



UNIVERSIDADE DA CORUÑA

Universidade de Vigo

Estimación no paramétrica de la distribución  
diamétrica de *Pinus radiata* D. Don en el noroeste  
de España

*Trabajo Fin de Máster*

Manuel Arias Rodil

*Universidade da Coruña*

*Facultade de Informática*

*Máster en Técnicas Estadísticas*

---

Curso 2012/2013

Directores

Ricardo Cao Abad  
Ulises Diéguez Aranda

A Coruña, julio de 2013



UNIVERSIDADE DA CORUÑA

Universidade de Vigo

Estimación no paramétrica de la distribución  
diamétrica de *Pinus radiata* D. Don en el noroeste  
de España

*Trabajo Fin de Máster*

Manuel Arias Rodil  
*Universidade da Coruña*  
*Facultade de Informática*  
*Máster en Técnicas Estadísticas*

---

Curso 2012/2013

Vº Bº  
Los directores

Ricardo Cao Abad

Ulises Diéguez Aranda

A Coruña, julio de 2013

## Índice

<b>Resumen</b>	<b>2</b>
<b>1. Introducción</b>	<b>3</b>
<b>2. Objetivos</b>	<b>4</b>
<b>3. Datos</b>	<b>4</b>
<b>4. Métodos</b>	<b>6</b>
4.1. Estimación no paramétrica . . . . .	6
4.1.1. Elección de los parámetros ventana . . . . .	8
4.2. Estimación paramétrica . . . . .	9
4.3. Estimación no paramétrica vs. paramétrica . . . . .	14
<b>5. Resultados</b>	<b>16</b>
5.1. Estimación no paramétrica . . . . .	16
5.2. Estimación paramétrica . . . . .	17
5.3. Estimación no paramétrica vs. paramétrica . . . . .	17
<b>6. Discusión</b>	<b>20</b>
<b>7. Conclusiones</b>	<b>24</b>
<b>Bibliografía</b>	<b>26</b>

## Resumen

En el campo forestal, conocer la distribución del diámetro normal de los árboles (diámetro del tronco medido a 1.3 m desde el suelo) resulta de gran interés para clasificar la producción (p. ej., en volumen de madera) por categorías de tamaños. Para ello, se pueden emplear metodologías paramétricas (las más habituales) o no paramétricas, que estiman la distribución del diámetro a partir de variables inherentes a la masa forestal (p. ej., número de árboles por hectárea, altura dominante o edad). El objetivo principal de este trabajo es la estimación no paramétrica de la distribución condicional del diámetro de la especie *Pinus radiata* D. Don en el noroeste de España, empleando una adaptación del estimador propuesto por Li y Racine (2008), basado en el de Nadaraya-Watson. Esta metodología se compara con la alternativa paramétrica de recuperación de parámetros mediante el método de los momentos, una de las más utilizadas en modelización forestal. Para la comparación entre métodos, se utilizan la distancia de Kolmogorov-Smirnov y el criterio de Cramér-von Mises, evaluando si existen diferencias significativas mediante el empleo del test de rangos con signo de Wilcoxon (1945). Los datos provienen de 557 parcelas instaladas en rodales regulares de edad conocida en las que se midió el diámetro normal y la altura total de todos los árboles, a partir de las cuales se calcularon variables de rodal. Según el test empleado, el método no paramétrico es mejor, por lo que se recomienda su empleo para estimar la distribución diamétrica de *Pinus radiata* en el noroeste de España. No obstante, en casos en los que la mayoría de los árboles son muy gruesos o muy delgados, el método paramétrico resulta más adecuado.

**Palabras o frases clave:** Diámetro normal, distribución condicional, no paramétrico, Nadaraya-Watson, método de los momentos, *Pinus radiata* D. Don.

## 1. Introducción

La necesidad de predecir el crecimiento y la producción de los árboles y las masas forestales, así como su respuesta a la aplicación de distintos tratamientos, ha sido y continúa siendo uno de los objetivos fundamentales de la investigación forestal, ya que un elemento fundamental para realizar una gestión correcta es conocer adecuadamente los procesos de desarrollo de las distintas especies (Diéguez-Aranda et al., 2009, p. 11). Para ello se emplean modelos que predicen variables de interés para la gestión, como son el volumen de madera en pie, la cantidad de biomasa o el carbono secuestrado. Estos modelos utilizan como *inputs* variables que se pueden medir fácilmente en árboles individuales o en grupos de árboles (rodales).

Las variables fundamentales que se miden en los árboles son el diámetro normal del tronco (a 1.30 m sobre el suelo) y la altura total. En ocasiones se miden también otras variables, como la altura de copa viva, la anchura de copa o el diámetro a diferentes alturas. Los diámetros normales se suelen agrupar en clases (normalmente de una amplitud de 1 a 5 cm) cuyo diámetro central se denomina marca de clase.

Generalmente, para desarrollar modelos de crecimiento y producción de una especie forestal en un área geográfica concreta, se instala una red de parcelas de superficie conocida que se distribuyen cubriendo el rango existente de edades, densidades y calidades de estación. La técnica más habitual para estimar la calidad de estación de un rodal se basa en el análisis de la evolución con la edad de la altura media de los árboles dominantes, que se denomina altura dominante. Para referenciar la calidad de estación utilizando la relación altura dominante-edad en rodales regulares (aquellos en los que al menos el 90 % de los árboles pertenecen a la misma clase de edad) se suele utilizar el denominado índice de sitio, que se define como el valor de su altura dominante a una determinada edad base o de referencia (Diéguez-Aranda et al., 2009, p. 50).

En el caso de los modelos de rodal, las variables de interés que se obtienen como *outputs* se pueden desagregar por clases de tamaño si se conoce la distribución de los diámetros normales, por lo que la estimación de dicha distribución tiene gran interés en la gestión forestal, permitiendo una valoración más fiable de los recursos forestales de una masa. En 1898, de Liocourt se interesó por primera vez en la estimación de la distribución de diámetros, concretamente en rodales forestales irregulares (aquellos que presentan árboles de todas las clases de edad), empleando para ello la función de distribución exponencial. A lo largo de los años, se han utilizado otras familias de distribuciones como la *gamma* (Nelson, 1964), la log-normal (Bliss y Reinker, 1964), la *beta* (Clutter y Bennett, 1965), la Weibull (Bailey y Dell, 1973; Cao, 2004) o la  $S_B$  de Johnson (Hafley y Schreuder, 1977), proporcionando mejores resultados las dos últimas.

Por otra parte, algunos autores también han utilizado métodos no paramétricos para la estimación de la distribución diamétrica de rodales forestales (Droessler y Burk, 1989; Maltamo y Kangas, 1998; Niggemeyer y Schmidt, 1999), para lo que han empleado técnicas basadas en el *k*-vecino más cercano, aunque en algunos casos se trata más bien de soluciones semiparamétricas. La estimación no paramétrica no emplea modelos y se basa en los propios datos, por lo que cabría esperar que si se dispone de una muestra representativa y suficientemente grande de la población, debería proporcionar mejores resultados que la estimación paramétrica. En ambos casos, la finalidad es poder predecir la distribución diamétrica del rodal sin tener que medir los diámetros de todos los árboles que lo componen.

La especie objeto de este trabajo es el pino insigne (*Pinus radiata* D. Don). Esta especie

procede del suroeste de América del Norte, y ha tenido una gran expansión fuera de su área natural, al ser introducida en plantaciones con fines productivos en otros países, hasta el punto de convertirse en la conífera exótica más plantada en todo el mundo (Lavery, 1986). Esta gran expansión se debe principalmente a su elevado crecimiento en climas húmedos, la versatilidad de su madera, su fácil propagación (posibilidad de recoger grandes cantidades de semilla), la relativa diversidad genética dentro de sus poblaciones naturales (proporciona genotipos adecuados para ambientes distintos) y su gran plasticidad y flexibilidad selvícola, que permite practicar distintos tratamientos sin que la producción se vea sensiblemente afectada Castedo, 2004. Los países con mayores superficies plantadas de pino insigne son Nueva Zelanda, Chile, Australia, España y Sudáfrica (Mead, 2013).

Este trabajo se centra en el área de distribución de la especie en Galicia y Asturias, en las que *Pinus radiata* es una de las coníferas de mayor importancia, ocupando un total 96,177 ha en Galicia (CIFOR-INIA, 2011, p. 16), lo que representa un 6.79 % de la superficie forestal arbolada de la comunidad, y 25,386 ha en Asturias (5.63 % de su superficie forestal arbolada) (CIFOR-INIA, 2012, p. 24).

## 2. Objetivos

El objetivo principal de este trabajo es el de proponer un método estadístico nuevo para la desagregación de variables de interés de una masa forestal por clases diamétricas. Para ello, se considera la distribución de probabilidad del diámetro condicionada a variables de rodal.

Desde el punto de vista forestal el objetivo es obtener una estimación más fiable de la distribución diamétrica que la que ofrecen los métodos más utilizados actualmente, permitiendo una mejor desagregación de variables de rodal.

Desde el punto de vista estadístico aparecen varios objetivos. El primero es el empleo de un método no paramétrico en la estimación de una distribución condicional, seleccionando los parámetros ventana óptimos. Por otra parte, se desea comparar el comportamiento de la alternativa propuesta con el método paramétrico más utilizado, por lo que es necesario ajustar los modelos paramétricos correspondientes y proponer un procedimiento de comparación entre métodos.

En conclusión, se definen dos objetivos fundamentales: la estimación no paramétrica de la distribución del diámetro de *Pinus radiata* en el noroeste de España y la comparación con uno de los métodos paramétricos más utilizados en el campo forestal, la función Weibull predicha por el método de los momentos.

## 3. Datos

Para la realización de este trabajo se han utilizado datos de 25,620 árboles de la especie *Pinus radiata*, que fueron medidos en 283 parcelas instaladas en rodales regulares de edad conocida en Galicia y Asturias por miembros de la *Unidade de Xestión Forestal Sostible* (Escuela Politécnica Superior, Universidade de Santiago de Compostela), del Departamento de Ingeniería y Ciencias Agrarias (Campus de Ponferrada, Universidad de León) y del Departamento de Biología de Organismos y Sistemas (Escuela Politécnica de Mieres, Universidad de Oviedo). En

cada árbol se realizaron como mínimo mediciones del diámetro normal (centímetros) y la altura total (metros). Se dispone de datos de al menos tres variables por parcela: edad (años), densidad (número de pies por hectárea) y altura dominante (metros).

