto de no cumplir este supuesto es que el estimador para la abundancia estará negativamente sesgado por un factor igual a la probabilidad de detectar a un individuo que claramente será menor que 1 si se viola este supuesto.

- **Los individuos se distribuyen uniformemente e independientemente en el área de estudio**

  Si no se cumple el supuesto de uniformidad, los intervalos de confianza basados en este supuesto, serán muy angostos si los individuos están en grupos, o muy anchos si son más escasos de lo que permite la distribución uniforme. Si la localización de individuos no es independiente, los intervalos de confianza serán muy angostos si existe una correlación positiva entre sus localizaciones, o muy ancho si la correlación es negativa.

Capítulo 2

Muestreo por distancias

2.1. Introducción

El muestreo por distancias es una metodología ampliamente utilizada para estimar la abundancia de poblaciones biológicas ya que a diferencia del muestreo por parcelas, este método incorpora la probabilidad de detección de un individuo lo cual permite que podamos obtener una buena estimación aún cuando no se detecten algunos individuos. El nombre de este método se deriva del hecho de que la información utilizada para hacer inferencia sobre la abundancia en un área, son las distancias a los individuos detectados medidas a partir de una línea o punto al que llamaremos **transecto,** el cual es **seleccionado aleatoriamente dentro del área de estudio.** En el caso de realizar el estudio sobre una línea (figura 2.2) se recopilarán las distancias perpendiculares al individuo (o en su defecto la distancia radial y el ángulo de observación), y en el caso de recopilar la información a partir de un punto fijo se guardarán las distancias radiales del transecto al individuo (figura 2.1).

Los dos métodos más comunes dentro del muestreo por distancias son el muestreo por transectos lineales y el muestreo por transectos puntuales, los cuales podemos pensarlos como una extensión del muestreo por parcelas con una importante diferencia: en los métodos de muestreo por distancias **solo detectamos una proporción de los individuos en la región cubierta**, mientras que en el muestreo por parcelas debemos contarlos en su totalidad.

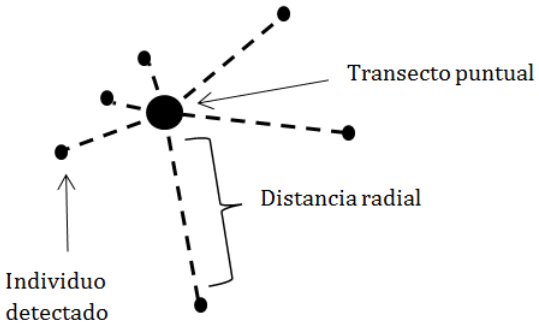

Figura 2.1: Diagrama de muestreo con un transecto puntual y 6 individuos detectados con sus respectivas distancias radiales.

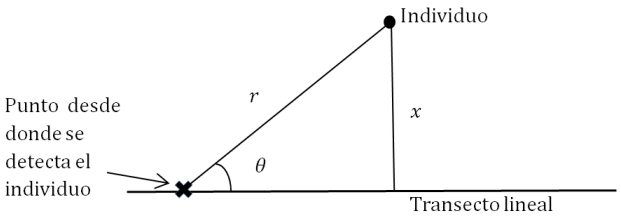

Figura 2.2: Diagrama de muestreo con un transecto lineal. Si se mide la distancia radial r a un individuo, también debe recopilarse el ángulo de observación θ para poder calcular la distancia perpendicular x como $x{=}r\,sen\,(\theta)$

Esta diferencia representa la principal ventaja del muestreo por distancias ya que nos permite estimar la abundancia cuando no podemos detectar todos los individuos dentro de la región cubierta, y podemos obtener buenas estimaciones incluso en casos donde una gran proporción de los individuos permanezca sin detectarse. Más adelante veremos que gran parte de la metodología del muestreo por distancias está enfocada en estimar la proporción de individuos detectados a partir de una **función de detección,** la cual representará **la probabilidad de detectar un individuo dado que se encuentra a una cierta distancia del transecto**. El rango de aplicación del muestreo por distancias es muy amplio, puede utilizarse para poblaciones de objetos inanimados como nidos de pájaros, árboles, madrigueras, en donde resultará un método práctico y eficiente en cuanto a costo y esfuerzo, o para poblaciones animales que se encuentran en constante movimiento como aves, mamíferos terrestres, peces y reptiles, para las

cuales el método de muestreo por parcelas sería inadecuado ya que no podríamos asegurarnos de contar a todos los individuos.

2.1.1. Función de detección

Como ya mencionamos, un concepto fundamental del muestreo por distancias es el de la función de detección que denotaremos por $g(y)$ y la definiremos como:

$g(y)$ =probabilidad de detectar un objeto, dado que se encuentra a distancia y del transecto (línea o punto) seleccionado aleatoriamente

$= \mathbb{P}[\text{detección} \mid \text{distancia } y]$.

En el caso de transectos lineales, la distancia y representará la distancia perpendicular medida desde el transecto y en el caso de transectos puntuales la distancia radial. Podemos observar que por la manera en que se toma la muestra en el muestreo por distancias la función de detección será decreciente conforme aumente la distancia y por ser una probabilidad tendremos que $0 \leq g(y) \leq 1$. El supuesto esencial que haremos para la función de detección será que $g(0)$=1,es decir, cualquier individuo que se encuentre en el transecto será siempre detectado. Generalmente el modelo elegido para la función de detección $g(y)$ dependerá de uno o más parámetros que tendrán que ser estimados a partir de la muestra de distancias obtenidas $y_1, \ldots, y_n$, es decir, una vez ajustado un modelo a los datos, solo conoceremos una función estimada $\hat{g}(y)$. En la figura 2.3 se muestran algunas de las formas más comunes que tienen las funciones de detección en el muestreo por distancias (suponiendo que las distancias varían entre 0 y 1).

Con frecuencia, solo nos será posible detectar una pequeña proporción de nuestros individuos de interés, sin embargo, una buena elección del modelo para la función de detección nos permitirá obtener estimadores confiables para la abundancia.

.

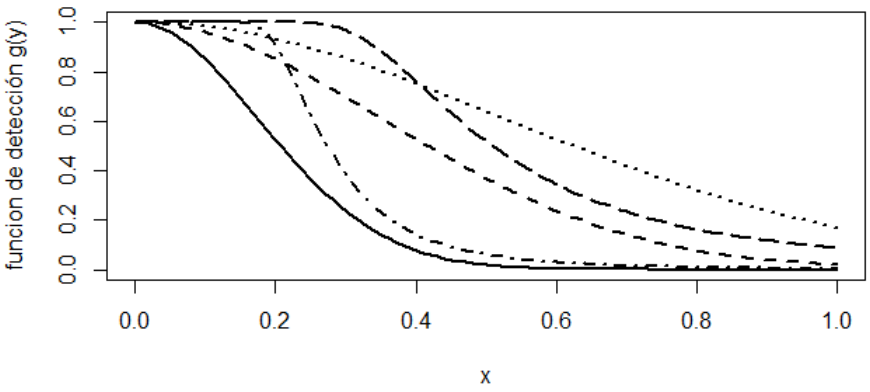

Figura 2.3: Algunos ejemplos de modelos para funciones de detección $g(y)$ comunes en el muestreo por distancias. La característica principal de estos modelos es que la probabilidad de detección decrece a medida que aumenta la distancia.

2.1.2. Truncamiento de datos

Al diseñar un muestreo por transectos lineales es posible establecer una distancia $y = w$ para la cual los individuos que se encuentren a una distancia mayor que w del transecto sean ignorados. De manera similar en el caso de transectos puntuales podemos establecer el radio w.

Si esta distancia no fue fijada antes de obtener la información en el campo, los datos obtenidos en el muestreo pueden truncarse antes de realizar cualquier análisis. Se puede elegir una distancia w tal que descartemos todas las distancias mayores a w para quitar observaciones atípicas del análisis, ya que estas podrían complicar la modelación de la función de detección. La elección de una distancia w adecuada dependerá del estudio que estemos realizando. Por ejemplo, podemos elegir w tal que la función estimada para $g(y)$, $\hat{g}(y)$, cumpla que $\hat{g}(w)$=.10. Un histograma de la distancias recopiladas es de gran ayuda para elegir una distrancia de truncamiento (figura 2.4).

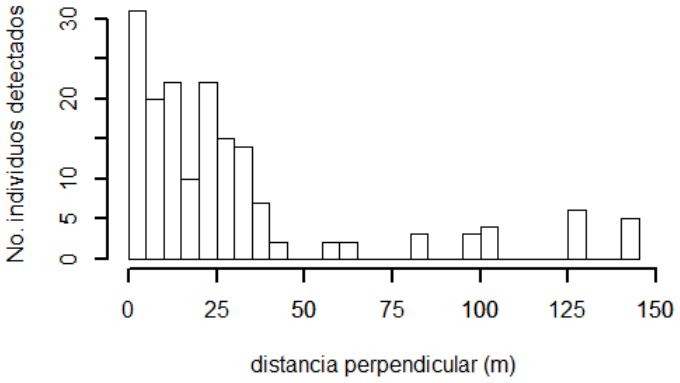

Figura 2.4: Histograma de distancias obtenidas utilizando transectos lineales, los datos sugieren la necesidad de truncar las observaciones después de $75m$

2.1.3. Poblaciones agrupadas

En el muestreo por distancias a menudo es preferible referirnos a un individuo como 'objeto de interés', ya que si bien los objetos de interés pueden ser individuos de una población, con frecuencia se estudiarán poblaciones en las que sus individuos se encuentran agrupados por naturaleza. En este texto haremos referencia a estos grupos como 'clusters', y estos pueden indicar por ejemplo: parvadas, bancos de peces, manadas de lobos, grupos de ballenas, etc.

Al estudiar poblaciones en las que los objetos de interés son clusters, existe una importante diferencia entre el muestreo por parcelas y el muestreo por transectos lineales o puntuales. En el muestreo por parcelas todos los individuos que se encuentran dentro de la parcela son contados y se ignora el hecho de que el individuo sea parte de un cluster. En el muestreo con distancias con una distancia de truncamiento fija w, se medirán la distancias del transecto al centro geométrico del cluster y registraremos un cluster como 'detectado' si el centro geométrico del cluster se encuentra a una distancia entre 0 y w del transecto, en este caso, para el tamaño del cluster se contarán todos los individuos que pertenezcan a él, incluso si incluye individuos que estén más allá de la distancia w. Por el contrario, si el centro geométrico del cluster se encuentra a una distancia mayor que w del transecto, el cluster no será registrado, incluso si algunos de sus individuos se encuentren a una distancia menor a w del transecto.

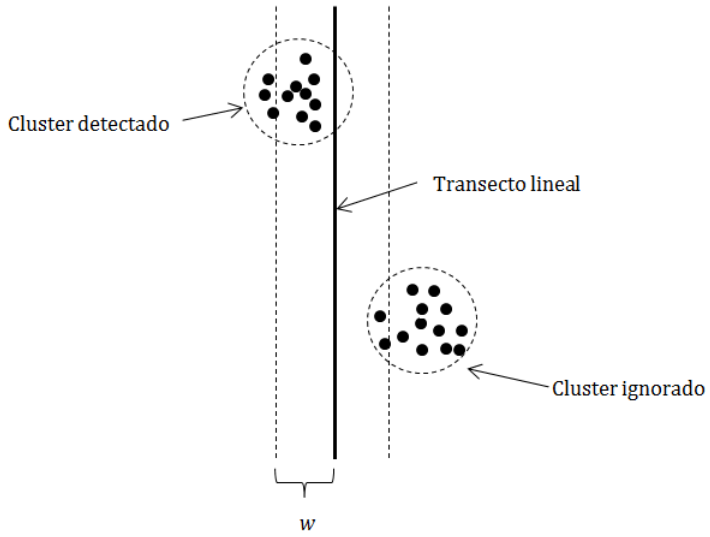

Figura 2.5: Diagrama de un transecto lineal en donde los objetos de interés son clusters. El tamaño del cluster detectado es de 11 individuos a pesar de que 3 de ellos se encuentran a una distancia del transecto mayor a w. El otro cluster es ignorado totalmente a pesar de tener 2 individuos dentro de la franja de ancho w.

Lo anterior se muestra en la figura 2.5.

Una vez obtenidas la distancias del transecto al centro geométrico de cada cluster detectado, la estimación de la abundancia de clusters se hace de igual manera que con individuos, y en este caso el tamaño de muestra n, será el número de clusters detectados durante el estudio. Si además de las distancias, se obtiene el número de individuos (s) de cada cluster detectado, la muestra de tamaños de los clusters $s_1, \ldots, s_n$, nos servirá para estimar el tamaño promedio de los clusters $\mathbb{E}(s)$. De este modo podemos calcular un estimador para la abundancia de los individuos como:

$$\hat{N} = \hat{N}_s * \widehat{\mathbb{E}(s)}$$

en donde $\hat{N}_s$ denota el estimador para la abundancia de los clusters y $\widehat{\mathbb{E}(s)}$ denota el estimador para el tamaño promedio de los clusters.

2.1.4. Constantes conocidas y parámetros

Constantes conocidas

En este capítulo se utilizarán varias constantes conocidas, algunas de las cuales ya fueron definidas anteriormente. Usaremos la siguiente notación:

A = área ocupado por la población de estudio
k = número de transectos utilizados
l_i = longitud del $i-$ésimo transecto
L = longitud total de los transectos $= \sum l_i$
w = distancia de truncamiento fijada en campo o antes del análisis

Parámetros

En el muestreo por distancias se definen los siguientes parámetros desconocidos:

D = densidad (por unidad de área)
N = tamaño total de la población o abundancia
$\mathbb{E}(s)$ = tamaño promedio de los clusters
$f(x)$ = función de densidad de probabilidad de distancias desde el transecto lineal
$g(0)$ = probabilidad de detectar un individuo dado que está sobre el transecto (suponemos igual a 1)

2.1.5. Supuestos

La inferencia realizada en el muestreo por distancias depende de la validez de algunos supuestos sobre las condiciones iniciales de la población de interés y la forma en que se realiza el estudio.

Como condición inicial es necesario que las líneas o puntos sean elegidas aleatoriamente con respecto a la distribución de los individuos sobre el área de interés. Esta aleatoriedad justificará las inferencias realizadas a partir de la muestra a la población. La otra consecuencia esencial de la aleatoriedad en la localización de los transectos es que aseguraremos que los individuos se distribuyan uniformemente con respecto a sus distancias del transecto.

Los siguientes tres supuestos son también esenciales para que se puedan obtener estimadores confiables de la abundancia o densidad usando transectos lineales o puntuales:

a) Los objetos sobre el transecto son siempre detectados

Supondremos que todos los objetos que se encuentran a distancia cero del transecto son detectados, esto es $g(0)=1$. En la práctica, al diseñar el estudio se debe considerar de que manera se cumplirá este supuesto.
Es posible realizar diferentes acciones en el campo para asegurar que $g(0)=1$. Por ejemplo en estudios aéreos y submarinos el uso de cámaras permite asegurar que se detecten los individuos que se encuentren en el transecto o muy cerca de él.

b) Los objetos son detectados en su posición inicial

Al estudiar animales, es posible que algunos se desplacen de su posición inicial antes de ser detectados, en este caso la distancia medida será del transecto a la posición al ser detectado y no a la inicial. Si el desplazamiento antes de la detección es aleatorio el efecto que tendrá en los estimadores será despreciable. Si el movimiento está relacionado con el observador se tendrán que modificar los procedimientos en el campo de manera que las detecciones se realicen a una distancia desde la cual no se provoque el desplazamiento del animal.

c) Las mediciones son exactas

Idealmente las distancias se deben recopilar sin errores de medición. A pesar de que se pueden reducir los efectos de mediciones inexactas con un análisis minucioso para agrupar los datos, es recomendable obtener mediciones precisas desde el campo. En particular es necesario que las mediciones cerca del transecto sean exactas ya que los errores de redondeo podrían producir una acumulación de datos que erróneamente presentaron distancia de detección cero. Errores graves en la medición de las distancias o ángulos de observación producirán estimadores sesgados.

Frecuentemente, cuando las distancias de observación son estimadas, el observador podría redondear las distancias a valores convenientes (e.g. 5, 10, 15,...). De esta manera el histograma de las distancias presentará frecuencias altas en sólo un pequeño número de distancias. De igual manera al medir ángulos es común que se redondee a valores como 0, 30, 45, 60 y 90 grados. A menudo agrupar las distancias en intervalos ayudará a modelar la función de detección.

2.2. Transectos lineales

En esta sección se describirá detalladamente toda la metodología necesaria para estimar la abundancia mediante transectos lineales.

2.2.1. Introducción

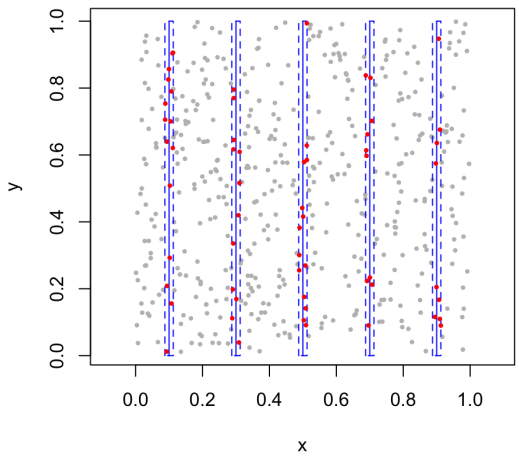

Figura 2.6: Población simulada con parcelas rectangulares (línea punteada) que tienen a los transectos como línea central (línea sólida), los puntos rojos representan los individuos contados.

Como mencionamos anteriormente, el muestreo por transectos lineales se puede entender como una extensión del muestreo por parcelas. Supongamos que contamos con k parcelas rectangulares (figura 2.6), cada rectángulo con

ancho $2w$, de tal manera que la línea central de cada rectángulo se extiende w unidades de cada lado y w es la distancia de truncamiento definida en la sección 2.1.4. Además supongamos que la longitud del $i-$ésimo rectángulo es l_i, con $i = 1, \ldots k$, y la longitud total de todos los rectángulos es $\sum_{i=1}^{k} l_i = L$. Entonces la región cubierta será de tamaño $a = 2wL$. En el caso de muestreo por parcelas el número total de n objetos en la región cubierta son contados y el estimador para la abundancia está dado por $\hat{N} = \frac{n}{\pi_c} = A \cdot \frac{n}{a} = A \cdot \hat{D}$.

En el muestreo por transectos lineales el observador se traslada sobre el transecto y registra las n distancias perpendiculares a los individuos detectados (o en su defecto los ángulos de observación) pero en este caso los n individuos detectados representan solamente una proporción del total de individuos en el área cubierta a. Es decir, la probabilidad de detección es distinta de uno y más aún supondremos que ésta decrece conforme aumenta la distancia del individuo al transecto. Denotemos a la proporción de objetos observados en el área cubierta como P_a. Como ejemplo, supongamos que para un estudio realizado conocemos esta proporción y $P_a = .5$. Es decir, si contamos $n = 100$ **objetos detectados**, podemos suponer que había $\frac{n}{P_a} = \frac{n}{0.5} = 200$ objetos en la región cubierta, lo que intuitivamente sugiere que nuestro estimador de la abundancia para el muestreo por transectos lineales debería ser $\hat{N} = \frac{200}{\pi_c} = \frac{n}{P_a \cdot \pi_c}$. Como era de esperarse, en la práctica no tendremos conocimiento de P_a pero podremos obtener un estimador $\hat{P}_a$ y el (por ahora intuitivo) estimador de la abundancia sería:

$$\hat{N} = \frac{n}{\hat{P}_a \cdot \pi_c} = \frac{n \cdot A}{\hat{P}_a \cdot a}$$

En la siguiente sección desarrollaremos formalmente el procedimiento para la obtención de $\hat{N}$.

2.2.2. Estimación por máxima verosimilitud

Es necesario notar que en el muestreo por distancias existen dos fuentes de aleatoriedad en el modelo, el primero es relativo a la distribución de los objetos en el área de estudio, (en este caso solo nos interesa suponer que la distribución es uniforme alrededor del transecto y no en toda el área de interés como en el muestreo por parcelas) y el segundo tiene que ver por supuesto con la probabilidad de detección. Por lo tanto, para obtener la función de verosimilitud tendremos que incorporar ambas fuentes de aleatoriedad.

Sabemos que P_a, denota la probabilidad de detectar un individuo que se encuentra en la región cubierta, y la probabilidad de que un individuo esté en la región cubierta de área a es $\pi_c = \frac{a}{A}$, entonces la probabilidad de que un individuo se encuentre en la región cubierta y sea detectado es $\pi_c \cdot P_a$. Por lo tanto, como los individuos son detectados independientemente, entonces n el número de objetos de interés detectados es una variable aleatoria binomial con parámetros N y $\pi_c \cdot P_a$, de modo que **si se conociera** P_a la verosimilitud para N estaría dada por:

$$L(N) = \binom{N}{n} (\pi_c P_a)^n (1 - \pi_c P_a)^{N-n} \tag{2.2.1}$$

y el EMV para N sería $\frac{n}{P_a \cdot \pi_c}$. Podemos observar que la ecuación (2.2.1) es igual a la ecuación (1.2.1) cuando $P_a = 1$. Sin embargo, al no conocer P_a la verosimilitud tiene dos parámetros desconocidos y será necesario obtener el EMV de P_a a partir de la muestra de distancias perpendiculares al transecto.
Para hacer esto notemos primero que :

$$\begin{aligned} P_a &= \mathbb{P}(\text{detección}) \\ &= \int_0^w \mathbb{P}(\text{detección} \mid \text{distancia } x) \cdot \frac{1}{w}\, dx \\ &= \int_0^w g(x) \cdot \frac{1}{w}\, dx \\ &= \frac{\mu}{w} \end{aligned} \tag{2.2.2}$$

donde $g(x)$ es la función de detección definida en la sección 2.1.1 y $\mu \coloneqq \int_0^w g(x)\, dx$. Es claro que $g(x)$ y por tanto μ dependen de un vector de parámetros desconocidos $\boldsymbol{\theta}$ que ahora habrá que estimar por máxima versosimilitud a partir de la muestra de distancias, denotadas por $x_1, \ldots, x_n$. Para esto desarrollaremos ahora la función de verosimilitud de $\boldsymbol{\theta}$.
Por la definición de la función de detección $g(x)$, la distribución uniforme de los objetos alrededor del transecto y el Teorema de Bayes se tiene que:

$$\begin{aligned} f(x) \coloneqq \mathbb{P}(\text{distancia } x \mid \text{detección}) &= \frac{\mathbb{P}(\text{detección} \mid \text{distancia } x) \cdot \frac{1}{w}}{\mathbb{P}(\text{detección})} \\ &= \frac{g(x) \cdot \frac{1}{w}}{\int_0^w g(x) \cdot \frac{1}{w}\, dx} = \frac{g(x)}{\mu} \end{aligned} \tag{2.2.3}$$

por lo tanto del hecho que $f(x)$ es una densidad condicional, la cantidad $\mu = \int_0^w g(x)\,dx$ es la constante necesaria para que $\int_0^w f(x)\,dx = 1$, es decir $f(x)$ y $g(x)$ son proporcionales.

Si los objetos son detectados independientemente, la densidad de las distancias dado que los objetos fueron detectados es $\prod_{i=1}^{n} f(x_i)$ y la verosimiltud para $\boldsymbol{\theta}$ dada la muestra de distancias es:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i) = (\mu)^{-n} \prod_{i=1}^{n} g(x_i) \tag{2.2.4}$$

Maximizando la ecuación (2.2.4) se obtiene $\hat{\boldsymbol{\theta}}$, el EMV de $\boldsymbol{\theta}$.

En resumen para estimar N usando transectos lineales debemos en primer lugar elegir un modelo adecuado para $g(x)$ analizando el histograma de distancias (véase sección 2.2.3) , después debemos obtener $\hat{\boldsymbol{\theta}}$ maximizando 2.2.4 (lo cual en la práctica casi siempre requiere métodos numéricos), posteriormente estimar μ como $\hat{\mu} = \int_0^w \hat{g}(x)\,dx$ y P_a como $\widehat{P_a} = \frac{\hat{\mu}}{w}$, lo que finalmente resulta en :

$$\hat{N} = \frac{n}{\widehat{P_a} \cdot \pi_c} = \frac{nA}{\frac{\hat{\mu}}{w} \cdot a} = \frac{nA}{\frac{\hat{\mu}}{w} \cdot 2wL} = \frac{nA}{2\hat{\mu}L} \tag{2.2.5}$$

La ecuación (2.2.5) es válida incluso si $w = \infty$, en este caso $\hat{\mu} = \int_0^\infty \hat{g}(x)\,dx$. Una vez que conocemos $\hat{f(x)}$ la forma más simple de obtener $\hat{\mu}$ es usando la ecuación (2.2.3) y el supuesto $g(0) = 1$, lo que resulta en $\hat{f}(0) = \frac{1}{\hat{\mu}}$.

Ahora mostraremos un ejemplo en donde se ilustra explícitamente el procedimiento anterior. Supongamos que el modelo que seleccionamos para la función de detección es $g(x) = e^{-\lambda x}$ y que w no está acotado. Entonces:

$$\mu = \int_0^\infty e^{-\lambda x}\,dx = \frac{1}{\lambda}$$

Usando la ecuación (2.2.4) tenemos

$$L(\lambda) = \left(\frac{1}{\lambda}\right)^{-n} \prod_{i=1}^{n} e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

$$\Rightarrow l(\lambda) = \ln\left(L(\lambda)\right) = n\ln(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

haciendo $\frac{d(l(\lambda))}{d\lambda} = 0$

$$\frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

de donde

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$$

por lo tanto el EMV para μ es

$$\hat{\mu} = \frac{1}{\hat{\lambda}} = \frac{\sum_{i=1}^{n} x_i}{n},$$

y por la ecuación (2.2.5)

$$\hat{N} = \frac{nA}{2\hat{\mu}L} = \frac{n^2 A}{2L\sum_{i=1}^{n} x_i}$$

En la siguiente sección se mostrarán los criterios para la elección de un modelo para la función de detección.

2.2.3. Estimación de la función de detección

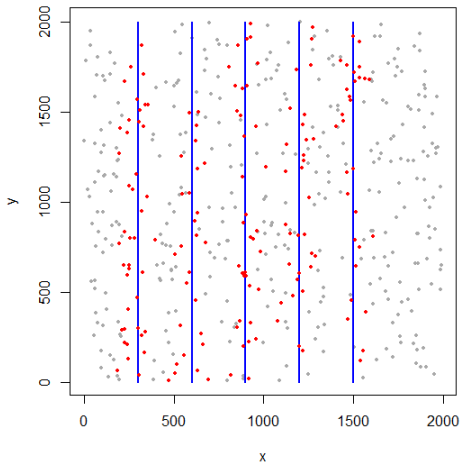

Figura 2.7: Población simulada con N=500 objetos, se simula el estudio con cinco transectos y se detectan 165 objetos (puntos rojos). En este caso A=4,000,000m^2 y L=$\sum_{k=1}^{5} l_i$=10,000m.

Como se mencionó en la sección anterior, el primer paso para obtener el estimador máximo verosímil de la abundancia por medio del muestreo por transectos lineales, es elegir un modelo parámetrico para la función de detección $g(x)$ y estimar sus parámetros a partir de la muestra de distancias perpendiculares a partir del tansecto.

La figura 2.7 es una representación de una simulación de muestreo por transectos lineales, misma que se utilizará como ejemplo para estimar la función de detección.

Una vez obtenidas la distancias perpendiculares a los objetos podemos graficar el histograma de distancias (figura 2.8).

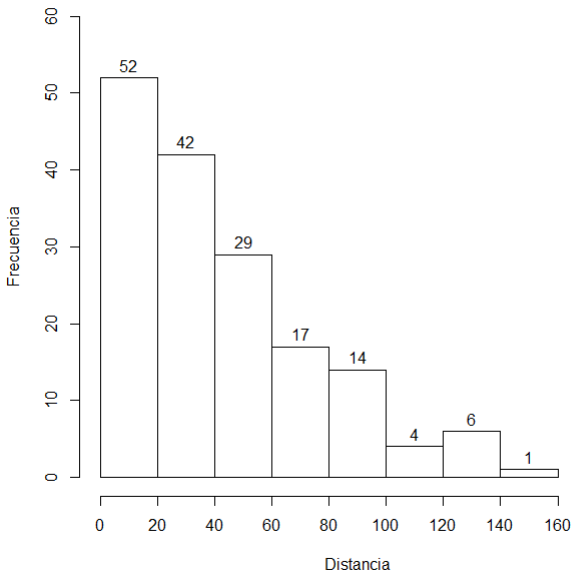

Figura 2.8: Histograma de las distancias perpendiculares obtenidas en la simulación de la figura 2.7. Se muestra la clara disminución de la frecuencia conforme aumenta la distancia de observación.

