UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

DEPARTAMENTO ACADÉMICO DE ESTADÍTICA E INFORMÁTICA

TRABAJO MONOGRÁFICO

DESCRIPCIÓN METODOLÓGICA DEL ANÁLISIS CLÚSTER UTILIZANDO EL ALGORITMO DE WARD

Presentado para optar el título de Ingeniero Estadístico e Informático

ANDIE BRYAN DONGO ROMÁN

Modalidad de Examen Profesional

LIMA-PERÚ

2017

A mis padres, con todo cariño, por su invaluable apoyo y por motivarme a ser mejor cada día, como profesional y como persona.

Gracias a mi querida universidad, donde aprendí tanto y conocí a grandes personas que participaron en mi proceso de formación.

ÍNDICE

ÍNDICE DE CUADROS

ÍNDICE DE FIGURAS

ÍNDICE DE ANEXOS

RESUMEN

El presenta trabajo tiene como objetivo principal describir la metodología que se debe seguir al realizar un análisis clúster utilizando el algoritmo de Ward, mostrando una serie de pasos para su correcta aplicación. Además, exponer cuáles son las características y las ventajas de elegir este algoritmo como criterio de agrupamiento.

El algoritmo de Ward es uno de los diversos métodos jerárquicos del análisis clúster, el cual viene a ser uno de lo más usados por tener un fundamento estadístico (mientras que los demás suelen ser heurísticos), pues se basa en el criterio de la suma de cuadrados para medir la proximidad entre clústeres durante el proceso de agrupamiento.

Como ejemplo aplicativo se planteó el caso de las Comunidades Autónomas de España, las cuáles se agruparon en base a la actividad de sus salas de proyección de cine. En este caso, siguiendo los pasos correspondientes, el algoritmo de Ward determinó que estas Comunidades se agrupaban en cuatro clústeres, los cuales mostraron características que los diferenciaban entre sí en función de las variables de estudio.

I. INTRODUCCIÓN

Las personas siempre han estado expuestas ante una infinidad de objetos, sucesos o datos en general. Es tanta la información que existe en el entorno que ha sido necesario idear mecanismos para resumirla o simplificarla y así poder entender la realidad con mayor facilidad. Uno de estos mecanismos es el de clasificar a los elementos en categorías, pues es más sencillo interpretar la realidad a través de un número reducido de grupos que analizar cada elemento de manera aislada.

El análisis clúster, también conocido como análisis conglomerado, taxonómico o tipológico; es un método de clasificación que tiene como objetivo agrupar objetos de tal manera que los grupos sean lo más homogéneos posible internamente y sean diferentes los unos de los otros.

Hay que tener en cuenta la diferencia que existe entre los métodos de clasificación no supervisada (análisis clúster) y los métodos de clasificación supervisada (discriminación). En los métodos de clasificación supervisada, a partir de un conjunto de datos clasificados a priori (conjunto de entrenamiento), se intenta asignar una clasificación a un segundo conjunto de datos. Por otro lado, en los métodos de clasificación no supervisada, no se dispone de un conjunto de datos previamente clasificados, sino que únicamente a partir de las propiedades de estos datos se les agrupará según su similaridad.

Existen diversas formas de clasificar a los métodos del análisis clúster, ya sea por la forma en que agrupa a los datos, por la cantidad de grupos a los que se puedan asignar los datos o por la cobertura en el agrupamiento de los datos.

Gallardo (s.f.) y Pedret et al. (2013) mencionan dos grandes categorías de métodos clúster: Los métodos jerárquicos, que se caracterizan porque las asignaciones de los individuos a los clústeres que se van creando permanecerán estables durante todo el proceso (destacan los criterios de Ward, enlace simple, enlace completo, entre otros); y los métodos no jerárquicos, donde el investigador fija el número de clústeres a conseguir y permite la reasignación de individuos a clústeres distintos si hubiera lugar a ello (comprende el método de K-medias, Forgy, el análisis factorial tipo Q, entre otros).

Los métodos jerárquicos del análisis clúster son de gran utilidad cuando se requiere agrupar objetos, pero no se sabe cuál es el número de clases o grupos que se deben formar; pues este tipo de técnicas muestra diferentes escenarios de agrupación con el mismo set de datos y permite determinar el número óptimo de clústeres a considerar. El algoritmo de Ward es una de las tantas opciones que ofrece el análisis clúster jerárquico; la ventaja de este algoritmo radica en que tiene fundamentos estadísticos, pues para agrupar los individuos utiliza la medida de la suma de cuadrados, buscando minimizar la dispersión dentro de los grupos formados.

El objetivo de esta monografía es explicar y describir la metodología del análisis clúster utilizando el algoritmo de Ward para poder clasificar un conjunto de datos identificando grupos homogéneos.

El cuerpo de la monografía consta de cuatro capítulos. En el capítulo II se abordará las definiciones y conceptos principales para entender el análisis clúster. El capítulo III describirá el algoritmo de Ward para entender sus fundamentos. El capítulo IV explicará la metodología para realizar el análisis clúster jerárquico con el algoritmo de Ward. En el capítulo V, como ejemplo ilustrativo, se aplicará el algoritmo de Ward siguiendo los pasos mencionados en el capítulo anterior.

Con esto, se busca que éste trabajo muestre cómo desarrollar un análisis clúster utilizando el algoritmo de Ward a través de cinco pasos: (1) el análisis preliminar de las variables, (2) la elección de la medida de asociación, (3) aplicación del algoritmo de Ward, (4) la selección de los clústeres y (4) la interpretación de los resultados.

Se encontró que, siguiendo los pasos mencionados sobre el caso de las Comunidades Autónomas de España, estas podrían agruparse en cuatro clústeres, los cuales poseían características diferentes entre sí respecto a la actividad del cine en cada Comunidad.

II. EL ANÁLISIS CLÚSTER

El análisis clúster corresponde una amplia variedad de procedimientos que pueden ser usados para crear una clasificación. Un método clúster es un procedimiento estadístico multivariante que comienza con un conjunto de datos conteniendo información sobre una muestra de entidades e intenta reorganizarlas en grupos relativamente homogéneos a los que se les conoce como clústeres (Gallardo s.f.).