La mayoría de las parcelas han sido remedidas al menos una vez, con un intervalo de tiempo respecto a la primera medición de como mínimo un año. Con el paso del tiempo, algunos árboles se mueren y otros se cortan, lo que implica una disminución de la densidad del rodal, mientras que los árboles vivos crecen, lo que supone un incremento en la altura dominante. Por ello, pese a que los datos de una misma parcela medidos en diferentes edades pueden presentar cierta dependencia, cada combinación parcela-inventario se ha considerado como un conjunto de datos con una distribución diamétrica independiente, debido a que el número de remediciones es pequeño en comparación con el número total de las observaciones (Castedo et al., 2006). Asumiendo este supuesto, el número de observaciones de diámetro normal asciende a 42,194 y el de combinaciones parcela-inventario a 557 (en adelante nos referiremos a ellas simplemente como parcelas). En la Tabla 1 se muestra un resumen de las principales variables de árbol y de rodal de los datos utilizados. En la Figura 1 se muestran los histogramas de las variables de rodal de la base de datos.

Tabla 1: Estadísticos descriptivos de la muestra de datos utilizados.

Variable	Mínimo	Media	Máximo	Desviación típica
Densidad (nº pies/ha)	200	1280	4864	515.6
Edad (años)	5	23.6	47	8.4
Altura dominante <sup>1</sup> (m)	5.8	21.0	35.8	5.7
Diámetro normal (cm)	1.0	19.6	81.0	9.7

<sup>1</sup> Media de las alturas de los 100 árboles más gruesos (de mayor diámetro normal) por hectárea.

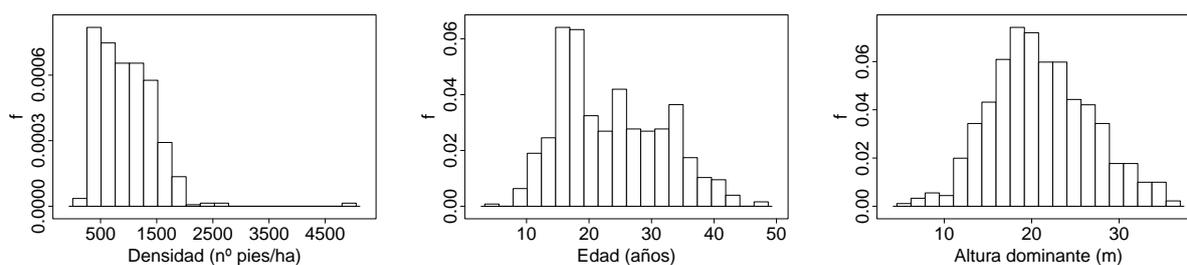


Figura 1: Histogramas de densidad, edad y altura dominante.

Las observaciones de diámetro han sido discretizadas en centímetros, ya que dicho grado de precisión es suficiente para la predicción del sistema de desagregación. En la Figura 2 se muestran, a modo de ejemplo, los histogramas de diámetro de cuatro parcelas seleccionadas aleatoriamente.

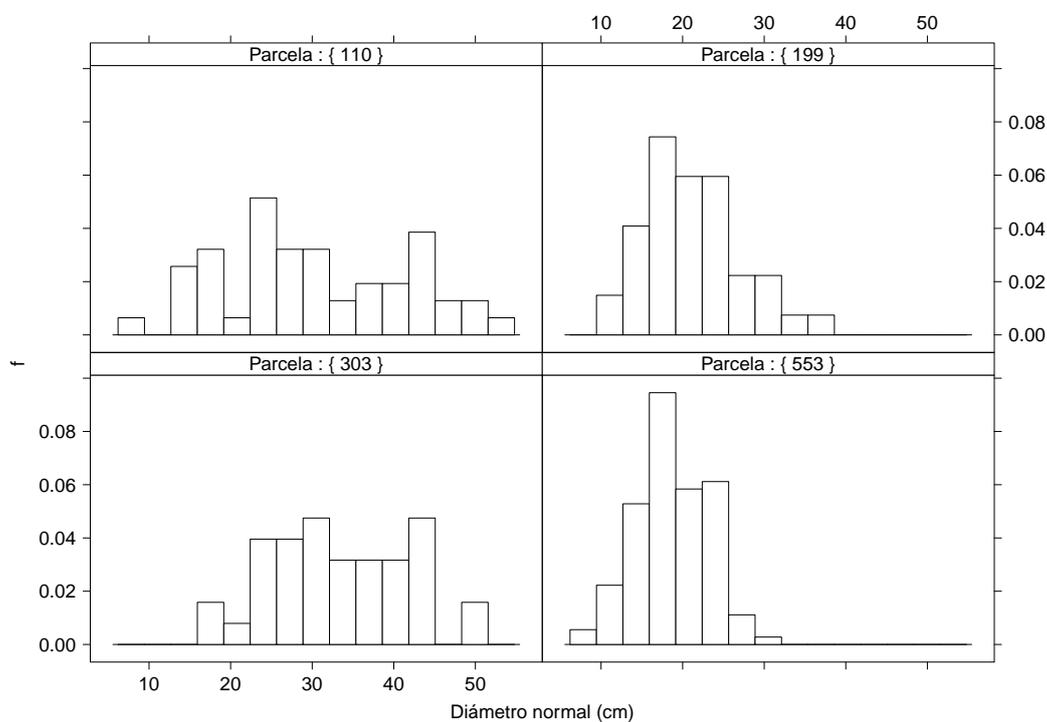


Figura 2: Histogramas de diámetro de cuatro parcelas seleccionadas aleatoriamente.

## 4. Métodos

### 4.1. Estimación no paramétrica

En el campo forestal, estimar la distribución del diámetro de una masa a partir de variables de rodal, es decir, sin tener que medir el diámetro de todos los árboles, resulta de gran interés para reducir el coste del inventario. En este sentido, se emplea la distribución condicional, que se basa en la estimación de la distribución de una variable condicionada a una o más covariables, ya sean discretas o continuas. La estimación de la distribución condicional ha sido tratada por diferentes autores a lo largo de los años (Hall et al., 1999; Cai, 2002; Hansen, 2004), utilizando el estimador de Nadaraya-Watson (Nadaraya, 1964; Watson, 1964) con una ponderación para cada observación basada en un estimador lineal local.

En este trabajo se estima la distribución del diámetro de *Pinus radiata* condicionada a covariables de rodal. El estimador empleado se basa en el propuesto por Li y Racine (2008), que a su vez se basa en el de Nadaraya-Watson, adaptándolo para el caso en que se utiliza más de una covariable y todas ellas son continuas. Su expresión es:

$$\hat{F}_{\mathbf{H}}(y | \mathbf{x}) = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} I_{\{y_{ij} \leq y\}} K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i)}{\sum_{i=1}^k n_i K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i)} \quad (1)$$

siendo  $y$  la variable para la que queremos estimar la distribución (en este caso el diámetro normal),  $\mathbf{x}$  el vector de covariables de rodal de la parcela objetivo,  $k$  el número de parcelas,  $n_i$  el número de árboles de la parcela  $i$  ( $n = \sum_{i=1}^k n_i$  es el tamaño muestral global),  $y_{ij}$  el valor del diámetro del árbol  $j$  de la parcela  $i$ ,  $\mathbf{x}_i$  el vector de covariables de rodal de la parcela  $i$ ,  $I$  la función indicadora (toma valor 1 si la condición del subíndice de  $I$  es cierta y 0 en el caso contrario) y  $K_{\mathbf{H}}$  el núcleo multivariante reescalado según la matriz de parámetros ventana  $\mathbf{H}$  definida positiva. Concretamente,  $K_{\mathbf{H}}(\mathbf{u}) = \frac{1}{\det(\mathbf{H})} K(\mathbf{H}^{-1}\mathbf{u})$ , siendo  $K$  una función núcleo de variable escalar (típicamente una función de densidad) y  $\mathbf{u}$  un vector. En el estimador (1), dicho núcleo otorga peso a cada parcela en función del valor de sus covariables y del de las de la parcela para la que se quiere estimar la distribución diamétrica.

En el estimador utilizado, la contribución de cada diámetro al estimador ( $\sum_{j=1}^{n_i} I_{\{y_{ij} \leq y\}}$ ) se pondera con el peso de la parcela a la que pertenece dicha observación, dentro del total de las parcelas, con arreglo al valor de las covariables  $\mathbf{x}$  de dicha parcela. Reordenando los términos de la expresión (1) se obtiene:

$$\hat{F}_{\mathbf{H}}(y | \mathbf{x}) = \sum_{i=1}^k \left( \sum_{j=1}^{n_i} I_{\{y_{ij} \leq y\}} \frac{K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i)}{\sum_{l=1}^k n_l K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_l)} \right) \quad (2)$$

El peso de la parcela  $i$  con respecto a la parcela objetivo ( $K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i)$ ) se obtiene con una función núcleo (*kernel*) multivariante típicamente no negativa, que debe cumplir:

$$\int K(\mathbf{x}) d\mathbf{x} = 1$$

El núcleo multivariante suele ser simétrico y unimodal (Cline, 1988). Una elección habitual es el núcleo producto que se construye multiplicando varios núcleos univariantes (Wand y Jones, 1995, p. 91):

$$K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i) = \prod_{r=1}^d \frac{1}{h_r} L\left(\frac{x_i - x_{i,r}}{h_r}\right) \quad (3)$$

siendo  $h_r$  el parámetro ventana correspondiente a la covariable  $r$  de la parcela  $i$  y  $L$  la función núcleo univariante, es decir,  $K(\mathbf{u}) = \prod_{r=1}^d L(u_r)$ . El término  $x_{i,r}$  de la expresión (3) denota el valor de la  $r$ -ésima covariable correspondiente a la parcela  $i$ -ésima.

En este caso, en lugar de un único parámetro ventana, se utiliza una matriz de parámetros ventana de dimensión  $d \times d$ , siendo  $d$  el número de covariables utilizadas. Por simplicidad, se ha empleado una matriz diagonal (ver expresión (4)), lo que significa que se considera un único parámetro ventana por covariable, omitiendo los correspondientes a las combinaciones de covariables.

$$\mathbf{H} = \begin{pmatrix} h_1 & 0 & \dots & 0 \\ 0 & h_2 & \vdots & \vdots \\ \vdots & \dots & \ddots & \vdots \\ 0 & \dots & \dots & h_d \end{pmatrix} \quad (4)$$

El núcleo univariante empleado en este trabajo ha sido la función de densidad normal estándar, aunque existen otros como el de Epanechnikov (1969), el triangular o el uniforme (Silverman, 1986, p. 40). Según el mismo Silverman (1986, p. 40), la elección del núcleo no es tan determinante como la de los parámetros ventana en la estimación de la distribución (en este trabajo se emplea para otorgar pesos a las parcelas).

#### 4.1.1. Elección de los parámetros ventana

La función núcleo multivariante con la que se obtienen los pesos asignados a cada parcela en función de las covariables correspondientes (ver expresión (3)), requiere la elección óptima de parámetros ventana, lo que constituye un aspecto clave en la estimación no paramétrica. Los valores de los parámetros ventana influyen en el grado de suavización de los pesos ( $K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i)$ ), de tal forma que valores de  $h$  grandes generarán un estimador muy suavizado (sesgo grande y varianza pequeña) y valores de  $h$  pequeños un estimador muy “ruidoso” (sesgo pequeño y varianza grande) (Härdle, 1991, p. 48). Partiendo de este supuesto, los valores de  $h$  óptimos deberían alcanzar un compromiso entre sesgo y varianza.