Mostraremos ahora dos alternativas que son comúnmente utilizadas para modelar a la función de detección en el muestreo por distancias y posteriormente veremos algunos criterios para elegir el modelo más adecuado para los datos del ejemplo.

El primer modelo que veremos es una caso particular de una normal truncada llamado en inglés ***half-normal,*** en este caso resulta adecuado ya que tenemos datos no agrupados y no usaremos una distancia de truncamiento. En este modelo la función de detección está dada por $g(x)=\exp(-x^2/2\sigma^2)$, $0 \leq x < \infty$, $\sigma > 0$. La figura 2.9 muestra la forma de la función half normal con dos valores distintos del parámetro.

Usualmente al elegir un modelo paramétrico para $g(x)$ es necesario usar métodos numéricos para obtener los estimadores máximo verosímiles de sus parámetros, pero en el caso del modelo half-normal siguiendo el procedimiento visto en la sección 2.2.2 podemos obtener explicitamente los estimadores para los parámetros del modelo.

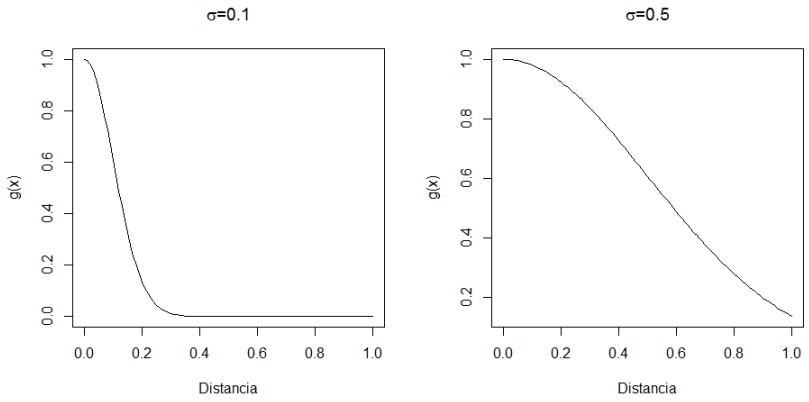

Figura 2.9: Funciones *half-normal* con $\sigma = 0.1$ y $\sigma = 0.5$, las curvas describen el comportamiento decreciente de las detecciones al incrementar la distancia, el mismo comportamiento que se observa en el histograma de la figura 2.8. Un valor grande de σ se traduce en una mayor probabilidad de observar objetos a grandes distancias.

Comenzamos obteniendo μ:

$$\mu = \int_0^\infty g(x)\,dx = \int_0^\infty \exp(-x^2/2\sigma^2)\,dx = \sqrt{\frac{\pi\sigma^2}{2}}$$

Dado n el número de detecciones y usando las ecuaciones 2.2.3 y 2.2.4 obtenemos la verosimilitud

$$L(\sigma) = \prod_{i=1}^{n} f(x_i) = \left(\sqrt{\frac{\pi\sigma^2}{2}}\right)^{-n} \prod_{i=1}^{n} \exp(-x_i^2/2\sigma^2)$$

de donde

$$l(\sigma) = \ln(L(\sigma)) = -\sum_{i=1}^{n}\left(\frac{x_i^2}{\sigma^2}\right) - n\,\ln\left(\sqrt{\frac{\pi\sigma^2}{2}}\right)$$

Derivando $l(\sigma)$ con respecto de σ^2 e igualando a cero, resulta:

$$\frac{dl(\sigma)}{d\sigma^2} = \sum_{i=1}^{n}\frac{x_i^2}{2\sigma^2} - \frac{n}{2\sigma^2} = 0$$

de tal manera que $\hat{\sigma}^2 = \sum_{i=1}^{n}\frac{x_i^2}{n}$ y $\hat{\mu} = \sqrt{\frac{\pi\sum_{i=1}^{n} x_i^2}{2n}}$. Aunque en este caso tomamos $w = \infty$, en la práctica si los datos no se truncan es común fijar w como

la máxima distancia detectada.

Utilizando los datos de la población simulada de la figura 2.7 y sustituyendo en la ecuación (2.2.5) tenemos:

$$\hat{N} = \frac{nA}{2\hat{\mu}L} = \frac{165 * 4,000,000}{2 * 67.43 * 10,000} = 489.33 \approx 489 \,\text{individuos}$$

lo cual es cercano a la abundancia real de 500 individuos.

La figura 2.10 muestra el ajuste de la curva estimada $\hat{g}(x)$ al histograma de la muestra de distancias perpendiculares.

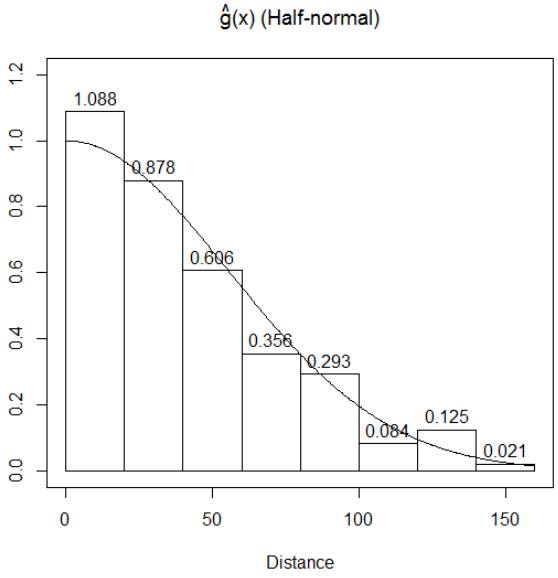

Figura 2.10: Ajuste half-normal de la función de detección estimada $\hat{g}(x)$ con parámetos obtenidos por máxima verosimilitud e histrograma de distancias. El histograma tiene escala de manera que su área coincida con el área bajo la curva $\hat{g}(x)$.

Es también común mostrar el histograma de densidad comparando este con la función $\hat{f}(x)$ (figura 2.11).

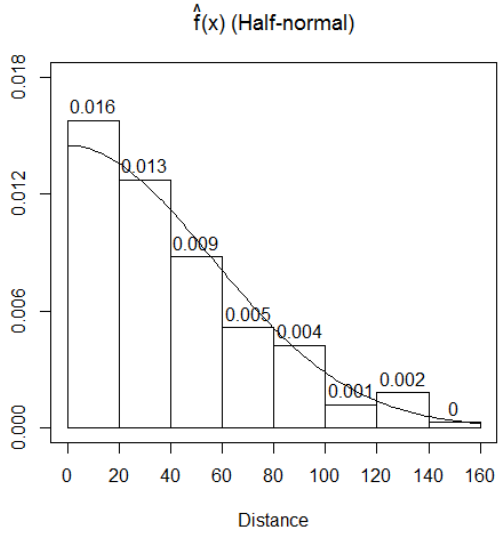

Figura 2.11: Ajuste half-normal de la función de densidad estimada $\hat{f}(x)$ con parámetos obtenidos por máxima verosimilitud e histrograma de distancias. El histograma tiene área igual a uno y por supuesto $\hat{f}(x)$ tiene la misma forma que $\hat{g}(x)$.

El siguiente modelo que veremos es el modelo ***hazard-rate***, el cual es un modelo más flexible pero requiere la estimación de dos parámetros. La función de detección para este modelo es $g(x) = 1 - \exp\left(-\left(\frac{x}{\sigma}\right)^{-b}\right)$, en donde b es un parámetro de forma y σ es un parámetro de escala. En la figura 2.12 se muestran las formas que toma esta función al variar ambos parámetros.

A diferencia del modelo half-normal, los estimadores máximo verosímiles de los parámetros para el modelo hazard-rate no tienen una forma cerrada por lo que habrá que calcularlos usando un paquete estadístico.

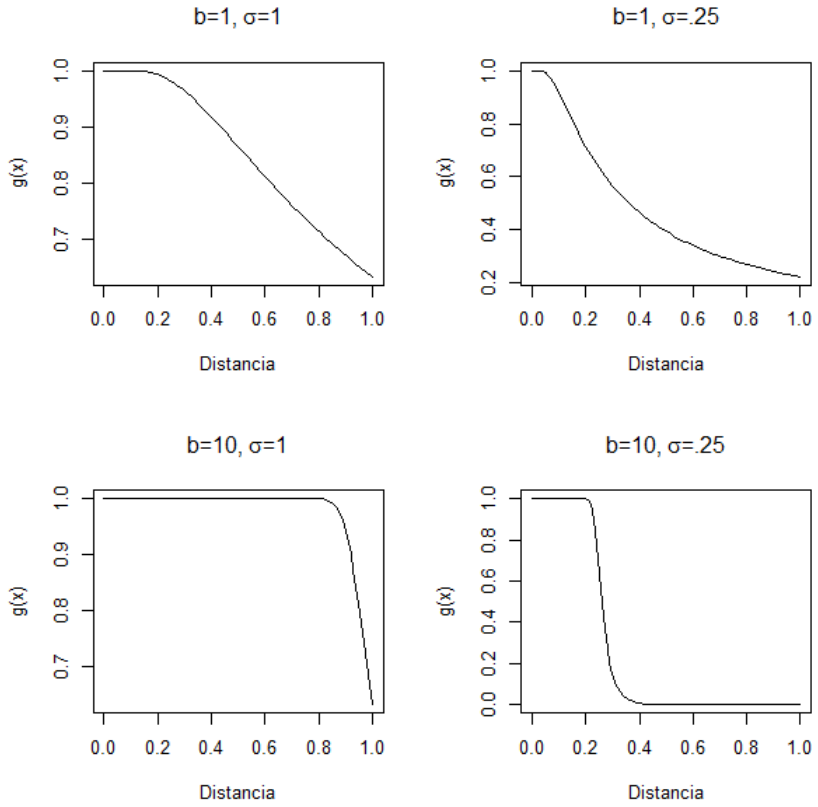

Figura 2.12: Funciones hazard-rate con distintos parámetros de escala y de forma. La función hazard-rate tiene un " hombro" más prominente que la función half normal y el parámetro de forma b, controla el tamaño del " hombro", es decir, que tan lejos la probabilidad de detección permanece igual a uno.

Tomando w como la máxima distancia detectada para la misma población ejemplo, obtenemos los valores estimados $\hat{b}$=2.16, $\hat{\sigma} = 46$ y $\hat{\mu} = 66.04$.

Utilizando de nuevo la ecuación (2.2.5) tenemos:

$$\hat{N} = \frac{nA}{2\hat{\mu}L} = \frac{165 * 4,000,000}{2 * 66.04 * 10,000} = 499.70 \approx 500 \,\text{individuos}$$

lo cual en este caso es el valor real de la abundancia de la población ejemplo y claramente es un mejor valor que el obtenido al ajustar el modelo half-normal.

Las figuras 2.13 y 2.14 muestran el ajuste del modelo hazard-rate a los datos.

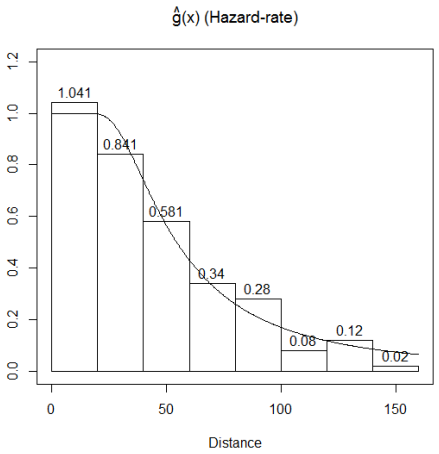

Figura 2.13: Ajuste hazard-rate de la función de detección estimada $\hat{g}(x)$ con parámetos obtenidos por máxima verosimilitud e histrograma de distancias. El histograma tiene escala de manera que su área coincida con el área bajo la curva $\hat{g}(x)$.

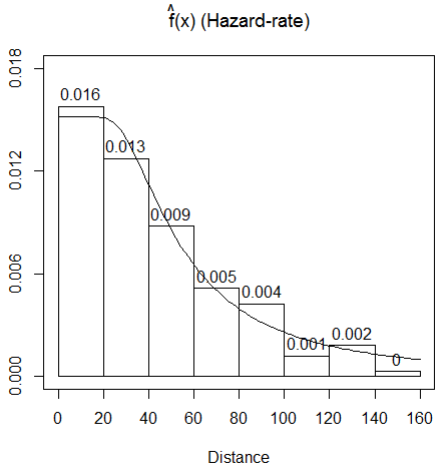

Figura 2.14: Ajuste hazard-rate de la función de densidad estimada $\hat{f}(x)$ con parámetos obtenidos por máxima verosimilitud e histrograma de distancias. El histograma tiene área igual a uno.

Si bien la figuras anteriores nos brindan una idea de la calidad del ajuste del modelo, es necesario analizar criterios más formales para poder elegir un modelo.

2.2.4. Elección del modelo

Para hacer una correcta elección del modelo para $g(x)$ necesitamos herramientas o criterios que nos permitan seleccionar de entre los modelos que hayamos ya ajustado.

Gráficos P-P

Un gráfico P-P es un método para comparar dos distribuciónes de probabilidad y evaluar que tan similares son. En nuestro caso, consideremos primero la función de distribución acumulada ajustada de las distancias perpendiculares, la cual denotaremos por $\hat{F}(x)$, recordando que para cada valor de x se tiene que:

$$\hat{F}(x) = \int_0^x \hat{f}(u)\, du = \int_0^x \frac{\hat{g}(u)}{\hat{\mu}}\, du$$

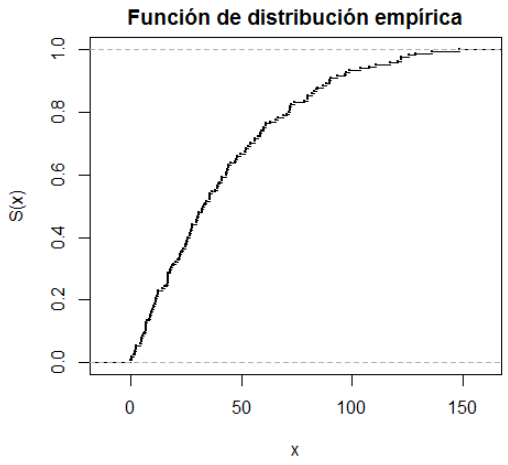

Figura 2.15: Función de distribución empírica de las distancias. Por definición será siempre constante en los intervalos $[x_i, x_{i+1})$ para $i = 1, \ldots, n-1$.

La función con la que debemos comparar a $\hat{F}(x)$ es la **función de distribución empírica** (denotada por $S(x)$) de la muestra de distancias perpendiculares $x_1, \ldots, x_n$, la cual como sabemos representa para cualquier valor t, la proporción

de observaciones menores o iguales que t. De esta manera podemos realizar el gráfico P-P el cual constará de todos los puntos $(S(x_i), \hat{F}(x_i))$ con $i = 1, \ldots, n$ y mostrará la similaridad de ambas funciones si la mayoría de los puntos están sobre la recta $y{=}x$. La figura 2.15 muestra la función de distribución empírica de las distancias simuladas.

La figura 2.16 muestra los gráficos P-P para los dos modelos ajustados en la sección anterior.

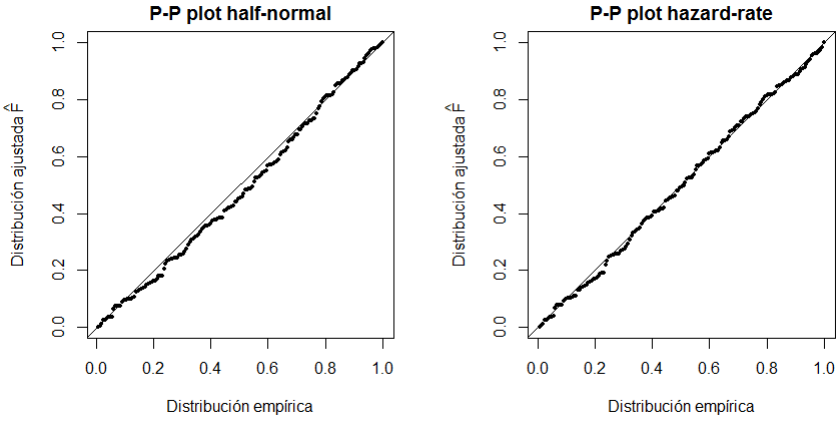

Figura 2.16: Gráficos P-P para el modelo half-normal y hazard-rate.

Observando el ajuste de ambos modelos notamos que en el caso half-normal la mayoría de los puntos no están sobre la recta, a diferencia del modelo hazard-rate que parece ajustar mejor a los datos.

Es posible profundizar las observaciones realizadas en los gráficos P-P con la ayuda de pruebas estadísticas para la bondad de ajuste.

Prueba de Kolmogorov-Smirnov

La prueba de Kolmogorov-Smirnov es una prueba no paramétrica que determina la bondad de ajuste entre dos distribuciónes de probabilidad y puede ser usada para comparar una muestra con una función de referencia. La estadística de prueba que se utiliza es el supremo de las distancias entre la función de distribución empírica $S(x)$ y la función de referencia que en este caso será $\hat{F}(x)$. Se probará la hipótesis nula de que la distribución de la muestra de distancias $x_1, \ldots, x_n$, es igual a $\hat{F}(x)$ y rechazaremos con un nivel de significancia α si

el p-value de la prueba es menor que α o equivalentemente si la estadística de prueba es mayor al cuantil $1-\alpha$ de su distribución.

Continuando con el ejemplo, la prueba Kolmogorov-Smirnof para el modelo half-normal produce una estadística de prueba de 0.056 y un p-value de 0.671, mientras que para la función hazard-rate obtenemos 0.039 y .95782 respectivamente. Aunque en ninguno de los dos casos se rechaza la hipótesis nula, la estadística de prueba, el p-value y el gráfico P-P sugieren que el ajuste del modelo hazard-rate es mejor.

Prueba de Cramér-von Mises

En lugar de tomar en cuenta solamente el supremo de las distancias entre $S(x)$ y $\hat{F}(x)$, podríamos preferir usar todas las distancias entre ambas funciones evaluadas en la muestra. La estadística de prueba de Cramér-von Mises considera la suma de todas estas distancias, sirve para probar la misma hipótesis nula que la prueba Kolmogorov-Smirnov y los criterios para rechazar son análogos.

Para nuestros datos simulados obtenemos una estadística de prueba de 0.122 y un p-value de 0.48 para el modelo half-normal. Con el modelo hazard-rate obtenemos 0.23 y 0.99 para la estadística y el p-value respectivamente. Utilizando esta prueba rechazamos con un nivel de significancia de 0.05 la hipótesis que la distribución de las distancias es half-normal, lo que confirma lo observado en los criterios anteriores.

Prueba de la χ^2

Como alternativa a las pruebas anteriores se muestra la prueba de la Ji-cuadrada para bondad de ajuste. Esta prueba mantiene la misma hipótesis nula que las anteriores pero su estadística de prueba se basa en los cuadrados de las diferencias entre las frecuencias esperadas en un intervalo y las observadas.

Supongamos que la muestra de distancias $x_1, \ldots, x_n$ la separamos en u grupos, con frecuencias absolutas $n_1, \ldots, n_u$. Sean $c_0 = 0, c_1, \ldots, c_u = w$ los puntos de corte que definen los u grupos y w la distancia de truncamiento, de tal forma que n_i será el número de observaciones en el intervalo $[c_{i-1}, c_i)$ con $i = 1, \ldots, u$. Supongamos también que se ajusta un modelo $g(x)$ con q parámetros para la función de detección, entonces podemos obtener la función estimada $\hat{f}(x)$. El área bajo esta función entre c_{i-1}y c_i está dada por:

$$\hat{\pi_i} = \int_{c_{i-1}}^{c_i} \hat{f}(x)\, dx$$

de modo que si la hipótesis nula es cierta, la cantidad $n \cdot \hat{\pi_i}$ representa la frecuencia esperada para el intervalo $[c_{i-1}, c_i)$ y n_i es la frecuencia observada, entonces la estadística:

$$T = \sum_{i=1}^{u} \frac{(n_i - n \cdot \hat{\pi_i})^2}{n \cdot \hat{\pi_i}}$$

se distribuye como χ^2 con $u - q - 1$ grados de libertad. Con un nivel de significancia α rechazamos la hipótesis nula si la estadística T es mayor que el cuantil $1 - \alpha$ de una $\chi^2_{(u-q-1)}$.

Para la población ejemplo tomemos $u = 8$, w como la máxima distancia observada (148.77), $c_i = 20 \cdot i$ con $i = 0, \ldots, u-1$ y $c_u = w$.

En el modelo half-normal estimamos $q = 1$ parámetro por lo que la estadística T tiene distribución $\chi^2_{(6)}$. En este caso la estadística $T = 5.79$ y con $\alpha = 0.05$ el cuantil del $1 - \alpha$ de una $\chi^2_{(6)}$ es $W^{(6)}_{0.95} = 12.6$, y se tiene que $T < W^{(6)}_{0.95}$ por lo que no se rechaza la hipótesis nula. Realizando la prueba con el modelo hazard-rate q=2 y T se distribuye $\chi^2_{(5)}$. La estadística T=3.28 y con un nivel de significancia de 0.05 el cuantil $1 - \alpha$ de la distribución de T es $W^{(5)}_{0.95} = 11.07$, de modo que no se rechaza la hipótesis nula. Equivalentemente se calcula para el modelo half-normal el p-value de 0.45 y para el hazard-rate 0.66, con lo que se decide no rechazar en ambos casos. El resultado anterior indica que ambos modelos tienen un ajuste aceptable a los datos pero el menor valor de la estadística de prueba para la función hazard- rate sugiere que este modelo es más adecuado.

Criterio de información de Akaike

El criterio de información de Akaike (AIC) es una medida cuantitativa para la selección de un modelo. Teniendo al menos dos modelos, el AIC estima la calidad relativa de un modelo con respecto a los demás. El AIC no prueba el ajuste de un modelo en el sentido de probar una hipótesis nula, de manera que el AIC no permite identificar si alguno o todos los modelos ajustan mal a los datos. El AIC se define como:

$$\text{AIC} = -2\ln L(\hat{\boldsymbol{\theta}}) + 2q$$

en donde $L(\hat{\boldsymbol{\theta}})$ es la verosimilitud de la ecuación (2.2.4) evaluada en el estimador máximo verosímil de $\boldsymbol{\theta}$ y q es el número de parámetros estimados en el modelo. Podemos interpretar la cantidad $-2\ln L(\hat{\boldsymbol{\theta}})$ como una medida de que tan bueno es el ajuste del modelo a los datos, mientras que $2q$ es una "penalización" por un incluir más parámetros en el modelo. El ajuste del modelo se puede siempre mejorar añadiendo parámetros al modelo, pero también aumenta la complejidad y la varianza. El AIC proporciona una medida equilibrada entre el ajuste del modelo y la complejidad.
Dada la muestra de distancias $x_1, \ldots, x_n$ se calcula el AIC para cada modelo candidato y se elige el de menor AIC. Continuando con el ejemplo, el modelo half-normal tiene un $\text{AIC}_{HN} = 1554.57$ mientras que para el modelo hazard-rate tenemos que $\text{AIC}_{HR} = 1553.16$. En este caso particular, si sólo nos enfocaramos en el AIC, tenemos que aunque el valor del AIC para el modelo hazard-rate es menor podríamos elegir el modelo half-normal por simplicidad, ya que el parámetro adicional de la función hazard-rate produce solo una leve mejora al AIC. Para elegir un modelo es recomendable considerar todos los métodos anteriores ya que por sus diferentes características proporcionan información distinta acerca de los modelos candidatos, utilizando sólo uno o dos podríamos tomar una decisión sin tener un panorama completo de las ventajas y desventajas de cada modelo. La elección de un modelo también dependerá de la situación particular, es decir, en ocasiones podríamos elegir un modelo de mejor ajuste a pesar de estimar una mayor (pero aún razonable) cantidad de parámetros y en otras situaciones quizá haya que priorizar la simplicidad del modelo sobre una calidad inferior del ajuste. En el caso de la población simulada podríamos por ejemplo, elegir el modelo hazard-rate el cual obtiene mejores resultados en todos las pruebas y criterios anteriores, y sólo requiere la estimación de dos parámetros, es decir, sigue siendo un modelo relativamente simple.

2.2.5. Estimación por intervalos

Para la estimación por intervalos en el muestreo por distancias, el método de normalidad asintótica depende fuertemente del modelo ajustado para la función de detección, ya que el estimador de la varianza de $\hat{N}$ ($\widehat{Var}(\hat{N})$), será una función del estimador de la varianza del número de detecciones n $\left(\widehat{Var}(n)\right)$ y del

estimador de la varianza de $\hat{f}(0)$ $\left(\widehat{Var}(\hat{f}(0))\right)$. De manera similar, el método de bootstrap paramétrico requiere generar objetos que suponemos que se distribuyen uniformemente sobre el área de estudio o alrededor del transecto, además de simular distancias de detección usando la función de detección estimada $\hat{g}(x)$. Por lo anterior, estos dos métodos para calcular intervalos de confianza suelen ser sensibles a poblaciones más o menos agrupadas de lo que permite la distribución uniforme o a ajustes inadecuados de la función de detección.

El método que permite evitar de mejor manera la dependencia en los supuestos de uniformidad y en la calidad del ajuste de la función de detección, es el método de bootstrap no-parámetrico, en el cual nos enfocaremos en esta sección. Utilizar el bootstrap no-parámetrico tomando como unidad de muestreo a los transectos, asegura que la información obtenida de diferentes transectos sea independiente, que los datos del remuestreo no se generen a partir del modelo ajustado a la función de detección y que no se use ningún supuesto distribucional de los objetos en el área de estudio.

Si todos los transectos tienen la misma longitud, el bootstrap no-paramétrico consistirá en remuestrear con reemplazo los transectos hasta que el tamaño de muestra sea igual a k, el número original de transectos en la muestra. Con transectos de distinta longitud, el remuestreo debe hacerse hasta que la longitud total de los transectos sea igual o casi igual a la longitud total original L. Para una estimación confiable es recomendable tener entre 15 y 20 transectos. Considerando lo anterior, el bootstrap no-paramétrico se implementa de manera muy similar a como se hace en el muestreo por parcelas (sección 1.3). Para la muestra bootstrap i (i=1,...,B) obtenemos una muestra con reemplazo de k transectos de la muestra original, donde k es el número de transectos inicial. Calculamos para cada muestra bootstrap el estimador $\hat{N}_i$ (ecuación (2.2.5)) obteniendo así la muestra $\hat{N}_1, \ldots, \hat{N}_B$. Calculamos también la media y la varianza muestra de $\hat{N}_1, \ldots, \hat{N}_B$:

$$\bar{N}_{BN} = \frac{\sum_{i=1}^{B} \hat{N}_i}{B} \tag{2.2.6}$$

$$\widehat{Var}_{BN}(\hat{N}_i) = \frac{\sum_{i=1}^{B} \left(\hat{N}_i - \bar{N}_{BN}\right)^2}{B-1} \tag{2.2.7}$$

Usando el método de los percentiles en la muestra $\hat{N}_1, \ldots, \hat{N}_B$, tenemos que el intervalo de confianza del $(1-\alpha)\times 100\,\%$ para N es $\left(\hat{N}_{(r)}, \hat{N}_{(s)}\right)$ donde $\hat{N}_{(r)}$ y $\hat{N}_{(s)}$ denotan los percentiles del $\frac{\alpha}{2}\times 100\,\%$ y del $\left(1-\frac{\alpha}{2}\right)\times 100\,\%$ respectivamente.

Aplicando el procedimiento anterior para la población simulada, usando el modelo half-normal y $B = 50,000$ obtenemos que el intervalo del 95 % de confianza para N es (435.95, 548.65) y $\widehat{Var}_{BN}(\hat{N})$=794.53 . Para el modelo hazard-rate y el mismo número de muestras bootstrap, el intervalo del 95 % de confianza para N es (445.18, 557.24) y $\widehat{Var}_{BN}(\hat{N}) = 826.89$. Ambos intervalos contienen al verdadero valor de la abundancia (500).

