Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales

Departamento de Matemática

Métodos para la segmentación de datos longitudinales. Aplicación a datos de rendimientos de cultivos en Argentina.

Cecilia Oliva

Tesis de Licenciatura en Ciencias Matemáticas

Directora: Dra. Liliana Orellana
Co-Director: Mg. Andrés Farall

Buenos Aires, Febrero 2015

Agradecimientos

Quiero dedicar esta tesis a mis abuelos Leonor y Juan Carlos, quienes llenaron de alegría mi infancia. También agradecer especialmente a todos los que participaron, colaboraron y me apoyaron para que sea posible la realización de esta tesis.

A mis papás, por darme la posibilidad de estudiar esta carrera.

A mis hermanas y mis amigos, por acompañarme en esta etapa.

A mis compañeros de cursada y de estudio, por todos los momentos compartidos.

A mis amigos y profes del CBC, con quienes compartí estudio y también trabajo.

A mis profesores, por su amplia dedicación y enseñanzas.

A Liliana Orellana y Andrés Farall, por su gran ayuda y compromiso en la elaboración de esta tesis.

A Jean-Philippe Boulanger, quien propuso el problema que plantea esta tesis, por haberme dado la posibilidad de llevarlo adelante.

A mis compañeros de trabajo, por el aprendizaje día a día.

A mis colegas de la Facultad de Ingeniería, quienes abrieron mi camino en la docencia.

A Diego, por el aguante, y por su apoyo incondicional.

Índice

Capítulo I

INTRODUCCION

En la discusión actual sobre cambio climático y su potencial influencia en el rendimiento de la producción agrícola resulta indispensable entender cómo diferentes escenarios de cambio climático pueden afectar el rendimiento de los cultivos. Para ello es necesario estimar el comportamiento esperado de un cultivo dado en una cierta región geográfica bajo diferentes condiciones climáticas. Por otra parte, una forma de optimizar la estrategia de siembra y manejo de cultivos, de un productor en una cierta campaña, sería disponer de una predicción sobre el rendimiento (predicción a mediano/largo plazo) dadas las condiciones climáticas que se esperan en la zona para los meses de la campaña. Para ello, nuevamente se requiere estimar el rendimiento esperado de un cierto cultivo en base a los datos disponibles.

Los datos de rendimiento de los diferentes tipos de cultivos, que en los países desarrollados corresponden a datos censales, abarcan muchos años y pueden estar disponibles a nivel de punto (nivel productor) o de agregados espaciales. Cuando se representan los datos de rendimientos de cultivos a lo largo del tiempo en distintas zonas geográficas, suelen observarse tendencias crecientes, pero además, una notable heterogeneidad en las trayectorias de rendimiento. Diferencias en el perfil de evolución de un cultivo para distintos agregados (o puntos) pueden deberse, entre otros, a: a) fenómenos puntuales ocurridos en un cierto momento del tiempo sólo en ciertas regiones (plagas, tornados, eventos extremos, etc.) ; b) diferencias en las condiciones climáticas (temperatura, precipitación, etc); c) diferencias en la incorporación de tecnología (nuevas semillas, fertilizantes, maquinaria, manejo de suelos, etc.); d) diferencias en políticas de promoción o innovación.

Una posible estrategia para estimar la curva de rendimientos para un cultivo en una cierta localización sería construir modelos de regresión que incluyan todos los determinantes de la evolución temporal del rendimiento. Desafortunadamente, no es usual disponer de información longitudinal de estos determinantes a nivel de punto o de aglomerado, por lo que esta propuesta resulta inviable.

Una segunda posibilidad sería intentar identificar grupos de aglomerados/puntos con patrones similares de rendimientos y ajustar la curva de rendimientos dentro de cada grupo. Los métodos de segmentación son algoritmos no supervisados, ubicados dentro de la rama del Aprendizaje Automático, que producen una partición del conjunto de objetos. Es de esperar que los factores descriptos anteriormente (clima, suelo, incorporación de tecnología, etc.) ocurran con cierta contigüidad espacial por lo que es razonable asumir que los grupos representarán regiones geográficas. Sin embargo, los límites de estas "regiones" no son impuestos por el analista sino que surgen del análisis.

El Ministerio de Agricultura de la Nación Argentina, recolecta información de siembra, cosecha, producción y rinde de cada campaña y la misma está disponible a nivel de departamentos. En esta tesis se usarán los datos anuales de cultivo de maíz, correspondientes al período 1969-2010. Es decir, la información relativa a cada departamento se puede representar en un vector de dimensión 42. En este contexto, no existe una "curva" de rendimientos, sino un "perfil" con una observación de rendimiento por año. A lo largo de esta tesis asumiremos que el comportamiento subyacente del vector de rendimientos esperados es suave, es decir, los rendimientos esperados para dos años sucesivos t y $t+1$ no pueden ser drásticamente diferentes. A pesar de que sólo se dispone de datos anuales de rendimiento, o sea un objeto discreto, usaremos polinomios para modelar los rendimientos esperados. El hecho de modelar con polinomios nos ayuda a trabajar con objetos fáciles de manejar y fáciles de codificar o de reproducir, es decir, con una pequeña cantidad de parámetros abarcamos una cantidad de comportamientos temporales muy diversos.

El *rendimiento observado* en un cierto año, para un cierto departamento, puede haber sido afectado por una plaga, una sequía, una inundación etc. que haga que éste sea muy distinto del *rendimiento esperado* para la zona a la que pertenece el departamento. El rendimiento esperado, por otro lado, es el rendimiento que “en promedio” debería ocurrir en una dada región, dadas las condiciones tecnológicas del momento.

El objetivo de esta tesis es proponer una metodología que permita: 1) identificar departamentos con perfiles de rendimientos similares (grupos) usando algoritmos de segmentación; 2) determinar la existencia de grupos para lo cual se propone una

medida de significatividad de la segmentación; y 3) evaluar si existe contigüidad espacial en el agrupamiento de perfiles, es decir, departamentos con patrones similares de rendimientos se ubican en zonas geográficamente cercanas o contiguas. En esta tesis se usarán datos anuales de cultivos de maíz, período 1969-2010, pero la metodología propuesta es de aplicación general, pudiendo aplicarse a cualquier tipo de cultivo, y a cualquier otro fenómeno indexado por el tiempo (datos longitudinales).

La tesis se organiza de la siguiente manera. En el capítulo 2 se describen los procedimientos estadísticos utilizados en la metodología propuesta. En particular se presentan las distintas técnicas de segmentación destacando las ventajas y desventajas de cada método. Se resumen los resultados del paper de Tibshirani, Walther y Hastie (2001) sobre el *estadístico Gap*, que permite seleccionar el número óptimo de clusters y que será usado para evaluar la significatividad de la segmentación. Se describe además el cálculo de componentes principales, un método de reducción de dimensión que utilizaremos al momento de simular trayectorias de rendimientos. Asimismo, se introduce la noción de test de hipótesis, que utilizaremos para evaluar la significatividad estadística en la contigüidad espacial de los patrones de rendimientos.

En el capítulo 3 se comienza describiendo el tipo de datos disponibles, su calidad y las características de las variables utilizadas. En este capítulo se presenta en detalle la propuesta metodológica, la que se ejemplifica aplicándola a los datos de trayectorias temporales de rendimiento de maíz en Argentina. Se discute el problema de determinar el número de agrupamientos, si existiese segmentación, y se propone una metodología para cuantificar la significatividad de la segmentación y un procedimiento para estudiar si hay contigüidad espacial en los perfiles de rendimientos.

El último capítulo presenta las conclusiones sobre el análisis de los datos de cultivos de maíz y sobre la propuesta metodológica. Finalmente se describen los temas abiertos para futuras investigaciones.

La lectura de datos de la información provista por el Ministerio de Agricultura y el Servicio Meteorológico, la limpieza y control de consistencia de datos, los gráficos, mapas, el ajuste de modelos estadísticos, y los métodos de segmentación fueron realizados mediante el lenguaje de programación **R.** Este lenguaje permite trabajar en un entorno donde se pueden implementar diversas técnicas estadísticas, las cuales se encuentran disponibles en las llamadas *bibliotecas* de **R**. A lo largo de la tesis se hace referencia a las funciones de **R** utilizadas.

Capítulo II

METODOLOGIA

En este capítulo se describen los procedimientos estadísticos utilizados en la propuesta metodológica, la presentación no pretende ser exhaustiva sino focalizada en las técnicas que resultan de interés para esta tesis. En primer lugar se presentan las distintas técnicas de segmentación destacando las ventajas y desventajas de cada método y comparando resultados. Se resumen los resultados del paper de Tibshirani, Walther y Hastie (2001) sobre el *estadístico Gap*, que permite seleccionar el número óptimo de clusters y que será usado para evaluar la significatividad de la segmentación. Se describe además el cálculo de componentes principales, un método de reducción de dimensiones que utilizaremos al momento de simular trayectorias de rendimientos. Finalmente se introduce la noción de test de hipótesis, utilizado para evaluar la significatividad de la contigüidad espacial de los patrones de rendimientos.

1. Segmentación

Se llama *segmentación* a la metodología que agrupa objetos similares a partir de características propias de los mismos. Esta técnica se conoce también como *análisis de clusters*, conglomerados o grupos. La segmentación es un método exploratorio cuyo objetivo es, dada una muestra de n objetos sobre los que se han registrado p variables, usando la información provista por las p variables, agrupar los objetos de modo que aquellos que sean más "similares" pertenezcan al mismo cluster o grupo. En general, se busca homogeneidad dentro de los grupos y heterogeneidad entre grupos. La técnica de segmentación es una herramienta útil en el proceso básico de investigación, en el cual interesa buscar patrones en los datos y tratar de encontrar leyes que expliquen estos patrones.

El análisis de clusters se usa para diferentes propósitos tales como: resumir una gran cantidad de datos en una pequeña cantidad de prototipos, buscar "grupos naturales" o generar criterios de clasificación (taxonómicos) o generar hipótesis.

Es importante distinguir entre técnicas de agrupamiento (*segmentación*) y métodos de clasificación (*discriminación*). En el caso de segmentación se dispone de una cierta cantidad de objetos, n, y se quiere encontrar grupos de objetos similares, sin conocer a qué grupos pertenecen las observaciones ni la cantidad de grupos. En el caso de clasificación, se dispone de observaciones clasificadas en grupos y el objetivo es definir un criterio que permita clasificar una nueva observación y asignarla a uno de los grupos existentes.

Una posible manera de generar grupos es enumerar todos los agrupamientos posibles y seleccionar "el mejor" en algún sentido. Sin embargo, el número total de agrupamientos es en general muy grande por lo que resultaría muy costoso encontrar el óptimo. La idea entonces es, aplicar algoritmos que busquen un buen agrupamiento, aunque no necesariamente el mejor.

El método de segmentación se basa en agrupar los objetos en base a medidas de similitud o disimilitud (distancia). Sea $\mathbf{x}_i' = (x_{i1}, \ldots, x_{ip})$ el vector $p-$dimensional que contiene las observaciones del $i-$ésimo objeto. Dos tipos de datos pueden usarse como input de un análisis de clusters:

1) El arreglo de $n \times p$ datos correspondientes a las observaciones de p variables en n objetos. Es decir, una matriz de datos $\mathbf{X}$, con columnas $\mathbf{x}_k$:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_i & \cdots & \mathbf{x}_p \end{pmatrix} = (x_{ik}) \quad i = 1, ..., n, \ k = 1, ..., p.$$

2) La matriz de proximidad de dimensión $n \times n$, (c_{rs}) o (d_{rs}), donde c_{rs} es una medida de similaridad y d_{rs} una medida de disimilaridad entre los objetos $r-$ésimo y $s-$ésimo.

Los algoritmos de segmentación más comunes se clasifican en dos clases: *jerárquica* y *particionada*. A su vez los métodos jerárquicos se subdividen en *aglomerativos,* también llamados 'ascendentes', y *divisivos*, o denominados 'descendentes'. Los

aglomerativos comienzan con tantos grupos o clusters como elementos tengamos en nuestra muestra, los que se unen en cada paso hasta obtener un único grupo que contiene el total de los elementos, de ahí el nombre 'ascendente'. Mientras que los algoritmos divisivos realizan el proceso inverso, partiendo de un conjunto global de elementos, que se va subdividiendo hasta llegar a grupos unitarios. En los métodos *particionantes*, a diferencia de los jerárquicos, el número de clases se establece previamente y el algoritmo asigna los elementos a las clases, partiendo de algunos valores iniciales y buscando optimizar algún criterio establecido de antemano.

En la siguiente sección se definen las medidas de disimilaridad y de similitud y en las subsiguientes se describen las características centrales de cada uno de los procedimientos.

1.1. Medidas de disimilaridad (distancia) o similitud

1.1.1. Medidas de distancia entre objetos

Recordemos que una medida de distancia, o métrica, en $\mathbb{R}^p$ es una función *d:* $\mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ que satisface para todo $\mathbf{x}, \mathbf{y}$ y $\mathbf{z}$ en $\mathbb{R}^p$:

1) $d(\mathbf{x}, \mathbf{y}) \geq 0$ para todo $\mathbf{x}$, $\mathbf{y}$ en $\mathbb{R}^p$

2) $d(\mathbf{x},\mathbf{y}) = 0$ si y sólo si $\mathbf{x} = \mathbf{y}$

3) $d(\mathbf{x},\mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ para todo $\mathbf{x}$, $\mathbf{y}$ en $\mathbb{R}^p$

4) $d(\mathbf{x},\mathbf{y}) \leq d(\mathbf{x},\mathbf{z}) + d(\mathbf{z},\mathbf{y})$ para todo $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ en $\mathbb{R}^p$

Las condiciones anteriores implican que $d(\mathbf{x}, \mathbf{y})$ es definida positiva (1, 2), simétrica (3) y cumple con la desigualdad triangular (4). Sin embargo, en esta tesis consideraremos también distancias que no verifican la desigualdad triangular.

La métrica de Minkowski

$$d(\mathbf{x},\mathbf{y}) \quad = \quad \left[\sum_{j=1}^{p} |x_j - y_j|^m \right]^{1/m}$$

define una familia de distancias que incluye a la métrica euclídea ($m = 2$) y a la métrica de "City Block" ($m = 1$).

Una desventaja de estas métricas es que cambios de escala pueden tener un efecto sustancial sobre el orden de las distancias. Por ejemplo, supongamos que tenemos tres objetos cuyos valores de peso y altura son:

Objeto	peso (gr)	altura (cm)
A	10	2
B	20	1
C	30	4

Las distancias euclídeas entre ellos son: $d(A, B) = 10{,}05$, $d(A, C) = 20{,}1$ y $d(B, C) = 10{,}44$. De donde se concluye que $d(A, B) < d(B, C) < d(A, C)$, es decir C está más cerca de B que de A. Sin embargo, cuando la altura viene dada en milímetros, tenemos $d(A, B) = 14{,}14$, $d(A, C) = 28{,}28$ y $d(B, C) = 31{,}62$. Luego, $d(A, B) < d(A, C) < d(B, C)$, el orden de las distancias cambia y ahora C está más cerca de A que de B.

Idealmente uno quisiera estandarizar los datos de modo que las varianzas "dentro" de los clusters fueran similares. Un modo de escalamiento posible sería usar la distancia de Mahalanobis o distancia Estadística

$$d(\mathbf{x},\mathbf{y}) \quad = \quad \left[(\mathbf{x}-\mathbf{y})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{y})\right]^{1/2}$$

donde $\mathbf{\Sigma}$ es la matriz de covarianzas.

Se han propuesto además modos de estandarizar la métrica L^1 de Minkowski, Gower (1971) ([4]) estandariza usando el rango de cada variable ($R = \underset{i}{max}(x_{ik}) - \underset{i}{min}(x_{ik})$), aunque las distancias estandarizadas resultan ser muy sensibles a outliers. Aplicar alguna estandarización es equivalente a pesar las variables y no hay acuerdo sobre lo que sería apropiado en general.

En el caso particular en que las variables son dicotómicas (presencia/ausencia), en las que la presencia o ausencia de una característica puede ser escrita como

$$x_{ik} \quad = \quad \begin{cases} 0 & si\ la\ característica\ k\ está\ ausente\ en\ el\ sujeto\ i \\ 1 & si\ la\ característica\ k\ está\ presente\ en\ el\ sujeto\ i \end{cases}$$

la distancia euclídea entre dos observaciones $\mathbf{x}_i$ y $\mathbf{x}_j$ resulta ser igual al número de desacuerdos,

$$d^2(\mathbf{x}_i - \mathbf{x}_j) \quad = \quad \sum_{i=1}^{p}(x_{ik} - x_{jk})^2 = cantidad\,de\,desacuerdos$$

ya que

$$(x_{ik} - x_{jk})^2 \quad = \quad \begin{cases} 0 & cuando\ hay\ acuerdo \\ 1 & cuando\ hay\ desacuerdo \end{cases}$$

Luego, d^2 aumenta con la cantidad de desacuerdos. Esta distancia tiene el problema de pesar del mismo modo los acuerdos $1-1$ y los $0-0$. En algunos casos $1-1$ es más indicativo de similitud que $0-0$. Por ejemplo, cuando se agrupa gente, el hecho que dos personas hayan leído los clásicos griegos es fuerte evidencia de similitud mientras que el no haberlos leído no lo es. En la sección siguiente se presentan medidas de similitud que pesan de manera diferencial los distintos tipos de acuerdo.

Un ejemplo es la distancia Chi-cuadrado, esta distancia es una noción de distancia entre individuos, i y j, con $1 \leq i,\ j \leq n$, que contempla pesos distintos para cada variable o atributo k, con $1 \leq k \leq p$. Para cada par de individuos la distancia se calcula como sigue*:*

$$d_{ij} \quad = \quad \sqrt{\sum_{k=1}^{p} \frac{(x_{ik} - x_{jk})^2}{\sum_{z=1}^{n} x_{zk}}}.$$

Esta métrica tiene en cuenta la frecuencia relativa de cada atributo, dando preponderancia a los atributos menos frecuentes.

1.1.2. Medidas de similitud entre objetos

Los coeficientes de similitud se conocen también como coeficientes de asociación y en general se usan para variables dicotómicas. Consideremos un estudio donde se miden p variables dicotómicas sobre n objetos. Para cada par de objetos

$\mathbf{x}_i$ y $\mathbf{x}_j$ podemos armar una tabla de contingencia resumiendo el número de acuerdos y de desacuerdos entre las p variables de los dos objetos:

objeto i

objeto j	1	0	
1	a	b	a+b
0	c	d	c+d
	a+c	b+d	p

Algunas medidas de asociación propuestas son:

1) $\frac{a+d}{p}$ igual peso para acuerdos $1-1$ y acuerdos $0-0$,

2) $\frac{2(a+d)}{2(a+d)+b+c}$ doble peso para acuerdos,

3) $\frac{a+d}{a+d+2(b+c)}$ doble peso para desacuerdos,

4) $\frac{a}{p}$ no considera los acuerdos tipo $0-0$ y

5) $\frac{a}{a+b+c}$ los acuerdos $0-0$ son irrelevantes (coeficiente de Jaccard).

La elección del coeficiente depende de los pesos relativos que se desee asignar a los acuerdos $1-1$ y a los acuerdos $0-0$. Claramente hay situaciones en las que la ausencia conjunta de una característica no da información y un coeficiente que da igual peso a ambos tipos de acuerdo no resulta apropiado.

Finalmente, notemos que siempre es posible construir una medida de similitud a partir de una distancia. Sea d_{ij} la distancia entre dos objetos i y j, y sea s_{ij} la medida de similitud entre ellos.

$$s_{ij} = \frac{1}{1+d_{ij}}$$

Observemos que, si $d_{ij}=0$, entonces $s_{ij}=1$, y si $d_{ij}\to\infty$ s_{ij} tiende a 0. Para el resto de los casos, $d_{ij}>0$, $s_{ij}\in(0,1)$. Luego, $0\leq s_{ij}\leq 1$. Sin embargo, no es siempre posible construir una medida de distancia a partir de similitud.

1.2. Métodos Jerárquicos de Agrupamiento

Como ya mencionamos, existen dos tipos de métodos jerárquicos. Los *aglomerativos*, que parten de n objetos, donde cada uno constituye un cluster, los que se van agrupando de acuerdo a similitudes hasta obtener un único cluster. Y los métodos *divisivos*, que comienzan con un único grupo que contiene todos los objetos, el que se divide en dos subgrupos de manera que los objetos en un subgrupo se encuentren "lejos" de los objetos del otro subgrupo. El proceso continúa hasta obtener n clusters. Estos algoritmos se denominan jerárquicos porque a cada nivel los nuevos agrupamientos (métodos aglomerativos) o divisiones (métodos divisivos) sólo pueden ocurrir entre los clusters definidos en el nivel anterior. Es decir, si en un determinado nivel del procedimiento aglomerativo se agrupan dos clusters, éstos quedan ya jerárquicamente agrupados para el resto de los niveles.

Los métodos jerárquicos se basan en una medida de distancia o similitud entre los elementos, y una métrica entre clusters. El tipo de medida que se utilice para definir la distancia entre grupos genera las distintas técnicas de unión o separación de clases que dan nombre a las metodologías de segmentación. El resultado de ambos tipos de métodos (aglomerativos y divisivos) se muestra en un diagrama conocido como *dendrograma* que ilustra la secuencia de fusiones o divisiones en los

sucesivos niveles. Cada rama del árbol representa un cluster. Las ramas se unen con nodos cuya posición representa la distancia (eje vertical) o nivel a la cual ocurrió la fusión.

Los métodos jerárquicos son muy útiles porque sugieren el número de clusters, y porque permiten establecer una jerarquía de agrupamientos. Además, el dendrograma facilita la visualización del proceso. Sin embargo, resulta costoso al aplicarlo en grandes bases de datos, y puede ser lento. Los métodos divisivos son menos populares que los ascendentes y se ha mostrado en estudios taxonómicos que tienen baja performance por lo que sólo describiremos los métodos ascendentes.

Antes de explicar los detalles de cada método es importante mencionar que ninguno de estos procedimientos brinda una solución óptima para un dado problema que se pueda plantear. En general distintos métodos producirán distintos resultados. El método más adecuado para cada caso resultará de una combinación entre la elección del algoritmo y la experiencia y el buen criterio del investigador.

Notemos que con la segmentación, jerárquica o particionada, se obtiene una partición de un conjunto de n elementos en k subconjuntos. Definimos una *partición* de un conjunto A de n elementos como la familia $\mathcal{P}_A$ de subconjuntos no vacíos de A, disjuntos dos a dos, cuya unión es A. Es decir,

$$\mathcal{P}_A \quad = \quad \{A_i : i \in \mathbb{N}_{\leq n}\},$$

donde se verifican las siguientes condiciones:

1. $A_i \subseteq A$, $A_i \neq \emptyset \ \forall\, i \in \mathbb{N}_{\leq n}$,
2. $A_i \cap A_j = \emptyset$ si $i \neq j$,
3. $A = \cup_{i \in \mathbb{N}_{\leq n}} A_i$.

En esta sección presentaremos diferentes algoritmos aglomerativos, los cuales comienzan con tantos clusters como elementos y los clusters se van combinando secuencialmente. En cada paso se combinan los dos clusters separados por la "mínima distancia". La definición de "mínima distancia" produce los diferentes métodos aglomerativos. En los procedimientos denominados *linkage* (unión-encadenamiento-amalgamiento) dos clusters se unen si alguna medida de distancia entre los clusters es mínima: la distancia entre vecinos más cercanos (*Single-linkage*), la distancia entre vecinos más lejanos (*Complete-linkage*), o el promedio de las distancias entre todos los pares de puntos de los dos clusters (*Average-linkage*). En los métodos del centroide y de la mediana se minimiza la distancia entre medidas resúmenes de los objetos que pertenecen a un cluster. Finalmente en el método de Ward se minimiza la varianza total dentro de clusters.

1.2.1. *Single Linkage*

En el método de *Single linkage*, la distancia entre dos clusters se define como la mínima distancia entre dos elementos (uno en cada cluster). El par de clusters que comparta los dos elementos más cercanos es el que se unirá en cada paso, por esta razón el método se denomina *de vecinos más cercanos*. Supongamos que nos encontramos en la etapa k-ésima, tenemos ya formados n-k clusters, la distancia entre los clusters C_i (con n_i elementos) y C_j (con n_j elementos) se define como

$$d(C_i, C_j) \quad = \quad \underset{\mathbf{x}_l \in C_i,\, \mathbf{x}_m \in C_j}{mín} \{d(\mathbf{x}_l, \mathbf{x}_m) : l = 1, ..., n_i; m = 1, ..., n_j\}.$$

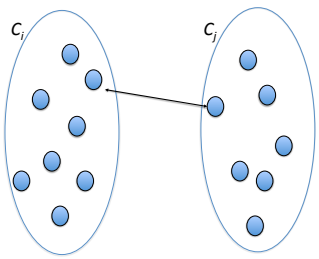

Figura 1: *Single Linkage considera la mínima distancia entre elementos, pertenecientes a distintos clusters, más cercanos entre sí, para determinar la próxima fusión.*

Ejemplo: supongamos que se dispone de información de siete objetos identificados como A, B, ..., G, y que se ha propuesto una medida de distancia a partir de la cual se obtuvo la siguiente matriz de distancias

	A	B	C	D	E	F	G
A	0						
B	0.6717854	0					
C	0.1439147	0.8054837	0				
D	0.3849284	0.4594972	0.5240015	0			
E	0.8305514	1.1500719	0.8777698	0.7066962	0		
F	0.2487760	0.9078208	0.1051530	0.6246905	0.9158099	0	
G	0.6997975	0.4561342	0.8396928	0.3157182	0.7801180	0.9402692	0

Los pasos a seguir utilizando la estrategia *Single linkage* son:

1. Nivel 1. Calculamos $\underset{i<j}{min}\{d(C_i, C_j)\} = d(C, F) = 0{,}1051530$, obtenemos el primer cluster no unitario (C, F).

2. Nivel 2. La distancia de un elemento Y al cluster (C, F) se calcula como $d\left(Y, (C, F)\right) = min\left\{d\left(Y, C\right), d\left(Y, F\right)\right\}$; y la matriz de distancias es

	A	B	(C,F)	D	E	G
A	0					
B	0.6717854	0				
(C,F)	0.1439147	0.8054837	0			
D	0.3849284	0.4594972	0.5240015	0		
E	0.8305514	1.1500719	0.8777698	0.7066962	0	
G	0.6997975	0.4561342	0.8396928	0.3157182	0.7801180	0

   En este paso $\underset{i<j}{min}\{d(C_i, C_j)\} = d(A, (C, F)) = 0{,}1439147$ obteniendo un cluster más grande que contiene al anterior $(A, (C, F))$.

3. Nivel 3. La matriz de distancias es

	(A,(C,F))	B	D	E	G
(A,(C,F))	0				
B	0.6717854	0			
D	0.3849284	0.4594972	0		
E	0.8305514	1.1500719	0.7066962	0	
G	0.6997975	0.4561342	0.3157182	0.7801180	0

y $\underset{i<j}{min}\{d(C_i, C_j)\} = d(D, G) = 0{,}3157182$ por lo que se forma el tercer cluster *(D,G).*

4. Nivel 4. La matriz de distancias en este paso es la siguiente y, la distancia mínima corresponde a $d((A,(C,F)),(D,G))$ y se forma un nuevo cluster.

	(A,(C,F))	B	(D,G)	E
(A,(C,F))	0			
B	0.6717854	0		
(D,G)	0.3849284	0.4561342	0	
E	0.8305514	1.1500719	0.7066962	0

5. Nivel 5. Se unen *((A,(C,F)),(D,G))* y *B.*

	((A,(C,F)),(D,G))	B	E
((A,(C,F)),(D,G))	0		
B	0.4561342	0	
E	0.7066962	1.1500719	0

6. Nivel 6. Se unen todos los elementos en un único cluster.

	(((A,(C,F)),(D,G)),B)	E
(((A,(C,F)),(D,G)),B)	0	
E	0.7066962	0

La Figura 2 muestra el dendrograma, que es un modo gráfico de representar la secuencia de fusiones entre los clusters. El eje vertical refleja la distancia a la cual dos clusters se unieron. El gráfico sugiere la existencia de al menos dos clusters en estos datos ya que la distancia asociada con la fusión de los clusters (((A,(C,F)),(D,G)),B) y E es notablemente mayor que las distancias a las que ocurrieron las demás fusiones.

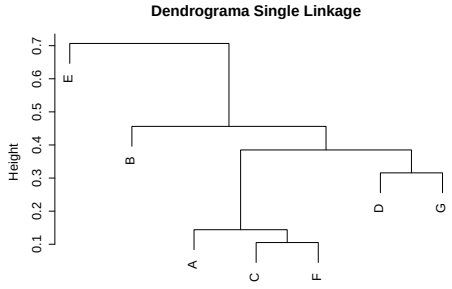

Figura 2: *Dendrograma para el ejemplo del método Single Linkage.*

1.2.2. *Complete Linkage*

Este algoritmo, también conocido como el método de los vecinos más lejanos, propone como medida de distancia entre dos clusters la distancia entre los dos elementos más alejados (uno en cada cluster). El par de cluster en que esta distancia sea mínima será el que se fusione en cada paso. En una dada etapa del procedimiento la distancia entre los clusters C_i (con n_i elementos) y C_j (con n_j elementos) es

$$d(C_i, C_j) \quad = \quad \underset{x_l \in C_i,\, x_m \in C_j}{máx} \{d(x_l, x_m) : l = 1, ..., n_i; m = 1, ..., n_j\}$$

y se unirán los clusters en los que esta medida de distancia sea mínima.

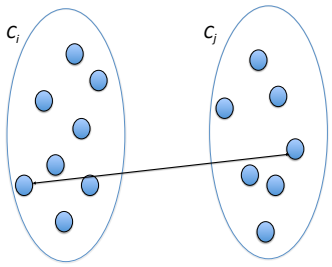

Figura 3: *Complete Linkage considera la mínima distancia entre elementos, pertenecientes a distintos clusters, más lejanos entre sí, para determinar la próxima fusión.*

Ejemplo: Aplicaremos el procedimiento a la matriz de distancias del ejemplo anterior.

