

# Aplicación de técnicas Clustering basadas en modelos con datos dendrocronológicos de *Pinus* spp.

**Cristina Valeriano Peñas**

Máster en Bioinformática y Bioestadística

Análisis de datos y técnicas *clustering*

**Daniel Fernández Martínez**

**Carles Ventura Royo**

Junio 2020



Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada [3.0 España de Creative Commons](https://creativecommons.org/licenses/by-nc-nd/3.0/es/)

## FICHA DEL TRABAJO FINAL

<b>Título del trabajo:</b>	<i>Model-based clustering approach in dendrochronology with Pinus spp.</i>
<b>Nombre del autor:</b>	<i>Cristina Valeriano Peñas</i>
<b>Nombre del consultor/tutor:</b>	<i>Daniel Fernández Martínez</i>
<b>Nombre del tutor externo:</b>	
<b>Nombre del PRA:</b>	<i>Carles Ventura Royo</i>
<b>Fecha de entrega (mm/aaaa):</b>	06/2020
<b>Titulación:</b>	Máster en Bioinformática y Bioestadística
<b>Área del Trabajo Final:</b>	Análisis de datos y técnicas <i>clustering</i>
<b>Idioma del trabajo:</b>	<i>Español</i>
<b>Key words:</b>	<i>Tree-ring with, clustering, finite mixture models, dendrochronology, EM algorithm</i>

**Resumen del Trabajo (máximo 250 palabras):** *Con la finalidad, contexto de aplicación, metodología, resultados y conclusiones del trabajo.*

El campo de la dendrocronología estudia series temporales a lo largo de gradientes ecológicos, los árboles que forman las cronologías pueden no responder de la misma manera al medio que les rodea, particularmente responden desigual al clima. Por ello la vulnerabilidad de los árboles al cambio climático puede depender de la respuesta específica de la especie o de la adaptación al sitio en el que se encuentre. Los análisis con técnicas multivariantes como el análisis clustering basado en modelos pueden ser instrumentos de gran utilidad para agrupar las diferencias existentes en la sensibilidad de los árboles y los bosques con el medio. En este trabajo analizamos el potencial de la técnica *clustering* basado en modelos, sus métodos y el algoritmo EM; para detectar y agrupar patrones de crecimiento comunes usando el ancho de los anillos de los árboles, de varias especies del género *Pinus* y en diferentes países de la Unión Europea. Los análisis se han desarrollado con tres paquetes del programa R (*mclust*, *funFEM/funHDDC* y *lcmm*) que disponen de funciones para realizar análisis de modelos de mezcla finita y que estiman los parámetros mediante la máxima verosimilitud con el algoritmo EM. Los análisis muestran agrupaciones de tres/cuatro clases de trayectorias, y en cada clase se observan diferentes características ecológicas.

**Abstract (in English, 250 words or less):**

Dendrochronology is a field of science be able to show spatial and temporal ecologic and climatic information based on tree-ring chronologies. These studies are usually integrated by chronologies along ecological gradients, though trees may not respond in the same way to the environment, particularly to climate. Therefore, this vulnerability of trees to climate change may depend on the species-specific distribution, or geographical adaptation prevails over species. Analysis with a multivariate technique such as model-based clustering can be a useful instrument when aiming at capturing and grouping these differences in trees and forests sensitivity to the environment. The aim of the study was analyze the potential of model-based clustering technique, their methods and the EM algorithm; to detect and cluster common growth patterns using the tree-ring width of several species for the genus *Pinus* and in different countries of the European Union. The analyses have been developed with three R packages (mclust, funFEM/funHDDC and lcmm) with the functions to perform finite mixture models and estimate the parameters using the maximum likelihood with the EM algorithm. The results show groupings of three/four trajectories classes, and in each class different ecological characteristics are observed.

# INDICE

<b>FICHA DEL TRABAJO FINAL.....</b>	<b>I</b>
<b>INDICE.....</b>	<b>III</b>
<b>1. INTRODUCCIÓN.....</b>	<b>1</b>
1.1 CONTEXTO Y JUSTIFICACIÓN.....	1
1.2 OBJETIVOS.....	2
<i>Objetivos generales:</i> .....	2
<i>Objetivos específicos:</i> .....	2
1.3 ENFOQUE Y MÉTODO SEGUIDO.....	3
1.4 PLANIFICACIÓN DEL TRABAJO.....	3
<i>Calendario:</i> .....	3
<i>Tareas:</i> .....	4
<i>Hitos:</i> .....	4
<i>Análisis de riesgos:</i> .....	4
1.5 BREVE SUMARIO DE PRODUCTOS OBTENIDOS.....	5
1.6 BREVE DESCRIPCIÓN DE LOS OTROS CAPÍTULOS DE LA MEMORIA.....	5
<b>2. CLUSTERING BASADO EN MODELOS.....</b>	<b>7</b>
2.1. TÉCNICA DE CLUSTERING O AGRUPAMIENTO.....	7
2.2. ALGORITMO EXPECTACIÓN-MAXIMIZACIÓN (EM).....	8
2.3. SELECCIÓN DEL MEJOR MODELO: CRITERIOS DE INFORMACIÓN Y OTROS MÉTODOS.....	9
2.4. FUNCIONES QUE IMPLEMENTAN EL ALGORITMO EM EN PAQUETES DE R.....	10
<b>3. DENDROCROLOGÍA.....</b>	<b>13</b>
3.1. DEFINICIONES Y METODOLOGÍA.....	13
3.2. CREACIÓN DE LA BASE DE DATOS.....	13
<b>4. RESULTADOS.....</b>	<b>17</b>
4.1. ANÁLISIS DESCRIPTIVO: CRONOLOGÍAS MAESTRAS.....	17
4.2. ANÁLISIS CON EL PAQUETE <i>MCLUST</i> .....	17
4.3. ANÁLISIS CON LOS PAQUETES <i>FUNFEM</i> Y <i>FUNHDDC</i> .....	18
4.4. ANÁLISIS CON EL PAQUETE <i>LCMM</i> .....	20
<b>5. CONCLUSIONES.....</b>	<b>22</b>
¿QUÉ LECCIONES SE HAN APRENDIDO DEL TRABAJO?.....	22
REFLEXIÓN CRÍTICA SOBRE EL LOGRO DE LOS OBJETIVOS.....	23
ANÁLISIS CRÍTICO DEL SEGUIMIENTO DE LA PLANIFICACIÓN Y METODOLOGÍA.....	23
LÍNEAS DE TRABAJO FUTURO QUE NO SE HAN PODIDO EXPLORAR.....	23
<b>6. GLOSARIO.....</b>	<b>24</b>
<b>7. BIBLIOGRAFÍA.....</b>	<b>25</b>
<b>ANEXOS.....</b>	<b>28</b>
PROGRAMA DE DESCARGA: CÓDIGO JAVA.....	28
CÓDIGO R DE LA CREACIÓN Y VALIDACIÓN DE LA BASE DE DATOS:.....	30
CÓDIGO R DEL ANÁLISIS CLUSTERING CON LOS DIFERENTES PAQUETES.....	52
<i>library(fda)</i> .....	52
<i>library(funHDDC)</i> .....	52
<i>library(funFEM)</i> .....	52
<i>library(lcmm)</i> .....	53



## Lista de Figuras

<i>Figura 1.- Cronograma del trabajo final de máster (TFM).....</i>	<i>3</i>
<i>Figura 2.- Ejemplo de dos distribuciones normales para explicar los pasos del algoritmo EM.....</i>	<i>9</i>
<i>Figura 3.- Ejemplo visual del resultado de un análisis de COFECHA.....</i>	<i>15</i>
<i>Figura 4.- Cronologías medias del grosor de los anillos de los árboles.....</i>	<i>17</i>
<i>Figura 5.- El criterio de la pendiente heurística (slope heuristic).....</i>	<i>19</i>
<i>Figura 6.- Valores medios de las cronologías agrupadas en tres clústers.....</i>	<i>19</i>
<i>Figura 7.- Representación de los valores predichos y observados.....</i>	<i>21</i>

# 1. Introducción

## 1.1 Contexto y justificación

Los estudios dendroecológicos están formados por cronologías a lo largo de gradientes ecológicos, los árboles que forman dichas cronologías pueden no responder de la misma manera al ambiente que les rodea, y particularmente no responden igual al clima. Por ejemplo, varios estudios en coníferas han mostrado impactos negativos en la performance de los árboles y en la formación del bosque debido al calentamiento global y las sequías extremas (Sánchez-Salguero et al. 2010; Heres et al. 2012; Dorado-Liñán et al. 2017), mientras que otros estudios debido al mismo efecto presentan una respuesta favorable en el crecimiento radial de los árboles (Shestakova et al. 2019). Por lo tanto, la vulnerabilidad de los árboles al cambio climático no es solo una función de la respuesta específica de la especie, sino que la adaptación al sitio también puede desempeñar un papel determinante (Dorado-Liñán et al. 2019). Los análisis con técnicas multivariantes son instrumentos de gran utilidad para agrupar las diferencias existentes en la sensibilidad de los árboles y los bosques con el medio. Una de estas técnicas es el análisis clustering, que agrupa las observaciones de acuerdo con las similitudes para una variable dada (Legendre & Legendre 1998). Esta técnica puede ser una herramienta útil en análisis de series temporales (Liao 2005) y se ha aplicado en algunos estudios dendroecológicos (García-González et al. 2008; Michael & Melnykov 2016).

Las técnicas clustering se pueden dividir en dos grandes grupos: jerárquico (*hierarchical*) y particional (*partitional*). La agrupación jerárquica se basa en distancias matemáticas y no tiene ningún modelo probabilístico asociado. Por otro lado, los métodos de agrupamiento particional utilizan modelos probabilísticos tales como: los basados en distancias, en densidad, en modelos o en cuadrículas (Saxena et al. 2017). Los estudios previos sobre dendroecología que aplicaron este análisis clúster utilizaron agrupamiento particional basado en la distancia *k-means* y/o el agrupamiento jerárquico, pero no mostraron resultados significativos y los grupos resultantes parecían arbitrarios (Haneca et al. 2005). Sin embargo, el potencial de usar agrupaciones probabilísticas basadas en modelos no se ha probado hasta ahora en estudios dendroecológicos (Michael & Melnykov 2016). Por ello, en este estudio usaremos un análisis de agrupación (*clustering*) basado en modelos (*model-based*), utilizando estimadores de máxima verosimilitud, como el algoritmo esperanza-maximización (*expectation-maximization*, EM; Banfield & Raftery 1993; Melnykov & Maitra 2010). Específicamente basándonos en modelos de mixtura finita (*finite mixture model*, FMM; McLachlan et al. 2019) y modelos mixtos usando clases de latencia (*latent classes mixed models*, LCMM; Proust-Lima et al. 2015).



En este trabajo se quiere estudiar el potencial de usar estos métodos, ya muy utilizados en el ámbito de la bioinformática (sobre todo en genética), para detectar y agrupar patrones de crecimiento comunes usando el ancho de los anillos de árboles (*Tree-Ring width*, TRW), en cinco especies de árboles que pertenecen al género *Pinus* (*P. halepensis*, *P. pinaster*, *P. pinea*, *P. nigra* y *P. sylvestris*).

### Justificación

Actualmente, el estudio del impacto del cambio climático en los sistemas naturales es un tema muy relevante en la ciencia. El uso de nuevas metodologías como el clustering basado en modelos en estudios dendroecológicos puede ser una ventaja para obtener una clasificación más precisa de las series de anillos de los árboles en diferentes grupos o clases.

## 1.2 Objetivos

Objetivos generales:

1. Ampliación de la metodología estadística utilizada para estudiar la distribución ecológica de una especie arbórea, *Pinus* spp., con técnicas de *clustering* basado en modelos.
2. Aplicación de diferentes técnicas de análisis *clustering* para obtener grupos o clústeres con un sentido ecológico.
3. Evaluación de los métodos *clustering* como herramienta estadística potencialmente útil en dendrocronología.

Objetivos específicos:

#### De formación

- Revisión bibliográfica de la técnica clustering y todos sus métodos.
- Diseño de la base de datos de anillos de árboles (TRW) del género *Pinus* spp.
- Estudio del software y desarrollo del código en R para el análisis clustering basado en modelos, particularmente utilizando los paquetes:

*mclust* (<https://cran.r-project.org/web/packages/mclust/mclust.pdf>)

*funHDDC* (<https://cran.r-project.org/web/packages/funHDDC/index.html>)

*funFEM* (<https://cran.r-project.org/web/packages/funFEM/index.html>)

*lcmm* (<https://cran.r-project.org/web/packages/lcmm/index.html>)

#### De desarrollo

- Procesamiento de datos: Diseño, creación y exploración de la base de datos. Análisis descriptivo y validación de los datos que forman la base de datos.
- Análisis clustering con los diferentes paquetes: Elección del mejor modelo con el criterio de información bayesiano (BIC) y el criterio de

información de Akaike (AIC). Representación gráfica y visualización de los resultados.

- Evaluación de los resultados obtenidos con los diferentes paquetes estudiados.

### 1.3 Enfoque y método seguido

El trabajo consta de dos partes, la primera es un estudio bibliográfico sobre el análisis clustering y sus diferentes métodos, específicamente desarrollando la técnica clustering basado en modelos probabilísticos. Mediante la revisión exhaustiva de la bibliografía disponible se identificaron los diferentes métodos de agrupación basados en modelos que puedan ser utilizados con los datos dendrocronológicos, así como los diferentes paquetes que pueden utilizarse para el análisis en R (R Core Team 2020).

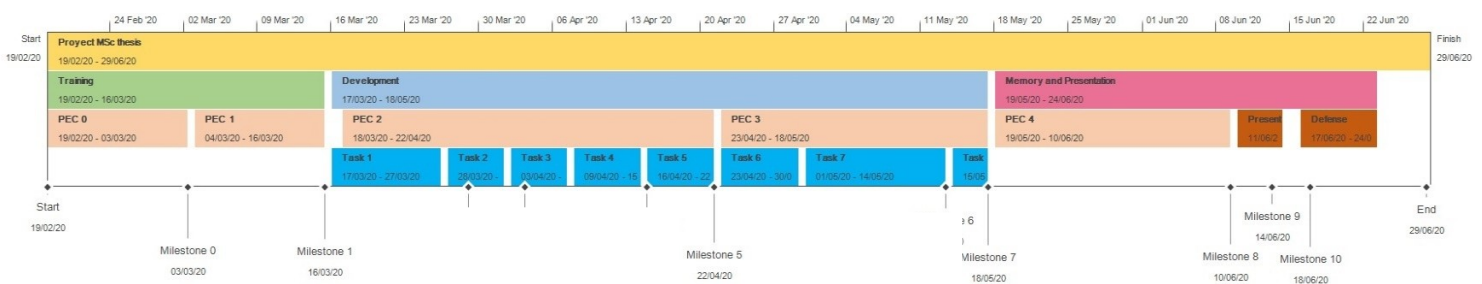
La segunda parte del trabajo se desarrolla en el entorno de RStudio (RStudio Team 2020) y desarrollará los análisis clustering con los paquetes seleccionados en la primera parte. Para el análisis se diseñó una base de datos de anillos de los árboles (TRW), extraídos de la web pública *International Tree-Ring Data Bank\** (ITRDB). Se seleccionaron diferentes cronologías que cubrieran toda Europa del género Pino.

\*Enlace a la base de datos ITRDB: <https://www.ncdc.noaa.gov/paleo-search/?dataTypeId=18>

### 1.4 Planificación del Trabajo

Se ha elaborado un calendario con Microsoft\_Project16 (**Fig. 1**) que recoge la temporización de las tareas, hitos, fases y PECs. A continuación del calendario se muestra detalladamente el listado de tareas e hitos que se han llevado a cabo para poder finalizar con éxito el TFM, junto con un análisis de riesgos.

#### Calendario



**Figura 1.-** Cronograma del trabajo final de máster (TFM). Se divide en tareas (*Task*), hitos (*Milestone*) y PECs.

#### Tareas

**Tarea 1-** Definición del análisis clustering en términos generales. Descripción de los diferentes métodos de clustering basado en modelos

**Tarea 2-** Diseño de la base de datos de anillos de árboles (datos dendrocronológicos) y preprocesamiento de los datos

**Tarea 3-** Investigación y desarrollo del código R para cada paquete

**Tarea 4-** Análisis clustering: estimación de parámetros

**Tarea 5-** Modelado e implementación del algoritmo EM con *mclust*

**Tarea 6-** Análisis con otros paquetes: *funHDDC*, *funFEM* y *lcmm*

**Tarea 7-** Evaluación del modelo. Representación gráfica y visualización

**Tarea 8-** Comparación con otra técnica de agrupamiento indirecto, CPCA

### Hitos

Inicio del TFM \_\_\_\_\_ 19/02/2020

#### **Fase de formación:**

Hito 0: Entrega de la propuesta de TFM, PEC0 02/03/2020

Hito 1: Entrega del Plan de trabajo, PEC1 16/03/2020

**Finalización de la fase de formación 31/03/2020**

**Fase de desarrollo:** 01/04/2020

Ejecución de las tareas 1, 2 y 3

Hito 2: Entrega del Informe de seguimiento fase 1, PEC2 22/04/2020

Ejecución de las tareas 4, 5, 6 y 7

**Finalización de la fase de desarrollo 31/05/2020**

Hito 3: Entrega del informe de seguimiento fase 2, PEC3 01/06/2020\*

Hito 4: Entrega de la Memoria, PEC4 24/06/2020\*

Hito 5: Preparación de la Presentación PowerPoint (del 25/06/2020 al 28/06/2020)\*

Hito 6: Presentación y Defensa del TFM (del 01/07/2020 al 08/07/2020)\*

*\*Las fechas han sufrido cambios.*

### Análisis de riesgos

Existen tres grandes secciones que pueden presentar riesgos en el desarrollo del trabajo:

- Creación de la base de datos: problemas derivados del acceso y la generación de la base de datos seleccionada, complicaciones derivadas de la transformación de la base de datos a un formato legible en R.

Necesario el desarrollo de un programa en lenguaje Java para poder descargar y unir todos los datos descargados de la web.

- Análisis: al ser un estudio exploratorio de diferentes paquetes de R aplicados a datos dendrocronológicos algunos de los paquetes no

tienen porque arrojar resultados significativos o reales desde el punto de vista ecológico. Habrá que realizar diferentes análisis para decidir cuál es el paquete óptimo con estos datos.

- Tiempo: problemas de compatibilidad con horas y carga de trabajo/tesis, viajes y enfermedades personales imprevistas o de familiares cercanos (pandemia global).

Retraso en el análisis debido a la complejidad de los métodos y al lento procesamiento de datos.

## 1.5 Breve resumen de productos obtenidos

### Plan de Trabajo

Descripción y cronograma de los objetivos, tareas y metodología que se llevarán a cabo durante el trabajo final de máster (TFM).

### Memoria final del TFM

Presentación detallada por escrito de los resultados, discusión y conclusiones obtenidas en la Tesis de Maestría (TFM). El documento final contendrá las siguientes secciones: Introducción, metodología, resultados, discusión, conclusiones y anexos.

### Presentación virtual

Diapositivas y video de 20 minutos que contendrá una presentación oral que resume los resultados y conclusiones más interesantes del trabajo.

### Productos obtenidos

Futuro artículo científico con el desarrollo de esta metodología con datos dendrocronológicos propios.

## 1.6 Breve descripción de los otros capítulos de la memoria

El apartado 2. **Clustering basado en modelos**: corresponde a un desarrollo teórico de todo lo estudiado en la revisión bibliográfica sobre la técnica de Clustering y más concretamente sobre los métodos basados en modelos. También se describirá el algoritmo EM y los criterios de información, y, por último, las funciones de los paquetes utilizados en el análisis.

El apartado 3. **Dendrocronología**: incluye definiciones básicas y describe la metodología utilizada para validar los datos, determinando las cronologías seleccionadas para el análisis.

El apartado 4. **Resultados y Discusión**: está formado por el análisis descriptivo de los datos dendrocronológicos y los resultados de los diferentes análisis clustering realizados con cada uno de los paquetes junto con la discusión de estos.

El apartado 5 incluirá las **Conclusiones** de todo el trabajo.

El apartado 6 y 7 recogen el **Glosario** y la **Bibliografía** de toda la memoria.

Por último, se encuentra un **ANEXO** con el código utilizado en este TFM, desarrollado en lenguaje JAVA y R. Y algunas figuras suplementarias.

## 2. Clustering basado en modelos

En este capítulo se definen los conceptos básicos de la técnica Clustering, profundizando en el método basado en modelos (subapartado 2.1.). También se describe el algoritmo EM y los criterios de información (subapartado 2.2. y 2.3.). En el subapartado 2.4. se analizan las funciones de los paquetes *mclust*, *funHDDC*, *funFEM* y *lcm* del software libre R (R Core Team 2020).

### 2.1. Técnica de Clustering o Agrupamiento

La técnica multivariante de **Clustering**, se basa en la agrupación de datos en clases o clústers, por lo general, a partir de una función de distancia o similitud (Legendre & Legendre 1998). El clustering es una técnica descriptiva de aprendizaje no supervisada (*unsupervised learning*), esto significa que no necesita datos de ejemplo para entrenar al algoritmo, sino que se extraen los grupos (o clases) a partir de los datos sin un conocimiento previo. Según Saxena et al. (2017) una aproximación taxonómica general divide los métodos de clustering en dos grandes categorías: jerárquicos y no jerárquicos o particionales. La **agrupación jerárquica** se subdivide en métodos aglomerativos y divisivos, y tiene como objetivo agrupar o dividir las clases basándose en las distancias matemáticas o medidas de similitud. La **agrupación particional** tiene muchas categorías dependiendo de en qué se basen para realizar los clústers, algunos de los más comunes son los basados: en distancias, en rejillas, en densidades o en modelos. En este TFM nos vamos a centrar en estos últimos: “*model-based*”.