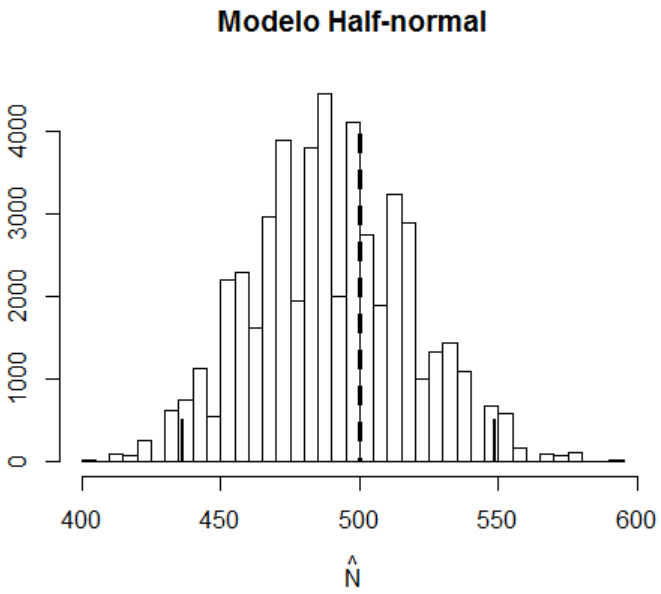

Figura 2.17: Distribución de $\hat{N}$, estimada con bootstrap no-paramétrico bajo el modelo Half-normal. La línea punteada representa el verdadero valor de $N = 500$ y las líneas sólidas verticales indican los extremos del intervalo del 95 % de confianza obtenido. El intervalo no es simétrico alrededor de N.

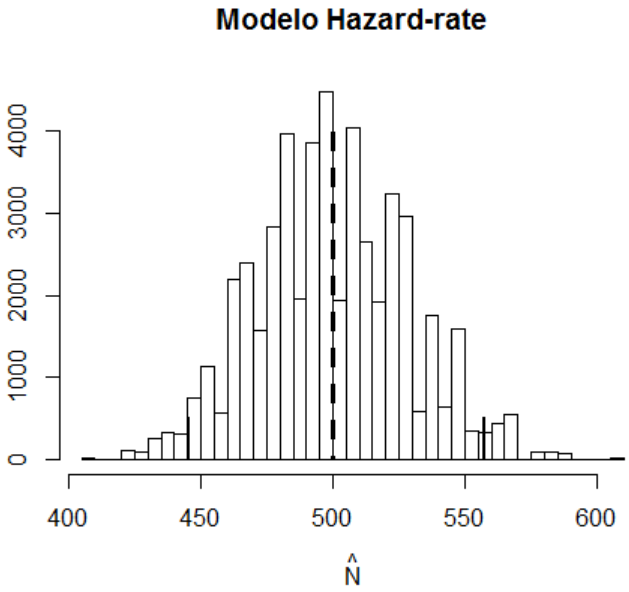

Figura 2.18: Distribución de $\hat{N}$, estimada con bootstrap no-paramétrico bajo el modelo Hazard-Rate. La línea punteada representa el verdadero valor de $N = 500$ y las líneas sólidas verticales indican los extremos del intervalo del 95 % de confianza obtenido. Para este modelo el intervalo obtenido está cercano a la simetría alrededor de N.

2.3. Transectos puntuales

En esta sección se describirá la metodología de transectos puntuales y se mencionarán algunas diferencias de uso e implementación que tiene con el método de transectos lineales.

2.3.1. Introducción

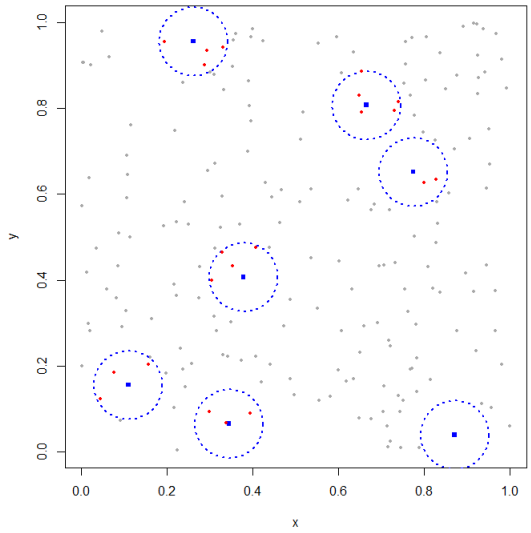

Figura 2.19: Población simulada con siete parcelas circulares localizadas aleatoriamente, los cuadrados representan los centros de las parcelas y los puntos rojos los individuos contados.

En algunas ocasiones, recorrer un transecto lineal para realizar las detecciones puede ser complicado debido a la naturaleza del terreno o de la población de interés. Como alternativa podríamos preferir permanecer en un punto fijo para ver u oír a los individuos de interés durante un periodo de tiempo. Este es precisamente el enfoque del muestreo por transectos puntuales. Para ver este método como una extensión del muestreo por parcelas, supongamos que dentro del área de estudio (de área A) se tienen k parcelas circulares localizadas aleatoriamente de radio w las cuales formaran la región cubierta con área $a{=}k\pi w^2$ y dentro de la cual se detecta el total de individuos, n (figura 2.19). En este caso $\pi_c = \frac{k\pi w^2}{A}$

y el estimador de la abundancia obtenido mediante el muestreo por parcelas será

$$\hat{N}=\frac{n}{\pi_c}=\frac{n\cdot A}{k\pi w^2}.$$

En el muestreo por tansectos puntuales se detectan n individuos dentro de los k círculos de radio w y se guardan los radios de observación $r_1,\ldots,r_n$. Al igual que en el muestreo por transectos lineales, n no representará la totalidad de individuos dentro de la región cubierta, sino una proporción desconocida P_a que al estimarla a partir de la muestra de radios de observación nos permitirá obtener el estimador de la abundancia:

$$\hat{N}=\frac{nA}{k\pi w^2\hat{P}_a}$$

donde w es la distancia de truncamiento. En la siguiente sección formalizaremos la obtención de este estimador.

2.3.2. Estimación por máxima verosimilitud

Al igual que en el muestreo por transectos lineales obtener el estimador por máxima verosimilitud usando transectos puntuales requiere considerar las dos fuentes de aleatoriedad ya mencionadas, la distribución de individuos en la región cubierta y la probabilidad de detección. La modelación de la función de detección se hará de igual forma que con transectos lineales pero la distribución de los individuos dentro de la región cubierta es la principal diferencia entre ambos enfoques. En el caso de transectos lineales era pertinente suponer que el número esperado de individuos era el mismo a cualquier distancia del transecto dentro de la región cubierta, es decir la función de densidad de las distancias para los individuos dentro de la región cubierta es $\frac{1}{w}$, donde w es la distancia de truncamiento, es decir, es constante para cualquier distancia. Para el caso de transectos puntuales, el número esperado de individuos se incrementa conforme crece la distancia al transecto, ya que el área de círculos concéntricos del transecto crece (de forma cuadrática) conforme nos alejamos de él, por lo que la densidad de las distancias a los individuos dentro de la región cubierta no será constante.

Supongamos una vez más que es igualmente probable que un individuo se encuentre en cualquier lugar dentro de la region de estudio de área A y sea w la distancia de truncamiento, de esta manera si R denota a la variable aleatoria de las distancias del transecto a un objeto dentro de un círculo de radio w, se

tiene que la función de distribución acumulativa de R es:

$$\Pi(r) = \mathbb{P}(R \leq r) = \frac{\pi r^2}{\pi w^2}, \text{ si } 0 \leq r < w$$

derivando se sigue que la densidad de distancias radiales para individuos dentro del círculo de radio w es:

$$\pi(r) = \frac{d\left(\Pi(r)\right)}{dr} = \frac{2\pi r}{\pi w^2} = \frac{2r}{w^2}, \text{ si } 0 \leq r < w$$

De la sección 2.2.2 sabemos que si P_a es la probabilidad de detectar un individuo dentro del área cubierta y π_c la probabilidad de que un individuo esté en la región cubierta, el número de detecciones n es una variable aleatoria binomial de párametros N y $\pi_c \cdot P_a$ y la verosimilitud de N está dada por la ecuación (2.2.1). Por lo que el EMV para N está dado por:

$$\hat{N} = \frac{n}{\pi_c \hat{P}_a} = \frac{nA}{k\pi w^2 \hat{P}_a}$$

donde $\hat{P}_a$ es el estimador máximo verosímil de P_a.
La obtención de $\hat{P}_a$ para transectos puntuales es igual a como se hizo para transectos lineales, pero claramente cambiando la densidad $\frac{1}{w}$ por $\pi(r) = \frac{2r}{w^2}$. Se tienen por lo tanto que:

$$\begin{aligned} P_a &= \mathbb{P}(\text{detección}) \\ &= \int_0^w \mathbb{P}(\text{detección} \mid \text{radio } r) \cdot \pi(r)\, dr \\ &= \int_0^w g(r) \cdot \frac{2\pi r}{\pi w^2} dr \\ &= \frac{\nu}{\pi w^2} \end{aligned} \tag{2.3.1}$$

donde $g(r)$ es la función de detección y $\nu \coloneqq \int_0^w 2\pi r g(r)\, dr$. Por supuesto $g(r)$ y ν dependen de un vector de parámetros desconocidos $\boldsymbol{\theta}$ que debemos estimar a partir de la muestra de radios, denotada por $r_1, \ldots, r_n$. Para desarrollar ahora la función de verosimilitud de $\boldsymbol{\theta}$ definimos primero a $f(x)$ como:

$$\begin{aligned} f(r) \coloneqq \mathbb{P}(\text{radio } r \mid \text{detección}) &= \frac{\mathbb{P}(\text{detección} \mid \text{radio } r) \cdot \pi(r)}{\mathbb{P}(\text{detección})} \\ &= \frac{g(r) \cdot \frac{2\pi r}{\pi w^2}}{\int_0^w g(r) \cdot \frac{2\pi r}{\pi w^2}\, dr} = \frac{2\pi r g(r)}{\nu} \end{aligned} \tag{2.3.2}$$

entonces, si los objetos son detectados independientemente, análogo a la ecuación (2.2.4) tenemos que la función de verosimilitud de $\boldsymbol{\theta}$ es:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i) = \left(\frac{2\pi}{\nu}\right)^n \prod_{i=1}^{n} r_i g(r_i) \tag{2.3.3}$$

Al igual que con transectos lineales podemos maximizar 2.3.3 y obtener el EMV $\hat{\boldsymbol{\theta}}$. Posteriormente calculamos $\hat{\nu} = \int_0^w 2\pi r \hat{g}(r)\, dr$ y $\hat{P}_a = \frac{\hat{\nu}}{\pi w^2}$, de donde se sigue que:

$$\hat{N} = \frac{n}{\hat{P}_a \cdot \pi_c} = \frac{nA}{\hat{P}_a \cdot a} = \frac{nA}{\frac{\hat{\nu}}{\pi w^2} \cdot k\pi w^2} = \frac{nA}{k\hat{\nu}} \tag{2.3.4}$$

Notemos que con k transectos, usando 2.3.1 se tiene que $a \cdot \hat{P}_a = k\pi w^2 \cdot \frac{\hat{\nu}}{\pi w^2} = k\hat{\nu}$ incluso si $w = \infty$. Notamos que mientras con transectos lineales $f(x)$ era proporcional a $g(x)$, en muestreo con transectos puntuales $f(r)$ es proporcional a $rg(r)$.

2.3.3. Estimación la función de detección

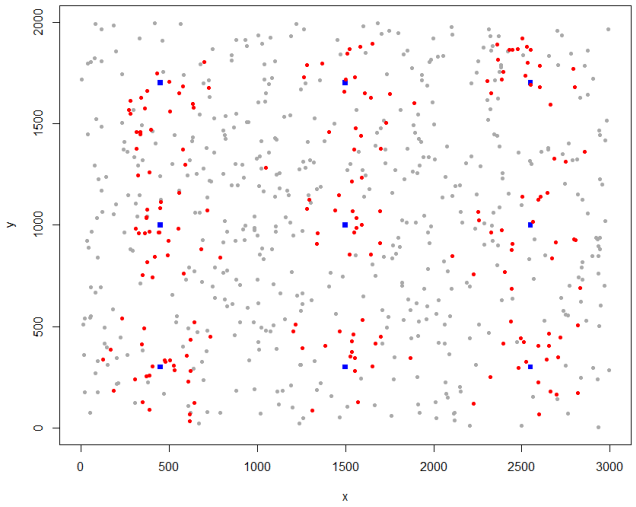

Figura 2.20: Población simulada con N=700 individuos y $k = 9$ transectos puntuales (cuadros), se detectan $n = 192$ individuos (puntos rojos). El área de la región de estudio es A=$6,000,000u^2$

La estimación de la función de detección y de la función de densidad de

distancias para el muestreo por transectos puntuales se llevará a cabo de forma casi idéntica que con transectos lineales, en este caso utilizaremos los resultados de la sección 2.3.2.

A lo largo de esta subsección se usarán los datos de la población simulada de la figura 2.20. Comenzamos por graficar el histograma de distancias detectadas (figura 2.21).

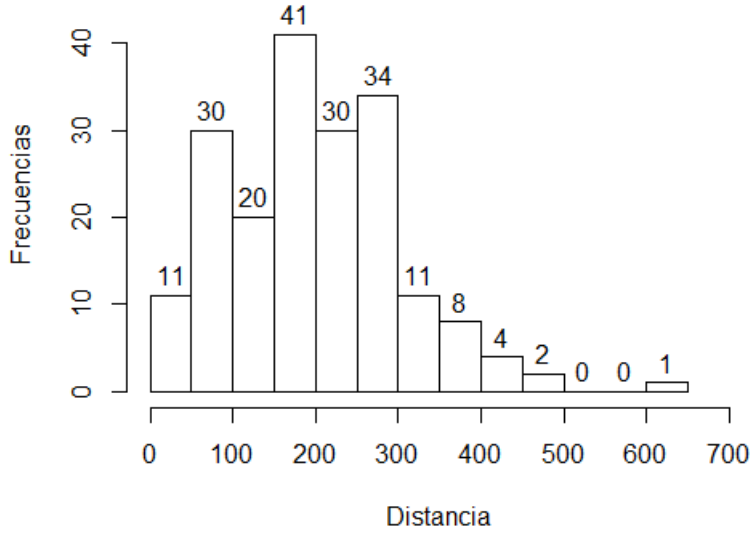

Figura 2.21: Histograma de las distancias detectadas usando transectos puntuales para estudiar la población simulada. Muestra un comportamiento diferente al muestreo con transectos lineales.

En el muestreo por transectos lineales las frecuencias de las distancias observadas tenían un comportamiento decreciente prácticamente en todo el rango de la muestra. Al estudiar la población simulada usando transectos puntuales, observamos que las frecuencias se incrementan hasta 200 metros y después decrecen conforme aumenta la distancia. Este comportamiento es natural al usar transectos puntuales. A medida que el radio a partir del transecto crece habrá más individuos para ser detectados, lo cual explica la tendencia en principio creciente del histograma. La parte decreciente del histograma, al igual que con

transectos lineales, se debe a la dificultad de detección de los individuos lejanos al transecto.

Con los resultados de la sección anterior ajustaremos a nuestros datos una función de detección half-normal, la cual sabemos está dada por $g(r)=\exp(-r^2/2\sigma^2)$, $0 \leq r < \infty$, $\sigma > 0$. Para obtener explícitamente el EMV para σ^2, comenzamos por calcular ν:

$$\begin{aligned}\nu &= \int_0^w 2\pi r g(r)\, dr = 2\pi \int_0^w r \cdot \exp\left(\frac{-r^2}{2\sigma^2}\right) dr \\ &= \left[-2\pi\sigma^2 \exp\left(\frac{-r^2}{2\sigma^2}\right)\right]\Bigg|_0^w \\ &= 2\pi\left(1 - \exp\left(\frac{-w^2}{2\sigma^2}\right)\right)\end{aligned}$$

suponiendo $w = \infty$ entonces $\nu = 2\pi\sigma^2$. Si se detectan independientemente n individuos, usando la ecuación (2.3.3) tenemos:

$$L(\sigma)=\prod_{i=1}^{n} f(x_i) = \left(\frac{2\pi}{\nu}\right)^n \prod_{i=1}^{n} r_i \exp\left(\frac{-r_i^2}{2\sigma^2}\right) = \left(\frac{1}{\sigma^{2n}}\right) \prod_{i=1}^{n} r_i \exp\left(\frac{-r_i^2}{2\sigma^2}\right)$$

de donde:

$$l(\sigma) = \ln(L(\sigma)) = \sum_{i=1}^{n}\left(\ln(r_i) - \frac{r_i^2}{2\sigma^2}\right) - n\ln(\sigma^2)$$

derivando con respecto a σ^2 e igualando a cero:

$$\frac{d(l(\sigma))}{d\sigma^2} = \sum_{i=1}^{n} \frac{r_i^2}{2\sigma^4} - \frac{n}{\sigma^2} = 0$$

y se sigue que $\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{r_i^2}{2n}$ y $\hat{\nu} = 2\pi\left(\sum_{i=1}^{n} \frac{r_i^2}{2n}\right)$. Con los datos de la población simulada y usando la ecuación (2.3.4) obtenemos:

$$\hat{N} = \frac{nA}{k\hat{\nu}} = \frac{192 \cdot 6,000,000}{9 * 154,798.5} = 826.88 \approx 827$$

La figura 2.22 muestra el ajuste del modelo half-normal de las funciones de detección y densidad.

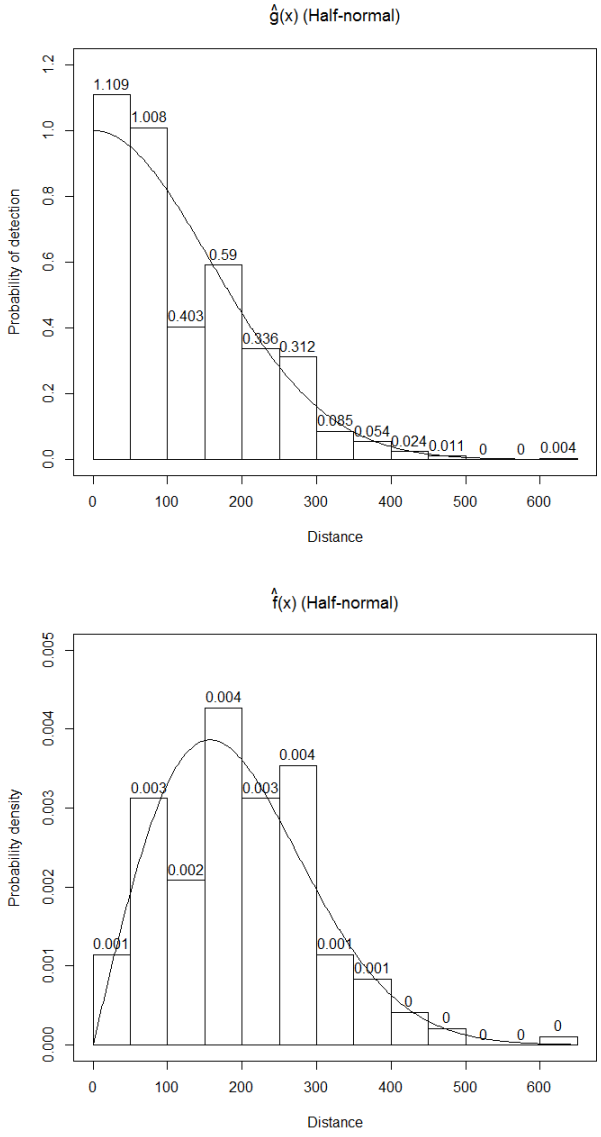

Figura 2.22: Funciones de detección (arriba) y densidad (abajo) ajustadas con el modelo half-normal.

Por la naturaleza del muestreo por transectos puntuales sabemos que las frecuencias de individuos observados es creciente para distancias pequeñas debido al crecimiento en área al alejarnos del transecto. Para examinar el ajuste de la función de detección estimada $\hat{g}(x)$, debemos enfocarnos solamente en cómo

decrece la probabilidad de detección conforme aumenta la distancia sin tomar en cuenta el incremento natural de las frecuencias debido al aumento de área. Por lo que para visualizar el ajuste de la función de detección estimada debemos "corregir" por el incremento de área, asignando a cada distancia r_i un peso de $\frac{1}{r_i}$ y graficando un histograma ponderado en donde la altura del histograma para cada intervalo de distancias, está dada por la suma de los pesos de las distancias en ese intervalo. Por último, este histograma ponderado se cambia de escala al igual que con transectos lineales, para que su área sea igual al área bajo $\hat{g}(x)$. La función de densidad estimada $\hat{f}(x)$, se grafica con el histograma de densidad de las distancias. El ajuste de ambas funciones se muestra en la figura 2.22, se observa la diferencia en la forma de $\hat{f}(x)$ y $\hat{g}(x)$ derivada de que estas funciones no son proporcionales como en el caso del muestreo por transectos lineales.

Ajustamos ahora el modelo hazard-rate a los datos para posteriormente poder comparar ambos ajustes. La función de detección hazard-rate está dada por $g(r) = 1 - \exp\left(-\left(\frac{r}{\sigma}\right)^{-b}\right)$, en donde b es un parámetro de forma y σ es un parámetro de escala. Sabemos que los estimadores máximo verosímiles para r y b no tienen una expresión cerrada por lo que se calculan mediante un software estadístico. Suponiendo $w = \infty$, para los datos de la población simulada obtenemos que $\hat{\sigma}$=239.67, $\hat{b} = 5.86$ y $\hat{\nu} = 244,465.3$.

Sustituyendo en la ecuación (2.3.4) obtenemos:

$$\hat{N} = \frac{nA}{k\hat{\nu}} = \frac{192 \cdot 6,000,000}{9 * 244,465.3} = 523.59 \approx 524$$

En este ejemplo los valores estimados para la abundancia son relativamente lejanos a la abundancia real de 700 individuos, lo cual se puede atribuir a que solo se simuló el estudio con 9 transectos y tanto el tamaño de muestra como la región cubierta son pequeños. El ajuste del modelo hazard-rate se muestra en la figura 2.23.

En la siguiente sección ocuparemos las herramientas ya conocidas para seleccionar alguno de los dos modelos.

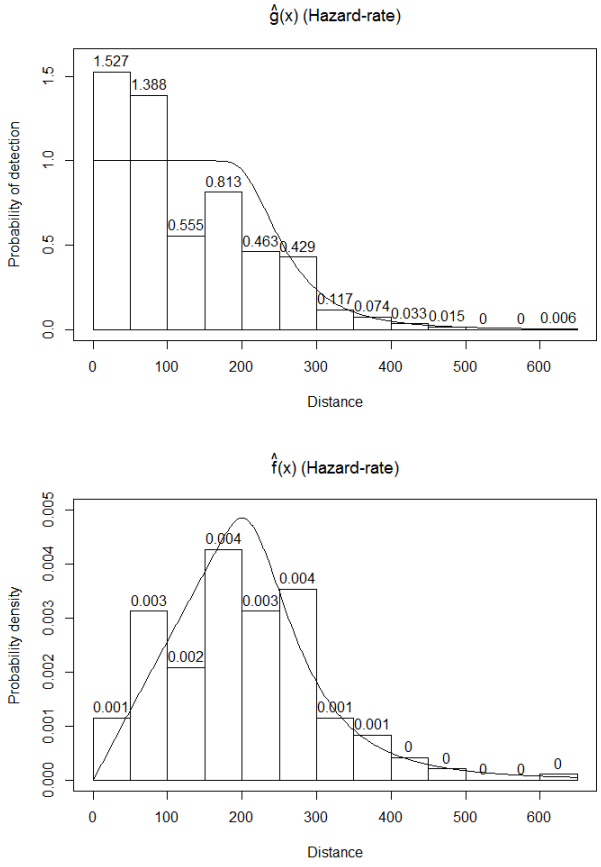

Figura 2.23: Funciones de detección (arriba) y densidad (abajo) ajustadas con el modelo hazard-rate.

2.3.4. Elección del modelo

Para elegir un modelo procederemos de igual forma que con transectos lineales, ya que las pruebas y criterios de la sección 2.2.4 se implementan de la misma manera al muestreo con transectos puntuales, la teoría y descripción de cada método puede ser revisada en esa sección. Nos enfocaremos principalmente en mostrar resultados de dichos métodos aplicados a la simulación ejemplo.

Gráficos P-P

Como primera inspección gráfica, comparamos las funciones $\hat{F}(x)$ y la función de distribución empírica de la muestra mediante gráficos P-P, para los modelos half-normal y hazard-rate ajustados en la sección anterior (figura 2.24).

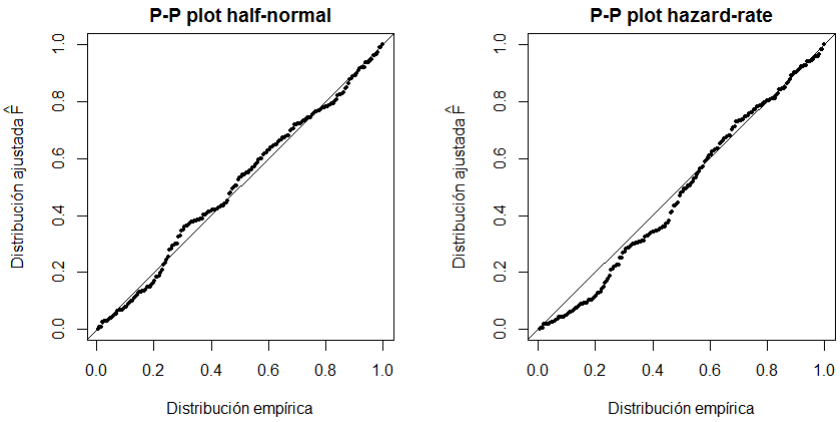

Figura 2.24: Gráficos P-P para ambos modelos ajustados.

Este método gráfico indica un mejor ajuste del modelo half-normal por presentar menos y menores desviaciones, el modelo hazard-rate parece tener un ajuste pobre para distancias cercanas a cero.

Prueba de Kolmogorov-Smirnov

Basada en la mayor diferencia absoluta entre la función de distribución empírica de la muestra $S(x)$ y $\hat{F}(x)$, esta prueba para bondad de ajuste sirve como complemento a la información obtenida del gráfico P-P.
Con el modelo half normal obtenemos un valor de la estadística de prueba de 0.057 y un p-value de 0.55. Para el modelo hazard-rate obtenemos 0.087 y 0.10 como estadística de prueba y p-value respectivamente.
Para ningún modelo se rechaza la hipótesis nula de esta prueba, pero el p-value considerablemente mayor con el modelo half-normal sugiere un mejor ajuste, confirmando lo visto en el gráfico P-P.

Prueba de Cramér-von Mises

La prueba de Cramér-von Mises produce para el modelo half-normal una estadístca de prueba y un p-value de 0.10 y 0.58. Para el modelo hazard-rate la estadística de prueba es 0.29 y el p-value 0.13. No rechazamos la hipótesis nula en ambos casos. Una vez más los valores obtenidos apuntan al modelo half-normal como mejor opción.

Prueba de la χ^2

Siguiendo el procedimiento visto para esta prueba en la sección 2.2.4, separamos la muestra de distancias $r_1, \ldots, r_n$, en u=5 celdas definidas por los puntos de corte $c_0 = 0, \ldots, c_5 = w$, de tal forma que $\hat{F}(c_i) = i*0.20$, para $i = 0, \ldots, 5$. Las frecuencias absolutas por celda serán $n_1, \ldots, n_5$, y los puntos de corten dependen por supuesto del modelo para el cual hagamos la prueba. Por la elección de los puntos de corte, las celdas serán en este caso equiprobables bajo la hipótesis nula. Para probar ambos modelos se usa un nivel de significancia de α=0.05.
Para el modelo half-normal obtenemos un p-value de 0.1, por lo que no se rechaza la hipótesis nula con este modelo. La prueba con el modelo hazard-rate produce un p-value de 0.024, por lo que se rechaza la hipótesis nula de que la muestra $r_1, \ldots, r_n$ tenga la distribución hazard-rate ajustada.