1. Nivel 1. $\underset{i<j}{min}\{d(C_i, C_j)\} = d(C, F) =$ 0,1051530, obtenemos el primer cluster (C, F).

2. Nivel 2. La distancia de un elemento Y al cluster (C, F) se calcula como $d\left(Y, (C, F)\right) = max\left\{d\left(Y, C\right), d\left(Y, F\right)\right\}$ y la nueva matriz de distancias es

	A	B	(C,F)	D	E	G
A	0					
B	0.6717854	0				
(C,F)	0.2487760	0.9078208	0			
D	0.3849284	0.4594972	0.6246905	0		
E	0.8305514	1.1500719	0.9158099	0.7066962	0	
G	0.6997975	0.4561342	0.9402692	0.3157182	0.7801180	0

   Notemos que $\underset{i<j}{min}\{d(C_i, C_j)\} = d(A, (C, F)) =$ 0,2487760 por lo que se forma el segundo cluster $(A, (C, F))$.

3. Nivel 3. La matriz de distancias es la siguiente y $\underset{i<j}{min}\{d(C_i, C_j)\} = d(D, G) =$ 0,3157182 obteniendo el tercer cluster(D, G).

	(A,(C,F))	B	D	E	G
(A,(C,F))	0				
B	0.9078208	0			
D	0.6246905	0.4594972	0		
E	0.9158099	1.1500719	0.7066962	0	
G	0.9402692	0.4561342	0.3157182	0.7801180	0

4. Nivel 4. La matriz de distancias se muestra a continuación y $\underset{i<j}{min}\{d(C_i, C_j)\} = d(B,(D,G)) = 0.4594972$ obteniendo un cluster más grande que contiene al anterior *(D,G).*

	(A,(C,F))	B	(D,G)	E
(A,(C,F))	0			
B	0.9078208	0		
(D,G)	0.9402692	0.4594972	0	
E	0.9158099	1.1500719	0.7801180	0

5. Nivel 5. En este caso se obtiene la matriz de distancias que sigue y $\underset{i<j}{min}\{d(C_i, C_j)\} = d((A,(C,F)),E) = 0{,}9158099$ generando el cluster $((A,(C,F)),E)$.

	(A,(C,F))	(B,(D,G))	E
(A,(C,F))	0		
(B,(D,G))	0.9402692	0	
E	0.9158099	1.1500719	0

6. Nivel 6. Finalmente, se unen todos los elementos en un único cluster que contiene a los 7 individuos.

	((A,(C,F)),E)	(B,(D,G))
((A,(C,F)),E)	0	
(B,(D,G))	1.1500719	0

Con el método *Complete linkage,* se obtiene el dendrograma que se presenta en la Figura 4 que difiere del obtenido usando Single linkage. En la mayoría de los casos cuando se aplican métodos de segmentación diferentes al mismo conjunto de datos se obtienen diferentes secuencias de agrupamiento.

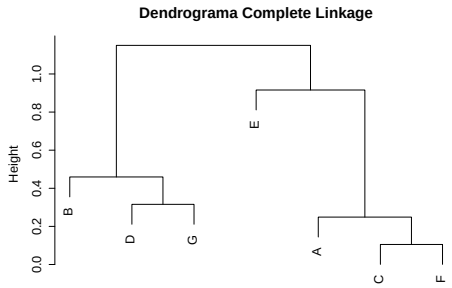

Figura 4: *Dendrograma para el ejemplo del método Complete Linkage.*

1.2.3. *Average linkage*

En el procedimiento *Average linkage* la distancia, o similitud entre clusters se define como la media aritmética de la distancia entre cada uno de los elementos del primer cluster respecto de cada uno de los elementos del segundo cluster.

Es decir, la distancia entre el cluster C_i con n_i elementos y el cluster C_j formado por n_j elementos se calcula como

$$d(C_i, C_j) = \frac{1}{n_i n_j} \sum_{l=1}^{n_i} \sum_{m=1}^{n_j} d(\mathbf{x}_{il}, \mathbf{x}_{jm}).$$

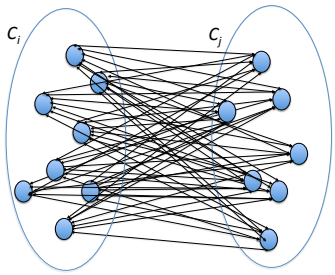

Figura 5: *Average Linkage considera el promedio de las distancias entre cada uno de los elementos del primer cluster respecto de cada uno de los elementos del segundo cluster para determinar la próxima fusión.*

Ejemplo: Aplicamos el procedimiento a la matriz de distancias anterior.

1. Nivel 1. $\min_{i<j}\{d(C_i, C_j)\} = d(C, F) = 0{,}1051530$, obtenemos el primer cluster (C, F).

2. Nivel 2. La nueva matriz de distancias es

	A	B	(C,F)	D	E	G
A	0					
B	0.6717854	0				
(C,F)	0.19634535	0.85665225	0			
D	0.3849284	0.4594972	0.574346	0		
E	0.8305514	1.1500719	0.89678985	0.7066962	0	
G	0.6997975	0.4561342	0.889981	0.3157182	0.7801180	0

   y $\min_{i<j}\{d(C_i, C_j)\} = d(A, (C, F)) = 0{,}19634535$, por lo que se forma el cluster $(A, (C, F))$ que contiene al anterior.

3. Nivel 3. La matriz de distancias es

	(A,(C,F))	B	D	E	G
(A,(C,F))	0				
B	0.79502996667	0			
D	0.5112068	0.4594972	0		
E	0.87471036667	1.1500719	0.7066962	0	
G	0.8265865	0.4561342	0.3157182	0.7801180	0

   y $\min_{i<j}\{d(C_i, C_j)\} = d(D, G) = 0{,}3157182$, obteniendo un nuevo cluster (D, G).

4. Nivel 4. La matriz de distancias en este paso es

	(A,(C,F))	B	(D,G)	E
(A,(C,F))	0			
B	0.79502996667	0		
(D,G)	0.66889665	0.4578157	0	
E	0.87471036667	1.1500719	0.7434071	0

y $\min_{i<j}\{d(C_i, C_j)\} = d(B,(D,G)) = 0{,}4578157$, obteniendo un cluster más grande que contiene al anterior *(D,G)*.

5. Nivel 5. La matriz de distancias es

	(A,(C,F))	(B,(D,G))	E
(A,(C,F))	0		
(B,(D,G))	0.71094108889	0	
E	0.87471036667	0.8789620333	0

y $\min_{i<j}\{d(C_i, C_j)\} = d((A,(C,F)),(B,(D,G))) = 0{,}71094108889$, generando el cluster $(A,(C,F)),(B,(D,G))$.

6. Nivel 6. La matriz de distancias es

	(A,(C,F)),(B,(D,G))	E
(A,(C,F)),(B,(D,G))	0	
E	0.8768362	0

y finalmente, se unen todos los elementos en un único cluster que contiene a los 7 individuos.

En este caso se obtiene el siguiente dendrograma.

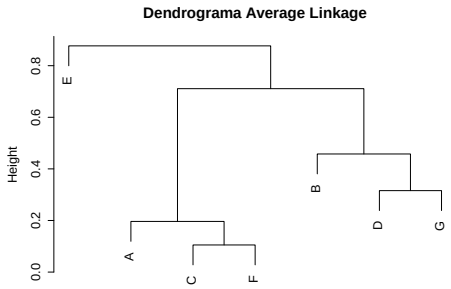

Figura 6: *Dendrograma para el ejemplo del método Average Linkage.*

Observemos que los tres procedimientos aplicados al mismo conjunto de datos condujeron a distintas secuencias de agrupamiento. Desde el primer nivel de fusión hasta el tercero, las tres estrategias unen los mismos objetos. Incluso, Complete linkage y Average linkage también realizan la misma fusión en el cuarto nivel. Sin embargo, Single linkage en el nivel cuatro fusiona $(A,(C,F))$ con (D,G), luego suma a este grupo el objeto B y finalmente se agrega el E. Complete y Average se diferencian en el quinto paso, Complete une $(A,(C,F))$ con E, mientras que Average fusiona $(A,(C,F))$ con $(B,(D,G))$. Ésto se debe a que cada método captura diferentes características de los datos. El método de Single linkage tiende a producir clusters elongados e irregulares, mientras que Complete linkage clusters de aproximadamente el mismo diámetro y Average linkage clusters con varianza similar. Este ejemplo muestra que es de suma importancia definir qué tipo de distancia se utilizará en un problema dado.

1.2.4. *Métodos del centroide y de la mediana*

En estos dos métodos cada cluster queda representado por una medida resumen de los elementos que lo componen, que no necesariamente coincide con uno de los elementos del cluster. La distancia o similaridad entre dos clusters se define como la distancia entre estas medidas resúmenes.

En el método del *centroide*, el centroide o media del cluster C_j (con n_j objetos) se define como

$$\mathbf{m}_j = \frac{1}{n_j} \sum_{\mathbf{x}_l \epsilon C_j} \mathbf{x}_l.$$

La distancia entre dos clusters C_i y C_j se define como

$$d(C_i, C_j) = d(\mathbf{m}_i, \mathbf{m}_j)$$

y en cada paso se fusionan los clusters en los que esta medida de distancia se minimiza.

El centroide de un nuevo cluster C_t resultante de la fusión de dos clusters, digamos C_i y C_j, se puede obtener simplificadamente como el promedio pesado de los centroides $\mathbf{m}_i$ y $\mathbf{m}_j$, es decir,

$$\mathbf{m}_t = \frac{n_i \mathbf{m}_i + n_j \mathbf{m}_j}{n_i + n_j}.$$

El método de la *mediana* es equivalente al del centroide, pero el centroide del nuevo cluster se define como

$$\mathbf{m}_t = \frac{\mathbf{m}_i + \mathbf{m}_j}{2}.$$

Es decir, no importa el tamaño de los clusters originales. Esta idea fue introducida para evitar que cuando un grupo pequeño se fusiona con uno grande, pierda su identidad y el nuevo centroide quede muy cerca del centroide del grupo grande. Por eso, este procedimiento es conocido también como método del centroide no ponderado.

Ejemplo:

Aplicaremos ambos métodos a los siguientes individuos sobre los cuales se han medido dos variables.

Individuo	x_1	x_2
A	10	10
B	10	20
C	20	15
D	25	25
E	30	5

Cuando se usa como medida de disimilaridad la distancia euclídea al cuadrado, obtenemos la siguiente matriz de distancias

	A	B	C	D	E
A	0				
B	100	0			
C	125	125	0		
D	450	250	125	0	
E	425	625	200	425	0

Aplicaremos en primer lugar el método del centroide.

1. Nivel 1. Obtenemos el primer cluster *(A,B)*, cuyo centroide es $(10, 15)'$.

2. Nivel 2. La matriz de distancias se transforma en:

	(A,B)	C	D	E
(A,B)	0			
C	100	0		
D	325	125	0	
E	500	200	425	0

y ahora, $\underset{i<,j}{min}\{d(C_i, C_j)\} = d((A, B), C) = 100$, por lo que se forma el cluster *((A,B),C)* y su centroide es $(13{,}33, 15)'$.

3. Nivel 3. La matriz de distancias se convierte en:

	((A,B),C)	D	E
((A,B),C)	0		
D	236.11	0	
E	377.77	425	0

donde $\underset{i,j}{min}\{d(C_i, C_j)\} = d(((A, B), C), D) = 236{,}11$, y como consecuencia se obtiene el nuevo cluster *(((A,B),C),D)* cuyo centroide es $(16{,}25, 17{,}50)'$.

4. Nivel 4. Las distancias son:

	(((A,B),C),D)	E
(((A,B),C),D)	0	
E	345.31	0

y finalmente el último centroide es $(19, 15)'$.

La Figura 7 muestra el dendrograma correspondiente a la estrategia del centroide o centroide ponderado.

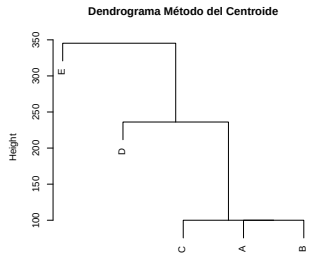

Figura 7: *Dendrograma para el ejemplo del método del Centroide.*

Aplicando el método de la mediana (centroide no ponderado) para los mismos datos tenemos:

1. Nivel 1. Como $\underset{i,j}{min}\{d(C_i, C_j)\} = d(A, B) = 100$, obtenemos el primer cluster *(A,B)*, donde el centroide es $(10, 15)'$.

2. Nivel 2. La matriz de distancias se transforma en:

	(A,B)	C	D	E
(A,B)	0			
C	100	0		
D	325	125	0	
E	500	200	425	0

y ahora, $\underset{i<,j}{min}\{d(C_i, C_j)\} = d((A, B), C) = 100$, por lo que se forma el cluster *((A,B),C)* y su centroide es $(15, 15)'$.

3. Nivel 3. La matriz de distancias se convierte en:

	((A,B),C)	D	E
((A,B),C)	0		
D	200	0	
E	325	425	0

donde $\underset{i<,j}{min}\{d(C_i, C_j)\} = d(((A, B), C), D) = 200$, y como consecuencia se obtiene el nuevo cluster *(((A,B),C),D)* cuyo centroide es $(20, 20)'$.

4. Nivel 4. Las distancias son:

	(((A,B),C),D)	E
(((A,B),C),D)	0	
E	325	0

y finalmente el último centroide es $(25, 12{,}5)'$.

La Figura 8 muestra el dendrograma correspondiente a la estrategia del centroide no ponderado. Éste es similar al dendrograma realizado con el método del centroide. En este ejemplo, la secuencia de agrupamientos producida por los dos métodos es la misma aunque varían las alturas ("Height"). Esto puede deberse a que el número de datos del ejemplo es pequeño y por lo tanto no se forman grupos de muchas observaciones y a que no hay datos extremadamente alejados del conjunto.

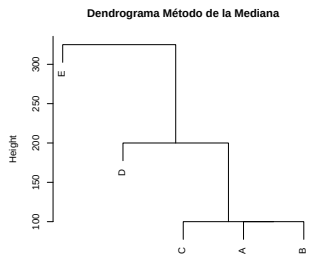

Figura 8: *Dendrograma para el ejemplo del método de la Mediana.*

1.2.5. *Método de Ward*

Este método, también conocido como "incremento en la suma de los cuadrados" o método de mínima varianza, es otro ejemplo de procedimiento jerárquico. El objetivo del método es unificar grupos de forma tal que la variabilidad dentro de los grupos no aumente dramáticamente. En cada paso se fusionan los dos clusters que producen la suma de cuadrados dentro de clusters (variabilidad within o intra-clusters) mínima entre todas las posibles particiones que se obtienen fusionando dos clusters del paso previo. En este contexto "suma de cuadrados dentro" refiere a la suma de las distancias al cuadrado de las observaciones del cluster respecto de la media de las observaciones del mismo cluster.

En el método de Ward la distancia entre dos clusters C_i y C_j, se mide como el aumento en la suma de cuadrados al unirlos. Sean C_i y C_j dos clusters con n_i y n_j elementos y medias $\mathbf{m}_i$ y $\mathbf{m}_j$. Sea $\mathbf{m}$ la media del cluster que se formaría al unir C_i

y C_j. En cada paso, la distancia o función objetivo a minimizar es I_{C_i,C_j}, el incremento en la suma de cuadrados dentro de clusters cuando se unen los dos clusters

$$I_{C_i,C_j} = \sum_{l\in C_i\cup C_j} \|\mathbf{x}_l - \mathbf{m}\|^2 - \left\{ \sum_{l\in C_i} \|\mathbf{x}_l - \mathbf{m}_i\|^2 + \sum_{l\in C_j} \|\mathbf{x}_l - \mathbf{m}_j\|^2 \right\} = \frac{n_i n_j}{n_i + n_j} \|\mathbf{m}_i - \mathbf{m}_j\|^2.$$

El siguiente lema demuestra que la segunda igualdad es válida.

Lema:

Sean C_i y C_j dos clusters con n_i y n_j elementos. Sean $\mathbf{m}_i$, $\mathbf{m}_j$ y $\mathbf{m}$ vectores con elementos m_{ik}, m_{jk} y m_k, $k = 1, ..., p$, que representan la media del cluster C_i, C_j y del cluster que se formaría al unir C_i y C_j, respectivamente. Entonces,

$$\sum_{l\in C_i\cup C_j} \|\mathbf{x}_l - \mathbf{m}\|^2 - \left\{ \sum_{l\in C_i} \|\mathbf{x}_l - \mathbf{m}_i\|^2 + \sum_{l\in C_j} \|\mathbf{x}_l - \mathbf{m}_j\|^2 \right\} = \frac{n_i n_j}{n_i + n_j} \|\mathbf{m}_i - \mathbf{m}_j\|^2.$$

Demostración:

sean $x_{lp}^{(i)}$ la componente $p-$ésima del vector $l-$ésimo del cluster C_i y $x_{lp}^{(i,j)}$ la componente $p-$ésima del vector $l-$ésimo del cluster $C_i \cup C_j$, entonces

$$\sum_{l\in C_i\cup C_j} \|\mathbf{x}_l - \mathbf{m}\|^2 - \left\{ \sum_{l\in C_i} \|\mathbf{x}_l - \mathbf{m}_i\|^2 + \sum_{l\in C_j} \|\mathbf{x}_l - \mathbf{m}_j\|^2 \right\} =$$

$$= \sum_{l=1}^{n_i+n_j} \sum_{s=1}^{p} \left(x_{ls}^{(i,j)} - m_s \right)^2 - \left\{ \sum_{l=1}^{n_i} \sum_{s=1}^{p} \left(x_{ls}^{(i)} - m_{is} \right)^2 + \sum_{k=1}^{n_j} \sum_{s=1}^{p} \left(x_{ls}^{(j)} - m_{js} \right)^2 \right\} =$$

$$= \sum_{l=1}^{n_i+n_j} \sum_{s=1}^{p} x_{ls}^{(i,j)\,2} - (n_i + n_j) \sum_{s=1}^{p} m_s^2 - \left\{ \sum_{l=1}^{n_i} \sum_{s=1}^{p} x_{ls}^{(i)2} - n_i \sum_{s=1}^{p} m_{is}^2 + \sum_{l=1}^{n_j} \sum_{s=1}^{p} x_{ls}^{(j)2} - n_j \sum_{s=1}^{p} m_{js}^2 \right\} =$$

$$= n_i \sum_{s=1}^{p} m_{is}^2 + n_j \sum_{s=1}^{p} m_{js}^2 - (n_i + n_j) \sum_{s=1}^{p} m_s^2 = \sum_{s=1}^{p} \left(n_i m_{is}^2 + n_j m_{js}^2 - (n_i + n_j)\, m_s^2 \right) \tag{1}$$

Notemos que $(n_i+n_j)m_s = n_i m_{is} + n_j m_{js}$, luego, $(n_i+n_j)^2 m_s^2 = (n_i m_{is})^2 + (n_j m_{js})^2 + 2 n_i n_j m_{is} m_{js}$. Además $2 m_{is} m_{js} = m_{is}^2 + m_{js}^2 - (m_{is} - m_{js})^2$. Por lo tanto,

$$(n_i + n_j)^2 m_s^2 \quad = n_i(n_i + n_j) \quad m_{is}^2 + n_j(n_i + n_j) m_{js}^2 - n_i n_j (m_{is} - m_{js})^2,$$

dividiendo por $(n_i + n_j)$ se obtiene

$$(n_i + n_j)\, m_s^2 = n_i m_{is}^2 + n_j m_{js}^2 - \frac{n_i n_j}{n_i + n_j} (m_{is} - m_{js})^2$$

y reemplazando en (1) resulta

$$I_{C_i,C_j} \quad = \quad \frac{n_i n_j}{n_i + n_j} \sum_{s=1}^{p} (m_{is} - m_{js})^2 = \frac{n_i n_j}{n_i + n_j} \|\mathbf{m}_i - \mathbf{m}_j\|^2. \qquad \blacksquare$$

Es decir, el criterio de similaridad es la distancia cuadrada entre los centroides pesada por un factor. El lema muestra adicionalmente que la fusión de dos clusters nunca puede reducir la suma de distancias intra-clusters. También puede apre-

ciarse que el método tenderá a unir clusters con centroides cercanos y con pocos elementos, o fuertemente desbalanceados. Este método tiende así a producir clusters esféricos y bien balanceados.

Ejemplo: Consideraremos el mismo caso que analizamos para ejemplificar los métodos del centroide:

Individuo	x_1	x_2
A	10	10
B	10	20
C	20	15
D	25	25
E	30	5

1. Nivel 1. Calculamos todas las posibles combinaciones de grupos, y los incrementos asociados

Agrupamientos	Centroides	Incrementos
((A,B),C,D,E)	$C_{AB} = (10, 15)$	$I_{AB} = 50$
((A,C),B,D,E)	$C_{AC} = (15, 12{,}5)$	$I_{AC} = 62{,}5$
((A,D),B,C,E)	$C_{AD} = (17{,}5, 17{,}5)$	$I_{AD} = 225$
((A,E),B,C,D)	$C_{AE} = (20, 7{,}5)$	$I_{AE} = 212{,}5$
((B,C),A,D,E)	$C_{BC} = (15, 17{,}5)$	$I_{BC} = 62{,}5$
((B,D),A,C,E)	$C_{BD} = (1{,}5, 22{,}5)$	$I_{BD} = 125$
((B,E),A,C,D)	$C_{BE} = (20, 12{,}5)$	$I_{BE} = 312{,}5$
((C,D),A,B,E)	$C_{CD} = (22{,}5, 20)$	$I_{CD} = 62{,}5$
((C,E),A,B,D)	$C_{CE} = (25, 10)$	$I_{CE} = 100$
((D,E),A,B,C)	$C_{DE} = (27{,}5, 15)$	$I_{DE} = 212{,}5$

   En este nivel se unirán los clusters A y B dado que su correspondiente incremento es mínimo.

2. Nivel 2. Los posibles agrupamientos son:

Agrupamientos	Centroides	Incrementos
(((A,B),C),D,E)	$C_{ABC} = (13{,}33, 15)$	$I_{ABC} = 116{,}66$
(((A,B),D),C,E)	$C_{ABD} = (15, 18{,}33)$	$I_{ABD} = 266{,}66$
(((A,B),E),C,D)	$C_{ABE} = (16{,}66, 11{,}66)$	$I_{ABE} = 383{,}33$
((A,B),(C,D),E)	$C_{AB} = (10, 15), C_{CD} = (22{,}5, 20)$	$I_{CD} = 62{,}5$
((A,B),(C,E),D)	$C_{AB} = (10, 15), C_{CE} = (25, 10)$	$I_{CE} = 100$
((A,B),C,(D,E))	$C_{AB} = (10, 15), C_{DE} = (27{,}5, 15)$	$I_{DE} = 212{,}5$

   En este nivel se unirán los clusters C y D dado que su correspondiente incremento es mínimo.

3. Nivel 3. Los posibles agrupamientos son:

Agrupamientos	Centroides	Incrementos
(((A,B),(C,D)),E)	$C_{ABCD} = (16{,}25, 17{,}5)$	$I_{ABCD} = 116{,}66$
(((A,B),E),(C,D))	$C_{ABE} = (16{,}66, 11{,}66), C_{CD} = (22{,}5, 20)$	$I_{ABE} = 382{,}92$
((A,B),(C,D,E))	$C_{AB} = (10, 15), C_{CDE} = (25, 15)$	$I_{CDE} = 250$

   Y, por lo tanto, se fusionarán los clusters (A,B) y (C,D).

4. Finalmente, en el nivel 4, se unen los dos clusters que quedaron y el incremento es $I_{ABCDE} = 116{,}66$.

La Figura 9 muestra el dendrograma correspondiente al método de Ward. Es interesante destacar que la secuencia de clusters que se obtiene con este criterio es diferente de la que obtuvimos al aplicar el método del centroide y el de la mediana. Comparemos la técnica de Ward con el algoritmo del centroide. En el primer nivel, las dos estrategias comienzan uniendo A y B, pero difieren a partir del segundo nivel. En el caso del método del centroide en ese nivel se fusionan el cluster (A,B) con el C, mientras que con el de Ward se unen C y D (Ver Figura 10).

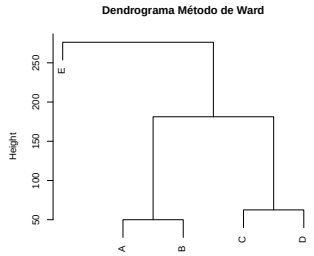

Figura 9: *Dendrograma para el ejemplo del método de Ward.*

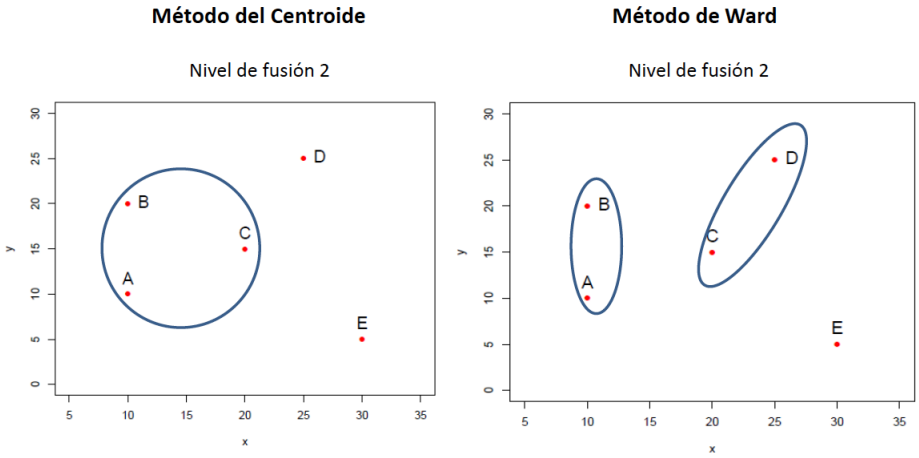

Figura 10: *Diferencia en el nivel 2 de fusión al aplicar el método del centroide y el de Ward en el ejemplo propuesto.*

1.3. Métodos Particionantes

Los métodos no jerárquicos están diseñados para agrupar objetos en un número predefinido de k clusters. Los métodos particionantes comienzan de una partición inicial de los objetos en k grupos, o bien, de un conjunto inicial de k puntos semillas que forman los núcleos de los clusters. La elección del conjunto de partida debería estar libre de sesgos, por lo que una opción es seleccionar aleatoriamente puntos semillas entre los objetos o particionar aleatoriamente los n objetos en k grupos. En cada paso se minimiza un criterio dado reubicando elementos entre los distintos clusters hasta que se alcanza una partición "óptima". En los algoritmos iterativos básicos, como el de k-medias o k-medianas la convergencia es local y la solución óptima global no está garantizada. Como no es necesario calcular una matriz de distancias (o similaridades) inicial y los datos no deben ser almacenados durante el procesamiento, los métodos no jerárquicos pueden aplicarse a conjuntos muy grandes de datos.

Discutiremos a continuación el método particionante más comúnmente usado.

1.3.1. *Método de k-medias*

El procedimiento tiene como objetivo minimizar la suma de cuadrados dentro de clusters, lo que equivale a asignar cada elemento al cluster en el que la distancia a la media del cluster es mínima. Como input se requiere el número k de clusters y la matriz de datos. El algoritmo consiste de los siguientes pasos:

1. Particionar los objetos en k clusters iniciales.
2. Reasignar cada ítem de la lista de a uno por vez al cluster de cuyo centroide (media) se encuentre más cerca. Recalcular el centroide del cluster que recibe el ítem y del cluster que pierde el ítem.
3. Repetir el paso 2 hasta que no haya más reasignaciones.

En vez de empezar con k clusters podemos especificar k centroides iniciales (puntos semilla) y comenzar a partir del paso 2. El agrupamiento final será dependiente, en general, de la partición inicial o de la selección inicial de semillas. La experiencia muestra que las reasignaciones más importantes ocurren en los primeros pasos.

Las ventajas de este tipo de métodos son la rapidez y la validez de su aplicación en grandes bases de datos. Una de las desventajas es que converge a un óptimo local, no global. Además la clusterización final depende de los centros iniciales. Y otra desventaja es que requiere fijar el número de clusters previamente, este número en general no se conoce. Hay tres argumentos fuertes para no fijar de antemano el número k de clusters:

1. si dos o más puntos semillas yacen dentro de un mismo cluster natural, entonces, los clusters resultantes tendrán poca definición.
2. la existencia de un outlier podría producir al menos un grupo con ítems muy dispersos.
3. aún si se supiera que la población consiste de k grupos, el método de muestreo podría ser tal que grupos raros no aparezcan en la muestra.

1.4. Comparación de los distintos procedimientos

Los distintos procedimientos que hemos descripto capturan distintas características de los datos. Milligan (1980) ([10]) resume la performance de cada método y el tipo de clusters que tiende a producir.

Single linkage no impone restricciones en la forma de grupos y sacrifica rendimiento en la recuperación de clusters compactos a cambio de la capacidad de detectar agrupaciones alargadas e irregulares. Además Single linkage tiende a cortar las colas de las distribuciones antes de la separación de los principales grupos (Hartigan, 1981) ([6]). Aunque tiene muchas propiedades teóricas deseables ha fallado en los estudios de Monte Carlo. Complete linkage está fuertemente sesgado hacia la producción de clusters con diámetros similares, y puede verse severamente distorsionado por valores atípicos moderados. Average linkage tiende a unir grupos con pequeña varianza, y es ligeramente sesgado hacia la producción de clusters con la misma varianza.