El enfoque más popular del clustering *model-based* (Fraley & Raftery 1998) es el **modelo de mezcla Gaussiana** (GMM, Banfield & Raftery 1993; Melnykov & Maitra 2010). Asume que el conjunto de datos utilizado en el análisis es una combinación (o “*mixtura*”) de dos o más componentes (clústers= $k$ ), y que cada componente o clúster está descrito por una distribución de probabilidad que determina su aportación al conjunto.

Aunque a cada componente se le puede asociar cualquier modelo de probabilidad, los GMM se basan en la distribución normal multivariante (Gaussiana) donde las variables  $n$  ( $X_1, X_2, \dots, X_n$ ) tienen una media  $\mu = (\mu_1, \mu_2, \dots, \mu_n)$  y una matriz de covarianzas ( $\Sigma_k$ ) que describe la variabilidad para cada variable. Los GMM determinan que los clústeres se pueden crear utilizando  $k$  distribuciones Gaussianas. La matriz de covarianza describe la geometría de los componentes con: el volumen, la forma y la orientación de los clústers. Como veremos en las descripciones de las funciones de los paquetes, la principal familia de modelos sigue una geometría en la parametrización de la covarianza elipsoidal, aunque también puede ser esférica o diagonal (Celeux & Govaert 1995). Los parámetros del modelo se estiman mediante la máxima verosimilitud de las distribuciones utilizando como herramienta una amplia familia de algoritmos iterativos refinados.

El más utilizado es el algoritmo de **Expectación-Maximización** (EM, Dempster et al. 1977) que describiremos en el siguiente apartado.

Para este estudio utilizaremos el modelo de mixtura finito (FMM; McNicholas 2016; McLachlan et al. 2019) y los modelos mixtos lineales de clase latente, que son similares a los modelos de mixtura de crecimiento (gmm; Ram & Grimm 2009). Estos modelos permiten diferenciar subpoblaciones homogéneas, pero con características heterogéneas entre ellas, así mismo posibilitan extraer características a nivel individuo, incluyéndolo como perteneciente a alguna de las subpoblaciones. Ambos tienen en cuenta que el estudio se repite a lo largo del tiempo (datos longitudinales).

## 2.2. Algoritmo Expectación-Maximización (EM)

El algoritmo EM fue descrito por Dempster et al. (1977) y se utiliza para estimar los parámetros del modelo, mejora al algoritmo de k-medias ya que en vez de usar distancias o similitudes aplica **distribuciones Gaussianas** (Banfield & Raftery 1993). También tiene la ventaja de que nos permite variar el número de clases (no son fijas) y que realiza agrupaciones no disjuntas, esto significa que cada muestra puede pertenecer en cierto grado a varios componentes o clases, todo lo contrario de otros métodos donde las clasificaciones son excluyentes. Este algoritmo exige variables numéricas que requieren de normalización y admite ruido suavizado.

Según McLachlan y Krishnan (1997) la formulación del algoritmo tiene dos etapas por cada iteración, los **pasos de expectación y maximización, EM**. El paso E: estima la probabilidad esperada o verosimilitud de las observaciones respecto a los parámetros del modelo, y el paso M (restimación de los parámetros del modelo): determina los parámetros óptimos del modelo para cada mezcla/mixtura utilizando las probabilidades de asignación como pesos.

Para describir los pasos más en profundidad nos vamos a ayudar de la **Fig. 2** como caso sencillo. El paso E determina el nivel de pertenencia de cada muestra (X) a los diferentes componentes (rojo y azul), para esto calcula la probabilidad que tiene la muestra (X) en cada uno de los modelos de distribución asociados a los componentes. Las probabilidades resultantes (Y1, Y2...) se utilizan como pesos que determinan el grado de pertenencia de la muestra X a cada componente o clase (K: clúster). En el paso M de maximización, se desplazan las distribuciones normales hasta que la media de la distribución ( $\mu_1$  y  $\mu_2$ ) queda centrada en el componente o clase (K).

Los pasos E y M se repiten alternativamente hasta que se cumple un criterio específico, definido previamente. Es muy importante el valor inicial de los parámetros, cuando no estamos seguros es mejor comenzar con 1 entre el número de variables totales.

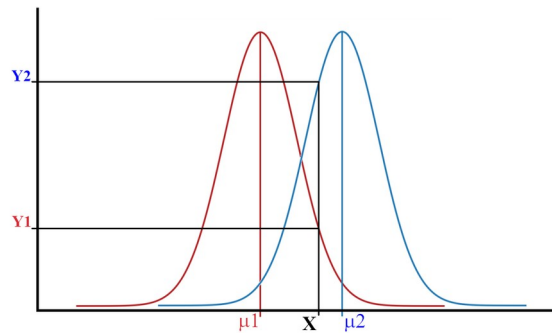


Figura 2.- Ejemplo de dos distribuciones normales para explicar los pasos del algoritmo EM.

### 2.3. Selección del mejor modelo: Criterios de información y otros métodos

La elección del número de componentes que constituyen un FMM o la selección del mejor modelo ha sido un factor de gran relevancia en los estudios multivariantes. Depende de dos factores: la distribución de los datos modelizados, y el número y forma de los componentes. Para esto existen diferentes técnicas y criterios que se adecuan a la función utilizada en el análisis.

Generalmente se emplean los criterios de información. En este estudio utilizaremos dos criterios muy populares: el **criterio de información de Akaike** (AIC, Akaike, 1974) y el **Bayesiano** (BIC, Schwarz 1978).

Según Akaike (1974) la expresión general del criterio la observamos en la fórmula 2.3.1.

$$AIC(\theta) = (-2) \ln \cdot L + 2k, \quad (2.3.1)$$

donde L es la función de máxima verosimilitud, k es el número de parámetros ajustados independientemente estimados dentro del modelo ( $\theta$ ). Se selecciona el modelo que presenta un menor valor de AIC. Hay que tener en cuenta que el AIC penaliza el sobreajuste que se utiliza en modelos grandes a través del principio de parsimonia.

Schwarz (1978) sugirió que el AIC podría no ser asintóticamente justificable y desarrolló un criterio alternativo con enfoque bayesiano (fórmula 2.3.2)

$$BIC(\theta) = -2 \ln \cdot L + k \ln(n), \quad (2.3.2)$$

L y k son igual que en AIC, mientras que n es el tamaño de la muestra. A diferencia del AIC puede utilizarse para modelos de muchas observaciones.

Existen otros métodos de selección que nos ofrecen algunos paquetes estadísticos y son menos conocidos. Como **la pendiente heurística** (*the slope heuristics*), o el **Test de Wald**. Birgé & Massart (2006) han propuesto la adopción de la pendiente heurística (método *data-driven*) como una forma de elegir una constante multiplicativa desconocida basándose en la noción de “penalización mínima” (Arlot 2019). Por otra parte, el Test de Wald nos muestra la coherencia de los parámetros en un modelo probabilístico seleccionado y ajustado.



El Test de Wald lo utilizaremos en los análisis con modelos mixtos. Es una prueba generalista que se aplica tras elegir el modelo, comparando si la estimación de los parámetros ( $\theta$ ) es adecuada a los parámetros propuestos ( $\theta_0$ ). Realiza un contraste de hipótesis para los parámetros del modelo (2.3.3).

$$T = (\theta - \theta_0)^2 / \text{var}(\theta) \quad (2.3.3)$$

#### 2.4. Funciones que implementan el algoritmo EM en paquetes de R

**Paquete *mclust*** (Scrucca et al. 2016): es uno de los paquetes más populares de clustering basado en modelos. Se utiliza en multitud de contextos, geoquímica, hidrología, energía eólica, genética, epidemiología, ciencia alimentaria, psicología clínica, antropología, ciencias políticas... entre otras.

La función *mclust* se basa en los GMM y es específico para estudios de clustering basados en modelos, aunque también realiza estimación de la densidad y análisis discriminante. Dispone de variedad de estructuras de covarianza y múltiples componentes ( $k$ ). Las funciones que proporciona este paquete estiman la máxima verosimilitud de los parámetros a través del algoritmo EM, incluyendo funciones para realizar cada paso por separado. Y validan el mejor modelo con el criterio BIC. Para iniciar la parametrización de EM se utilizan las particiones obtenidas en un análisis clustering jerárquico subtipo aglomerativo. La parametrización parsimoniosa de la matriz se obtiene mediante la descomposición de la componente (fórmula 2.4.4).

$$\Sigma_k = \lambda_k \cdot D_k \cdot A_k \cdot D_k^T, \quad (2.4.4)$$

donde la matriz de covarianza ( $\Sigma_k$ ) se descompone en:  $\lambda_k$  que es el escalar que controla el volumen del elipsoide (ya que los grupos generados son elipsoidales),  $A_k$  es la matriz diagonal que especifica los contornos de la densidad con el  $\det(A_k) = 1$ , y  $D_k$  es la matriz ortogonal que determina la orientación del elipsoide correspondiente (Banfield & Raftery 1993). Para una dimensión tendremos dos modelos: E si hay varianzas iguales y V para varianzas variables. A escala multivariable tendremos un total de 14 posibles modelos con diferentes características geométricas (EII, VII, EEI, VEI, EVI, VVI, EEE, EVE, VEE, VVE, EEV, VEV, EVV y VVV) descrito en Scrucca et al. (2016).

Uno de los grandes hándicaps de este paquete (específicamente para este trabajo) es que no tiene en cuenta el factor tiempo. Los resultados que ofrece nos devuelven modelos con diferentes clases, pero sin tener en cuanto el paso del tiempo.

**Paquete *funHDDC*** (Bouveyron et al. 2017) y ***funFEM*** (Bouveyron 2014)

Estos paquetes realizan análisis de datos funcionales basados en modelos por lo que para trabajar con ambos es necesario describir la función de los datos que queramos estudiar.

Para ello se utiliza el **paquete *fda*** (Ramsay & Silverman 1997) con las funciones: *create.basis* y *smooth.basis*. A partir de esta “transformación” de los datos se realiza el análisis clúster con los dos algoritmos (*funFEM* y *funHDDC*).

La función *funFEM* más generalista que *HDDC* nos permite agrupar series temporales, basándose en un GMM funcional discriminativo. La ventaja que presenta es que al ser parsimonioso puede manejar series largas. La función *funHDDC* (más desarrollada y compleja que la anterior) se basa en un GMM funcional de mezclas latentes (GLM) y se ajusta a datos funcionales grupo-específicos. Restringe los parámetros del modelo dentro y entre los grupos. Como la función *funFEM*, forma modelos parsimoniosos capaces de adaptarse a muchos tipos de estudio. Al igual que el paquete *mclust*, el FLM de esta función se descompone en los parámetros (formula 2.4.5) requeridos para estimar las matrices  $Q_k$  (Bouveyron et al. 2011).

$$\text{FLM } [a_{kj} \cdot b_k \cdot Q_k \cdot d_k] = \rho + \tau + 2K + D, \quad (2.4.5)$$

donde  $a$  y  $b$  son los valores propios (*eigenvalues*) que regularizan la estimación de  $k$ . Y las propiedades del modelo son:  $\rho = Kp + K - 1$  es el número de parámetros requeridos para la estimación de medias y proporciones,  $\tau$  es el número de parámetros requeridos para la estimación de la matriz de orientación  $Q_k$  y  $D = \sum_{k=1} d_k$ . Al restringir los parámetros del modelo FLM dentro o entre grupos obtenemos 6 submodelos parsimoniosos correctos (fórmula 2.4.6).

$$\begin{aligned} a_{kj} \cdot b_k \cdot Q_k \cdot d_k &= \rho + \tau + 2K + D \\ a_{kj} \cdot b \cdot Q_k \cdot d_k &= \rho + \tau + K + D + 1 \\ a_k \cdot b_k \cdot Q_k \cdot d_k &= \rho + \tau + 3K \\ ab_k \cdot Q_k \cdot d_k &= \rho + \tau + 2K + 1 \\ a_k \cdot b \cdot Q_k \cdot d_k &= \rho + \tau + 2K + 1 \\ abQ_k \cdot d_k &= \rho + \tau + K + 2 \end{aligned} \quad (2.4.6)$$

Ambas funciones utilizan un procedimiento de estimación basado la teoría clásica del algoritmo EM. Y se validan los resultados con los criterios: BIC y la *slope heuristic*.

### Paquete *lcmm* (Proust-Lima et al. 2015)

Este paquete proporciona varias funciones para estimar modelos estadísticos basados en modelos lineales mixtos (LMM). En nuestro caso utilizaremos la función *hlme* para modelos mixtos lineales de clase latente (LCMM) con resultados longitudinales gaussianos. La formulación de los modelos mixtos está compuesta por variables de efectos fijos y de efectos aleatorios, los efectos fijos son idénticos para todas las observaciones del estudio mientras que los aleatorios son diferentes para cada observación. Por lo tanto, en la fórmula general para los LMM (2.4.7) la variable respuesta  $Y_{ij}$  ( $i$  es cada variable y  $j$  es la medición que se realiza en el tiempo,  $t$ ) es igual a la suma de los efectos fijos ( $\beta$ ) por su vector de covariables ( $X_{t_{ij}}$ ) más los efectos aleatorios ( $\alpha$ ) por su vector de covariables ( $Z_{t_{ij}}$ ) y más el error ( $\epsilon$ ) (McCulloch 2003).

$$Y_{ij} = X_{t_{ij}} \cdot \beta + Z_{t_{ij}} \cdot \alpha + \epsilon_{ij} \quad (2.4.7)$$

El submodelo LCMM, que utiliza este paquete, amplía la fórmula general (2.4.7) con las características de las trayectorias de la variable ( $Y_{ij}$ ) para cada clase latente ( $l$ ) añadiendo un proceso estocástico Gaussiano evaluado en el tiempo ( $B_{t_{ij}}$ ). Asume que cada clase latente se caracteriza por un modelo de medición, además de un modelo estructural, permitiendo explorar la existencia de una posible estructura latente en los datos (Nagin et al. 2018). Por lo tanto, este análisis proporciona las diferentes trayectorias que siguen las variables en relación con el tiempo, suponiendo que el crecimiento individual dentro de cada clase es homogéneo y donde las estimaciones de varianza y covarianza para los factores de crecimiento dentro de cada clase se fija en cero (Jung & Wickrama 2008).

A diferencia de los otros paquetes, el estimador de máxima verosimilitud se obtienen utilizando un algoritmo denominado: *Marquardt* -modificado (Marquardt 1963), que pertenece a la familia del algoritmo EM, y se utiliza para resolver problemas de mínimos cuadrados no lineales (Levenberg 1944). Mediante este algoritmo la función del paquete va actualizando de forma iterativa el vector de los parámetros hasta que converge según los criterios seleccionados.

El algoritmo de estimación se puede iniciar de cuatro formas diferentes: automáticamente definiendo los clústers, especificando los valores iniciales, por defecto o de forma aleatoria (*random*), este último extrae los valores iniciales de la distribución asintótica del MLE. Se validan los resultados con los criterios: BIC y AIC. Para validar los parámetros seleccionados utiliza el contraste de hipótesis del Test de Wald.

## 3. Dendrocronología

En este capítulo se definen los conceptos y la metodología básica en el campo de la dendrocronología (subapartado 3.1.). También se desarrollan los pasos utilizados para la creación de la base de datos y su posterior validación (subapartado 3.2.)

### 3.1. Definiciones y Metodología

La dendrocronología es la ciencia que estudia los anillos de los árboles. Etimológicamente tiene su origen en los términos griegos *dendron* - árbol-, *crono* -tiempo- y *logía* -ciencia-. Esta disciplina científica está dotada de un conjunto de principios, técnicas y métodos que permiten datar los anillos de crecimiento que se encuentran en el tronco de los árboles. El objetivo principal es establecer las denominadas cronologías o series maestras de crecimiento (Gutiérrez 2009). Por otro lado, la dendroecología es la subdisciplina que aplica los principios de la ciencia dendrocronológica para resolver problemas ecológicos (Amoroso et al. 2017).

En la actualidad existe un archivo público de datos de grosor de los anillos de los árboles denominado ITRDB y lo administra el equipo de paleoclimatología del NCEI's (*National Centers for Environmental Information*). Es el archivo público más grande del mundo y dispone de datos de más de 4000 sitios en 6 continentes, incluye medidas de ancho del anillo (TRW), densidad de madera, medidas de isótopos etc. (Grissino-Mayeri & Fritts 1997). Debido a la imposibilidad de realizar el estudio con datos propios se decidió utilizar cronologías de uso público de esta base de datos.

### 3.2. Creación de la base de datos

Inicialmente se descargaron 27 cronologías del género *Pinus* spp. de diferentes países de toda la Unión Europea. Seleccionando las cinco especies más características de Pino en Europa: *Pinus pinea*, *Pinus nigra*, *Pinus halepensis*, *Pinus sylvestris* y *Pinus pinaster*. Gracias a un programa simple desarrollado en lenguaje JAVA (código del programa en ANEXO) se automatizó la descarga de los datos que cumplieran con una serie de requisitos iniciales, y así construir la base de datos para el análisis estadístico. Los requisitos fueron:

- Que las cronologías sean de árboles adultos o incluso viejos (pero no senescentes, centenarios o milenarios). Esto es necesario para que el crecimiento sea homogéneo entre las muestras de cada sitio. Los árboles jóvenes suelen formar anillos muy grandes en sus primeros años, y los árboles senescentes muestran una tendencia final de crecimiento negativa, ambas situaciones distorsionan el crecimiento.

Excepción: Existe un caso de *Pinus sylvestris* que no cumple este requisito, porque es centenario, pero se ha dejado para tener una muestra de Pino Silvestre en España.

- Que los últimos años de crecimiento estén en la década del 2000 (específicamente 2001), y que tengan una edad suficiente (más de 90 años). Este punto, es muy importante, ya que para el análisis clúster necesitamos un conjunto de datos con un periodo igual para cada una de las variables (evitar NAs).
- Que las especies se distribuyan a lo largo del territorio europeo, para que no pertenezcan todas a un mismo clima o zona climática específica (como, por ejemplo, la Mediterránea). Queremos analizar cronologías de diferentes ambientes climáticos para observar si las asociaciones o clústeres se crean debido al clima de la zona.

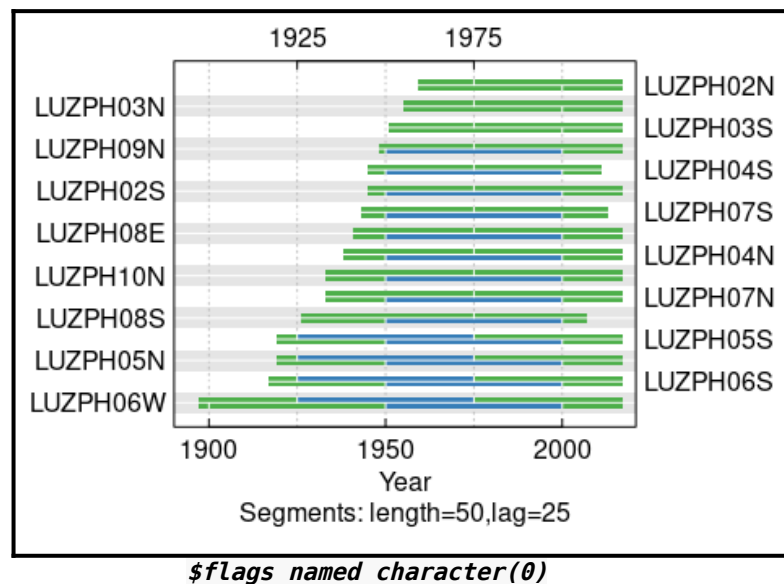
La localización y características de los datos que cumplían con los requisitos se puede observar en la **Tabla 1** y son un total de 17 cronologías. Hay que señalar que de la especie *Pinus pinaster* no hay ninguna cronología ya que todas las series encontradas en la web ninguna cumplía con el requisito de la edad. Ocurriendo lo mismo para la especie *Pinus halepensis* se solucionó utilizando dos cronologías de datos propios (Valeriano-Peñas 2017) para tener variabilidad interespecífica (al menos cuatro especies diferentes).

**Tabla 1.-** Localización y características de los sitios de estudio. Las unidades de la altitud son metros sobre el nivel del mar y las coordenadas se muestran en combinación de latitud y longitud. Datos descargados de la web de: *International Tree-Ring Data Bank (ITRDB)*.