Los selectores de parámetros ventana se pueden dividir, *grosso modo*, en dos clases: una en la que se emplean fórmulas simples, sin garantías matemáticas de que el valor de  $h$  sea el óptimo pero rápidas en lo que a computación se refiere, y otra en la que el valor de  $h$  óptimo se obtiene basándose en algún criterio de error a minimizar (p. ej., error cuadrático medio integrado, MISE), por lo que son más fiables pero a costa de un mayor tiempo de computación (Wand y Jones, 1995, p. 59).

Se han desarrollado métodos efectivos para la obtención de parámetros ventana óptimos para la estimación de densidades incondicionales, funciones de media condicional (es decir, funciones de regresión) y densidades condicionales (ver, p. ej., Jones et al., 1996; Hall et al., 2004; Hall et al., 2007). Con respecto a la distribución condicional, Li y Racine (2008) mencionan que la obtención de parámetros ventana seguía siendo un tema abierto, aunque posteriormente Li et al. (2013) proponen un método automático basado en los datos. En él, el valor de  $h$  óptimo se calcula minimizando el error medio cuadrático (MSE) ponderado, estimado mediante la siguiente función de validación cruzada:

$$CV(\mathbf{H}) = n^{-1} \sum_{i=1}^k \sum_{j=1}^{n_i} \int \left\{ I(y_{ij} \leq y) - \hat{F}_{\mathbf{H}}^{-i}(y | \mathbf{x}_i) \right\}^2 dy \quad (5)$$

siendo  $\hat{F}_{\mathbf{H}}^{-i}(y | \mathbf{x}_i)$  la estimación de la distribución de  $y$  condicionada a la parcela  $i$ , a partir de la información de todas las parcelas excepto la  $i$ .

Dado que la expresión (5) debe calcularse por integración numérica, su minimización puede resultar muy lenta computacionalmente. Para reducir el tiempo de computación del proceso de obtención de los valores óptimos de los parámetros ventana, se discretizó la integral en tantos puntos como diámetros diferentes existen en la base de datos, que representan los puntos de cambio tanto de  $I(y_{ij} \leq y)$  como de  $\hat{F}_{\mathbf{H}}^{-i}(y | \mathbf{x}_i)$ .

En las expresiones (1) y (2) se podría emplear una función núcleo en lugar de la función indicadora, con lo que la estimación de la distribución para una nueva parcela estaría suavizada, es decir, no habría saltos en cada punto de cambio. En este contexto, sería necesario también obtener un parámetro ventana óptimo para la variable de interés (diámetro normal), y al suavizar la estimación, no sería posible la discretización que se ha comentado en el párrafo anterior, por variar en cualquier valor y no solamente para los diferentes diámetros de la base de datos.

## 4.2. Estimación paramétrica

El método de estimación paramétrica empleado en este trabajo se basa en la función de distribución de Weibull, que fue utilizada por primera vez en la estimación de la distribución diamétrica de árboles por Bailey y Dell (1973). La expresión de su función de densidad es:

$$f(y) = \frac{c}{b} \left( \frac{y-a}{b} \right)^{c-1} \exp \left( - \left( \frac{y-a}{b} \right)^c \right), \text{ si } y \geq a, 0 \text{ en otro caso} \quad (6)$$

siendo  $y$  la variable de interés,  $a$  un parámetro de localización,  $b$  un parámetro de escala y  $c$  un parámetro de forma.

Si se fija a cero el parámetro de localización ( $a = 0$ ), se obtiene la función de densidad Weibull biparamétrica, lo que facilita la estimación de los parámetros  $b$  y  $c$  sin afectar a la precisión de las estimaciones (Maltamo et al., 1995):

$$f(y) = \frac{c}{b} \left( \frac{y}{b} \right)^{c-1} \exp \left( - \left( \frac{y}{b} \right)^c \right), \text{ si } y \geq 0, 0 \text{ en otro caso.} \quad (7)$$

Integrando la función de densidad biparamétrica se obtiene la siguiente función de distribución:

$$F(y) = 1 - \exp \left( - \left( \frac{y}{b} \right)^c \right), \text{ si } y \geq 0, 0 \text{ en otro caso.} \quad (8)$$

El método paramétrico empleado se basa en “recuperar” los parámetros  $b$  y  $c$  a partir de los momentos de la distribución diamétrica, y se conoce como método de los momentos, habiendo sido empleado con éxito por varios autores (Newby, 1980; Burk y Newberry, 1984; Cao, 2004). En los siguientes párrafos se explica este método en detalle.

Sea  $D$  una variable aleatoria con distribución Weibull, la esperanza (E) de la distribución Weibull se obtiene integrando la función de densidad de la expresión 7:

$$E(D) = \int_{-\infty}^{\infty} Df(D)dD = \int_{-\infty}^{\infty} D \frac{c}{b} \left( \frac{D}{b} \right)^{c-1} \exp \left( - \left( \frac{D}{b} \right)^c \right) dD \quad (9)$$

siendo  $b$  y  $c$  parámetros de la función de densidad de Weibull.

Si se realiza el siguiente cambio de variable:  $z = (D/b)^c$  (expresión (10)), se llega a la expresión (11), de la esperanza.

$$D = bz^{1/c} \text{ y } dz = \frac{c}{b} \left(\frac{D}{b}\right)^{c-1} dD \quad (10)$$

$$E(D) = b \int_{-\infty}^{\infty} z^{1/c} \exp(-z) dz \quad (11)$$

La integral de la expresión (11) tiene ciertas similitudes con la función *gamma* ( $\Gamma(k)$ ):

$$\Gamma(k) = \int_0^{\infty} w^{k-1} \exp(-w) dw \quad (12)$$

siendo  $w$  la variable de interés y  $k$  la variable de la que depende la función.

El límite inferior de integración es 0, al contrario que en la expresión (11), pero la función de densidad biparamétrica de Weibull solamente toma valores positivos, por lo que los límites son equivalentes, llegando finalmente a la siguiente expresión de esperanza de una variable aleatoria  $D$  con distribución Weibull:

$$E(D) = b \Gamma\left(1 + \frac{1}{c}\right) \quad (13)$$

En cuanto a la varianza (Var) de dicha variable, se obtiene como la diferencia entre la esperanza del cuadrado de la variable y el cuadrado de la esperanza de la variable:

$$\text{Var}(D) = E(D^2) - E(D)^2 \quad (14)$$

La expresión de  $E(D)^2$  se obtiene de manera simple, elevando al cuadrado la expresión (13). Por su parte,  $E(D^2)$  resulta de aplicar el mismo procedimiento realizado para  $D$  (ver expresiones (9), (10), (11), (12) y (13)) pero para la variable  $D^2$ , obteniendo la siguiente expresión:

$$E(D^2) = b^2 \Gamma\left(1 + \frac{2}{c}\right) \quad (15)$$

Se llega así a la expresión de varianza (Var) de una variable aleatoria  $D$  con distribución Weibull:

$$\text{Var}(D) = b^2 \Gamma\left(1 + \frac{2}{c}\right) - E(D)^2 \quad (16)$$

Si sustituimos  $E(D)$  por su expresión (13) se obtiene:

$$\text{Var}(D) = b^2 \left( \Gamma\left(1 + \frac{2}{c}\right) - \Gamma^2\left(1 + \frac{1}{c}\right) \right) \quad (17)$$

A su vez, si despejamos el parámetro  $b$  de la expresión (13), substituyéndolo posteriormente en la expresión (17), se llega a una forma en la que la varianza depende solamente del valor de  $c$ :

$$\text{Var}(D) = \frac{E(D)^2}{\Gamma^2\left(1 + \frac{1}{c}\right)} \left( \Gamma\left(1 + \frac{2}{c}\right) - \Gamma^2\left(1 + \frac{1}{c}\right) \right) \quad (18)$$

Los momentos poblacionales son desconocidos, por lo que es preciso aproximarlos mediante los correspondientes estimadores muestrales: la media ( $\bar{d}$ ) y la varianza ( $s_d^2$ ) muestrales. Así, substituyendo los valores muestrales en las expresiones (13) y (18) respectivamente, se llega a las siguientes expresiones:

$$\bar{d} = b \Gamma\left(1 + \frac{1}{c}\right) \quad (19)$$

$$s_d^2 = \frac{\bar{d}^2}{\Gamma^2\left(1 + \frac{1}{c}\right)} \left( \Gamma\left(1 + \frac{2}{c}\right) - \Gamma^2\left(1 + \frac{1}{c}\right) \right) \quad (20)$$

Fijémonos que la esperanza del cuadrado de la variable ( $E(D^2)$ ) se correspondería con el estimador muestral del cuadrado de la media cuadrática ( $d_c^2$ ). La varianza muestral se podría obtener como la diferencia entre el cuadrado de la media cuadrática ( $d_c$ ) y aritmética ( $\bar{d}$ ), como se ha comentado anteriormente para los valores poblacionales. Así, se pueden obtener estos dos estimadores (expresiones (21) y (22)) y a partir de ellos la varianza (expresión (23)).

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n} \quad (21)$$

$$d_c = \sqrt{\frac{\sum_{i=1}^n d_i^2}{n}} \quad (22)$$

$$s_d^2 = d_c^2 - \bar{d}^2 \quad (23)$$

siendo  $n$  el número de observaciones y  $d_i$  el valor de la variable  $d$  medido en el individuo  $i$ .

Este desarrollo se puede adaptar directamente a la variable objeto de este trabajo, en el que se hablará de diámetro medio cuadrático ( $d_g$ ) y aritmético ( $d_m$ ) (se emplean  $d_g$  y  $d_m$  por ser la notación habitual en el campo forestal para dichas variables). Así, estimando  $s_d^2$  a partir de  $d_g$  y  $d_m$ , se llega a un sistema de dos ecuaciones con dos incógnitas (parámetros  $b$  y  $c$ ).

$$\begin{cases} d_m = b \Gamma\left(1 + \frac{1}{c}\right) \\ s_d^2 = \frac{d_m^2}{\Gamma^2\left(1 + \frac{1}{c}\right)} \left( \Gamma\left(1 + \frac{2}{c}\right) - \Gamma^2\left(1 + \frac{1}{c}\right) \right) \end{cases} \quad (24)$$

En el sistema de ecuaciones de la expresión (24) se puede obtener el valor del parámetro  $c$  mediante un procedimiento numérico como el algoritmo de Brent (1973), que combina el método de la bisección, el método secante e interpolación inversa cuadrática. Este algoritmo está implementado en la función `uniroot` de R (R Core Team, 2012). Posteriormente, y substituyendo el valor obtenido de  $c$  en la primera expresión del sistema de ecuaciones, se recupera el valor del parámetro  $b$ .