En el análisis clúster, poca o ninguna información es conocida sobre la estructura de las categorías, lo cual lo diferencia de los métodos multivariantes de asignación y discriminación, donde esta información se conoce a priori. De todo lo que se dispone es de una colección de observaciones, siendo el objetivo operacional en este caso, descubrir la estructura de las categorías en la que se encajan las observaciones.

Como señala Oliva (2015) en su tesis; es importante distinguir entre técnicas de clasificación no supervisada (agrupamiento o segmentación) y métodos de clasificación supervisada (discriminación). En el caso de segmentación, se dispone de una cierta cantidad de objetos 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 discriminación, se dispone de observaciones clasificadas en grupos (a priori) y el objetivo es definir un criterio que permita clasificar una nueva observación y asignarla a uno de los grupos existentes.

Aunque poco o nada se conoce sobre la estructura de las categorías de manera a priori, se tiene con frecuencia algunas nociones sobre características deseables e inaceptables a la hora de establecer un determinado esquema de clasificación (Gallardo s.f.).

El objetivo es ordenar las observaciones en grupos tales que el grado de asociación natural sea alto entre los miembros del mismo grupo y bajo entre miembros de grupos diferentes.

Kumar et al. (2006:490) señala que:

> "El análisis clúster agrupa los objetos basándose únicamente en la información que se encuentra en los datos que describen a los objetos y sus relaciones. El

objetivo es que los objetos dentro de un grupo sean similares (o estén relacionados) entre sí y diferentes de (o no estén relacionados con) los objetos de otros grupos. Cuanto mayor sea la similitud (u homogeneidad) dentro de un grupo y mayor sea la diferencia entre los grupos, mejor será la clasificación."

Existen diversas formas de clasificar a las técnicas clúster. Kumar et al. (2006) propone tres tipos de clasificación de estas técnicas:

- **Jerárquicos y No jerárquicos (particionales):** Se diferencian en el proceso de agrupamiento de datos, considerando si los elementos se van anidando o no.
- **Exclusivos, Superpuestos y Difusos**: Son exclusivos cuando los objetos son asignados a un solo grupo. En el caso de los superpuestos, los objetos pueden ser asignados en más de un grupo. En los difusos, todos los objetos pueden pertenecer a todos los grupos, asignándoles una probabilidad o peso para la membresía a cada grupo.
- **Completo y Parcial:** Son completos cuando se logra asignar a todos los objetos a algún grupo, mientas que en los parciales algunos objetos no serán asignados a algún grupo específico.

Los usos del análisis clúster pueden ser resumidos en cuatro objetivos principales:

- Desarrollar una tipología o clasificación (el más usado).
- Investigar esquemas conceptuales útiles para agrupar entidades.
- Generar hipótesis a través de la exploración de los datos.
- Contrastar hipótesis o intentar determinar si tipos definidos por otros procedimientos están de hecho presentes en un conjunto de datos.

Una vez considerado que el objetivo del análisis clúster que se requiere, es necesario definir qué se entiende por agrupaciones naturales y, por lo tanto, con arreglo a qué criterio se puede decir que dos grupos son más o menos similares. Esta cuestión conlleva otras dos:

- ¿Cómo se puede medir la cercanía entre dos individuos de la muestra? Medidas de asociación o índices de proximidad.
- ¿Cómo se puede evaluar cuándo dos clúster pueden ser o no agrupados? Criterios o métodos de agrupación.

2.1 Medidas de asociación o proximidad

Para poder unir variables o individuos es necesario tener algunas medidas numéricas que caractericen las relaciones entre las variables o los individuos.

Cada medida refleja asociación en un sentido particular y se debe elegir una medida apropiada para el problema concreto que se esté tratando.

De La Fuente (2011) señala que la medida de asociación puede ser una distancia o una similaridad:

- Cuando se elige una distancia como medida de asociación (por ejemplo, la distancia euclídea) los grupos formados contendrán individuos parecidos de forma que la distancia entre ellos tiene que ser pequeña.
- Cuando se elige una medida de similaridad (por ejemplo, el coeficiente de correlación) los grupos formados contendrán individuos con una similaridad alta entre ellos. La correlación de Pearson y los coeficientes de Spearman y de Kendall son índices de similitud.

Medidas de asociación o proximidad para datos binarios (dicotómicos):

Cuadro 1: Relación de dos variables binarias

X_i / X_j	1	0	Totales
1	a	b	a+b
0	c	d	c+d
Totales	a+c	b+d	m=a+b+c+d

Fuente: Tomado de De La Fuente, 2011.

- Jaccard:

$$\frac{a}{a+b+c}$$

- Russell y Rao:

$$\frac{a}{a+b+c+d} = \frac{a}{m}$$

- Rogers y Tanimoto:

$$\frac{a+d}{a+d+2(b+c)}$$

- Medida de parejas simples:

$$\frac{a+d}{a++b+c+d}=\frac{a+d}{m}$$

Medidas de asociación o proximidad para variables continuas:

- Distancia de euclídea:

$$d(x_i,x_j)=\sqrt{\sum_{c=1}^{p}(x_{ic}-x_{jc})^2}$$

- Distancia de Minkowsky:

$$d_q(x_i,x_j)=\left(\sum_{c=1}^{p}|x_{ic}-x_{jc}|^q\right)^{1/q} \quad donde\ q\geq 1$$

- Distancia métrica de Chebyshev:

$(q=\infty)$: $d_\infty e(x_i,x_j)=máx(c=1,\dots,p)|x_{ic}-x_{jc}|$

- Bloque, Manhattan o City-Block

$$d(x_i,x_j)=\sum_{c=1}^{p}|x_{ic}-x_{jc}|$$

Para medir similaridad entre variables, pueden usarse el coseno de vectores, la correlación de Pearson, entre otras medidas.

Pedret et al. (2013) señala que el índice de proximidad más utilizado es el de la distancia, y dentro de ellas, la distancia euclidiana (aunque es válido aplicar cualquier otra).

2.2 Métodos de agrupamiento

Pedret et al. (2003) y Gallardo (s.f.) distinguen, a grandes rasgos, dos grandes categorías de métodos de agrupamiento para el análisis clúster: métodos jerárquicos y métodos no jerárquicos, las cuales fueron tomadas como referencia en este trabajo.

2.2.1 Métodos Jerárquicos

Estos métodos tienen por objetivo agrupar clústeres para formar uno nuevo o bien separar alguno ya existente para dar origen a otros dos, de tal forma que se minimice alguna función de distancia o bien se maximice alguna medida de similitud.