	Sitio (ID)	Autor(es)	Especie	Nº Árboles	Periodo tiempo	Altitud	Coordenadas
1	Divjaka, Albania (ALB005)	Toromani, E.; Pasho, E.; Alla, A.	<i>Pinus pinea</i> L.	10	1862-2010	3	(41.00, 19.48)
2	San Rossore, Italia (ITAL035)	Cherubini, P.	<i>Pinus pinea</i> L.	95	1910-2002	10	(43.70, 10.30)
3	Dehesa del Peral, España (SPAI057)	Briongos, J.M.; Del Cerro Barja, A.	<i>Pinus pinea</i> L.	22	1874-2001	1055	(40.67, -2.77)
4	La Pasadilla, España (SPAI059)	Briongos, J.M.; Cerro-Barja, A.	<i>Pinus pinea</i> L.	31	1907-2001	705	(39.28, -1.35)
5	Mount Etna, Italia (ITAL046)	Seiler, R. et al.	<i>Pinus nigra</i>	104	1785-2013	1850	(37.80, 15.04)
6	Scotida Forest Kastoria, Grecia (GREE008)	Kuniholm, P.I.	<i>Pinus nigra</i>	42	1751-2003	1500	(40.3, 20.9)
7	Ennenda GL Sitenwald, Suiza (SWIT187)	Bigler, C.; Hobi, M.	<i>Pinus nigra</i>	17	1893-2007	620	(47.3, 9.08)
8	Boumort, España (PNS-PBOU)	Gutiérrez, E.; Valeriano, C.	<i>Pinus nigra</i>	14	1900-2013	1100	(42.14, 1.80)
9	El Valle, España (PHValle)	Gutiérrez, E.; Valeriano, C.	<i>Pinus halepensis</i>	14	1860-2016	42-100	(37.59, 1.59)
10	Sierra Espuña, España (PHSE)	Gutiérrez, E.; Valeriano, C.	<i>Pinus halepensis</i>	78	1886-2015	600-1100	(37.52, 1.34)
11	Ketomella PK, Finlandia (FINL072)	Frank, D.C. et al.	<i>Pinus sylvestris</i>	82	1724-2006	300	(68.22, 24.05)
12	Nordrhein-Westfalen, Alemania (GERM188)	Neuwirth, B.	<i>Pinus sylvestris</i>	13	1854-2011	415	(50.6, 6.48)
13	Kologrivsky F. N. Reserve, Rusia (RUSS276)	Hughes, M.K. et al.	<i>Pinus sylvestris</i>	36	1829-2009	150	(58.89, 44.07)
14	Khakassia, Rusia-Siberia (PSRUSE)	Gutiérrez, E.	<i>Pinus sylvestris</i>	73	1832-2012	300	(53.3, 90.0)
15	Pinar de Lillo, España (SAPI083)	Gutiérrez, E.	<i>Pinus sylvestris</i>	32	1511-2002	1600	(43.07, -5.25)
16	Ralla Tall, Suecia (SWED313)	Axelsson, T.	<i>Pinus sylvestris</i>	29	1858-2006	30	(56.77, 16.55)
17	Felsberg GR Brennten, Suiza (SWIT269)	Bigler, C.; Vanoni, M.	<i>Pinus sylvestris</i>	19	1816-2009	1300	(46.85, 9.46)

A continuación, es necesario validar las dataciones de las 17 cronologías que cumplían los requisitos. Para ello se utilizó un paquete específico de R denominado *dplR* (Bunn et al. 2020). Este paquete agrupa las herramientas de dos antiguos programas que se ejecutan a través de línea de comandos: COFECHA y ARSTAN (Holmes 1983; Cook & Krusic 2005).

Con el primero (COFECHA, código R en el ANEXO) se analizan las dataciones de las series individuales (de cada árbol) que forman la cronología media final, consideramos válidas las dataciones si la correlación entre series es significativa ( $p < 0,01$ ) y si no se encuentra ningún *flag* o error de datación (**Fig.3**).



**Figura 3.-** Ejemplo visual del resultado de un análisis de COFECHA, cronologías de la población de *Pinus halepensis* en España (PHValle). Se observa que todos los individuos tienen correlaciones significativas al 95% (color verde) y al 99% (color azul), no se encuentra ningún error de datación (*flags=0*) pues sería la línea de color rojo.

Después de confirmar que las dataciones de los árboles que forman la cronología son correctas se genera la cronología media que se utilizará en los análisis. Las series de crecimiento brutas individuales no tienen media constante, y presentan una disminución del grosor de los anillos debido a la edad. Por lo tanto, para que las series sean comparables unas con otras, es decir que el crecimiento de cada año de cada serie tenga el mismo peso en la cronología final las convertimos (o estandarizamos) en series estacionarias (de media constante) mediante el ajuste de una función crecimiento (Fritts 1976; Cook 1987). El promedio de estas series nos dará la **cronología maestra** para cada sitio cuya fiabilidad depende del número de series que intervengan y del grado de sincronía entre ellas.

Para determinar la fiabilidad y generar la cronología maestra utilizamos el programa (ARSTAN) y el estadístico EPS (*Expressed Population Signal*), cuyo umbral es 0,85 (Wigley et al. 1984). Este estadístico es utilizado para aceptar si una cronología es fiable respecto al número de individuos y durante que periodo de tiempo mantiene la fiabilidad.

En la **Tabla 2** podemos ver los resultados de esta validación, de los 17 sitios analizados, dos sitios no podrán utilizarse en el análisis pues sus cronologías no son fiables ya que no cumplen con el estadístico EPS.

**Tabla 2.-** Descripción de los datos dendrocronológicos. Correlación entre la serie de crecimiento individual con la cronología maestra de cada sitio (datación correcta). Edad media de los árboles en años. Y valores de EPS para cada parcela. EPS = señal de la población expresada por la muestra, significativamente más de 0.8 Resultados obtenidos con el paquete dplR. \*Sitios eliminados para el análisis

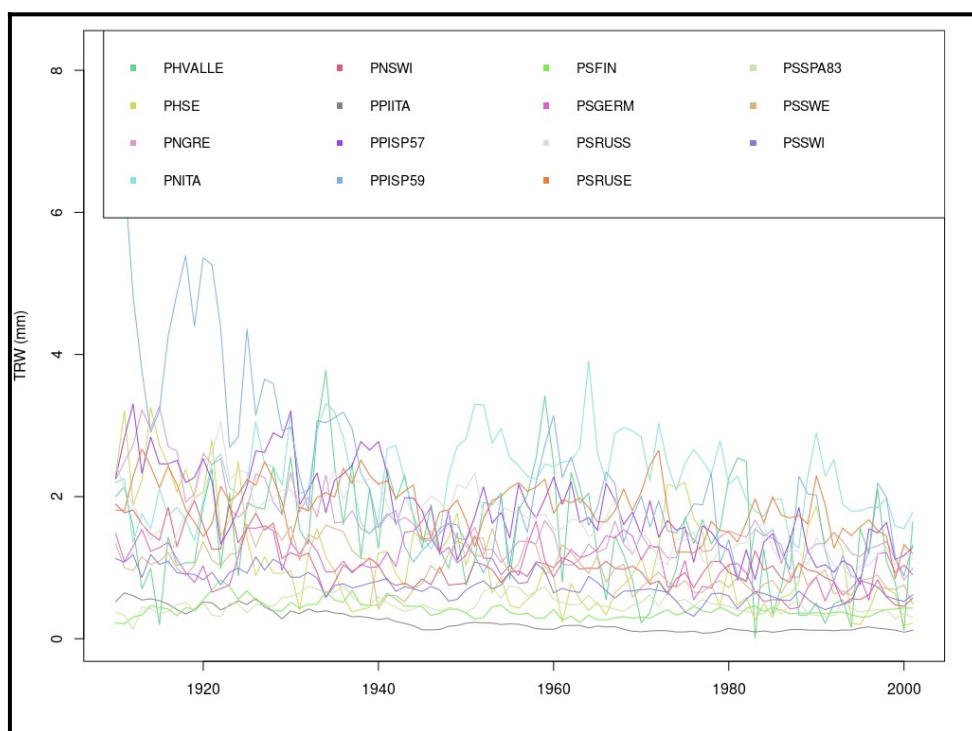
	ID	Nº Series individuales	Correlación entre series	Edad media	EPS > 0.8
1*	ALB005 -> PPIALB	10	0.454	102.1	1970-2010
2	ITAL035 -> PPIITA	95	0.511	75.8	1928-2002
3	SPAI057 -> PPISP57	22	0.629	88.4	1902-2001
4	SPAI059 -> PPISP59	31	0.644	62.9	1907-2001
5	ITAL046 -> PNITA	104	0.478	97	1889-2013
6	GREE008 -> PNGRE	42	0.505	100.2	1879-2003
7	SWIT187 -> PNSWI	17	0.435	97.7	1908-2007
8*	PNSPBOU	13	0.467	109.7	EPS=0.69
9	PHValle	16	0.710	82	1942-2016
10	PHSE	182	0.827	95.6	1891-2015
11	FINL072 -> PSFIN	82	0.590	167.4	1757-2006
12	GERM188 -> PSGERM	13	0.649	150.3	1862-2011
13	RUSS276 -> PSRUSS	36	0.499	113.8	1885-2009
14	PSRUSE	73	0.570	96.1	1838-2012
15	SPAI083 -> PSSPA83	32	0.596	351.2	1603-2002
16	SWED313 -> PSSWE	29	0.658	130	1871-2009
17	SWIT269 -> PSSWI	19	0.560	159.6	1835-2009
15	<b>Cronologías SELECCIONADAS</b>	<b>793 Árboles</b>	<b>&gt; 0.5</b>	<b>92 años</b>	<b>1910-2001</b>

La correlación entre series tampoco llega al 0.5, pero esto ocurre también en otros casos y no es tan importante debido a que, en ocasiones, aunque las dataciones son correctas los árboles presentan baja intercorrelación. Esto puede ser debido a enfermedades de árboles concretos, incendios o simplemente variabilidad entre los crecimientos de árboles de un mismo sitio. A partir de las 15 cronologías maestras validadas se genera la base de datos (código R del procedimiento en ANEXO).

## 4. Resultados

### 4.1. Análisis descriptivo: Cronologías maestras

La base de datos definitiva para el estudio queda compuesta por 15 cronologías medias robustas con un periodo de crecimiento común de 92 años, 1910-2001 (**Fig. 4**). Cada cronología tiene 92 observaciones que corresponden al grosor del anillo (TRW) para cada año. Hay un total de 4 especies diferentes en 8 países distintos. Podemos observar en la **Figura 4** que las series no siguen una tendencia de crecimiento muy marcada y que sobresalen dos cronologías de la especie *P. pinea*: la cronología de España muestra crecimientos muy grandes en la década de los 20 (PPISP59) y la de Italia tiene un crecimiento global muy pequeño y sin mucha variabilidad (PPIITA).



**Figura 4.-** Cronologías medias del grosor de los anillos de los árboles (TRW) para cada sitio validado en milímetros, período seleccionado 92 años (1910-2001). Códigos de leyenda: especie + sitio.

### 4.2. Análisis con el paquete *mclust*

Los análisis realizados con este paquete no nos arrojaron ningún resultado significativo o coherente. Esto es debido a que no tenía en cuenta el factor tiempo. Por lo tanto, al ignorar uno de los dos componentes que forman nuestra matriz de datos no podemos utilizar los resultados obtenidos. Los resultados previos (sin tener en cuenta el tiempo) agrupaban las cronologías en 9 clústers. Dicha clasificación no sigue ninguna lógica ecológica, ya que no agrupa ni por especie, ni por procedencia, ni tipo de crecimiento. Incluso dos de las cronologías de pino silvestre (PS) quedan en grupos individuales (**Tabla 3**).



**Tabla 3.-** Clúster asignado para cada cronología, resultado del FMM realizado con el paquete *mclust*.

Clúster	PH					PPI			PS			PS		
	PH-VA-LL	PH-SE	PN-GRE	PNI-TA	PN-SWI	PPI-ITA	PPI-S-PA5	PPI-SPA-59	PS-FIN	PS-GERM	PS-RUSS	PS-RUSE	PS-SPA-83	PS-SWE
1	2	3	4	4	3	3	5	6	7	4	3	8	9	3

### 4.3. Análisis con los paquetes *funFEM* y *funHDDC*

Antes de realizar el análisis se utilizaron las funciones de *create.fourier.basis* y *smooth.basis* del paquete *fda*, se ha seleccionado el estadístico de Fourier ya que es el que mejor se ajusta al tipo de datos que estamos utilizando (**Fig.1A**, en ANEXO). El número de bases utilizado es la mitad de las observaciones menos uno (Código R en ANEXO). A partir del objeto del paso anterior (*fdaobj*) se llevó a cabo un análisis con la función *funHDDC* aplicando el modelo FLM, y usando las 45 bases de la función Fourier. Se evaluaron todos los submodelos de la función, con un número de clústers de entre 2 y 6. El algoritmo se inició con un preanálisis de *k-means*, puesto que el inicio mediante mini-EM no funcionaba correctamente con los datos.

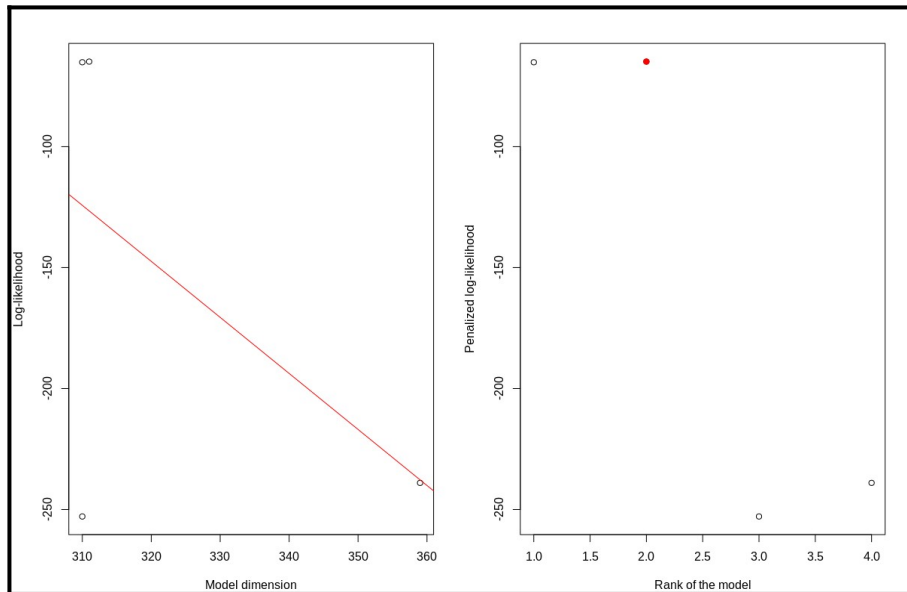
Después del primer análisis se redujo el número de clústers a 2 o 3 y se repitió el análisis (afinando el número de clústers) para obtener el modelo adecuado. Los criterios de información (BIC) resultantes para cada modelo y clúster se pueden ver en la **Tabla 4**. El mejor modelo es  $[ab_kQ_kD_k]$  con dos clústers. En este modelo el parámetro del valor propio 'a' no es dependiente de los clústers (k) y como vimos en la fórmula 2.4.6 los parámetros que lo forman son:  $\rho + \tau + 2K + 1$ . Este mismo análisis con la función *funFEM* determina que el valor óptimo de clústers es también entre 2 y 3, dependiendo del modelo seleccionando.

**Tabla 4.-** Mejor modelo y número de clústers seleccionado por el criterio de información BIC de los dos algoritmos.

Función	Modelo	Nº de clústers (BIC)	
		2	3
funHDDC	$[akjbkQkDk]$	-973	<b>-1003</b>
funHDDC	$[akjbQkDk]$	-1346	-1453
funHDDC	$[akbkQkDk]$	-1345	-
funHDDC	$[akbQkDk]$	-971	-1450
funHDDC	<b><math>[abkQkDk]</math></b>	<b>-969</b>	-
funHDDC	$[abQkDk]$	-1343	-1447
funFEM	$DkBk$	<b>-601</b>	-
funFEM	$DkB$	-710	-525
funFEM	$DBk$	-807	-747
funFEM	$DB$	-916	-957
funFEM	$AkjBk$	<b>-601</b>	<b>-313</b>
funFEM	$AkB$	-710	-523
funFEM	$AkBk$	-601	<b>-310</b>
funFEM	$AkB$	-710	-521
funFEM	$AjBk$	-807	-748
funFEM	$AjB$	-916	-938
funFEM	$ABk$	-807	-749
funFEM	$AB$	-916	-960

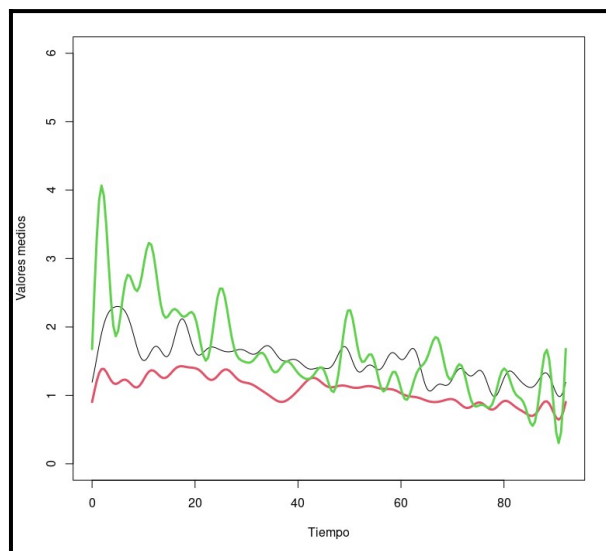
Si validamos los resultados con la pendiente heurística vemos que nos sugiere dos grupos (**Figura 5**).

Estas dos gráficas están relacionadas, la línea roja de la figura de la izquierda se utiliza para calcular la función de probabilidad de verosimilitud penalizada (*penalized log-likelihood*) del eje de la gráfica de la derecha.



**Figura 5.-** Criterio de la pendiente heurística (*slope heuristic*).

El resultado final del análisis lo vemos en la **Figura 6**, se observan los valores medios de los datos agrupados en tres clústers con el modelo general [akjbkQkDk]. La línea negra asocia las poblaciones del clúster 1, formado por: *P. halepensis* de SE (España), *P. nigra* de Italia y *P. pinea* de Italia y España (57). El clúster 2 (color rojo) incluye: el *P. nigra* de Suiza junto con todos los *P. Sylvestris* (a excepción de los Rusos). Por último, el clúster 3 (color verde) liga los *P. sylvestris* de Rusia, *P. nigra* de Grecia, *P. pinea* de España (59) y *P. halepensis* del Valle (España).



**Figura 6.-** Valores medios de las cronologías agrupadas en tres clústers.

Estos resultados, mucho más interesantes que los generados con el paquete *mclust*, nos describen unas agrupaciones con algo de sentido

ecológico y tendencias negativas del crecimiento. Por ejemplo, uno de los clústers está formado principalmente por pino silvestre (a excepción de los árboles rusos los cuales viven en unas condiciones climáticas muy diferentes al resto) y el p. *nigra* de las montañas suizas, también todos ellos comparten edades avanzadas en sus cronologías. El clúster 1, no se decanta por una especie mayoritaria ni un periodo de edad concreto, pero todas son poblaciones Mediterráneas a una altitud mayor de 1000 m. s.n.m. (menos el P. *pinea* de Italia), destacar que las dos cronologías distintas que comentábamos en la Figura 4 se encuentran en este grupo. Por último, el tercer clúster tiene las cuatro especies y no se encuentran características similares entre ellas, por lo que se piensa que con dos clústers la agrupación sería más lógica. Estos resultados no llegan a ser concluyentes y por ello se realizó un tercer análisis con un paquete de modelos mixtos (lcmm).

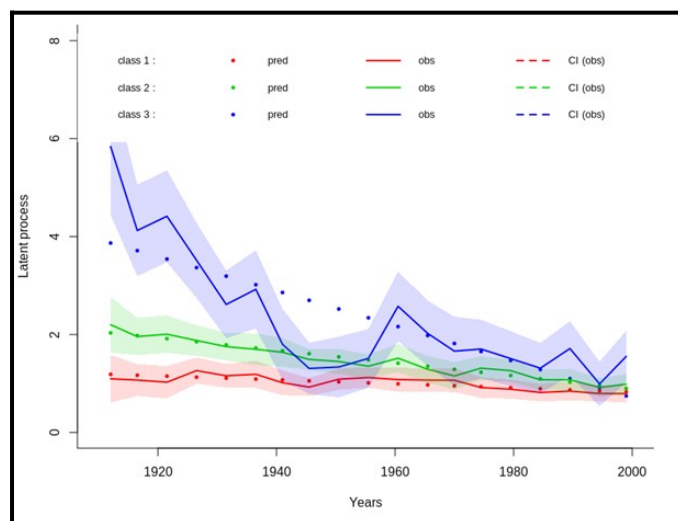
#### 4.4. Análisis con el paquete *lcmm*

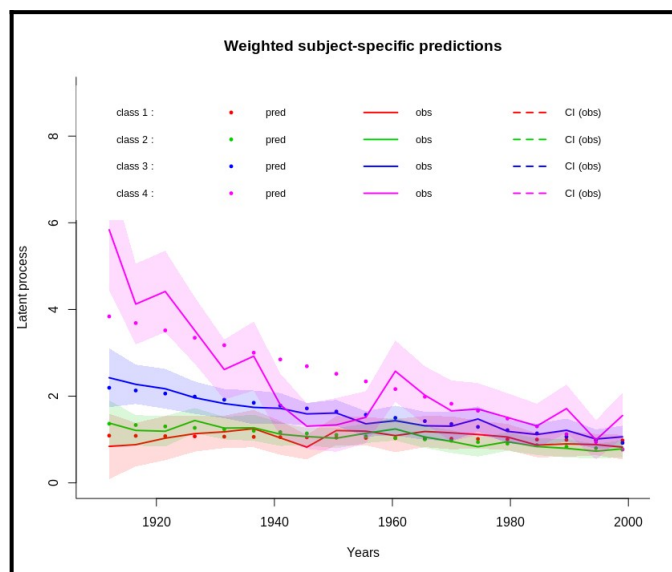
Para realizar el análisis con la función *hlme*, la base de datos tiene que estar configurada por columnas, en nuestro caso tres. En la primera columna se encuentra la variable aleatoria (el tiempo), en la segunda columna esta la variable fija o longitudinal (TRW) y por último en la tercera columna tenemos un identificador único para cada cronología (ID), el código R del análisis se encuentra en el ANEXO.