El método del centroide es más robusto a los valores extremos (outliers) que la mayoría de otros métodos jerárquicos pero en otros aspectos puede no funcionar tan bien como el método de Ward o Average linkage. El método de Ward tiende a unir grupos con un pequeño número de observaciones, y está fuertemente sesgado hacia la producción de clusters con aproximadamente el mismo número de observaciones. También es muy sensible a valores atípicos.

2. Propuestas para seleccionar el número de clusters

En la segmentación jerárquica, como ya aclaramos, no se requiere a priori la cantidad de agrupamientos en el conjunto de datos a analizar. Ésto es consistente con el hecho de que no se conoce fácticamente acerca de la existencia de grupos, y en caso de existir, cuál es la cantidad de grupos. Muchas veces no interesa la jerarquía completa, sino sólo un subconjunto de particiones obtenidas a partir del dendrograma resultante. Es decir, para el estudio en cuestión puede alcanzar con hacer un recorte del dendrograma o elegir qué ramas se consideran más representativas de acuerdo a los conocimientos y experiencia de la persona que realiza el trabajo. Otras veces, el problema a resolver es la determinación de la cantidad de grupos, problema para el cual no existe una única solución.

Si quisiéramos calcular para un conjunto A de n elementos la cantidad de particiones en k subconjuntos, teniendo en cuenta la definición de *partición* dada en la sección 1.2 de este capítulo, deberíamos recurrir a la fórmula de *los números de Stirling de segunda especie S(n,k)*:

$$S(n,\, k) \quad = \quad \frac{1}{k!}\sum_{j=1}^{k}(-1)^{k-j}\binom{k}{j}j^n.$$

Los números de Stirling de segunda especie se definen como la cantidad de maneras distintas que existen de generar una partición de un conjunto de n elementos en k subconjuntos. Luego podríamos calcular para n fijo y para $k \in \mathbb{N}_{\leq n}$ el número total de particiones, y seleccionar la mejor. Ese número es conocido como *número de Bell,* y se calcula como $\sum_{k=1}^{n} S(n,\, k)$, es decir, es el número de subconjuntos no vacíos que se pueden formar con n elementos. Sin embargo, esta manera exhaustiva resulta impracticable si n es muy grande porque, para cada k, $S(n,\, k)$ se hace mucho más grande aún. De modo que no es conveniente en la práctica intentar buscar la cantidad óptima de clusters. De hecho, se ha demostrado en [13] que $S(n,\, k) \geq \frac{1}{2}(k^2 + k + 2)k^{n-k-1} - 1$, por lo que considerando $n = 100$ y $k = 10$ se obtiene que $S(100,\, 10) \geq 56\,10^{89} - 1$. Esta cota inferior es superior al número estimado de átomos en todo el universo!

Para enfrentar esta situación han surgido algunas propuestas. En muchos casos, en las ciencias aplicadas el objeto de estudio es entender algunos patrones generales, y para ello basta con observar el dendrograma y definir en base a conocimiento experto el número de grupos. En casos más generales se pueden emplear métodos heurísticos o reglas basadas en tests de hipótesis formales. En la sección 3 de este capítulo, presentamos una de las propuestas: el estadístico gap. Aquí podemos enumerar otros procedimientos.

Primero, la técnica de recortar el dendrograma que si bien es subjetiva y depende de la opinión del investigador, resulta sencilla y práctica para la persona que tiene suficiente conocimiento sobre el fenómeno y no requiere conocer el número exacto de clusters. Otro método heurístico pero más formal es representar gráficamente la cantidad de grupos en cada nivel del dendrograma frente a los niveles de fusión a los que los clusters se unen. Los saltos o discontinuidades en los niveles de fusión sugerirán si la nueva unión de clusters aporta o no información a la unión anterior y así hasta definir un número apropiado. Otro procedimiento también basado en los saltos relativos en los valores de fusión fue el que propuso Mojena en 1977 ([12]). En este caso se compara el valor de fusión de cada etapa con el promedio de los valores de fusión sumado con el producto de una cierta constante por la cuasidesviación típica de los valores de fusión. Cuando un valor de fusión supera dicha cantidad se concluye que el nivel precedente es el que origina la solución óptima. Mojena sugirió que el valor de la constante debía de estar comprendido en el rango de 2.75 a 3.50; si bien Milligan (1985, ver [11]) tras una detallada investigación de valores de fusión, en función del número de clusters, establece que el valor óptimo para dicha constante debe ser 1.25. Una cuarta propuesta de Beale (1969, ver [1]) utiliza un test basado en la distribución *F de Snedecor* para testear la hipótesis de la existencia de k_2 clusters frente a la existencia de k_1 clusters, siendo $k_2 > k_1$. Para ello se consideran la suma, para cada partición, de las desviaciones cuadráticas medias de los elementos de cada cluster a su centroide, llamémoslas DC_1 y DC_2

$$DC_1 = \frac{1}{n-k_1} \sum_{i=1}^{k_1}\sum_{j=1}^{n_i} \|\mathbf{x}_{ij} - \mathbf{m}_i\|^2$$

$$DC_2 = \frac{1}{n-k_2} \sum_{i=1}^{k_2}\sum_{j=1}^{n_i} \|\mathbf{x}_{ij} - \mathbf{m}_i\|^2$$

donde se ha supuesto que el cluster $i-$ésimo posee n_i elementos y n es el total de la muestra. El estadístico

$$F = \frac{\frac{DC_1 - DC_2}{DC_2}}{\left[\frac{n-k_1}{n-k_2}\left(\frac{k_2}{k_1}\right)^{\frac{2}{p}} - 1\right]},$$

tiene distribución aproximada *F de Snedecor* $\mathcal{F}_{p(k_2-k_1),p(n-k_2)}$ cuando el incremento en la cantidad de clusters de k_1 a k_2 no provee una mejora significativa de la segmentación.

Por otra parte hay métodos que usan las nociones de sumas de cuadrados de procedimientos del análisis multivariado paramétrico. Sean las siguientes matrices

$$\mathbf{T} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \mathbf{m})(\mathbf{x}_{ij} - \mathbf{m})'$$

$$\mathbf{W} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \mathbf{m}_i)(\mathbf{x}_{ij} - \mathbf{m}_i)'$$

$$\mathbf{B} = \sum_{i=1}^{k} n_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})'$$

donde $\mathbf{T}$ es una matriz que contiene las sumas de cuadrados de todos los individuos respecto de su centroide, $\mathbf{W}$ es la matriz que contiene las sumas de cuadrados dentro de los grupos y $\mathbf{B}$ entre los grupos (suma de cuadrados entre clusters), siendo k el número total de clusters y n el tamaño total de la muestra de manera que $n = n_1 + \cdots + n_k$. Notemos que vale la igualdad $\mathbf{T} = \mathbf{W} + \mathbf{B}$. Dicha igualdad es la extensión al caso multivariado de la conocida descomposición de la suma de cuadrados del análisis de la varianza de un factor.

Varios criterios para determinar el número de clusters se basan en elegir el valor de k que minimice una medida de la varianza intra-clusters o, equivalentemente, maximice una medida de la varianza entre-clusters. Por ejemplo, las dos propuestas que se detallan a continuación:

- Marriot (1971) ([9]) sugiere seleccionar el k que minimice el estadístico $k^2|\mathbf{W}|$, es decir usa como medida de variabilidad total el determinante de la matriz $\mathbf{W}$.
- Calinski y Harabasz (1974) ([3]) proponen definir el número óptimo de clusters como el valor k que maximiza el estadístico usando como noción de variabilidad las trazas de la matriz de covarianzas.

$$CH = \frac{\frac{tr[\mathbf{B}]}{k-1}}{\frac{tr[\mathbf{W}]}{n-k}}.$$

3. Método gap

Tibshirani, Walther y Hastie (2001) propusieron, en [17], un método para estimar el número óptimo de clusters. La técnica se puede emplear luego de aplicar cualquier algoritmo de segmentación. En esta sección resumiremos el procedimiento propuesto para encontrar la cantidad de grupos en una población, conocido como "estadístico gap".

3.1. El estadístico gap

Supongamos que tenemos n observaciones independientes, cada una con p características. Sea $\mathbf{x}_i$ con elementos (x_{ij}), $i = 1, 2, ..., n$ y $j = 1, 2, ..., p$ y sea $d(\mathbf{x}_i, \mathbf{x}_{i'})$ la distancia entre las observaciones $\mathbf{x}_i$ y $\mathbf{x}_{i'}$. Consideremos como distancia el cuadrado de la distancia euclídea, es decir

$$d(\mathbf{x}_i, \mathbf{x}_{i'}) \quad = \quad \sum_j (x_{ij} - x_{i'j})^2.$$

Supongamos que como resultado de un análisis hemos obtenido k clusters, $C_1, C_2, ..., C_k$; donde C_r denota el conjunto de índices de las observaciones en el cluster r y $n_r = |C_r|$ es la cantidad de elementos del conjunto C_r. Sea

$$D_r = \sum_{i,i' \epsilon C_r} d(\mathbf{x}_i, \mathbf{x}_{i'}) \tag{2}$$

la suma de las distancias entre los elementos del cluster r tomados de a pares (notemos que se cuenta dos veces cada distancia) y sea

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r.$$

W_k es función de la cantidad de clusters k y mide la dispersión dentro de los clusters. Observemos que esta medida, bajo cualquier método razonable de segmentación, es monótona decreciente a medida que k crece, y se va "achatando" acercándose cada vez más hacia el valor 0, lo cual es razonable ya que la varianza intra-cluster va disminuyendo cuando aumenta la cantidad de grupos. Cuando d es la distancia Euclídea, W_k es la suma de cuadrados dentro de clusters; una medida "pooled" de la dispersión de las observaciones, dentro de los clusters.

La idea de la propuesta es comparar, para cada k, el valor de $log(W_k)$ con su esperanza bajo una distribución nula de referencia en la que asumimos que no hay clusters. Se define

$$Gap_n(k) = E_n^* \{log(W_k)\} - log(W_k)$$

donde E_n^* es la esperanza para muestras de tamaño n extraídas de la distribución de referencia. Un primer estimador del número de clusters en la distribución de la que proviene la muestra es el valor $\hat{k}$, que maximiza $Gap_n(k)$ después de tener en cuenta la distribución de referencia.

El problema entonces consiste en proponer una distribución de referencia apropiada y construir la distribución de muestreo del estadístico gap.

3.2. La distribución de referencia

Asumimos un modelo nulo en el que no existen clusters ($k = 1$) y queremos rechazarlo en favor de un modelo de k componentes o clusters ($k > 1$) si para algún $k > 1$ existe suficiente evidencia.

Tibshirani y otros (2001) proponen modelar las distribuciones de un solo cluster como densidades log-cóncavas en vez de usar densidades unimodales y denotan S^p el conjunto de tales distribuciones en $\mathbb{R}^p$ con un único cluster o componente. Ejemplos de estas distribuciones son la normal y la uniforme con soporte convexo.

Para mostrar cómo buscar una adecuada distribución de referencia, los autores consideran la versión poblacional del estadístico $Gap_n(k)$ para el caso del algoritmo de k-medias:

$$g(k) \quad = \quad log\left\{\frac{MSE_{X^*}(k)}{MSE_{X^*}(1)}\right\} - log\left\{\frac{MSE_X(k)}{MSE_X(1)}\right\},$$

donde

$$MSE_X(k) \quad = \quad E(min_{\mu \epsilon A_k} \|X-\mu\|^2),$$

con $A_k \subset \mathbb{R}^p$ el conjunto de k puntos que minimiza $MSE_X(k)$. Notar que los autores dividen por $MSE(1)$ en ambos términos para asegurarse que $g(1) = 0$. Con ésto lo que se busca es la distribución de referencia de un único cluster para X^* menos favorable, tal que $g(k) \leq 0$ para todo $X \in S^p$ y todo $k \geq 1$. Es decir, interesa hallar la distribución unimodal con mayor probabilidad de producir grupos espurios cuando se usa el estadístico gap. Tibshirani, Walther y Hastie (2001) mostraron que para el caso univariado, dicha distribución es la uniforme $U[0,1]$.

En el caso de dimensión $p > 1$ los mismos autores probaron que no existe distribución $U \in S^p$ tal que

$$\underset{X \in S^p}{inf}\left\{\frac{MSE_X(k)}{MSE_X(1)}\right\} \quad = \quad \left\{\frac{MSE_U(k)}{MSE_U(1)}\right\},$$

a menos que su soporte sea degenerado y concluyeron que no es posible para el caso multivariado elegir una distribución de referencia que resulte aplicable en general ya que la geometría de la distribución nula importa.

3.3. Cálculo del estadístico gap

Tibshirani, Walther y Hastie (2001) proponen dos maneras de elegir la distribución de referencia:

1. Generar distribuciones uniformes dentro del rango de valores observados para cada una de las p variables;
2. Generar distribuciones uniformes a partir de las componentes principales de los datos. Sea $\mathbf{X} \in \mathbb{R}^{nxp}$, asumiendo que $\mathrm{E}(\mathbf{X}) = 0$, en otro caso se resta a $\mathbf{X}$ la media para centrar los datos, y calculando la descomposición en valores singulares $\mathbf{X} = \mathbf{UDV}'$ donde $\mathbf{U}$ es una matriz de dimensión $n \times n$, $\mathbf{D}$ es una matriz rectangular diagonal de dimensión $n \times p$ con números no negativos en la diagonal, denominados valores singulares de $\mathbf{X}$, y $\mathbf{V}'$ es una matriz de dimensión $p \times p$. Las n columnas de $\mathbf{U}$ y las p columnas de $\mathbf{V}$ se denominan vectores singulares izquierdos y derechos de $\mathbf{X}$. La descomposición en valores singulares está relacionada con la descomposición en autovectores y autovalores; los vectores singulares izquierdos y derechos de $\mathbf{X}$ son los autovectores de $\mathbf{XX}'$ y $\mathbf{X}'\mathbf{X}$ respectivamente. Transformamos $\mathbf{X}^{\#} = \mathbf{XV}$ y seleccionamos datos aleatorios $\mathbf{Z}^{\#}$ a partir de uniformes sobre los rangos de columnas de $\mathbf{X}^{\#}$. Finalmente retransformamos $\mathbf{Z} = \mathbf{Z}^{\#}\mathbf{V}'$ para obtener una muestra de datos de la distribución de referencia.

El primer método es más simple, el segundo tiene en cuenta la forma de la distribución empírica.

En cada caso se extraen B muestras aleatorias de n observaciones de la distribución de referencia, se calcula $log(W_k^*)$ en cada muestra y se estima $E_n^*\{log(W_k)\}$ promediando sobre las B repeticiones. Luego, nos interesa la distribución de muestreo del estadístico gap. Sea $sd(k)$ el desvío estándar de $log(W_k^*)$ en las B copias, y considerando el error en la simulación, el desvío estándar de $E_n^*\{log(W_k)\}$ resulta ser $s_k = \sqrt{1+1/B}sd(k)$.

Tibshirani y otros (2001) proponen como criterio elegir $\hat{k}$ como el más pequeño de los k que verifique que $Gap(k) \geq Gap(k+1) - s_{k+1}$ y prueban a través de simulaciones que el criterio funciona razonablemente bien.

3.3.1. Ejemplos

A continuación se presenta un análisis en el que datos bidimensionales provienen de una distribución en la que no existen grupos. La Figura 11 presenta la muestra de $n = 100$ datos observados. La muestra original fue segmentada usando el procedimiento de Ward y se calculó $log(W_k)$ para valores de $k = 1, .., 10$. La Figura 12 muestra el gráfico de $log(W_k)$ en función de k. Finalmente, se generaron $B = 1000$ muestras usando como distribución de referencia el soporte definido por la transformación con vectores singulares, es decir, a partir de las componentes principales de los datos. En cada muestra se repitió la secuencia de pasos utilizados en la muestra original de datos. La Figura 13 muestra las curvas $log(W_k^*)$ resultantes de las 1000 repeticiones y su promedio calculado a partir de las B muestras aleatorias. Finalmente la Figura 14 muestra los valores del estadístico $Gap_n(k)$ en función de k. Claramente, el criterio de Tibshirani y otros, estima el número de clusters en la población como $\hat{k} = 1$ ya que es el menor k que cumple el criterio $Gap(k) \geq Gap(k+1) - s_{k+1}$.

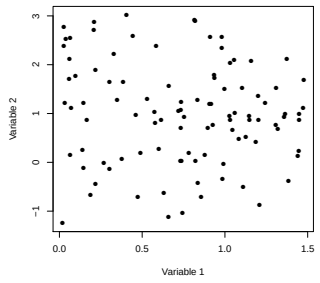

Figura 11: *Datos sin estructura de grupos.*

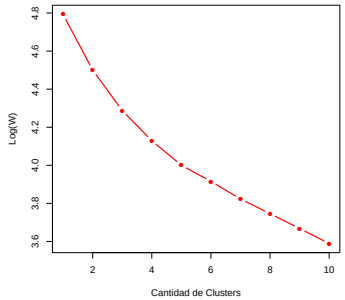

Figura 12: *Log(W) para ejemplo sin estructura en los datos.*

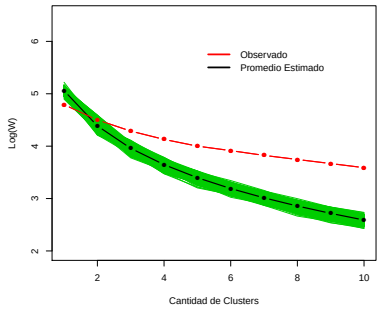

Figura 13: *Logaritmo de W observadas y promedio de logaritmo de W^* estimadas con distribución de referencia para el caso sin estructura de grupos en los datos.*

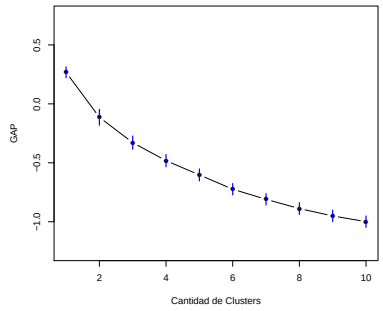

Figura 14: *Curva gap con segmentos de desvíos (en azul) para el ejemplo sin estructura de grupos.*

Consideremos ahora un conjunto de 200 datos bidimensionales en el que existen dos agrupamientos bien distinguibles (Figura 15). La Figura 16 muestra el dendrograma obtenido después de aplicar el método de Ward, donde se puede apreciar el notable salto relativo en la altura o "height", que corresponde al incremento en la suma de los cuadrados dentro de clusters, cuando se agrupan los dos últimos clusters. Finalmente, la Figura 17 muestra el comportamiento del estadístico W_k, suma de los cuadrados dentro de los clusters, como función del número de clusters, por simplicidad sólo se representa $k \leq 10$. Nótese la fuerte caída en W_k para $k = 2$.

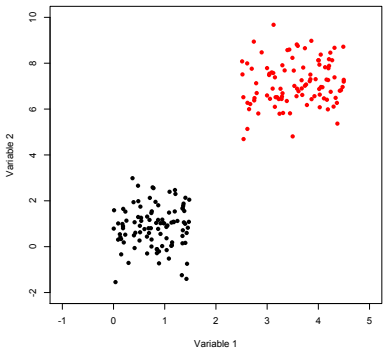

Figura 15: *Ejemplo de datos con dos grupos.*

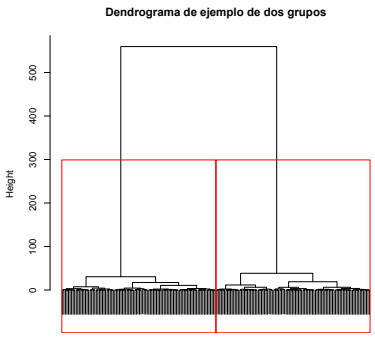

Figura 16: *Dendrograma para el ejemplo con dos grupos.*

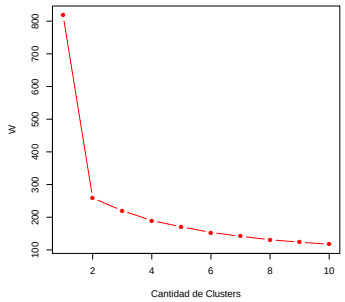

Figura 17: *W para la segmentación de dos grupos.*

En el gráfico de la Figura 18 se muestran, para $1 \leq k \leq 10$, la curva $log(W)$ (en color rojo), las curvas $log(W^*)$ (en color verde), con W^* estimadas a partir de la distribución de referencia, y su promedio (en color negro). Mientras que la Figura 19 muestra la curva gap para este ejemplo de dos grupos, y, en azul, los segmentos verticales de desvíos para cada valor de k entre 1 y 10. Se observa que $k = 2$ es el mínimo k que satisface la condición $Gap(k) \geq Gap(k+1) - s_{k+1}$.

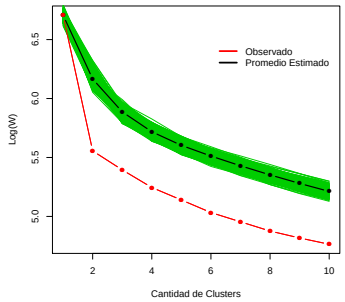

Figura 18: *Logaritmo de W observadas y media de logaritmos de W^* estimadas con distribución de referencia.*

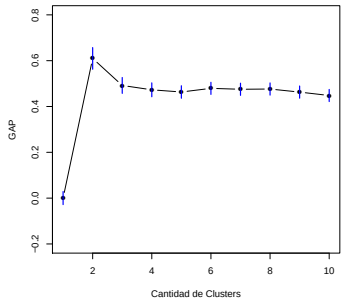

Figura 19: *Curva gap con segmentos de desvíos (en azul).*

3.4. Cálculo del estadístico W_k bajo el método de segmentación de Ward

En la subsección 1.2.5 de este capítulo, hemos presentado el método de segmentación de Ward. Recordemos que en este procedimiento se fusiona el par de clusters que minimiza el incremento en la suma de los cuadrados dentro de clusters,

entre todas las posibles particiones que se obtienen uniendo un par de grupos del paso previo. Este incremento se calcula como

$$I_{C_i,C_j} = \sum_{l\in C_i\cup C_j} \|\mathbf{x}_l - \mathbf{m}\|^2 - \left\{ \sum_{l\in C_i} \|\mathbf{x}_l - \mathbf{m}_i\|^2 + \sum_{l\in C_j} \|\mathbf{x}_l - \mathbf{m}_j\|^2 \right\},$$

siendo C_i y C_j dos clusters con n_i y n_j elementos y medias $\mathbf{m}_i$ y $\mathbf{m}_j$ respectivamente y $\mathbf{m}$ la media del nuevo cluster que se formaría al unir C_i y C_j.

A continuación demostramos que existe una relación directa entre la función objetivo a minimizar en el método de Ward y la noción de variabilidad utilizada por el método gap. En primer lugar se presentan dos lemas con resultados que se usarán para demostrar el teorema.

Lema 1: Sean $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^p$ y sea $\mathbf{m} = \dfrac{\sum_{l=1}^{n} \mathbf{x}_l}{n}$, entonces

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \|\mathbf{x}_i - \mathbf{x}_j\|^2 \quad = \quad 2n\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{m}\|^2 .$$

Demostración: sabemos que

$$\mathbf{x}_i - \mathbf{m} = \frac{n\mathbf{x}_i - \sum_{j=1}^{n} \mathbf{x}_j}{n} = \frac{1}{n}\sum_{j=1}^{n}(\mathbf{x}_i - \mathbf{x}_j). \tag{3}$$

Además,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \|\mathbf{x}_i - \mathbf{x}_j\|^2 \quad = \quad \sum_{i=1}^{n}\sum_{j=1}^{n} (\mathbf{x}_i - \mathbf{x}_j)' (\mathbf{x}_i - \mathbf{x}_j) =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \left\{ \mathbf{x}_i' (\mathbf{x}_i - \mathbf{x}_j) - \mathbf{x}_j' (\mathbf{x}_i - \mathbf{x}_j) \right\} \quad = \quad \sum_{i=1}^{n}\sum_{j=1}^{n} \left\{ \mathbf{x}_i' (\mathbf{x}_i - \mathbf{x}_j) + \mathbf{x}_j' (\mathbf{x}_j - \mathbf{x}_i) \right\} =$$

$$= 2\sum_{i=1}^{n}\sum_{j=1}^{n} \left\{ \mathbf{x}_i' (\mathbf{x}_i - \mathbf{x}_j) \right\} = 2\sum_{i=1}^{n} \mathbf{x}_\mathbf{i}' \sum_{j=1}^{n} (\mathbf{x}_i - \mathbf{x}_j) .$$

De (3), se sigue que

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \|\mathbf{x}_i - \mathbf{x}_j\|^2 = 2n\sum_{i=1}^{n} \mathbf{x}_i' (\mathbf{x}_i - \mathbf{m})$$

Por otra parte,

$$\sum_{i=1}^{n} \mathbf{m}' (\mathbf{x}_i - \mathbf{m}) = \quad \mathbf{m}' \textstyle\sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{m}) = 0,$$

por lo tanto

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \|\mathbf{x}_i - \mathbf{x}_j\|^2 = 2n\sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{m})' (\mathbf{x}_i - \mathbf{m}) = \quad 2n \textstyle\sum_{i=1}^{n} \quad \|\mathbf{x}_i - \mathbf{m}\|^2$$

que es lo que queríamos demostrar. ■

Es interesante notar que la noción de variabilidad utilizada en el método Ward es equivalente al concepto de "variación total" de un conjunto de vectores multivariados, definido por la traza de la matriz de covarianza muestral S, lo que se prueba en el siguiente lema.

Lema 2: Sean $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^p$, $\mathbf{m} = \frac{\sum_{l=1}^{n} \mathbf{x}_l}{n}$ y $S = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})'$ la matriz de covarianza muestral, entonces,

$$\frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{m}\|^2 = Tr(S).$$

Demostración:

$$\frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{m}\|^2 = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{m})'(\mathbf{x}_i - \mathbf{m}) = \frac{1}{n}\sum_{i=1}^{n} Tr\left[(\mathbf{x}_i - \mathbf{m})'(\mathbf{x}_i - \mathbf{m})\right] =$$

$$= \frac{1}{n}\sum_{i=1}^{n} Tr\left[(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})'\right] = \frac{1}{n} Tr\left[\sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})'\right] = \frac{1}{n} Tr(nS) = Tr(S). \quad ■$$

A continuación, mostraremos la relación entre el estadístico gap y el método de Ward, mostrando en el siguiente teorema que el incremento del método de Ward se puede obtener a partir de los estadísticos W_k y W_{k+1}.

Teorema: Sea I_{C_i,C_j} la función objetivo a minimizar en el método de Ward, en la etapa en que se pasa de $k+1$ a k grupos cuando se fusiona C_i con C_j

$$I_{C_i,C_j} = \sum_{l \in C_i \cup C_j} \|\mathbf{x}_l - \mathbf{m}\|^2 - \left\{ \sum_{l \in C_i} \|\mathbf{x}_l - \mathbf{m}_i\|^2 + \sum_{l \in C_j} \|\mathbf{x}_l - \mathbf{m}_j\|^2 \right\}.$$

Sea $W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r$ con D_r definido en (2) entonces

$$I_{C_i,C_j} = W_k - W_{k+1}.$$

Demostración: por el Lema 1 vale que

$$\sum_{l \in C_i \cup C_j} \|\mathbf{x}_l - \mathbf{m}\|^2 = \frac{1}{2(n_i + n_j)} \sum_{l \in C_i \cup C_j} \sum_{l' \in C_i \cup C_j} \|\mathbf{x}_l - \mathbf{x}_{l'}\|^2,$$

$$\sum_{l \in C_i} \|\mathbf{x}_l - \mathbf{m}_i\|^2 = \frac{1}{2n_i} \sum_{l \in C_i} \sum_{l' \in C_i} \|\mathbf{x}_l - \mathbf{x}_{l'}\|^2 \quad \text{y}$$

$$\sum_{l \in C_j} \|\mathbf{x}_l - \mathbf{m}_j\|^2 = \frac{1}{2n_j} \sum_{l \in C_j} \sum_{l' \in C_j} \|\mathbf{x}_l - \mathbf{x}_{l'}\|^2$$

donde $\mathbf{m}_i$, $\mathbf{m}_j$ y $\mathbf{m}$ son las medias de los clusters C_i, C_j y $C_i \cup C_j$ respectivamente.

Luego

$$I_{C_i,C_j} = \frac{1}{2(n_i+n_j)} \sum_{l\in C_i\cup C_j} \sum_{l'\in C_i\cup C_j} \|\mathbf{x}_l - \mathbf{x}_{l'}\|^2 - \frac{1}{2n_i} \sum_{l\in C_i} \sum_{l'\in C_i} \|\mathbf{x}_l - \mathbf{x}_{l'}\|^2 - \frac{1}{2n_j} \sum_{l\in C_j} \sum_{l'\in C_j} \|\mathbf{x}_l - \mathbf{x}_{l'}\|^2. \tag{4}$$

En el paso $n-(k+1)$ se unirán dos clusters de modo que en el nivel siguiente habrá k. Sin pérdida de generalidad, podemos suponer que se unen C_i y C_j, y que esa fusión genera el cluster C_h en el nivel de fusión $n-k$.

Consideremos ahora las siguientes definiciones. Sea C_α^k el α-ésimo cluster en el nivel de fusión $n-k$, es decir en una partición que contiene exactamente k clusters, y sea n_α^k el número de objetos en el cluster C_α^k. Definimos a D_α^k como la suma de las distancias al cuadrado de los elementos del cluster C_α^k. Del mismo modo definimos n_α^{k+1} y D_α^{k+1} en el nivel de fusión previo $(n-(k+1))$ con $k+1$ grupos.