Los métodos jerárquicos se subdividen a su vez en aglomerativos y disociativos. Los aglomerativos comienzan el análisis con tantos grupos como individuos haya en el estudio. A partir de ahí se van formando grupos de forma ascendente, hasta que, al final del proceso, todos los casos están englobados en un mismo conglomerado. Los métodos disociativos o divisivos realizan el proceso inverso al anterior. Empiezan con un conglomerado que engloba a todos los individuos. A partir de este grupo inicial se van formando, a través de sucesivas divisiones, grupos cada vez más pequeños. Al final del proceso se tienen tantos grupos como individuos en la muestra estudiada (Gallardo s.f.).

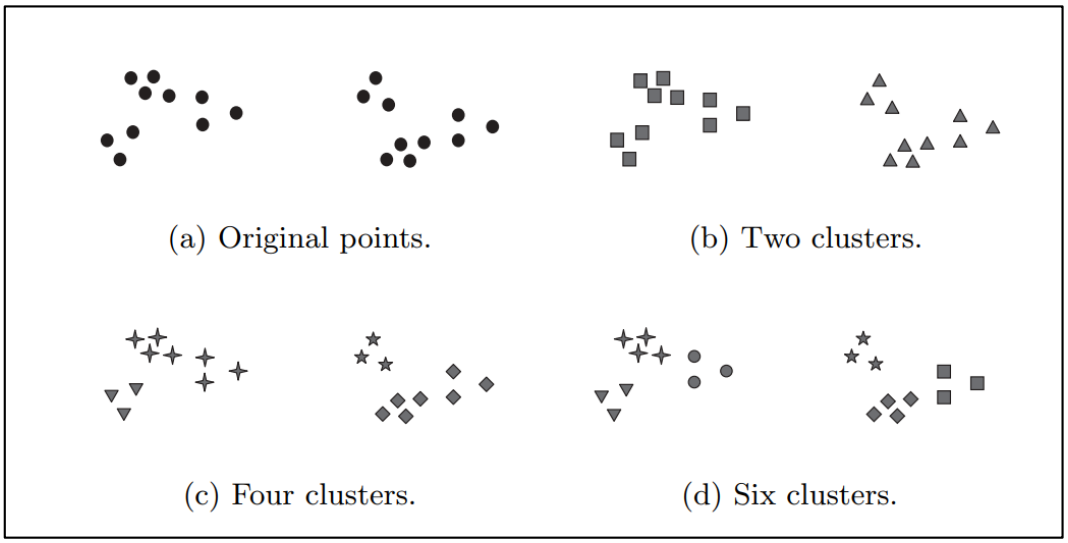

Figura 1: Diferentes formas de agrupar el mismo conjunto de puntos (elementos).
Fuente: Tomado de Kumar et al., 2006.

Independientemente del proceso de agrupamiento, hay diversos criterios para ir formando los conglomerados; todos estos criterios se basan en una matriz de distancias o similitudes. Por ejemplo, dentro de los métodos aglomerativos destacan (Gallardo s.f.):

- Algoritmo del enlace simple (vecino más cercano).
- Algoritmo del enlace completo (vecino más lejano).
- Algoritmo del promedio entre grupos.

- Algoritmo del centroide.
- Algoritmo de la mediana.
- Algoritmo de Ward.

Dentro de los métodos disociativos, destacan, además de los anteriores, que siguen siendo válidos:

- El análisis de asociación.
- El detector automático de interacción.

Para Pedret et al. (2003), estos métodos son apropiados principalmente para clasificar productos, marcas, empresas, etc. (número reducido de datos) en función de sus similitudes sobre un conjunto de atributos o características.

Sin embargo, en una etapa exploratoria de los datos, cuando se desconoce o no se tiene idea del número de grupos que se puedan formar, es recomendable aplicar un método jerárquico para encontrar el número de clústeres apropiados a partir de estos datos. (Lebart, citado por Del Campo 2007).

2.2.2 Métodos No Jerárquicos

En cuanto a los métodos no jerárquicos, también conocidos como partitivos o de optimización, tienen por objetivo realizar una sola partición de los individuos en K grupos. Ello implica que el investigador debe especificar a priori los grupos que deben ser formados, siendo ésta, posiblemente, la principal diferencia respecto de los métodos jerárquicos, (no obstante hay que señalar que hay diversas versiones de estos procedimientos que flexibilizan un tanto el número final de clúster a obtener). La asignación de individuos a los grupos se hace mediante algún proceso que optimice el criterio de selección. Otra diferencia de estos métodos respecto a los jerárquicos reside en que trabajan con la matriz de datos original y no precisan su conversión en una matriz de distancias o similitudes (Gallardo s. f.).

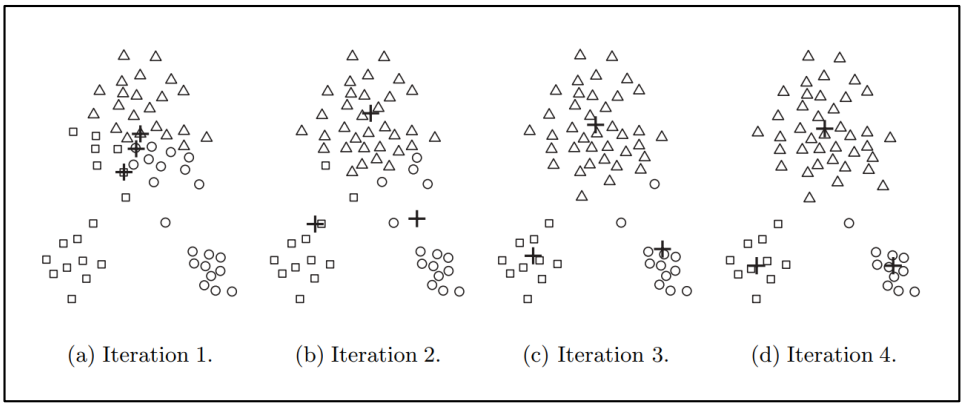

Figura 2: Diferentes formas de encontrar tres grupos con el mismo conjunto de datos. Fuente: Tomado de Kumar et al., 2006.