El modelo diseñado para este estudio fue:

TRW ~ poly (TIME, degree=2, raw=TRUE) ...

En el análisis anterior vimos que posiblemente dos clústers eran mejor que tres, por ello se comenzó estimando el modelo para dos clases latentes (K=2). Al graficar los valores predictivos (por el modelo) frente a los observados una de las dos clases no respondía adecuadamente ([Fig. 2A](#), en el ANEXO). Por lo tanto, se repitió el análisis para tres y cuatro clases (**Figura 7**). En ambos casos (3 y 4 clústers), se observa que dos (tres) clases ajustan bien, y que una de las clases no encaja del todo.





**Figura 7.-** Representación de los valores predichos y observados para dos modelos con diferentes clústers, figura arriba K=3 y figura abajo K=4.

Analizando todos los modelos con los criterios AIC, BIC y el Test de Wald validamos el más correcto (dependiendo de los clústers). En la **Tabla 5** vemos que las agrupaciones de 3 y 4 clases son las más adecuadas en función de su BIC y AIC, la agrupación de 4 sería la más ajustada de todas por unas décimas. Ninguno de los p-valores del test de Wald es menor a 0.05 lo que significa que los modelos no serían útiles para representar la relación entre los valores estimados y propuestos.

**Tabla 5.-** Comparación y selección del mejor modelo, mediante los criterios de información y el Test de Wald.

Modelo	Loglik	Nº de Pa- rámetros	BIC	AIC	Porcentaje de Cronologías por clase				Test de Wald P-valor
					Clase 1	Clase 2	Clase 3	Clase 4	
K=1	-1738.0	10	3505.1	3497.0	100				1
K=2	-912.7	11	1855.2	1847.4	73.3	26.6			1
K=3	-864.2	15	1768.9	1758.3	53.3	40.0	6.6		0.42254
K=4	-858.7	19	1768.8	1755.4	20	46.6	26.6	6.6	0.88397
K=5	-871.8	23	1805.0	1789.7	53.3	40	0	0	0.97305

Si analizamos los resultados del modelo de cuatro clústers vemos que la clase 4 sólo tiene una muestra, que es la cronología que no ajusta bien en la **Figura 7**.

Este análisis es muy revelador en comparación con el estudio anterior, ya que concreta de forma mucho más ajustada el número de clases de latencia que se encuentran en los datos. Vemos que la clase 4 tiene unos crecimientos decrecientes al final y muy deformados al inicio, lo cual es difícil de modelizar, la clase 3 también tiende a disminuir y las clases 1 y 2 se mantienen constantes. Sería interesante predecir las tendencias de las clases 1 a la 3 e incluir covariables de altitud y clima.

## 5. Conclusiones

Después del estudio bibliográfico llevado a cabo en la primera parte del trabajo se han obtenido una gran cantidad de conocimientos sobre el análisis clustering basado en modelos. A partir de los resultados se observa que no es fácil encontrar el paquete mas adecuado a la hora de realizar análisis basados en modelos y que, aunque dos de los paquetes arrojaron resultados matemáticamente lógicos, no podemos decir que sean resultados significativos ecológicamente hablando. Sería muy interesante utilizar el paquete de modelos mixtos para analizar datos dendrocronológicos de estudios propios, debido a que la posibilidad de incluir covariables como la altitud, el clima o alguna característica ecológica nos puede generar diferentes clases latentes muy útiles en el estudio dendroecológico.

### ¿Qué lecciones se han aprendido del trabajo?

La totalidad del desarrollo de este TFM ha sido un aprendizaje continuo, desde el repaso de las bases de la técnica clustering ahondando en el inmenso mundo de los clustering basados en modelos y los modelos mixtos. Toda esta metodología se ha tenido que aprender desde sus bases y en profundidad para poder entender cada una de las funciones de los paquetes utilizados para el análisis. También cada paquete ha sido un aprendizaje en sí mismo, gracias a estudios desarrollados con esos paquetes se han podido llevar a cabo pruebas antes de los análisis con los datos de este TFM. Específicamente he aprendido que el paquete *mclust* aunque sea uno de los más utilizados para este tipo de clustering no tiene en cuenta el tiempo, y los paquetes funcionales (*funFEM* y *fun HDDC*) no trabajaban directamente con la base de datos necesitan de otro paquete para crear las bases y la función que describen los datos del estudio. Del último paquete (*lcmm*) el aprendizaje ha sido menor pues sólo he tenido tiempo de analizar la función concreta (*hlme*) que necesitaba para mis datos, pero es un paquete con muchas funciones y utilidades en el campo del análisis de datos.

A nivel dendrocronología, la utilización de datos de la web pública también fue una novedad. No sólo el hecho de aprender a bajarse gran cantidad de datos de esta web, sino también descubrir la baja calidad de los datos que están allí colgados y lo poco actualizados que están. En el apartado tres de este TFM se ha mostrado que de 27 cronologías posibles (¡Cinco especies diferentes y en toda Europa!) sólo 15 pudieron ser utilizadas para este estudio, sin contar con que dos de las 15 son añadidas de otros trabajos. En este aspecto la lección más valiosa que obtengo es la importancia de los datos y de ser impecable cuando generamos datos y los publicamos.

## Reflexión crítica sobre el logro de los objetivos

Los objetivos generales se han cumplido de forma parcial ya que se han realizado los objetivos 2 y 3 pero los resultados no son satisfactorios. En el caso del objetivo 1, creo que, sí que se ha logrado, pero hace falta mucho más tiempo y unos datos mejores para afirmar que esta metodología puede ser utilizada dentro de los estudios dendroecológicos.

Para los objetivos específicos, la sección de formación se ha cumplido a la perfección, incluyendo a lo largo del estudio una ampliación en el número de paquetes para lograr resultados correctos (inicialmente sólo estaba *mclust*). La sección de desarrollo se ha seguido paso por paso y únicamente la tarea 8 no ha podido llevarse a cabo por falta de tiempo al incluir los nuevos paquetes.

## Análisis crítico del seguimiento de la planificación y metodología

Pese a que se inició el TFM con una planificación muy correcta e interesante (PEC 0), los acontecimientos obligaron a buscar un plan B para poder llevar a cabo el estudio, cerraron los laboratorios y los datos propios se tuvieron que descartar y se decidió utilizar datos públicos.

Del análisis de riesgos definido, han ocurrido todos los problemas posibles. A posteriori, se han incluido en ese apartado algunas de las soluciones utilizadas, pero en general todo lo que podía salir mal salió mal.

En lo que tiene que ver con la metodología y seguimiento, se han cumplido las expectativas y se ha profundizado mucho en nuevas metodologías bastante innovadoras. Pero globalmente el tiempo ha sido un hándicap muy grande, ya que intentar realizar todos los análisis profundizando en su funcionalidad ha supuesto una carrera de fondo.

## Líneas de trabajo futuro que no se han podido explorar

Lo primero de todo, sería muy interesante realizar estos análisis con datos fiables y con más variabilidad de especies o sitios. También me gustaría realizar la tarea 8, ya que el paquete CPCA de análisis de componentes principales, me parece muy prometedor. Del paquete de modelos mixtos habría que profundizar en muchas de las funcionalidades, de la función utilizada (*hlme*) sería interesante incluir covariables en el análisis y realizar predicciones.

## AGRADECIMIENTOS

Me gustaría agradecer el apoyo de mi tutor Daniel Fernández, que pese a todos los problemas y retrasos que hemos sufrido siempre ha estado ahí intentando avanzar y en contacto directo conmigo.

## 6. Glosario

**AIC:** criterio de información de Akaike (*Akaike information criterion*)

**BIC:** criterio de información bayesiano (*bayesian information criterion*)

**EM:** esperanza-maximización (*expectation-maximization*)

**EPS:** señal expresada de la población (*Expressed Population Signal*)

**FLM:** modelo funcional de mezclas latentes (*functional latent mixture model*)

**FMM:** modelos de mezcla/mixtura finita (*finite mixture model*)

**gmm:** modelos de mezcla/mixtura de crecimiento (*growth mixture model*)

**GMM:** modelo de mezcla Gaussiana (*Gaussian mixture model*)

**ITRDB:** base de datos internacional de anillos de los árboles (*International Tree-Ring Data Bank*)

**lcmm:** modelos mixtos usando clases de latencia (*latent classes mixed models*)

**NA:** valores ausentes/no disponibles (*not available*)

**R:** lenguaje de programación, software gratuito y de código abierto

**RStudio:** entorno de desarrollo integrado para el lenguaje R

**TRW:** grosor del anillo de los árboles (*Tree-Ring width*)

## 7. Bibliografía

- Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions On Automatic Control*,19: 716-23.
- Amoroso, M. M., Daniels, L. D., Baker, P. J., & Camarero, J. J. (Eds.). (2017). *Dendroecology: tree-ring analyses applied to ecological studies* (Vol. 231). Springer.
- Arlot, S. (2019). Minimal penalties and the slope heuristics: a survey. *arXiv preprint arXiv:1901.07277*.
- Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 803-821.
- Birgé, L. & Massart, P. (2006). Minimal penalties for gaussian model selection. *Probability Theory and Related Fields*, 138(1-2):33–73
- Bouveyron, C.; Côme, E.; and Jacques, J. (2014). The discriminative functional mixture model for the analysis of bike sharing systems, Preprint HAL n.01024186, University Paris Descartes.
- Bouveyron, C., & Jacques, J. (2011). Model-based clustering of time series in group-specific functional subspaces. *Advances in Data Analysis and Classification*, 5(4), 281-300.
- Bouveyron, C.; Cheze, L.; Jacques, J.; Martin, P. and Schmutz, A. (2017). Clustering multivariate functional data in group-specific functional subspaces, Preprint HAL 01652467.
- Bunn A, Korpela M, Biondi F, Campelo F, Mérian P, Qeadan F, Zang C (2020). *dplR: Dendrochronology Program Library in R*. R package version 1.7.1, <https://CRAN.R-project.org/package=dplR>.
- Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern recognition*, 28(5):781–793.
- Cook, E.R. (1987). The decomposition of tree-ring series for environmental studies. *Tree-Ring Bulletin* 47:37-59.
- Cook, E.R. & Krusic, P.J. (2005). *Program ARSTAN: a tree-ring standardization program based on detrending and autoregressive time series modeling, with interactive graphics*. Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood for incomplete data via the EM algorithm. *J. R. Stat. Soc., B*, 39, 1–38.
- Dorado-Liñán, I., Cañellas, I., Valbuena-Carabaña, M., Gil, L., & Gea-Izquierdo, G. (2017). Coexistence in the Mediterranean-Temperate transitional border: Multi-century dynamics of a mixed old-growth forest under global change. *Dendrochronologia*, 44, 48–57.
- Dorado-Liñán, I., Piovesan, G., Martínez-Sancho, E., Gea-Izquierdo, G., Zang, C., Cañellas, I., ... & Fernández-de-Uña, L. (2019). Geographical adaptation prevails over species-specific determinism in trees' vulnerability to climate change at Mediterranean rear-edge forests. *Global change biology*, 25(4), 1296-1314.



- Fernández, D., Arnold, R., & Pledger, S. (2016). Mixture-based clustering for the ordered stereotype model. *Computational Statistics & Data Analysis*, 93, 46-75.
- Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The computer journal*, 41(8), 578-588.
- Fritts, H. C. (1976). *Tree rings and climate*. Academic, San Diego, California.
- García-González, I. (2008). Comparison of different distance measures for cluster analysis of tree-ring series. *Tree-Ring Research* 64(1):27-37.
- Gutiérrez, E. (2009). La dendrocronología: métodos y aplicaciones. En *Arqueología náutica mediterránea*. 303-315.
- Grissino-Mayer, H. D. & Fritts, H. C. (1997). The International Tree-Ring Data Bank: an enhanced global database serving the global scientific community. *The Holocene*, 7(2), 235-238.
- Haneca, K., Wazny, T., Van Acker, J., and Beeckman, H. (2005), "Provenancing Baltic Timber from Art Historical Objects: Success and Limitations", *Journal of Archaeological Science*, 32, 261–271.
- Heres, A. M.; Martínez-Vilalta, J. & López, B. C. (2012). Growth patterns in relation to drought-induced mortality at two Scots pine (*Pinus sylvestris* L.) sites in NE Iberian Peninsula. *Trees*, 26(2), 621-630.
- Holmes, R.L. (1983). Computer-assisted quality control in tree-ring dating and measurement. *Tree-Ring Bulletin*, 43: 68–78.
- Jung, T., & Wickrama, K. A. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and personality psychology compass*, 2(1), 302-317.
- Legendre, P., and L. Legendre, 1998. *Numerical Ecology*. Elsevier Scientific Publishing Co., New York. Chapter 8, pages 303-385.
- Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2(2), 164-168.
- Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern recognition*, 38(11), 1857-1874.
- Marquardt, D. (1963) An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *SIAM Journal on Applied Mathematics*, 11, 431-441.
- McCulloch, C. E. (2003, January). Generalized linear mixed models. In NSF-CBMS regional conference series in probability and statistics (pp. i-84). Institute of Mathematical Statistics and the American Statistical Association.
- McLachlan, G. y Krishnan, T. 1997. *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics, New York.
- McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite mixture models. *Annual review of statistics and its application*, 6, 355-378.
- McNicholas, P. D. (2016). Model-based clustering. *Journal of Classification*, 33(3), 331-373.
- Melnykov, V., & Maitra, R. (2010). Finite mixture models and model-based clustering. *Statistics Surveys*, 4, 80-116.

- Michael, S., & Melnikov, V. (2016). Finite mixture modelling of Gaussian regression time series with application to dendrochronology. *Journal of Classification*, 33(3), 412-441.
- Nagin, D. S., Jones, B. L., Passos, V. L., & Tremblay, R. E. (2018). Group-based multi-trajectory modeling. *Statistical Methods in Medical Research*, 27(7), 2015-2023.
- Proust-Lima, C.; Philipps, V. & Liquet, B. (2015). Estimation of extended mixed models using latent classes and latent processes: the R package lcmm. arXiv preprint arXiv:1503.00890.
- Ram, N., & Grimm, K. J. (2009). Methods and measures: Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups. *International journal of behavioral development*, 33(6), 565-576.
- Ramsay, J. O., & Silverman, B. W. (1997). Principal differential analysis. In *Functional data analysis* (pp. 239-256). Springer, New York, NY.
- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>. R version 4.0.1 (2020-06-06).
- RStudio Team (2020). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL <http://www.rstudio.com/>. RStudio version 1.3.959.
- Sánchez-Salguero, R., Navarro-Cerrillo, R. M., Camarero, J. J., & Fernández-Cancio, Á. (2012). Selective drought-induced decline of pine species in southeastern Spain. *Climatic Change*, 113(3-4), 767-785.
- Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., ... & Lin, C. T. (2017). A review of clustering techniques and developments. *Neurocomputing*, 267, 664-681.
- Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461-4.
- Shestakova, T. A., Gutiérrez, E., Valeriano, C., Lapshina, E., & Voltas, J. (2019). Recent loss of sensitivity to summer temperature constrains tree growth synchrony among boreal Eurasian forests. *Agricultural and forest meteorology*, 268, 318-330.
- Scrucca L, Fop M, Murphy TB, Raftery AE (2016). "mclust 5: clustering, classification and density estimation using Gaussian finite mixture models." *The R Journal*, 8(1), 289–317. <https://doi.org/10.32614/RJ-2016-021>.
- Valeriano-Peñas, C. (2017). Plasticidad del *Pinus halepensis* Mill. frente al cambio climático. Trabajo Final de Máster, Universitat de Barcelona. Dirigido por la Doctora Emilia Gutiérrez.
- Wigley, T. M., Briffa, K. R., & Jones, P. D. (1984). On the average value of correlated time series, with applications in dendroclimatology and hydrometeorology. *Journal of climate and Applied Meteorology*, 23(2), 201-213.

# Anexos

## Programa de descarga: Código Java

```
import java.io.BufferedInputStream;
import java.io.File;
import java.io.FileNotFoundException;
import java.io.FileOutputStream;
import java.io.FileReader;
import java.io.IOException;
import java.net.URL;
import org.json.simple.JSONArray;
import org.json.simple.JSONObject;
import org.json.simple.parser.JSONParser;
import org.json.simple.parser.ParseException;
import org.jsoup.Jsoup;
import org.jsoup.nodes.Document;
import org.jsoup.nodes.Element;
import org.jsoup.select.Elements;
# Acceso a la web de descarga
public class CrawlerMain {
    private static String base = "c:/database/";
    public static void main(String[] args) {
        String url = "https://www.ncdc.noaa.gov/paleo-
search/study/search.json?
dataPublisher=NOAA&dataTypeId=18&locations=Continent%3EEurope%3EWester
n%20Europe%3EFrance&species=PISY&cvWhats=physical%20property%3Elength
%3Etotal%20ring%20width
h&headersOnly=true";
# Método para descargar el resultado de la consulta al web-service
        download(url, "search");
        try {
# Web-service devuelve una estructura de datos JSON y se analiza para
acceder a las partes del objeto que nos interesa
            readJSON("search");
        } catch (FileNotFoundException e) {
            // TODO Auto-generated catch block
            e.printStackTrace();
        } catch (IOException e) {
            // TODO Auto-generated catch block
            e.printStackTrace();
        } catch (ParseException e) {
            // TODO Auto-generated catch block
            e.printStackTrace();
        }
    }
# Método que descarga un recurso web a partir de una url válida
    private static void download(String url, String name) {
        try (BufferedInputStream in = new BufferedInputStream(new
URL(url).openStream());
FileOutputStream outputStream = new FileOutputStream(base + name))
        {
            byte dataBuffer[] = new byte[1024];
            int bytesRead;
            while ((bytesRead = in.read(dataBuffer, 0, 1024)) != -1) {
                outputStream.write(dataBuffer, 0, bytesRead);
            }
        } catch (IOException e) {
            // handle exception
        }
    }
}
```

```

# Método que analiza la estructura de datos JSON para obtener las urls
de los archivos RWL y descargar aquellos que cumplen los criterios de
busqueda: años, especie, país...
private static void readJSON(String namer) throws
FileNotFoundException, IOException,
ParseException {
    JSONParser parser = new JSONParser();
    JSONArray catalog = (JSONArray)((JSONObject)parser.parse(new
    FileReader(base +
    namer))).get("study");
    for (Object o : catalog)
    {
        JSONObject entry = (JSONObject) o;
        String xmlId = (String) entry.get("xmlId");
        download("https://www.ncdc.noaa.gov/paleo-search/study/search.json?
        xmlId=" + xmlId,
        xmlId);
        JSONArray studies = (JSONArray)((JSONObject)parser.parse(new
        FileReader(base +
        xmlId))).get("study");
        for (Object oo : studies)
        {
            JSONObject study = (JSONObject) oo;
            String studyName = (String) study.get("studyName");
            JSONArray sites = (JSONArray)study.get("site");
            for (Object s : sites)
            {
                JSONObject site = (JSONObject) s;
                JSONArray paleoDatas = (JSONArray)site.get("paleoData");
                for (Object p : paleoDatas) {
                    JSONObject data = (JSONObject) p;
                    Long earliestYearCE = (Long) data.get("earliestYearCE");
                    Long mostRecentYearCE = (Long) data.get("mostRecentYearCE");
                    if(earliestYearCE >= 1500 && earliestYearCE <= 1890 &&
                    mostRecentYearCE >=1995) {
                        new File(base + studyName).mkdirs();
                        System.out.println("***** " + earliestYearCE + " - "
                        + mostRecentYearCE + " [" + studyName + "]" );
                        JSONArray files = (JSONArray)data.get("dataFile");
                        for (Object f : files)
                        {
                            JSONObject file = (JSONObject) f;
                            String fURL = (String) file.get("fileUrl");
                            String n = (String) file.get("linkText");
                            download(fURL, studyName + "/" + n);
                        }
                    } else {
                        System.out.println(earliestYearCE + " - " +
                        mostRecentYearCE + " [" + studyName + "]" );
                    }
                }
            }
        }
        System.out.printf("");
    }
    System.out.printf("");
}