Los diámetros medios aritmético y cuadrático muestrales se pueden calcular directamente si se ha medido el diámetro normal de todos los árboles de una parcela, lo que permite la recuperación de parámetros para dicha parcela. Sin embargo, la metodología de desagregación tiene sentido para estimar la distribución del diámetro sin haber medido esta variable, reduciendo así el coste de los trabajos de inventario, por lo que es necesario relacionar los diámetros medios con variables de rodal fácilmente medibles en campo.

El diámetro medio cuadrático se explica a partir de diferentes variables de rodal, como pueden ser la edad ( $t$ ), la altura dominante ( $H_0$ ), el espaciamiento medio ( $\frac{100}{\sqrt{N}}$ ) o el índice de sitio ( $IS$ ). En este contexto, se propone un modelo paramétrico que relacione la variable dependiente con las variables de rodal. Para ello, se han analizado diferentes modelos, y tras un análisis gráfico de la nube de puntos de variable dependiente frente a las independientes (Figura 3), se propone un modelo alométrico (expresión (25)). El espaciamiento medio está relacionado con la densidad ( $N$ ) y representa la distancia media que separa a todos los árboles de una masa.

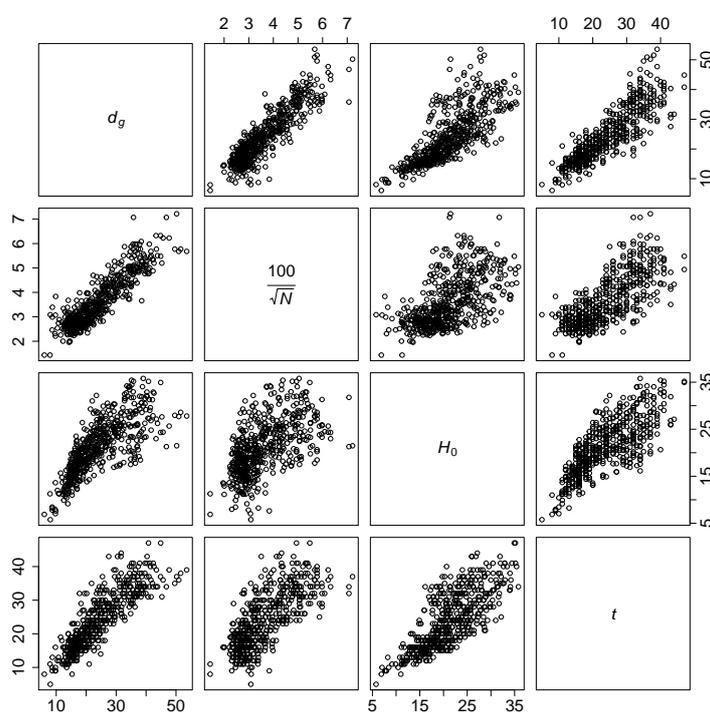


Figura 3: Gráficos de dispersión del diámetro medio cuadrático frente a  $\frac{100}{\sqrt{N}}$ ,  $H_0$  y  $t$ .

$$d_g = b_0 \prod_{m=1}^p x_m^{b_m} \quad (25)$$

siendo  $b_0$  un parámetro del modelo,  $x_m$  la variable de rodal  $m$ ,  $b_m$  el parámetro correspondiente a dicha variable y  $p$  el número de variables independientes consideradas en el modelo.

Por otra parte, considerando la elevada correlación que existen entre el diámetro medio aritmético y el diámetro medio cuadrático, y teniendo en cuenta que las estimaciones del primero deben ser siempre menores que las del segundo, se plantea un modelo de la forma (Frazier, 1981):

$$d_m = d_g - \exp(\mathbf{x}\beta) \quad (26)$$

siendo  $\mathbf{x}$  un vector de covariables y  $\beta$  los parámetros del modelo.

El vector  $\mathbf{x}$  puede estar compuesto por covariables de rodal ( $t$ , edad;  $H_0$ , altura dominante;  $N$ , densidad;  $IS$ , índice de sitio). De la misma forma que para el diámetro medio cuadrático, se ha analizado gráficamente la nube de puntos de la variable dependiente frente a varias variables independientes (Figura 4).

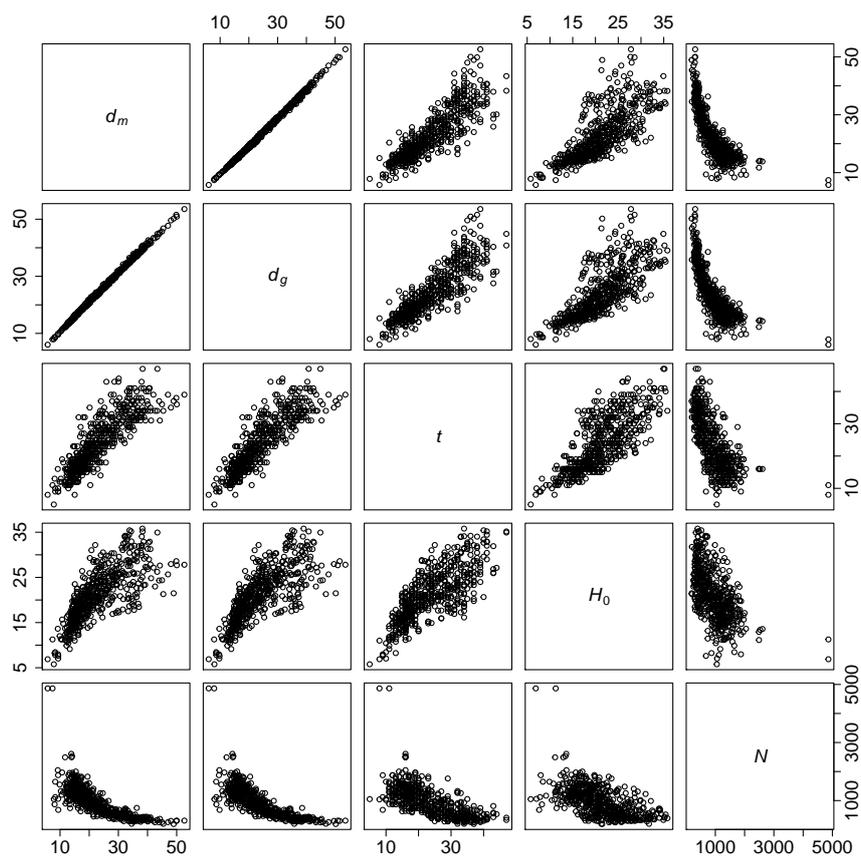


Figura 4: Gráficos de dispersión del diámetro medio cuadrático frente a  $d_g$ ,  $t$ ,  $H_0$  y  $N$ .

Los parámetros de los dos modelos propuestos se obtienen mediante regresión no lineal empleando la función `nlsLM` del paquete `minpack.lm` de R (R Core Team, 2012), que utiliza el algoritmo de Levenberg-Marquardt (Levenberg, 1944; Marquardt, 1963) para minimizar la suma de cuadrados del error. Este algoritmo interpola entre el algoritmo de Gauss-Newton (éste es el que utiliza la función `nls`, implementada en R) y el de descenso de gradiente. Este proceso es más robusto que el algoritmo de Gauss-Newton para alcanzar una solución, cuando se parte de parámetros de inicio alejados de los óptimos.

El análisis de la capacidad de ajuste de los modelos se basó en comparaciones numéricas y gráficas de los residuos. Se utilizaron dos criterios estadísticos: el coeficiente de determinación ( $R^2$ ) y la raíz del error medio cuadrático (REMC). Aunque existen opiniones fundamentadas que plantean dudas en relación con el empleo del  $R^2$  en la selección de modelos, este estadístico da una idea intuitiva de la variabilidad que explican. No obstante, nunca debe utilizarse como el único criterio para elegir el modelo que mejor predice entre un conjunto de modelos candidatos (Myers, 1990, p. 166). Además, pese a los inconvenientes asociados al uso del  $R^2$  en regresión no lineal, la utilidad general de emplear alguna medida de la adecuación global del modelo supera dichas limitaciones (Ryan, 1997, p. 424). El estadístico REMC resulta útil porque está expresado en las mismas unidades que la variable dependiente, por lo que da una idea del error medio que se comete con el modelo. Las expresiones de estos estadísticos son:

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (27)$$

$$\text{REMC} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - p}} \quad (28)$$

siendo  $Y_i$ ,  $\hat{Y}_i$  y  $\bar{Y}$  respectivamente los valores real, estimado y promedio la variable dependiente,  $n$  el número de observaciones y  $p$  el número de parámetros del modelo.

### 4.3. Estimación no paramétrica vs. paramétrica

En el método no paramétrico, la estimación de la distribución del diámetro requiere obtener el valor óptimo de los parámetros ventana. Por su parte, la alternativa paramétrica requiere ajustar los modelos de diámetros medios aritmético y cuadrático. Una vez que se ha realizado esto, para comparar ambas alternativas se necesita utilizar alguna medida de distancia o discrepancia entre la distribución empírica y las estimadas para cada parcela, que permita evaluar qué metodología ofrece mejores resultados.

En este trabajo se han elegido como medidas de discrepancia la distancia de Kolmogorov-Smirnov (KS) y el criterio de Cramér-von Mises (CvM), porque proporcionan respectivamente información de la bondad de la estimación en términos de máxima discrepancia y de diferencia acumulada. A menor valor de dichas medidas, mejor se ajustará la estimación a la distribución empírica. Las expresiones empleadas para cada medida de discrepancia se muestran a continuación:

$$\text{KS} = \sup_y \left| F_n(y) - \hat{G}(y) \right| \quad (29)$$

$$\text{CvM} = \int \left( F_n(y) - \hat{G}(y) \right)^2 dF_n(y) \quad (30)$$

siendo  $F_n(y)$  la distribución empírica de una parcela y  $\hat{G}(y)$  la distribución estimada.

Las dos medidas de discrepancia se computan por parcela, mediante validación cruzada, realizando 50 iteraciones<sup>1</sup> del siguiente proceso:

- Se seleccionan aleatoriamente el 20% del total de parcelas (en este caso 169), que constituyen la base de datos para validación.
- Las parcelas restantes constituyen la base de datos de entrenamiento, que se emplea para obtener los parámetros ventana óptimos ( $\mathbf{H}$ ) y ajustar los modelos paramétricos (expresiones (25) y (26)).
- Para cada parcela de validación, se calculan la distancia de Kolmogorov-Smirnov y el criterio de Cramér-von Mises de cada método, empleando los parámetros ventana óptimos y los modelos, obtenidos y ajustados respectivamente con la base de datos de entrenamiento.

El número de valores obtenidos de cada medida de discrepancia no es el mismo para todas las parcelas, ya que en cada iteración las parcelas de validación se eligen aleatoriamente. Por tal motivo, se promedian los resultados por parcela de cada medida de discrepancia, lo que permite otorgar el mismo peso a todas las parcelas en la comparación.