Pedret et al. (2003) agrupa los métodos no jerárquicos en cuatro familias:

Métodos de Reasignación: Permiten que un individuo asignado a un grupo en un determinado paso del proceso sea reasignado a otro grupo en un paso posterior, si ello optimiza el criterio de selección. El proceso acaba cuando no quedan individuos cuya reasignación permita optimizar el resultado que se ha conseguido. Dentro de estos métodos están:

- El método K-Medias.
- El Quick-Clúster análisis.
- El método de Forgy.
- El método de las nubes dinámicas.

Métodos de búsqueda de la densidad: Dentro de estos métodos están los que proporcionan una aproximación tipológica y una aproximación probabilística.

En el primer tipo, los grupos se forman buscando las zonas en las cuales se da una mayor concentración de individuos. Entre ellos destacan:

- El análisis modal de Wishart.
- El método Taxmap.
- El método de Fortin.

En el segundo tipo se parte del postulado de que las variables siguen una ley de probabilidad según la cual los parámetros varían de un grupo a otro. Se trata de encontrar los individuos que pertenecen a la misma distribución. Entre los métodos de este tipo destaca el método de las combinaciones de Wolf.

Métodos directos: Busca clasificar en simultáneo a individuos y variables.

- Block Clustering.

Métodos de reducción de dimensiones: Busca factores en el espacio de los individuos, cada factor corresponde a un grupo.

- Análisis Factorial tipo Q.

Pedret et al. (2003) señala que estos métodos son recomendados para la agrupación de grandes conjuntos de datos como pueden ser clasificaciones de individuos, consumidores, compradores, etc. En función de características de comportamiento, actitudes, sociodemográficas, entre otras.

No obstante, en una etapa exploratoria, una vez que se tenga una idea del número de grupos que pueda haber en los datos (puede ser con ayuda de un método jerárquico), se puede aplicar un método no jerárquico para encontrar la clasificación final y confirmar la pertinencia de los grupos encontrados. (Lebart, citado por Del Campo 2007).

2.3 Precauciones en el uso del análisis clúster

Gallardo (s. f.) menciona algunas precauciones que hay tener sobre los métodos clúster:

- Algunos de los métodos de análisis clúster son procedimientos que, en la mayor parte de los casos, no están soportados por un cuerpo de doctrina estadística teórica. En otras palabras, la mayor parte de los métodos son heurísticos.
- Distintos procedimientos clúster pueden generar soluciones diferentes sobre el mismo conjunto de datos. Una razón para ello radica en el hecho de que los métodos clúster se han desarrollado a partir de diversas disciplinas que han dado origen a reglas diferentes de formación de grupos. De esta manera, lógicamente, es necesaria la existencia de técnicas que puedan ser usadas para determinar qué método produce los grupos naturalmente más homogéneos en los datos.

III. EL ALGORITMO DE WARD

El método de agrupamiento jerárquico de la suma de cuadrados de errores de Ward ha sido ampliamente utilizado desde su primera descripción por Ward en una publicación de 1963.

Ward es el único entre los algoritmos de agrupamiento aglomerativo que se basa en un criterio clásico de suma de cuadrados, produciendo grupos que minimizan la dispersión dentro del grupo en cada fusión binaria. Además, el método de Ward es interesante porque busca clústeres en el espacio euclidiano multivariado. Éste es también el espacio de referencia en los métodos de ordenación multivariante, y en particular en el análisis de componentes principales (Legendre y Murtagh 2014).

Oliva (2015:20) señala lo siguiente:

> ‘‘Este método, también conocido como ‘incremento en la suma de los cuadrados’ o método de mínima varianza, (…) tiene como objetivo unificar grupos de forma tal que la variabilidad dentro de los grupos no aumente dramáticamente. En cada paso se fusionan los dos clústeres que producen la suma de cuadrados dentro de clúster (variabilidad *within* o intra-clústeres) mínima entre todas las posibles particiones que se obtienen fusionando dos clústeres del paso previo. En este contexto ‘suma de cuadrados dentro’ refiere a la suma de las distancias al cuadrado de las observaciones del clúster respecto de la media de las observaciones del mismo clúster. ’’

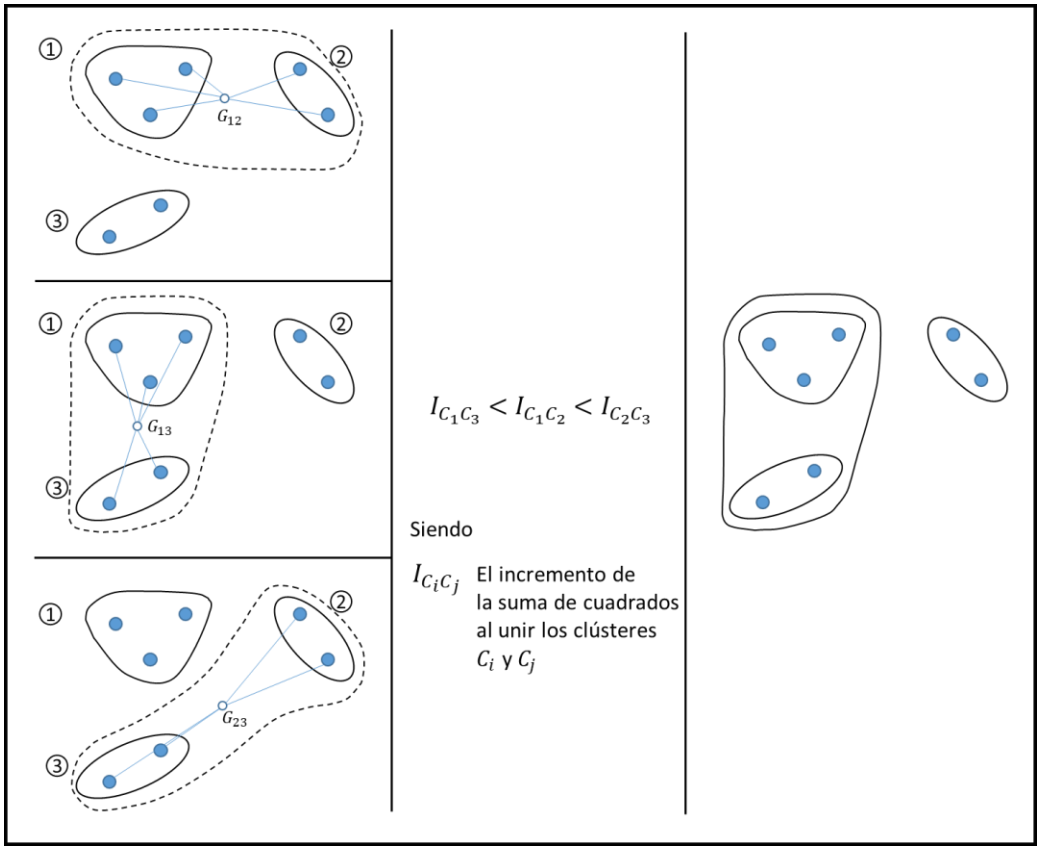

Figura 3: Representación gráfica del método de Ward.
Fuente: Adaptado de Predret 2003.