Por lo tanto, el primer término de (4) se puede reescribir como

$$\frac{1}{2(n_i+n_j)} \sum_{l\in C_i\cup C_j} \sum_{l'\in C_i\cup C_j} \|\mathbf{x}_l - \mathbf{x}_{l'}\|^2 = \frac{1}{2n_h^k} \sum_{l\in C_h^k} \sum_{l'\in C_h^k} \|\mathbf{x}_l - \mathbf{x}_{l'}\|^2 = \frac{1}{2n_h^k} D_h^k. \tag{5}$$

Mientras que el segundo y tercer término de (4) serán

$$-\frac{1}{2n_i} \sum_{l\in C_i} \sum_{l'\in C_i} \|\mathbf{x}_l - \mathbf{x}_{l'}\|^2 - \frac{1}{2n_j} \sum_{l\in C_j} \sum_{l'\in C_j} \|\mathbf{x}_l - \mathbf{x}_{l'}\|^2 = -\frac{1}{2n_i^{k+1}} \sum_{l\in C_i^{k+1}} \sum_{l'\in C_i^{k+1}} \|\mathbf{x}_l - \mathbf{x}_{l'}\|^2 - \frac{1}{2n_j^{k+1}} \sum_{l\in C_j^{k+1}} \sum_{l'\in C_j^{k+1}} \|\mathbf{x}_l - \mathbf{x}_{l'}\|^2 =$$

$$= -\frac{1}{2n_i^{k+1}} D_i^{k+1} - \frac{1}{2n_j^{k+1}} D_j^{k+1}. \tag{6}$$

Notemos además que por la estructura jerárquica y anidada de las particiones, para cada cluster, en el nivel de fusión $n-k$, C_α^k con $\alpha \neq h$, $\exists\ \alpha'$ tal que $C_\alpha^k = C_{\alpha'}^{k+1}$ con $\alpha' \neq i,\ j$, es decir existe un cluster igual en el nivel de fusión anterior $n-(k+1)$. Análogamente, para cada D_α^k con $\alpha \neq h$, $\exists\ \alpha'$ tal que $D_\alpha^k = D_{\alpha'}^{k+1}$ con $\alpha' \neq i,\ j$. Por lo tanto,

$$\sum_{\alpha\neq i,j} \frac{1}{2n_\alpha^{k+1}} D_\alpha^{k+1} \quad = \quad \sum_{\alpha\neq h} \frac{1}{2n_\alpha^k} D_\alpha^k.$$

Observemos ahora que si sumamos $\sum_{\alpha\neq h} \frac{1}{2n_\alpha^k} D_\alpha^k$ en (5) obtenemos

$$\frac{1}{2n_h^k} D_h^k + \quad \sum_{\alpha\neq h} \tfrac{1}{2n_\alpha^k} D_\alpha^k = \sum_{s=1}^k \tfrac{1}{2n_{\alpha_s}^k} D_{\alpha_s}^k = \quad W_k.$$

Por otro lado si restamos la misma expresión en (6), pero en términos de D_α^{k+1}, es decir, restamos $\sum_{\alpha\neq i,j} \frac{1}{2n_\alpha^{k+1}} D_\alpha^{k+1}$

$$-\tfrac{1}{2n_i^{k+1}} D_i^{k+1} - \tfrac{1}{2n_j^{k+1}} D_j^{k+1} - \quad \sum_{\alpha\neq i,j} \frac{1}{2n_\alpha^{k+1}} D_\alpha^{k+1} = -\sum_{s=1}^{k+1} \frac{1}{2n_{\alpha_s}^{k+1}} D_{\alpha_s}^{k+1} = -W_{k+1}.$$

Luego, sumando (5) y (6) recuperamos la igualdad (4), que es el incremento de Ward. Es decir,

$$I_{C_i,C_j} = W_k - W_{k+1}. \qquad \blacksquare$$

De esta manera, logramos demostrar que el incremento en la suma de las distancias al cuadrado (distancias de cada observación respecto al centroide del grupo) derivado del método de Ward, es igual al cambio en el estadístico gap $(W_k - W_{k+1})$ con W_k y W_{k+1} definidos en Tibshirani y otros (2001). Luego W_k es la suma acumulada de los incrementos desde el nivel de fusión 0 hasta el nivel $n-k$.

Notemos que tal como se demostró en el lema de la sección 1.2.5 el incremento $I_{C_i,C_j} = W_k - W_{k+1}$ es un número positivo $\left(\frac{n_i n_j}{n_i+n_j} \|\mathbf{m}_i - \mathbf{m}_j\|^2\right)$ por lo que deducimos que W_k es una función monótona estrictamente decreciente. Es interesante además destacar que a la hora de buscar el k óptimo no tiene sentido minimizar W_k, pues en todos los casos $k = n$ sería el óptimo, pero sí tiene sentido minimizar la diferencia, o el incremento, que es exactamente lo que plantea la estrategia de Ward. Por otro lado, al minimizar el incremento, lo que busca el método de Ward es minimizar la cantidad $n_i n_j$ (fijados los centroides y el total de objetos), por lo que este método tiende a unir clusters con menos elementos entre sí y producir clusters de aproximadamente el mismo número de elementos.

4. Componentes principales

El método de componentes principales se usa para explicar la estructura de varianza-covarianzas de una muestra a partir de unas pocas combinaciones lineales de las variables originales, denominadas *componentes principales*. Dado un conjunto de observaciones en $\mathbb{R}^p$ la transformación se realiza de modo tal que la primera componente principal representa la dirección con máxima varianza y cada componente siguiente tiene la máxima varianza posible bajo la restricción de ser ortogonal a todas las componentes anteriores. Geométricamente las componentes principales representan un nuevo sistema de coordenadas obtenido rotando el sistema de coordenadas original. Los nuevos ejes representan las direcciones de máxima variabilidad por lo que en general la mayor parte de la variabilidad de los datos originales es retenida por un número de componentes menor que p. En este sentido, el análisis de componentes principales se considera un método de reducción de dimensión.

4.1. Componentes principales poblacionales

Sea $X^{\intercal} = (X_1, X_2, ..., X_p)$ un vector aleatorio en $\mathbb{R}^p$, sea $\mathbf{\Sigma}$ su matriz de covarianza con autovalores $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. Consideremos las siguientes combinaciones lineales

$$\begin{aligned} Y_1 &= \mathbf{w}_1^{\intercal} X = w_{11}X_1 + w_{12}X_2 + ... + w_{1p}X_p \\ Y_2 &= \mathbf{w}_2^{\intercal} X = w_{21}X_1 + w_{22}X_2 + ... + w_{2p}X_p \\ &\vdots \\ Y_p &= \mathbf{w}_p^{\intercal} X = w_{p1}X_1 + w_{p2}X_2 + ... + w_{pp}X_p. \end{aligned}$$

Notemos que

$$Var\left(Y_i\right) = \mathbf{w}_i^{\intercal} \mathbf{\Sigma} \mathbf{w}_i \qquad i = 1, ..., p \tag{7}$$

$$Cov\left(Y_i, Y_j\right) = \mathbf{w}_i^{\intercal} \mathbf{\Sigma} \mathbf{w}_j \qquad i, j = 1, ..., p \tag{8}$$

Definimos las *componentes principales* de X como las combinaciones lineales Y_1, Y_2, ..., Y_p no correlacionadas, cuya varianza (7) es máxima, que satisfacen $\mathbf{w}_i^{\intercal}\mathbf{w}_i = 1$ $i = 1, 2, ...p$. La primera componente principal se elige como la combinación lineal $\mathbf{w}_1^{\intercal} X$, que maximice $Var\left(\mathbf{w}_1^{\intercal} X\right)$ sujeto a la restricción $\mathbf{w}_1^{\intercal}\mathbf{w}_1 = 1$. La segunda componente se elige como la combinación lineal $\mathbf{w}_2^{\intercal} X$, que maximice $Var\left(\mathbf{w}_2^{\intercal} X\right)$ sujeto a $\mathbf{w}_2^{\intercal}\mathbf{w}_2 = 1$ y $\mathbf{w}_1^{\intercal}\mathbf{w}_2 = 0$ o $Cov\left(\mathbf{w}_1^{\intercal} X, \mathbf{w}_2^{\intercal} X\right) = 0$. La $i-$ésima componente se elige como la combinación lineal $\mathbf{w}_i^{\intercal} X$, que maximice $Var\left(\mathbf{w}_i^{\intercal} X\right)$ sujeto a $\mathbf{w}_i^{\intercal}\mathbf{w}_i = 1$, $\mathbf{w}_i^{\intercal}\mathbf{w}_j = 0$ para $j < i$.

Resultado: Sea $\mathbf{\Sigma}$ la matriz de covarianza asociada al vector aleatorio $X^{\intercal} = (X_1, X_2, ..., X_p)$ y sean $(\mathbf{v}_i, \lambda_i)\,, i = 1,\ ...,\ p$ los pares de autovectores y autovalores asociados a $\mathbf{\Sigma}$ con $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$. La $i-$ésima componente principal de X está dada por

$$Y_i = \mathbf{v}_i^{\intercal} X = v_{i1}X_1 + v_{i2}X_2 + ... + v_{ip}X_p \qquad i = 1, 2, ..., p.$$

Bajo esta elección

$$Var\,(Y_i) = \mathbf{v}_i^{\intercal}\mathbf{\Sigma}\mathbf{v}_i = \lambda_i \qquad i = 1, ..., p$$

$$Cov\,(Y_i, Y_j) = \mathbf{v}_i^{\intercal}\mathbf{\Sigma}\mathbf{v}_j = 0 \qquad i \neq j.$$

A partir del resultado anterior puede demostrarse que

$$tr\,(\mathbf{\Sigma}) = \sum_{i=1}^{p} Var\,(X_i) = \sum_{i=1}^{p} Var\,(Y_i) = \sum_{i=1}^{p} \lambda_i.$$

En consecuencia, la contribución de la componente Y_i a la varianza total, definida como $tr\,(\mathbf{\Sigma})$, es λ_i y la proporción de variabilidad explicada por las m primeras componentes principales es $P_m = \frac{\lambda_1 + ... + \lambda_m}{\lambda_1 + ... + \lambda_p}$.

4.2. Componentes principales muestrales

Consideremos ahora datos $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ correspondientes a una muestra de n observaciones en p variables obtenida a partir de una distribución con media $\boldsymbol{\mu}$ y matriz de covarianzas $\mathbf{\Sigma}$. Estimamos $\boldsymbol{\mu}$ y $\mathbf{\Sigma}$ con el vector de medias $\bar{\mathbf{x}}$ y la matriz de covarianzas muestral $\mathbf{S}$. Las componentes principales muestrales son los autovectores de la matriz $\mathbf{S}$, y la contribución de cada componente principal se mide a través de los autovalores de $\mathbf{S}$.

En la mayoría de las situaciones prácticas las p variables están correlacionadas por lo que cabe esperar que las primeras componentes expliquen un elevado porcentaje de la variabilidad total. Por ejemplo, si $m = 2 < p$, y $P_2 = 90\,\%$, las dos primeras componentes explican una gran parte de la variabilidad de las variables. Entonces podremos sustituir X_1, X_2, . . . ,X_p por las componentes principales Y_1, Y_2. En muchas aplicaciones, es posible encontrar una interpretación para las primeras componentes.

A modo de ejemplo calcularemos las componentes principales de una muestra de $n = 100$ observaciones en $\mathbb{R}^2$, en las que las variables presentan una fuerte correlación positiva. En la Figura 20 panel superior se presentan gráficos de los datos en el espacio original, mientras que los gráficos del panel inferior corresponden al espacio de las componentes principales. Los gráficos de la izquierda muestran las direcciones principales (flechas en rojo) en escala de los autovalores. En los gráficos de la derecha se muestra el soporte de cada conjunto de datos en línea verde punteada. En el espacio de las variables transformadas el soporte de los datos es rectangular ya que las variables no están correlacionadas.

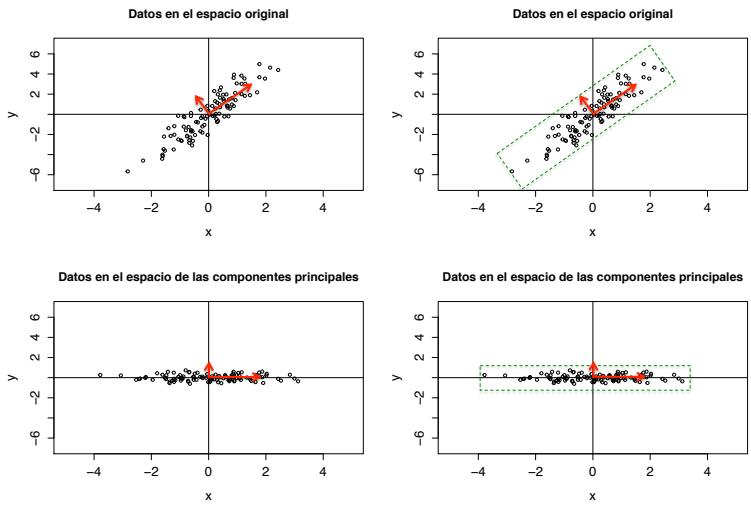

Figura 20: *Ejemplo de datos correlacionados y su representación en el espacio de componentes principales.*

En esta tesis utilizaremos componentes principales para simular curvas de rendimientos bajo la hipótesis nula de no existencia de grupos (sección 4.1 del capítulo 3). La propuesta consiste en simular observaciones a partir de una densidad uniforme dentro del soporte rectangular de las componentes principales y luego retransformarlas al espacio original para obtener observaciones que preserven las propiedades de asociación de los datos originales.

5. Test de hipótesis

El test de hipótesis es una regla formal que permite decidir entre dos hipótesis aquella que es más verosímil dado un cierto conjunto de datos. Más específicamente, es un procedimiento para distinguir entre dos distribuciones de probabilidad en base a variables aleatorias generadas por una de estas distribuciones.

Sea un vector aleatorio $\mathbf{X} = (X_1, X_2, ..., X_n)$ y supongamos que cada variable X_i tiene función de distribución perteneciente a la familia $F(x, \theta)$ con $\theta \in \Theta \subseteq \mathbb{R}^q$. Consideremos Θ_1 y Θ_2 tales que $\Theta_1 \cap \Theta_2 = \emptyset$ y $\Theta_1 \cup \Theta_2 = \Theta$. Un test para este problema será una regla basada en $\mathbf{X}$ para decidir entre dos hipótesis

$$H_0 : \theta \in \Theta_1 \quad \textit{contra} \quad H_1 : \theta \in \Theta_2,$$

donde H_0 se denomina hipótesis nula, y H_1 hipótesis alternativa.

A continuación presentamos algunas definiciones básicas que necesitamos tener presentes para entender cómo formular un test de hipótesis y de qué manera interpretarlo.

Definición 5.1: Un *test* es una función de decisión. Se dice que un test es *no aleatorizado* si toma solamente los valores 0 ó 1. Cuando $\varphi(\mathbf{X}) = 1$ se rechaza la hipótesis H_0 y por lo tanto, se acepta H_1. Por el contrario, si $\varphi(\mathbf{X}) = 0$ no se rechaza H_0.

Si el test toma valores distintos de 0 y 1, se dice que es un test *aleatorizado*. En este caso, el valor $\varphi(\mathbf{X})$ indica con qué probabilidad se rechaza H_0 si se observa $\mathbf{X} = x$, es decir, $\varphi(\mathbf{X}) = P(\textit{rechazar}\, H_0|\mathbf{X})$.

En la mayor parte de los casos, los tests se definen como función de un estadístico, llamado *estadístico del test*. En general, el estadístico del test mide la diferencia entre los datos observados y lo que se esperaría observar bajo H_0.

Definición 5.2: Se define la *región crítica* $\mathcal{R}$ de un test φ, como el conjunto de valores del estadístico que llevan a la decisión de rechazar H_0 y la *región de aceptación* $\mathcal{A}$ como el conjunto de valores del estadístico que llevan a aceptar H_0.

Definición 5.3: Al aplicar un test es posible que la decisión sea errónea. Llamamos *error de tipo 1* al que se comete al rechazar la hipótesis nula cuando ésta es verdadera, y *error de tipo 2* al que se comete al no rechazar la hipótesis nula cuando es falsa.

Para un test no aleatorizado, la probabilidad de cometer un error de tipo 1 será $P_\theta(\mathcal{R})$, cuando $\theta \in \Theta_1$. Mientras que la probabilidad de error de tipo 2 será $P_\theta(\mathcal{A}) = 1 - P_\theta(\mathcal{R})$, cuando $\theta \in \Theta_2$.

Definición 5.4: Se llama *función de potencia del test* $\varphi(\mathbf{X})$ a la función

$$\beta_\varphi(\theta) \quad = \quad P_\theta(\varphi(\mathbf{X}) = 1),$$

donde P_θ indica la probabilidad cuando la distribución que genera los datos es $F(x, \theta)$.

Podemos expresar la probabilidad de cada tipo de error en términos de $\beta_\varphi(\theta)$: la probabilidad de que ocurra un error de tipo 1 será $\beta_\varphi(\theta)$ para $\theta \in \Theta_1$ y la de que ocurra un error de tipo 2 será $(1 - \beta_\varphi(\theta))$ para $\theta \in \Theta_2$.

Un buen test debería tener probabilidad de errores de tipo 1 y 2 pequeñas. Es decir, la función de potencia $\beta_\varphi(\theta)$ debería tomar valores cercanos a 0 para $\theta \in \Theta_1$ y valores cercanos a 1 para $\theta \in \Theta_2$. Sin embargo, no podemos hacer simultáneamente ambas probabilidades de errores pequeñas, para un tamaño de muestra fijo, ya que cuando disminuye la probabilidad de error de tipo 1 aumenta la probabilidad de error de tipo 2, y viceversa. Para lograr que ambas probabilidades de error sean pequeñas lo que debemos hacer es aumentar la cantidad de observaciones.

La teoría clásica de test de hipótesis considera que el error de tipo 1 es mucho más grave que el error de tipo 2, es decir, es peor rechazar H_0 cuando es cierta que no rechazarla cuando es falsa. Por ésto, se quiere lograr tener mucha evidencia para decidir rechazar H_0.

Definición 5.5: El *nivel de significación* de un test φ está definido por

$$\alpha \quad = \quad \sup_{\theta \in \Theta_1} \beta_\varphi(\theta),$$

es decir, α es el supremo de la probabilidad de cometer un error de tipo 1.

Definición 5.6: Definimos el *nivel crítico* o *p-valor* asociado a una muestra de observaciones x como el menor valor de significación para el cual se rechaza la hipótesis H_0.

Así definido, el p-valor es una medida cuantitativa del nivel de evidencia que los datos presentan en contra de la hipótesis nula. Una vez fijado el nivel de significación α, y evaluado el p-valor p del test utilizado, rechazaremos H_0 si $p < \alpha$.

Capítulo III

APLICACION A DATOS DE RENDIMIENTOS AGRICOLAS

En este capítulo se describe la metodología que proponemos para identificar departamentos con patrones de rendimientos similares (grupos) a través del uso de algoritmos de segmentación. Para el desarrollo de la propuesta usaremos como ejemplo de aplicación los datos anuales de cultivos de maíz, en el período 1969-2010, provistos por el Ministerio de Agricultura de la Nación. Sin embargo, es importante remarcar que la metodología propuesta es de aplicación general a un problema en el que interese segmentar datos longitudinales.

1. Datos disponibles

El Ministerio de Agricultura de la Nación Argentina, recolecta información de siembra, cosecha, producción y rinde de cada campaña y la misma está disponible a nivel de departamentos. En esta tesis se usarán los datos anuales de los rendimientos de maíz en Argentina en las últimas décadas. El rendimiento de un cultivo para una dada campaña se calcula como el total de producción (en kilogramos) dividido por el área de la superficie cosechada (en hectáreas). Los valores corresponden a "campañas" que usualmente involucran el final de un año calendario con el comienzo del año siguiente, por lo que se acostumbra decir, por ejemplo, "el rendimiento de maíz en la campaña 2009-2010 en el departamento de Junín es de 10000 kg/ha". La información de rendimiento de cultivos está disponible a nivel de departamentos.

Debido a que, para un estudio futuro, nos interesa relacionar el clima con los rendimientos agrícolas, se seleccionaron los departamentos donde encontramos datos disponibles de cultivos de maíz, pero que además se encuentran cerca de estaciones meteorológicas importantes de modo de disponer de mediciones de temperatura y precipitación. Esta información es suministrada por el Servicio Meteorológico Nacional.

Del total de departamentos con cultivo de maíz en Argentina durante el período 1969-2010, consideramos aquellos en los que se hubiera cultivado maíz en la campaña 2010 y que tuvieran a lo sumo 10 valores faltantes de rendimiento (kg por ha.). Asignamos a cada departamento la estación meteorológica más cercana, con distancia inferior a 250 km. Contamos con 65 estaciones meteorológicas principales, lo que nos permitió analizar 264 departamentos que cubren el noreste de la provincia de La Pampa, las provincias de Bs. As., San Luis, Entre Ríos, Santa Fé, Córdoba, parte de Catamarca, región mesopotámica, centro de Tucumán, Santiago del Estero, Chaco, y regiones de Salta y Jujuy. El mapa de la Figura 21 muestra los departamentos que fueron seleccionados para este análisis, donde cada color representa los departamentos de una provincia.

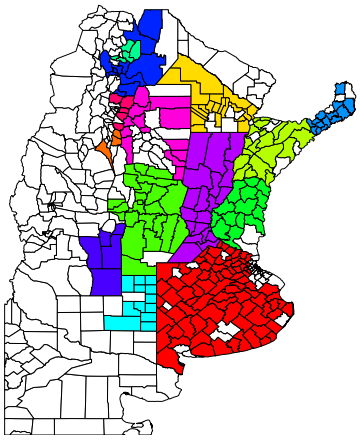

Figura 21: *Departamentos de Argentina con cultivo de Maíz.*

Vamos a llamar n a la cantidad total de objetos, es decir al número de departamentos, $n = 264$, y T a la cantidad de campañas realizadas entre 1969 y 2010 , entonces $T = 42$. Así, cada departamento queda caracterizado por un vector de 42 observaciones, es decir, para cada i tal que $1 \leq i \leq 264$ consideraremos

$$\mathbf{y}_i \quad = \quad (y_{i1}, y_{i2}, \cdots, y_{i42}).$$

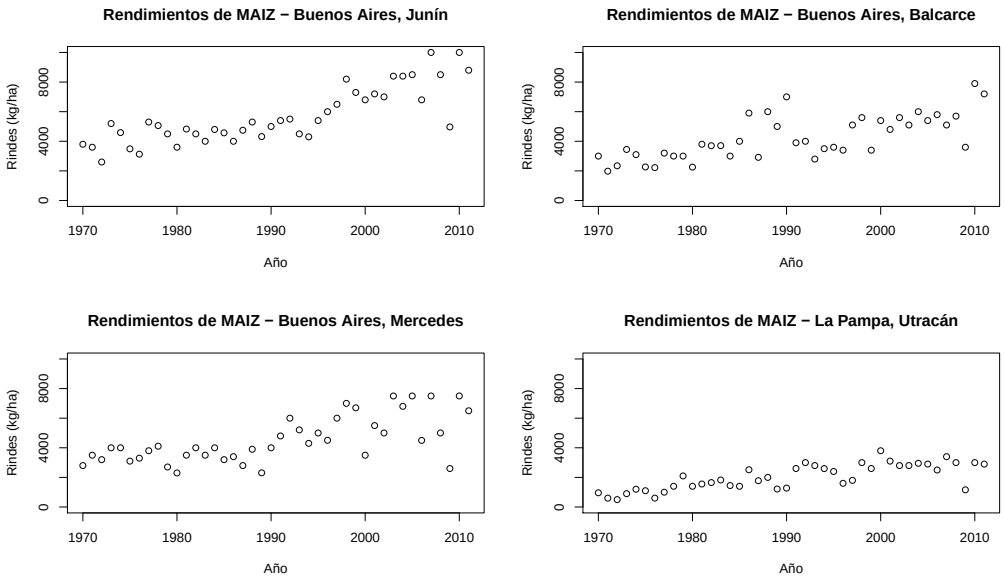

Figura 22: *Rendimientos de maíz en cuatro departamentos de Argentina 1969-2010.*

La Figura 22 muestra ejemplos de los datos de rendimiento en cuatro departamentos de Argentina para el período 1969-2010. Los departamentos fueron elegidos para mostrar algunas características típicas de este tipo de datos. En primer lugar, notemos que podría ocurrir que (no es el caso de los gráficos) para algunas campañas no se dispone de datos, sea porque no se cultivó maíz ese año o porque el Ministerio de Agricultura no dispone de la información. En segundo lugar, notemos que las "curvas" o "patrones" de rendimientos pueden diferir sustancialmente, por ejemplo, el perfil de rendimiento de Junín crece notablemente a partir de los años 90, mientras que en los departamentos de Balcarce y Mercedes la tendencia de crecimiento es aproximadamente lineal. En el departamento de Utracán, en la provincia de La Pampa, sin embargo,

aunque se observa un leve aumento en los rindes, el comportamiento es bastante más estable.

Nuestro objetivo consiste en segmentar el conjunto de 264 departamentos (objetos) en grupos de departamentos en los que el patrón de rendimientos sea homogéneo. Una posibilidad es aplicar un procedimiento de segmentación a los datos crudos de rendimientos. Sin embargo, el hecho de que tengamos datos faltantes para algunas campañas limita esta opción por la dificultad que representan éstos en un análisis multivariado. Aún si tuviéramos datos completos, el vector de rendimientos observados en un departamento en general incluye datos que representan situaciones atípicas tales como siniestros, catástrofes o condiciones climáticas extremas y puntuales. En general no estamos interesados en estos comportamientos atípicos sino en identificar la tendencia de los rendimientos a lo largo del tiempo, como un perfil suave. En la sección siguiente se propone una manera de estimar este perfil suave.

2. Cálculo de la curva suave de rendimientos

Dados datos de rendimientos obtenidos a lo largo del tiempo, el objetivo es obtener para cada unidad de observación (departamento) una curva suave que represente el comportamiento de los rendimientos en ese departamento, es decir, queremos trabajar con funciones continuas y derivables para que resulten trayectorias suaves. La idea es transformar el vector de datos originales en otro vector de la misma dimensión que represente los valores de rendimientos "predichos" bajo un modelo suave. De este modo lograremos: 1) evitar el problema de los datos faltantes ya que los valores de rendimientos para las campañas faltantes pueden ser "predichos" bajo el modelo de la curva suave y, 2) evitar el efecto de los casos 'raros', de índole climatológica o natural, como una sequía o plaga en el algoritmo de segmentación. A continuación se presenta cómo obtener esta curva de rendimiento suave para cada departamento, usando un método paramétrico, es decir, ajustando un polinomio de grado g.

Consideremos el espacio $\mathbb{R}^T$, en nuestro ejemplo de aplicación $T = 42$, y el producto escalar usual, dado por

$$\langle \mathbf{x}, \mathbf{y} \rangle \quad = \quad \sum_{t=1}^{T} x_t y_t.$$

Sea y_t el rendimiento observado en la campaña t-ésima y sea x_t la variable que indica la campaña $t-$ésima. Sea $y_t = f(x_t) + \epsilon_t$, donde f es una función continua que desconocemos y que representa la tendencia de los rendimientos del departamento en el período considerado, y ϵ_t es la distancia entre la función y el rendimiento observado en la campaña t. Restringimos el espacio de todas las funciones continuas y derivables al espacio de los polinomios con coeficientes reales de grado menor o igual a g, con g suficientemente pequeño. Nos interesa minimizar la distancia entre el vector de datos observados $\mathbf{y}^{\intercal} = (y_1, y_2, ..., y_T)$ y la función polinómica evaluada en las campañas correspondientes ($p(x_1)$, $p(x_2)$, ... , $p(x_T)$) en la norma asociada al producto escalar. En otras palabras, queremos ajustar un modelo polinómico (grado $g > 1$) minimizando la suma de los cuadrados de los errores o residuos ϵ_t.

Si $p(x) = a_g x^g + a_{g-1} x^{g-1} + \ ... \ + a_1 x + a_0$ entonces podemos escribir

$$\begin{pmatrix} p(x_1) \\ p(x_2) \\ \vdots \\ p(x_T) \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^g \\ 1 & x_2 & x_2^2 & \cdots & x_2^g \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_T & x_T^2 & \cdots & x_T^g \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_g \end{pmatrix}.$$

Sea $\mathbf{A} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^g \\ 1 & x_2 & x_2^2 & \cdots & x_2^g \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_T & x_T^2 & \cdots & x_T^g \end{pmatrix}$ y $\mathbf{a}^{\intercal} = (a_0, a_1, ..., a_g)$ nuestro problema se reduce a encontrar $\mathbf{a} \in \mathbb{R}^{g+1}$ que minimice $\|\mathbf{y} - \mathbf{A}\mathbf{a}\|$.

Esta minimización se conoce como método de mínimos cuadrados y es el método clásico de estimación en un problema de regresión lineal. Llamaremos $\tilde{\mathbf{a}}^{\intercal} = (\tilde{a}_0, ..., \tilde{a}_g)$ al vector de coeficientes resultante de la minimización e $\tilde{\mathbf{y}} = \mathbf{A}.\tilde{\mathbf{a}}$ al vector de valores del polinomio "estimado" para los distintos años.

Un problema usual al ajustar una curva suave usando el método de mínimos cuadrados es la gran influencia que pueden tener observaciones atípicas. Ya hemos mencionado anteriormente que es frecuente encontrar en las series de rendimientos campañas con resultados atípicos, usualmente debido a fenómenos naturales. Por esta razón, "estimaremos" la función suave utilizando un método robusto, en el que se reduce la influencia de los datos atípicos. En particular, obtendremos los coeficientes del vector utilizando M-estimadores (ver [8]), los cuales tienen la ventaja de combinar robustez y eficiencia bajo el modelo de regresión; ajustaremos el modelo polinomial usando la función *rlm* de **R** (regresión robusta).

En cada aplicación particular, será necesario elegir el grado del polinomio que se propondrá como modelo de curva suave. Una manera de seleccionar el grado del polinomio es analizar los residuos. En el ejemplo de aplicación a rendimientos de maíz, decidimos trabajar con polinomios de grado $g = 5$, ya que resulta una buena relación de compromiso entre flexibilidad y parsimonia. Es decir, no permite demasiadas oscilaciones, y cubre la gama de patrones de rendimiento observados en los departamentos.