Si bien se dispone de las medidas de discrepancia entre métodos y se podría realizar la comparación directamente para cada parcela, es necesario emplear un test que permita evaluar si las diferencias entre ambos métodos son estadísticamente significativas. Dado que las medidas de discrepancia empleadas para diferentes iteraciones de la validación cruzada y diferentes parcelas pueden no seguir una distribución normal, se utilizó el test no paramétrico de los rangos con signo de Wilcoxon (1945).

Las hipótesis nula y alternativa para el test de Wilcoxon son:

$$H_0 : \mathbf{D}_{NP} \geq \mathbf{D}_P$$

$$H_1 : \mathbf{D}_{NP} < \mathbf{D}_P$$

siendo  $\mathbf{D}$  la medida de discrepancia a considerar (Kolmogorov-Smirnov o Cramér-von Mises) para el método no paramétrico ( $NP$ ) y paramétrico ( $P$ ).

La hipótesis nula corresponde a afirmar que la distancia del método no paramétrico es mayor o igual a la del método paramétrico. El estadístico de este test se obtiene de la siguiente forma:

- Se calculan las diferencias en valor absoluto entre dos muestras pareadas (en este caso las medidas de discrepancia para ambos métodos) de  $N$  observaciones:  $|\mathbf{D}_{NP} - \mathbf{D}_P|$ .

---

<sup>1</sup>El coste computacional de la validación cruzada es elevado debido a que en cada iteración se han de obtener los parámetros ventana óptimos, por lo que el número de iteraciones realizado no es muy alto.

- Se obtiene el rango de las diferencias en valor absoluto ( $R_i$ ): orden en la clasificación de menor a mayor valor absoluto.
- Se multiplica el rango de cada diferencia por el signo de ésta:  $R_i \text{ signo}(\mathbf{D}_{NP,i} - \mathbf{D}_{P,i})$ .
- El estadístico se obtiene sumando los rangos positivos:

$$W^+ = \sum R_i^+, \text{ siendo } R_i^+ = \begin{cases} R_i & \text{si } D_{NP,i} - D_{P,i} > 0 \\ 0 & \text{si } D_{NP,i} - D_{P,i} \leq 0 \end{cases}$$

El valor de estadístico  $W^+$  puede aproximarse por una distribución normal cuya media y varianza (bajo la hipótesis nula) son:

$$\mu_W = \frac{N(N+1)}{4} \quad (31)$$

$$\sigma_W^2 = \frac{N(N+1)(2N+1)}{24} \quad (32)$$

La hipótesis nula se rechazará para valores bajos del estadístico  $W^+$  ya que representa la cantidad de rangos positivos, y si este valor es bajo para la distribución normal considerada, quiere decir que la suma de rangos positivos es significativamente diferente de la suma de rangos negativos. El  $p$ -valor se extrae a partir del valor de  $w_{obs}^+$ , calculando  $P(w_{obs}^+ \leq W^+)$  para  $W^+$  que sigue una distribución  $N(\mu_W, \sigma_W)$ .

## 5. Resultados

En cada iteración del procedimiento de comparación se obtuvieron unos parámetros ventana óptimos y se realizó un ajuste de los modelos paramétricos, para cada combinación de parcelas de entrenamiento y validación. Se calculó un valor promedio de cada medida de discrepancia por método y parcela (en total 557 valores de cada medida), y se aplicó el test de rangos con signo de Wilcoxon para evaluar si las diferencias entre los dos métodos eran estadísticamente significativas. Los resultados de parámetros ventana óptimos y ajuste de modelos corresponden a la base de datos completa.

### 5.1. Estimación no paramétrica

En la Tabla 2 se muestran los parámetros ventana que se obtuvieron al minimizar la función de validación cruzada de la expresión (5). El valor mínimo de la función  $CV$  resultó 3.692.

Tabla 2: Parámetros ventana óptimos obtenidos mediante validación cruzada.

Covariable	Parámetro ventana	Valor
Altura dominante (metros)	$h_1$	1.614
Edad (años)	$h_2$	3.232
Densidad (número de pies/hectárea)	$h_3$	95.04

Para minimizar la función  $CV$  se empleó la función `constrOptim` de R (R Core Team, 2012), que permite realizar la optimización restringiendo los posibles valores de la variable decisoria (en este caso el tamaño de ventana  $h$ , que se restringió entre 0.001 y un valor cercano al máximo de la correspondiente variable de rodal). Se utilizó el algoritmo de Nelder-Mead (1965), que viene implementado por defecto.

Los valores de los parámetros ventana obtenidos guardan relación con la magnitud de las variables de rodal a las que afectan (Tabla 1). A mayor valor de la desviación típica de la covariable de rodal (Tabla 1), mayor es el valor del parámetro ventana correspondiente.

## 5.2. Estimación paramétrica

En los modelos finalmente ajustados, se emplearon las combinaciones de variables de rodal que proporcionaron mejores resultados en términos de error, descartando aquellos modelos en los que algún parámetro no fue significativo. El ajuste mediante regresión no lineal de los modelos de diámetro medio cuadrático ( $d_g$ , expresión (25)) y diámetro medio aritmético ( $d_m$ , expresión (26)) proporcionó los siguientes resultados:

$$d_g = 1.457 \left( \frac{100}{\sqrt{N}} \right)^{0.7137} t^{0.3038} H_0^{0.3018} \quad (33)$$

$$d_m = d_g - \exp \left( \frac{-9.127}{t} + 0.02057 H_0 \right) \quad (34)$$

Las estimaciones de todos los parámetros resultaron significativamente distintas de cero a un nivel de 0.00001. En lo relativo a los criterios de error obtenidos, el modelo ajustado de diámetro medio cuadrático proporcionó un  $R^2$  de 0.926 y un error medio cuadrático de 2.40 cm, mientras que el modelo ajustado correspondiente al diámetro medio aritmético explica un 99.8% de la variabilidad, con un REMC de 0.35 cm.

Las expresiones (33) y (34) se utilizaron para recuperar los parámetros  $c$  y  $b$  de la función de distribución Weibull biparamétrica, con el objetivo de comparar las estimaciones que proporciona con las del método no paramétrico propuesto.

## 5.3. Estimación no paramétrica vs. paramétrica

La primera comparación entre ambos métodos se realiza en términos de tiempo de computación. En el método no paramétrico, la selección de parámetros ventana óptimos es un proceso con un alto coste computacional (239,610.73 segundos en un ordenador con procesador Intel® Core™ 2 Duo a 3.00 GHz y 4.0 GB de RAM), tal y como se ha comentado, ya que está asociada a una función de validación cruzada. Por su parte, el ajuste de los modelos paramétricos es muy rápido (0.05 segundos en el mismo procesador). En este aspecto, existen diferencias claras entre ambos métodos, aunque lo realmente interesante es la comparación del tiempo de computación en el momento de aplicarlos, ya que los procesos comentados anteriormente solamente es necesario realizarlos una vez para un conjunto de datos. Para ello, se muestra en la Tabla 3 el resumen de los tiempos de ejecución (en segundos), por método, empleados en la estimación de la distribución de una parcela, aplicándola a toda la base de datos. Gráficamente, se pueden observar los

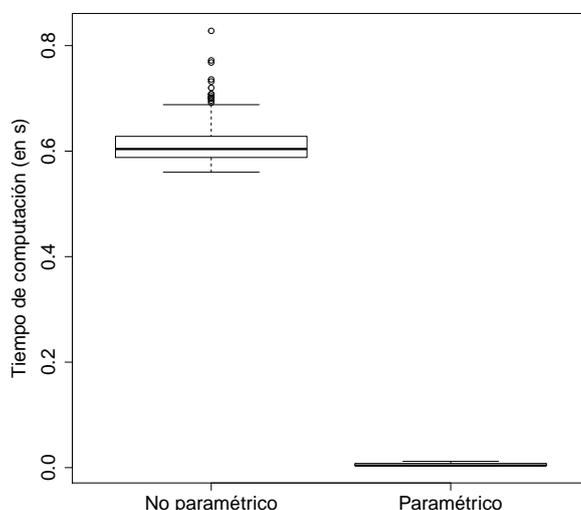


Figura 5: Gráfico de cajas de los tiempos de computación (en segundos) empleados por los métodos paramétrico y no paramétrico en la estimación de la distribución de todas las parcelas.

resultados obtenidos en la Figura 5. Los tiempos de ejecución se han obtenido en un ordenador con procesador Intel® Core™ i7 Q720 a 1.60 GHz y 8.0 GB de RAM.

Tabla 3: Resumen de los tiempos de computación (en segundos) empleados por los métodos paramétrico y no paramétrico en la estimación de la distribución de todas las parcelas.

Método	Mínimo	Media	Máximo
No paramétrico	0.560	0.612	0.828
Paramétrico	0.004	0.006	0.012

Las diferencias en el tiempo de ejecución de la aplicación de ambos métodos son altas, aunque los valores obtenidos son bajos y todos inferiores a un segundo, por lo que la eficiencia computacional de ambos métodos no es determinante para decantarse por uno u otro.

Para comparar la bondad del método propuesto con la alternativa paramétrica utilizada habitualmente, se emplean las medidas de discrepancia calculadas, seleccionando como mejor para estimar la distribución diamétrica el que menor valor proporcione de estas medidas. Los resultados obtenidos (por parcela) de las medidas de discrepancia se resumen, por método, en la Tabla 4. En la Figura 6 se representan dichos resultados en un gráfico de cajas, diferenciando también por método y medida de discrepancia. En el gráfico se ha empleado una escala logarítmica para facilitar la comparación entre métodos, debido a que la presencia de parcelas con valores muy altos de las medidas de discrepancia genera una desfiguración del conjunto de puntos.

Los valores medios y máximos obtenidos para el método no paramétrico son menores, mientras que con los mínimos sucede lo contrario. Observando el gráfico de cajas (Figura 6), las medidas de discrepancia del método no paramétrico son menores que las proporcionadas por la

Tabla 4: Resumen de los valores obtenidos para la distancia de Kolmogorov-Smirnov y Cramér-von Mises.