3.1 El algoritmo de Ward para agrupamiento

En este algoritmo, la distancia entre dos clústeres se mide como el incremento en la suma de cuadrados dentro del nuevo clúster que sea forma al unirlos. Sean C_i y C_j dos clústeres con n_i y n_j elementos y medias $\boldsymbol{m}_i$ y $\boldsymbol{m}_j$. Sea $\boldsymbol{m}$ la media del clúster que se formaría al unir C_i y C_j. En cada paso, la distancia o función objetivo a minimizar es $I_{C_iC_J}$, que representa el incremento en la suma de cuadrados dentro del clúster cuando se unen los dos clústeres.

$$I_{C_iC_J} = \sum_{l\in C_i\cup C_j} ||x_l - \boldsymbol{m}||^2 - \left\{\sum_{l\in C_i} ||x_l - \boldsymbol{m_i}||^2 + \sum_{l\in C_j} ||x_l - \boldsymbol{m_j}||^2\right\}$$
$$= \frac{n_i n_j}{n_i + n_j} ||\boldsymbol{m}_i - \boldsymbol{m}_j||^2$$

Oliva (2015) en sus estudios demuestra la segunda igualdad:

$$\sum_{l\in C_i\cup C_j} ||x_l - \boldsymbol{m}||^2 - \left\{\sum_{l\in C_i} ||x_l - \boldsymbol{m_i}||^2 + \sum_{l\in C_j} ||x_l - \boldsymbol{m_j}||^2\right\} = \frac{n_i n_j}{n_i + n_j} ||\boldsymbol{m}_i - \boldsymbol{m}_j||^2$$

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

Entonces, sean $x_{lp}^{(i)}$ la componente p−ésima del vector l−ésimo del clúster C_i y $x_{lp}^{(i,j)}$ la componente p−ésima del vector l−ésimo del clúster $C_i \cup C_i$, entonces:

$$\sum_{l\in C_i\cup C_j} \left|\left|x_l - \boldsymbol{m}\right|\right|^2 - \left\{\sum_{l\in C_i} ||x_l - \boldsymbol{m_i}||^2 + \sum_{l\in C_j} ||x_l - \boldsymbol{m_j}||^2\right\} =$$

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

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

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

…(1)

Se observa que $(n_i + n_j)m_s = n_i m_{is} + n_j m_{js}$.

Luego, $(n_i + n_j)^2 m_s^2 = (n_i m_{is})^2 + (n_j m_{js})^2 + 2n_i n_j m_{is} m_{js}$.

Además, $2m_{is}m_{js} = m_{is}^2 + m_{js}^2 - (m_{is} - m_{js})^2$.

Por lo tanto:

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

Dividiendo por $(n_i + n_j)$ se obtiene:

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

Reemplazando en (1), resulta:

$$I_{C_i C_J} = \frac{n_i n_j}{n_i + n_j} \sum_{s=1}^{p} (m_{is} + m_{js})^2 = \frac{n_i n_j}{n_i + n_j} ||\boldsymbol{m}_i - \boldsymbol{m}_j||^2$$

Es decir, la medida de proximidad es la distancia cuadrada entre los centroides pesada por un factor. El lema muestra adicionalmente que la fusión de dos clústeres nunca puede reducir la suma de distancias intra-clústers. También puede apreciarse que el método tenderá a unir clústeres con centroides cercanos y con pocos elementos. Además, éste método tiende así a producir clústeres esféricos y bien balanceados.

A continuación, un ejemplo de cómo agrupa el algoritmo de Ward, en el cual a cinco individuos se les han medido dos variables:

Cuadro 2: Datos de ejemplo para el algoritmo de Ward

Individuo	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$
A	10	5
B	20	20
C	30	10
D	30	15
E	5	10

Fuente: Tomado de Gallardo, s.f.

- Nivel 1: Se calculan todas las posibles combinaciones de grupos, y los incrementos asociados.

Cuadro 3: Agrupamiento en el nivel 1

Agrupamientos	Centroides	Incrementos
(A,B)	(15,12.5)	162.5
(A,C)	(20,7.5)	212.5
(A,D)	(20,10)	250
(A,E)	(7.5,7.5)	25
(B,C)	(25,15)	100
(B,D)	(25,17.5)	62.5
(B,E)	(12.5,15)	162.5
(C,D)	(30,12.5)	12.5
(C,E)	(17.5,10)	312.5
(D,E)	(17.5,12.5)	325

Fuente: Elaboración propia.

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

- Nivel 2: Los posibles agrupamientos son:

Cuadro 4: Agrupamiento en el nivel 2

Agrupamientos	Centroides	Incrementos
((C,D),A)	(23.33,10)	304.16
((C,D),B)	(26.66,15)	104.16
((C,D),E)	(21.66,11.66)	420.83
(A,B)	(15,12.5)	162.5
(A,E)	(7.5,7.5)	25
(B,E)	(12.5,15)	162.5

Fuente: Elaboración propia.

En este nivel se unirán los clúster A y E dado que su incremento es el mínimo (62.5). La configuración en este nivel sería (C,D),(A,E),B.

- Nivel 3: Los posibles agrupamientos son:

Cuadro 5: Agrupamiento en el nivel 3

Agrupamientos	Centroides	Incrementos
((C,D),(A,E))	(18.75,10)	531.25
((C,D),B)	(26.66,15)	104.16
((A,E),B)	(11.66,11.66)	208.3

Fuente: Elaboración propia.

Luego, se fusionarán los clústeres B y (C,D) con un incremento de 104.16. La configuración actual sería (B,C,D),(A,E).

- Nivel 4: Finalmente, se agruparán los clústeres que quedaron formando uno solo (A,B,C,D,E), con incremento de distancias igual a 508.34.