Cuando se ajusta un polinomio de grado 5 a los datos de rendimiento de maíz en cada departamento, usando regresión robusta, se obtienen vectores de rendimientos ajustados $\tilde{\mathbf{y}}_i^{\intercal} = (\tilde{y}_{i1}, \tilde{y}_{i2}, \cdots, \tilde{y}_{i42})$ con $i = 1, ..., 264$.

A continuación, presentaremos algunos gráficos de las trayectorias temporales ajustadas (curvas de color azul), junto con los valores observados de rinde (puntos negros). Se puede ver que los departamentos de Junín y Monte (Figuras 23 y 24) tienen un comportamiento similar con un crecimiento de tipo exponencial a partir de los años 90. Los más altos rendimientos de maíz, cercanos a 10000 kg/ha, se alcanzan en los últimos años. Por otro lado, los departamentos de Balcarce, San Justo y Mercedes, Figuras 25, 26 y 27, presentan una tendencia de crecimiento prácticamente lineal, con rindes inferiores a 8000 kg/ha. El departamento de Utracán, Figura 28, se diferencia de los anteriores ya que el rendimiento de maíz creció en forma mucho más lenta. Nótese que el valor ajustado apenas supera los 2500 kg/ha, y que la variabilidad de los datos es notablemente menor que en los otros cinco departamentos. Estos ejemplos muestran que los rindes de maíz presentan comportamientos muy variados tanto en referencia a los valores absolutos de producción por hectárea cosechada, como al tipo de trayectoria o la evolución de la tendencia a lo largo de las campañas.

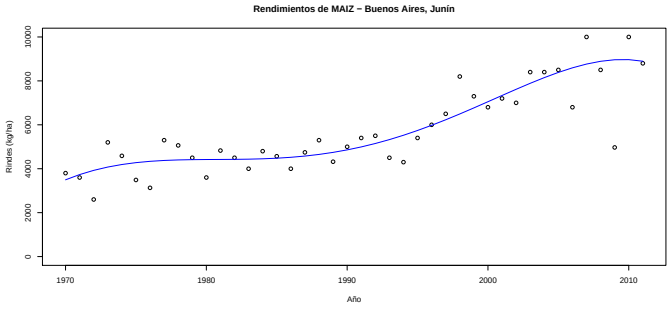

Figura 23: *Curva ajustada de rendimientos de maíz para el departamento de Junín.*

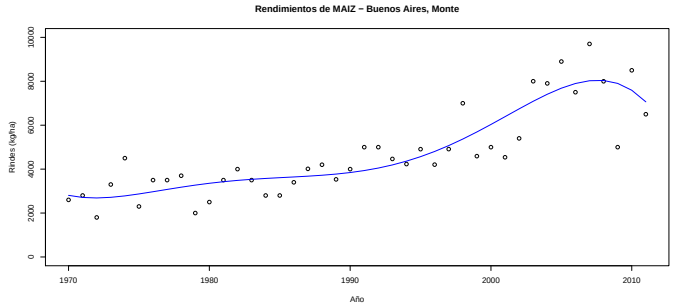

Figura 24: *Curva ajustada de rendimientos de maíz para el departamento de Monte.*

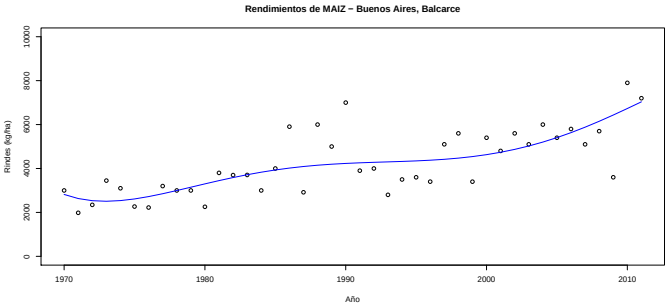

Figura 25: *Curva ajustada de rendimientos de maíz para el departamento de Balcarce.*

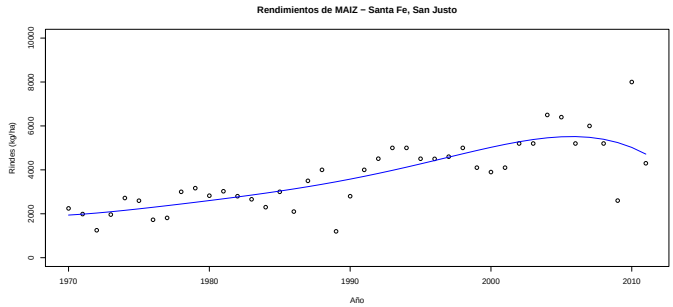

Figura 26: *Curva ajustada de rendimientos de maíz para el departamento de San Justo.*

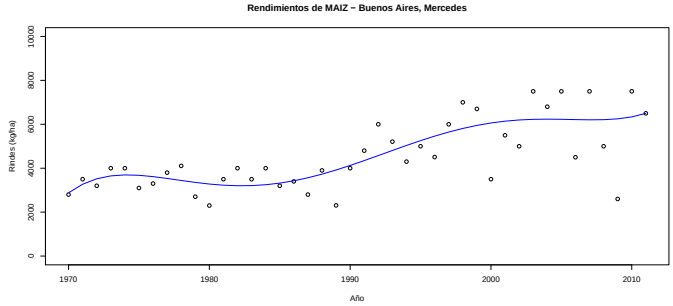

Figura 27: *Curva ajustada de rendimientos de maíz para el departamento de Mercedes.*

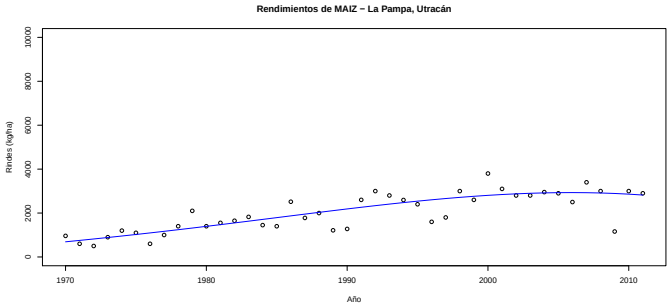

Figura 28: *Curva ajustada de rendimientos de maíz para el departamento de Utracán.*

El crecimiento observado en los rindes a partir de los años 90 para una gran parte de los departamentos del país puede ser explicado por el uso de nueva tecnología y los avances científicos en nuestro país. En un artículo de la revista *Agromensajes* (ver [15]), el ingeniero agrónomo Daniel Rossi (Cátedra de Mejoramiento Vegetal y Producción de Semillas, Facultad de Ciencias Agrarias, Universidad Nacional de Rosario) comenta acerca del cultivo de maíz en Argentina:

"En los años 90 se logran importantes incrementos en la productividad y además son destacables los avances en materia de calidad. Entre las causas del crecimiento operado en la producción nacional debe citarse el aumento de la superficie cultivada, la disponibilidad de nuevos HS (híbridos simples) de mayor potencial y mejor resistencia a enfermedades y plagas, el incremento en el área fertilizada, la creciente adopción de la siembra directa, la incorporación del riego complementario, el recambio del parque de cosechadoras y, a partir del ciclo agrícola 1998/1999, el inicio de la difusión de materiales transgénicos...

... En el primeros años de los 2000 el incremento en el uso de fertilizantes, la siembra de precisión y la mayor superficie regada, junto con la disponibilidad de mejor germoplasma y la adopción de híbridos transgénicos con resistencia a insectos (RI o Bt) o tolerancia a herbicidas (TH) determinó un nuevo y significativo cambio en la tendencia creciente de los rendimientos por unidad de superficie."

3. Procedimiento de segmentación

Nuestro objetivo es encontrar grupos en el conjunto de datos de rendimientos de maíz. Como se explicitó en la sección anterior, el vector de datos originales de rindes para cada departamento será reemplazado por el vector de valores predichos de rendimientos bajo un ajuste polinómico; el que será referido a partir de ahora como "curva de rendimientos". El paso siguiente será utilizar un método de segmentación para agrupar las curvas de rendimientos.

En primer lugar, decidimos elegir un método jerárquico por la propiedad fundamental que tiene esta clase de segmentación de que los grupos están anidados. Formalmente esta propiedad puede expresarse del siguiente modo:

Sea α el *nivel de fusión* en un procedimiento jerárquico, donde $\alpha = 0$ representa n clusters individuales, y $\alpha = n - 1$ un único cluster conteniendo todos los elementos. Sean, ahora, los conjuntos de las particiones $\mathcal{P}^{n-\alpha}$ formadas por los $n - \alpha$ agrupamientos según sea el nivel de fusión α. Luego, tomemos dos niveles de fusión l y m, tales que $l < m$, $l, m \in \{0, 1, 2, \ldots, n-1\}$. Dos clusters cualesquiera $C_{\tilde{l}}$ y $C_{\tilde{m}}$, que verifican que $C_{\tilde{l}} \in \mathcal{P}^{n-l}$ y $C_{\tilde{m}} \in \mathcal{P}^{n-m}$, cumplen que

$$C_{\tilde{l}} \subseteq C_{\tilde{m}}, \text{ o } \quad C_{\tilde{l}} \cap C_{\tilde{m}} = \emptyset.$$

Notemos que $C_{\tilde{l}}$ y $C_{\tilde{m}}$ son dos clusters cualesquiera que pertenecen respectivamente a las particiones $\mathcal{P}^{n-l}$ y $\mathcal{P}^{n-m}$.

Del conjunto de procedimientos jerárquicos ascendentes seleccionamos el método de Ward, el que fue descripto en la sección 1.2.5 del capítulo 2. Este procedimiento tiende a producir clusters compactos o esféricos, lo que es beneficioso en el

contexto de curvas de rendimientos, ya que nos interesa agrupar curvas con patrones similares y resumirlas en una curva "promedio"o "típica". Un método de segmentación que produce grupos esféricos se corresponde con una densidad en el espacio multivariado que es bien resumida por una medida de centralidad (la curva típica). Además este método tiende a formar grupos de aproximadamente el mismo tamaño y tal como se mostró en la sección 3.4 del capítulo 2, la medida de disimilitud usada por Ward está directamente relacionada al estadístico gap propuesto por Tibshirani y otros (2001). Este estadístico será evaluado en la sección 5 de este capítulo.

La función *hclust* de **R** proporciona diferentes métodos de agrupamiento y permite la elección de la medida de disimilaridad o de distancia a ser usada por el método. En este caso, consideramos la función *hclust* con parámetro de método "Ward" y distancia euclídea. Cuando se aplica el método Ward a los datos de las trayectorias suaves de rendimientos de maíz en Argentina, $\tilde{\mathbf{y}}_i' = (\tilde{y}_{i1}, \tilde{y}_{i2}, \cdots, \tilde{y}_{i42})$ con $i = 1, ..., 264$, obtenemos el dendrograma de la Figura 29.

El dendrograma es un árbol que representa las relaciones de similaridad/disimilitud entre los objetos y la secuencia en la que ocurren las fusiones. El eje horizontal representa los objetos originales. Las alturas verticales (height) representan la distancia o el grado de disimilaridad entre los grupos que se unen en cada paso, mientras más alta la línea vertical, mayor es la diferencia entre los clusters que se unen. En el caso del método de Ward la altura vertical corresponde a la diferencia en la suma de cuadrados intra-cluster en los dos pasos. El dendrograma es una herramienta útil para determinar el número de clusters. Cualquier salto importante en la altura es indicativo de un número de clusters a ser considerado. Por ejemplo el dendrograma de la Figura 29 sugiere la existencia de un número entre 3 y 10 grupos dependiendo cuán importante sea la disimilitud (salto en altura) entre clusters que estamos dispuestos a tolerar. En la sección siguiente presentaremos el problema de seleccionar el número de clusters.

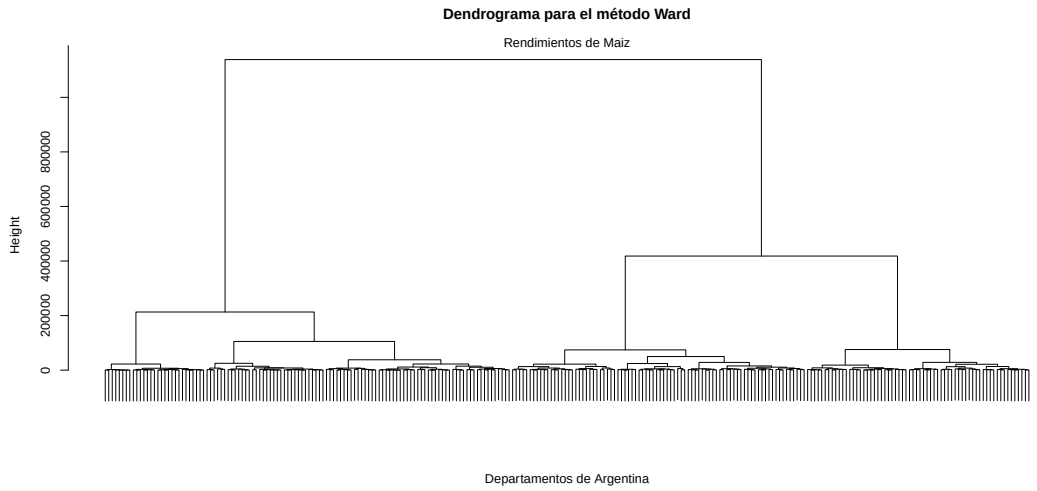

Figura 29: *Dendrograma según método de Ward aplicado a curvas ajustadas de rendimientos de maíz. Datos de Argentina, 1969-2010.*

4. El problema de determinar el número de clusters

Notemos que cualquiera sea la distribución de los objetos en el espacio multivariado, un procedimiento de segmentación jerárquico producirá un dendrograma con una secuencia de clusterización, aún cuando no hubiera grupos. Es decir, si los objetos se distribuyeran uniformemente en cierto soporte del espacio $\mathbb{R}^p$, el método de segmentación necesariamente producirá un dendrograma a partir del cual podemos definir clusters. Por lo tanto, en un problema de segmentación es importante preguntarse si efectivamente en la población de objetos bajo estudio existen grupos, o dicho de otro modo, si existen zonas del espacio $\mathbb{R}^p$ en las que hay mayor concentración de objetos.

Como se comentó en la sección anterior el dendrograma es una herramienta gráfica que permite evaluar la existencia de grupos y proponer la cantidad de grupos, pero el método es más bien subjetivo ya que no permite cuantificar la

incertidumbre o dar una noción de significatividad estadística. En la sección 2 del capítulo 2 hemos presentado algunas de las propuestas para seleccionar el número de clusters. La característica más destacable del método gap es que propone cómo obtener la distribución empírica del estadístico simulando datos a partir de una distribución nula en la que no existen clusters, denominada distribución de referencia.

Cuando los datos corresponden a "curvas ajustadas de rendimientos", como es el caso que nos ocupa, elegir la distribución de referencia implica consideraciones adicionales. En la sección que sigue presentaremos estas consideraciones y una propuesta de cómo generar la distribución de referencia. Mostraremos además el comportamiento de los distintos métodos usando conjuntos de datos ficticios en los que se conoce el número de clusters en la población ($k = 2$, $k = 3$ y $k = 7$). Finalmente, en la subsección 4.3 proponemos una modificación del estadístico gap y sugerimos considerar dos nuevos criterios.

4.1. Una propuesta de cómo generar la distribución de referencia cuando los datos corresponden a "curvas ajustadas"

El objetivo de esta tesis es encontrar agrupamientos en el conjunto de datos de rendimientos de maíz en Argentina en el período 1969-2010. Recordemos que los datos que usaremos para el procedimiento de segmentación son vectores de dimensión T, cuyas componentes resultan de evaluar el polinomio de grado g ajustado $\tilde{\mathbf{y}}^{\intercal} = (\tilde{y}_1, \tilde{y}_2, \cdots, \tilde{y}_T) = (p(x_1), p(x_2), \ldots, p(x_T))$. En nuestro ejemplo de aplicación $T = 42$ y $g = 5$.

Un problema por considerar es cómo simular observaciones $\tilde{\mathbf{y}}^{\intercal}$ a partir de una distribución bajo la hipótesis nula de no existencia de clusters. Es decir, a partir de cierto soporte en el que los objetos (vectores representando las curvas ajustadas) tengan distribución uniforme. Un polinomio queda unívocamente definido por el conjunto de $g + 1$ coeficientes, así que una opción sería considerar el conjunto de coeficientes. Notemos además que lo que importa son las curvas o trayectorias de rendimientos evaluadas en los años de interés y no las funciones polinómicas que definen estas trayectorias.

Dado que el objetivo es agrupar departamentos en base a su comportamiento en un subconjunto del dominio natural, años 1969-2010, y no en todo el dominio, no usaremos la métrica definida en el espacio de los coeficientes, $\mathbb{R}^{g+1}$, sino que restringiremos la métrica de comparación a los valores de las imágenes en el período de interés. Recordemos que, de todos modos, en $\mathbb{R}^n$, ambas métricas son equivalentes.

Para entender por qué es conveniente comparar las imágenes (y no los coeficientes) tomamos un polinomio $\mathcal{P}_0$ de grado 5 con coeficientes $\mathbf{a}_0^{\intercal} = (a_1, a_2, a_3, a_4, a_5, a_6) \in \mathbb{R}^6$ y generamos 200 nuevos vectores de coeficientes perturbando ligeramente el vector $\mathbf{a}_0$. Así obtenemos los polinomios $\mathcal{P}_j$ de coeficientes $\mathbf{a}_j^{\intercal} = (a_{j1}, a_{j2}, a_{j3}, a_{j4}, a_{j5}, a_{j6}) \in \mathbb{R}^6$, $1 \leq j \leq 200$. Definimos el conjunto de los coeficientes de los 200 polinomios generados como $A = \{\mathbf{a}_j^{\intercal} = (a_{j1}, a_{j2}, a_{j3}, a_{j4}, a_{j5}, a_{j6}) \in \mathbb{R}^6, 1 \leq j \leq 200\}$, y sea la distancia $d_{Coef}(\mathbf{a}_j) = \|\mathbf{a}_0 - \mathbf{a}_j\|$.

A partir de estos 200 polinomios seleccionamos un conjunto de 50 polinomios usando dos criterios:

1. Aquellos en que la distancia $d_{Coef}(\mathbf{a}_j) = \|\mathbf{a}_0 - \mathbf{a}_j\|$ es mínima y que definen el subconjunto $A_{Coef} \subseteq A$.

2. Aquellos en que la imagen es más próxima a la imagen de $\mathcal{P}_0$. Tomamos como dominio el intervalo $[1, 42]$, y consideramos 1000 valores equiespaciados en él, es decir, para $h = 1, 2, \ldots, 1000$ nos quedamos con los elementos $c_h = \left(1 + (h-1)\frac{(42-1)}{999}\right) \in [1, 42]$. Calculamos la distancia $d_{Im}(\mathbf{a}_j) = \|\mathbf{b}_0 - \mathbf{b}_j\|$ donde $\mathbf{b}_0 = (b_{0\,1}, b_{0\,2}, \ldots, b_{0\,1000}) \in \mathbb{R}^{1000}$, con $b_{0\,h} = \mathcal{P}_0(c_h)$; y $\mathbf{b}_j = (b_{j\,1}, b_{j\,2}, \ldots, b_{j\,1000}) \in \mathbb{R}^{1000}$, con $b_{j\,h} = \mathcal{P}_j(c_h)$, $1 \leq j \leq 200$. Definimos el subconjunto $A_{Im} \subseteq A$ que contiene los 50 vectores de coeficientes con menor distancia d_{Im}.

La Figura 30 muestra el polinomio original $\mathcal{P}_0$ (curva negra) y el gráfico de los 50 polinomios obtenidos bajo estas dos diferentes propuestas. En el primer caso se graficaron los polinomios con coeficientes en A_{Coef} (métrica de los coeficientes, panel izquierdo), y en el segundo caso, los polinomios con coeficientes en A_{Im} (métrica de las imágenes, panel derecho).

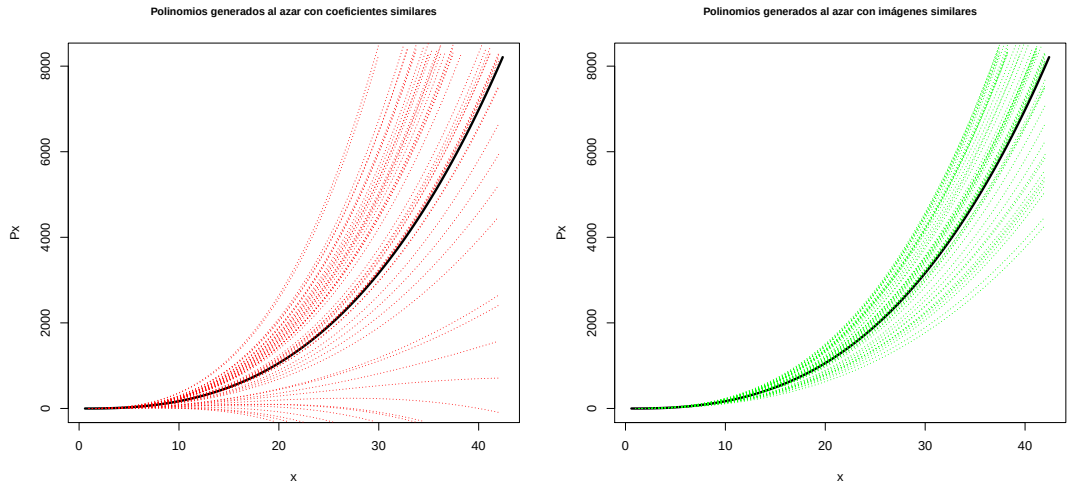

Figura 30: *Ejemplo de generación al azar de polinomios a partir de otro polinomio (curva negra), considerando los más próximos en cuanto a coeficientes (curvas rojas en gráfico izquierdo), y los más similares en términos de su imagen (curvas verdes en gráfico derecho).*

Notemos que generar polinomios a partir de los coeficientes, aún usando perturbaciones muy pequeñas, produce una gran variabilidad de perfiles. Algunos de estos perfiles no son compatibles con el fenómeno que estamos estudiando, por ejemplo curvas con valores negativos de rendimientos. Por otro lado, la gran variabilidad en el aspecto de las curvas no representa la hipótesis nula de interés: la no existencia de grupos de curvas. Por estas razones, proponemos generar la distribución de referencia a partir del espacio de las imágenes.

Sean $\tilde{\mathbf{y}}_i \in \mathbb{R}^T, i = 1, ..., n$ las observaciones que representan las curvas ajustadas. Generamos datos bajo la distribución de referencia (hipótesis nula de no existencia de clusters de curvas) siguiendo la propuesta de Tibshirani (2001) descripta en la sección 3.2 del capítulo 2. El procedimiento es el siguiente: 1) calculamos las componentes principales para el conjunto de curvas $\tilde{\mathbf{y}}$; 2) generamos aleatoriamente cada componente a partir de uniformes en el rango en que se mueven las componentes principales (ver Figura 20 en capítulo 2, sección 4.2); y 3) retransformamos los vectores obtenidos para obtener una muestra de datos en el espacio original, es decir, una trayectoria de rendimientos.

4.2. Estudio del comportamiento de las propuestas para seleccionar el número de clusters

En esta sección estudiaremos la performance del estadístico gap y del estadístico CH para detectar el número de clusters en situaciones de complejidad creciente. Comenzaremos considerando un problema en el que existen claramente dos grupos de curvas de rendimiento. Estos grupos corresponden a dos conjuntos de curvas obtenidos a partir de la clusterización jerárquica del total de departamentos elegidos *ad-hoc* como el par de grupos de curvas que a simple vista se diferencian más. Con el mismo criterio seleccionamos 3 conjuntos de curvas. Finalmente y para obtener 7 grupos, consideramos los tres grupos anteriores y generamos otros 4 grupos a partir de rectas visualmente separadas y diferenciadas de los agrupamientos anteriores.

En todos los ejemplos, siguiendo el procedimiento descripto en la sección anterior, se generaron $B = 1000$ curvas de rendimiento bajo la hipótesis de no existencia de clusters para estudiar el comportamiento de los estadísticos que permiten estimar el número de clusters.

4.2.1. Dos grupos de curvas

Comenzamos con un ejemplo de una situación con dos grupos claramente distinguibles, como se observa en la Figura 31, los cuales corresponden a dos de los clusters de curvas obtenidos en el dendrograma de la Figura 29. Las curvas rojas

representan departamentos con los más altos niveles de rindes con ubicación geográfica en la zona núcleo de nuestro país, en el norte de la provincia de Buenos Aires y en el sur de Santa Fé ($n = 30$), en tanto que las amarillas representan departamentos con los niveles más bajos, ubicados en departamentos de la zona norte de la provincia de Corrientes y en la provincia de Misiones ($n = 29$).

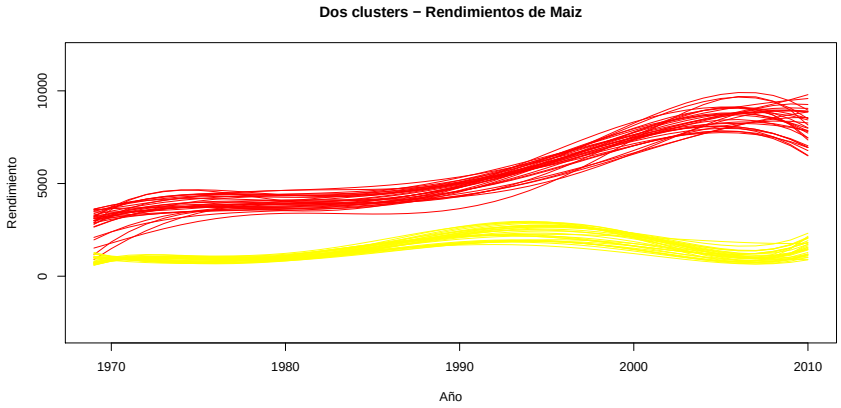

Figura 31: *Ejemplo de 2 grupos de trayectorias de rendimientos de maíz.*

Cuando aplicamos el método de segmentación de Ward a este conjunto de datos, se obtiene el dendrograma de la Figura 32, donde se observan claramente dos agrupamientos. Una separación tan abrupta no es común en problemas de segmentación.

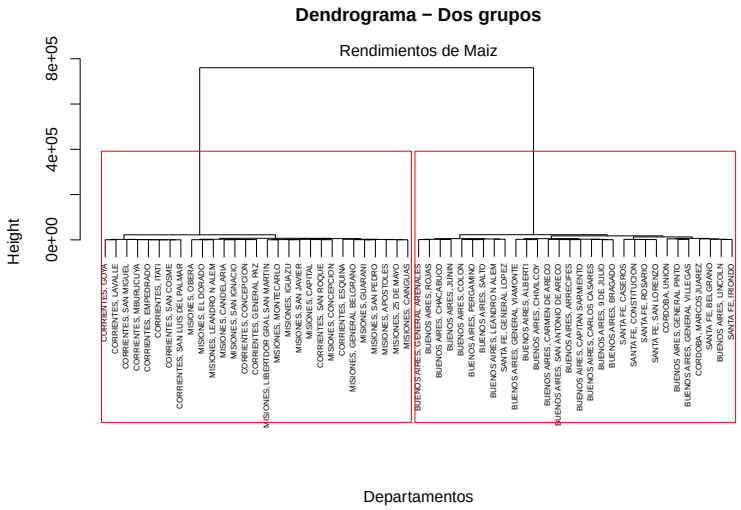

Figura 32: *Dendrograma correspondiente al ejemplo de dos grupos de trayectorias de rendimientos de maíz (Figura 31).*

El estimador del número de clusters en la distribución de la que proviene la muestra es el valor $\hat{k}$, que maximiza $Gap_n(k)$. Recordemos que, el estadístico gap, presentado en la sección 3 del capítulo 2 se define como

$$Gap_n(k) \quad = \quad E_n^* \left\{log(W_k)\right\} - log(W_k),$$

donde E_n^* es la esperanza de una muestra de tamaño n bajo la distribución de referencia y W_k es la suma de cuadrados dentro de clusters al obtener un agrupamiento de k clusters. El valor $\hat{k}$ que maximice $Gap_n(k)$ será la cantidad de grupos óptima. Tibshirani y otros (2001) proponen alternativamente definir $\hat{k}$ como el menor k que verifique $Gap(k) \geq$

$Gap(k+1) - s_{k+1}$ siendo $s_k = \sqrt{1+1/B}sd(k)$, B la cantidad de repeticiones y $sd(k)$ el desvío estándar de las B copias de $log(W_k^*)$.