Medida de discrepancia	Método	Mínimo	Media	Máximo
Kolmogorov-Smirnov	Paramétrico	0.031	0.171	0.614
	No paramétrico	0.048	0.161	0.526
Cramér-von Mises	Paramétrico	0.0002	0.0150	0.2054
	No paramétrico	0.0004	0.0128	0.1622

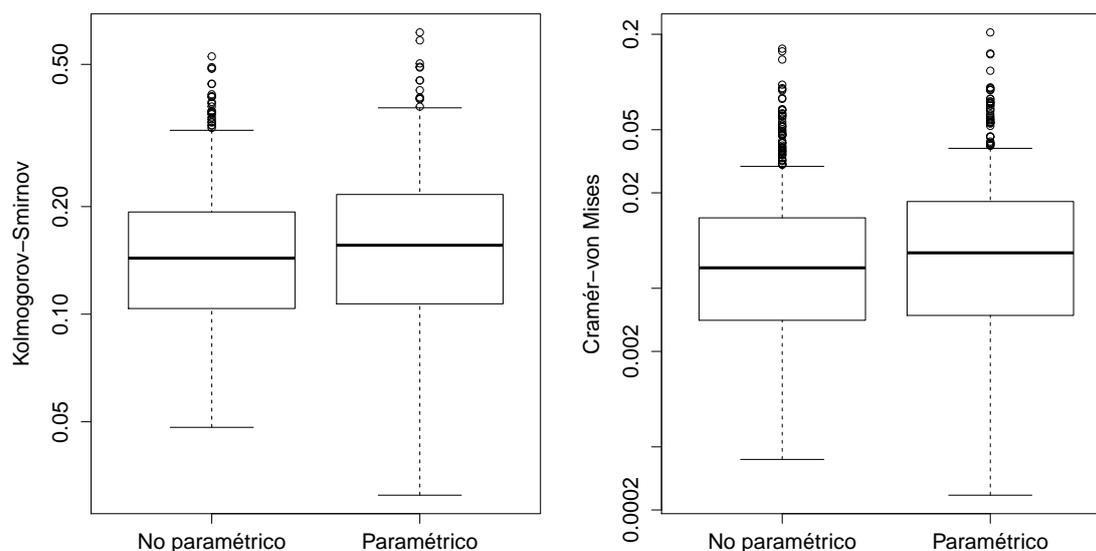


Figura 6: Gráfico de cajas por medidas de discrepancia (izquierda: Kolmogorov-Smirnov; derecha: Cramér-von Mises) y por método.

alternativa paramétrica. Por otra parte, en la Figura 7 se muestran los valores obtenidos de las medidas de discrepancia del método no paramétrico frente al paramétrico. También en este caso se ha empleado una escala logarítmica.

La nube de puntos es ligeramente más densa por debajo de la recta 1:1 en los dos gráficos, lo que significa que existe un mayor número de parcelas con valores más bajos de las medidas de discrepancia en el caso no paramétrico, situación que ya habíamos advertido en la Tabla 4. Tanto para la distancia de Kolmogorov-Smirnov como para el criterio de Cramér-von Mises, el método no paramétrico es mejor en aproximadamente el 57% de las parcelas.

En la Figura 8 se representan, a modo de ejemplo, la distribución empírica y las estimaciones no paramétrica y paramétrica para varias parcelas seleccionadas al azar.

Finalmente, se aplicó el test de rangos con signo de Wilcoxon a las medidas de discrepancia calculadas, para evaluar si las diferencias entre alternativas de estimación son significativas. En la Tabla 5 se muestra el estadístico obtenido y su  $p$ -valor asociado.

En ambas realizaciones del test, el  $p$ -valor asociado es muy bajo, por lo que a un nivel de

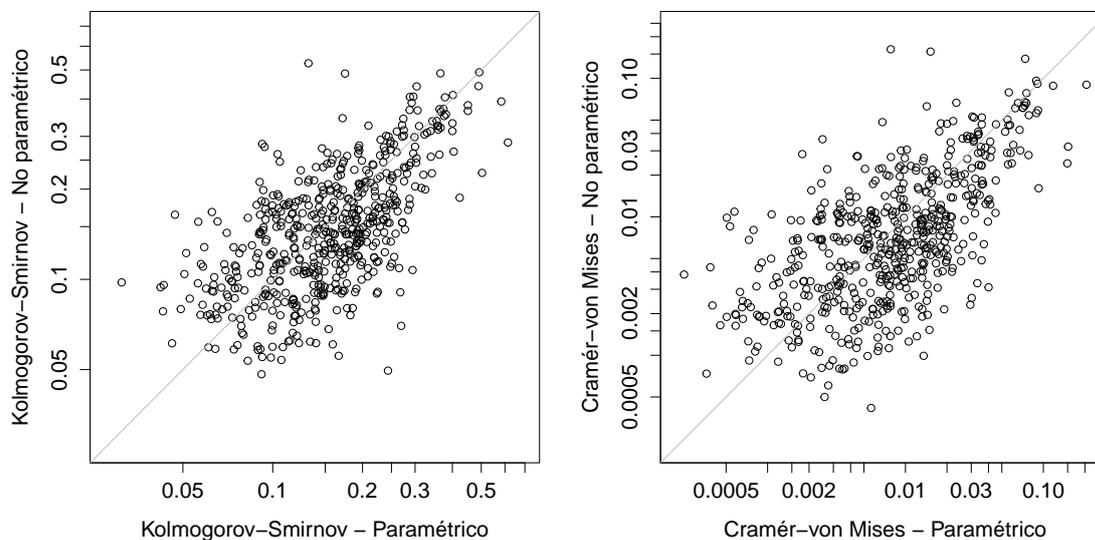


Figura 7: Dispersión de la distancia de Kolmogorov-Smirnov (izquierda) y el criterio de Cramér-von Mises (derecha) para los métodos no paramétrico (eje Y) y paramétrico (eje X) en escala logarítmica; recta 1:1 superpuesta.

Tabla 5: Resultados del test de rangos con signo de Wilcoxon para las medidas de discrepancia.

Medida de discrepancia	$W^+$	$p$ -valor
Kolmogorov-Smirnov	63,361	$8.04 \cdot 10^{-5}$
Cramér-von Mises	59,482	$8.15 \cdot 10^{-7}$

significación del 0.01 % se rechaza la hipótesis nula de que el método paramétrico sea mejor o igual que la alternativa no paramétrica.

## 6. Discusión

En el presente trabajo se ha comparado un método no paramétrico frente a otro paramétrico para estimar la distribución de diámetros de *Pinus radiata* en el noroeste de España. Según los resultados de las medidas de discrepancia utilizadas (Tabla 5), el método no paramétrico ha resultado mejor en un 57 % de los casos.

La estimación no paramétrica utiliza directamente la base de datos para estimar la distribución diamétrica. Entonces, las parcelas formadas mayoritariamente por árboles muy delgados o muy gruesos se encuentran en los extremos de los rangos de los diámetros disponibles. Así, la distribución estimada ofrece diámetros más altos o más bajos a los de la distribución empírica, según sean parcelas con árboles muy delgados o muy gruesos respectivamente, debido a la carencia de información en los extremos comentados. En la mayoría de los casos, la dimensión de los diámetros de los árboles está relacionada con la edad, excepto en algunos casos en los que la densidad es muy alta, y diámetros bajos responden a un efecto conjunto de la edad y la

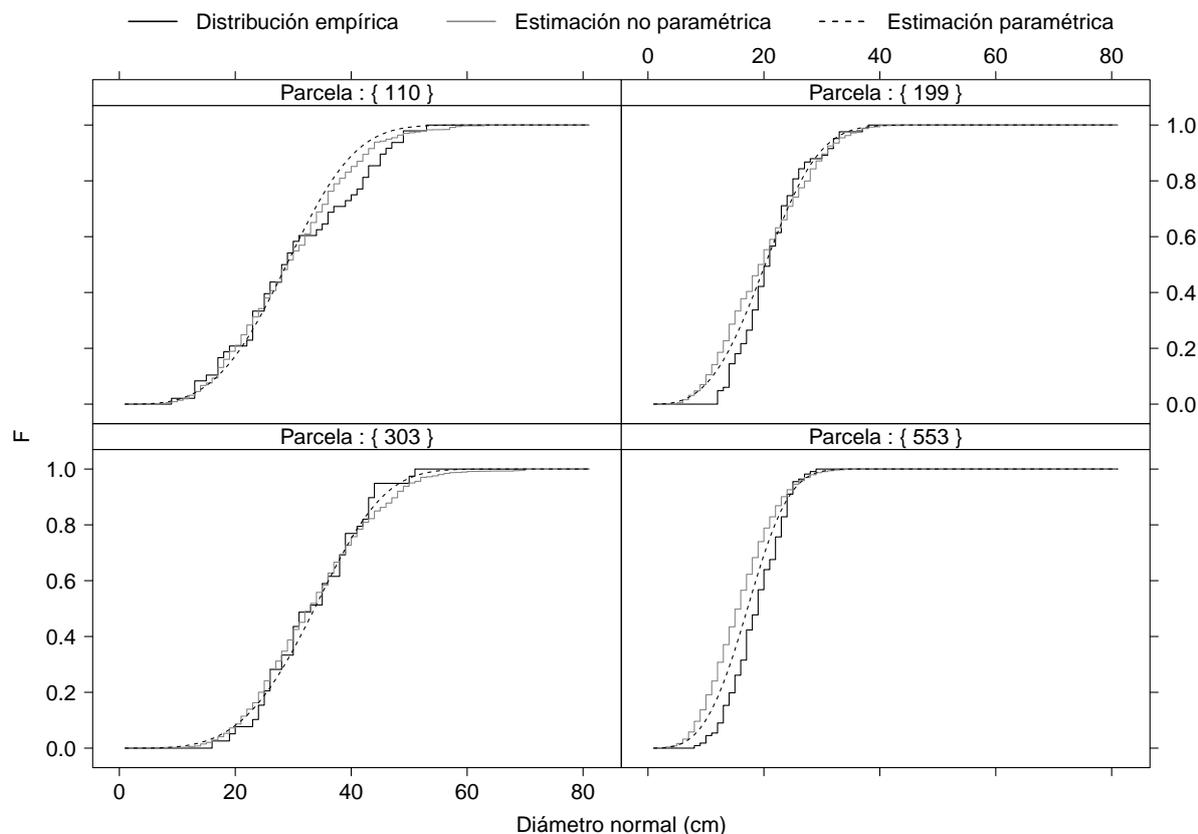


Figura 8: Ejemplos de las estimaciones no paramétrica y paramétrica superpuestas sobre la distribución empírica para las parcelas 110, 199, 303 y 553.

densidad.

En la Figura 9 se representan gráficamente las funciones de distribución empíricas de todas las parcelas, resaltando aquellas que ofrecen medidas de discrepancias altas para el método no paramétrico.

Los modelos paramétricos ajustados para obtener el diámetro medio cuadrático y el diámetro medio aritmético son la base del método paramétrico empleado. Así, la capacidad de predicción del método paramétrico viene determinada por la de estos modelos. En este sentido, para parcelas en las que la relación entre el diámetro medio cuadrático y las variables de rodal no se ajusta a la respuesta media del modelo, esta alternativa no ofrecerá estimaciones fiables de la distribución. El modelo ajustado que relaciona diámetro medio aritmético con diámetro medio cuadrático, altura dominante y edad proporciona buenos resultados ( $R^2 = 0.998$ , REMC = 0.35 cm), y el error cometido por éste es muy inferior al que proporciona el modelo que explica el diámetro medio cuadrático a partir de variables de rodal (REMC = 2.40 cm). Entonces, al evaluar el comportamiento del método paramétrico, será más conveniente analizar los resultados de la predicción del diámetro medio cuadrático.