3.2 Relación del algoritmo de Ward con otras técnicas

El Análisis de Componente Principales (ACP) es otro camino para representar la varianza entre las observaciones, esta vez en un diagrama de ordenación, que puede ser visto como una representación "espacial" de las relaciones entre las observaciones. ACP es una descomposición de la varianza total de la tabla de datos, seguida de la selección de los ejes que representan la mayor parte de la varianza; Estos ejes se utilizan para la representación de las observaciones en unas pocas dimensiones, usualmente dos (dependiendo del criterio del investigador). A partir de este razonamiento, se puede ver que los métodos espaciales (por ejemplo, ACP) y de agrupamiento (por ejemplo, Ward) implican modelos espaciales y de agrupamientos diferentes pero complementarios que se ajustan a los datos usando el mismo principio matemático. Esta es la razón por la cual en la práctica los resultados del agrupamiento aglomerado de Ward son propensos a delinear grupos que visualmente corresponden a regiones de altas densidades de puntos en la ordenación ACP.

El método de Ward comparte el criterio de suma total de cuadrados con el método K-medias, que es ampliamente utilizado para agrupar directamente las observaciones en el espacio euclidiano, por lo tanto para crear una partición del conjunto de observación. Este agrupamiento se realiza sin ninguna restricción estructural, como la integración de clústeres representada por una jerarquía. Debido a que la partición K-medias es un problema de cálculo complejo, a menudo se busca una solución aproximada usando múltiples inicios aleatorios del algoritmo y escogiendo la solución que minimiza el criterio de suma total de cuadrados. Un enfoque más directo y eficiente en la computadora es aplicar primero el agrupamiento aglomerado de mínima varianza de Ward a los datos, identificar la partición de los objetos en K grupos en el dendrograma y luego utilizar esa partición como la aproximación inicial para el particionamiento del K-medias, ya que está cerca de la solución que uno está buscando. Esta solución puede entonces ser mejorada por iteraciones del algoritmo de K-medias (Legendre y Murtagh 2014).

IV. METODOLOGÍA PARA EL ANÁLISIS CLÚSTER USANDO EL ALGORITMO DE WARD

4.1 Paso 1: Análisis preliminar de las variables

La elección inicial del conjunto concreto de características usadas para describir a cada individuo constituye un marco de referencia para establecer las agrupaciones o clústeres; dicha elección, posiblemente, refleje la opinión del investigador acerca de su propósito de clasificación. Consecuentemente, la primera cuestión a responder sobre la elección de variables es si son relevantes para el tipo de clasificación que se va buscando. Es importante tener en cuenta que la elección inicial de variables es, en sí misma, una categorización de los datos, para lo cual sólo hay limitadas directrices matemáticas y estadísticas (Gallardo s.f.).

Una vez seleccionadas las variables a considerar en el análisis, se debe verificar que éstas no estén afectadas por la presencia de casos atípicos.

De La Fuente (2011) plantea dos soluciones para abordar este problema, si se presenta:

- Cambiar datos iniciales por datos promedios. Por ejemplo, en lugar de número de salas de cine en una ciudad, se podría usar el número de salas de cine por cada mil habitantes de una ciudad.
- Realizar transformaciones de la distribución de datos utilizando la escalera de Tukey.
    - La *asimetría positiva* se corrige con raíces cuadradas y logaritmos naturales cuando tienen valores bajos, y con funciones inversas o inversos cuadráticos cuando los valores son elevados.
    - La *asimetría negativa* se corrige mediante elevaciones cúbicas y cuadráticas cuando es suave, y con antilogaritmos cuando es muy elevada.

Debido a que el análisis clúster estudia las características estructurales de un conjunto de observaciones con el fin de agruparlas en conjuntos homogéneos, y al no ser propiamente

una técnica de inferencia estadística, las exigencias de normalidad, linealidad, entre otros supuestos; no son fundamentales como sí lo son en procedimientos de inferencia.

Sin embargo, una correcta aplicación del análisis clúster requiere que los datos cumplan con las siguientes condiciones básicas:

- **Ausencia de correlación entre las variables:**

  La existencia de correlación entre las variables implica que unas variables son combinaciones lineales de otras, que comparten información con otras variables; lo que implicaría que esta información compartida tenga una mayor importancia (ponderación).

  Además, cuando las variables están correlacionadas, se corre el peligro de incluir información redundante en el modelo, incumpliendo con el principio de parsimonia.

- **Número de variables no muy elevado.**

  En muchas aplicaciones es probable que el investigador se equivoque tomando demasiadas medidas, lo cual puede dar origen a diversos problemas, bien sea a nivel computacional o bien porque dichas variables oscurezcan la estructura de los grupos.

- **Las variables no deben estar medidas en escalas diferentes.**

  El requisito de que las variables no estén medidas en unidades diferentes se soluciona mediante la estandarización (o tipificación) de todas las unidades a tratar.

  Sin embargo, existe cierta controversia sobre si la tipificación debe ser un procedimiento a utilizar en todo análisis clúster: Everitt y Edelborck, citados por De La Fuente (2011), son algunos de los autores que no defienden el proceso de estandarización, y plantean tres posibles soluciones para el problema de tener variables con distinta unidad: (1) Recategorizar todas las variables en binarias, y aplicar a éstas una distancia apropiada para este tipo de medidas. (2) Realizar distintos análisis clúster con grupos de variables homogéneas (en cuanto a su medida), y sintetizar después los diferentes resultados. (3) Utilizar la distancia de Gower, que es aplicable con cualquier tipo de métrica.

  Pese a la falta de consenso y las diversas alternativas que surgen a partir de este problema, la mayoría de expertos aconsejan realizar el análisis con variables estandarizadas (De La Fuente 2011).

4.2 Paso 2: Elección de la medida de asociación o proximidad

La mayor parte de los métodos clúster requieren establecer una medida de asociación que permita medir la proximidad de los objetos en estudio. Cuando se realiza un análisis clúster de individuos, la proximidad suele venir expresada en términos de distancias, mientras que el análisis clúster por variables involucra generalmente medidas del tipo coeficiente de correlación, algunas de las cuales tienen interpretaciones en distintos sentidos mientras que otras son difíciles de describir, dado el carácter subjetivo de las mismas (De La Fuente 2011).