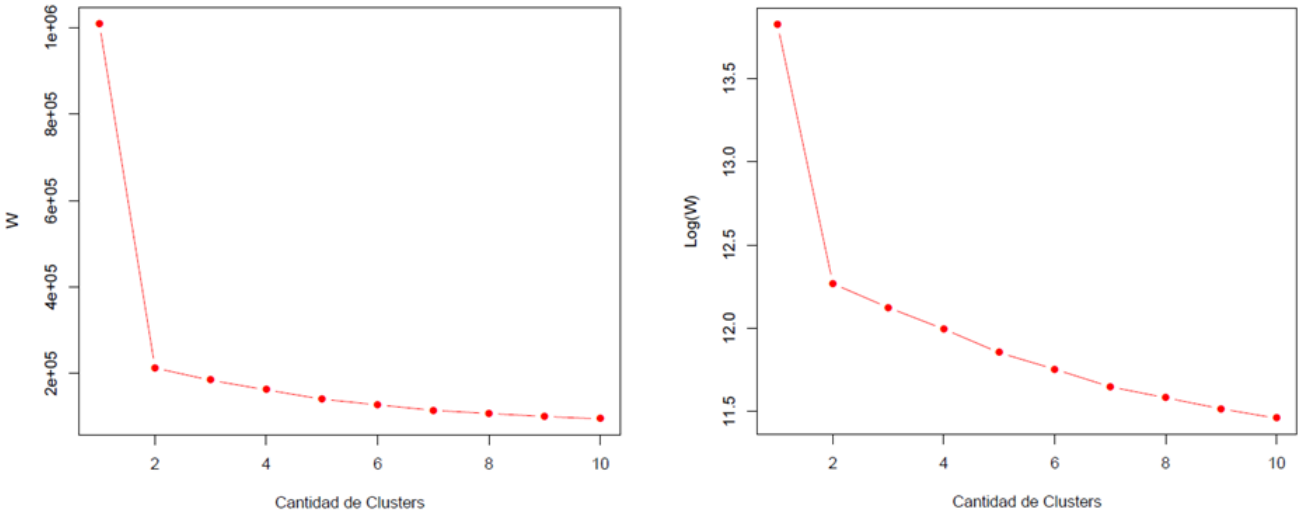

Figura 33: *W observada (izquierda) y logaritmos de W observados (derecha) en el caso de dos grupos.*

La Figura 33 muestra los gráficos de W y de $log(W)$ obtenidos para los datos originales en nuestro ejemplo de dos grupos. Como se puede observar el cambio en W o $log(W)$ es abrupto al pasar de dos grupos a uno solo, lo que genera un "codo" en el gráfico. Recordemos que, en los procedimientos de segmentación jerárquica aglomerativa como el que estamos utilizando, el algoritmo parte de una cantidad de clusters igual a la cantidad de elementos los que se van uniendo hasta obtener un único grupo que contiene todos los elementos. El "precio" que se paga por cada unión está dado por W.

El paso siguiente es obtener $B = 1000$ curvas simuladas desde una distribución bajo la hipótesis nula de no existencia de grupos y calcular el estadístico gap bajo cada repetición. Para ello se siguió el procedimiento que incluye los pasos:

- Calcular las componentes principales de los vectores correspondientes a las curvas rojas y amarillas.
- Generar $B = 1000$ vectores de dimensión 42 a partir de una distribución uniforme en el rango de los datos, previo centrado de los mismos, transformados al espacio de las componentes principales (distribución de referencia). La Figura 34 muestra las 20 primeras curvas de rendimientos generadas con este procedimiento.

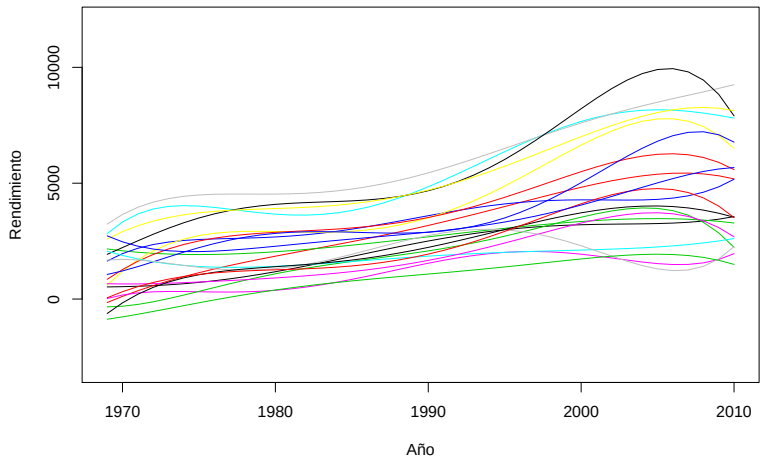

Figura 34: *Curvas de rendimientos generadas a partir de la distribución de referencia.*

- Aplicar el método de segmentación jerárquica de Ward bajo la distancia euclídea a cada muestra simulada.
- Calcular W, y $log(W)$ para cada muestra.

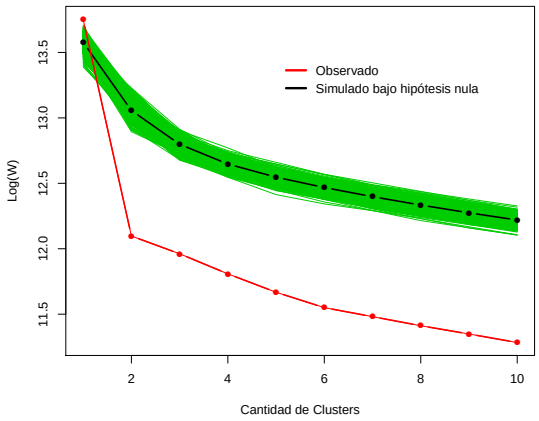

Figura 35: *Logaritmos de W observado y simulado bajo la distribución de referencia. Datos Figura 31 (dos grupos).*

En la Figura 35 se representa $log(W)$ observado (curva roja) junto con $log(W^*)$ (curvas verdes) para las 1000 muestras simuladas a partir de la distribución de referencia que supone la no existencia de clusters. Se puede notar la diferencia entre la curva de $log(W)$ proveniente de la segmentación de los datos originales (Figura 31) y el promedio de las curvas de $log(W^*)$ de las segmentaciones realizadas para las repeticiones bajo la distribución de referencia (curva negra). Es notable el incremento en la varianza intra-clusters al pasar de 2 a 1 cluster con los datos originales, en comparación con el incremento que se observa bajo la hipótesis de no existencia de grupos. Notemos además que a partir de $k \geq 2$ la curva de $log(W)$ observada es prácticamente paralela a la curva de $log(W^*)$ bajo la hipótesis nula. Esto significa que antes de alcanzar el valor de k óptimo el incremento en variabilidad al unir clusters es comparable en la muestra observada con el que se observa bajo la hipótesis nula (que se observaría en una situación de no existencia de clases). Dicho de otro modo, la "curva" de $log(W)$ observado es prácticamente paralela a la de $log(W^*)$ bajo la hipótesis nula a partir de $k \geq 2$.

En la Figura 36 se muestra el gráfico del estadístico gap para los datos originales (Figura 31) y los desvíos estándares de

las B copias de $log(W_k^*)$ calculados en base a las 1000 curvas simuladas. El valor óptimo de k corresponde al menor k tal que $Gap(k) - Gap(k+1) \geq -s_{k+1}$, que nuevamente resulta ser $k = 2$.

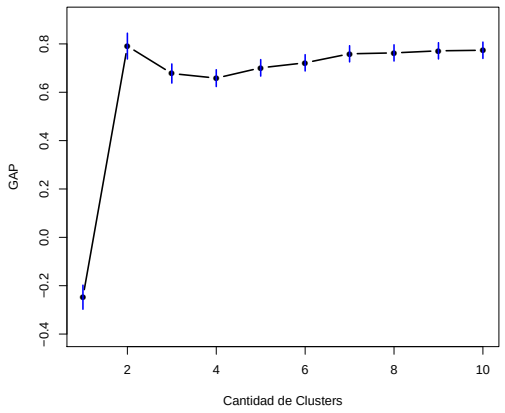

Figura 36: *Gráfico del estadístico gap para los datos de la Figura 31. Los segmentos azules representan los desvíos estándares de* $log(W^*)$ *bajo la distribución de referencia.*

Tengamos en cuenta ahora el criterio de Calinski y Harabasz presentado en la sección 2 del capítulo 2. En este caso, se busca maximizar el estadístico

$$CH(k) = \frac{\frac{tr[\mathbf{B}]}{k-1}}{\frac{tr[\mathbf{W}]}{n-k}} \qquad (k > 1)$$

donde $tr(\mathbf{B})$y $tr(\mathbf{W})$ son las sumas de cuadrados entre y dentro de clusters. Por definición k debe ser mayor que uno, por lo que la curva CH comienza en $k = 2$. La Figura 37 muestra los valores del estadístico $CH(k)$ para distintos valores de k para el ejemplo de la Figura 31. Este método obtiene $k = 2$ como cantidad óptima de agrupamientos ya que el máximo de la curva CH se da en este valor de k.

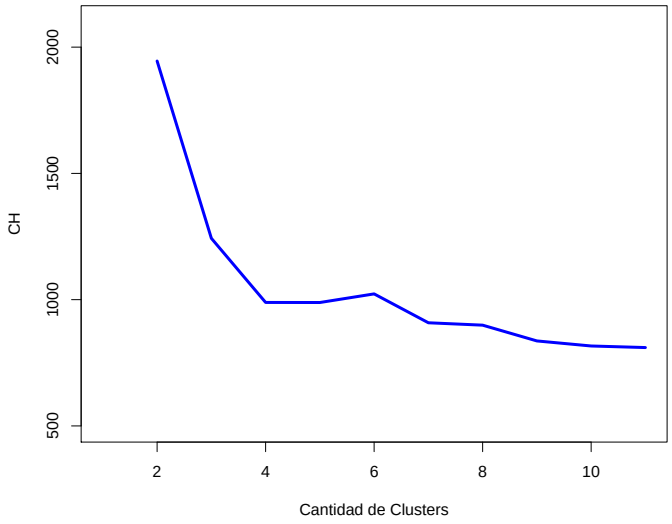

Figura 37: *Estadístico* CH *para el caso de dos grupos de trayectorias de rendimiento.*

4.2.2. Tres grupos de curvas

Consideremos a continuación los datos que se muestran en la Figura 38, en los que se observan los dos grupos de curvas analizados en el ejemplo anterior y un nuevo grupo de curvas de rindes intermedios, correspondientes a departamentos ubicados en el sur de la provincia de Buenos Aires, noroeste de Córdoba, norte de Santa Fé, este de Entre Ríos y en Tucumán ($n = 37$).

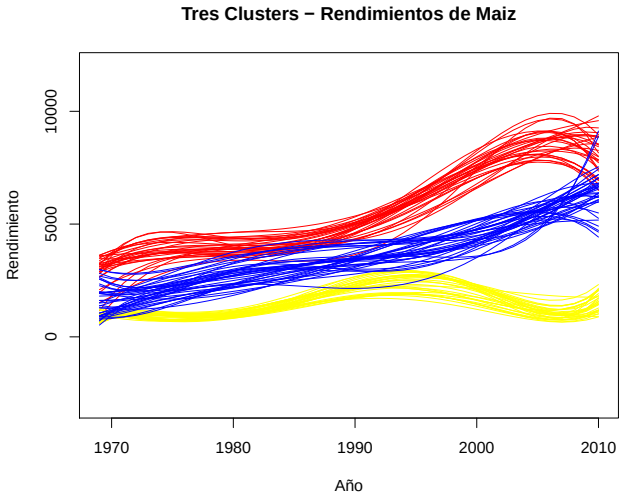

Figura 38: *Ejemplo de tres grupos de trayectorias "distintas" de rendimientos de maíz.*

El dendrograma (Figura 39) obtenido aplicando el método de Ward con distancia euclídea permite decidir con certeza que existen tres clusters ya que los saltos en altura al unir cualquiera de estos clusters es notablemente mayor que los saltos en los pasos anteriores.

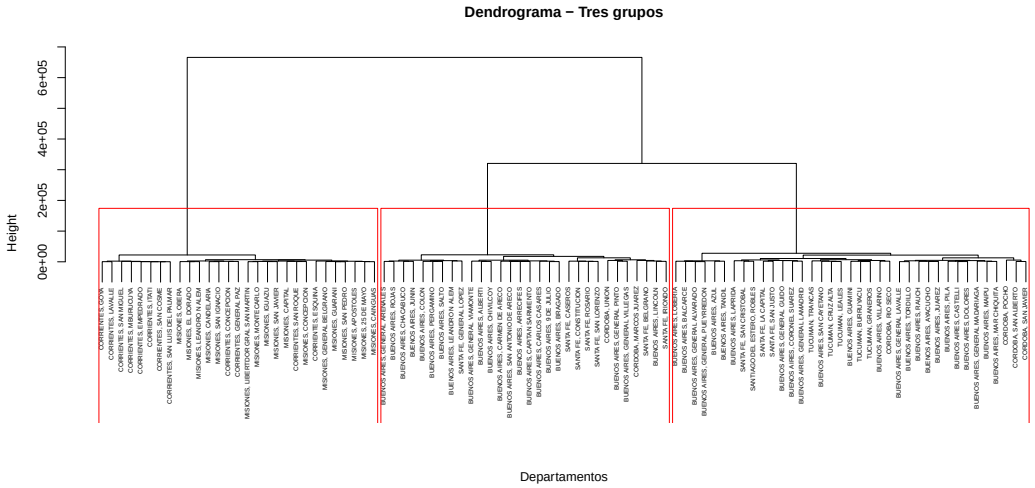

Figura 39: *Dendrograma correspondiente al ejemplo de tres clusters de trayectorias de rendimientos de maíz.*

La Figura 40, panel izquierdo, presenta $log(W)$ observado (curva roja) junto con $log(W^*)$ (curvas verdes) para las 1000 muestras simuladas a partir de la distribución de referencia que supone la no existencia de clusters. En $k = 3$ se observa un "codo" muy pronunciado; y a partir de este k los perfiles observado y simulado son paralelos.

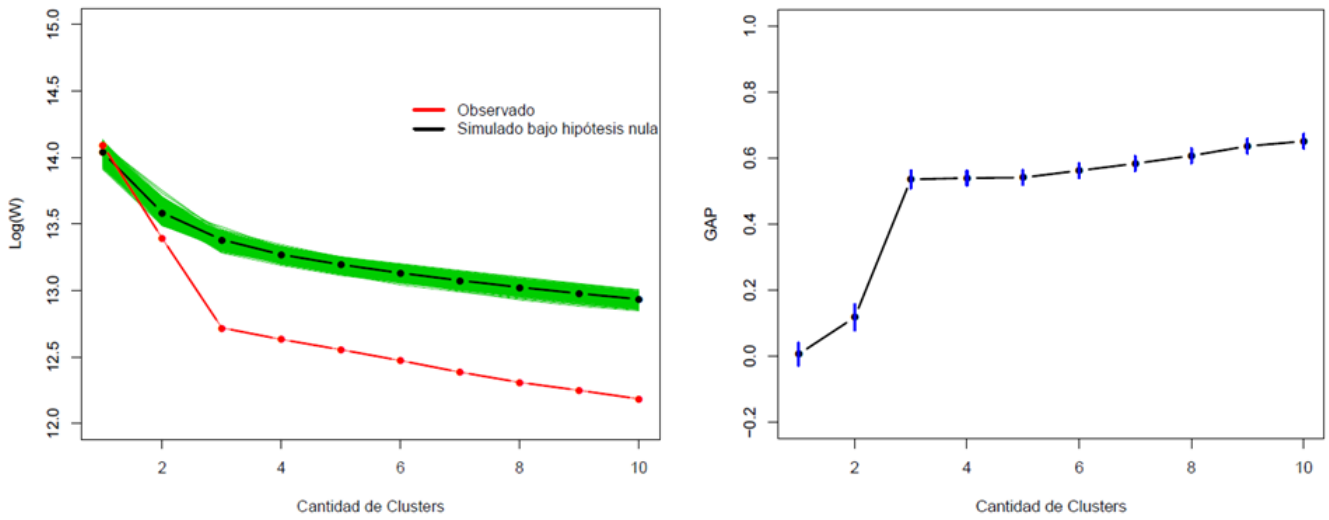

Figura 40: *Logaritmos de W observados y simulados bajo la distribución de referencia (izquierda) y gráfico del estadístico gap para los datos de la Figura 38 (tres grupos). Los segmentos azules representan los desvíos estándares de* $log(W^*)$ *bajo la distribución de referencia.*

La Figura 40, panel derecho, muestra el gráfico del estadístico gap para los datos originales (Figura 38) y los desvíos estándares de las B copias de $log(W_k^*)$. Según el criterio que considera el menor k tal que $Gap(k+1) - Gap(k) \leq s_{k+1}$ obtenemos nuevamente $k = 3$. Notemos que el criterio selecciona el último valor de k (cuando se recorre el gráfico de derecha a izquierda) en el que el salto en el gap no es importante.

La Figura 41 muestra los valores del estadístico $CH(k)$ para el ejemplo de tres grupos. Puede observarse que el máximo ocurre en $k = 3$.

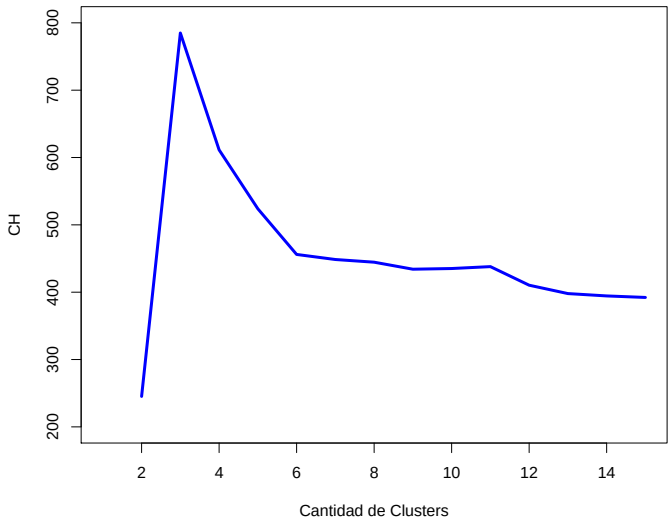

Figura 41: *Estadístico CH para el caso de tres grupos de trayectorias de rendimiento.*

A partir del criterio $CH(k)$ es posible proponer un nuevo criterio con una idea similar al método gap. La propuesta es seleccionar el **mayor** valor de k en el que el estadístico $CH_{modificado}(k)$ presente un máximo absoluto o local, que se

encuentre fuera de la banda de valores del estadístico calculados bajo la distribución de referencia. Proponemos "*CH modificado*" como

$$CH_{modificado}(k) \quad = \quad \frac{CH(k) - CH(k-1)}{CH(k)}$$

definido para $k \geq 3$, dado que $CH(k)$ está definido para $k \geq 2$. La Figura 42 muestra los valores de $CH_{modificado}(k)$ para los datos observados (azul) y las 1000 replicaciones bajo la distribución de referencia. Se puede ver en este caso que el criterio selecciona $k = 3$.

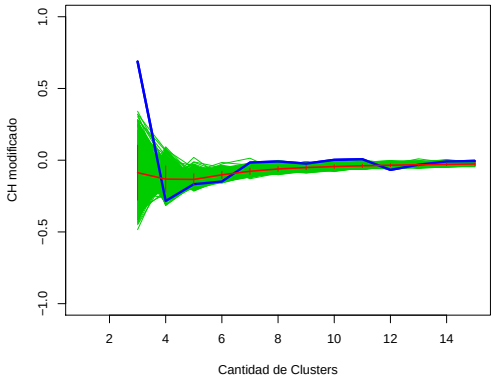

Figura 42: *CH_{modif} para datos originales (azul) y simulados bajo la distribución de referencia (verdes) para el caso de tres grupos de trayectorias de rendimiento.*

4.2.3. Siete grupos de trayectorias

Finalmente analicemos el conjunto de curvas que se muestran en la Figura 43, donde las curvas rojas, amarillas y azules, corresponden a rendimientos de maíz, y los otros cuatro grupos fueron generados con cuatro rectas con pequeñas perturbaciones de los coeficientes que las definen.

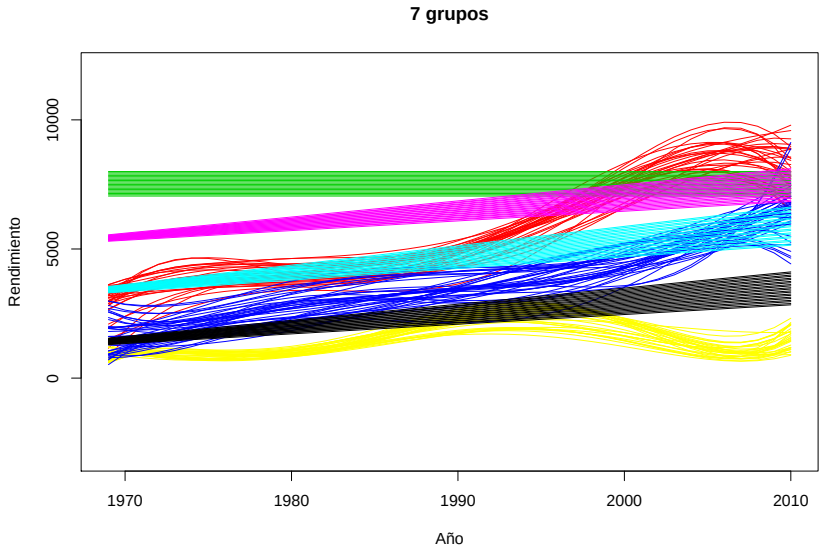

Figura 43: *Ejemplo, construido a propósito, de siete grupos de trayectorias "distintas" de rendimientos de maíz.*

Aplicaremos los tres criterios analizados anteriormente, gap, CH y CH modificado a estos datos para evaluar su performance en este ejemplo más complejo.

Al estudiar el comportamiento de $log(W)$ (Figura 44, panel izquierdo) se observa el "codo" en $k = 7$ y desde ahí en adelante las curvas de $log(W)$ y $log(W^*)$ son prácticamente paralelas.

Sin embargo, cuando se considera el criterio basado en el estadístico gap (Figura 44, panel derecho) el menor k tal que verifica la condición $Gap(k+1) - Gap(k) \leq s_{k+1}$ es $k = 1$, lo que se interpretaría como ausencia de grupos en los datos. Llama la atención que habiendo tantos saltos importantes en el valor del estadístico, el criterio seleccione un número de clusters que correspondería a una situación patológica. Cuando se excluye $k = 1$, es claro que el menor valor de k que satisface la condición de Tibshirani es $k = 7$.

Para salvar ésto, en la sección siguiente proponemos modificar levemente el criterio propuesto por el autor.

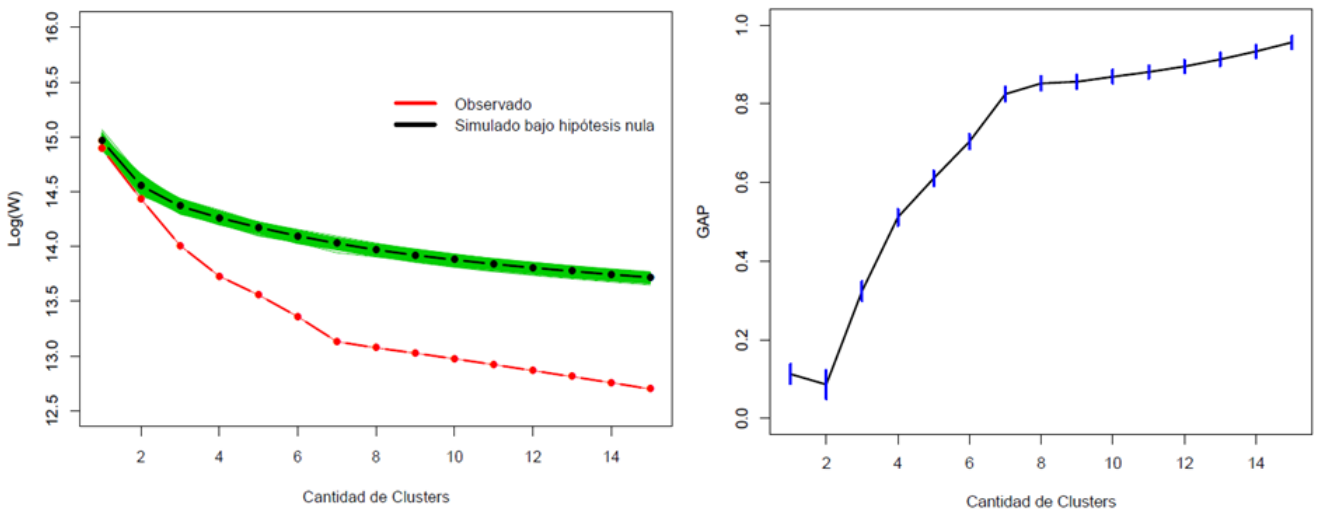

Figura 44: *Logaritmo de W observados y media de logaritmos de W^* estimadas con distribución de referencia para 7 grupos de curvas de rendimientos (izquierda). Estadístico gap para el ejemplo de 7 grupos de trayectorias de rendimientos (derecha). Los segmentos azules representan los desvíos estándares de $log(W^*)$ bajo la distribución de referencia.*

El criterio $CH(k)$ selecciona correctamente $k = 7$ como el número óptimo de clusters (Figura 45, panel izquierdo). El criterio del $CH_{modificado}$ (Figura 45, panel derecho) también conduce a seleccionar $k = 7$ como el número óptimo de clusters, ya que 7 es el mayor valor de k para el que ocurre un máximo que queda por fuera de la banda de variabilidad bajo la hipótesis nula.

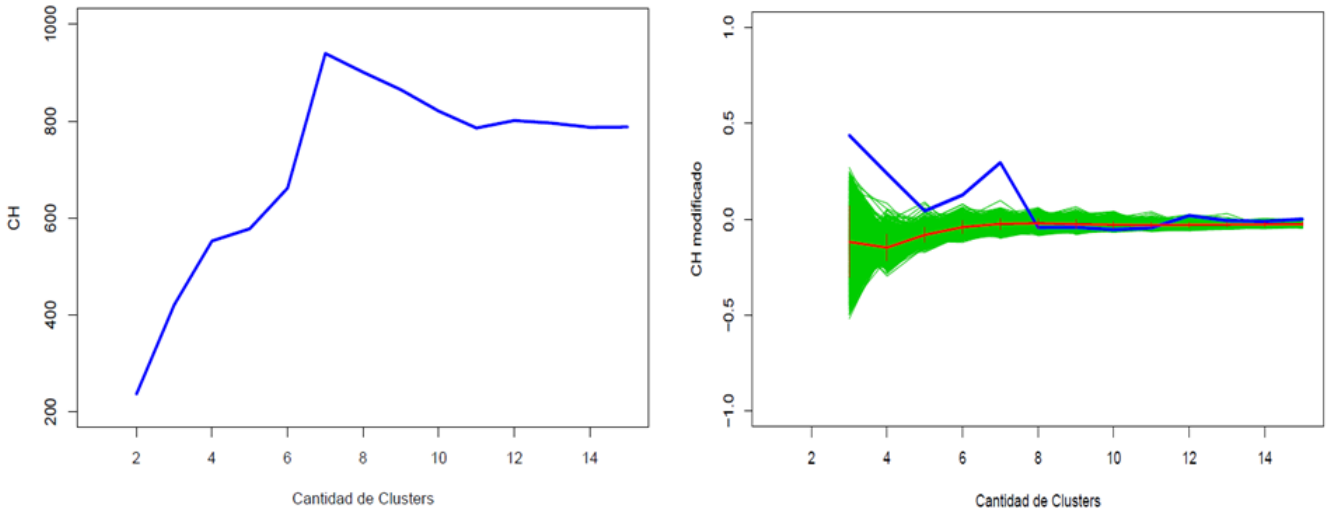

Figura 45: *Estadísticos CH (izquierda) y CH_{modif} (derecha) para el caso de 7 grupos de trayectorias de rendimiento. En azul, datos originales y en verde, datos generados desde la distribución de referencia.*

4.3. Nuevos criterios para seleccionar el número de clusters basados en el estadístico gap

En la sección anterior se mostró que el criterio propuesto por Tibshirani y otros (2001) no siempre resulta apropiado. De hecho, con el ejemplo de siete grupos mostramos empíricamente que, aún cuando existían grupos de curvas distintas, el criterio basado en el estadístico gap no detecta grupos. En esta sección presentaremos dos nuevas propuestas basadas en el estadístico gap.

La primera resulta de modificar el criterio de selección del valor óptimo de k, redefiniéndolo de manera de evitar situaciones como la que se presentó en el ejemplo anterior: *Criterio gap modificado 1*: "seleccione el **mayor** valor de k tal que $Gap(k) - Gap(k-1) \geq 2s_k$". Este criterio intenta identificar el primer "salto" realmente significativo en la curva gap (cuando ésta se recorre de derecha a izquierda), por eso proponemos duplicar s_k. Además provee al analista del número máximo de clusters a partir del cual sería razonable comenzar a caracterizar los grupos. Notemos que este criterio también selecciona como óptimo número de clusters el número correcto de grupos en los dos primeros ejemplos presentados en la sección anterior (2 grupos, Figura 36, y 3 grupos, Figura 40 (panel derecho)). Lo mismo ocurre en el caso de 7 grupos, pero no es tan claro al observar la Figura 44 (panel derecho)), por eso mostramos en la siguiente Figura 46 la curva gap con segmentos azules representando ahora dos veces el desvío.

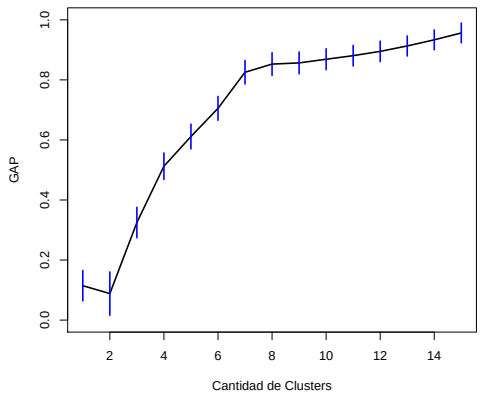

Figura 46: *Estadístico gap para el ejemplo de 7 grupos "construidos" de trayectorias de rendimientos. Los segmentos azules representan el doble de los desvíos estándares de* $log(W^*)$ *bajo la distribución de referencia.*

La segunda propuesta, consiste nuevamente en cambiar el criterio de selección de la cantidad de grupos. Consideremos la medida

$$Q = \frac{GAP(k) - GAP(k-1)}{\sqrt{\frac{sd^2(k)+sd^2(k-1)}{2}}}$$

definida para $k \geq 2$, que mide el salto entre $k-1$ y k, en términos de desvío estándar. Proponemos seleccionar el valor de k en donde Q alcanza el máximo absoluto. Llamaremos a este método *Criterio gap modificado 2*. Este criterio intenta encontrar el valor de k que produce el máximo salto relativo del estadístico gap.