El valor del diámetro medio cuadrático es indicativo del grosor o delgadez de los árboles de

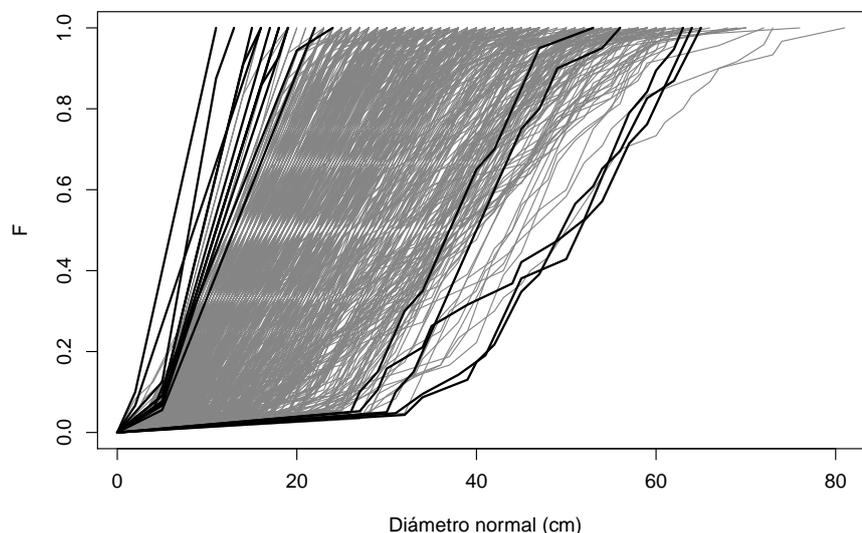


Figura 9: Funciones de distribución empíricas de todas las parcelas (en gris), resaltando las de las parcelas con medidas de discrepancias altas para el método no paramétrico (en negro).

una parcela, y existen algunas tendencias que cabría esperar en relación con las variables de rodal. En cuanto a la densidad, a mayor valor de ésta, los árboles tienden a ser más delgados por efecto de la competencia en altura, que provoca que los árboles destinen la mayoría de los recursos disponibles para el crecimiento en altura, en detrimento del crecimiento en grosor. Por otra parte, la altura dominante está relacionada directamente con el diámetro, de tal forma que árboles altos tienden a ser gruesos. Finalmente, la edad se relaciona con el grosor de los árboles de la misma forma a como lo hace la altura dominante.

Existen muchos casos en los que las variables no se ajustan exactamente a las tendencias comentadas anteriormente, observando, por ejemplo, árboles más delgados de lo que sería lógico para una densidad determinada (caso más común en la base de datos empleada). Esta situación puede venir provocada por haber realizado una clara en la masa en la que se encontraba la parcela. Este tratamiento consiste en eliminar algunos árboles de la masa, para reducir la competencia y favorecer el crecimiento de otros árboles, denominados de porvenir. El efecto inmediato es la reducción de la densidad de la masa, a lo que árboles deberían haber reaccionado creciendo en grosor, si bien es posible que en el momento de la toma de datos la clara hubiera sido realizada recientemente (reducción de densidad) y la reacción de los árboles a dicho tratamiento aún no se hubiese hecho patente (árboles no engrosados).

En la Figura 10 se representan gráficamente el valor del diámetro medio cuadrático frente a la densidad para todas las parcelas, resaltando aquellas que ofrecen una medida de discrepancia más alta para la alternativa paramétrica. Se observa que algunas de las parcelas muestran una relación entre diámetro medio cuadrático y densidad que sigue la respuesta media del modelo paramétrico ajustado, pero se alejan de ésta si consideramos las otras variables de rodal (altura

dominante y edad).

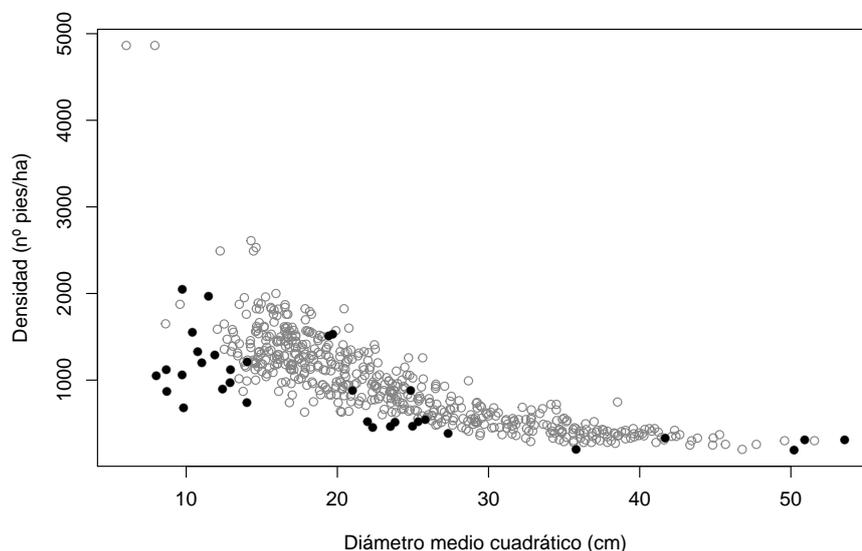


Figura 10: Densidad frente a diámetro medio cuadrático de todas las parcelas (en gris), superponiendo información de las que ofrecen una medida de discrepancia alta para el método paramétrico (en negro).

Observando la Figura 7 algunas parcelas se comportan de forma atípica, pudiendo distinguirse cuatro grupos:

1. Parcelas que pertenecen a la nube de puntos con valores de las medidas de discrepancia bajas y sin diferencias altas entre métodos.
2. Parcelas en las que tanto la estimación no paramétrica como la paramétrica se alejan de la distribución real.
3. Parcelas en las que la estimación no paramétrica proporciona medidas de discrepancia altas, siendo bajas para el método paramétrico.
4. Parcelas en las que la estimación paramétrica proporciona medidas de discrepancia altas, siendo bajas para el método no paramétrico.

La diferenciación entre grupos por su comportamiento en los dos métodos se muestra gráficamente en la Figura 11. Se trata de un gráfico de dispersión de los valores del criterio de Cramér-von Mises para ambos métodos. Esta medida ofrece información de la bondad de la estimación para todo el rango de la variable de interés, frente a la distancia de Kolmogorov-Smirnov que solamente proporciona información acerca de la diferencia máxima. Por ello, éste es el criterio empleado para diferenciar entre comportamientos.

La división de parcelas según su comportamiento se basa en lo expuesto en los párrafos anteriores. Si las parcelas no están formadas en su mayoría por árboles muy delgados o muy gruesos, y la relación del diámetro medio cuadrático con las variables de rodal se encuentra cerca

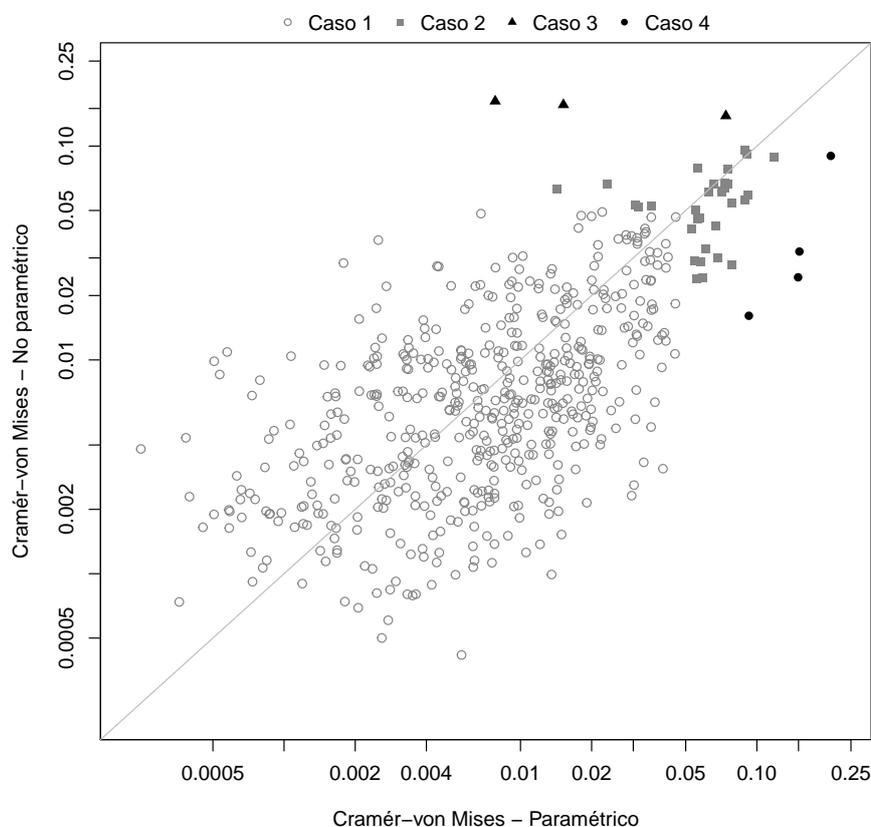


Figura 11: Representación gráfica de la medida de discrepancia de Cramér-von Mises de las parcelas para el método no paramétrico frente al paramétrico, clasificadas según su comportamiento para ambos métodos.

de la respuesta media de los modelos paramétricos ajustados, estas parcelas corresponderán a las incluidas en el caso 1. En cambio, si no se cumpliera cualquiera de las dos anteriores premisas, estaríamos ante parcelas del caso 3 o 4 respectivamente. Finalmente, si una parcela además de estar formada en su mayoría por árboles muy delgados o muy gruesos, su relación entre diámetro medio cuadrático y variables de rodal se alejara de la tendencia descrita por los modelos paramétricos, la parcela correspondería a las incluidas en el caso 2.

En la Figura 12 se muestran la distribución empírica y las estimaciones no paramétrica y paramétrica de una parcela de ejemplo perteneciente a cada uno de los grupos considerados.