Criterio de Información de Akaike

Por último calculamos el AIC de ambos modelos. Para el modelo half-normal obtenemos AIC_{HN}=2306.35 y para el hazard-rate AIC_{HR}=2314.71.
En este caso todos las pruebas y criterios sugieren elegir el modelo half-normal para el muestreo simulado.

2.3.5. Estimación por intervalos

Por las razones expuestas en la sección 2.2.4, al igual que con transectos lineales, se usa el bootstrap no-paramétrico para calcular los intervalos de confianza. El procedimiento es idéntico al de transectos lineales, usando como unidad de muestra a los transectos puntuales. Aunque mencionamos que debemos selec-

cionar el modelo half-normal, obtenemos el intervalo de confianza también para el modelo hazard-rate para comparar.

Para el modelo half-normal, el bootstrap no-paramétrico con 5,000 remuestreos, da un intervalo del 95 % de confianza de (757.97, 900.09) con $\widehat{Var}_{BN}(\widehat{N}) = 1397.54$, con el modelo hazard-rate obtenemos (479.95, 572.67) y $\widehat{Var}_{BN}(\widehat{N}) = 570.25$. Notamos que en ambos casos el intervalo obtenido no contiene al verdadero valor de la abundancia. Para una estimación por intervalos confiable se busca un diseño con al menos 15 transectos.

Capítulo 3

Vecino más cercano

3.1. Introducción

En el método del vecino más cercano suponemos que el número de individuos en una región de estudio de área A es una realización de una distribución Poisson con media AD, donde D se interpreta en este caso como el número promedio de individuos por unidad de área. También supondremos que los individuos se distribuyen uniformemente y de forma independiente en la región de estudio.
La distribución Poisson es usada en este método por convención, conveniencia y por la consecuencia de algunos supuestos básicos. En ecología, es común interpretar la aleatoriedad con una distribución Poisson. La conveniencia será aparente en la siguiente sección al desarrollar el estimador para la abundancia. Si la ocurrencia de un individuo en una región dentro del área de estudio la consideramos como un ensayo Bernoulli con una probabilidad de éxito pequeña, entonces el número de ocurrencias se aproxima a una distribución Poisson.
Para obtener la muestra se eligen J puntos aleatoriamente dentro de la región de estudio y para cada uno de ellos se mide la distancia al individuo más cercano, obteniendo así una muestra de distancias $x_1, \ldots x_J$. El método se extiende naturalmente al caso en donde se mida la distancia a los K-vecinos más cercanos. En áreas de alta densidad este enfoque con $K > 1$, da una estimación más eficiente.
El método del vecino más cercano es raramente utilizado en la práctica por varias razones:

- Si existen áreas escasamente pobladas, puede ser complicado o ineficiente

localizar al individuo más cercano.

- El método es ineficiente porque al intentar localizar al individuo más cercano tenemos la oportunidad de probablemente localizar a muchos más individuos, es decir, recopilamos poca información a un gran costo.
- Para poblaciones móviles este método es muy propenso a obtener estimadores sesgados.
- El supuesto de uniformidad suele ser poco realista en la práctica y la estimación con este método es muy sensible al incumplimiento de este supuesto.

En la siguiente sección veremos que en efecto, si la distribución real de los individuos está más agrupada que en un patrón aleatorio, la abundancia será subestimada.

3.2. Estimación por máxima verosimilitud

En primera instancia el método del vecino más cercano parece similar al muestreo por parcelas, pensando cada parcela como un círculo con radio igual a la distancia al individuo más cercano. La complicación, a diferencia del muestreo por parcelas, es que la distancia al vecino más cercano a partir de un punto aleatorio es una variable aleatoria.

Por lo tanto debemos obtener una función de verosimilitud ocupando el supuesto de la distribución Poisson para el número de individuos en el área de estudio.

Sea X la variable aleatoria de la distancia al objeto más próximo. Supongamos que se elige aleatoriamente un punto dentro de la región de estudio y se mide la distancia x entre este punto y el individuo más cercano. Veremos que podemos obtener la función de densidad de X y usar la observación x para estimar D.

Notemos que debido al supuesto de uniformidad, el número de individuos en un área πx^2 sigue una distribución Poisson con media $\pi x^2 D$, entonces:

$$\mathbb{P}(X > x) = \mathbb{P}(\text{el área circular de radio } x \text{ alrededor del transecto, esté vacía})$$

$$\begin{aligned} &= \frac{\left(\pi x^2 D\right)^0 \exp\left(-\pi x^2 D\right)}{0!} \\ &= \exp\left(-\pi x^2 D\right) \end{aligned} \tag{3.2.1}$$

entonces:

$$F(x) = \mathbb{P}(X \leq x) = 1 - \exp\left(-\pi x^2 D\right) \tag{3.2.2}$$

derivando obtenemos la función de densidad:

$$f(x) = 2\pi x D \exp\left(-\pi x^2 D\right) \tag{3.2.3}$$

Si se eligen J puntos aleatorios y se obtienen las distancias $x_1, \ldots x_J$, la función de verosimilitud de D está dada por:

$$L(D) = \prod_{j=1}^{J} f(x_i) = (2\pi D)^J \left(\prod_{j=1}^{J} x_j\right) \exp\left(-\pi D \sum_{j=1}^{J} x_j^2\right) \tag{3.2.4}$$

de donde:

$$l(D) = \ln(L(D)) = J \ln(2\pi D) + \sum_{j=1}^{J} \ln(x_j) - \pi D \sum_{j=1}^{J} x_j^2$$

derivando e igualando a cero:

$$\frac{d(l(D))}{dD} = \frac{J}{D} - \pi \sum_{j=1}^{J} x_j^2 = 0$$

por lo tanto:

$$\hat{D} = \frac{J}{\pi \sum_{j=1}^{J} x_j^2} \tag{3.2.5}$$

y:

$$\hat{N} = \hat{D} \cdot A = \frac{JA}{\pi \sum_{j=1}^{J} x_j^2} \tag{3.2.6}$$

Si para cada uno de los J puntos se mide la distancia a los K individuos más cercanos obtenemos que:

$$\hat{N} = \frac{JKA}{\pi \sum_{j=1}^{J} \sum_{k=1}^{K} x_{jk}^2} \tag{3.2.7}$$

donde x_{jk} denota la distancia del punto aleatorio j al $k-$ésimo individuo más próximo.

Si la población de interés presenta una distribución agrupada, las distancias medidas a partir de los puntos aleatorios a los individuos más cercanas serán más grandes que las esperadas con una distribución uniforme, por lo que el denominador de la ecuación (3.2.6) será grande y la abundancia subestimada.

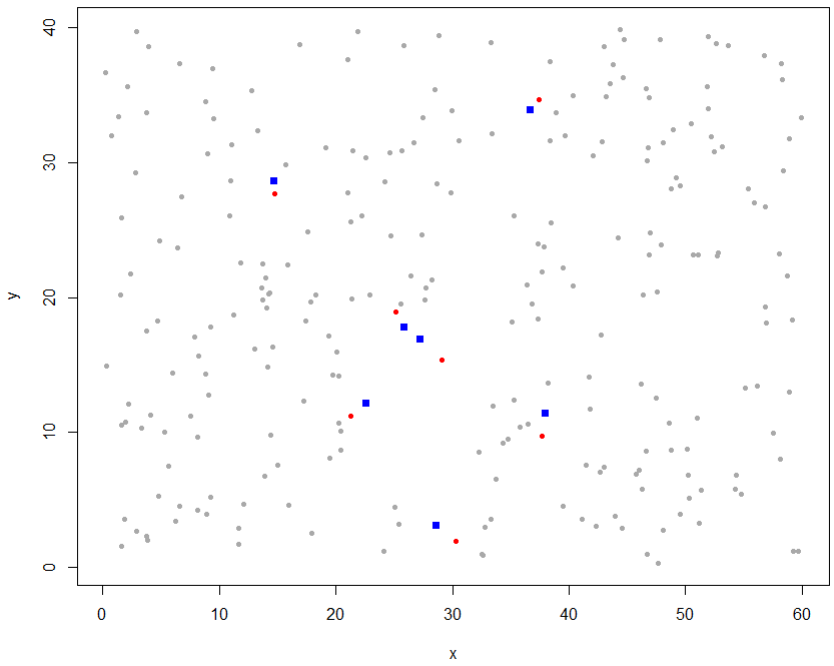

Figura 3.1: Población simulada con N=250 individuos, se muestran J=7 puntos aleatorios (cuadros) junto con los objeto más cercanos a ellos (puntos rojos).

Con la simulación de la figura 3.1 obtenemos $\hat{N}$=271.92 $\approx$ 272. En este caso el valor estimado es relativamente cercano ya que la distribución de los objetos es cercana a la uniforme.

3.3. Estimación por intervalos

El **bootstrap parámetrico** se implementa sustituyendo $\hat{D}$ por D en la ecuación (3.2.2) y simulando J observaciones de esta distribución, las cuales se sustituyen en la ecuación 4.2.6 para obtener un estimador bootstrap. Repitiendo este procedimiento B veces, obtenemos $\hat{N}_1, \ldots, \hat{N}_B$ y calculamos el intervalo del (1-α) $\times 100\,\%$ de confianza como $\left(\hat{N}_{(r)}, \hat{N}_{(s)}\right)$ donde $\hat{N}_{(r)}$ y $\hat{N}_{(s)}$ denotan los percentiles del $\frac{\alpha}{2} \times 100\,\%$ y del $\left(1 - \frac{\alpha}{2}\right) \times 100\,\%$.
Para simular las J observaciones de $\hat{F}(x) = 1 - \exp\left(-\pi x^2 \hat{D}\right)$ usamos el método de la transformada inversa, el cual consiste en generar J observaciones $(u_1, \ldots u_J)$ de una distribución uniforme en el intervalo [0,1] y haciendo $x_j = \hat{F}^{-1}(u_j) = \sqrt{\frac{-\ln(1-u_j)}{\pi \hat{D}}}$ para $j = 0, \ldots, J$, obtenemos $x_1, \ldots x_j$, las cuales serán observaciones simuladas de $\hat{F}(x)$.
Para la población de la figura 3.1, el intervalo del 95 % de confianza usando bootstrap paramétrico con B=10,000 es de (144.3, 683.8).

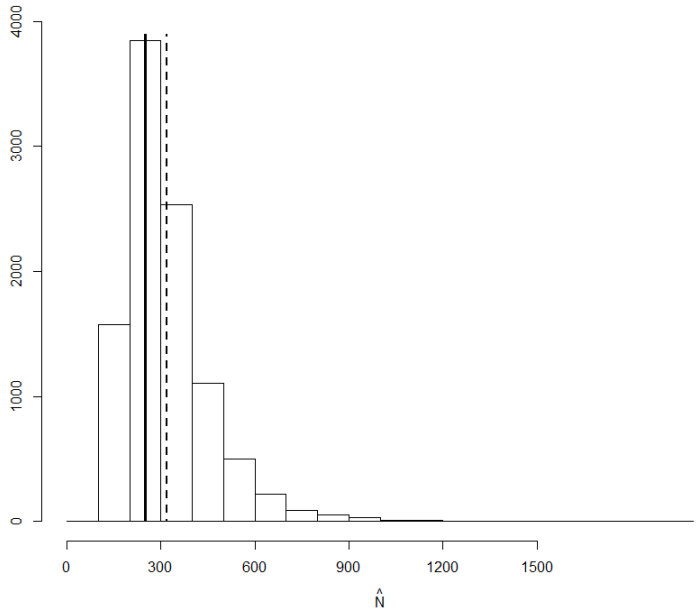

Figura 3.2: Distribución bootstrap del estimador $\hat{N}$. La línea negra indica el verdadero valor de N=250, la línea punteada es la media estimada de $\hat{N}$ (318.2).

Al igual que en muestreo por distancias, los intervalos de confianza obtenidos usando el **bootstrap no-paramétrico** son menos sensibles a poblaciones con distribución distinta a la uniforme. Lo anterior debido a que las remuestras son generadas sin usar el supuesto de uniformidad.

El bootstrap no paramétrico se implementa de manera muy similar a como se hace en el muestreo por transectos puntuales. Se obtienen B muestras con reemplazo de tamaño J de la muestra original de distancias $x_1, \ldots x_j$. Para cada muestra bootstrap se calcula $\hat{N}$ obteniendo $\hat{N}_1, \ldots, \hat{N}_B$. Posteriormente el intervalo del (1-α) $\times$ 100 % de confianza se obtiene de igual manera que en el bootrstrap paramétrico.

Para la población de la figura 3.1, el intervalo del 95 % de confianza usando bootstrap no-paramétrico con B=10,000 es de (187.5, 447.9), el cual es considerablemente más angosto que el obtenido con bootstrap paramétrico.

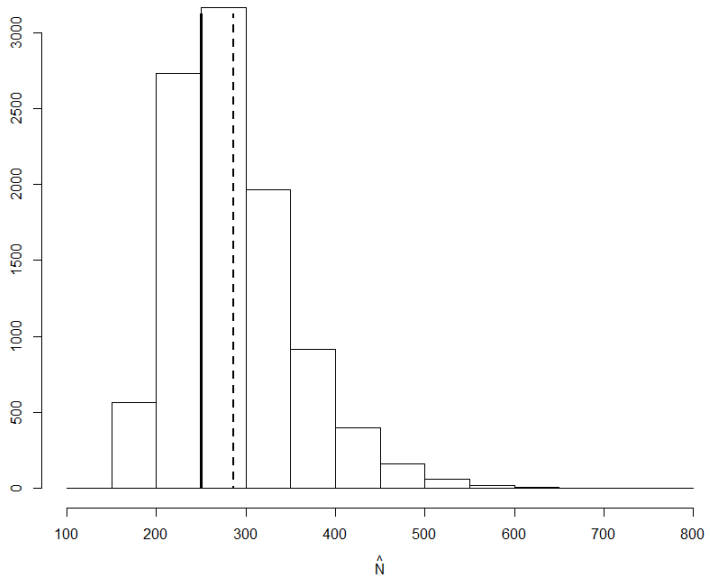

Figura 3.3: Distribución bootstrap (no-paramétrico) del estimador $\hat{N}$. La línea negra indica el verdadero valor de N=250, la línea punteada es la media estimada de $\hat{N}$ (284).

Capítulo 4

Aplicaciones

4.1. Introducción

En este capítulo aplicaremos los métodos presentados anteriormente a dos poblaciones simuladas, en las cuales la distribución de los individuos no es uniforme en el área de estudio. Al aplicar los métodos a las mismas poblaciones podremos comparar los estimadores de cada uno y observar cual es más robusto al violarse el supuesto de uniformidad. Las aplicaciones sobre poblaciones simuladas nos permiten también obtener la distribución aproximada de los estimadores al simular el estudio un gran número de veces, lo que permite medir la proporción en la que subestimamos o sobreestimamos la abundancia.

4.2. Primera aplicación

La primera población simulada presenta una tendencia en la distribución de los individuos sobre el área de estudio, la abundancia real es de N=1598 y el área de estudio es A=3.84 Km^2. Se muestra esta población en la figura 4.1.

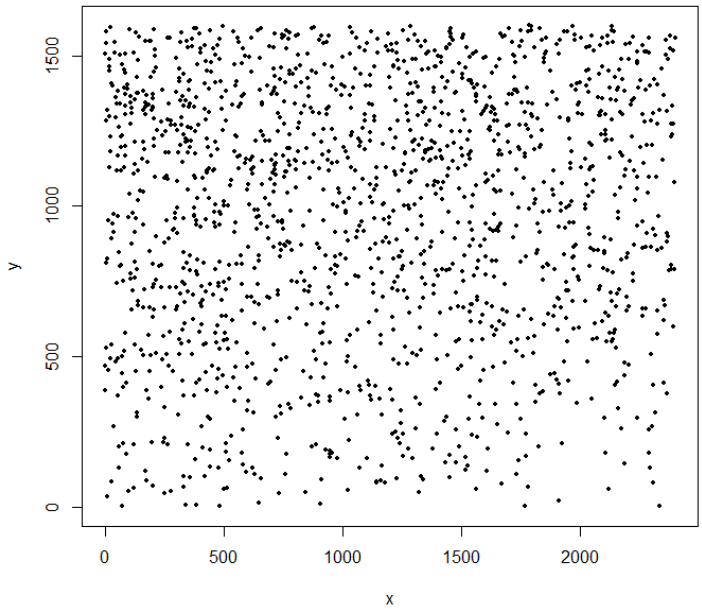

Figura 4.1: Población simulada con tendencia sureste-noroeste y N=1598 individuos.

4.2.1. Muestreo por parcelas

Para aplicar el método de muestreo por parcelas se seleccionan aleatoriamente 21 parcelas rectangulares cada una con 38,400m^2 de superficie, por lo que $\pi_c = \frac{a}{A}$=.21. Se cuentan $n = 369$ individuos dentro de las parcelas (figura 4.2) . El valor estimado para la abundancia es:

$$\hat{N} = \frac{n}{\pi_c} = \frac{369}{.21} \approx 1757$$

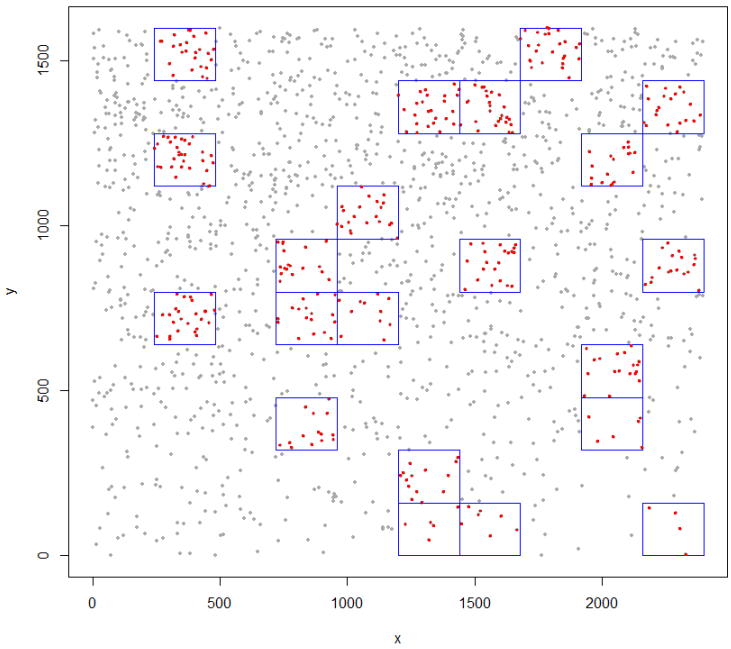

Figura 4.2: Muestreo por parcelas. La región cubierta representa el 21 % del área de estudio y se detectan 369 individuos.

A pesar de que los individuos no se distribuyen uniformemente sobre el área de estudio, la selección aleatoria de las parcelas permite obtener un valor estimado cercano al de la abundancia real de 1598.

Utilizando los procedimientos de la sección 1.3 podemos obtener los distintos intervalos de confianza. En todos los casos el nivel de confianza será del 95 %. El intervalo exacto para la población simulada es (1606, 1905). Suponiendo la normalidad del estimador $\hat{N}$, obtenemos el intervalo (1598, 1917). Haciendo bootstrap paramétrico, suponiendo que $n \sim \text{Bin}(\hat{N}, \pi_c)$, obtenemos con el método de percentiles el intervalo (1609, 1919). En cualquiera de los tres intervalos anteriores suponemos que la distribución de los individuos es uniforme, lo cual no se cumple para la población simulada.

En la sección 1.3 mencionamos que el bootstrap no-paramétrico obteniendo muestras con reemplazo de una población con $\hat{N}$ individuos y el bootstrap paramétrico producen aproximadamente el mismo intervalo de confianza. Alter-

nativamente podemos hacer bootrstrap no-paramétrico tomando como unidad muestral a las parcelas, de esta manera se toma en cuenta la varianza adicional de la distribución no uniforme de los individuos, por lo que es una estimación más robusta. Realizando el bootstrap no-paramétrico de esta forma, obtenemos mediante el método de percentiles, el intervalo (1433, 2057). Si la distribución de los individuos fuera uniforme, el número de individuos en cualquier parcela sería aproximadamente el mismo y el intervalo obtenido mediante bootstrap no-paramétrico sería similar al del bootstrap paramétrico.

De la ecuación (1.3.2) podemos obtener un valor estimado para la desviación estándar de $\hat{N}$ como $\widehat{SE}(\hat{N}) = \sqrt{\frac{\hat{N}\pi_c(1-\pi_c)}{\pi_c^2}} = 77.53$. Claramente este estimador tiene el mismo inconveniente que los intervalos de confianza exactos, normales y de bootstrap paramétrico, por lo tanto es recomendable estimar la desviación estándar mediante el bootstrap no-paramétrico, en este caso $\widehat{SE}_{BN}(\hat{N}) = 165.1$.

Sabemos de la ecuación (1.3.1) que el estimador $\hat{N}$ es insesgado cuando la distribucíón de los individuos es uniforme, para el caso de nuestra población podemos estimar el sesgo simulando observaciones de $\hat{N}$.

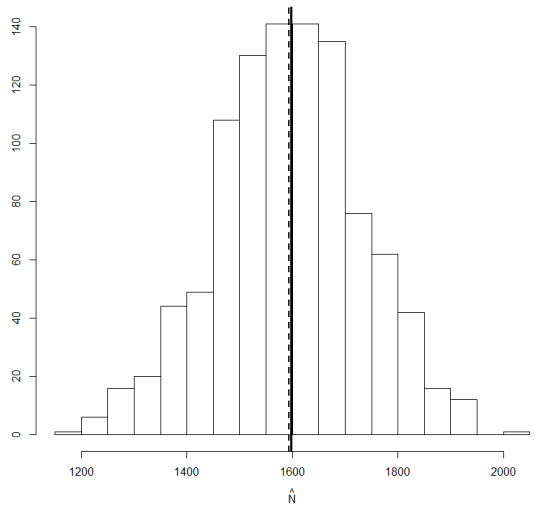

Figura 4.3: Distribución del EMV $\hat{N}$, obtenida de 1,000 simulaciones de muestreo por parcelas. La línea sólida es la abundancia real (1,598) y la línea punteada es la media de las observaciones simuladas (1,594).

La figura 4.3 muestra que a pesar de que los individuos no se distribuyen

uniformemente, el estimador basado en este supuesto es en promedio menor a la abundancia real por sólo 4 individuos.

4.2.2. Muestreo por distancias

4.2.2.1. Muestreo por transectos lineales

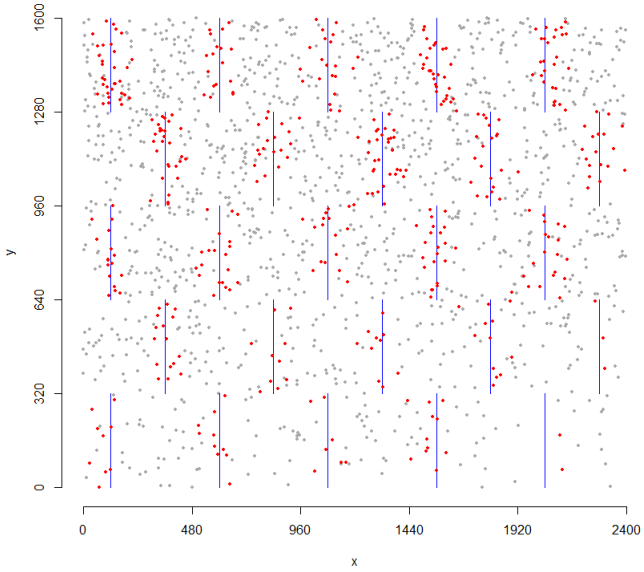

Figura 4.4: Muestreo por transectos lineales con k=25 transectos representados por las líneas azules, los puntos rojos son los individuos detectados.

Para la aplicación del muestreo por transectos lineales, se usa una red de k=25 transectos lineales igualmente espaciados que cubren el área de estudio, todos con una longitud de 320 metros por lo que la longitud total es de L=8,000 metros. Se detectan $n = 433$ individuos. Lo anterior se muestra en la figura 4.4.

Ajustando el modelo half-normal, se obtiene $\hat{\sigma} = 50.18$, $\hat{\mu} = 62.7$, con lo que finalmente $\hat{N} \approx 1658$.

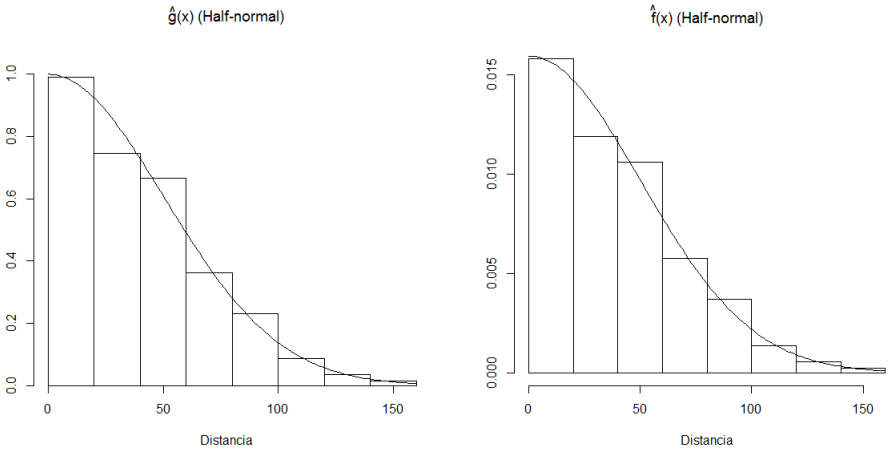

Figura 4.5: Histograma de las distancias detectadas mediante transectos lineales con el ajuste half-normal de la función de detección ajustada (izquierda) y la función de densidad ajustada (derecha).

Para el ajuste del modelo hazard-rate obtenemos $\hat{\sigma} = 58.5$, $\hat{b} = 3.41$, $\hat{\mu} = 72.7$ y $\hat{N} \approx 1429$.

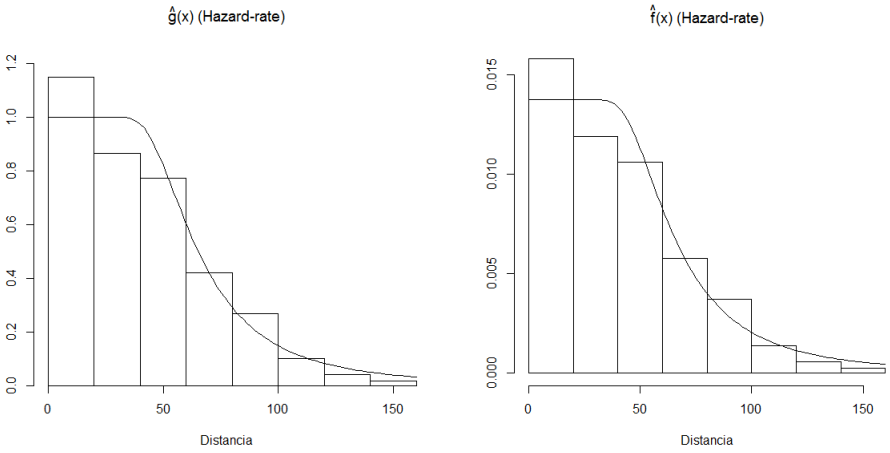

Figura 4.6: Histograma de las distancias detectadas con el ajuste hazard-rate de la función de detección ajustada (izquierda) y la función de densidad ajustada (derecha).

El ajuste de ambos modelos se muestra en la figuras 4.5 y 4.6.
Para seleccionar un modelo utilizamos los criterios vistos en la sección 2.2.4.

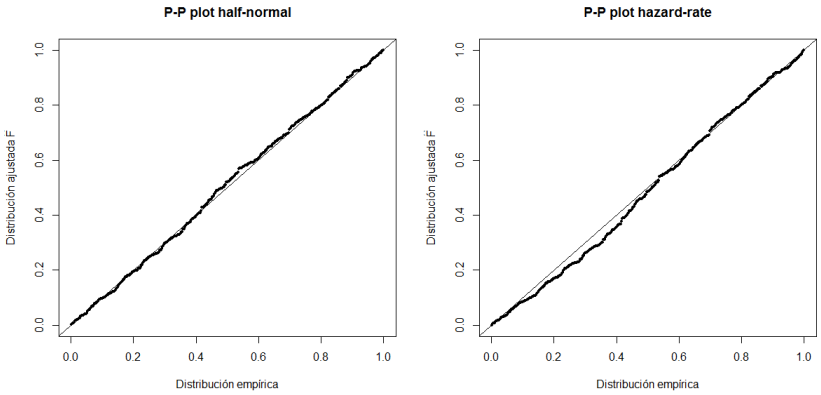

Figura 4.7: Gráficos P-P del ajuste de ambos modelos.