Los dos criterios presentados no necesariamente coincidirán en un caso particular ya que apuntan a características diferentes. El primero definiría el máximo número de clusters que el analista debería tener en cuenta mientras que el segundo indicaría el número de clusters a partir del cual las fusiones generan un cambio extremadamente alto en el estadístico gap. Este último caso podría pensarse como el mínimo valor de k a ser considerado por el analista.

Las Figuras 47, 48 y 49 muestran el perfil del estadístico Q para los ejemplos de dos, tres y siete grupos respectivamente. En el caso de dos y tres grupos el máximo se produce en $k = 2$ y $k = 3$ respectivamente.

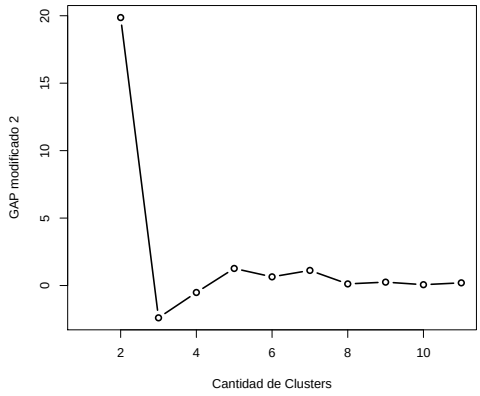

Figura 47: *Curva* $GAP_{modif-2}$ *para el ejemplo de dos grupos de trayectorias de rendimientos.*

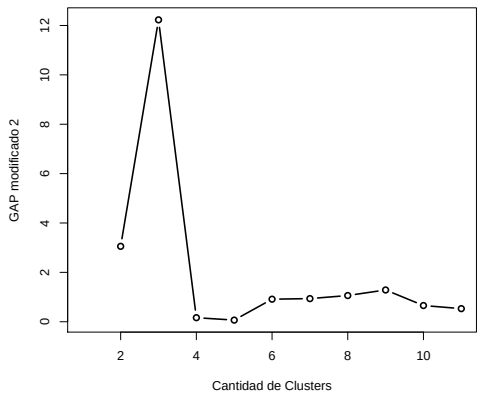

Figura 48: *Curva $GAP_{modif-2}$ para el ejemplo de tres grupos de trayectorias de rendimientos.*

En el ejemplo de 7 grupos (Figura 49) se observa el máximo absoluto en $k = 4$. A partir de los dos criterios la recomendación entonces sería considerar algún número de clusters entre cuatro (criterio 2) y siete (criterio 1) grupos, en este caso, tener cuatro se asocia a los cuatro "niveles" de rendimientos, el cual finalmente debería ser definido en base a conocimiento experto en el tema.

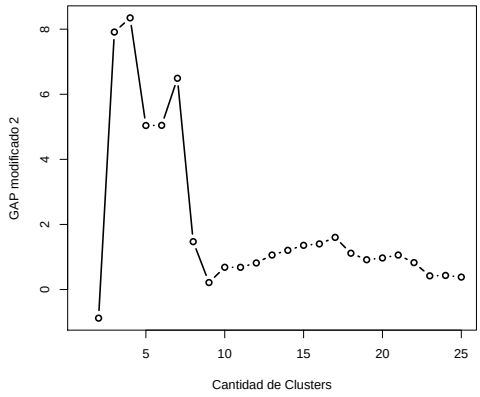

Figura 49: *Curva $GAP_{modif-2}$ para el ejemplo de 7 grupos de trayectorias.*

Observemos que en la siguiente Figura 50 se muestra, para el ejemplo construido con 7 grupos "distintos" de trayectorias, que al considerar 4 clases se prioriza el nivel o altura de las curvas, por lo que si deseamos además diferenciar "forma o tendencia" necesitamos tener más clusters y por ende tenemos que aumentar la cantidad de grupos. Ello requiere el conocimiento experto del analista quien en última instancia determinará cual es la partición que provee la cantidad de agrupamientos razonable para los datos.

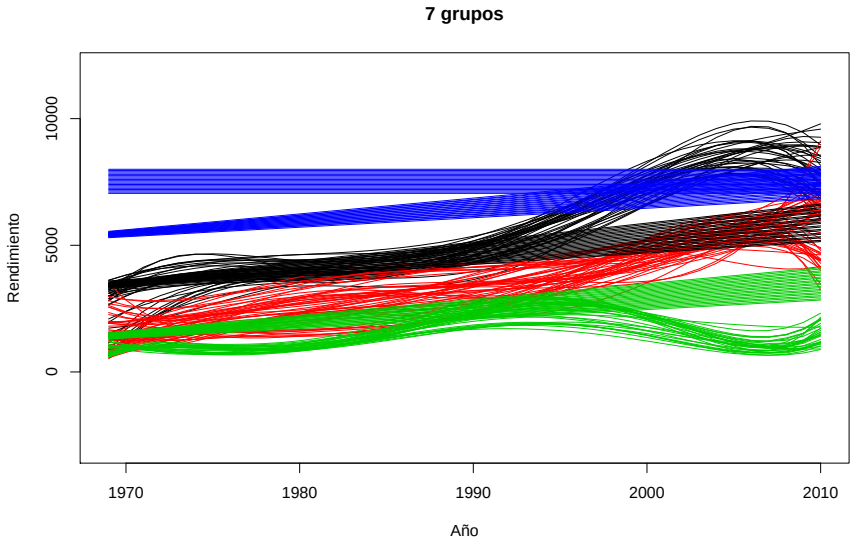

Figura 50: *Segmentación en 4 grupos en el ejemplo, construido a propósito, de siete grupos de trayectorias "distintas" de rendimientos de maíz.*

5. Análisis de los datos de rendimiento de maíz en Argentina, 1969-2010

5.1. Segmentación y selección del número de clusters

En esta sección aplicaremos la metodología propuesta para seleccionar el número de clusters a los datos de curvas de rendimientos de maíz. La Figura 51 muestra el dendrograma que se obtiene como salida del procedimiento de Ward. A la izquierda se muestran particiones definidas por $k = 2$ y $k = 3$; a la derecha la partición de $k = 7$ grupos. No es claro a partir del dendrograma cuántos grupos elegir.

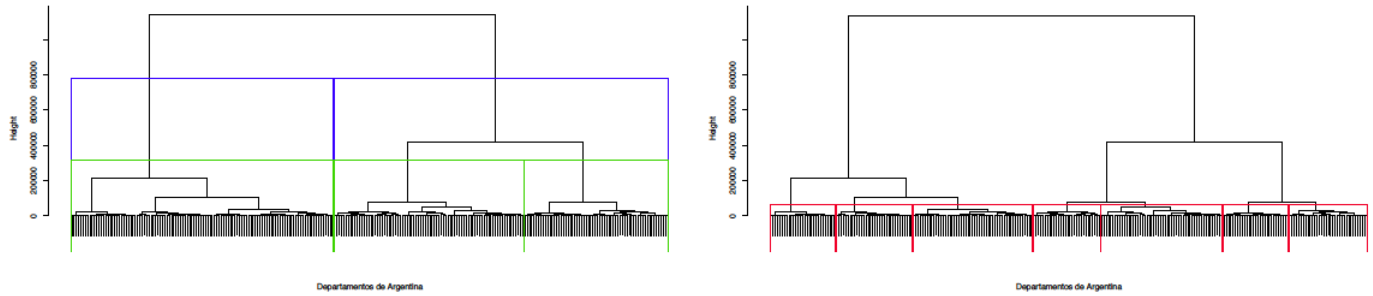

Figura 51: *Dendrograma según método de Ward aplicado a curvas ajustadas de rendimientos de maíz. Argentina 1969-2010. Izquierda: se muestran dos particiones: $k = 2$ (azul) y $k = 3$ (verde). Derecha: se muestra una partición en $k = 7$ grupos.*

Aplicamos el procedimiento descripto en las secciones anteriores para generar datos desde la distribución de referencia. La Figura 52, panel izquierdo, muestra el $log(W)$ para los datos de rendimiento de maíz cuando el número de clusters es $k = 1, ..., 15$. A partir de este gráfico no es posible identificar un cambio pronunciado en $log(W)$ en el rango de k considerado, aunque pareciera que el perfil observado y el perfil del $log(W)$ simulado bajo hipótesis nula son razonablemente paralelos a partir de $k = 8$.

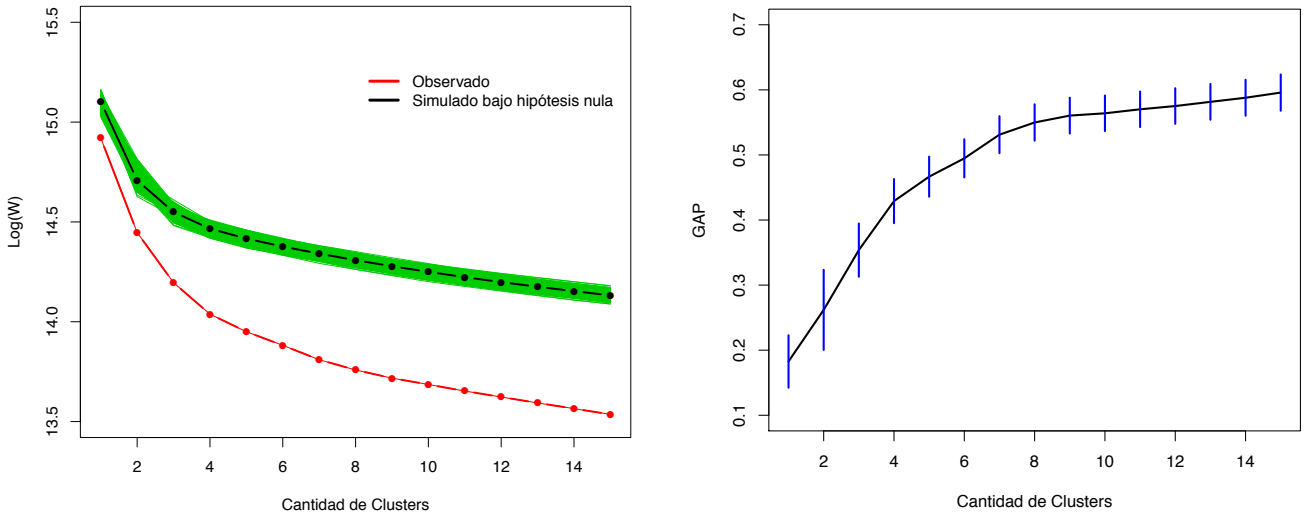

Figura 52: *Izquierda: $Log(W)$ observados y simulados bajo distribución de referencia. Derecha: Estadístico gap. Los segmentos azules representan los desvíos estándares de $log(W^*)$ bajo la distribución de referencia. Datos: Trayectorias de rendimientos de maíz, Argentina 1969-2010.*

El panel derecho de la Figura anterior, muestra los valores del estadístico gap y los desvíos estándares estimados a partir de $B = 1000$ réplicas de $n = 264$ curvas generadas aleatoriamente a partir de la distribución de referencia. Recordemos que el criterio propuesto por Tibshirani y otros para seleccionar el número óptimo de clusters es el menor k que verifica la condición $Gap(k+1) - Gap(k) \leq s_{k+1}$. En este caso la decisión sería seleccionar $k = 8$.

Cuando se utiliza el *criterio gap modificado 1* propuesto en la sección anterior "seleccione el **mayor** valor de k tal que $Gap(k) - Gap(k-1) \geq 2s_k$", el k óptimo nuevamente resulta ser 7. Notemos que, por otro lado, siendo más exigentes en el criterio de Tibshirani, y considerando la condición $Gap(k+1) - Gap(k) \leq 2s_{k+1}$ nuevamente obtenemos $k = 7$.

La Figura 53 muestra el estadístico correspondiente al nuevo *criterio gap modificado 2* presentado en la sección anterior. Este criterio identifica como máximo global a $k = 4$. En consecuencia, los criterios modificados 1 y 2 nos indican que deberíamos considerar entre 4 y 7 grupos para segmentar estos datos.

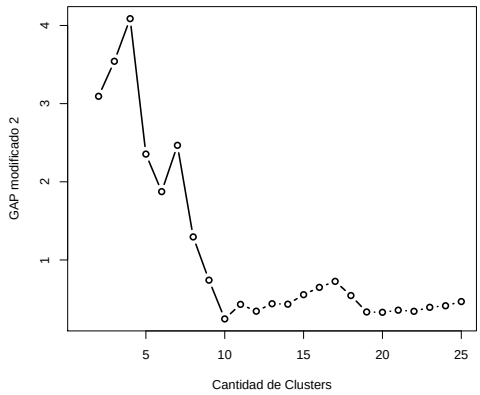

Figura 53: *Curva $GAP_{modif-2}$ para el ejemplo de trayectorias originales de rendimientos.*

La curva del método CH, en la Figura 54, panel izquierdo, alcanza el máximo en $k = 3$. Mientras que en el panel derecho

de la misma Figura 54 el estadístico CH_{modif} identifica $k = 7$ como el mayor valor de k en el que se observa un máximo no incluido en la envoltura de valores generadas bajo la hipótesis nula.

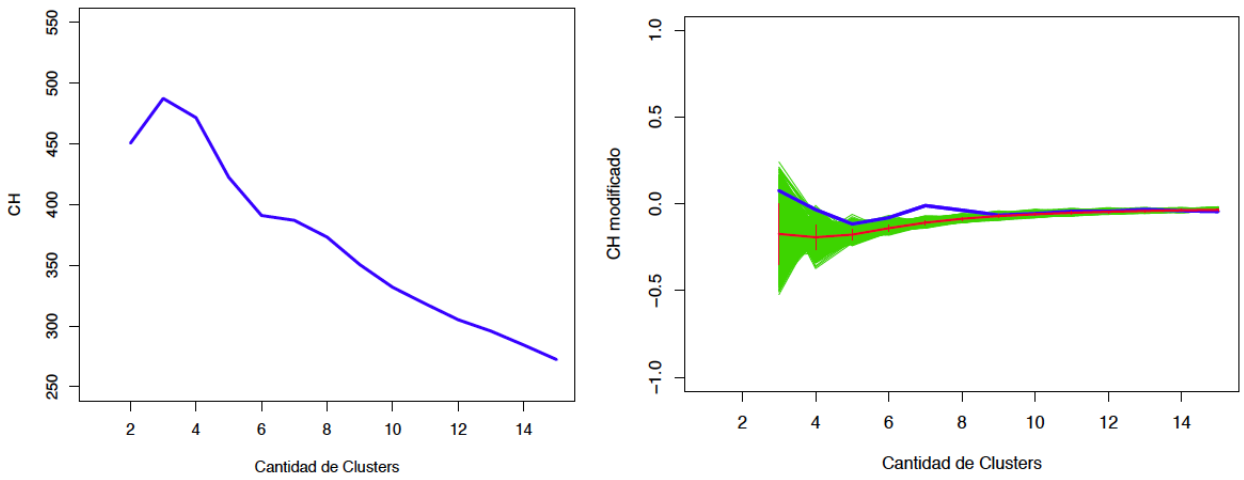

Figura 54: *Izquierda: estadístico CH para el caso de trayectorias originales de rendimiento. Derecha: estadístico* CH_{modif} *para el caso de trayectorias originales de rendimiento. En azul, datos originales, y en verde, datos generados desde la distribución de referencia.*

En resumen, el criterio gap obtiene $k = 8$, pero duplicando el desvío se selecciona $k = 7$. Los criterios gap modificado 1 y CH modificado devuelven $k = 7$ como número óptimo de clusters. CH identifica $k = 3$. Y finalmente el criterio gap modificado 2 identifica a $k = 4$ como el mínimo número de clusters a considerar. En un problema de segmentación real los criterios sirven para orientar en cómo seleccionar el número de clusters. En nuestro caso la evidencia apunta a seleccionar entre un mínimo de 4 y un máximo de 7 grupos, con gran coincidencia de los distintos criterios en que el valor de k es 7. En la práctica, el analista toma la decisión en base al conocimiento experto en el tema y a partir de un análisis en profundidad de los grupos que devuelve la segmentación.

5.2. Una mirada más detallada de los datos de trayectorias de rendimientos de maíz

La discusión con expertos en el tema revela que el número de agrupamientos de patrones de datos de rendimientos de maíz (Argentina) no debería ser muy pequeño, porque se sabe que: 1) ha habido zonas del país con crecimiento exponencial en la producción por hectárea, mientras que en otras zonas el rendimiento ha crecido a tasa lineal; 2) cuando se considera una dada campaña del período considerado, existe una gran variabilidad en los rendimientos, especialmente en los últimos 10 años del período bajo estudio (por ejemplo en el año 2005 los rendimientos van desde 1000 a 9800 kg/ha.). Por otra parte, estamos interesados en un número no demasiado grande de grupos que nos permita entender qué estrategias se pueden aplicar en cada grupo de acuerdo a sus características principales y por lo tanto brindar una solución simple para la posterior toma de decisiones en la práctica. En consecuencia, decidimos considerar 7 grupos de curvas. En la Figura 55 se presentan todas las curvas ajustadas de los siete grupos.

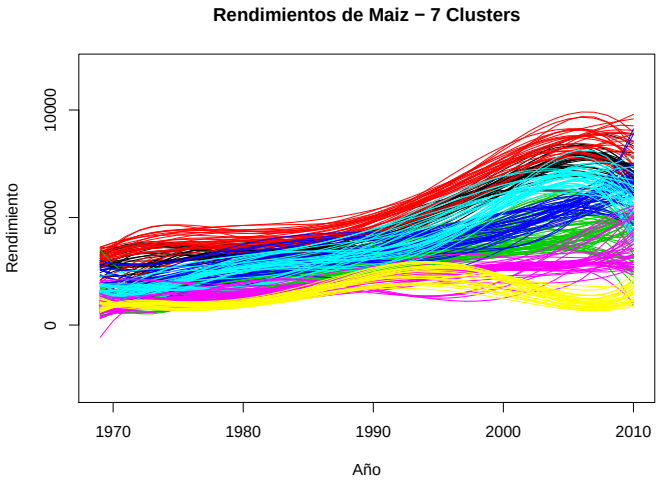

Figura 55: *Trayectorias originales coloreadas según pertenencia a cada grupo.*

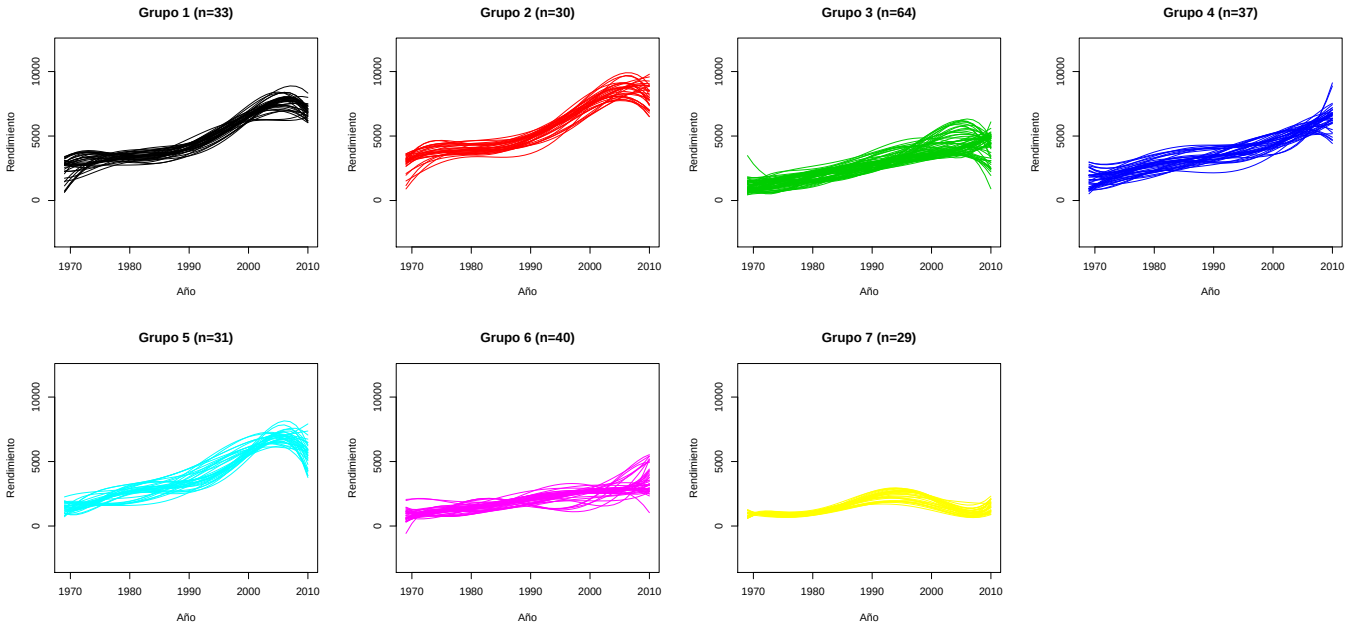

Figura 56: *Resultado de la segmentación de trayectorias de rendimiento de maíz, siete grupos.*

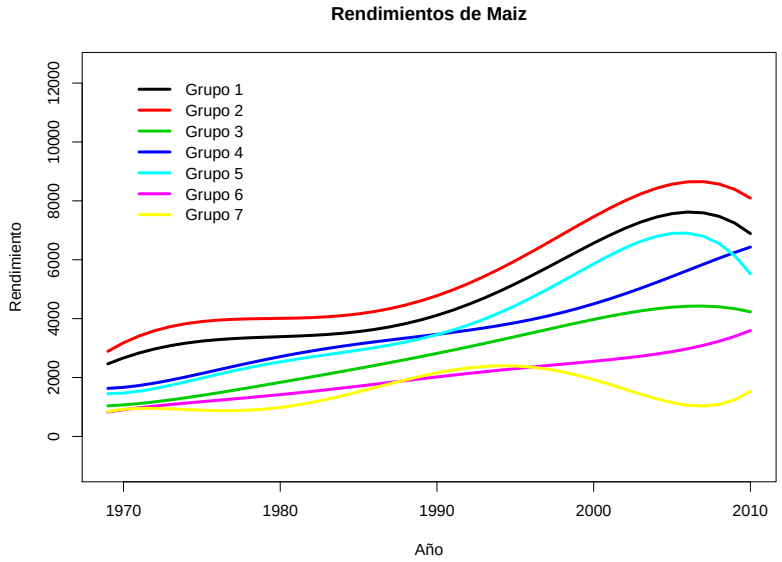

Figura 57: *Trayectoria promedio de rendimientos para cada grupo.*

La Figura 56 muestra las trayectorias de rendimiento de maíz de los 264 departamentos para una partición en siete grupos, mientras que la Figura 57 muestra la trayectoria promedio o la curva representativa de cada uno de los siete grupos. Se pueden observar diferencias tanto en la forma o tendencia de las curvas como en el nivel (altura). Así, por ejemplo, los departamentos agrupados en el cluster 1 (negro) y el 2 (rojo) presentan el mismo perfil típico, pero una diferencia en el nivel de rendimientos. El cluster 5, presenta un perfil similar al de los dos anteriores pero recién a partir de 1990, nuevamente la altura de las curvas difiere, indicando departamentos con menos rinde. Los departamentos que componen el cluster 4 (azul) y el 6 (magenta) muestran la misma tendencia pero con distinta velocidad de crecimiento. La trayectoria promedio para los departamentos agrupados en el cluster 3 (verde) muestra crecimiento lineal, mientras que la del cluster 7 (amarillo) se diferencia claramente de todas las demás porque corresponde a departamentos con bajo rendimiento y que no han tenido un incremento sustancial del rinde en el período considerado.

Si decidiéramos considerar 4 o 3 grupos, la secuencia de fusiones a partir del paso en que se tienen 7 clusters resulta ser: 1) se unen los grupos 1 y 5; 2) se unen los grupos 6 y 7; 3) se unen los grupos 3 y 4 y 4) se unen los grupos 2 y el grupo resultante de unir 1 y 5. De lo que deducimos que parece importar más la altura que la forma de la curva pues la segmentación prioriza los niveles por sobre la tendencia al momento de unir dos grupos. Este hecho ya se había observado al realizar la segmentación para el ejemplo ficticio de siete grupos, donde al intentar separar en 4 clusters, se confirma que se privilegia el nivel o altura de la curva en lugar de su forma.

5.3. Nuevo análisis basado en los rendimientos relativos

En esta sección proponemos aplicar la misma metodología para identificar grupos en base a la "tendencia" de los rendimientos. Entendemos por *tendencia o rendimiento relativo* del departamento $i-ésimo$ al vector que se obtiene al restar la media de los rendimientos ajustados

$$\mathbf{y}_i^* \quad = \quad \widetilde{\mathbf{y}}_i - \sum_{t=1}^{42} \frac{\widetilde{y}_{it}}{42},$$

con $\widetilde{\mathbf{y}}_i^{\mathrm{t}} = (\tilde{y}_{i1}, \tilde{y_{i2}}, \cdots, \tilde{y}_{i42}) \in \mathbb{R}^{42}$ el vector de rendimientos ajustados para el departamento $i-$ésimo en los 42 años del período bajo estudio.

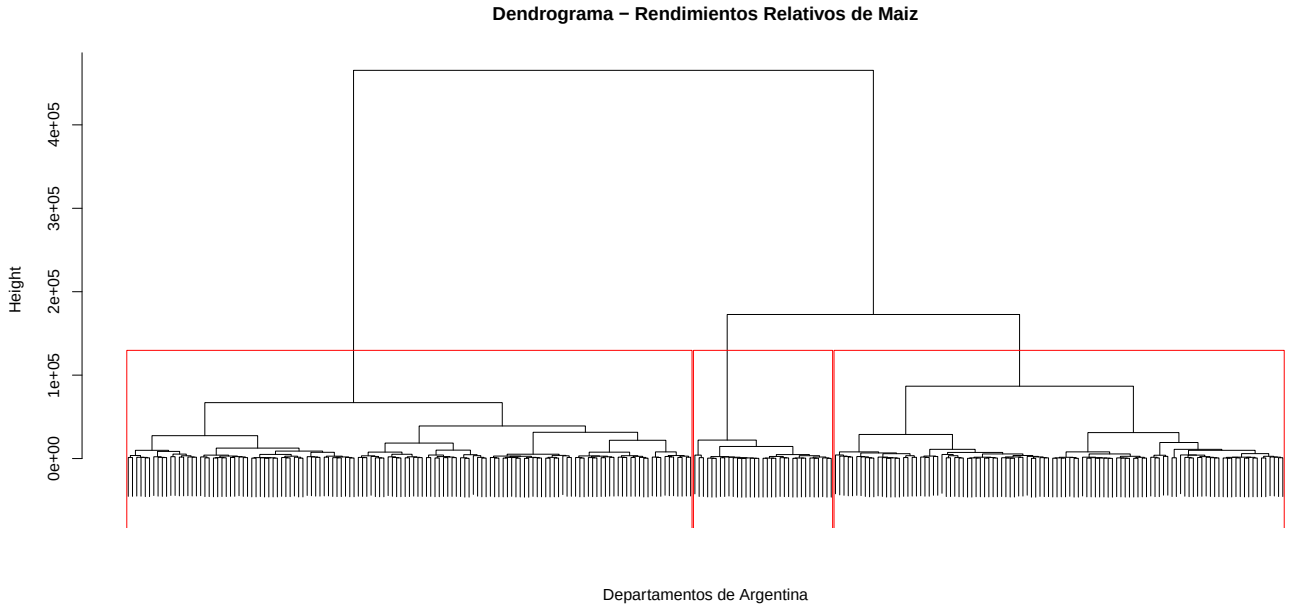

Figura 58: *Dendrograma según método de Ward aplicado a rendimientos relativos de maíz. Argentina 1969-2010. Se muestra una partición de $k = 3$ grupos.*

La Figura 58 muestra el dendrograma que se obtiene al aplicar el método de segmentación jerárquica de Ward a los datos de rendimientos relativos. Cuando se aplican los métodos descriptos anteriormente para seleccionar el número de clusters se obtiene: $k = 3$ con el criterio de gap modificado 1 (Figura 59), $k = 2$ con el criterio de gap modificado 2 (Figura 60, panel izquierdo), y $k = 2$ con el método CH (Figura 60, panel derecho). La estrategia de CH modificado no alcanza máximo fuera de la envolvente de variabilidad bajo la hipótesis nula, por lo que no consideramos este análisis. En consecuencia, podemos decir que como mínimo existen 2 o 3 clases. Dado que el dendrograma muestra 3 grandes grupos y que ésto se asocia con la detección inicial de tres tendencias de crecimiento, exponencial, lineal y estable, parece razonable decidir trabajar con 3 clusters.