La medida de asociación o proximidad permite obtener la matriz de distancia, proximidad o similaridad, la cual muestra la distancia entre los individuos de estudio. En una primera instancia se juntarán los dos objetos más cercanos, formando un clúster. En los pasos posteriores, el agrupamiento de estos conjuntos dependerá del criterio o técnica de agrupamiento escogido (Pedret 2013).

Cuadro 6: Esquema de la matriz de proximidad.

Caso	1	2	3	4	5	…	n
1	0	d(2,1)	d(3,1)	d(1,4)	d(1,5)	…	d(n,1)
2	d(2,1)	0	d(2,3)	d(2,4)	d(2,5)	…	d(n,2)
3	d(3,1)	d(3,2)	0	d(3,4)	d(3,5)	…	d(n,3)
4	d(4,1)	d(4,2)	d(4,3)	0	d(4,5)	…	d(n,4)
5	d(5,1)	d(5,2)	d(5,3)	d(5,4)	0	…	d(n,5)
…	…	…	…	…	…	…	…
n	d(n,1)	d(n,2)	d(n,3)	d(n,4)	d(n,5)	…	0

Fuente: Elaboración propia.

La matriz mostrará $\frac{n(n-1)}{2}$ medidas de proximidad entre los (n) casos o individuos tomados de dos en dos. Los primeros casos que se agruparán serán aquellos que presentan una menor distancia.

En el caso del método de Ward, éste considera como índice de proximidad la distancia euclidiana al cuadrado (Oliva 2015).

4.3 Paso 3: Elección del criterio de agrupamiento (Algoritmo de Ward)

Luego de juntar los dos primeros elementos según la matriz de proximidad, existen una gran cantidad de criterios que permiten agrupar dos objetos cuya distancia entre ellos sea mínima.

Según cuál sea el criterio que se aplique (ver capítulo 2.2.1), los objetos que conformen cada grupo y su jerarquía serán distintos.

Cuadro 7: Esquema de la matriz de pertenencia.

Col. A	Col B.	Col C.	Col. D	Col. E	Col F.	Col. G
	Clúster combinado			Primera aparición del clúster de etapa		
Etapa	Clúster 1	Clúster 2	Coeficientes	Clúster 1	Clúster 2	Etapa siguiente
1						
2						
3						
4						
5						
…						
n-1						

Fuente: Elaboración propia.

La matriz de pertenencia muestra los agrupamientos en cada etapa del análisis y puede presentarse de diferentes formas. La Tabla 7 muestra la matriz de pertenencia que reporta el programa SPSS. La columna A indica cada una de las (n-1) etapas de agrupamiento, siendo (n) la cantidad de observaciones. Las columnas B y C indican cuáles son los clústeres que se agrupan en cada etapa; en caso que uno de los clústeres que se agrupen esté formado por dos o más individuos, el programa solo nombrará al menor de ellos. La columna D indica la distancia, de manera acumulada, a la que se agrupan los clústeres en determinada etapa (en caso del método de Ward, indicará el incremento de varianza acumulado). La columna E indica la etapa en la que el clúster de la columna B fue agrupado, si fuera el caso. La columna F indica lo mismo que la columna E pero para el clúster de la columna C. La Columna G indica cuál será la próxima etapa en la que el clúster formado vuelva a ser utilizado.

Gallardo (s.f.) señala que es conveniente, a la hora de las aplicaciones prácticas, no elegir un sólo procedimiento, sino abarcar un amplio abanico de posibilidades y contrastar los resultados obtenidos con cada una de ellas. De este modo, si los resultados finales son parecidos, se podrán obtener unas conclusiones mucho más válidas sobre la estructura natural de los datos. En caso contrario, de existir grandes diferencias en los resultados obtenidos, esto daría un indicio de que tal vez los datos con los que se está trabajando no obedecen a una estructura bien definida.

En este caso, el método o criterio a utilizar será el Ward, el cual minimiza el incremento de la varianza intra-grupo al juntar dos grupos. Este método es uno de los más utilizados en la

práctica, pues su procedimiento de cálculo se adapta mejor a los objetivos planteados por los investigadores (Pedret 2013).

De La Fuente (2011) menciona que una investigación llevada a cabo por Kuiper y Fisher probó que el método de Ward era capaz de acertar mejor con la clasificación óptima que otros métodos, como el enlace simple, enlace completo, media y centroide.

4.4 Paso 4: Selección de los clústeres

Dado el criterio seleccionado, la matriz de pertenencia indicará, en cada etapa, cómo se van agrupando todos los elementos o individuos del estudio.

Para cualquiera de los métodos, los resultados de los sucesivos agrupamientos se verán plasmados gráficamente en el árbol de clasificación o dendograma. Los distintos niveles de similitud, es decir los índices de fusión o separación que aparezcan en el dendograma, permiten visualizar la matriz de pertenencia. Esta indexación del árbol jerárquico (dendograma) permite visualizar rápidamente la pertenencia de un objeto a un grupo, para cada nivel de agregación (Pedret et al. 2013).

De La Fuente (2011) propone unas pautas para determinar el número de clústeres que se debe considerar a partir de un dendograma:

Para identificar cuáles son los clúster formados en un dendograma, habría que cortar el árbol con una línea horizontal. De esta forma, se determina el número de clústeres en que se divide el conjunto de objetos (ver figura 4).