La figura 4.7 muestra un mejor ajuste del modelo half-normal. El modelo hazard-rate presenta más desviaciones contra los datos.

Con la prueba de Kolmogorov-Smirnov para el modelo half-normal obtenemos un p-value de 0.95 y una estadística de prueba de 0.02. Para el modelo hazard-rate obtenemos 0.15 y 0.05 respectivamente. Tomando $\alpha = 0.05$, para ningún modelo rechazamos la hipótesis nula, pero el menor valor de la estadística de prueba indica menores desviaciones en el ajuste del modelo half-normal.

La prueba de Cramér-von Mises nos da un p-value de 0.96 para el modelo half-normal, y 0.21 para el modelo hazard-rate. Con $\alpha = 0.05$, no se rechaza la hipótesis nula en ambos casos, el mayor p-value del modelo half-normal indica un mejor ajuste.

Realizando la prueba de la Ji-cuadrada para el modelo half-normal con $u = 5$ celdas equiprobables se tiene que $T \sim \chi^2_{(3)}$, en este caso T=0.98. Tomando $\alpha = 0.05$, el cuantil $1 - \alpha$ de la distribución de T es $W^{(3)}_{1-\alpha}$=7.81 por lo que no se rechaza la hipótesis nula. Para el modelo hazard-rate, $T \sim \chi^2_{(2)}$y se obtiene T=3.91, el p-value es de 0.14 por lo que tampoco rechazamos la hipótesis nula. El valor más pequeño de la estadística de prueba para el modelo half-normal sugiere una vez más un mejor ajuste.

Finalmente obtenemos el AIC para ambos modelos: $\text{AIC}_{HN} = 4018.6$ y $\text{AIC}_{HR} = 4026.4$. Con el menor valor del AIC, las pruebas anteriores y lo observado en los gráficos P-P , elegimos el modelo half-normal.

Una vez elegido el modelo, utilizamos bootstrap no-parámetrico sobre los transectos y el método de percentiles, el intervalo del 95 % de confianza es (1340,

1972). Con bootstrap no-paramétrico obtenemos también $\widehat{SE}_{BN}(\hat{N}) = 159.37$.

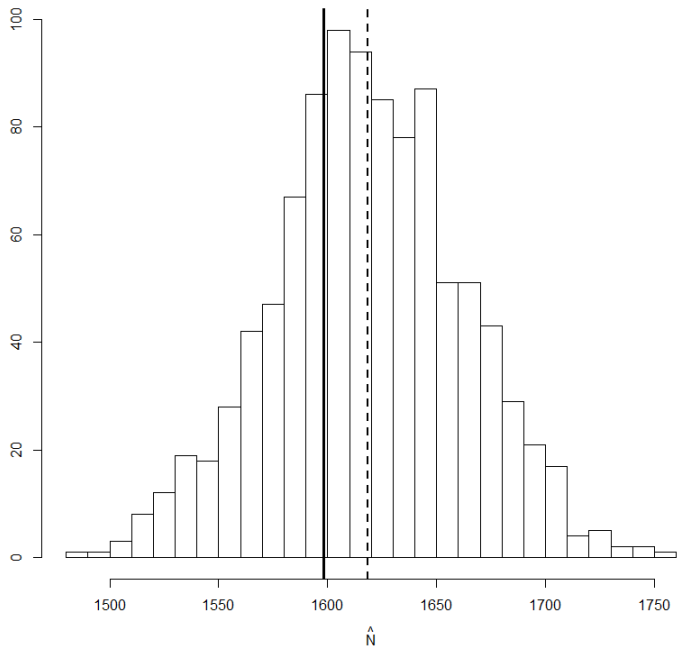

Figura 4.8: Distribución de $\hat{N}$, obtenida de 1,000 simulaciones de muestreo por transectos lineales. La línea sólida es la abundancia real (1,598) y la línea punteada es la media de la observaciones simuladas (1,618).

La figura 4.8 muestra que el EMV es cercano en promedio al verdadero valor de la abundancia. El sesgo aproximado del estimador es de 20 individuos.

4.2.2.2. Muestreo por transectos puntuales

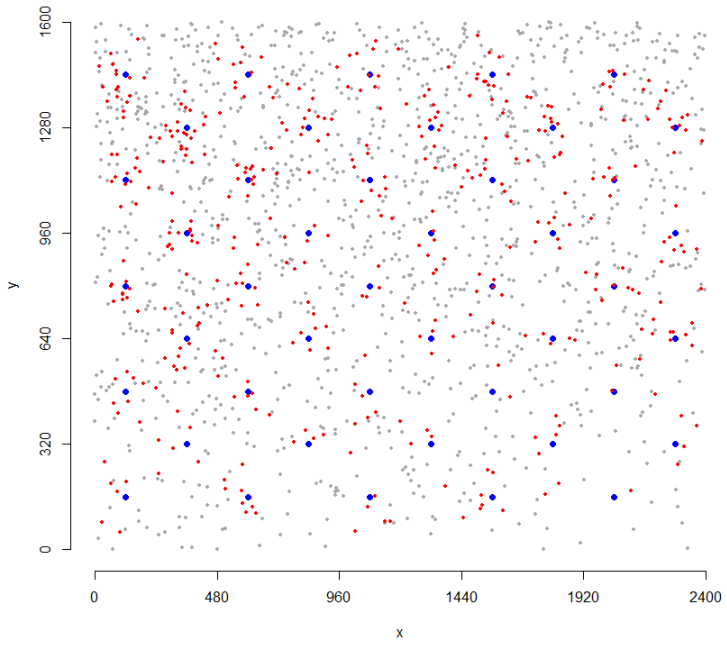

Figura 4.9: Localización de los k=45 transectos puntuales, los puntos rojos son los individuos detectados.

La figura 4.9 muestra una simulación del muestreo por transectos puntuales con una red de 45 transectos igualmente espaciados que cubren el área de estudio. Se detectan $n = 426$ individuos.

Haciendo el ajuste del modelo half-normal obtenemos: $\hat{\sigma} = 58.9$, $\hat{\nu} = 21,695.57$ y $\hat{N} \approx 1,658$. Ajustando el modelo hazard-rate obtenemos: $\hat{\sigma} = 70.34$, $\hat{b} = 3.73$, $\hat{\nu} = 26,554.72$ y $\hat{N} \approx 1,369$. El ajuste de ambos modelos se muestra en las figuras 4.10 y 4.11

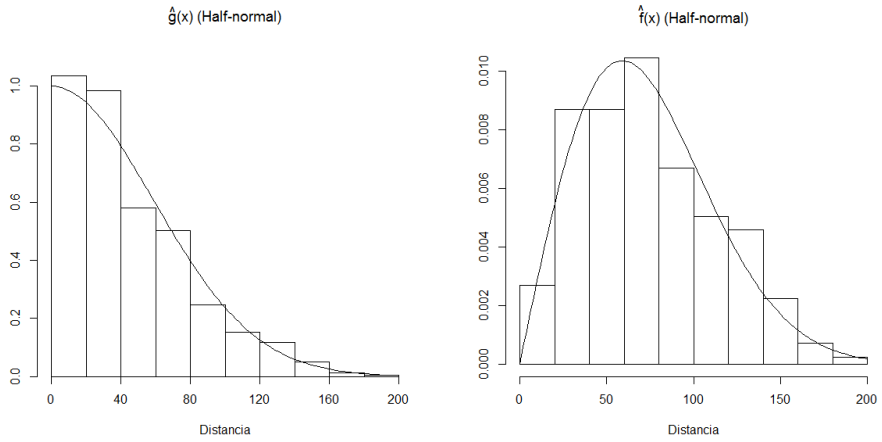

Figura 4.10: Histograma de las distancias detectadas mediante transectos puntuales con el ajuste half-normal de la función de detección (izquierda) y la función de densidad (derecha).

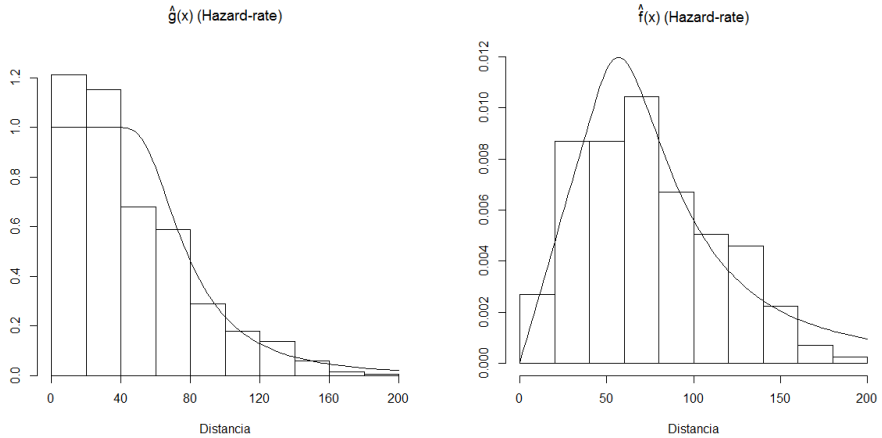

Figura 4.11: Histograma de las distancias detectadas mediante transectos puntuales con el ajuste hazard-rate de la función de detección ajustada (izquierda) y la función de densidad ajustada (derecha).

Los gráficos P-P mostrados en la figura 4.12 no proporcionan suficiente información para elegir un modelo, ambos ajustes presentan desviaciones y se deben analizar las otras pruebas de bondad de ajuste.

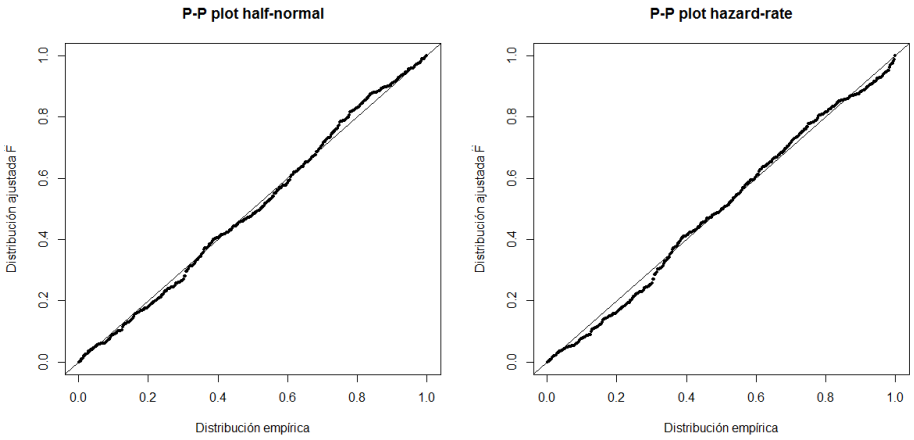

Figura 4.12: Gráficos P-P del ajuste de ambos modelos.

Se toma un nivel de signifcancia del 0.05 en todas las pruebas. Con la prueba Kolmogorov-Smirnov se obtienen los valores de p-value de 0.51 y 0.31 para el modelo half-normal y hazard-rate respectivamente, con la prueba de Cramér-von Mises se obtienen 0.43 y 0.32, y por último, en la prueba de la Ji-cuadrada con u=5 celdas equiprobables, obtenemos 0.06 y 0.003. Solamente se rechaza la hipótesis nula con el modelo hazard-rate en la prueba de la Ji-cuadrada.

Finalmente se obtiene un valor de 4,284 para el AIC del modelo half-normal, el cual es menor que el calculado para el modelo hazard-rate de 4,300. Se selecciona el modelo half-normal.

Con bootstrap no-paramétrico y el método de percentiles se obtiene un intervalo del 95 % de confianza de (1443, 1959), también se obtiene el valor estimado para la desviación estándar: $\widehat{SE}_{BN}(\hat{N}) = 125.84$.

La figura 4.13 muestra la distribución de $\hat{N}$ obtenida a partir de 1,000 simulaciones del muestreo por transectos puntuales. Se observa que el sesgo del estimador por transectos puntuales es mayor al de muestreo por parcelas y muestreo por transectos lineales. En este caso se sobreestima la población en promedio por 104 individuos.

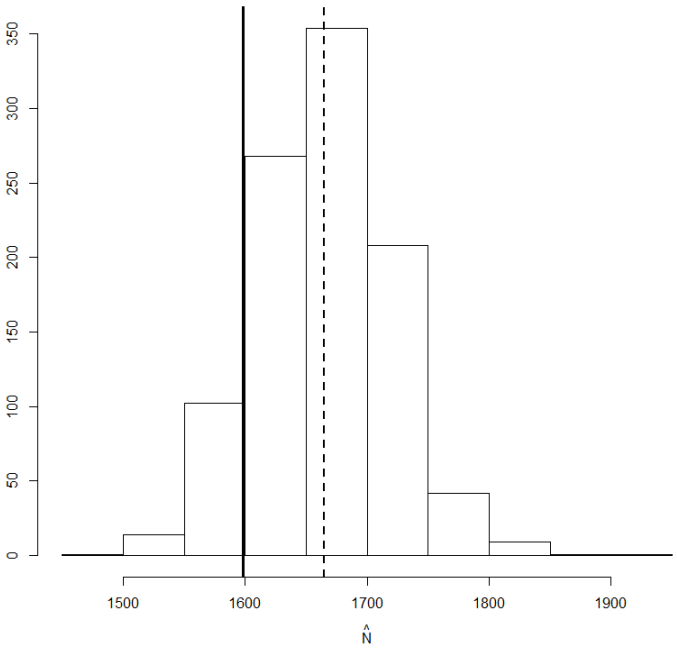

Figura 4.13: Distribución de $\hat{N}$, obtenida de 1,000 simulaciones de muestreo por transectos puntuales. La línea sólida es la abundancia real (1,598) y la línea punteada es la media de la observaciones simuladas (1,702).

4.2.3. Vecino más cercano

Para la aplicación de este método seleccionamos aleatoriamente 30 puntos dentro del área de estudio, y medimos la distancia al individuo más cercano (figura 4.14).

Usando las ecuaciones 3.2.5 y 3.2.6 obtenemos que $\hat{N} = 1,698$.

Una vez obtenido el valor estimado para la abundancia (y por tanto para la densidad) podemos calcular un intervalo de confianza simulando muestras de tamaño 30 de la función de distribución de la ecuación (3.2.2), es decir, usando bootstrap paramétrico. Con este procedimiento y el método de percentiles se calcula un intervalo del 95 % de confianza de (1,222, 2,542).

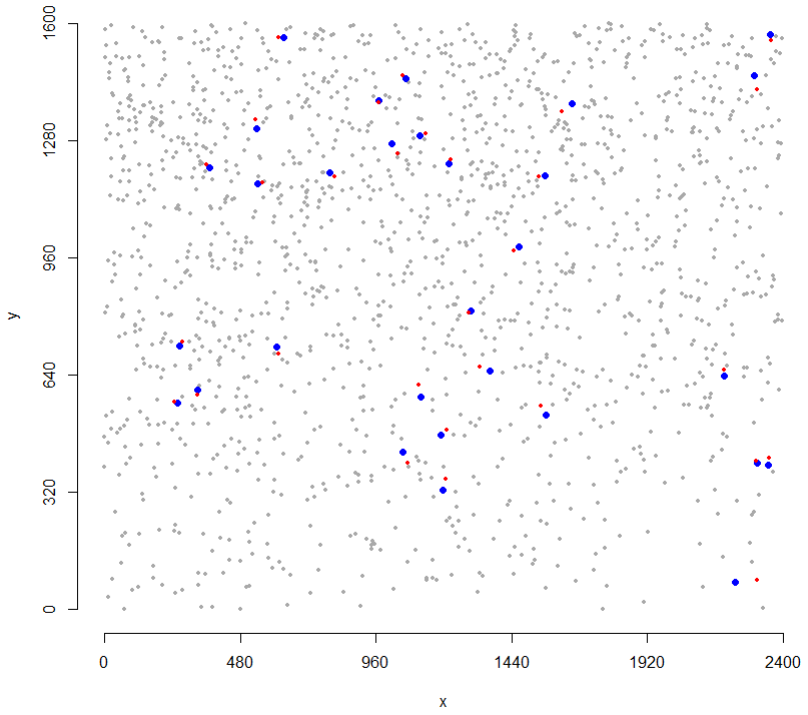

Figura 4.14: Población simulada con los 30 puntos aleatorios seleccionados (azul) y los individuos más cercanos (rojo).

Para el bootstrap-no paramétrico se generan muestras con reemplazo de tamaño 30, de la muestra original de distancias y se calcula para cada una el estimador para la abundancia. Se obtiene así, un intervalo del 95 % de confianza de (1,066, 2,884) y el valor estimado para la desviación estándar $\widehat{SE}_{BN}(\hat{N}) = 476.4$.
Para observar que tan sesgado está el estimador del vecino más cercano en esta población, simulamos el estudio 1000 veces, cada una por supuesto con 30 puntos aleatorios. La figura 4.15 muestra la distribución obtenida.
El estimador del vecino más cercano tiene un sesgo aproximado de 437 individuos, el cual es mayor a cualquiera de los métodos anteriores e indica que este método es muy sensible a la no uniformidad.

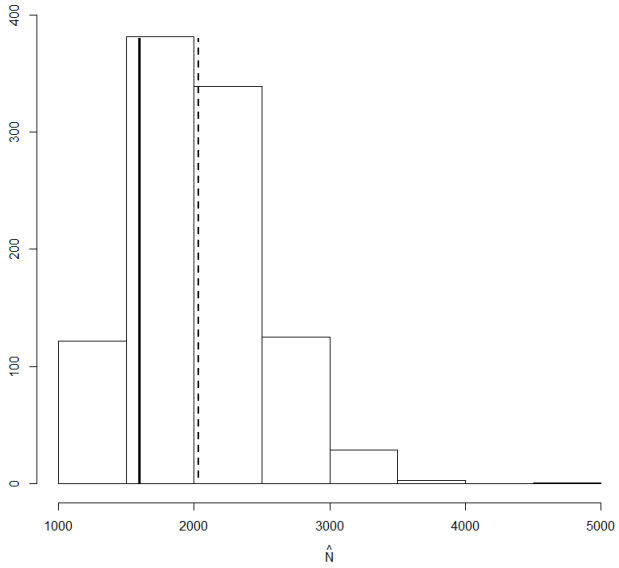

Figura 4.15: Distribución simulada del EMV para el método del vecino más cercano. La línea sólida representa el verdadero valor de la abundancia (1,598), la línea punteada es la media de las simulaciones (2,035).

4.2.4. Conclusiones

La siguiente tabla muestra un resumen con los valores estimados e intervalos de confianza obtenidos para la población simulada con cada método. Se muestran los intervalos del 95 % de confianza y el valor estimado de la desviación estándar, ambos calculados mediante bootstrap no-paramétrico. Para el muestreo por transectos lineales y transectos puntuales se muestra el valor estimado para la abundancia obtenido a partir del modelo elegido en 4.2.2.1 y 4.2.2.2. El sesgo estimado se calculó después de simular el estudio 1000 veces para cada método, claramente en la práctica no es posible calcularlo de esta manera. A partir de las simulaciones, podemos también calcular el coeficiente de variación estimado, $\widehat{CV}$, el cual es una medida relativa que muestra a la desviación estándar como proporción de la media y sirve para una mejor interpretación de la variabilidad del estimador. Se calcula como $\widehat{CV} = \frac{SE_S}{\bar{N}_S}$, en donde $\bar{N}_s$ y SE_S denotan la

media y desviación estándar de la simulaciones.

	Abundancia estimada	Intervalo de confianza	$\widehat{SE}_{BN}$	Sesgo estimado	$\widehat{CV}$
Muestreo por parcelas	1757	(1433, 2057)	165.1	-4	0.09
Transectos lineales	1658	(1340, 1972)	159.4	20	0.026
Transectos puntuales	1658	(1143, 1959)	125.8	104	0.04
Vecino más cercano	1698	(1066, 2884)	476.4	437	0.233

Podemos observar que los valores estimados para la abundancia en cualquiera de los métodos son relativamente cercanos al verdadero valor de 1598. El método de muestreo por parcelas tiene en promedio un sesgo menor que los demás métodos, pero en la práctica es complicado e ineficiente contar a todos los individuos dentro de una parcela. El sesgo menor de este estimador se debe a la selección aleatoria de las parcelas, ya que con esto, el estimador puntual no depende fuertemente de la uniformidad.

La estimación puntual del muestreo por transectos lineales coincide en este caso con la de transectos puntuales, los intervalos de confianza obtenidos también son similares pero el calculado para transectos lineales tiene menor longitud. El sesgo estimado sugiere que el muestreo por transectos lineales es más adecuado que el muestreo por transectos puntuales para esta población.

El método del vecino más cercano tiene el sesgo más grande de entre todos los métodos y la desviación estimada mayor, lo que indica que es el método más sensible a poblaciones no uniformes. También presenta el coeficiente de variación más grande, es decir, su variabilidad relativa es mayor y por supuesto ésta característica no es deseable. La mayor desviación estándar se deriva del hecho de solo medir la distancia al individuo más cercano, por lo que para diferentes estudios estas mediciones tienen mucha varianza cuando la población no es uniforme, por supuesto esto ocasiona que el intervalo de confianza sea muy ancho y de poca utilidad. El sesgo indica que sobreestimamos la abundancia en promedio por 437 individuos, lo cual sugiere que en el área de estudio existen más zonas con densidad mayor a la permitida por la distribución uniforme. Una población más escasamente distribuida produciría el efecto contrario.

Exceptuando el método del vecino más cercano, los métodos aplicados para la estimación dan resultados buenos a pesar de la no uniformidad de la población. Los resultados indican que los mejores métodos son, en este caso, el muestreo por parcelas y el muestreo por transectos lineales. La elección de un método dependerá de la naturaleza de la población, el terreno en donde se encuentra y los recursos disponibles para hacer el estudio entre otras variables, por ejemplo, el muestreo por transectos lineales requiere un menor esfuerzo para obtener la muestra que el muestreo por parecelas, además de que no tiene ninguna restricción acerca de los individuos que se deben detectar, por ello en la práctica es mucho más utilizado.

4.3. Segunda aplicación

En esta segunda aplicación aplicaremos los métodos sobre una población simulada con agrupaciones muy marcadas, el tamaño de la población es N=870 y el área de la región de estudio es $A = 1.5\text{Km}^2$ (figura 4.16).

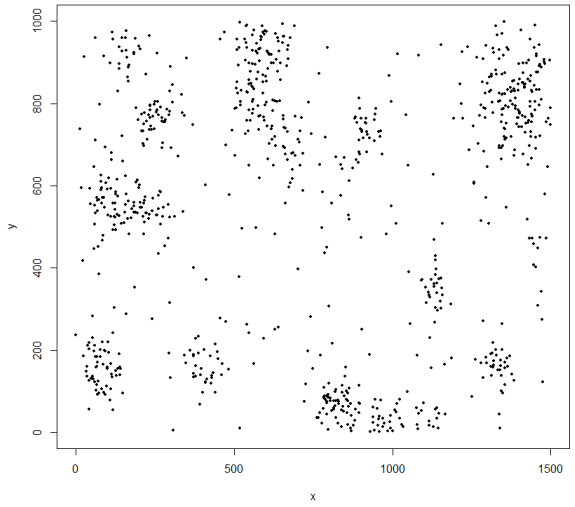

Figura 4.16: Población simulada con N=870 individuos.

4.3.1. Muestreo por parcelas

Se seleccionan aleatoriamente 19 parcelas rectangulares dentro de la región de estudio como se muestra en la figura 4.17, cada una de área de 0.015 Km^2, por lo que la región cubierta tiene área $a = 19 * 0.015\,\mathrm{Km}^2 = 0.285\,\mathrm{Km}^2$ y $\pi_c = \frac{0.285}{1.5} = 0.19$. Se detectan n=117 individuos por lo que $\hat{N} = \frac{117}{0.19} \approx 616$ individuos.

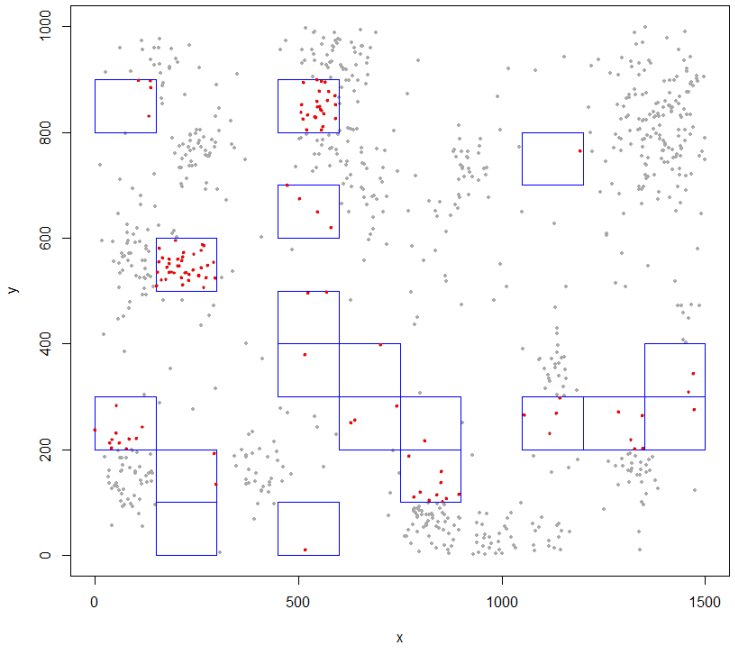

Figura 4.17: Muestreo por parcelas. Se detectan 117 individuos. Notemos que en 2 parcelas se cuentan el 54 % de los individuos detectados.

Con un nivel de confianza del 95 % para todos los intervalos, se calcula el intervalo exacto de (524, 726), suponiendo normalidad se obtiene (515, 716), usando bootstrap parámetrico y el método de percentiles se obtiene el intervalo (516, 716). Notemos que ninguno de los intervalos anteriores contiene al verdadero valor de la abundancia, esto se debe a que claramente la población no es uniforme y el número de individuos detectados es aproximadamente 30 % menor a lo que sería con una distribución uniforme (165) por lo que los estimadores, puntual y por intervalos, bajo el supuesto de uniformidad, están negativamente sesgados.

Como ya se sabe, calcular el intervalo con bootstrap no-paramétrico produce una estimación más robusta. El intervalo obtenido con bootstrap no-paramétrico es (252, 1089), el cual es muy ancho debido a que la varianza del número de individuos dentro de las diferentes parcelas es muy grande. La desviación estándar estimada con bootstrap no-paramétrico es $\widehat{SE}_{BN}(\hat{N}) = 216.43$.
Para ver si la subestimación de 616 individuos no fue ocasionada puramente por el azar, se simulan 1000 estudios de muestreo por parcelas en la misma población y se muestra la distribución del estimador en la figura 4.18.

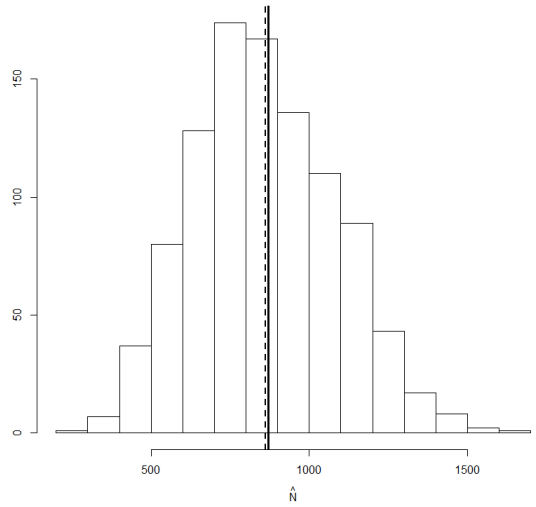

Figura 4.18: Distribución del EMV $\hat{N}$, obtenida de 1,000 simulaciones de muestreo por parcelas. La línea sólida es la abundancia real (870) y la línea punteada es la media de la observaciones simuladas (861).