```

## Código R de la creación y validación de la Base de Datos:

```
library(dplyr)
setwd("/media/cvp/Seagate Expansion
Drive/Universidad-Certificados/UOC-UB-Master BIOESTADISTICA y
BIOINFORMATICA/MASTER UOC 5ºsem TFM/Database/All")
# Lectura de los Ficheros trw
# Phalepensis
PHALBA<- read.rwl("PH-alb004.rwl")

## Attempting to automatically detect format.
## Assuming a Tucson format file.
## There appears to be a header in the rwl file
## There are 40 series
## 1      PHDVA01A      1961      2010      0.01
## 2      PHDVA02A      1967      2010      0.01
## 3      PHDVA03B      1963      2010      0.01
## 4      PHDVA04A      1955      2010      0.01
## 5      PHDVA04B      1966      2010      0.01
## 6      PHDVA05A      1961      2010      0.01
## 7      PHDVA05B      1956      2010      0.01
## 8      PHDVA06A      1953      2010      0.01
## 9      PHDVA06B      1949      2010      0.01
## 10     PHDVA07B      1972      2010      0.01
## 11     PHDVA08A      1961      2010      0.01
## 12     PHDVA08B      1957      2010      0.01
## 13     PHDVA09A      1953      2010      0.01
## 14     PHDVA09B      1949      2005      0.01
## 15     PHDVA10A      1963      2010      0.01
## 16     PHDVA10B      1955      2004      0.01
## 17     PHDVA11A      1967      2010      0.01
## 18     PHDVA12A      1978      2010      0.01
## 19     PHDVA12B      1968      2009      0.01
## 20     PHDVA13A      1955      2010      0.01
## 21     PHDVA14A      1960      2010      0.01
## 22     PHDVA14B      1959      2010      0.01
## 23     PHDVA15A      1956      2009      0.01
## 24     PHDVA15B      1953      2010      0.01
## 25     PHDVA16A      1955      2009      0.01
## 26     PHDVA16B      1958      2009      0.01
## 27     PHDVA17A      1955      2006      0.01
## 28     PHDVA17B      1956      2009      0.01
## 29     PHDVA18A      1958      2010      0.01
## 30     PHDVA18B      1959      2010      0.01
## 31     PHDVA19X      1970      2010      0.01
## 32     PHDVA20B      1966      2010      0.01
## 33     PHDVA21A      1960      2010      0.01
## 34     PHDVA21B      1958      2010      0.01
## 35     PHDVA22A      1960      2010      0.01
## 36     PHDVA22B      1966      2010      0.01
## 37     PHDVA23A      1957      2010      0.01
## 38     PHDVA23B      1991      2010      0.01
## 39     PHDVA24A      1963      2010      0.01
## 40     PHDVA24B      1959      2008      0.01

PHFRAN<- read.rwl("PH-fran008.rwl")

## Attempting to automatically detect format.
## Assuming a Tucson format file.
## There appears to be a header in the rwl file
## There are 21 series
## 1      101011      1833      1973      0.01
```

```

## 2      101012      1875      1973      0.01
## 3      101013      1883      1973      0.01
## 4      101061      1807      1955      0.01
## 5      101062      1816      1954      0.01
## 6      101063      1808      1954      0.01
## 7      101071      1830      1973      0.01
## 8      101072      1846      1973      0.01
## 9      101073      1843      1973      0.01
## 10     101081      1839      1973      0.01
## 11     101082      1838      1973      0.01
## 12     101083      1842      1973      0.01
## 13     101121      1845      1942      0.01
## 14     101122      1873      1973      0.01
## 15     101123      1877      1967      0.01
## 16     101761      1835      1973      0.01
## 17     101762      1844      1973      0.01
## 18     101763      1861      1973      0.01
## 19     101771      1824      1973      0.01
## 20     101772      1845      1973      0.01
## 21     101773      1833      1973      0.01

```

```
PHVALLE<-read.rwl("PH-Spain_ElValleCVP.rwl")
```

```

## Attempting to automatically detect format.
## Assuming a Tucson format file.
## There does not appear to be a header in the rwl file
## There are 16 series
## 1      LUZPH02N      1958      2016      0.01
## 2      LUZPH02S      1944      2016      0.01
## 3      LUZPH03N      1954      2016      0.01
## 4      LUZPH03S      1949      2016      0.01
## 5      LUZPH04N      1937      2016      0.01
## 6      LUZPH04S      1942      2010      0.01
## 7      LUZPH05N      1918      2016      0.01
## 8      LUZPH05S      1917      2016      0.01
## 9      LUZPH06W      1895      2016      0.01
## 10     LUZPH06S      1915      2016      0.01
## 11     LUZPH07N      1930      2016      0.01
## 12     LUZPH07S      1943      2012      0.01
## 13     LUZPH08E      1939      2016      0.01
## 14     LUZPH08S      1923      2006      0.01
## 15     LUZPH09N      1942      2016      0.01
## 16     LUZPH10N      1931      2016      0.01

```

```
PHSE<-read.rwl("PH-Spain_SECV.rwl")
```

```

## Attempting to automatically detect format.
## Assuming a Tucson format file.
## There does not appear to be a header in the rwl file
## There are 182 series
## 1      Mu60011      1917      2015      0.01
## 2      Mu60012      1941      2015      0.01
## 3      Mu60032      1920      2015      0.01
## 4      Mu60031      1942      2015      0.01
## 5      Mu60041      1911      2015      0.01
## 6      Mu60042      1930      2015      0.01
## 7      Mu60091      1915      1993      0.01
## 8      Mu60092      1919      2007      0.01
## 9      Mu60151      1918      2007      0.01
## 10     Mu60152      1919      2006      0.01
## 11     Mu60211      1912      2007      0.01
## 12     Mu60212      1912      1992      0.01
## 13     Mu60391      1940      2007      0.01
## 14     Mu60392      1940      2007      0.01
## 15     Mu60701      1931      2007      0.01
## 16     Mu60811      1915      2007      0.01

```

## 17	Mu60812	1916	2007	0.01
## 18	Mu60991	1922	2007	0.01
## 19	Mu60992	1915	2007	0.01
## 20	Mu61061	1912	2007	0.01
## 21	Mu61062	1916	2007	0.01
## 22	Mu61151	1940	2007	0.01
## 23	Mu61152	1940	2007	0.01
## 24	Mu61381	1915	2007	0.01
## 25	Mu61382	1914	2007	0.01
## 26	Mu61391	1915	2007	0.01
## 27	Mu61392	1916	2007	0.01
## 28	Mu61561	1919	2007	0.01
## 29	Mu61562	1918	2007	0.01
## 30	Mu7010-1	1912	2015	0.01
## 31	Mu7010-2	1913	2015	0.01
## 32	Mu7031-1	1912	2007	0.01
## 33	Mu7031-2	1912	2007	0.01
## 34	Mu7057-1	1925	2015	0.01
## 35	Mu7057-2	1917	2015	0.01
## 36	Mu7060-1	1919	2007	0.01
## 37	Mu7060-2	1922	2007	0.01
## 38	Mu7069-1	1925	2015	0.01
## 39	Mu7069-2	1918	2015	0.01
## 40	Mu7070-1	1918	2007	0.01
## 41	Mu7070-2	1910	1996	0.01
## 42	Mu7074-1	1910	2007	0.01
## 43	Mu7074-2	1915	2007	0.01
## 44	Mu7080-1	1912	2007	0.01
## 45	Mu7080-2	1912	2007	0.01
## 46	Mu7096-1	1911	2007	0.01
## 47	Mu7096-2	1909	2007	0.01
## 48	Mu7097-1	1911	2007	0.01
## 49	Mu7097-2	1914	2007	0.01
## 50	Mu7108-1	1911	2015	0.01
## 51	Mu7108-2	1908	2015	0.01
## 52	Mu7110-1	1913	2007	0.01
## 53	Mu7110-2	1913	2007	0.01
## 54	Mu7113-1	1913	2007	0.01
## 55	Mu7113-2	1912	2007	0.01
## 56	Mu7116-1	1916	2007	0.01
## 57	Mu7116-2	1916	2007	0.01
## 58	Mu7128-1	1922	2015	0.01
## 59	Mu7128-2	1919	2015	0.01
## 60	Mu8-005N	1913	2015	0.01
## 61	Mu8-005S	1930	2015	0.01
## 62	Mu8-0251	1915	2007	0.01
## 63	Mu8-0252	1925	2007	0.01
## 64	Mu8-032N	1910	2015	0.01
## 65	Mu8-032E	1924	2015	0.01
## 66	Mu8-0431	1940	2007	0.01
## 67	Mu8-0432	1916	2007	0.01
## 68	Mu8-0461	1913	2007	0.01
## 69	Mu8-0462	1915	2007	0.01
## 70	Mu8-0531	1910	1995	0.01
## 71	Mu8-0532	1913	2007	0.01
## 72	Mu8-056N	1913	2015	0.01
## 73	Mu8-056S	1913	2015	0.01
## 74	Mu8-0671	1909	2007	0.01
## 75	Mu8-0672	1907	2003	0.01
## 76	Mu8-0701	1909	2007	0.01
## 77	Mu8-0702	1906	2007	0.01
## 78	Mu8-0801	1906	2007	0.01
## 79	Mu8-0802	1912	2007	0.01
## 80	Mu8-0861	1908	2007	0.01
## 81	Mu8-0921	1911	2007	0.01
## 82	Mu8-0922	1909	2006	0.01

## 83	Mu8-0951	1912	2007	0.01
## 84	Mu8-0952	1911	2007	0.01
## 85	Mu8-103N	1918	2015	0.01
## 86	Mu8-103S	1906	2015	0.01
## 87	Mu8-1101	1921	2007	0.01
## 88	Mu8-1102	1911	2007	0.01
## 89	Mu8-1161	1914	2007	0.01
## 90	Mu8-1162	1911	2007	0.01
## 91	Mu8-1211	1908	2007	0.01
## 92	Mu8-1212	1906	2007	0.01
## 93	Mu8-1291	1908	2007	0.01
## 94	Mu8-1292	1912	2007	0.01
## 95	Mu8-1301	1906	2007	0.01
## 96	Mu8-1302	1909	2007	0.01
## 97	Mu9038N	1902	2015	0.01
## 98	Mu9038S	1902	2015	0.01
## 99	Mu90231	1908	2007	0.01
## 100	Mu90232	1907	2007	0.01
## 101	Mu90581	1891	2007	0.01
## 102	Mu90681	1895	2015	0.01
## 103	Mu90682	1900	2015	0.01
## 104	Mu90741	1896	2007	0.01
## 105	Mu90742	1896	1990	0.01
## 106	Mu90761	1900	2007	0.01
## 107	Mu90762	1890	2007	0.01
## 108	Mu90791	1896	2007	0.01
## 109	Mu90792	1896	2007	0.01
## 110	Mu90831	1889	2007	0.01
## 111	Mu90832	1894	2007	0.01
## 112	Mu90961	1899	2007	0.01
## 113	Mu90962	1899	2007	0.01
## 114	Mu91211	1886	2007	0.01
## 115	Mu91212	1898	2007	0.01
## 116	Mu91241	1900	2015	0.01
## 117	Mu91242	1900	2015	0.01
## 118	Mu91271	1896	2007	0.01
## 119	Mu91272	1897	2007	0.01
## 120	Mu91281	1901	2007	0.01
## 121	Mu91282	1901	2007	0.01
## 122	Mu91291	1904	2004	0.01
## 123	Mu91292	1912	2007	0.01
## 124	Mu11001N	1960	2015	0.01
## 125	Mu11001S	1919	2015	0.01
## 126	Mu110091	1923	2007	0.01
## 127	Mu110092	1928	2007	0.01
## 128	Mu110111	1922	2007	0.01
## 129	Mu110112	1921	2007	0.01
## 130	Mu110141	1922	2007	0.01
## 131	Mu110171	1942	2007	0.01
## 132	Mu110361	1921	2006	0.01
## 133	Mu110362	1922	2006	0.01
## 134	Mu110381	1917	2007	0.01
## 135	Mu110382	1943	2007	0.01
## 136	Mu11047N	1919	2015	0.01
## 137	Mu11047S	1923	2015	0.01
## 138	Mu110651	1934	2007	0.01
## 139	Mu110652	1920	2003	0.01
## 140	Mu110801	1920	2007	0.01
## 141	Mu110802	1920	2007	0.01
## 142	Mu110811	1917	2007	0.01
## 143	Mu11089W	1932	2015	0.01
## 144	Mu11089N	1922	2015	0.01
## 145	Mu111091	1922	2007	0.01
## 146	Mu11094N	1930	2015	0.01
## 147	Mu11094S	1921	2015	0.01
## 148	Mu11112N	1920	2015	0.01



```

## 149 Mu111125 1920 2015 0.01
## 150 Mu111152 1920 2007 0.01
## 151 Mu120011 1916 2007 0.01
## 152 Mu120012 1917 2007 0.01
## 153 Mu120041 1915 2007 0.01
## 154 Mu120042 1924 2007 0.01
## 155 Mu120241 1912 2007 0.01
## 156 Mu120242 1909 1993 0.01
## 157 Mu120381 1908 2007 0.01
## 158 Mu120382 1912 2007 0.01
## 159 Mu120471 1906 2007 0.01
## 160 Mu120472 1910 2007 0.01
## 161 Mu120711 1910 2007 0.01
## 162 Mu120712 1907 2007 0.01
## 163 Mu120801 1913 2007 0.01
## 164 Mu120802 1910 2007 0.01
## 165 Mu120921 1911 2015 0.01
## 166 Mu120922 1904 2015 0.01
## 167 Mu121101 1906 2015 0.01
## 168 Mu121102 1910 2015 0.01
## 169 Mu121151 1907 2007 0.01
## 170 Mu121152 1917 1993 0.01
## 171 Mu121751 1910 2007 0.01
## 172 Mu121752 1902 2007 0.01
## 173 Mu121801 1904 2015 0.01
## 174 Mu121802 1907 2015 0.01
## 175 Mu121811 1907 2007 0.01
## 176 Mu121812 1905 2007 0.01
## 177 Mu122081 1912 1997 0.01
## 178 Mu122082 1909 1996 0.01
## 179 Mu122091 1907 2007 0.01
## 180 Mu122092 1904 2007 0.01
## 181 Mu122151 1910 2015 0.01
## 182 Mu122152 1910 2015 0.01

```

```
PHGA<-read.rwl("PH-Spain_GAEMG.rwl")
```

```

## Attempting to automatically detect format.
## Assuming a Tucson format file.
## There does not appear to be a header in the rwl file
## There are 26 series
## 1 PH01Nea 1958 2014 0.01
## 2 PH01Se 1958 2014 0.01
## 3 PH02E 1933 2014 0.01
## 4 PH02W 1934 2014 0.01
## 5 PH03NWe 1957 2014 0.01
## 6 PH03SEe 1957 2014 0.01
## 7 PH04Ee 1942 2014 0.01
## 8 PH04We 1952 2014 0.01
## 9 PH05SWe 1942 2014 0.01
## 10 PH05NEe 1942 2014 0.01
## 11 PH06NE 1936 2014 0.01
## 12 PH06SWe 1936 2014 0.01
## 13 PH07W 1922 2014 0.01
## 14 PH07Ee 1922 2014 0.01
## 15 PH08Sa 1936 2014 0.01
## 16 PH08Ne 1936 2014 0.01
## 17 PH09Se 1934 2014 0.01
## 18 PH09Ne 1933 2014 0.01
## 19 PH11Se 1950 2014 0.01
## 20 PH11Ne 1940 2014 0.01
## 21 PH12Ne 1938 2014 0.01
## 22 PH12Se 1938 2014 0.01
## 23 PH13Ne 1928 2014 0.01
## 24 PH13Se 1928 2014 0.01

```

```
## 25 PH14No 1926 2014 0.01
## 26 PH014Se 1926 2014 0.01
```

### # *Pnigra*

```
PNAUS<- read.rwl("PN-aust109.rwl")
```

```
## Attempting to automatically detect format.
## Assuming a Tucson format file.
## There appears to be a header in the rwl file
## There are 38 series
```

## 1	BiePN01a	1890	1995	0.01
## 2	BiePN01b	1853	1995	0.01
## 3	BiePN02a	1856	1995	0.01
## 4	BiePN02b	1845	1995	0.01
## 5	BiePN03a	1883	1995	0.01
## 6	BiePN03b	1857	1995	0.01
## 7	BiePN04a	1880	1995	0.01
## 8	BiePN04b	1868	1995	0.01
## 9	BiePN05a	1906	1995	0.01
## 10	BiePN05b	1887	1995	0.01
## 11	BiePN06a	1857	1995	0.01
## 12	BiePN06b	1873	1995	0.01
## 13	BiePN07a	1861	1995	0.01
## 14	BiePN07b	1850	1995	0.01
## 15	BiePN08a	1861	1995	0.01
## 16	BiePN08b	1855	1995	0.01
## 17	BiePN09a	1849	1995	0.01
## 18	BiePN09b	1854	1995	0.01
## 19	BiePN10a	1850	1995	0.01
## 20	BiePN10b	1845	1995	0.01
## 21	BiePN11a	1850	1995	0.01
## 22	BiePN11b	1843	1995	0.01
## 23	BiePN12a	1873	1995	0.01
## 24	BiePN12b	1886	1995	0.01
## 25	BiePN13a	1843	1995	0.01
## 26	BiePN13b	1840	1995	0.01
## 27	BiePN14a	1852	1995	0.01
## 28	BiePN14b	1856	1995	0.01
## 29	BiePN15a	1862	1995	0.01
## 30	BiePN15b	1850	1995	0.01
## 31	BiePN16a	1892	1995	0.01
## 32	BiePN16b	1883	1995	0.01
## 33	BiePN17a	1859	1995	0.01
## 34	BiePN17b	1853	1995	0.01
## 35	BiePN18a	1872	1995	0.01
## 36	BiePN18b	1881	1995	0.01
## 37	BiePN19a	1918	1995	0.01
## 38	BiePN19b	1904	1995	0.01

```
PNGRE<- read.rwl("PN-gree008.rwl")
```

```
## Attempting to automatically detect format.
## Assuming a Tucson format file.
## There does not appear to be a header in the rwl file
## There are 42 series
```

## 1	622011	1850	1978	0.01
## 2	622012	1862	1978	0.01
## 3	622031	1751	1978	0.01
## 4	622041	1893	1978	0.01
## 5	622051	1856	1977	0.01
## 6	622061	1878	1978	0.01
## 7	622071	1895	1978	0.01
## 8	622091	1761	1978	0.01
## 9	622101	1896	1978	0.01
## 10	622111	1913	1978	0.01

```

## 11      622121      1890      1978      0.01
## 12      622231      1920      2002      0.01
## 13      622232      1920      2003      0.01
## 14      622241      1907      2002      0.01
## 15      622242      1907      2002      0.01
## 16      622251      1905      2002      0.01
## 17      622252      1905      2002      0.01
## 18      622261      1893      2002      0.01
## 19      622262      1893      2002      0.01
## 20      622271      1922      2003      0.01
## 21      622272      1922      2003      0.01
## 22      622281      1919      2003      0.01
## 23      622282      1919      2003      0.01
## 24      622291      1922      2000      0.01
## 25      622292      1922      2002      0.01
## 26      622301      1927      2003      0.01
## 27      622302      1927      2003      0.01
## 28      622311      1912      2003      0.01
## 29      622312      1912      2003      0.01
## 30      622321      1918      2003      0.01
## 31      622322      1918      2003      0.01
## 32      633331      1901      2002      0.01
## 33      622332      1901      2002      0.01
## 34      622341      1889      2002      0.01
## 35      622342      1889      2002      0.01
## 36      622351      1893      2003      0.01
## 37      622352      1893      2003      0.01
## 38      622361      1920      2003      0.01
## 39      622362      1920      2003      0.01
## 40      622371      1913      2003      0.01
## 41      622372      1913      2003      0.01
## 42      622381      1903      2003      0.01

```

```
PNITA<- read.rwl("PN-ital046.rwl")
```

```

## Attempting to automatically detect format.
## Assuming a Tucson format file.
## There appears to be a header in the rwl file
## There are 104 series
## 1      PINI1      1906      2002      0.01
## 2      PINI2      1899      2001      0.01
## 3      PINI3      1903      2009      0.01
## 4      PINI4      1904      2009      0.01
## 5      PINI5      1957      2010      0.01
## 6      PINI6      1959      2010      0.01
## 7      PINI7      1897      2010      0.01
## 8      PINI8      1908      2010      0.01
## 9      PINI9      1927      1991      0.01
## 10     PINI10     1914      2010      0.01
## 11     PINI11     1921      2010      0.01
## 12     PINI12     1924      2010      0.01
## 13     PINI13     1919      2010      0.01
## 14     PINI14     1903      2010      0.01
## 15     PINI15     1895      2010      0.01
## 16     PINI16     1925      2009      0.01
## 17     PINI17     1907      2009      0.01
## 18     PINI18     1923      2010      0.01
## 19     PINI19     1910      2002      0.01
## 20     PINI20     1908      2010      0.01
## 21     PINI21     1884      2010      0.01
## 22     PINI22     1891      2010      0.01
## 23     PINI23     1928      2010      0.01
## 24     PINI24     1906      2002      0.01
## 25     PINI25     1904      2010      0.01
## 26     PINI26     1901      2010      0.01
## 27     PINI27     1925      2010      0.01