## 7. Conclusiones

En este trabajo, se ha realizado la estimación no paramétrica de la distribución del diámetro de *Pinus radiata* D. Don en el noroeste de España, obteniendo como parámetros ventana óptimos  $h_1 = 1.614$  m,  $h_2 = 3.232$  años y  $h_3 = 95.037$  pies/ha, utilizando una adaptación del estimador

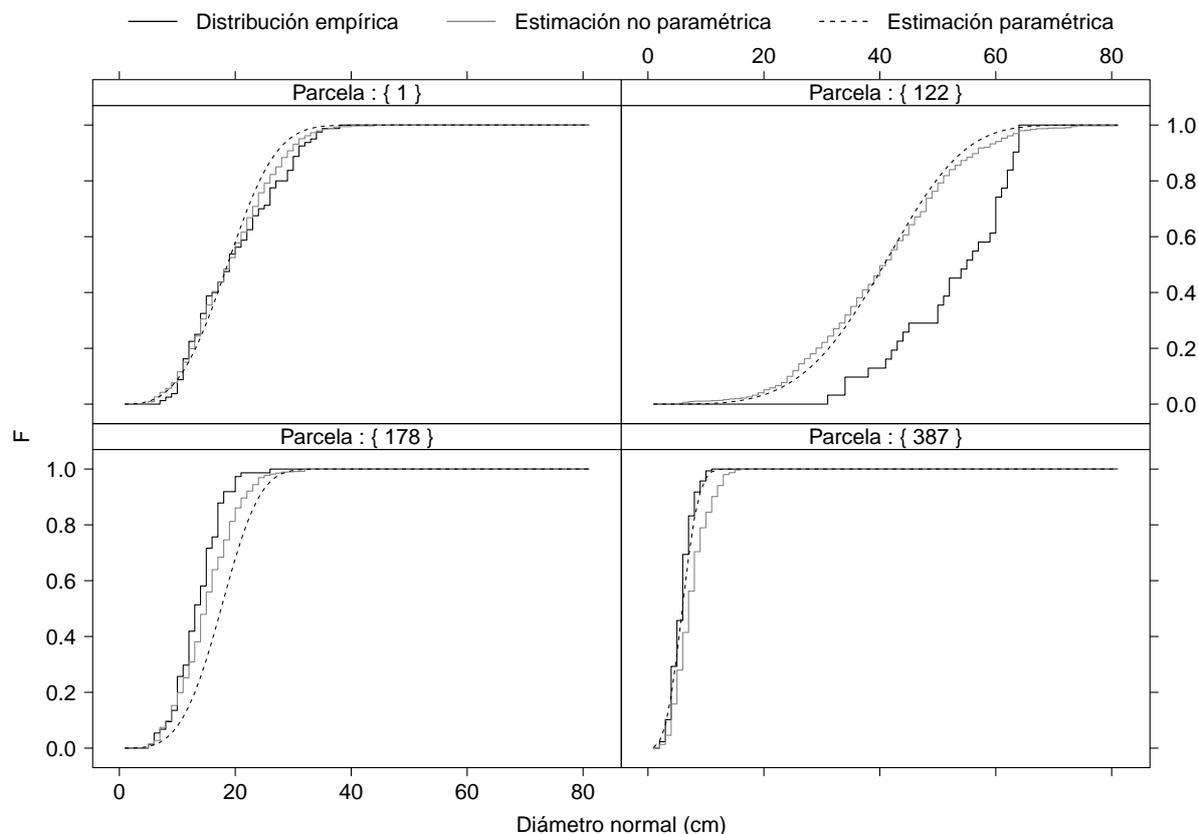


Figura 12: Representación gráfica de la distribución empírica y las estimaciones no paramétrica y paramétrica del diámetro para las parcelas 1 (caso 1), 122 (caso 2), 178 (caso 4) y 387 (caso 3).

propuesto por Li y Racine (2008), basado en el de Nadaraya-Watson.

El método no paramétrico se ha mostrado superior a la metodología paramétrica más utilizada en la modelización de distribuciones diamétricas, que es la basada en ajustes condicionales de tipo Weibull, tras la modelización del diámetro medio y el diámetro medio cuadrático a partir de variables de rodal. Esta afirmación se sustenta en los resultados del test de Wilcoxon aplicado a las medidas de discrepancia (Kolmogorov-Smirnov y Cramér-von Mises) empleadas para comparar las alternativas paramétrica y no paramétrica.

Se observa que, para parcelas con la mayoría de árboles muy delgados o muy gruesos, la estimación no paramétrica proporciona malos resultados, mientras que el método paramétrico es peor en la estimación de la distribución real en los casos en los que la relación entre el diámetro medio cuadrático y las variables de rodal de la parcela se alejan de la respuesta media del modelo paramétrico de diámetro medio cuadrático ajustado.

Como recomendación práctica, se recomienda el empleo del método no paramétrico, excepto para parcelas formadas mayoritariamente por árboles muy delgados o muy gruesos, en cuyo caso se aconseja el empleo de la alternativa paramétrica.

En futuros trabajos se contempla evaluar cómo mejoraría la estimación no paramétrica pro-

puesta al incrementar el número de parcelas de la base de datos y compararla con otras metodologías no paramétricas empleadas en la estimación de la distribución diamétrica de rodales forestales.

## Bibliografía

- Bailey, R. y Dell, T. 1973. Quantifying diameter distributions with the Weibull function. *Forest Science* 19, 97-104.
- Bliss, C. y Reinker, K. 1964. A log-normal approach to diameter distributions in even-aged stands. *Forest Science* 10, 350-360.
- Brent, R. P. 1973. Algorithms for Minimizing without Derivatives. *Prentice-Hall, Englewood Cliffs, NJ*.
- Burk, T. E. y Newberry, J. D. 1984. A Simple Algorithm for Moment-Based Recovery of Weibull Distribution Parameters. *Forest Science* 30, 329-332.
- Cai, Z. 2002. Regression quantiles for time series. *Econometric Theory* 18, 169-192.
- Cao, Q. 2004. Predicting parameters of a Weibull function for modeling diameter distribution. *Forest Science* 50, 682-685.
- Castedo, F. 2004. Modelo dinámico de crecimiento para las masas de *Pinus radiata* D. Don en Galicia. Simulación de alternativas selvícolas con inclusión del riesgo de incendio. Tesis doct. Universidad de Santiago de Compostela.
- Castedo, F., Diéguez-Aranda, U., Barrio, M., Sánchez, M. y Gadow, K. 2006. A generalized height-diameter model including random components for radiata pine plantations in north-western Spain. *Forest Ecology and Management* 229, 202-213.
- CIFOR-INIA 2011. *Cuarto Inventario Forestal Nacional - Galicia*. Ed. por D. G. de Medio Natural y Política Forestal. Ed. por y. M. R. y M. Ministerio de Medio Ambiente.
- 2012. *Cuarto Inventario Forestal Nacional - Principado de Asturias*. Ed. por D. G. de Desarrollo Rural y Política Forestal. Ed. por A. y. M. A. Ministerio de Agricultura.
- Cline, D. 1988. Admissible kernel estimators of a multivariate density. *The Annals of Statistics* 16, 1421-1427.
- Clutter, J. y Bennett, F. 1965. Diameter distributions in old-field slash pine plantations. *Georgia Forest Research Council* 13.
- Diéguez-Aranda, U., Rojo, A., Castedo-Dorado, F., Álvarez, J., Barrio-Anta, M., Crecente-Campo, F., González, J., Pérez-Cruzado, C., Rodríguez, R., López-Sánchez, C., Balboa-Murias, M., Gorgoso, J. y Sánchez, F. 2009. *Herramientas selvícolas para la gestión forestal sostenible en Galicia*. Ed. por X. d. G. Dirección Xeral de Montes Consellería do Medio Rural. Xunta de Galicia. 259 pp.
- Droessler, T. y Burk, T. 1989. A test of nonparametric smoothing of diameter distributions. *Scandinavian Journal of Forest Research* 4, 407-415.
- Epanechnikov, V. A. 1969. Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications* 14, 153-158.
- Frazier, J. R. 1981. Compatible whole-stand and diameter distribution models for loblolly pine plantations. Tesis doct. Virginia Polytechnic Institute and State University.

- Hafley, W. y Schreuder, H. 1977. Statistical distributions for fitting diameter and height data in even-aged stands. *Canadian Journal of Forest Research* 7, 481-487.
- Hall, P., Wolff, R. y Yao, Q. 1999. Methods for estimating a conditional distribution function. *Journal of the American Statistical Association* 94, 154-163.
- Hall, P., Racine, J. y Li, Q. 2004. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association* 99, 1015-1026.
- Hall, P., Li, Q. y Racine, J. S. 2007. Nonparametric estimation of regression functions in the presence of irrelevant regressors. *The Review of Economics and Statistics* 89, 784-789.
- Hansen, B. 2004. Nonparametric estimation of smooth conditional distributions. *Technical Report: Department of Economics, University of Wisconsin*.
- Härdle, W. 1991. *Smoothing techniques: with implementation in S*. Springer Verlag.
- Jones, M. C., Marron, J. S. y Sheather, S. J. 1996. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association* 91, 401-407.
- Lavery, P. B. 1986. Plantation forestry with *Pinus radiata*. *Paper, School of Forestry, University of Canterbury*.
- Levenberg, K. 1944. A method for the solution of certain problems in least squares. *Quarterly of applied mathematics* 2, 164-168.
- Li, Q. y Racine, J. 2008. Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. *Journal of Business and Economic Statistics* 26, 423-434.
- Li, Q., Lin, J. y Racine, J. S. 2013. Optimal bandwidth selection for nonparametric conditional distribution and quantile functions. *Journal of Business & Economic Statistics* 31, 57-65.
- Maltamo, M. y Kangas, A. 1998. Methods based on  $k$ -nearest neighbor regression in the prediction of basal area diameter distribution. *Canadian Journal of Forest Research* 28, 1107-1115.
- Maltamo, M., Puumalainen, J. y Päivinen, R. 1995. Comparison of beta and Weibull functions for modelling basal area diameter distribution in stands of *Pinus sylvestris* and *Picea abies*. *Scandinavian Journal of Forest Research* 10, 284-295.
- Marquardt, D. W. 1963. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics* 11, 431-441.
- Mead, D. 2013. *Sustainable management of Pinus radiata plantations*. Ed. por FAO.
- Myers, R. H. 1990. *Classical and modern regression with applications*. Vol. 2. Duxbury Press Belmont, CA.
- Nadaraya, E. A. 1964. On estimating regression. *Theory of Probability & Its Applications* 9, 141-142.
- Nelder, J. A. y Mead, R. 1965. A Simplex Method for Function Minimization. en. *The Computer Journal* 7, 308-313.
- Nelson, T. 1964. Diameter distribution and growth of loblolly pine. *Forest Science* 10, 105-115.
- Newby, M. 1980. The Properties of Moment Estimators for the Weibull Distribution Based on the Sample Coefficient of Variation. *Technometrics* 22, 187-194.
- Niggemeyer, P. y Schmidt, M. 1999. Estimating diameter distributions using non-parametric methods. *INCO meeting*.
- R Core Team 2012. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.
- Ryan, T. P. 1997. *Modern regression analysis*. John Wiley & Sons.

- Silverman, B. W. 1986. *Density estimation for statistics and data analysis*. Vol. 26. Chapman & Hall.
- Wand, M. y Jones, M. 1995. *Kernel Smoothing*. Chapman & Hall.
- Watson, G. S. 1964. Smooth regression analysis. *Sankhya: The Indian Journal of Statistics, Series A*, 359–372.
- Wilcoxon, F. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin* 1, 80–83.