A pesar de las agrupaciones en la población el estimador de muestreo por parcelas tiene un sesgo estimado de -9 individuos solamente, esto no es suficiente para considerar que este método es adecuado, ya que la varianza del estimador es muy grande.

4.3.2. Muestreo por distancias

4.3.2.1. Muestreo por transectos lineales

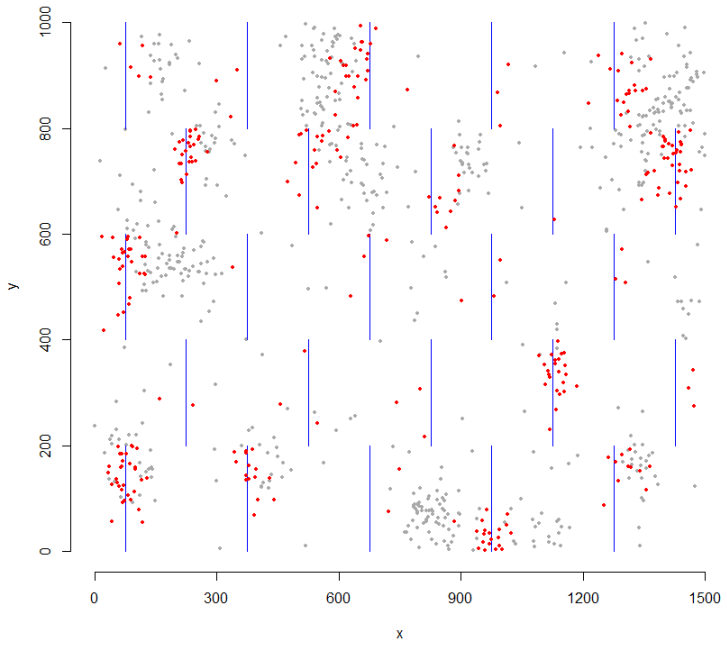

Figura 4.19: Muestreo por transectos lineales con k=25 transectos representados por las líneas azules, los puntos rojos son los individuos detectados.

Para la aplicación de este métodos se utiliza una red de k=25 transectos lineales, cada uno uno con una longitud de 200 metros por lo que la longitud total es L=5 Km. Se detectan $n = 294$ individuos que representan el 34 % de la población. Lo anterior se muestra en la figura 4.19.
Se ajusta el modelo half-normal y se obtiene $\hat{\sigma} = 35.35$, $\hat{\mu} = 44.04$, con lo que finalmente se calcula el valor estimado para la abundancia de $\hat{N} \approx 1002$ individuos, aproximadamente 15 % mayor a la abundancia real.

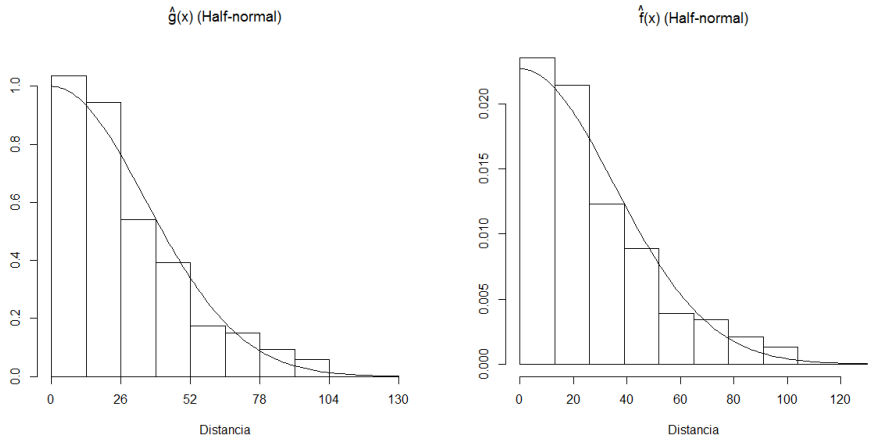

Figura 4.20: Histograma de las distancias detectadas mediante transectos lineales con el ajuste half-normal de la función de detección ajustada (izquierda) y la función de densidad ajustada (derecha).

Ajustando un modelo hazard-rate, se obtiene $\hat{\sigma} = 28.7$, $\hat{b} = 2.1$, $\hat{\mu} = 41.9$ y la abundancia estimada de $\hat{N} \approx 1053$, mayor que la abundancia real en 21 %.

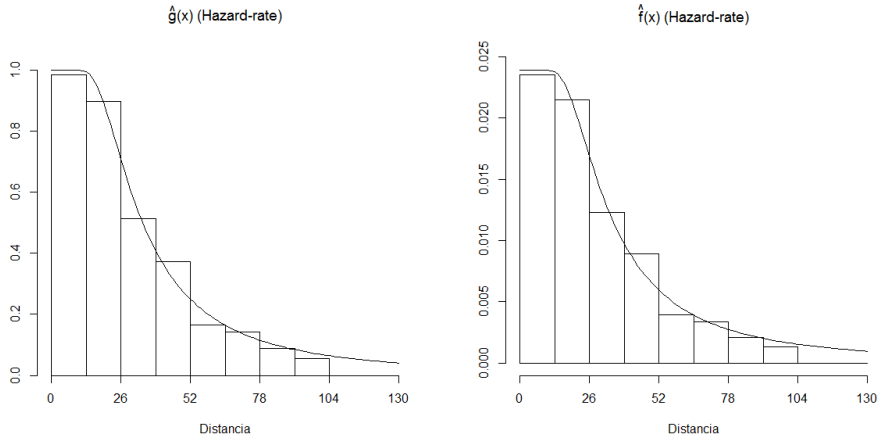

Figura 4.21: Histograma de las distancias detectadas con el ajuste hazard-rate de la función de detección ajustada (izquierda) y la función de densidad ajustada (derecha).

El ajuste de ambos modelos se muestra en las figuras 4.20 y 4.21. Para seleccionar un modelo, comenzamos por observar los gráficos P-P en la figura 4.22.

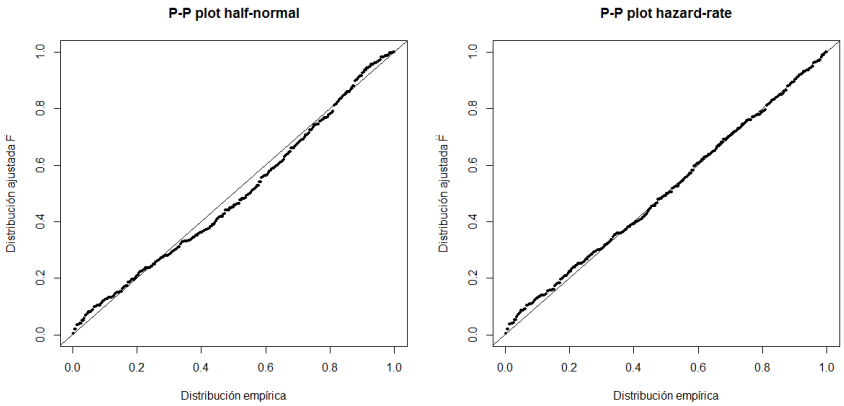

Figura 4.22: Gráficos P-P del ajuste de ambos modelos.

En este caso se observa que el modelo hazard-rate ajusta mejor que el half-normal. Veamos si las pruebas de bondad de ajuste y el AIC confirman lo observado en los gráficos P-P.

Con la prueba Kolmogorov-Smirnov obtenemos para el modelo half-normal un p-value de 0.15 y para el modelo hazard-rate de 0.76. Con la prueba de Cramér-von Mises se obtiene 0.10 para el modelo half-normal y 0.83 para el hazard-rate. Finalmente con la prueba de la Ji-cuadrada con $u = 5$ celdas equiprobables obtenemos 0.19 y 0.17 como p-value para el modelo half-normal y hazard-rate respectivamente. En las tres pruebas, no se rechaza la hipótesis nula en ningún caso, pero se obtieneN valores más grandes del p-value para el modelo hazard-rate, indicando un mejor ajuste.

Por último calculamos el AIC para ambos modelos y obtenemos: $\text{AIC}_{HN} = 2521.7$ y $\text{AIC}_{HR} = 2512.1$. Se selecciona el modelo hazard-rate.

Utilizando bootstrap no-paramétrico obtenemos el intervalo del 95 % de confianza de (692, 1458) y la desviación estándar estimada $\widehat{SE}_{BN}(\hat{N}) = 194.02$.

Simulando 1000 veces el estudio y calculando el estimador bajo el modelo hazard-rate obtenemos que el sesgo estimado es de 121 individuos (13 %).

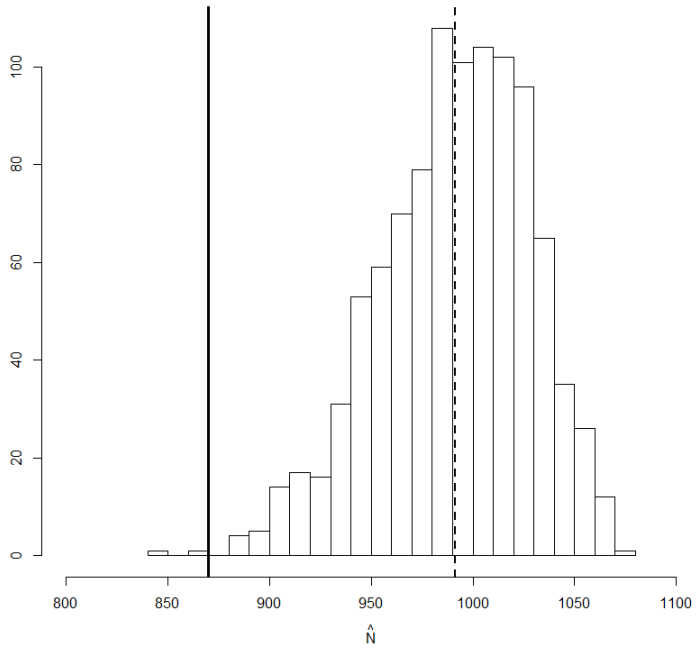

Figura 4.23: Distribución de $\hat{N}$, obtenida de 1,000 simulaciones de muestreo por transectos lineales. La línea sólida es la abundancia real (870) y la línea punteada es la media de las observaciones simuladas (991).

4.3.2.2. Muestreo por transectos puntuales

Para la aplicación de este método utilizamos una red de $k = 45$ transectos puntuales igualmente espaciados, con los cuales se detectan n=188 individuos.

Ajustando el modelo half-normal se obtiene: $\hat{\sigma} = 35.4$, $\hat{\nu} = 7,865.3$ y una abundancia estimada de $\hat{N} \approx 798$ individuos. El ajuste de las funciones de detección y de densidad se muestra en la figura 4.24.

Recordemos que por la escala utilizada para graficar la función de detección y el histograma ponderado de las distrancias, es mejor juzgar la bondad de ajuste con la gráfica de la densidad ajustada y el histograma de densidad.

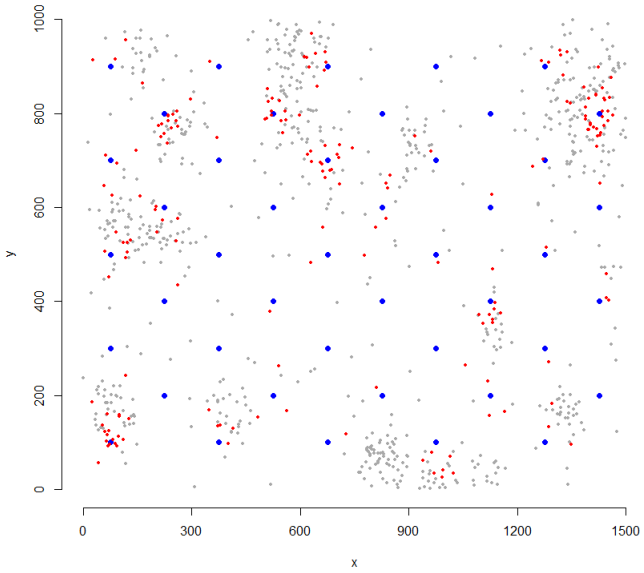

Figura 4.24: Localización de los k=45 transectos puntuales, los puntos rojos son los individuos detectados.

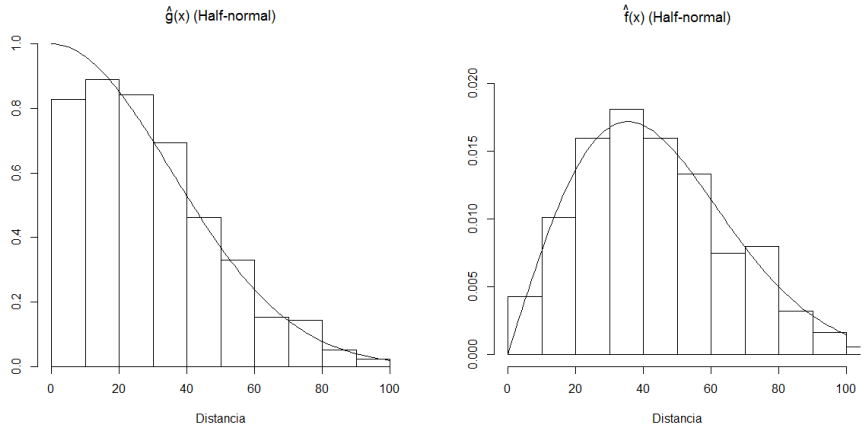

Figura 4.25: Histograma de las distancias detectadas mediante transectos puntuales con el ajuste half-normal de la función de detección (izquierda) y la función de densidad (derecha).

Con el ajuste del modelo hazard-rate obtenemos $\hat{\sigma} = 46.8$, $\hat{b} = 4.73$ $\hat{\nu} =$

$10,360$, y se estima que la abundancia es de $\hat{N} = 605$ individuos.

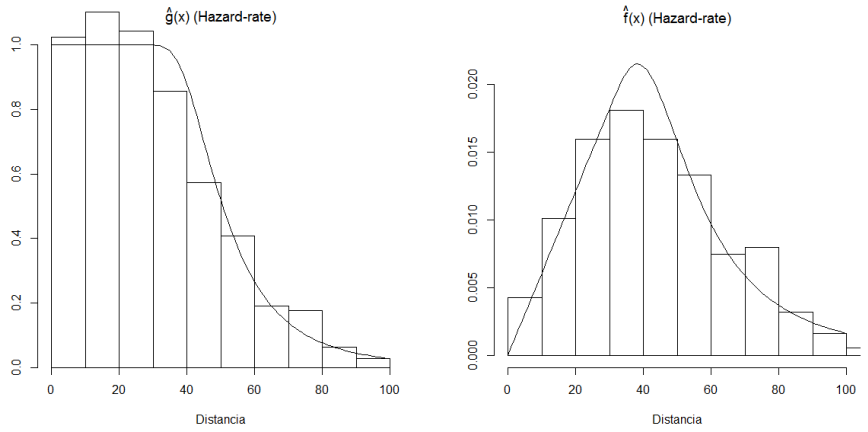

Figura 4.26: Histograma de las distancias detectadas usando transectos puntuales con el ajuste hazard-rate de la función de detección (izquierda) y la función de densidad (derecha).

Debemos ahora seleccionar un modelo.

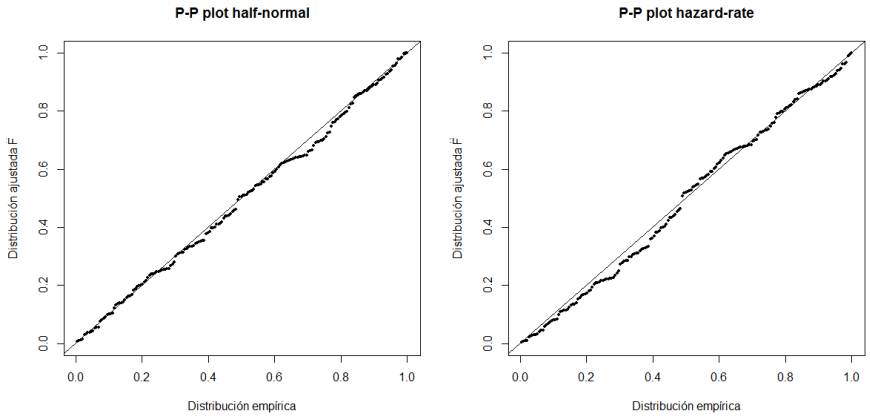

Figura 4.27: Gráficos P-P del ajuste de ambos modelos.

Con los gráficos P-P se observa que hay menos desviaciones con el modelo half-normal, pero no es totalmente claro que sea mejor el ajuste.

Con la prueba Kolmogorov-Smirnov se obtienen 0.73 y 0.60 como p-value para los modelos half-normal y hazard-rate, en el mismo orden, con la prueba

de Cramér-von Mises obtenemos 0.86 y 0.61. La prueba de la Ji-cuadrada con 5 celdas equiprobables produce el p-value de 0.76 para el modelo half-normal y 0.28 para el hazard-rate. No rechazamos en nigún caso la hipótesis nula.

Calculando el AIC para el modelo half-normal obtenemos $\text{AIC}_{HN} = 1698.8$ y para el modelo hazard-rate, $\text{AIC}_{HR} = 1700$. En este caso los resultados obtenidos en las pruebas de bondad de ajuste y en el AIC son muy similares para ambos modelos, pero por simplicidad, los p-value ligeramente mayores, y el menor valor del AIC, elegimos el modelo half-normal.

El intervalo del 95 % de confianza obtenido con bootrstrap-no paramétrico y el método de percentiles es (583, 1252) y la desviación estándar estimada es $\widehat{SE}_{BN}(\hat{N}) = 185.06$.

Veamos en la siguiente figura la distribución del estimador puntual, obtenida a partir de 1000 simulaciones de muestreo por transectos puntuales.

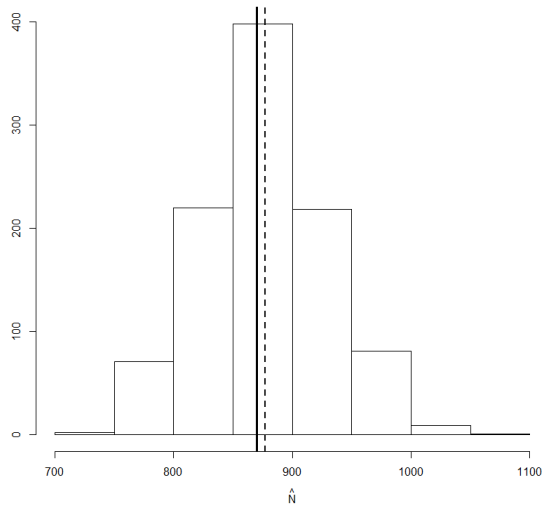

Figura 4.28: Distribución de $\hat{N}$, obtenida de 1,000 simulaciones de muestreo por transectos puntuales. La línea sólida es la abundancia real (870) y la línea punteada es la media de la observaciones simuladas (877).

El sesgo estimado para el estimador obtenido mediante transectos puntuales es de tan sólo 7 individuos, pese a que la población tiene grandes agrupaciones.

4.3.3. Vecino más cercano

En la sección 3.2 mencionamos brevemente las consecuencias que tienen las poblaciones agrupadas en el estimador obtenido con el método del vecino más cercano. La aplicación a la población simulada en esta sección debería confirmar lo mencionado anteriormente, es decir, que el estimador debe estar sesgado negativamente.

Aplicamos este método seleccionando 30 puntos aleatorios dentro de la región de estudio y midiendo las distancias al individuo más cercano, como se muestra en la figura 4.29.

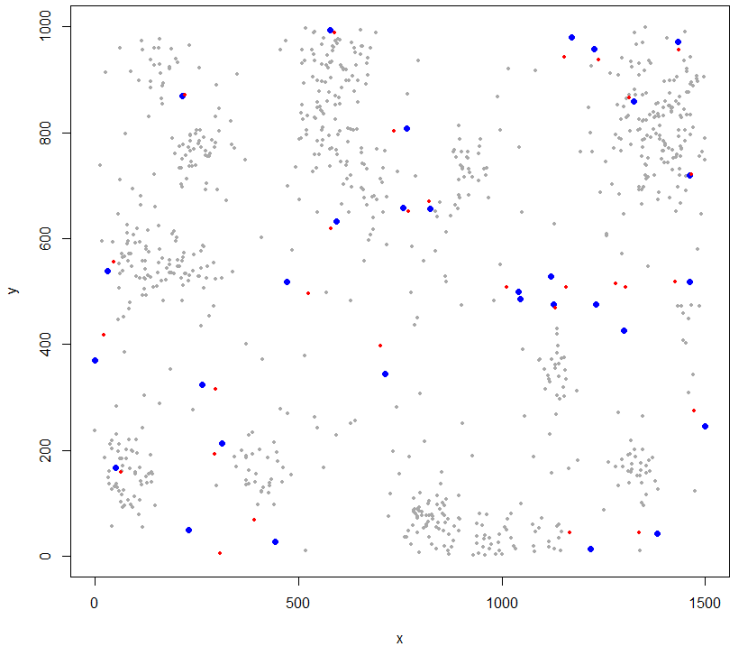

Figura 4.29: Población simulada con 870 individuos, 30 puntos aleatorios seleccionados (azul) y los individuos más cercanos (rojo).

Se obtiene un valor estimado para la abundancia de $\hat{N} \approx 276$ individuos, 68 % menor que el valor real de 870.

Es claro que las distancias al individuo más cercano son mucho más grandes de lo que serían con una distribución uniforme de los individuos y ciertamente esto ocasionará que el intervalo de confianza obtenido con bootstrap paramétrico

sea demasiado angosto. Veamos que en efecto, el intervalo del 95 % de confianza de (200, 409) obtenido mediante bootstrap paramétrico, está lejos de contener al verdadero valor de la abundancia. En este caso incluso el intervalo obtenido mediante bootstrap no-paramétrico de (189, 434) es muy angosto y no incluye al tamaño real de la población. La desviación estándar estimada es $\widehat{SE}_{BN}(\hat{N}) = 64.8$.

Para ilustrar que tan sesgado resulta el estimador, se muestra en la figura 4.30 el histograma de su distribución obtenida a partir de 1000 simulaciones.

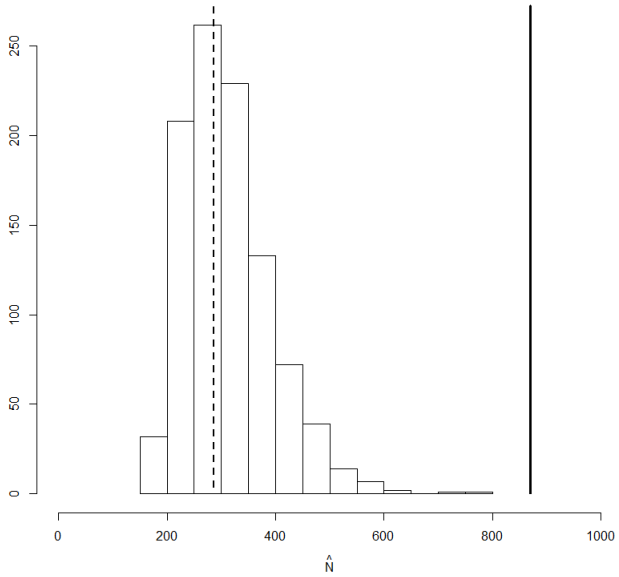

Figura 4.30: Distribución de $\hat{N}$, obtenida de 1,000 simulaciones del método del vecino más cercano con 30 puntos aleatorios. La línea sólida es la abundancia real (870) y la línea punteada es la media de la observaciones simuladas (312).

Si bien, la población simulada está muy agrupada, sirve para ejemplificar el riesgo de utilizar este método cuando la distribución de individuos no es al menos cercana a la uniforme. En este caso el sesgo estimado es de -558 individuos, es decir, en promedio se subestima la abundancia en 64 %.

4.3.4. Conclusiones

Se muestra a continuación la tabla resumen con las estimaciones obtenidas para cada método. Los intervalos del 95 % de confianza y la desviación estándar estimada son los obtenidos con bootstrap no-paramétrico. Se presenta también el coeficiente de variación estimado a partir de las simulaciones realizadas.

	Abundancia estimada	Intervalo de confianza	$\widehat{SE}_{BN}$	Sesgo estimado	$\widehat{CV}$
Muestreo por parcelas	616	(252, 1098)	216.43	-9	0.354
Transectos lineales	1053	(692, 1458)	192.02	121	0.037
Transectos puntuales	798	(583, 1252)	185.06	7	0.060
Vecinos más cercano	276	(189, 434)	64.8	-558	0.241

En esta segunda aplicación, el muestreo por parcelas parece un buen estimador en el sentido de que su sesgo estimado es de sólo -9 individuos, pero su coeficiente de variación es mucho más grande que el obtenido, por ejemplo, con transectos puntuales. Por esta razón y las mencionadas anteriormente, relacionadas con el poco rango de aplicación y la ineficiencia del método, podríamos descartar este método para la población simulada en esta aplicación.

Los métodos de muestreo por distancias presentan ambos, un coeficiente de variación pequeño, pero claramente el estimador puntual del muestreo por transectos puntuales es bastante mejor en el sentido de que tiene mucho menor sesgo estimado. El valor estimado obtenido con transectos puntuales más cercano a la abundancia real que el obtenido con transectos lineales. La principal razón de la diferencia en el funcionamiento entre estos dos métodos, para la población simulada, es que la varianza resultante de las agrupaciones en la población afecta más al método de transectos lineales: la varianza de las detecciones por transecto es 36.98 para transectos puntuales y 118.6 para transectos lineales, aproximadamente tres veces mayor para transectos lineales. Por supuesto, la menor varianza en las detecciones entre transectos es una característica

que asemeja de mejor manera el supuesto de uniformidad. Es claro que el mejor método para esta población resulta ser el de muestreo por transectos puntuales.

Las desventajas del método del vecino más cercano son muy evidentes con esta población. Las distancias en la muestra son demasiado grandes, por lo que se subestima la población en una cantidad inaceptable. Lo anterior ocasiona también que el intervalo de confianza no contenga al valor real de la abundancia. La aplicación del método del vecino más cercano a las dos poblaciones en este capítulo, nos indican lo sensible que es este método a las distribuciones no uniformes. Por tal motivo existen métodos estadísticos basados en la distribución de las distancias al vecino más cercano que son utilizados para analizar la uniformidad de ciertos datos.

Si bien, en este caso el estimador obtenido con muestreo por transectos puntuales, produce un buen estimador puntual a pesar de las agrupaciones, no es recomendable la aplicación de los métodos presentados a poblaciones que estén demasiado alejadas de los supuestos necesarios en cada método.

Existen métodos alternativos cuyos modelos son también función de la localización espacial de los individuos, es decir, permiten estimar la abundancia en una menor escala espacial. Lo anterior permite estimar la abundancia diferenciando áreas con distinta densidad de población y así obtener estimadores más precisos.