```

## 28	PINI28	1923	2010	0.01
## 29	PINI29	1904	2010	0.01
## 30	PINI30	1901	2003	0.01
## 31	PINI31	1926	2010	0.01
## 32	PINI32	1933	2010	0.01
## 33	PINI33	1919	2010	0.01
## 34	PINI34	1925	2010	0.01
## 35	PINI35	1961	2010	0.01
## 36	PINI36	1809	2011	0.01
## 37	PINI37	1785	2011	0.01
## 38	PINI38	1952	2010	0.01
## 39	PINI39	1958	2010	0.01
## 40	PINI40	1803	2012	0.01
## 41	PINI41	1790	2012	0.01
## 42	PINI42	1908	2010	0.01
## 43	PINI43	1912	2010	0.01
## 44	PINI44	1953	2008	0.01
## 45	PINI45	1952	2008	0.01
## 46	PINI46	1908	2007	0.01
## 47	PINI47	1931	2006	0.01
## 48	PINI48	1930	2007	0.01
## 49	PINI49	1906	2007	0.01
## 50	PINI50	1952	2007	0.01
## 51	PINI51	1954	2006	0.01
## 52	PINI52	1899	2007	0.01
## 53	PINI53	1798	2007	0.01
## 54	PINI54	1957	2007	0.01
## 55	PINI55	1932	2007	0.01
## 56	PINI56	1929	2007	0.01
## 57	PINI57	1932	2007	0.01
## 58	PINI58	1976	2007	0.01
## 59	PINI59	1902	2007	0.01
## 60	PINI60	1958	2007	0.01
## 61	PINI61	1965	2007	0.01
## 62	PINI62	1964	2006	0.01
## 63	PINI63	1932	2007	0.01
## 64	PINI64	1904	2002	0.01
## 65	PINI65	1916	2007	0.01
## 66	PINI66	1906	2007	0.01
## 67	PINI67	1958	2007	0.01
## 68	PINI68	1915	2007	0.01
## 69	PINI69	1894	2007	0.01
## 70	PINI70	1907	2007	0.01
## 71	PINI71	1914	2007	0.01
## 72	PINI72	1966	2013	0.01
## 73	PINI73	1972	2012	0.01
## 74	PINI74	1917	2012	0.01
## 75	PINI75	1948	2012	0.01
## 76	PINI76	1961	2013	0.01
## 77	PINI77	1898	2003	0.01
## 78	PINI78	1941	2013	0.01
## 79	PINI79	1892	2013	0.01
## 80	PINI80	1901	2010	0.01
## 81	PINI81	1898	2011	0.01
## 82	PINI82	1897	2013	0.01
## 83	PINI83	1902	2013	0.01
## 84	PINI84	1910	2013	0.01
## 85	PINI85	1908	2013	0.01
## 86	PINI86	1907	2010	0.01
## 87	PINI87	1942	2010	0.01
## 88	PINI88	1905	2013	0.01
## 89	PINI89	1935	2013	0.01
## 90	PINI90	1929	2010	0.01
## 91	PINI91	1934	2010	0.01
## 92	PINI92	1911	2011	0.01
## 93	PINI93	1909	2011	0.01

```

## 94 PINI94 1816 2013 0.01
## 95 PINI95 1811 2011 0.01
## 96 PINI96 1887 2013 0.01
## 97 PINI97 1906 2012 0.01
## 98 PINI98 1903 2011 0.01
## 99 PINI99 1915 2011 0.01
## 100 PINI100 1902 2011 0.01
## 101 PINI101 1897 2011 0.01
## 102 PINI102 1928 2013 0.01
## 103 PINI103 1896 2003 0.01
## 104 PINI104 1900 2009 0.01

```

```
PNSP17<-read.rwl("PN-spai017.rwl")
```

```

## Attempting to automatically detect format.
## Assuming a Tucson format file.
## There appears to be a header in the rwl file
## There are 14 series
## 1 GR501EPN 1769 1991 0.01
## 2 GR501SPN 1754 1991 0.01
## 3 GR5020PN 1832 1991 0.01
## 4 GR502SPN 1824 1991 0.01
## 5 GR5030PN 1785 1991 0.01
## 6 GR503SPN 1830 1991 0.01
## 7 GR5040PN 1852 1991 0.01
## 8 GR504EPN 1838 1991 0.01
## 9 GR505SPN 1852 1991 0.01
## 10 GR5050PN 1859 1991 0.01
## 11 GR204NPN 1818 1987 0.01
## 12 GR204SPN 1815 1987 0.01
## 13 GR206XPN 1812 1986 0.01
## 14 GR206YPN 1818 1987 0.01

```

```
PNSPBOU<-read.rwl("PN-Spain_BoumortEMG.RWL")
```

```

## Attempting to automatically detect format.
## Assuming a Tucson format file.
## There does not appear to be a header in the rwl file
## There are 13 series
## 1 PNB002NE 1900 2013 0.01
## 2 PNB002SW 1900 1980 0.01
## 3 PNB002W 1900 2013 0.01
## 4 PNB002E 1900 2013 0.01
## 5 PNB003NW 1900 2013 0.01
## 6 PNB004N 1900 2012 0.01
## 7 PNB004W 1900 2013 0.01
## 8 PNB005W 1900 2013 0.01
## 9 PNB006N 1900 2013 0.01
## 10 PNB006W 1917 2013 0.01
## 11 PNB011N 1905 2013 0.01
## 12 PNB011W 1900 2013 0.01
## 13 PNB012W 1900 2013 0.01

```

```
PNSWI<-read.rwl("PN-swit187.rwl")
```

```

## Attempting to automatically detect format.
## Assuming a Tucson format file.
## There appears to be a header in the rwl file
## There are 17 series
## 1 GES01GL 1897 2007 0.01
## 2 GES02GL2 1897 2007 0.01
## 3 GES05GL 1907 2007 0.01
## 4 GES06GL 1897 2007 0.01
## 5 GES07GL 1903 2007 0.01
## 6 GES08GL 1923 2007 0.01

```

```

## 7      GES09GL      1923      2007      0.01
## 8      GES10GR      1933      2007      0.01
## 9      GES11GR      1903      2007      0.01
## 10     GES12GL      1928      2007      0.01
## 11     GES13GL      1925      2007      0.01
## 12     GES14GL      1893      2007      0.01
## 13     GES15GR      1938      2007      0.01
## 14     GES16GR      1900      2007      0.01
## 15     GES17GL      1911      2007      0.01
## 16     GES18GL      1897      2007      0.01
## 17     GES19GL      1900      2007      0.01

```

### # *Ppinea*

```
PPIALB<-read.rwl("PPI-alb005.rwl")
```

```

## Attempting to automatically detect format.
## Assuming a Tucson format file.
## There appears to be a header in the rwl file
## There are 10 series
## 1      PPDIV04A      1882      2010      0.01
## 2      PPDIV04B      1891      2010      0.01
## 3      PPDIV11A      1890      2009      0.01
## 4      PPDIV11B      1927      2009      0.01
## 5      PPDIV02B      1920      2010      0.01
## 6      PPDIV05A      1950      2010      0.01
## 7      PPDIV10B      1906      2010      0.01
## 8      PPDIV10A      1950      2010      0.01
## 9      PPDIV03A      1862      2010      0.01
## 10     PPDIV03B      1909      2010      0.01

```

```
PPIITA<-read.rwl("PPI-ital035.rwl")
```

```

## Attempting to automatically detect format.
## Assuming a Tucson format file.
## There appears to be a header in the rwl file
## There are 95 series
## 1      110a          1916      2002      0.01
## 2      110b          1913      2002      0.01
## 3      111a          1912      2000      0.01
## 4      112a          1912      2002      0.01
## 5      112b          1917      2002      0.01
## 6      114a          1911      2002      0.01
## 7      114b          1914      2002      0.01
## 8      115a          1914      2002      0.01
## 9      115b          1910      2002      0.01
## 10     116a          1912      2002      0.01
## 11     116b          1925      2002      0.01
## 12     117a          1914      2002      0.01
## 13     117b          1912      2002      0.01
## 14     118a          1912      2002      0.01
## 15     118b          1921      2001      0.01
## 16     119a          1910      2002      0.01
## 17     119b          1916      2001      0.01
## 18     11xa          1911      2002      0.01
## 19     11xb          1912      2002      0.01
## 20     11ya          1914      2002      0.01
## 21     11yb          1914      2002      0.01
## 22     120a          1921      2001      0.01
## 23     120b          1919      1977      0.01
## 24     121a          1916      2002      0.01
## 25     121b          1919      2002      0.01
## 26     122a          1918      2002      0.01
## 27     122b          1921      2002      0.01
## 28     123a          1917      2002      0.01
## 29     123b          1911      2002      0.01

```

## 30	124a	1923	1990	0.01
## 31	124b	1924	2002	0.01
## 32	125a	1910	2002	0.01
## 33	125b	1924	2002	0.01
## 34	126a	1928	2002	0.01
## 35	126b	1921	2002	0.01
## 36	127a	1913	2002	0.01
## 37	127b	1933	2002	0.01
## 38	128b	1921	2002	0.01
## 39	129a	1913	2002	0.01
## 40	12xb	1921	2002	0.01
## 41	12ya	1921	2002	0.01
## 42	12yb	1925	2002	0.01
## 43	130a	1949	2002	0.01
## 44	130b	1949	2002	0.01
## 45	131a	1959	2002	0.01
## 46	131b	1948	2002	0.01
## 47	132a	1947	2002	0.01
## 48	132b	1951	2002	0.01
## 49	133a	1955	2002	0.01
## 50	134a	1950	2002	0.01
## 51	135a	1947	2002	0.01
## 52	135b	1949	2002	0.01
## 53	136a	1949	2002	0.01
## 54	137a	1947	2002	0.01
## 55	137b	1949	2002	0.01
## 56	138a	1945	2002	0.01
## 57	138b	1949	2002	0.01
## 58	139a	1950	2002	0.01
## 59	139b	1947	2002	0.01
## 60	13ya	1950	2002	0.01
## 61	13yb	1946	2002	0.01
## 62	150a	1926	2002	0.01
## 63	150b	1932	2002	0.01
## 64	151a	1921	2002	0.01
## 65	151b	1925	2002	0.01
## 66	152a	1922	2002	0.01
## 67	152b	1924	2002	0.01
## 68	153a	1925	2002	0.01
## 69	154a	1924	2001	0.01
## 70	154b	1925	1997	0.01
## 71	155a	1923	2002	0.01
## 72	156a	1928	2002	0.01
## 73	156b	1946	2002	0.01
## 74	158a	1923	1995	0.01
## 75	158b	1922	2001	0.01
## 76	159a	1931	2002	0.01
## 77	159b	1933	2002	0.01
## 78	15ya	1924	2002	0.01
## 79	15yb	1922	2002	0.01
## 80	172a	1923	2002	0.01
## 81	172b	1943	2002	0.01
## 82	173a	1935	2002	0.01
## 83	173b	1927	2002	0.01
## 84	174a	1921	2002	0.01
## 85	175a	1921	2002	0.01
## 86	175b	1923	1997	0.01
## 87	176a	1923	1998	0.01
## 88	177b	1924	2002	0.01
## 89	178a	1920	2002	0.01
## 90	178b	1927	2002	0.01
## 91	179b	1920	2002	0.01
## 92	17xa	1923	2002	0.01
## 93	17xb	1927	2002	0.01
## 94	17ya	1921	2002	0.01
## 95	17yb	1920	2002	0.01

```
PPISP57<- read.rwl("PPI-spai057.rwl")
```

```
## Attempting to automatically detect format.  
## Assuming a Tucson format file.  
## There appears to be a header in the rwl file  
## There are 22 series  
## 1      BUD02      1928      2000      0.01  
## 2      BUD04      1931      1999      0.01  
## 3      BUD05      1934      2000      0.01  
## 4      BUD06      1874      1999      0.01  
## 5      BUD07      1918      2000      0.01  
## 6      BUD07B     1929      2000      0.01  
## 7      BUD40A     1934      2001      0.01  
## 8      BUD41B     1934      2001      0.01  
## 9      BUD42A     1911      2001      0.01  
## 10     BUD42B     1909      2001      0.01  
## 11     BUD43A     1937      2001      0.01  
## 12     BUD43B     1931      2001      0.01  
## 13     BUD44A     1889      2001      0.01  
## 14     BUD44B     1899      2001      0.01  
## 15     BUD45A     1892      2001      0.01  
## 16     BUD46A     1891      2001      0.01  
## 17     BUD46B     1891      2001      0.01  
## 18     BUD47A     1907      2001      0.01  
## 19     BUD48A     1926      2001      0.01  
## 20     BUD48B     1917      1997      0.01  
## 21     BUD50A     1900      2001      0.01  
## 22     BUD53M     1905      2001      0.01
```

```
PPISP59<- read.rwl("PPI-spai059.rwl")
```

```
## Attempting to automatically detect format.  
## Assuming a Tucson format file.  
## There appears to be a header in the rwl file  
## There are 31 series  
## 1      CDV02A     1960      2001      0.01  
## 2      CDV02B     1956      2001      0.01  
## 3      CDV03A     1929      2001      0.01  
## 4      CDV03B     1933      2001      0.01  
## 5      CDV04A     1953      2001      0.01  
## 6      CDV05A     1959      2001      0.01  
## 7      CDV05B     1963      2001      0.01  
## 8      CDV06A     1980      2001      0.01  
## 9      CDV06B     1980      1995      0.01  
## 10     CDV08B     1930      2001      0.01  
## 11     CDV09B     1937      2001      0.01  
## 12     CDV10A     1923      2001      0.01  
## 13     CDV10B     1929      2001      0.01  
## 14     CDV11A     1950      2001      0.01  
## 15     CDV11B     1948      2001      0.01  
## 16     CDV12AM    1919      2000      0.01  
## 17     CDV12B     1918      2001      0.01  
## 18     CDV14A     1939      2001      0.01  
## 19     CDV15A     1917      2001      0.01  
## 20     CDV15B     1919      2001      0.01  
## 21     CDV16A     1926      2001      0.01  
## 22     CDV16BM    1926      2001      0.01  
## 23     CDV17A     1954      2001      0.01  
## 24     CDV18A     1953      2001      0.01  
## 25     CDV19      1949      2001      0.01  
## 26     CDV20A     1952      2001      0.01  
## 27     CDV20B     1952      2001      0.01  
## 28     CDV21      1913      2001      0.01  
## 29     CDV22AM    1907      2001      0.01
```



```
## 30      CDV23      1907      2001      0.01
## 31      CDV24AM    1923      2001      0.01
```

```
PPISPREM<-read.rwl("PPI-Spain_RemEMG.rwl")
```

```
## Attempting to automatically detect format.
## Assuming a Tucson format file.
## There does not appear to be a header in the rwl file
## There are 32 series
## 1      ERPP01N      1931      2016      0.01
## 2      PPER01S      1924      2014      0.01
## 3      ERPP02N      1935      2016      0.01
## 4      ERPP02S      1935      2003      0.01
## 5      PPER02SW     1936      2014      0.01
## 6      ERPP03NW     1931      2016      0.01
## 7      PPER03NW     1924      2014      0.01
## 8      ERPP03SE     1931      2016      0.01
## 9      ERPP04NW     1925      2016      0.01
## 10     PPER04SE     1925      2014      0.01
## 11     ERPP05N      1936      2016      0.01
## 12     PPER05SW     1936      2014      0.01
## 13     ERPP06N      1931      2016      0.01
## 14     PPER06N      1935      2014      0.01
## 15     PPER06S      1935      2014      0.01
## 16     ERPP07S      1922      2016      0.01
## 17     PPER07SE     1926      2014      0.01
## 18     PPER08NW     1921      2014      0.01
## 19     ERPP09S      1924      2016      0.01
## 20     PPER09N      1923      2014      0.01
## 21     PPER10N      1929      2014      0.01
## 22     ERPP10S      1926      2016      0.01
## 23     ERPP11N      1931      2016      0.01
## 24     PPER11SE     1935      2014      0.01
## 25     ERPP12N      1926      2016      0.01
## 26     ERPP12S      1917      2016      0.01
## 27     PPER12N      1962      2014      0.01
## 28     PPER12S      1950      2014      0.01
## 29     ERPP13S      1925      2016      0.01
## 30     PPER13N      1917      2014      0.01
## 31     ERPP14SE     1930      2016      0.01
## 32     PPER14W      1933      2014      0.01
```

```
#Psylvestris
```

```
PSFIN<-read.rwl("PS-finl072.rwl")
```

```
## Attempting to automatically detect format.
## Assuming a Tucson format file.
## There does not appear to be a header in the rwl file
## There are 82 series
## 1      120902a      1816      2006      0.001
## 2      120902b      1816      2006      0.001
## 3      120903b      1818      1989      0.001
## 4      120904a      1837      2005      0.001
## 5      120904b      1837      2006      0.001
## 6      120905a      1803      1985      0.001
## 7      120905b      1801      1958      0.001
## 8      120906a      1813      2006      0.001
## 9      120906b      1813      2006      0.001
## 10     120907a      1821      2006      0.001
## 11     120907b      1760      2004      0.001
## 12     120907c      1820      2006      0.001
## 13     120908b      1825      2005      0.001
## 14     120911b      1724      1867      0.001
## 15     120912a      1807      2006      0.001
## 16     120912b      1809      2001      0.001
```

## 17	120913a	1758	1954	0.001
## 18	120913b	1762	1950	0.001
## 19	120914a	1798	2006	0.001
## 20	120914b	1787	2006	0.001
## 21	120915a	1814	2006	0.001
## 22	120915b	1830	2006	0.001
## 23	120916a	1909	2006	0.001
## 24	120916b	1909	2006	0.001
## 25	120917as	1806	1911	0.001
## 26	120917bs	1804	1918	0.001
## 27	120918a	1764	2006	0.001
## 28	120918b	1773	2006	0.001
## 29	120919a	1804	1998	0.001
## 30	120919b	1802	2006	0.001
## 31	120920a	1818	2006	0.001
## 32	120920b	1817	2006	0.001
## 33	120922a	1758	2006	0.001
## 34	120922b	1774	2006	0.001
## 35	120923a	1831	2006	0.001
## 36	120923b	1830	2006	0.001
## 37	120924a	1763	2006	0.001
## 38	120924b	1760	2006	0.001
## 39	120925a	1880	2006	0.001
## 40	120925b	1882	2006	0.001
## 41	120926a	1744	2006	0.001
## 42	120926b	1746	2006	0.001
## 43	120927a	1833	2006	0.001
## 44	120927b	1862	2006	0.001
## 45	120927zs	1814	1875	0.001
## 46	120931a	1781	2006	0.001
## 47	120931b	1774	2006	0.001
## 48	120932a	1761	1883	0.001
## 49	120932b	1759	1920	0.001
## 50	120933b	1778	1900	0.001
## 51	120935a	1847	1997	0.001
## 52	120935b	1842	1950	0.001
## 53	120936a	1838	2004	0.001
## 54	120936b	1837	1963	0.001
## 55	120937a	1964	2006	0.001
## 56	120937b	1965	2003	0.001
## 57	120938b	1943	2006	0.001
## 58	120939a	1976	2006	0.001
## 59	120939b	1978	2006	0.001
## 60	120940a	1799	1971	0.001
## 61	120940b	1798	1946	0.001
## 62	120941a	1858	2006	0.001
## 63	120941b	1911	2006	0.001
## 64	120942a	1834	2006	0.001
## 65	120942b	1834	2003	0.001
## 66	120943a	1755	2006	0.001
## 67	120943b	1761	2006	0.001
## 68	120944a	1790	1952	0.001
## 69	1209101a	1877	1980	0.001
## 70	1209101b	1877	2006	0.001
## 71	1209102a	1848	2006	0.001
## 72	1209102b	1849	2006	0.001
## 73	1209112a	1763	2006	0.001
## 74	1209112b	1764	1973	0.001
## 75	1209113a	1818	2006	0.001
## 76	1209113b	1816	2006	0.001
## 77	1209114a	1839	2006	0.001
## 78	1209114b	1839	2005	0.001
## 79	1209219a	1965	2005	0.001
## 80	1209219b	1962	2006	0.001
## 81	1209320a	1753	1960	0.001
## 82	1209320b	1748	1963	0.001

```
PSGERM<-read.rwl("PS-germ188.rwl")
```

```
## Attempting to automatically detect format.  
## Assuming a Tucson format file.  
## There appears to be a header in the rwl file  
## There are 13 series  
## 1      dre2913m      1861      2011      0.01  
## 2      dre2901m      1884      2011      0.01  
## 3      dre2902m      1856      2011      0.01  
## 4      dre2903m      1859      2011      0.01  
## 5      dre2904m      1854      2011      0.01  
## 6      dre2905m      1857      2011      0.01  
## 7      dre2906m      1856      1990      0.01  
## 8      dre2907m      1856      2011      0.01  
## 9      dre2908m      1861      2011      0.01  
## 10     dre2909m      1865      2011      0.01  
## 11     dre2910m      1856      2011      0.01  
## 12     dre2911m      1856      2011      0.01  
## 13     dre2912m      1860      2011      0.01
```

```
PSRUSS<-read.rwl("PS-russ276.rwl")
```

```
## Attempting to automatically detect format.  
## Assuming a Tucson format file.  
## There does not appear to be a header in the rwl file  
## There are 36 series  
## 1      ksl10a      1920      2009      0.001  
## 2      ksl11a      1890      2009      0.001  
## 3      ksl12a      1897      2009      0.001  
## 4      ksl13a      1910      2009      0.001  
## 5      ksl14a      1906      2009      0.001  
## 6      ksl16a      1898      2009      0.001  
## 7      ksl18a      1896      2009      0.001  
## 8      ksl1a      1895      2009      0.001  
## 9      ksl20a      1905      2009      0.001  
## 10     ksl21a      1931      2009      0.001  
## 11     ksl22a      1906      2009      0.001  
## 12     ksl23a      1933      2009      0.001  
## 13     ksl3a      1902      2009      0.001  
## 14     ksl4a      1920      2009      0.001  
## 15     ksl5a      1907      2009      0.001  
## 16     ksl6a      1910      2009      0.001  
## 17     ksl7a      1934      2009      0.001  
## 18     ksl8a      1933      2009      0.001  
## 19     ksl9a      1932      2009      0.001  
## 20     kyk10a      1878      2009      0.001  
## 21     kyk11a      1871      2009      0.001  
## 22     kyk12a      1883      2009      0.001  
## 23     kyk13a      1874      2009      0.001  
## 24     kyk14a      1865      2009      0.001  
## 25     kyk15a      1940      2009      0.001  
## 26     kyk16a      1872      2009      0.001  
## 27     kyk17a      1857      2009      0.001  
## 28     kyk18a      1862      2009      0.001  
## 29     kyk19a      1829      2009      0.001  
## 30     kyk20a      1881      2009      0.001  
## 31     kyk3a      1890      2009      0.001  
## 32     kyk4a      1888      2009      0.001  
## 33     kyk5a      1891      2009      0.001  
## 34     kyk6a      1887      2009      0.001  
## 35     kyk7a      1883      2009      0.001  
## 36     kyk9a      1888      2009      0.001
```

```
PSRUSE<-read.rwl("PS-russEMG.RWL")
```

```

## Attempting to automatically detect format.
## Assuming a Tucson format file.
## There does not appear to be a header in the rwl file
## There are 73 series
## 1      RKHPS011    1927    2012    0.01
## 2      RKHPS012    1927    2012    0.01
## 3      RKHPS015    1918    2012    0.01
## 4      RKHPS016    1918    2012    0.01
## 5      RKHPS021    1925    2012    0.01
## 6      RKHPS022    1925    2012    0.01
## 7      RKHPS023    1912    2012    0.01
## 8      RKHPS024    1912    2012    0.01
## 9      RKHPS031    1924    2012    0.01
## 10     RKHPS032    1924    2012    0.01
## 11     RKHPS033    1920    2012    0.01
## 12     RKHPS034    1920    2012    0.01
## 13     RKHPS041    1904    2012    0.01
## 14     RKHPS042    1904    2012    0.01
## 15     RKHPS044    1908    2012    0.01
## 16     RKHPS051    1934    2012    0.01
## 17     RKHPS052    1934    2012    0.01
## 18     RKHPS053    1930    2012    0.01
## 19     RKHPS054    1930    2012    0.01
## 20     RKHPS061    1833    2012    0.01
## 21     RKHPS062    1833    2012    0.01
## 22     RKHPS063    1832    2012    0.01
## 23     RKHPS064    1832    2012    0.01
## 24     RKHPS071    1905    2012    0.01
## 25     RKHPS072    1909    2012    0.01
## 26     RKHPS073    1923    2012    0.01
## 27     RKHPS074    1918    2012    0.01
## 28     RKHPS081    1947    2012    0.01
## 29     RKHPS082    1947    2012    0.01
## 30     RKHPS083    1941    2012    0.01
## 31     RKHPS084    1941    2012    0.01
## 32     RKHPS091    1908    2012    0.01
## 33     RKHPS092    1908    2012    0.01
## 34     RKHPS093    1913    2012    0.01
## 35     RKHPS101    1917    2012    0.01
## 36     RKHPS102    1914    2012    0.01
## 37     RKHPS103    1927    2012    0.01
## 38     RKHPS104    1927    2012    0.01
## 39     RKRPS011    1919    2012    0.01
## 40     RKRPS012    1919    2012    0.01
## 41     RKRPS015    1924    2012    0.01
## 42     RKRPS016    1924    2012    0.01
## 43     RKRPS021    1917    2012    0.01
## 44     RKRPS022    1917    2012    0.01
## 45     RKRPS023    1927    2012    0.01
## 46     RKRPS024    1927    2012    0.01
## 47     RKRPS031    1917    2012    0.01
## 48     RKRPS032    1917    2012    0.01
## 49     RKRPS033    1935    2012    0.01
## 50     RKRPS034    1935    2012    0.01
## 51     RKRPS041    1936    2012    0.01
## 52     RKRPS042    1936    2012    0.01
## 53     RKRPS043    1918    2012    0.01
## 54     RKRPS044    1918    2012    0.01
## 55     RKRPS051    1914    2012    0.01
## 56     RKRPS052    1914    2012    0.01
## 57     RKRPS053    1912    2012    0.01
## 58     RKRPS054    1912    2012    0.01
## 59     RKRPS061    1925    2012    0.01
## 60     RKRPS062    1925    2012    0.01
## 61     RKRPS063    1915    2012    0.01
## 62     RKRPS064    1915    2012    0.01