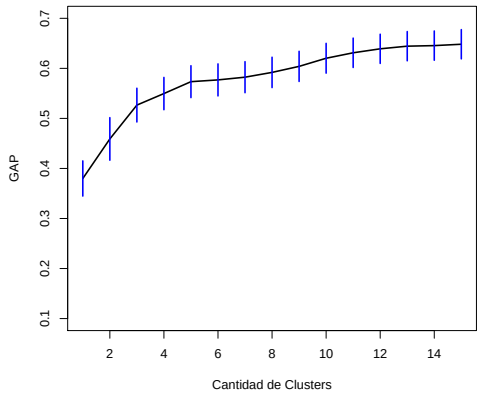

Figura 59: *Estadístico gap para trayectorias de rendimientos relativos de maíz. Argentina 1969-2010. Los segmentos azules representan el doble de los desvíos estándares de $log(W^*)$ bajo la distribución de referencia.*

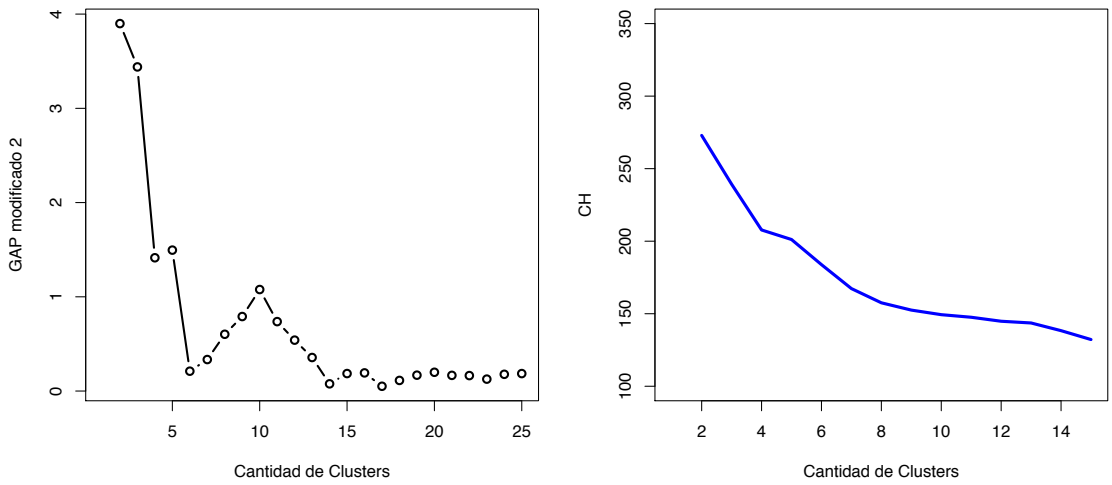

Figura 60: *Curva $GAP_{modif-2}$ (izquierda) y estadístico CH (derecha) para el ejemplo de trayectorias de rendimientos relativos.*

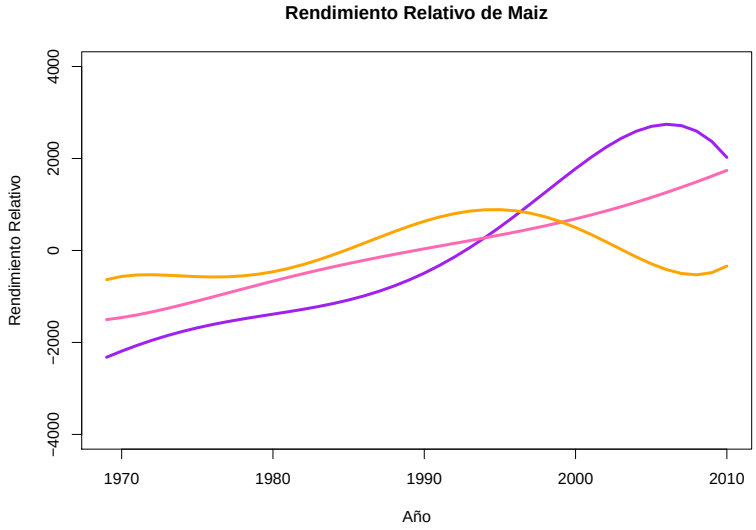

Figura 61: *Prototipo de cada grupo de trayectorias de rendimientos relativos.*

La Figura 61 muestra las curvas de rendimientos relativos promedios para cada cada una de las tres clases. Notemos que se observan dos tendencias crecientes, una con comportamiento lineal (curva rosa), y la otra con un crecimiento parecido al de una función exponencial a partir del año 1990 (curva violeta). Por otra parte, la trayectoria promedio del tercer grupo (curva naranja) con una tendencia más estable y presentando decrecimiento respecto de los últimos años.

5.4. Interpretación de los grupos de curvas obtenidos en los dos análisis

La Figura 62 muestra los resultados de los dos análisis. Es interesante resaltar que en Argentina los patrones de rendimiento de maíz en las últimas cuatro décadas presentan cambios notables de nivel, es decir el rendimiento promedio por departamento es muy variable. Notemos que el perfil de rendimientos de los grupos 1, 2 y 5 (curvas coloreadas en negro, rojo y celeste respectivamente) obtenidos a partir de la segmentación basada en rendimientos ajustados se corresponden con el

grupo 1 (curva violeta) definido a partir de la segmentación basada en rendimientos relativos (tendencias). Mientras que, la tendencia aproximadamente lineal observada en los grupos 3, 4 y 6 (curvas verdes, azules y magentas respectivamente) se corresponde con la clase 2 (curva rosa) del análisis de rendimientos relativos. Finalmente, el grupo 7 (curvas amarillas) del análisis de rendimientos ajustados corresponde al grupo de la clase 3 (curva naranja) del análisis de rendimientos relativos.

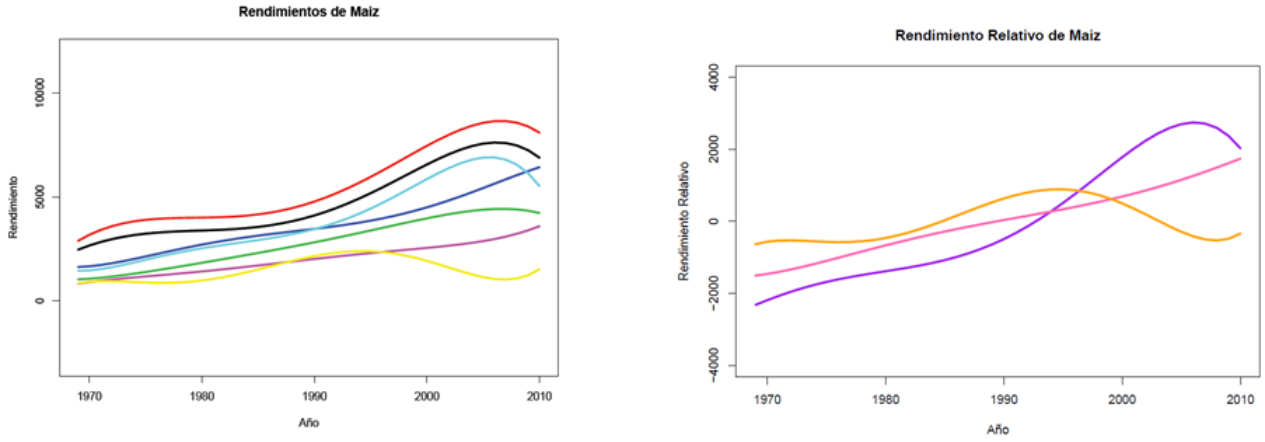

Figura 62: *Curvas promedio basadas en rendimientos (izquierda) y en rendimientos relativos (derecha) de maíz en Argentina, 1969-2010.*

La Figura 63 muestra el mapa de Argentina con los departamentos considerados en este análisis coloreados según la pertenencia a los grupos identificados en la segmentación. El mapa de la izquierda corresponde a los siete grupos obtenidos usando rendimientos ajustados y el mapa de la derecha a los tres grupos obtenidos en base a rendimientos relativos. Los colores representados en el mapa se corresponden con los de las curvas promedio de la Figura 62. Los departamentos coloreados en rojo (Grupo 2), negro (Grupo 1) y celeste (Grupo 5) presentan los rendimientos de mayor nivel en las dos últimas décadas y se asocian con regiones donde el crecimiento fue exponencial. Estos departamentos se ubican en la zona núcleo de la provincia de Buenos Aires, gran parte de las provincias de Córdoba, Santa Fé y Entre Ríos (departamentos coloreados en violeta en el mapa de la derecha). Los departamentos que evolucionaron de acuerdo a un crecimiento lineal moderado (Grupo 3, verde; Grupo 4, azul y Grupo 6, magenta) corresponden a departamentos del sur de Buenos Aires, San Luis, La Pampa, Jujuy, Salta, Tucumán, Santiago del Estero y Chaco. Por último, el grupo de departamentos en el que los rendimientos presentan tendencia estable o decreciente, corresponde a la zona norte de la región mesopotámica (Grupo 7, amarillo). Es interesante destacar que la mayoría de las inversiones tecnológicas que ocurrieron a partir de los años 90 se realizó en la zona núcleo (área negra y roja del mapa de la derecha), que es la zona más propicia para el desarrollo del cultivo de maíz. Esto explicaría la tendencia de tipo exponencial que se observa en estas zonas.

Notemos que los grupos resultantes del análisis de segmentación muestran una clara asociación con la ubicación geográfica de los departamentos. Este hecho es consistente con lo que esperábamos observar, ya que departamentos cercanos geográficamente comparten características de suelo, clima, innovaciones tecnológicas, políticas agropecuarias, etc. Es decir, se espera que haya cierta contigüidad espacial en el comportamiento del fenómeno bajo estudio. El hecho que el método de segmentación haya producido clusters de departamentos que presentan un razonable nivel de contigüidad espacial o muy fuerte contigüidad espacial (Figura 63, mapa de la izquierda y la derecha respectivamente) es indicativo de que los agrupamientos obtenidos correlacionan fuertemente con características de los departamentos, las cuales no fueron tenidas en cuenta en este análisis.

En la sección siguiente proponemos una metodología para estudiar el grado de contigüidad espacial de un agrupamiento.

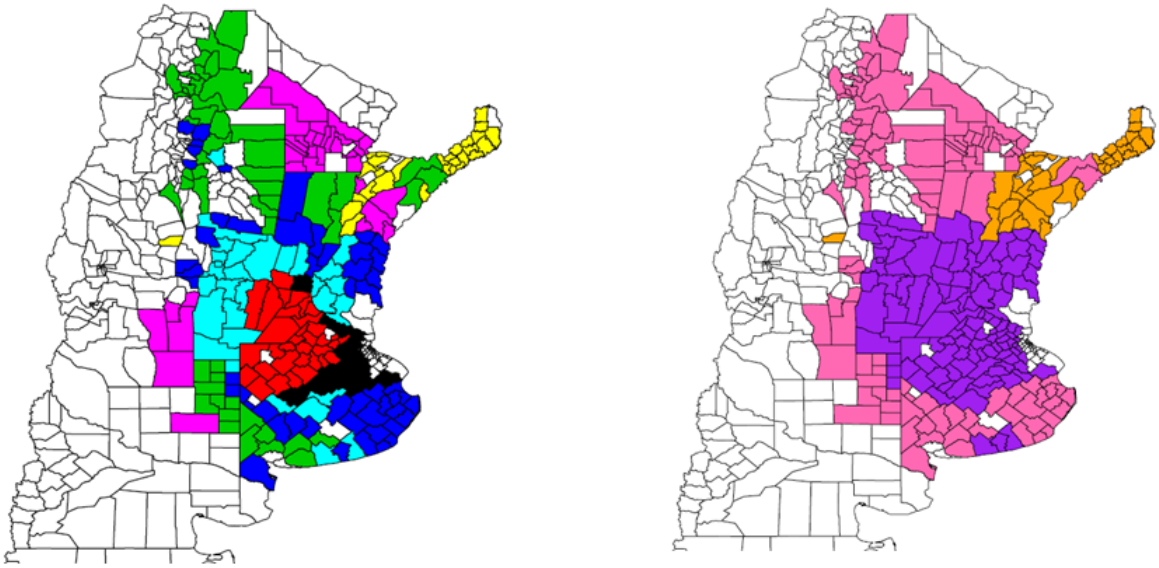

Figura 63: *Departamentos con rendimientos de maíz estimados coloreados de acuerdo al grupo de nivel (izquierda) y según su clase de tendencia (derecha).*

5.5. Propuesta para evaluar contigüidad espacial

Diremos que dos departamentos son contiguos si son limítrofes, es decir, si comparten un "límite" de acuerdo a la división político-administrativa del territorio. La Figura 63 (mapa de la izquierda) muestra que los departamentos de un dado cluster presentan cierto grado de contigüidad espacial ya que en general se ubican en una región geográfica. El mapa sugiere que los departamentos pertenecientes a un cluster, están esencialmente cerca, más cerca de lo que uno esperaría si la distribución geográfica de las curvas de rendimientos fuera al azar. Dicho de otra manera, las distancias geográficas entre departamentos de un mismo cluster, son mucho más chicas que las distancias entre departamentos de clusters distintos. Queremos probar que ésto, que es un concepto absolutamente intuitivo, porque uno espera que los rendimientos esperados de un departamento sean similares a los rendimientos esperados de un departamento cercano o de la misma zona geográfica, efectivamente es así. Es decir, queremos demostrar que en los grupos de departamentos, obtenidos a partir de las trayectorias de rindes, la contigüidad geográfica es mucho mayor que la que se esperaría si dicha caracterización fuera al azar.

Es fundamental destacar que en la metodología de clusterización, la noción de contigüidad geográfica no fue utilizada. El conocimiento experto indica que los factores que afectan las tendencias de rendimientos poseen un fuerte componente espacial. Por lo que confirmar la existencia de contigüidad espacial es una verificación indirecta de la validez de la información de rendimientos, así como de la metodología de segmentación propuesta.

Proponemos usar un test estadístico para estudiar la contigüidad espacial. Basaremos la presentación en el caso de $k = 7$ clusters y representaremos la distancia entre dos departamentos, contiguos o no, como la distancia entre sus centroides geográficos.

Sean C_k con $1 \leq k \leq 7$ los clusters definidos por el algoritmo de segmentación y sea n_k es el número de departamentos en el cluster $k-$ésimo. Sea d_{ij} la distancia geográfica en km entre el departamento $i-$ésimo y el departamento $j-$ésimo, $1 \leq i < j \leq 264$. Sea $\mathbf{u}^{\mathrm{t}} = (u_1, u_2, \cdots, u_{264})$ el vector con las etiquetas asociadas a cada departamento, donde u_i indica el cluster al que pertenece el departamento $i-$ésimo. Usaremos, en lo que sigue, el supraíndice d para indicar la noción de distancia "dentro" de los clusters, y el supraíndice e para distancia "entre" clusters.

Sean $\mathbf{d}^d$ un vector de dimensión N^d que contiene las distancias entre los departamentos dentro del mismo cluster, $N^d = \sum_{k=1}^{7} \begin{pmatrix} n_k \\ 2 \end{pmatrix}$ y $\mathbf{d}^e$ un vector de dimensión N^e que contiene las distancias entre departamentos pertenecientes a distintos

clusters $N^e = \sum_{k=1}^{7} \sum_{l \neq k} n_k n_l$.

Sea $\mathbf{u}_b$ una permutación del vector $\mathbf{u}$, $b = 1, ..., 1000$. Cada permutación produce un nuevo agrupamiento de los departamentos, ya que lo que hacemos es permutar el rótulo del cluster. De este modo aleatorizamos la pertenencia a los clusters de los departamentos, respetando la cantidad de elementos en cada cluster, y manteniendo el número de clusters. Para cada permutación $\mathbf{u}_b$ calculamos el nuevo vector de distancias, $\mathbf{d}_b^d$ y $\mathbf{d}_b^e$.

Consideremos en primer lugar las distancias entre elementos de un mismo cluster. La Figura 64 muestra la distribución acumulada de las distancias calculadas bajo la segmentación original (curva roja) y las distribuciones acumuladas de las distancias mínimas (curva celeste), medias (curva azul) y máximas (curva magenta) obtenidas en las 1000 permutaciones. Además en línea punteada se presentan los percentiles 0.05 y 0.95. Notemos que la distribución empírica (curva roja) se encuentra absolutamente alejada de lo esperado bajo la hipótesis nula de no contigüidad espacial en los elementos que componen los clusters, en consecuencia la hipótesis de no contigüidad debería rechazarse con fuerza. En particular, la mediana de las distancias entre departamentos de un cluster, bajo la distribución de aleatorización se mueve en una pequeña banda alrededor de 600 km, en tanto que la mediana de las distancias en los clusters originales es aproximadamente 400 km.

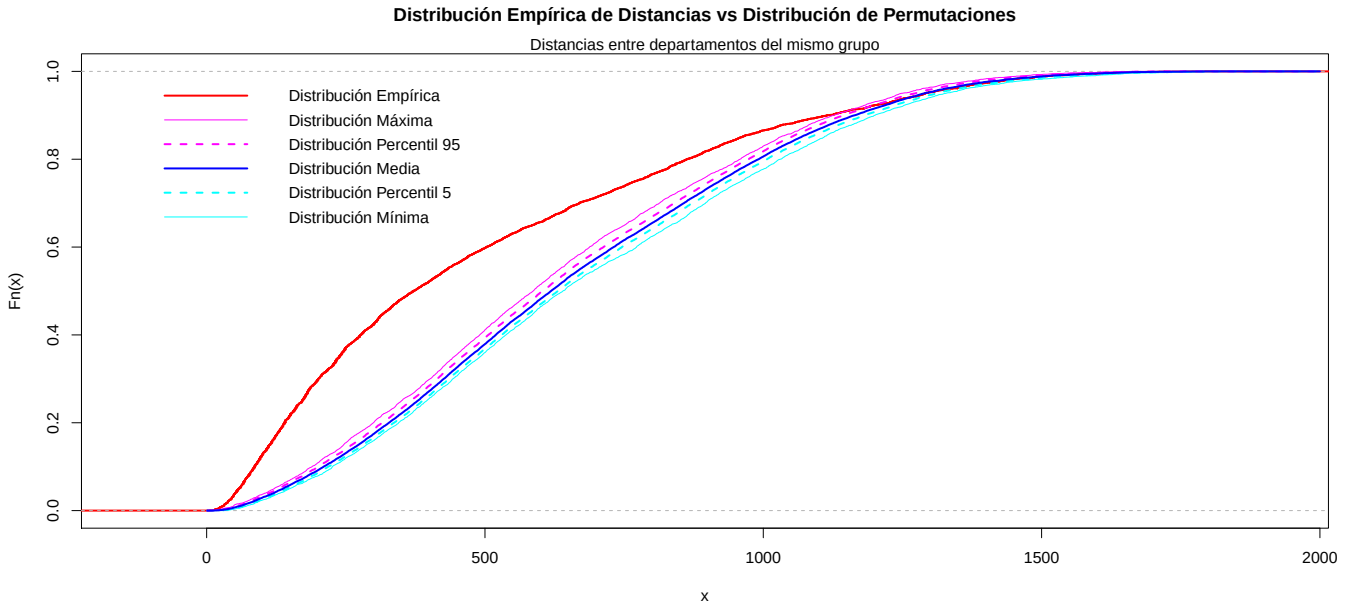

Figura 64: *Distribución acumulada empírica de distancias entre departamentos de cada grupo (curva roja) y distribuciones acumuladas resultado de las 1000 permutaciones.*

Consideremos ahora una comparación entre las distancias geográficas entre departamentos del mismo cluster $\mathbf{d}^d$ y las distancias entre departamentos pertenecientes a distintos clusters $\mathbf{d}^e$. En la Figura 65 se presenta la distribución acumulada de las distancias dentro de los clusters y de las distancias entre los clusters para la clasificación resultante de la segmentación (gráfico de la izquierda) y para las 1000 permutaciones (gráfico de la derecha), considerando distancias entre departamentos de distintos clusters. Nuevamente se concluye que la distribución observada de distancias (izquierda) es claramente diferente de la distribución esperada bajo la hipótesis nula de no contigüidad espacial.

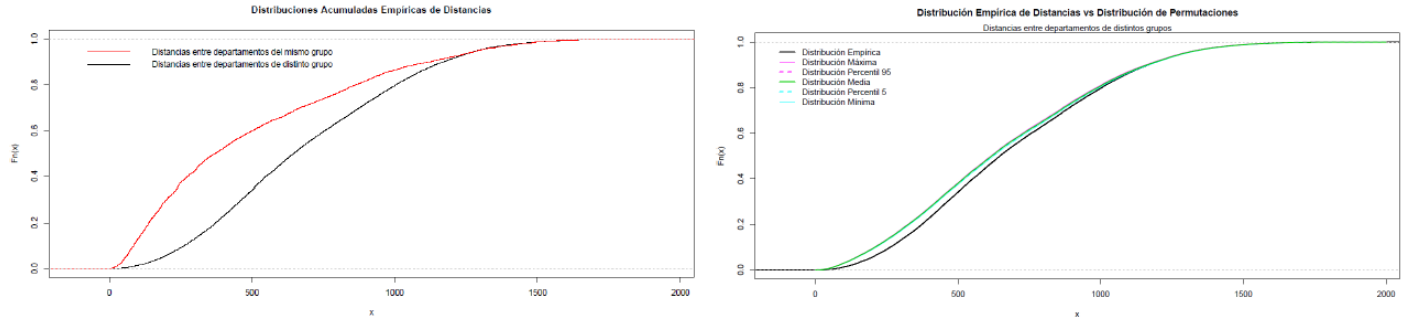

Figura 65: *Izquierda: Distribuciones acumuladas de distancias entre departamentos de cada grupo y entre departamentos pertenecientes a clusters distintos (clasificación resultante de la segmentación). Derecha: resultado de las 1000 permutaciones para distancias entre departamentos de distintos grupos.*

Es importante notar que cuanto más extenso sea el territorio de un departamento, mayor será la distancia desde su centroide al centroide de cualquier otro departamento (Figura 66). Es decir, las distancias entre departamentos no son independientes. Por este motivo se utiliza un test estadístico de aleatorización basado en los valores observados empíricos, y no otro test que pudiera requerir independencia de los datos. A continuación proponemos un test de diferencia de medias, basado en la distribución de aleatorización, para evaluar la hipótesis nula de no contigüidad espacial en los clusters.

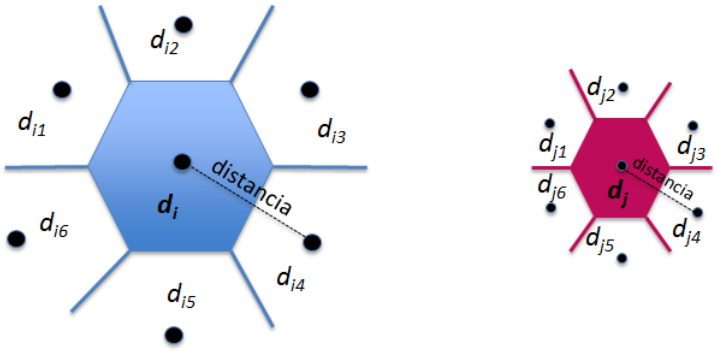

Figura 66: *La distancia entre departamentos depende de la extensión de sus territorios.*

Recordemos que las componentes de $\mathbf{d}^d$ y $\mathbf{d}^e$ son las distancias geográficas entre departamentos del mismo cluster y entre departamentos de distintos clusters respectivamente. Sean D^d y D^e las correspondientes variables aleatorias. Formulamos las hipótesis

$$H_0 : E\left(D^d\right) = E\left(D^e\right) \qquad H_1 : E\left(D^d\right) < E\left(D^e\right).$$

La idea consiste en que bajo la hipótesis nula de falta de contigüidad espacial, la media de las distancias de los elementos dentro de los clusters es igual a la media de las distancias de los elementos entre clusters. Proponemos el estadístico $T = \bar{D}^e - \bar{D}^d$ para evaluar la hipótesis de contigüidad espacial.

Sean $\bar{D}_b^d$ y $\bar{D}_b^e$ la media de las correspondientes distancias obtenidas en la permutación $b-$ésima. El valor observado del estadístico T es $T_{observado} = \bar{D}^e - \bar{D}^d = 688{,}77 - 497{,}92 = 190{,}85$. La Figura 67 muestra la distribución del estadístico T bajo la distribución de aleatorización, equivalente a la hipótesis nula de no contigüidad espacial. Observemos que la distribución del estadístico T cuando la hipótesis nula es verdadera se encuentra en el rango (-20, +30), mientras que el valor del estadístico obtenido para la segmentación original es 190,85. Concluimos entonces que la probabilidad de observar un valor de 190,85 o mayor es cero, es decir, $p - valor = 0$, lo que nos garantiza suficiente evidencia para rechazar H_0 y

aceptar entonces H_1. Esto significa que $E\left(D^d\right) < E\left(D^e\right)$ y por lo tanto, en los clusters obtenidos por el procedimiento propuesto, las distancias geográficas entre los departamentos dentro de cada cluster son significativamente menores que las distancias entre departamentos pertenecientes a grupos distintos. Concluimos entonces que existe contigüidad espacial entre departamentos en un mismo cluster.

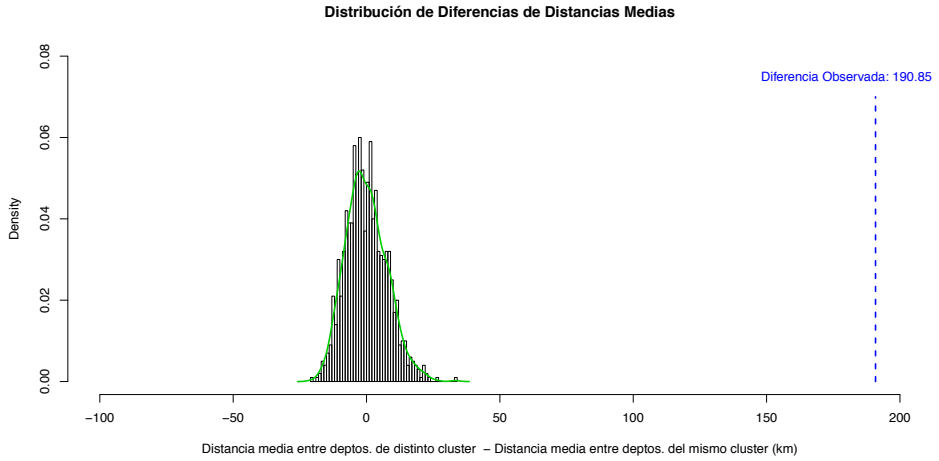

Figura 67: *Distribución de diferencia de medias de distancias entre departamentos pertenecientes a distintos clusters y medias de distancias entre departamentos del mismo cluster.*

Capítulo IV

CONCLUSIONES Y DISCUSION

En esta tesis se proponen métodos para la identificación de grupos de series de rindes de cultivo de maíz en Argentina. El problema presenta características especiales ya que los objetos a segmentar son observaciones longitudinales.

La propuesta incluye procedimientos de segmentación, algoritmos para cuantificar la significatividad de la segmentación y una propuesta para testear la contigüidad espacial de las observaciones intra-cluster. La propuesta fue ejemplificada usando datos de rinde de maíz pero es de aplicación general y puede usarse para datos longitudinales en general. Se identificaron situaciones en las que el criterio propuesto por Tibshirani para la elección del número de clusters no resulta apropiado y se propusieron criterios alternativos. Por otra parte, se demostró la relación intrínseca entre el estadístico de Ward y el estadístico gap, cuando ambos se basan en la noción de distancia euclídea. Un problema abierto es evaluar la performance de los criterios propuestos en diferentes escenarios más generales.

Es interesante destacar que aunque la información de la cual partimos concierne a datos agregados de superficies y rendimientos, proveniente de promedios calculados en base a datos provistos por los productores agropecuarios, con potenciales sesgos y/u omisiones, hemos llegado a resultados muy razonables y coherentes. La información es consistente, inclusive con toda la variabilidad que puede haber, para un año, una región y un cultivo. Quizás el rasgo mas importante para la verificación de la consistencia de los resultados, se halle en el hecho de obtener grupos de rendimientos con un alto grado de contigüidad espacial. Dicha contigüidad espacial es un emergente del análisis, ya que la misma no fue impuesta en la metodología de segmentación utilizada.

Finalmente, queda como tema abierto avanzar en la resolución del problema en que se enmarca esta tesis, que sería incorporar información climatológica para mejorar la predicción de rendimientos en una campaña/año en particular.

Capítulo V

BIBLIOGRAFIA

[1] Beale, E. M. L., *Euclidean clusters analysis*, Bulletin of the International Statistical Institute, p. 92–94, 1969.

[2] Boente Graciela, Yohai Víctor, *Notas de Estadística.* **www.dm.uba.ar/materias/estadistica_teorica_Mae/2010/2/Notas%20de%20Estadistica.pdf**

[3] Calinski T., Harabasz, J., *A dendrite method for cluster analysis*, Communications in Statistics, pages 1–27, 1974.

[4] Gower, J. C., *A general coefficient of similarity and some of its properties*, Biometrics, 27, p. 857-874, 1971.

[5] Hastie Trevor, Tibshirani Robert, Friedman Jerome, *The Elements of Statistical Learning,* Springer Series in Statistics, 2009, California, Estados Unidos.

[6] Hartigan J. A., *Consistency of single linkage for high density clusters*, J.Amer. Statist. Assoc., 76, p. 388-394, 1981.

[7] Inzenman Alan Julian, *Modern Multivariate Statistical Techniques - Regression, Classification and Manifold Learning,* Springer Science+Business Media, LLC, 2008, Nueva York, Estados Unidos.

[8] Maronna R.A., Martin R.D. y Yohai V.J., *Robust Statistics: Theory and Methods*, John Wiley & Sons, Ltd., 2006, Sussex Occidental, Inglaterra.

[9] Marriot, F. H. C., *Practical problems in a method of cluster analysis*, Biometrics, 27, p. 501-514, 1971.

[10] Milligan, G. W., Mahajan, V., *A note on procedures for testing the quality of a clustering of a set of objects*, Decision Sciences, 11, p. 669-677, 1980.

[11] Milligan G. W., Cooper M. C., A*n examination of procedures por determining the number of clusters in a data set*, Psychometrika 50 (2), p. 159-179 , 1985.

[12] Mojena, *Hierarchical grouping methods and stopping rules: An evaluation*, The Computer Journal, 20 (4), p. 359–363, 1977.

[13] Rennie B. C. y Dobson A. J., *On Stirling Numbers of the Second Kind*, Journal of Combinatorial Theory 7, p. 116-121, 1969, University College of Townsville, Queensland, Australia.

[14] Richard A. Johnson, Dean W. Wichern, *Applied Multivariate Statistical Analysis*, Pearson Education, Inc., 2007, Nueva Jersey, Estados Unidos.

[15] Rossi Daniel, *Evolución de los cultivares de maíz utilizados en la Argentina*, Revista Agromensajes, Nº 22, 2007, Facultad de Ciencias Agrarias, Universidad Nacional de Rosario.

[16] Seber G. A. F, *Multivariate Observations,* John Wiley & Sons, Inc., 2004, Nueva Jersey, Estados Unidos.

[17] Tibshirani Robert, Walther Guenther y Hastie Trevor, *Estimating the number of clusters in a data set via the gap statistic*, Journals of the Royal Statistical Society, Serie B, N° 63, 2001.