Apéndice A

Códigos en R

Este apéndice contiene los códigos para el software estadístico R con los que se implementaron los métodos presentados para las aplicaciones del **Cápitulo 4.**

A.1. Primera aplicación

```
##Instalar y cargar paquetes
install.packages("devtools")
library(devtools)
install_github("dill/wisp")
install_github("dill/Distance2")
install.packages('mgcv'); install.packages('plotrix')
install.packages('GoFKernel'); library(wisp); library(mgcv)
library(Distance2); library(plotrix); library(GoFKernel)
###4.2 PRIMERA APLICACIÓN
#Se genera la población
myreg <- generate.region(x.length=2400, y.width=1600)
mydens <- generate.density(myreg, southwest=30,
    southeast=10,northwest=90,nint.x=25, nint.y=10)
mypop.pars<-setpars.population(mydens, number.groups=1598,
      size.method="poisson")
mypop <- generate.population(mypop.pars, seed=123)
x<-mypop$posx;y<-mypop$posy
plot(x,y,type="points",pch=19,col='black',cex=.5)#FIGURA 4.1
## 4.2.1 MUESTREO POR PARCELAS
#figura 4.2
set.seed(157)
plot(x,y,type="points",pch=19,col='dark gray',cex=.5)
quadrat <- list(x=c(0,0,240,240,0),y=c(0,160,160,0,0))
n <- 0;sx<-c();sy<-c();sxy<-c();detxparc<-c()
for(i in 1:21){
  # Elegir parcela aleatoria y graficar
  this.quadrat <- quadrat
  sx<-sample(0:9,1);sy<-sample(0:9,1)
```

```
  band<-0;cont<-1
  while(band==0){
    if(length(which(sxy==sx+sy*10))==0){
      band<-1
      sxy<-c(sxy,sx+sy*10)}
    else{
      sx<-sample(seq(0,9,by=1),1)
      sy<-sample(seq(0,9,by=1),1)} }
  this.quadrat$x <- this.quadrat$x + sx*240
  this.quadrat$y <- this.quadrat$y + sy*160
  polygon(this.quadrat, lty=1, border="blue")
  # ver que puntos están dentro y contarlos
  inout <- inSide(this.quadrat, x, y)
  detxparc<-c(detxparc,sum(inout))
  n <- n + sum(inout)
  points(x[inout], y[inout], pch=15, cex=0.4,
  col="red",xlab=NULL,ylab=NULL,axes=FALSE,axis=2)}
n #individuos contados
#Intervalos de confianza exactos
NF<- function(q,p,n){
  vec<-n:4000
  y=abs(pbinom(n,vec,p)-q)
  tab<-cbind(vec,y)
  return(tab[y==min(y)][1])}
CIEMP<-c(NF(0.975,.21,n),NF(0.025,.21,n));CIEMP
#Intervalos normales
segmp<-sqrt(n/.21*(1-.21)/.21)
CINMP<-c((n/.21)-qnorm(.975,0,1)*segmp,(n/.21)+qnorm(.975,0,1)*segmp);CINMP
#Intervalos bootstrap paramétrico
set.seed(154)
MBP<-rbinom(1000,1757,.21)/.21
CIBPMP<-quantile(MBP,c(0.025,.975));CIBPMP
sd(MBP)#desviación estándar
#Intervalos bootstrap no-paramétrico
MBNP<-c()
for (i in 1:1000){
rem<-sample(detxparc,21,replace = T)
MBNP<-c(MBNP,sum(rem)/.21)}
CIBNMP<-quantile(MBNP,c(0.025,.975));CIBNMP
sd(MBNP)#desviación estándar
#Simulación del estudio 1000 veces
#función para simular
EMVMP<-function(){quadrat <- list(x=c(0,0,240,240,0),y =c(0,160,160,0,0))
n <- 0;sx <- c();sy <- c();sxy <- c()
for (i in 1:21) {
  this.quadrat <- quadrat
  sx <- sample(0:9, 1)
  sy <- sample(0:9, 1)
  band <- 0
  cont <- 1
  while (band == 0) {
    if (length(which(sxy == sx + sy * 10)) == 0) {
      band <- 1; sxy <- c(sxy, sx + sy * 10)}
    else{sx <- sample(seq(0, 9, by = 1), 1)
      sy <- sample(seq(0, 9, by = 1), 1)}}
  this.quadrat$x <- this.quadrat$x + sx * 240
  this.quadrat$y <- this.quadrat$y + sy * 160
```

```
  inout <- inSide(this.quadrat, x, y)
  n <- n + sum(inout)}n}
Nhatmp<-c()
for (i in 1:1000){Nhatmp[i]<-EMVMP()/.21}
hist(Nhatmp,breaks=20,prob=F,xlab=expression(hat(N)),ylab='',main='')
abline(v=c(1598,mean(Nhatmp)),lty=c(1,2),lwd=c(3,2))# figura 4.3
mean(Nhatmp);sd(Nhatmp)
sd(Nhatmp)/mean(Nhatmp)#coeficiente de variación

## 4.2.2 MUESTREO POR DISTANCIAS
##4.2.2.1 MUESTREO POR TRANSECTOS LINEALES
P<-as.data.frame( cbind(x,y))
par(mfrow=c(1,1));set.seed(157)
plot(x,y,type="points",pch=19,col='dark gray',cex=.5,axes=F)
axis(side=2, at=seq(0, 1600, by=320))
axis(side=1, at=seq(0, 2400, by=480))
cx<-c();cy<-c();xcons<-c();ycons<-c()
dist<-c();det_trans<-c();det<-c()
det_dist<-c()
for (i in 1:5){
  cy<-320*(i-1)
  for (j in 1:5){
    if ((i%%2)==0){ cx<-120*3}
    else {cx<-120 }
    if((cx+(j-1)*480)<2400){segments(cx+(j-1)*480,cy,cx+(j-1)*480,cy+320,col='blue')
    xcons<-P$x[(P$y<=cy+320)&(P$y>cy)]
    ycons<-P$y[(P$y<=cy+320)&(P$y>cy)]
    dist<-abs((cx+(j-1)*480)-xcons)
    det <- exp(-dist^2/(2*50^2)) > runif(length(dist))
    det_trans<-c(det_trans,sum(det[det=TRUE]))
    det_dist<-c(det_dist,dist[det])
    points(xcons[det], ycons[det], pch=19, cex=.5, col="red")}}}#figura 4.4
n<-length(det_dist);n
###ajuste half normal
sighn<-sqrt(sum(det_dist^2)/n);sighn
ghn<-function(x){exp(-x^2/(2*(sighn^2)))}##g estimada
mghn<-integrate(ghn,lower=0,upper = max(det_dist))$value;mghn
abghn<-(length(det_dist)*(2400*1600))/(2*mghn*25*320);abghn #abundancia estimada
fhn<-function(x){exp(-x^2/(2*(sighn)^2))/mghn}##f estimada}
FHN<-function(x){integrate(fhn,lower = 0,upper = x)$value}#F estimada
IFHN<-inverse(FHN,0,max(det_dist))
par(mfrow=c(1,2))
weighted.hist(det_dist,rep(mghn/(length(det_dist)*20),length(det_dist)),
breaks= seq(0,160,by=20),xlim=c(0,155),xaxis=F,
main=expression(paste(hat(g),'(x) (Half-normal)')),xlab='Distancia',ylab='')
axis(side=1, at=seq(0, 150, by=50))
curve(ghn,add=T,from = 0,to=160)
hist(det_dist,prob  = T,breaks=8,axes = T,main=expression
(paste(hat(f),'(x) (Half-normal)')),xlab='Distancia',ylab='')
curve(fhn,from = 0,to=160,add = T)#figura 4.5
par(mfrow=c(1,1))
##ajuste hazard rate (USAMOS LIBRERIA DISTANCE)
DAT<-as.data.frame(det_dist)
names(DAT)<-'distance'
modhrl<-ds(DAT,model=df(model=~hr))
sighr<-exp(modhrl$par[1]);sighr
bghr<-exp(modhrl$par[2]);bghr
```

```
ghr<-function(x){1-exp(-(x/sighr)^(-bghr))}
mghr<-integrate(ghr,lower=0,upper = max(det_dist))$value;mghr
abghr<-(length(det_dist)*(2400*1600))/(2*mghr*25*320);abghr
fhr<-function(x){(1-exp(-(x/sighr)^(-bghr)))/mghr}
FHR<-function(x){integrate(fhr,lower = 0,upper = x)$value}
IFHR<-inverse(FHR,0,max(det_dist))
par(mfrow=c(1,2))
weighted.hist(det_dist,rep(mghr/(length(det_dist)*20),length(det_dist)),
breaks= seq(0,160,by=20),xlim=c(0,155),xaxis=F,
main=expression(paste(hat(g),'(x) (Hazard-rate)')),xlab='Distancia',ylab='')
axis(side=1, at=seq(0, 150, by=50))
curve(ghr,add=T,from = 0,to=160)
hist(det_dist,prob  = T,breaks=8,axes = T,main=expression
(paste(hat(f),'(x) (Hazard-rate)')),xlab='Distancia',ylab='')
curve(fhr,from = 0,to=160,add = T)#figura 4.6
par(mfrow=c(1,1))
#Elección modelo
#Gráficos P-P
emp<-ecdf((det_dist));ord<-sort(det_dist)
a1<-emp(ord);bhn<-c();bhr<-c()
for (i in 1:n){bhn[i]<-FHN(ord[i]);bhr[i]<-FHR(ord[i])}
par(mfrow=c(1,2))
plot(a1,bhn,ylim=c(0,1),cex=.6,pch=19,xlab='Distribución empírica',
     ylab=expression(paste("Distribución ajustada  ",hat(F))),
     main='P-P plot half-normal')
abline(a=0,b=1,lwd=1)
plot(a1,bhr,ylim=c(0,1),cex=.6,pch=19,xlab='Distribución empírica',
     ylab=expression(paste("Distribución ajustada  ",hat(F))),
     main='P-P plot hazard-rate')
abline(a=0,b=1,lwd=1)#figura 4.7
par(mfrow=c(1,1))
#Prueba Kolmogorov-Smirnov y Cramér-von Mises
modhnl<-ds(DAT)
kehn<-gof_tests(modhnl,plot = F)$kolmogorov_smirnov$Dn
kphn<-gof_tests(modhnl,plot = F)$kolmogorov_smirnov$p
kehn;kphn
cehn<-gof_tests(modhnl,plot = F)$cramer_vonMises$W
cphn<-gof_tests(modhnl,plot = F)$cramer_vonMises$p
cehn;cphn
kehr<-gof_tests(modhrl,plot = F)$kolmogorov_smirnov$Dn
kphr<-gof_tests(modhrl,plot = F)$kolmogorov_smirnov$p
kehr;kphr
cehr<-gof_tests(modhrl,plot = F)$cramer_vonMises$W
cphr<-gof_tests(modhrl,plot = F)$cramer_vonMises$p
cehr;cphr
#Prueba Ji-cuadrada
#Half normal
obs<-c();esp<-c();cut<-c()
for (i in 1:6){cut[i]<-IFHN(.20*(i-1))}
for (i in 1:5){
  obs[i]<-length(det_dist[(det_dist>=cut[i])&(det_dist<=cut[i+1])])
  esp[i]<-(FHN(cut[i+1])-FHN(cut[i]))*n}
ECH<-sum(((obs-esp)^2)/esp); ECH
qchisq(.95,length(obs)-1-1)
pv<-1-pchisq(ECH,length(obs)-1-1);pv
#Hazard rate
obs<-c();esp<-c();cut<-c()
```

```
for (i in 1:6){cut[i]<-IFHR(.20*(i-1))}
for (i in 1:5){
  obs[i]<-length(det_dist[(det_dist>=cut[i])&(det_dist<=cut[i+1])])
  esp[i]<-(FHR(cut[i+1])-FHR(cut[i]))*n}
ECH<-sum(((obs-esp)^2)/esp); ECH
qchisq(.95,length(obs)-1-2)
pv<-1-pchisq(ECH,length(obs)-1-2);pv
#AIC
#Half normal
hnaic<- -2*sum(log(fhn(det_dist)))+2
#hazard rate
hraic<- -2*sum(log(fhr(det_dist)))+4
hnaic;hraic
#se elige el modelo half normal
OBSTL<-function(){#función para simular
  cx<-c();cy<-c();xcons<-c();ycons<-c()
  dist<-c();det_trans<-c();det<-c();det_dist<-c()
  for (i in 1:5){
    cy<-320*(i-1)
    for (j in 1:5){
      if ((i%%2)==0){ cx<-120*3 }
      else {cx<-120 }
      if((cx+(j-1)*480)<2400){
        xcons<-P$x[(P$y<=cy+320)&(P$y>cy)]
        ycons<-P$y[(P$y<=cy+320)&(P$y>cy)]
        dist<-abs((cx+(j-1)*480)-xcons)
        det <- exp(-dist^2/(2*50^2)) > runif(length(dist))
        det_trans<-c(det_trans,sum(det[det=TRUE]))
        det_dist<-c(det_dist,dist[det])}}}
n<-length(det_dist)
return(list(n,det_dist))}
#Estimacion por intervalos
#bootstrap no paramétrico
MBNP<-c()
for (i in 1:1000){
  rem<-sample(det_trans,25,replace = T)
  MBNP<-c(MBNP,(sum(rem)*(2400*1600))/(2*mghn*25*320))}
CIBNMP<-quantile(MBNP,c(0.025,.975));CIBNMP
sd(MBNP)#desviación estándar
#1000 simulaciones del estudio
Nhattl<-c()
for (i in 1:1000){
  m<-OBSTL()
  sighn<-sqrt(sum(m[[2]]^2)/m[[1]])
  ghn<-function(x){exp(-x^2/(2*(sighn^2)))}
  mghn<-integrate(ghn,lower=0,upper = max(m[[2]]))$value
  Nhattl[i]<-(m[[1]]*(2400*1600))/(2*mghn*25*320)}
hist(Nhattl,breaks=20,prob=F,xlab=expression(hat(N)),ylab='',main='')
abline(v=c(1598,mean(Nhattl)),lwd=c(3,2),lty=c(1,2))#figura 4.8
mean(Nhattl);sd(Nhattl)
sd(Nhattl)/mean(Nhattl)#coeficiente de variación

##4.2.2.2 MUESTREO POR TRANSECTOS PUNTUALES
set.seed(157)
plot(x,y,type="points",pch=19,col='dark gray',cex=.5,axes=F)
axis(side=2, at=seq(0, 1600, by=320))
axis(side=1, at=seq(0, 2400, by=480))
```

```
cx<-c();cy<-c();xcons<-c();ycons<-c()
dist<-c();det_trans<-c();det<-c();det_dist<-c()
for (i in 1:9){
  cy<-160*(i)
  for (j in 1:5){
    if ((i%%2)==0){ cx<-120*3}
    else {cx<-120 }
    if((cx+(j-1)*480)<2400){points(cx+(j-1)*480,cy,col='blue',pch=19)
      xcons<-P$x[(P$y<=cy+160)&(P$y>cy-160)]
      ycons<-P$y[(P$y<=cy+160)&(P$y>cy-160)]
      dist<-sqrt((cx+(j-1)*480 - xcons)^2+(cy-ycons)^2)
      det <- exp(-dist^2/(2*60^2)) > runif(length(dist))
      det_trans<-c(det_trans,sum(det[det=TRUE]))
      det_dist<-c(det_dist,dist[det])
      points(xcons[det], ycons[det], pch=19, cex=.5, col="red")}}}#figura 4.9
n<-length(det_dist);n
#ajuste half normal
sighn<-sqrt(sum(det_dist^2)/(2*n));sighn#sigma est.
ghn<-function(x){exp(-x^2/(2*(sighn^2)))}#g est.
integ<-function(x){2*pi*x*exp(-x^2/(2*(sighn^2)))}
mughn<-integrate(ghn,lower = 0,upper = max(det_dist))$value
nughn<-integrate(integ,lower=0,upper = max(det_dist))$value;nughn
Pahn<-nughn/(pi*(max(det_dist)^2))
abghn<-(n*(2400*1600))/(45*nughn);abghn
fhn<-function(x){(2*pi*x*exp(-x^2/(2*(sighn^2))))/nughn}#f est.
FHN<-function(x){integrate(fhn,lower = 0,upper = x)$value}
IFHN<-inverse(FHN,0,max(det_dist))
par(mfrow=c(1,2))
weighted.hist(det_dist,(1/det_dist)*(mughn/(sum(1/det_dist)*20)),
              breaks = seq(0,200,by=20),xlim=c(0,200),xaxis=F,
              main=expression(paste(hat(g),'(x) (Half-normal)')),xlab='Distancia',ylab='')
axis(side=1, at=seq(0, 200, by=40))
curve(ghn,add=T,from = 0,to=200)
hist(det_dist,prob  = T,axes = T,
     main=expression(paste(hat(f),'(x) (Half-normal)')),xlab='Distancia',ylab='')
curve(fhn,from = 0,to=200,add = T)# figura 4.10
par(mfrow=c(1,1))
#ajuste hazard rate (usamos libreria distance)
DAT<-as.data.frame(det_dist)
names(DAT)<-'distance'
mhrp<-ds(DAT,transect="point",model=df(model=~hr),truncation = max(det_dist))
sighr<-exp(mhrp$par[1]);sighr
bghr<-exp(mhrp$par[2]);bghr
ghr<-function(x){1-exp(-(x/sighr)^(-bghr))}
integ1<-function(x){2*pi*x*(1-exp(-(x/sighr)^(-bghr)))}
mughr<-integrate(ghr,lower=0,upper = max(det_dist))$value
nughr<-integrate(integ1,lower=0,upper = max(det_dist))$value;nughr
abghr<-(n*(2400*1600))/(45*nughr);abghr
fhr<-function(x){2*pi*x*(1-exp(-(x/sighr)^(-bghr)))/nughr}
FHR<-function(x){integrate(fhr,lower = 0,upper = x)$value}
IFHR<-inverse(FHR,0,max(det_dist))
par(mfrow=c(1,2))
weighted.hist(det_dist,(1/det_dist)*(mughr/(sum(1/det_dist)*20)),
              breaks = seq(0,200,by=20),xlim=c(0,200),xaxis=F,
              main=expression(paste(hat(g),'(x) (Hazard-rate)')),xlab='Distancia',ylab='')
axis(side=1, at=seq(0, 200, by=40))
curve(ghr,add=T,from = 0,to=200)
```

```
hist(det_dist,prob  = T,axes = T,
     main=expression(paste(hat(f),'(x) (Hazard-rate)')),
     xlab='Distancia',ylab='',ylim = c(0,0.012))
curve(fhr,from = 0,to=200,add = T)#figura 4.11
par(mfrow=c(1,1))
#Elección del modelo
#Gráficos P-P
emp<-ecdf((det_dist));ord<-sort(det_dist)
a1<-emp(ord);bhn<-c();bhr<-c()
for (i in 1:n){bhn[i]<-FHN(ord[i]);bhr[i]<-FHR(ord[i])}
par(mfrow=c(1,2))
plot(a1,bhn,ylim=c(0,1),cex=.6,pch=19,xlab='Distribución empírica',
     ylab=expression(paste("Distribución ajustada  ",hat(F))),
     main='P-P plot half-normal')
abline(a=0,b=1,lwd=1)
plot(a1,bhr,ylim=c(0,1),cex=.6,pch=19,xlab='Distribución empírica',
     ylab=expression(paste("Distribución ajustada  ",hat(F))),
     main='P-P plot hazard-rate')
abline(a=0,b=1,lwd=1)#figura 4.12
par(mfrow=c(1,1))
#Prueba Kolmogorov-Smirnov y Cramér-von Mises
mhnp<-ds(DAT,transect = "point",truncation = max(det_dist))
kehn<-gof_tests(mhnp,plot = F)$kolmogorov_smirnov$Dn
kphn<-gof_tests(mhnp,plot = F)$kolmogorov_smirnov$p
kehn;kphn
kehr<-gof_tests(mhrp,plot = F)$kolmogorov_smirnov$Dn
kphr<-gof_tests(mhrp,plot = F)$kolmogorov_smirnov$p
kehr;kphr
cehn<-gof_tests(mhnp,plot = F)$cramer_vonMises$W
cphn<-gof_tests(mhnp,plot = F)$cramer_vonMises$p
cehn;cphn
cehr<-gof_tests(mhrp,plot = F)$cramer_vonMises$W
cphr<-gof_tests(mhrp,plot = F)$cramer_vonMises$p
cehr;cphr
#Prueba Ji-cuadrada
#half normal
obs<-c();esp<-c();cut<-c()
for (i in 1:6){cut[i]<-IFHN(.20*(i-1))}
for (i in 1:5){
  obs[i]<-length(det_dist[(det_dist>=cut[i])&(det_dist<=cut[i+1])])
  esp[i]<-(FHN(cut[i+1])-FHN(cut[i]))*n}
ECH<-sum(((obs-esp)^2)/esp); ECH
qchisq(.95,length(obs)-1-1)
pv<-1-pchisq(ECH,length(obs)-1-1);pv
#hazard rate
obs<-c();esp<-c();cut<-c()
for (i in 1:6){cut[i]<-IFHR(.20*(i-1))}
for (i in 1:5){
  obs[i]<-length(det_dist[(det_dist>=cut[i])&(det_dist<=cut[i+1])])
  esp[i]<-(FHR(cut[i+1])-FHR(cut[i]))*n}
ECH<-sum(((obs-esp)^2)/esp); ECH
qchisq(.95,length(obs)-1-2)
pv<-1-pchisq(ECH,length(obs)-1-2);pv
#AIC
#half normal
hnaic<- -2*sum(log(fhn(det_dist)))+2
#hazard rate
```

```
hraic<- -2*sum(log(fhr(det_dist)))+4
hnaic;hraic
#elegimos el modelo half normal
OBSTP<-function(){cx<-c()#función para simular
cy<-c();xcons<-c();ycons<-c();dist<-c()
det_trans<-c();det<-c();det_dist<-c()
for (i in 1:9){
  cy<-160*(i)
  for (j in 1:5){
    if ((i%%2)==0){ cx<-120*3}
    else {cx<-120 }
    if((cx+(j-1)*480)<2400){
      xcons<-P$x[(P$y<=cy+160)&(P$y>cy-160)]
      ycons<-P$y[(P$y<=cy+160)&(P$y>cy-160)]
      dist<-sqrt((cx+(j-1)*480 - xcons)^2+(cy-ycons)^2)
      det <- exp(-dist^2/(2*60^2)) > runif(length(dist))
      det_trans<-c(det_trans,sum(det[det=TRUE]))
      det_dist<-c(det_dist,dist[det])}}}
n<-length(det_dist)
return(list(n,det_dist))}
#Estimación por intervalos
#bootstrap no paramétrico
MBNP<-c()
for (i in 1:1000){
  rem<-sample(det_trans,45,replace = T)
  MBNP<-c(MBNP,(sum(rem)*(2400*1600))/(45*nughn))}
CIBNMP<-quantile(MBNP,c(0.025,.975));CIBNMP
sd(MBNP)#desviación estándar
#1000 simulaciones
Nhattp<-c()
for (i in 1:1000){
  m<-OBSTP()
  sighn<-sqrt(sum(m[[2]]^2)/(2*m[[1]]));sighn
  ghn<-function(x){exp(-x^2/(2*(sighn^2)))}##g estimada
  integ<-function(x){2*pi*x*exp(-x^2/(2*(sighn^2)))}
  mughn<-integrate(ghn,lower = 0,upper = max(det_dist))$value
  nughn<-integrate(integ,lower=0,upper = max(det_dist))$value
  Nhattp[i]<-(m[[1]]*(2400*1600))/(45*nughn)}
hist(Nhattp,prob=F,xlab=expression(hat(N)),ylab='',main='')
abline(v=c(1598,mean(Nhattp)),lwd=c(3,2),lty=c(1,2))#figura 4.13
mean(Nhattp);sd(Nhattp)
sd(Nhattp)/mean(Nhattp)#coeficiente de variación

##4.2.3 VECINO MÁS CERCANO
set.seed(1447)
h<-runif(30)*2400;k<-runif(30)*1600
plot(x,y,type="points",pch=19,col='dark gray',cex=.5,axes=F)
axis(side=2, at=seq(0, 1600, by=320))
axis(side=1, at=seq(0, 2400, by=480))
cx<-c();cy<-c();xcons<-c();ycons<-c();dist<-c();det_dist<-c()
for (i in 1:length(h)){
  cy<-k[i]
      points(h[i],k[i],col='blue',pch=19)
      xcons<-P$x[(P$y<=cy+160)&(P$y>cy-160)]
      ycons<-P$y[(P$y<=cy+160)&(P$y>cy-160)]
      dist<-sqrt((h[i] - xcons)^2+(k[i]-ycons)^2)
      points(xcons[which(dist==min(dist))],ycons[which(dist==min(dist))],
```

```
        col='red',pch=19,cex=.5)
        det_dist<-c(det_dist,min(dist))}#figura 4.14
ngnn<-30*2400*1600/(pi*sum(det_dist^2));ngnn #abundancia estimada
dgnn<-ngnn/(2400*1600);dgnn #densidad estimada
#estimación por intervalos
#bootstrap paramétrico
finv<-function(x){sqrt((-log(1-x))/(pi*dgnn))}
estim<-function(x){(30*2400*1600)/(pi*sum(x^2))}
munif<-c();Nbp<-c()
for (i in 1:10000){
  munif<-runif(30)
  Nbp[i]<-estim(finv(munif))}
quantile(Nbp,probs = c(0.025,.975))#intervalo bp
#bootstrap no-paramétrico
Nbnp<-c();mubnp<-c();ddbnp<-c()
for(i in 1:10000){
  mubnp<-sample(1:30,30,replace = T)
  ddbnp<-det_dist[mubnp]
  Nbnp[i]<-estim(ddbnp)}
quantile(Nbnp,probs = c(0.025,.975))#intervalo bnp
mean(Nbnp)
sd(Nbnp)#desviación estándar
#1000 simulaciones
OBSVC<-function(){h<-runif(30)*2400 #función para simular
k<-runif(30)*1600;cx<-c();cy<-c();xcons<-c();ycons<-c()
dist<-c();det_dist<-c()
for (i in 1:length(h)){
  cy<-k[i]
  xcons<-P$x[(P$y<=cy+160)&(P$y>cy-160)]
  ycons<-P$y[(P$y<=cy+160)&(P$y>cy-160)]
  dist<-sqrt((h[i] - xcons)^2+(k[i]-ycons)^2)
  det_dist<-c(det_dist,min(dist))}
det_dist}
NVC<-c()
for (i in 1:1000){
  NVC[i]<-45*2400*1600/(pi*sum(OBSVC()^2))}
hist(NVC,ylim = c(0,400),main='',ylab='',xlab=expression(hat(N)))
clip(0,2500,0,380)
abline(v=c(1598,mean(NVC)),lwd=c(3,2),lty=c(1,2))#figura 4.15
mean(NVC)
sd(NVC)/mean(NVC)#coeficiente de variación
```