```

```

## 63      RKRPS071      1917      2012      0.01
## 64      RKRPS072      1917      2012      0.01
## 65      RKRPS073      1914      2012      0.01
## 66      RKRPS074      1914      2012      0.01
## 67      RKRPS081      1933      2012      0.01
## 68      RKRPS082      1933      2012      0.01
## 69      RKRPS083      1923      2012      0.01
## 70      RKRPS093      1944      2012      0.01
## 71      RKRPS094      1924      2012      0.01
## 72      RKRPS101      1914      2012      0.01
## 73      RKRPS102      1914      2012      0.01

```

```
PSSPA83<- read.rwl("PS-spai083.rwl")
```

```

## Attempting to automatically detect format.
## Assuming a Tucson format file.
## There does not appear to be a header in the rwl file
## There are 32 series
## 1      Es2Ps02a      1594      2002      0.01
## 2      Es2Ps02b      1676      2002      0.01
## 3      Es2Ps10a      1671      2002      0.01
## 4      Es2Ps10b      1679      2002      0.01
## 5      Es2Ps12a      1725      2002      0.01
## 6      Es2Ps12b      1738      2002      0.01
## 7      Es2Ps15a      1680      2002      0.01
## 8      Es2Ps15b      1690      2002      0.01
## 9      Es2Ps16a      1646      1897      0.01
## 10     Es2Ps16b      1633      1907      0.01
## 11     Es2Ps17a      1577      1884      0.01
## 12     Es2Ps17b      1615      1891      0.01
## 13     Es2Ps18a      1650      2002      0.01
## 14     Es2Ps18b      1640      2002      0.01
## 15     Es2Ps19a      1708      2002      0.01
## 16     Es2Ps19b      1749      2002      0.01
## 17     Es2Ps20a      1667      2002      0.01
## 18     Es2Ps20b      1686      2002      0.01
## 19     Es2Ps22a      1640      2002      0.01
## 20     Es2Ps22b      1638      2002      0.01
## 21     Es2Ps23a      1613      2002      0.01
## 22     Es2Ps23b      1589      2002      0.01
## 23     Es2Ps23c      1598      2002      0.01
## 24     Es2Ps24a      1545      2002      0.01
## 25     Es2Ps24b      1511      2002      0.01
## 26     Es2Ps24c      1511      2002      0.01
## 27     Es2Ps25a      1652      2002      0.01
## 28     Es2Ps25b      1669      2002      0.01
## 29     Es2Ps26a      1625      2002      0.01
## 30     Es2Ps26b      1634      2002      0.01
## 31     Es2Ps27a      1590      2002      0.01
## 32     Es2Ps27b      1590      2002      0.01

```

```
PSSWE<- read.rwl("PS-swed313.rwl")
```

```

## Attempting to automatically detect format.
## Assuming a Tucson format file.
## There does not appear to be a header in the rwl file
## There are 29 series
## 1      ORT11a      1869      2003      0.01
## 2      ORT11b      1858      2003      0.01
## 3      ORT12a      1860      2003      0.01
## 4      ORT12b      1865      2003      0.01
## 5      ORT13a      1878      2003      0.01
## 6      ORT13b      1876      2003      0.01
## 7      ORT14a      1862      2003      0.01
## 8      ORT14b      1866      2003      0.01
## 9      ORT15a      1859      2003      0.01

```

```

## 10   ORT15b   1860   2003   0.01
## 11   ORT16b   1864   2003   0.01
## 12   ORT17a   1865   2003   0.01
## 13   ORT17b   1861   2003   0.01
## 14   ORT18a   1866   2003   0.01
## 15   ORT18b   1864   2003   0.01
## 16   ORT19a   1872   2006   0.01
## 17   ORT19b   1871   1930   0.01
## 18   ORT20a   1862   2006   0.01
## 19   ORT20b   1858   1910   0.01
## 20   ORT21a   1859   2006   0.01
## 21   ORT21b   1860   2005   0.01
## 22   ORT22a   1866   2005   0.01
## 23   ORT22b   1865   2006   0.01
## 24   ORT23a   1862   1969   0.01
## 25   ORT23b   1862   1980   0.01
## 26   ORT24a   1868   1960   0.01
## 27   ORT24b   1868   2006   0.01
## 28   ORT25a   1863   2006   0.01
## 29   ORT25b   1895   2006   0.01

```

```
PSSWI<- read.rwl("PS-swit269.rwl")
```

```

## Attempting to automatically detect format.
## Assuming a Tucson format file.
## There appears to be a header in the rwl file
## There are 19 series

```

```

## 1     FB.W01     1891     2009     0.01
## 2     FB.W05     1842     2009     0.01
## 3     FB.W06B    1842     2009     0.01
## 4     FB.W07     1841     2009     0.01
## 5     FB.W08     1840     2009     0.01
## 6     FB.W09     1845     2009     0.01
## 7     FB.W10     1834     2009     0.01
## 8     FB.W11     1847     2009     0.01
## 9     FB.W12     1855     2009     0.01
## 10    FB.W13     1870     2009     0.01
## 11    FB.W14     1847     2009     0.01
## 12    FB.W15     1848     2009     0.01
## 13    FB.W17     1839     2009     0.01
## 14    FB.W21     1876     2009     0.01
## 15    FB.W22     1863     2009     0.01
## 16    FB.W23     1851     2009     0.01
## 17    FB.W24     1856     2009     0.01
## 18    FB.W25     1816     2009     0.01
## 19    FB.W26     1855     2009     0.01

```

```
# EJEMPLO DE COFECHA con la cronología de Pinus halepensis
corr.rwl.seg(PHVALLE)
```

```

## $avg.seg.rho
## 1900.1949 1925.1974 1950.1999
##          NaN 0.4292917 0.6385504

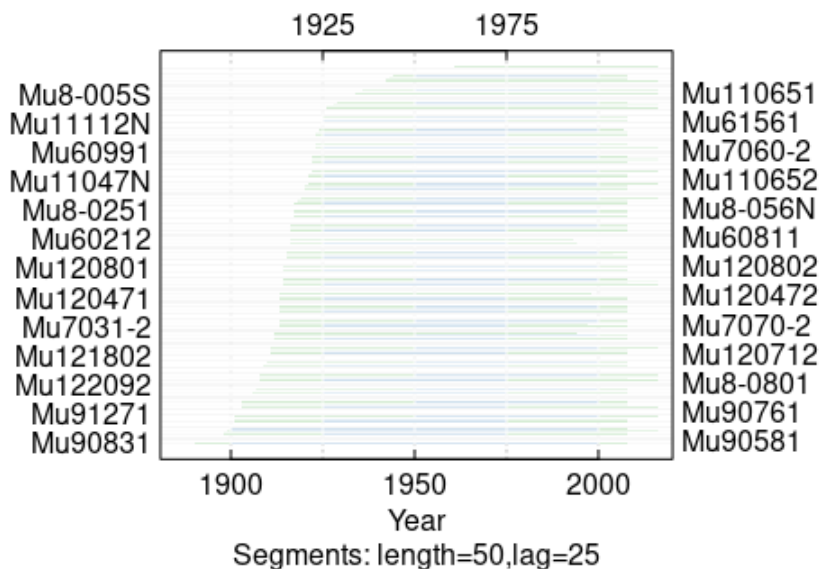
```

```

##
## $flags
## named character(0)
##

```

```
corr.rwl.seg(PHSE)
```



```
## $avg.seg.rho
## 1900.1949 1925.1974 1950.1999
## 0.7608932 0.7768951 0.7853418
##
## $flags
## named character(0)
##
# Pnigra
corr.rwl.seg(PNGRE)
corr.rwl.seg(PNITA) # flags en 12 árboles
corr.rwl.seg(PNSPBOU)
corr.rwl.seg(PNSWI)
# Ppinea
corr.rwl.seg(PPIALB)
corr.rwl.seg(PPIITA, seg.length = 40, bin.floor = 20)
corr.rwl.seg(PPISP57)
corr.rwl.seg(PPISP59, seg.length = 40, bin.floor = 20)
# Psylvestris
corr.rwl.seg(PSFIN)
corr.rwl.seg(PSGERM)
corr.rwl.seg(PSRUSS)
corr.rwl.seg(PSRUSE)
corr.rwl.seg(PSSPA83)
corr.rwl.seg(PSSWE)
corr.rwl.seg(PSSWI)

# CRONOLOGÍA MAESTRA & ESTADÍSTICOS
PHALBA.crn <- chron((PHALBA))
PHFRAN.crn <- chron((PHFRAN))
PHVALLE.crn <- chron((PHVALLE))
PHSE.crn <- chron((PHSE))
PHGA.crn <- chron((PHGA))
min_year<-min(c(rownames(PHALBA.crn),
                rownames(PHFRAN.crn),
                rownames(PHVALLE.crn), rownames(PHSE.crn), rownames(PHGA.crn)))
max_year<-max(c(rownames(PHALBA.crn),
                rownames(PHFRAN.crn),
                rownames(PHVALLE.crn), rownames(PHSE.crn), rownames(PHGA.crn)))
PHVALLE.crn2 <- (PHVALLE)
PHSE.crn2 <- (PHSE)

gp.rwl.ids<-read.ids(PHSE.crn2, stc = c(2, 5, 1))
```

```

gp.rwi.stats.running<-rwi.stats.running(PHSE.crn2,ids = gp.rwl.ids)
gp.rwi.stats.running

PNAUS.crn <- chron((PNAUS))
PNGRE.crn <- chron((PNGRE))
PNITA.crn <- chron((PNITA))
PNSP17.crn <- chron((PNSP17))
PNSPBOU.crn <- chron((PNSPBOU))
PNSWI.crn <- chron((PNSWI))
min_year<-min(c(min_year, rownames(PNAUS.crn), rownames(PNGRE.crn),
rownames(PNITA.crn), rownames(PNSP17.crn), rownames(PNSPBOU.crn),
rownames(PNSWI.crn)))
max_year<-max(c(max_year, rownames(PNAUS.crn), rownames(PNGRE.crn),
rownames(PNITA.crn), rownames(PNSP17.crn), rownames(PNSPBOU.crn),
rownames(PNSWI.crn)))
PNGRE.crn2 <- (PNGRE)
PNITA.crn2 <- (PNITA)
PNSPBOU.crn2 <- (PNSPBOU)
PNSWI.crn2 <- (PNSWI)

gp.rwl.ids<-read.ids(PNSPBOU.crn2,stc = c(3, 3, 1))
gp.rwi.stats.running<-rwi.stats.running(PNSPBOU.crn2,ids = gp.rwl.ids)
gp.rwi.stats.running

PPIALB.crn <- chron((PPIALB))
PPIITA.crn <- chron((PPIITA))
PPISP57.crn <- chron((PPISP57))
PPISP59.crn <- chron((PPISP59))
PPISPREM.crn <- chron((PPISPREM))
min_year<-min(c(min_year, rownames(PPIALB.crn), rownames(PPIITA.crn),
rownames(PPISP57.crn), rownames(PPISP59.crn), rownames(PPISPREM.crn)))
max_year<-max(c(max_year, rownames(PPIALB.crn), rownames(PPIITA.crn),
rownames(PPISP57.crn), rownames(PPISP59.crn), rownames(PPISPREM.crn)))
PPIALB.crn2 <- (PPIALB)
PPIITA.crn2 <- (PPIITA)
PPISP57.crn2 <- (PPISP57)
PPISP59.crn2 <- (PPISP59)
PPISPREM.crn2 <- (PPISPREM)

gp.rwl.ids<-read.ids(PPISP59.crn2,stc = c(3, 2, 1))
gp.rwi.stats.running<-rwi.stats.running(PPISP59.crn2,ids = gp.rwl.ids)
gp.rwi.stats.running

PSFIN.crn<- chron((PSFIN))
PSGERM.crn<- chron((PSGERM))
PSRUSS.crn<- chron((PSRUSS))
PSRUSE.crn<- chron((PSRUSE))
PSSPA83.crn<- chron((PSSPA83))
PSSWE.crn<- chron((PSSWE))
PSSWI.crn<- chron((PSSWI))
min_year<-min(c(min_year, rownames(PSFIN.crn), rownames(PSGERM.crn),
rownames(PSRUSS.crn), rownames(PSRUSE.crn), rownames(PSSPA83.crn),
rownames(PSSWE.crn), rownames(PSSWI.crn)))
max_year<-max(c(max_year, rownames(PSFIN.crn), rownames(PSGERM.crn),
rownames(PSRUSS.crn), rownames(PSRUSE.crn), rownames(PSSPA83.crn),
rownames(PSSWE.crn), rownames(PSSWI.crn)))
PSFIN.crn2<- (PSFIN)
PSGERM.crn2<- (PSGERM)
PSRUSS.crn2<- (PSRUSS)

```



```

PSRUSE.crn2<- (PSRUSE)
PSSPA83.crn2<- (PSSPA83)
PSSWE.crn2<- (PSSWE)
PSSWI.crn2<- (PSSWI)

gp.rwl.ids<-read.ids(PSSWI.crn2,stc = c(4, 2, 1))
gp.rwi.stats.running<-rwi.stats.running(PSSWI.crn2,ids = gp.rwl.ids)
gp.rwi.stats.running

```

```

#DATA FRAME DE CADA ESPECIE
PHALBA.crn[, "code"]<- "PHALBA"
PHFRAN.crn[, "code"]<- "PHFRAN"
PHVALLE.crn[, "code"]<- "PHVALLE"
PHSE.crn[, "code"]<- "PHSE"
PHGA.crn[, "code"]<- "PHGA"

PNAUS.crn[, "code"]<- "PNAUS"
PNGRE.crn[, "code"]<- "PNGRE"
PNITA.crn[, "code"]<- "PNITA"
PNSP17.crn[, "code"]<- "PNSP17"
PNSPBOU.crn[, "code"]<- "PNSPBOU"
PNSWI.crn[, "code"]<- "PNSWI"

PPIALB.crn[, "code"]<- "PPIALB"
PPIITA.crn[, "code"]<- "PPIITA"
PPISP57.crn[, "code"]<- "PPISP57"
PPISP59.crn[, "code"]<- "PPISP59"
PPISPREM.crn[, "code"]<- "PPISPREM"

PSFIN.crn[, "code"]<- "PSFIN"
PSGERM.crn[, "code"]<- "PSGERM"
PSRUSS.crn[, "code"]<- "PSRUSS"
PSRUSE.crn[, "code"]<- "PSRUSE"
PSSPA83.crn[, "code"]<- "PSSPA83"
PSSWE.crn[, "code"]<- "PSSWE"
PSSWI.crn[, "code"]<- "PSSWI"

```

*#Creación del data frame con el que vamos a trabajar*

```

DB<-data.frame(YEAR=seq(min_year,max_year))
rownames(DB)<-DB$YEAR

DB<-cbind(DB, PHALBA=PHALBA.crn[, "xxxstd"][match(rownames(DB),
rownames(PHALBA.crn))])
DB<-cbind(DB, PHFRAN=PHFRAN.crn[, "xxxstd"][match(rownames(DB),
rownames(PHFRAN.crn))])
DB<-cbind(DB, PHVALLE=PHVALLE.crn[, "xxxstd"][match(rownames(DB),
rownames(PHVALLE.crn))])
DB<-cbind(DB, PHSE=PHSE.crn[, "xxxstd"][match(rownames(DB),
rownames(PHSE.crn))])
DB<-cbind(DB, PHGA=PHGA.crn[, "xxxstd"][match(rownames(DB),
rownames(PHGA.crn))])

DB<-cbind(DB, PNAUS=PNAUS.crn[, "xxxstd"][match(rownames(DB),
rownames(PNAUS.crn))])
DB<-cbind(DB, PNGRE=PNGRE.crn[, "xxxstd"][match(rownames(DB),
rownames(PNGRE.crn))])
DB<-cbind(DB, PNITA=PNITA.crn[, "xxxstd"][match(rownames(DB),
rownames(PNITA.crn))])
DB<-cbind(DB, PNSP17=PNSP17.crn[, "xxxstd"][match(rownames(DB),
rownames(PNSP17.crn))])
DB<-cbind(DB, PNSWI=PNSWI.crn[, "xxxstd"][match(rownames(DB),
rownames(PNSWI.crn))])

```

```

DB<-cbind(DB, PPIITA=PPIITA.crn[, "xxxstd"][match(rownames(DB),
rownames(PPIITA.crn))])
DB<-cbind(DB, PPISP57=PPISP57.crn[, "xxxstd"][match(rownames(DB),
rownames(PPISP57.crn))])
DB<-cbind(DB, PPISP59=PPISP59.crn[, "xxxstd"][match(rownames(DB),
rownames(PPISP59.crn))])
DB<-cbind(DB, PPISPREM=PPISPREM.crn[, "xxxstd"][match(rownames(DB),
rownames(PPISPREM.crn))])

DB<-cbind(DB, PSFIN=PSFIN.crn[, "xxxstd"][match(rownames(DB),
rownames(PSFIN.crn))])
DB<-cbind(DB, PSGERM=PSGERM.crn[, "xxxstd"][match(rownames(DB),
rownames(PSGERM.crn))])
DB<-cbind(DB, PSRUSS=PSRUSS.crn[, "xxxstd"][match(rownames(DB),
rownames(PSRUSS.crn))])
DB<-cbind(DB, PSRUSE=PSRUSE.crn[, "xxxstd"][match(rownames(DB),
rownames(PSRUSE.crn))])
DB<-cbind(DB, PSSPA83=PSSPA83.crn[, "xxxstd"][match(rownames(DB),
rownames(PSSPA83.crn))])
DB<-cbind(DB, PSSWE=PSSWE.crn[, "xxxstd"][match(rownames(DB),
rownames(PSSWE.crn))])
DB<-cbind(DB, PSSWI=PSSWI.crn[, "xxxstd"][match(rownames(DB),
rownames(PSSWI.crn))])

DB_F<-DB[DB$YEAR>= 1910 & DB$YEAR<= 2001,]
DB_F$YEAR<-NULL
DB_F <- DB_F[ , colSums(is.na(DB_F)) == 0]

```

## Código R del análisis Clustering con los diferentes paquetes

```
## DATA-FRAME CON LOS SITIOS SELECCIONADOS (DB_F)
```

```
data<-DB_F # 15 columnas-variables y 92 filas-observaciones
```

```
d<-as.matrix(data)
```

### library(fda)

```
## Create a minimal Fourier basis for the analyse:
```

```
# nbasis = la mitad de los datos menos 1
```

```
basis<-create.fourier.basis(rangeval = c(0,92),nbasis =(92/2-1))
```

```
fdobj <-smooth.basis(argvals=(1:92), y=d, fdParobj=basis)$fd
```

```
fdobj
```

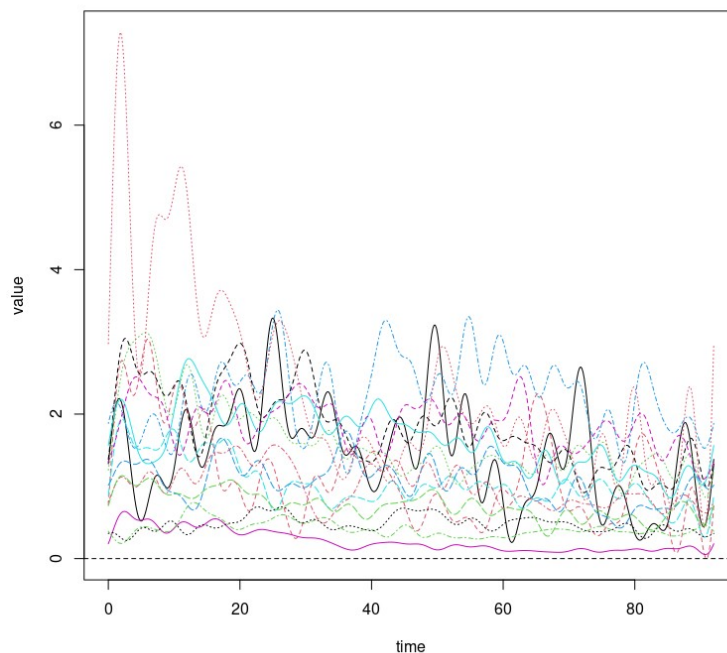


Figura 1A.- Datos funcionales creados con el paquete fda, función Fourier, a partir de nuestros datos.