A.2. Segunda aplicación

```
##Se utilizan las mismas librerías que en la primer aplicación
###4.2 PRIMERA APLICACIÓN
#Se genera la población
r<- generate.region(x.length =1500, y.width = 1000)
de <- generate.density(r,nint.x = 80, nint.y = 80 ,southwest = 10,
                       southeast = 10, northwest = 10)
set.seed(7854)
de<-add.hotspot(de, runif(1)*1500,runif(1)*1000, 250,40)
de<-add.hotspot(de, runif(1)*1500,runif(1)*1000, 400,25)
de<-add.hotspot(de, runif(1)*1500,runif(1)*1000, 400,40)
de<-add.hotspot(de, runif(1)*1500,runif(1)*1000, 400,80)
de<-add.hotspot(de, runif(1)*1500,runif(1)*1000, 250,40)
```

```
de<-add.hotspot(de, runif(1)*1500,runif(1)*1000, 400,40)
de<-add.hotspot(de, runif(1)*1500,runif(1)*1000, 250,40)
de<-add.hotspot(de, runif(1)*1500,runif(1)*1000, 250,40)
de<-add.hotspot(de, runif(1)*1500,runif(1)*1000, 400,40)
de<-add.hotspot(de, runif(1)*1500,runif(1)*1000, 120,25)
de<-add.hotspot(de, runif(1)*1500,runif(1)*1000, 400,40)
de<-add.hotspot(de, runif(1)*1500,runif(1)*1000, 250,50)
de<-add.hotspot(de, runif(1)*1500,runif(1)*1000, 250,40)
de<-add.hotspot(de, runif(1)*1500,runif(1)*1000, 400,30)
de<-add.hotspot(de, runif(1)*1500,runif(1)*1000, 250,40)
de<-add.hotspot(de, runif(1)*1500,runif(1)*1000, 400,40)
de<-add.hotspot(de, runif(1)*1500,runif(1)*1000, 300,25)
de<-add.hotspot(de, runif(1)*1500,runif(1)*1000, 250,40)
de<-add.hotspot(de, runif(1)*1500,runif(1)*1000, 120,40)
de<-add.hotspot(de, runif(1)*1500,runif(1)*1000, 120,25)
de<-add.hotspot(de, runif(1)*1500,runif(1)*1000, 250,40)
de<-add.hotspot(de, runif(1)*1500,runif(1)*1000, 400,40)
de<-add.hotspot(de, runif(1)*1500,runif(1)*1000, 120,25)
de<-add.hotspot(de, runif(1)*1500,runif(1)*1000, 250,40)
de<-add.hotspot(de, runif(1)*1500,runif(1)*1000, 120,40)
pop.pars<-setpars.population(density.pop = de, number.groups = 870,
                             size.method = "user", size.values = c(10,15,30),
                            size.prob=c(.4,.4,.2))
mipob <- generate.population(pop.pars, seed=1191)
x<-mipob$posx;y<-mipob$posy
P<-as.data.frame( cbind(x,y))
plot(x,y,type="points",pch=19,col='black',cex=.5)#figura 4.16
###MUESTREO POR PARCELAS
set.seed(1157)
plot(x,y,type="points",pch=19,col='dark gray',cex=.5)
quadrat <- list(x=c(0,0,150,150,0),y=c(0,100,100,0,0))
n <- 0;sx<-c();sy<-c();sxy<-c();detxparc<-c()
for(i in 1:19){
  this.quadrat <- quadrat
  sx<-sample(0:9,1);sy<-sample(0:9,1)
  band<-0;cont<-1
  while(band==0){
    if(length(which(sxy==sx+sy*10))==0){
      band<-1;sxy<-c(sxy,sx+sy*10)}
    else{
      sx<-sample(seq(0,9,by=1),1)
      sy<-sample(seq(0,9,by=1),1)}}
  this.quadrat$x <- this.quadrat$x + sx*150
  this.quadrat$y <- this.quadrat$y + sy*100
  polygon(this.quadrat, lty=1, border="blue")
  inout <- inSide(this.quadrat, x, y)
  detxparc<-c(detxparc,sum(inout))
  n <- n + sum(inout)
  points(x[inout], y[inout], pch=15, cex=0.4, col="red",
         xlab=NULL,ylab=NULL,axes=FALSE,axis=2)}#figura 4.17
n
n/.19 #abundancia est.
#Estimación por intervalos
#Intervalos exactos
NF<- function(q,p,n){
  vec<-n:4000
  y=abs(pbinom(n,vec,p)-q)
```

```
  tab<-cbind(vec,y)
  return(tab[y==min(y)][1])}
CIEMP<-c(NF(0.975,.19,n),NF(0.025,.19,n));CIEMP
#Intervalos normalidad
segmp<-sqrt(n/.19*(1-.19)/.19)
CINMP<-c((n/.19)-qnorm(.975,0,1)*segmp,(n/.19)+qnorm(.975,0,1)*segmp)
CINMP
#Intervalos bootstrap paramétrico
set.seed(157)
MBP<-rbinom(1000,round(n/.19,0),.19)/.19
CIBPMP<-quantile(MBP,c(0.025,.975));CIBPMP
#Intervalo bootstrap no paramétrico
MBNP<-c()
for (i in 1:1000){
  rem<-sample(detxparc,19,replace = T)
  MBNP<-c(MBNP,sum(rem)/.19)}
CIBNMP<-quantile(MBNP,c(0.025,.975));CIBNMP
sd(MBNP)#desviación estándar
mean(MBNP)
#1000 simulaciones
EMVMP<-function(){quadrat <- list(x=c(0,0,150,150,0),y=c(0,100,100,0,0))
n <- 0;sx<-c();sy<-c();sxy<-c();detxparc<-c()
for(i in 1:19){
  this.quadrat <- quadrat
  sx<-sample(0:9,1)
  sy<-sample(0:9,1)
  band<-0;  cont<-1
  while(band==0){
    if(length(which(sxy==sx+sy*10))==0){
      band<-1
      sxy<-c(sxy,sx+sy*10)}
    else{
      sx<-sample(seq(0,9,by=1),1)
      sy<-sample(seq(0,9,by=1),1)}}
  this.quadrat$x <- this.quadrat$x + sx*150
  this.quadrat$y <- this.quadrat$y + sy*100
  inout <- inSide(this.quadrat, x, y)
  detxparc<-c(detxparc,sum(inout))
  n <- n + sum(inout)}
n}
Nhatmp<-c()
for (i in 1:1000){Nhatmp[i]<-EMVMP()/.19}
hist(Nhatmp,breaks=20,prob=F,main='',ylab='',xlab=expression(hat(N)))
abline(v=c(870,mean(Nhatmp)),lty=c(1,2),lwd=c(3,2))#figura 4.18
mean(Nhatmp)
sd(Nhatmp)
sd(Nhatmp)/mean(Nhatmp) #coeficiente de variación

###4.3.2 MUESTREO POR DISTANCIAS
##4.3.2.1 MUESTREO POR TRANSECTOS LINEALES
set.seed(147)
plot(x,y,type="points",pch=19,col='dark gray',cex=.5,axes=F)
axis(side=2, at=seq(0, 1000, by=200))
axis(side=1, at=seq(0, 1500, by=300))
cx<-c();cy<-c();xcons<-c();ycons<-c();dist<-c()
det_trans<-c();det<-c();det_dist<-c()
for (i in 1:5){
```

```
  cy<-200*(i-1)
  for (j in 1:5){
    if ((i%%2)==0){ cx<-75*3}
    else {cx<-75 }
    if((cx+(j-1)*300)<1500){segments(cx+(j-1)*300,cy,cx+(j-1)*300,cy+200,col='blue')
      xcons<-P$x[(P$y<=cy+200)&(P$y>cy)]
      ycons<-P$y[(P$y<=cy+200)&(P$y>cy)]
      dist<-abs((cx+(j-1)*300)-xcons)
      det <- exp(-dist^2/(2*40^2)) > runif(length(dist))
      det_trans<-c(det_trans,sum(det[det=TRUE]))
      det_dist<-c(det_dist,dist[det])
      points(xcons[det], ycons[det], pch=19, cex=.5, col="red")}}}#figura 4.19
n<-length(det_dist);n
#ajuste half normal
sighn<-sqrt(sum(det_dist^2)/n);sighn
ghn<-function(x){exp(-x^2/(2*(sighn^2)))}##g est
mghn<-integrate(ghn,lower=0,upper = max(det_dist))$value;mghn
abghn<-(n*(1500*1000))/(2*mghn*25*200);abghn#abundancia est
ghn<-function(x){exp(-x^2/(2*(sighn^2)))}##g est
fhn<-function(x){exp(-x^2/(2*(sighn)^2))/mghn}###f est}
FHN<-function(x){integrate(fhn,lower = 0,upper = x)$value}
IFHN<-inverse(FHN,0,max(det_dist))
par(mfrow=c(1,2))
weighted.hist(det_dist,rep(mghn/(length(det_dist)*13),length(det_dist)),
              breaks = seq(0,130,by=13),xlim=c(0,130),xaxis=F,
              main=expression(paste(hat(g),'(x) (Half-normal)'))
              ,xlab='Distancia',ylab='')
axis(side=1, at=seq(0, 130, by=26))
curve(ghn,add=T,from = 0,to=130)
hist(det_dist,prob  = T,breaks=seq(0,130,by=13),axes = T,
     main=expression(paste(hat(f),'(x) (Half-normal)')),
     xlab='Distancia',ylab='')
curve(fhn,from = 0,to=130,add = T)#figura 4.20
par(mfrow=c(1,1))
#ajuste hazard rate
DAT<-as.data.frame(det_dist)
names(DAT)<-'distance'
modhrl<-ds(DAT,model=df(model=~hr),truncation = max(det_dist))
sighr<-exp(modhrl$par[1]);sighr
bghr<-exp(modhrl$par[2]);bghr
ghr<-function(x){1-exp(-(x/sighr)^(-bghr))}
mghr<-integrate(ghr,lower=0,upper = max(det_dist))$value ;mghr
abghr<-(length(det_dist)*(1500*1000))/(2*mghr*25*200);abghr
fhr<-function(x){(1-exp(-(x/sighr)^(-bghr)))/mghr}
FHR<-function(x){integrate(fhr,lower = 0,upper = x)$value}
IFHR<-inverse(FHR,0,max(det_dist))
par(mfrow=c(1,2))
weighted.hist(det_dist,rep(mghr/(length(det_dist)*13),
                           length(det_dist)),breaks = seq(0,130,by=13),
              xlim=c(0,130),xaxis=F,main=expression(paste(hat(g),'(x) (Hazard-rate)')),
              xlab='Distancia',ylab='')
axis(side=1, at=seq(0, 130, by=26))
curve(ghr,add=T,from = 0,to=130)
hist(det_dist,prob  = T,breaks=seq(0,130,by=13),axes = F,
     main=expression(paste(hat(f),'(x) (Hazard-rate)')),
     xlab='Distancia',ylab='',xlim = c(0,130),ylim=c(0,0.025))
axis(side=2, at=seq(0, 0.025, by=0.005))
```

```
184 axis(side=1, at=seq(0, 130, by=26))
    curve(fhr,from = 0,to=130,add = T)#figura 4.21
    par(mfrow=c(1,1))
    #Elección del modelo
188 #Gráficos P-P
    emp<-ecdf((det_dist));ord<-sort(det_dist)
    a1<-emp(ord);bhn<-c();bhr<-c()
    for (i in 1:n){bhn[i]<-FHN(ord[i]);bhr[i]<-FHR(ord[i])}
192 par(mfrow=c(1,2))
    plot(a1,bhn,ylim=c(0,1),cex=.6,pch=19,xlab='Distribución empírica',
         ylab=expression(paste("Distribución ajustada ",hat(F))),
         main='P-P plot half-normal')
196 abline(a=0,b=1,lwd=1)
    plot(a1,bhr,ylim=c(0,1),cex=.6,pch=19,xlab='Distribución empírica',
         ylab=expression(paste("Distribución ajustada ",hat(F))),
         main='P-P plot hazard-rate')
200 abline(a=0,b=1,lwd=1)
    par(mfrow=c(1,1))#figura 4.22
    #Kolmogorov-Smirnov y Cramér-von Mises
    modhnl<-ds(DAT)
204 kehn<-gof_tests(modhnl,plot = F)$kolmogorov_smirnov$Dn
    kphn<-gof_tests(modhnl,plot = F)$kolmogorov_smirnov$p
    kehn;kphn
    cehn<-gof_tests(modhnl,plot = F)$cramer_vonMises$W
208 cphn<-gof_tests(modhnl,plot = F)$cramer_vonMises$p
    cehn;cphn
    kehr<-gof_tests(modhrl,plot = F)$kolmogorov_smirnov$Dn
    kphr<-gof_tests(modhrl,plot = F)$kolmogorov_smirnov$p
212 kehr;kphr
    cehr<-gof_tests(modhrl,plot = F)$cramer_vonMises$W
    cphr<-gof_tests(modhrl,plot = F)$cramer_vonMises$p
    cehr;cphr
216 #Prueba de la Ji-cuadrada
    #half normal
    obs<-c();esp<-c();cut<-c()
    cut[6]<-max(det_dist)
220 for (i in 1:5){cut[i]<-IFHN(.20*(i-1))}
    for (i in 1:5){
      obs[i]<-length(det_dist[(det_dist>=cut[i])&(det_dist<=cut[i+1])])
      esp[i]<-(FHN(cut[i+1])-FHN(cut[i]))*n}
224 ECH<-sum(((obs-esp)^2)/esp); ECH
    qchisq(.95,length(obs)-1-1)
    pv<-1-pchisq(ECH,length(obs)-1-1);pv
    #hazard rate
228 obs<-c();esp<-c();cut<-c()
    cut[6]<-max(det_dist)
    for (i in 1:5){ cut[i]<-IFHR(0.20*(i-1))}
    for (i in 1:5){
232   obs[i]<-length(det_dist[(det_dist>=cut[i])&(det_dist<=cut[i+1])])
      esp[i]<-(FHR(cut[i+1])-FHR(cut[i]))*n
    }
    ECH<-sum(((obs-esp)^2)/esp); ECH
236 qchisq(.95,length(obs)-1-2)
    pv<-1-pchisq(ECH,length(obs)-1-2);pv
    #AIC
    #half normal
240 hnaic<- -2*sum(log(fhn(det_dist)))+2
```

```
#hazard rate
hraic<- -2*sum(log(fhr(det_dist)))+4
hnaic;hraic
#elegimos el modelo hazard rate
OBSTL<-function(){
  cx<-c();cy<-c();xcons<-c(); ycons<-c()
  dist<-c();det_trans<-c(); det<-c();det_dist<-c()
  for (i in 1:5){
    cy<-200*(i-1)
    for (j in 1:5){
      if ((i%%2)==0){ cx<-75*3}
      else {cx<-75 }
      if((cx+(j-1)*300)<1500){
        xcons<-P$x[(P$y<=cy+200)&(P$y>cy)]
        ycons<-P$y[(P$y<=cy+200)&(P$y>cy)]
        dist<-abs((cx+(j-1)*300)-xcons)
        det <- exp(-dist^2/(2*40^2)) > runif(length(dist))
        det_trans<-c(det_trans,sum(det[det=TRUE]))
        det_dist<-c(det_dist,dist[det])}}}
  n<-length(det_dist)
  return(list(n,det_dist))}
#Estimación por intervalos
#bootstrap no paramêtrico
MBNP<-c()
for (i in 1:1000){
  rem<-sample(det_trans,25,replace = T)
  MBNP<-c(MBNP,(sum(rem)*(1500*1000))/(2*mghr*25*200))}
CIBNMP<-quantile(MBNP,c(0.025,.975));CIBNMP
sd(MBNP)#desviación estándar
#1000 simulaciones
Nhattl<-c()
for (i in 1:1000){
  m<-OBSTL()
  DAT<-as.data.frame(m[[2]])
  names(DAT)<-'distance'
  modhrl<-ds(DAT,model=df(model=~hr),truncation = max(m[[2]]))
  sighr<-exp(modhrl$par[1])
  bghr<-exp(modhrl$par[2])
  ghr<-function(x){1-exp(-(x/sighr)^(-bghr))}
  mghr<-integrate(ghr,lower=0,upper = max(m[[2]]))$value
  abghr<-(m[[1]]*(1500*1000))/(2*mghr*25*200)
  Nhattl[i]<-abghr}
hist(Nhattl,breaks=20,xlim=c(800,1100),xlab = expression(hat(N)),ylab='',main='')
abline(v=c(870,mean(Nhattl)),lwd=c(3,2),lty=c(1,2))#figura 4.23
sd(Nhattl)
mean(Nhattl)
sd(Nhattl)/mean(Nhattl)#coeficiente de variación

##4.3.2.2 MUESTREO POR TRANSECTOS PUNTUALES
set.seed(1958)
plot(x,y,type="points",pch=19,col='dark gray',cex=.5,axes=F)
axis(side=2, at=seq(0, 1000, by=200))
axis(side=1, at=seq(0, 1500, by=300))
cx<-c();cy<-c();xcons<-c();ycons<-c()
dist<-c();det_trans<-c();det<-c();det_dist<-c()
for (i in 1:9){
  cy<-100*(i)
```

```
  for (j in 1:5){
    if ((i%%2)==0){ cx<-75*3}
    else {cx<-75}
    if((cx+(j-1)*300)<1500){points(cx+(j-1)*300,cy,col='blue',pch=19)
      xcons<-P$x[(P$y<=cy+100)&(P$y>cy-100)]
      ycons<-P$y[(P$y<=cy+100)&(P$y>cy-100)]
      dist<-sqrt((cx+(j-1)*300 - xcons)^2+(cy-ycons)^2)
      det <- exp(-dist^2/(2*35^2)) > runif(length(dist))
      det_trans<-c(det_trans,sum(det[det=TRUE]))
      det_dist<-c(det_dist,dist[det])
      points(xcons[det], ycons[det], pch=19, cex=.5, col="red")}}}#figura 4.24
n<-length(det_dist);n
#ajuste half normal
sighn<-sqrt(sum(det_dist^2)/(2*n));sighn
ghn<-function(x){exp(-x^2/(2*(sighn^2)))}##g est
integ<-function(x){2*pi*x*exp(-x^2/(2*(sighn^2)))}
mughn<-integrate(ghn,lower = 0,upper = max(det_dist))$value;mughn
nughn<-integrate(integ,lower=0,upper = max(det_dist))$value;nughn
Pahn<-nughn/(pi*(max(det_dist)^2))
abghn<-(n*(1500*1000))/(45*nughn);abghn
fhn<-function(x){(2*pi*x*exp(-x^2/(2*(sighn^2))))/nughn}#f est
FHN<-function(x){integrate(fhn,lower = 0,upper = x)$value}
IFHN<-inverse(FHN,0,max(det_dist))
par(mfrow=c(1,2))
weighted.hist(det_dist,(1/det_dist)*(mughn/(sum(1/det_dist)*10)),
              breaks = seq(0,100,by=10),xlim=c(0,100),xaxis=F,
              main=expression(paste(hat(g),'(x) (Half-normal)')),
              xlab='Distancia',ylab='',ylim = c(0,1))
axis(side=1, at=seq(0, 100, by=20))
curve(ghn,add=T,from = 0,to=100)
hist(det_dist,prob  = T,main=expression(paste(hat(f),'(x) (Half-normal)')),breaks = 10,
     xlab='Distancia',ylab='',xlim=c(0,100),ylim=c(0,0.022),axes=F)
axis(side=2, at=seq(0, 0.022, by=0.005))
axis(side=1, at=seq(0, 100, by=20))#figura 4.25
curve(fhn,from = 0,to=100,add = T)
par(mfrow=c(1,1))
#ajuste hazard rate (libreria Distance2)
DAT<-as.data.frame(det_dist)
names(DAT)<-'distance'
mhrp<-ds(DAT,transect="point",model=df(model=~hr),truncation = max(det_dist))
sighr<-exp(mhrp$par[1]);sighr
bghr<-exp(mhrp$par[2]);bghr
ghr<-function(x){1-exp(-(x/sighr)^(-bghr))}
integ1<-function(x){2*pi*x*(1-exp(-(x/sighr)^(-bghr)))}
mughr<-integrate(ghr,lower=0,upper = max(det_dist))$value;mughr
nughr<-integrate(integ1,lower=0,upper = max(det_dist))$value;nughr
abghr<-(n*(1500*1000))/(45*nughr);abghr
fhr<-function(x){2*pi*x*(1-exp(-(x/sighr)^(-bghr)))/nughr}
FHR<-function(x){integrate(fhr,lower = 0,upper = x)$value}
IFHR<-inverse(FHR,0,max(det_dist))
par(mfrow=c(1,2))
weighted.hist(det_dist,(1/det_dist)*(mughr/(sum(1/det_dist)*10)),
              breaks = seq(0,100,by=10),xlim=c(0,100),xaxis=F,
              main=expression(paste(hat(g),'(x) (Hazard-rate)')),
              xlab='Distancia',ylab='',ylim = c(0,1))
axis(side=1, at=seq(0, 100, by=20))
curve(ghr,add=T,from = 0,to=100)
```

```
hist(det_dist,prob  = T,main=expression(paste(hat(f),'(x) (Hazard-rate)'))
     ,xlab='Distancia',ylab='',xlim=c(0,100),ylim=c(0,0.022),axes=F,breaks = 10)
axis(side=2, at=seq(0, 0.022, by=0.005))
axis(side=1, at=seq(0, 100, by=20))
curve(fhr,from = 0,to=100,add = T)#figura 4.26
par(mfrow=c(1,1))
#Elección del modelo
#Gráficos P-P
emp<-ecdf((det_dist));ord<-sort(det_dist)
a1<-emp(ord);bhn<-c();bhr<-c()
for (i in 1:n){bhn[i]<-FHN(ord[i]);bhr[i]<-FHR(ord[i])}
par(mfrow=c(1,2))
plot(a1,bhn,ylim=c(0,1),cex=.6,pch=19,xlab='Distribución empírica',
     ylab=expression(paste("Distribución ajustada ",hat(F))),
     main='P-P plot half-normal')
abline(a=0,b=1,lwd=1)
plot(a1,bhr,ylim=c(0,1),cex=.6,pch=19,xlab='Distribución empírica',
     ylab=expression(paste("Distribución ajustada ",hat(F))),
     main='P-P plot hazard-rate')
abline(a=0,b=1,lwd=1)#figura 4.27
par(mfrow=c(1,1))
###KS TEST Y CRAMER VONMISES
mhnp<-ds(DAT,transect = "point",truncation = max(det_dist))
kehn<-gof_tests(mhnp,plot = F)$kolmogorov_smirnov$Dn
kphn<-gof_tests(mhnp,plot = F)$kolmogorov_smirnov$p
kehn;kphn
kehr<-gof_tests(mhrp,plot = F)$kolmogorov_smirnov$Dn
kphr<-gof_tests(mhrp,plot = F)$kolmogorov_smirnov$p
kehr;kphr
cehn<-gof_tests(mhnp,plot = F)$cramer_vonMises$W
cphn<-gof_tests(mhnp,plot = F)$cramer_vonMises$p
cehn;cphn
cehr<-gof_tests(mhrp,plot = F)$cramer_vonMises$W
cphr<-gof_tests(mhrp,plot = F)$cramer_vonMises$p
cehr;cphr
##prueba ji cuadrada
##HALF NORM
obs<-c();esp<-c();cut<-c()
cut[6]<-max(det_dist)
for (i in 1:5){cut[i]<-IFHN((.20)*(i-1))}
for (i in 1:5){
  obs[i]<-length(det_dist[(det_dist>=cut[i])&(det_dist<=cut[i+1])])
  esp[i]<-(FHN(cut[i+1])-FHN(cut[i]))*n}
ECH<-sum(((obs-esp)^2)/esp); ECH
qchisq(.95,length(obs)-1-1)
pv<-1-pchisq(ECH,length(obs)-1-1);pv
###HAZARD RATE
obs<-c();esp<-c();cut<-c()
cut[6]<-max(det_dist)
for (i in 1:5){cut[i]<-IFHR((.20)*(i-1))}
for (i in 1:5){
  obs[i]<-length(det_dist[(det_dist>=cut[i])&(det_dist<=cut[i+1])])
  esp[i]<-(FHR(cut[i+1])-FHR(cut[i]))*n}
ECH<-sum(((obs-esp)^2)/esp); ECH
qchisq(.95,length(obs)-1-2)
pv<-1-pchisq(ECH,length(obs)-1-2);pv
##AIC
```

```
##HALF NORMAL
hnaic<- -2*sum(log(fhn(det_dist)))+2
##hazard rate
hraic<- -2*sum(log(fhr(det_dist)))+4
hnaic;hraic
#elegimos el modelo half normal
OBSTP<-function(){cx<-c()
cy<-c();xcons<-c();ycons<-c()
dist<-c();det_trans<-c();det<-c();det_dist<-c()
for (i in 1:9){
  cy<-100*(i)
  for (j in 1:5){
    if ((i%%2)==0){ cx<-75*3}
    else {cx<-75}
    if((cx+(j-1)*300)<1500){
      xcons<-P$x[(P$y<=cy+100)&(P$y>cy-100)]
      ycons<-P$y[(P$y<=cy+100)&(P$y>cy-100)]
      dist<-sqrt((cx+(j-1)*300 - xcons)^2+(cy-ycons)^2)
      det <- exp(-dist^2/(2*35^2)) > runif(length(dist))
      det_trans<-c(det_trans,sum(det[det=TRUE]))
      det_dist<-c(det_dist,dist[det])}}}
n<-length(det_dist)
return(list(n,det_dist))}
#Estimación por intervalos
#bootstrap no paramétrico
MBNP<-c()
for (i in 1:1000){
  rem<-sample(det_trans,45,replace = T)
  MBNP<-c(MBNP,(sum(rem)*(1500*1000))/(45*nughn))}
CIBNMP<-quantile(MBNP,c(0.025,.975));CIBNMP
sd(MBNP)#desviación estándar
mean(MBNP)
#1000 simulaciones
Nhattp<-c()
for (i in 1:1000){
  m<-OBSTP()
  sighn<-sqrt(sum(m[[2]]^2)/(2*m[[1]]));sighn
  ghn<-function(x){exp(-x^2/(2*(sighn^2)))}##g estimada
  integ<-function(x){2*pi*x*exp(-x^2/(2*(sighn^2)))}
  mughn<-integrate(ghn,lower = 0,upper = max(m[[2]]))$value
  nughn<-integrate(integ,lower=0,upper = max(m[[2]]))$value
  Nhattp[i]<-(m[[1]]*(1500*1000))/(45*nughn)}
hist(Nhattp,prob=F,xlab=expression(hat(N)),ylab='',main='')
abline(v=c(870,mean(Nhattp)),lwd=c(3,2),lty=c(1,2))#figura 4.28
sd(Nhattp)
mean(Nhattp)
sd(Nhattp)/mean(Nhattp)#coeficiente de variación

##4.3.3 VECINO MÁS CERCANO
set.seed(157)
plot(x,y,pch=19,col='dark grey',cex=0.5)
h<-runif(30)*1500;k<-runif(30)*1000
cx<-c();cy<-c();xcons<-c();ycons<-c()
dist<-c();det_dist<-c()
for (i in 1:length(h)){
    cy<-k[i]
    points(h[i],k[i],col='blue',pch=19)
```

```
      xcons<-P$x[(P$y<=cy+160)&(P$y>cy-160)]
      ycons<-P$y[(P$y<=cy+160)&(P$y>cy-160)]
      dist<-sqrt((h[i] - xcons)^2+(k[i]-ycons)^2)
      points(xcons[which(dist==min(dist))],ycons[which(dist==min(dist))],
             col='red',pch=19,cex=.5)
      det_dist<-c(det_dist,min(dist)) }# figura 4.29
ngnn<-30*1500*1000/(pi*sum(det_dist^2));ngnn #abundancia est
dgnn<-ngnn/(1500*1000);dgnn #densidad est
#Estimación por intervalos
#bootstrap paramétrico
finv<-function(x){sqrt((-log(1-x))/(pi*dgnn))}
estim<-function(x){(30*1500*1000)/(pi*sum(x^2))}
munif<-c(); Nbp<-c()
  for (i in 1:10000){
    munif<-runif(30)
    Nbp[i]<-estim(finv(munif))}
quantile(Nbp,probs = c(0.025,.975))
#bootstrap no paramêtrico
Nbnp<-c();mubnp<-c();ddbnp<-c()
  for(i in 1:1000){
    mubnp<-sample(1:30,30,replace = T)
    ddbnp<-det_dist[mubnp]
    Nbnp[i]<-estim(ddbnp)}
quantile(Nbnp,probs = c(0.025,.975))
mean(Nbnp)
sd(Nbnp)#desviación estándar
#1000 simulaciones
  OBSVC<-function(){h<-runif(30)*1500
  k<-runif(30)*1000
  cx<-c();cy<-c();xcons<-c(); ycons<-c()
  dist<-c();det_dist<-c()
  for (i in 1:length(h)){
    cy<-k[i]
    xcons<-P$x[(P$y<=cy+160)&(P$y>cy-160)]
    ycons<-P$y[(P$y<=cy+160)&(P$y>cy-160)]
    dist<-sqrt((h[i] - xcons)^2+(k[i]-ycons)^2)
    det_dist<-c(det_dist,min(dist))}
  det_dist}
  NVC<-c()
  for (i in 1:1000){
    NVC[i]<-30*1000*1500/(pi*sum(OBSVC()^2))}
hist(NVC,xlim=c(0,1000),main='',ylab='',xlab=expression(hat(N)))
clip(0,2000,0,3120)
abline(v=c(870,mean(Nbnp)),lwd=c(3,2),lty=c(1,2))# figura 4.30
mean(NVC)
sd(NVC)/mean(NVC) #coeficiente de variación
```

Bibliografía

[1] Borchers, D. L., Buckland, S. T., & Zucchini, W. (2010). *Estimating animal abundance: Closed populations*. London: Springer.

[2] Buckland, S. T., Anderson, D. R., & Burnham, K. P. (2010). *Advanced distance sampling*. Oxford: Oxford University Press.

[3] Buckland, S. T., Anderson, D. R., Burnham, K. P., & Laake, J. L. (2014). *Distance sampling: Estimating abundance of biological populations*. New York: Springer-Science Business Media.

[4] Casella, G., & Berger, R. L. (2002). *Statistical infence*. Australia: Duxbury.

[5] Efron, B. (2000). *An introduction to the bootstrap*. Chapman & Hall/CRC: Boca Raton, Florida.

[6] Fox, G. A., Negrete-Yankelevich, S., & Sosa, V. J. (2015). *Ecological statistics: Contemporary theory and application*. Oxford: Oxford University Press.

[7] Gibbons, J. D., & Chakraborti, S. (2011). *Nonparametric statistical inference*. Boca Raton, FL: CRC Press.

[8] Gore, A., & Paranjpe, S. (2001). *A course in mathematical and statistical ecology*. Dordrecht: Kluwer Academic.

[9] Pastor, J. (2008). *Mathematical ecology of populations and ecosystems*. Oxford: Wiley-Blackwell.

[10] Sprent, P., & Smeeton, N. C. (2007). *Applied nonparametric statistical methods*. Boca Raton: Chapman et Hall/CRC.