### library(funHDDC)

```
# Selection Nº of clusters and model (6 different models depend on the number of the parameters)
```

```
# With the BIC criterion and the slope heuristic criterion
```

```
res.classif<-funHDDC(fdobj,K=2:6,c("AkjBkQkDk","AkjBQkDk","AkBkQkDk",  
"AkBQkDk","ABkQkDk","ABQkDk"),init="kmeans")
```

```
SELECTED: model ABKQKDK with 2 clusters.
```

```
Selection Criterion: BIC.
```

### library(funFEM)

```
# Choice of K with the other algorithm. 12 different models depend on the number of the parameters)
```

```
res = funFEM(fdobj,K=2:3,model='all',init='kmeans',lambda=0,disp=TRUE)
```

The best model is **AkBk** with **K = 3** ( $bic = -310.8632$ ) (TABLA 3)

```
slopeHeuristic(res.classif)
```

```
[1] 2
```

```
## Graphical representation with 3 clusters
```

```
res.uni<-funHDDC(fdobj,K=3,"[akjbkQkDk]",
```

```
+          init="random",threshold=0.2)
```

```
# random: is run 20 times with the same model options and the  
solution which maximises the log-likelihood is printed
```

```
funHDDC:
```

```
model K threshold complexity BIC
```

```
1 ABKQKDK 3 0.2 316 -35.18
```

```
SELECTED: model ABKQKDK with 3 clusters.
```

```
Selection Criterion: BIC.
```

```
> res.uni$class
```

```
[1] 1 1 3 1 3 2 3 1 2 2 3 3 2 2 2
```

```
## Graphical representation of groups means curves
```

```
par(mfrow=c(1,1))
```

```
plot(fdobj, col=res.uni$class, lwd=2, lty=2)
```

```
select1 <- fd(fdobj$coefs[,which(res.uni$class==1)], fdobj$basis)
```

```
select2 <- fd(fdobj$coefs[,which(res.uni$class==2)], fdobj$basis)
```

```
select3 <- fd(fdobj$coefs[,which(res.uni$class==3)], fdobj$basis)
```

```
plot(mean.fd(select1),col=1, ylim=c(0,6), lty=1, lwd=3) #negro=1
```

```
lines(mean.fd(select2),col=2, lty=1,lwd=3) #rojo=2
```

```
lines(mean.fd(select3),col=3, lty=1,lwd=3) #verde=3
```

```
# Statistics pca: 73% de varianza explica en el primer componente
```

```
par(mfrow=c(1,1))
```

```
res.pca<-mfpca(fdobj)
```

```
plot.mfpca(res.pca)
```

## library(lcmm)

```
dataok<-read.csv("dataok.csv",header = T)
```

```
str(dataok)
```

```
'data.frame': 1380 obs. of 3 variables:
```

```
$ TIME: int 1910 1911 1912 1913 1914 1915 1916 1917 1918 1919 ...
```

```
$ TRW : num 2 2.14 1.56 0.7 0.99 ...
```

```
$ ID : Factor w/ 15 levels "PHSE","PHVALLE",...: 2 2 2 2 2 2 2 2 2 ...
```

```
data<-transform(dataok, ID=as.numeric(ID))
```

```
str(data)
```

```
'data.frame': 1380 obs. of 3 variables:
```

```
$ TIME: int 1910 1911 1912 1913 1914 1915 1916 1917 1918 1919 ...
```

```
$ TRW : num 2 2.14 1.56 0.7 0.99 ...
```

```
$ ID : num 2 2 2 2 2 2 2 2 2 ...
```

```

# Estimation of latent class linear mixed models. K=2
m2<- hlme(fixed=TRW ~ poly(TIME, degree = 2, raw = TRUE),
          random=~TIME, mixture=~poly(TIME, degree = 2, raw = TRUE),
          subject="ID",ng=2, data=data,maxiter=1000)
Be patient, hlme is running ...
The program took 701.93 seconds
> summary(m2)
Heterogenous linear mixed model
  fitted by maximum likelihood method
Statistical Model:
  Dataset: data
  Number of subjects: 15
  Number of observations: 1380
  Number of latent classes: 2
  Number of parameters: 11
Iteration process:
  Maximum number of iteration reached without convergence
  Number of iterations: 1000
  Convergence criteria: parameters= 2.5e-13
                      : likelihood= 1.4e-05
                      : second derivatives= 1
Goodness-of-fit statistics:
  maximum log-likelihood: -912.69
  AIC: 1847.38
  BIC: 1855.17
Maximum Likelihood Estimates:
Fixed effects in the class-membership model:
(the class of reference is the last class)

      coef Se Wald p-value
Intercept class1 0.13721
Fixed effects in the longitudinal model:
      coef Se Wald p-value
intercept class1          0.63699
intercept class2          1.91920
poly(TIME, degree = 2, raw = TRUE)1 class1 0.00628
poly(TIME, degree = 2, raw = TRUE)1 class2 0.02040
poly(TIME, degree = 2, raw = TRUE)2 class1 0.00000
poly(TIME, degree = 2, raw = TRUE)2 class2 -0.00001
Variance-covariance matrix of the random-effects:
      intercept TIME
intercept 2.32674
TIME      -0.00103 0
                                coef Se
Residual standard error: 0.44990
> postprob(m2)
Posterior classification:
  class1 class2
N 11.00 4.00
% 73.33 26.67
Posterior classification table:
--> mean of posterior probabilities in each class
      prob1 prob2
class1 0.9960 0.0040
class2 0.0009 0.9991
Posterior probabilities above a threshold (%):
      class1 class2

```

```

prob>0.7  100  100
prob>0.8  100  100
prob>0.9  100  100

```

**#Goodness of fit 2: plot of residuals**

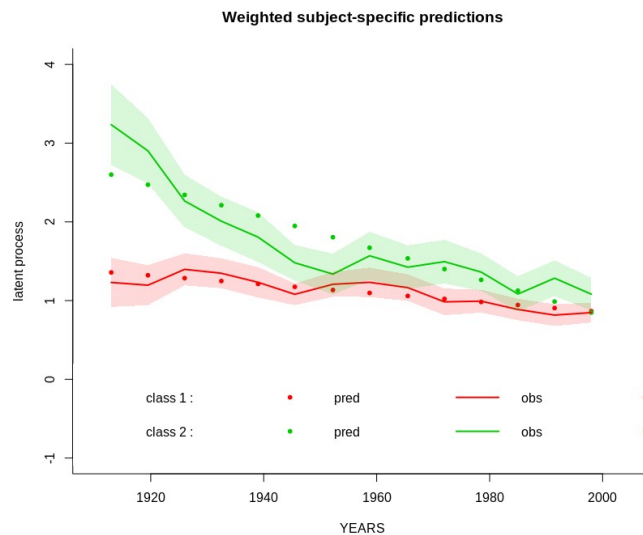
```
plot(m2, cex.main=0.9)
```

**#Goodness of fit 2: plot of predictions versus observations**

```

plot(m2, which="fit", var.time="TIME",
     bty="l", xlab="YEARS", break.times=15, ylab="latent process",
     lwd=2, marg=FALSE, ylim=c(-1,4),xlim=c(1910,2005), shades=TRUE,
     col=c(2,3))

```



**Figura 2A.-** Gráfico de valores predichos vs. observados para 2 clases latentes.

**# Estimation of latent class linear mixed models. K=3**

```

m3<- hlme(fixed=TRW ~ poly(TIME, degree = 2, raw = TRUE),
          random=~TIME,
          mixture=~poly(TIME, degree = 2, raw = TRUE),
          subject="ID",
          ng=3, data=data,
          maxiter=1000)

```

Be patient, hlme is running ...

**The program took 1191.3 seconds**

```
> summary(m3)
```

Heterogenous linear mixed model  
fitted by maximum likelihood method

Statistical Model:

```

Dataset: data
Number of subjects: 15
Number of observations: 1380
Number of latent classes: 3
Number of parameters: 15

```

Iteration process:

```

Maximum number of iteration reached without convergence
Number of iterations: 1000
Convergence criteria: parameters= 1.2e-11
                    : likelihood= 0.00048
                    : second derivatives= 1

```

Goodness-of-fit statistics:  
 maximum log-likelihood: -864.16  
 AIC: 1758.33  
 BIC: 1768.95

Maximum Likelihood Estimates:  
 Fixed effects in the class-membership model:  
 (the class of reference is the last class)  
 coef Se Wald p-value  
 intercept class1 0.07777  
 intercept class2 0.02502

Fixed effects in the longitudinal model:

	coef	Se	Wald	p-value
intercept class1				0.00045
intercept class2				1.27792
intercept class3				2.55592
poly(TIME, degree = 2, raw = TRUE)1 class1				0.00513
poly(TIME, degree = 2, raw = TRUE)1 class2				0.01324
poly(TIME, degree = 2, raw = TRUE)1 class3				0.03556
poly(TIME, degree = 2, raw = TRUE)2 class1				0.00000
poly(TIME, degree = 2, raw = TRUE)2 class2				-0.00001
poly(TIME, degree = 2, raw = TRUE)2 class3				-0.00002

Variance-covariance matrix of the random-effects:

	intercept	TIME
intercept	0.81174	
TIME	-0.00027	0

Residual standard error: 0.43580

**> postprob(m3)**

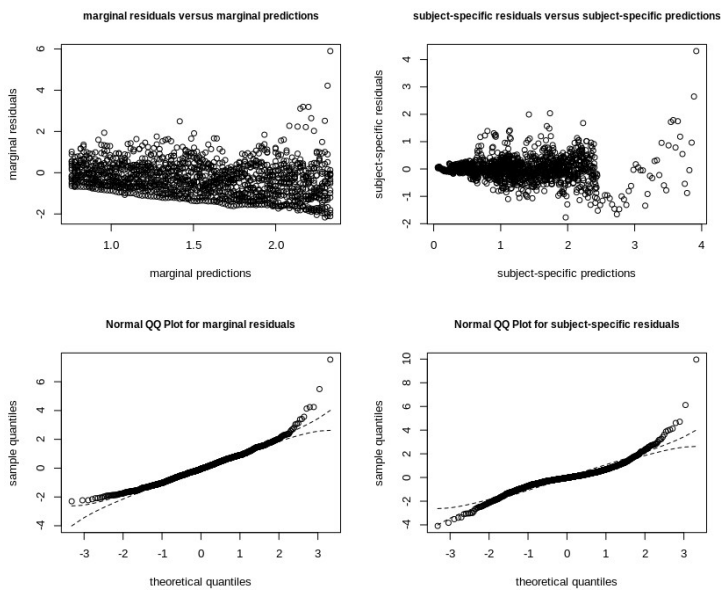
Posterior classification:  
 class1 class2 class3  
 N 8.00 6 1.00  
 % 53.33 40 6.67

Posterior classification table:  
 --> mean of posterior probabilities in each class  
 prob1 prob2 prob3  
 class1 0.9964 0.0036 0  
 class2 0.0170 0.9830 0  
 class3 0.0000 0.0000 1

Posterior probabilities above a threshold (%):  
 class1 class2 class3  
 prob>0.7 100 100 100  
 prob>0.8 100 100 100  
 prob>0.9 100 100 100

**# Goodness of fit 3: plot of residuals**

```
plot(m3, cex.main=0.9)
```



**# Goodness of fit 3: plot of predictions versus observations**

```
plot(m3, which="fit", var.time="TIME",
     bty="l", xlab="YEARS", break.times=15, ylab="latent process",
     lwd=2, marg=FALSE, ylim=c(-1,4),xlim=c(1910,2005), shades=TRUE,
     col=c(2,3,4))
```

**# Estimation of latent class linear mixed models. K=4**

```
m4<- hlme(fixed=TRW ~ poly(TIME, degree = 2, raw = TRUE),
          random=~TIME,
          mixture=~poly(TIME, degree = 2, raw = TRUE),
          subject="ID",
          ng=3, data=data,
          maxiter=1000)
```

Be patient, hlme is running ...

The program **took 1787.26 seconds**

```
> summary(m4)
```

Heterogenous linear mixed model

fitted by maximum likelihood method

Statistical Model:

Dataset: data

Number of subjects: 15

Number of observations: 1380

Number of latent classes: 4

Number of parameters: 19

Iteration process:

Maximum number of iteration reached without convergence

Number of iterations: 1000

Convergence criteria: parameters= 3.5e-11

: likelihood= 8e-04



: second derivatives= 1

Goodness-of-fit statistics:

maximum log-likelihood: -858.7

AIC: 1755.4

BIC: 1768.85

Maximum Likelihood Estimates:

Fixed effects in the class-membership model:  
(the class of reference is the last class)

	coef	Se	Wald	p-value
intercept class1	0.00662			
intercept class2	0.04266			
intercept class3	0.00706			

Fixed effects in the longitudinal model:

	coef	Se	Wald	p-value
intercept class1				-0.63915
intercept class2				0.63834
intercept class3				1.91666
intercept class4				3.19436
poly(TIME, degree = 2, raw = TRUE)1 class1	0.00299			0.00299
poly(TIME, degree = 2, raw = TRUE)1 class2	0.00732			0.00732
poly(TIME, degree = 2, raw = TRUE)1 class3	0.01428			0.01428
poly(TIME, degree = 2, raw = TRUE)1 class4	0.03408			0.03408
poly(TIME, degree = 2, raw = TRUE)2 class1	0.00000			0.00000
poly(TIME, degree = 2, raw = TRUE)2 class2	0.00000			0.00000
poly(TIME, degree = 2, raw = TRUE)2 class3	-0.00001			-0.00001
poly(TIME, degree = 2, raw = TRUE)2 class4	-0.00002			-0.00002

Variance-covariance matrix of the random-effects:

	intercept	TIME	
intercept	1.00074		
TIME	-0.00025	0	

coef Se

Residual standard error: 0.43258

> **postprob(m4)**

Posterior classification:

	class1	class2	class3	class4
N	3	7.00	4.00	1.00
%	20	46.67	26.67	6.67

Posterior classification table:

--> mean of posterior probabilities in each class

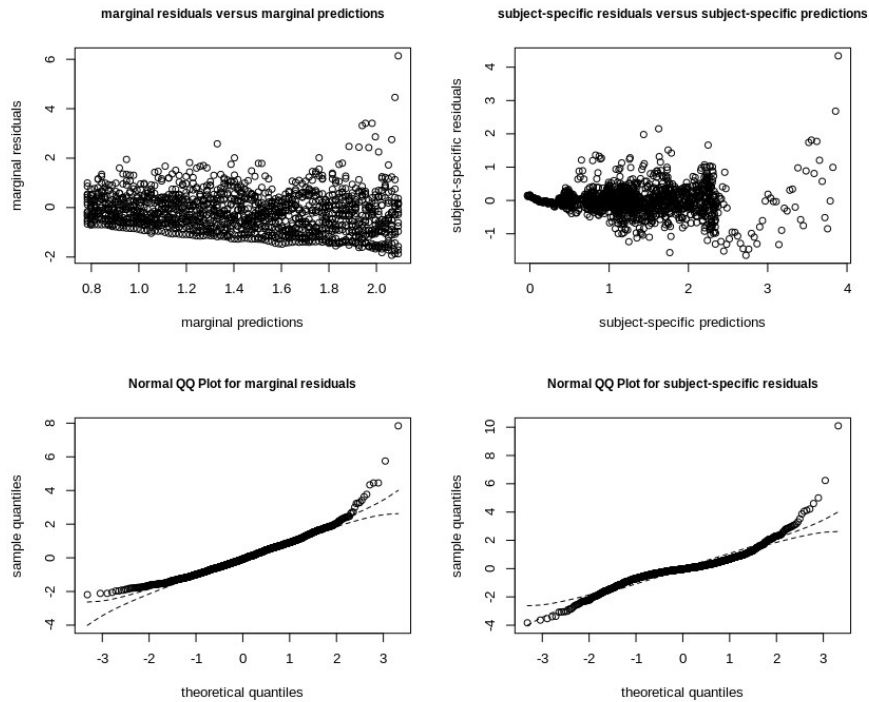
	prob1	prob2	prob3	prob4
class1	0.9944	0.0056	0.0000	0
class2	0.0300	0.9543	0.0157	0
class3	0.0000	0.0033	0.9967	0
class4	0.0000	0.0000	0.0000	1

Posterior probabilities above a threshold (%):

```

class1 class2 class3 class4
prob>0.7 100 100 100 100
prob>0.8 100 100 100 100
prob>0.9 100 100 100 100
#Goodness of fit 4: plot of residuals
plot(m4, cex.main=0.9)

```



```

#Goodness of fit 4: plot of predictions versus observations
plot(m4, which="fit", var.time="TIME",
      bty="l", xlab="YEARS", break.times=15, ylab="latent process",
      lwd=2, marg=FALSE, ylim=c(-1,4),xlim=c(1910,2005), shades=TRUE,
      col=c(2,3,4,5))

```

```

> summarytable(m1,m2, m3,m4 ,m5,which = c("loglik", "conv", "npm",
"AIC"))

```

	loglik	conv	npm	AIC
m1	-1738.9924	2	10	3497.985
m2	-912.6885	2	11	1847.377
m3	-864.1645	2	15	1758.329
m4	-858.6999	2	19	1755.400
m5	-871.8418	2	23	1789.684

```

> summarytable(m1,m2, m3,m4 ,m5)

```

	G	loglik	npm	BIC	%class1	%class2	%class3	%class4	%class5
m1	1	-1738.9924	10	3505.065	100.00000				
m2	2	-912.6885	11	1855.166	73.33333	26.66667			
m3	3	-864.1645	15	1768.950	53.33333	40.00000	6.66667		
m4	4	-858.6999	19	1768.853	20.00000	46.66667	26.66667	6.66667	

```
m5 5 -871.8418 23 1805.969 53.33333 40.00000 0.000000 0.000000  
6.666667
```

**# Test de wald**

```
> WaldMult(m4, pos = c(5, 6),  
+         name = "4 clusters")  
      Wald Test p_value  
4 clusters 0.24667 0.88397  
> WaldMult(m3, pos = c(5, 6), name = "3 clusters")  
      Wald Test p_value  
3 clusters 1.72292 0.42254  
> WaldMult(m2, pos = c(5, 6), name = "2 clusters")  
      Wald Test p_value  
2 clusters 0      1
```