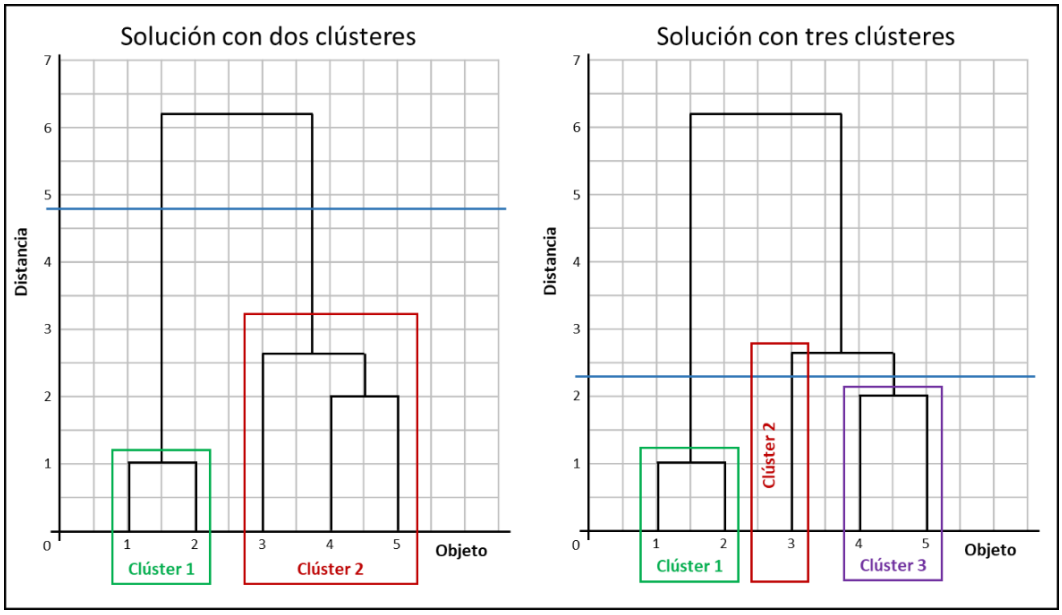

Figura 4: Ejemplos de diferentes soluciones con un dendograma.
Fuente: Adaptado de De La Fuente 2011.

La decisión del número óptimo de clústeres es subjetiva, especialmente cuando se incrementa el número de objetos pues si se seleccionan pocos, los clústeres resultantes son heterogéneos y artificiales, mientras que si se seleccionan demasiados, la interpretación de los mismos suele resultar complicada.

Para tomar una decisión sobre el número de clústeres se suelen representar los distintos pasos del algoritmo y la distancia a la que se produce la fusión (ver figura 5).

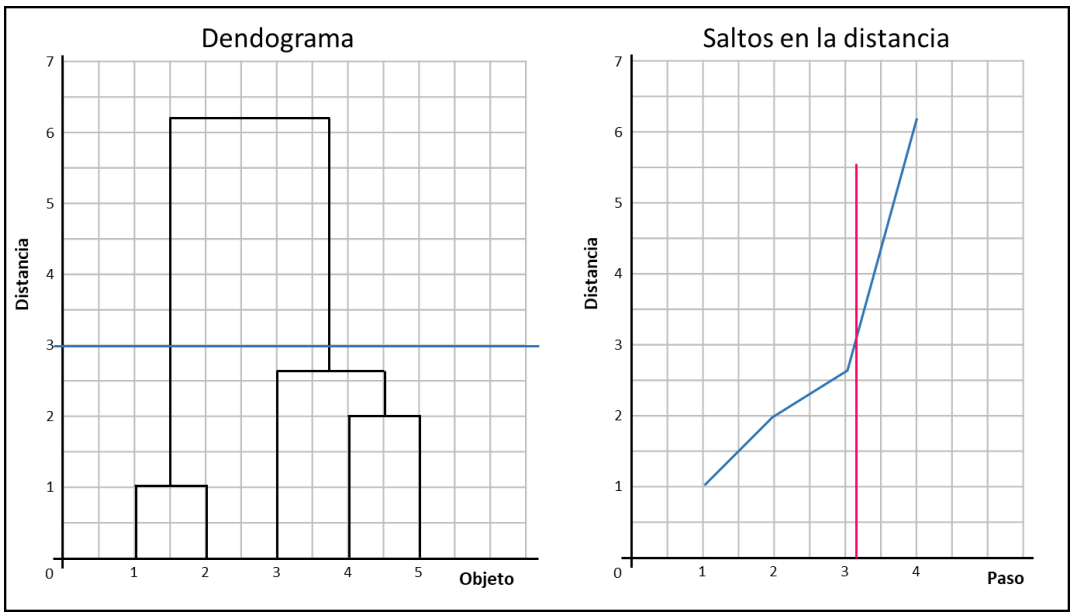

Figura 5: Selección del número de clústeres.
Fuente: Adaptado de De La Fuente 2011.

En los primeros pasos el salto de las distancias es pequeño, mientras que en los últimos el salto entre pasos es mayor. Considerando que las distancias pequeñas indican clústeres homogéneos y que grandes distancias definen clústeres heterogéneos, el punto de corte será aquel paso en el que empiezan a producirse los saltos más bruscos (De La Fuente 2011).

4.5 Paso 5: Interpretación de los resultados

Finalmente, una vez identificado cuáles serán los clústeres con los que se trabajará, se tendrá que dar una interpretación a la clasificación obtenida. Para esto, se deben analizar las variables del estudio y determinar cuál es la diferencia del comportamiento de estas variables entre los clústeres identificados, por ejemplo, a través de estadísticos descriptivos. El objetivo es encontrar las características que diferencien a cada clúster.

V. APLICACIÓN DEL ALGORITMO DE WARD

4.1 Descripción del problema

El objetivo de este ejemplo aplicativo es identificar los grupos homogéneos que se puedan formar a partir de la actividad de las salas de proyección de cine en las Comunidades Autónomas (CCAA) de España, utilizando el algoritmo de Ward.

Para esto, se trabajó en base a la actividad de las salas de cine en las CCAA de España con información obtenida en el año 1998 por el Instituto Nacional de Estadística de España (ver Anexo 1).

Los criterios para realizar la agrupación de las Comunidades, están en relación a las siguientes variables:

- Número de cines (Cines).
- Número de películas proyectadas (Películas).
- Número de espectadores de películas españolas (Pelis_Españ).
- Número de espectadores de películas extranjeras (Pelis_Extran).
- Recaudación obtenida en miles de pesetas (Recaudación).

4.2 Paso 1: Análisis preliminar de las variables

Para comprender el fenómeno en estudio, primero se analizaron las variables de manera descriptiva.

Cuadro 8: Estadísticos descriptivos por variable.

	N	Mínimo	Máximo	Media	Desviación estándar	Asimetría	
	Estadístico	Estadístico	Estadístico	Estadístico	Estadístico	Estadístico	Error estándar
Número de cines	17	22	468	164.94	149.72	1.14	0.55
Número de películas	17	234	649	377.24	97.59	1.43	0.55
Espect_Español	17	120,135	2,710,431	728,227.76	700,616.83	1.71	0.55
Espect_Extranj	17	769,674	13,976,149	4,414,753.59	4,161,240.76	1.56	0.55
Recaudación	17	526,496	9,963,937	2,940,129.41	2,636,639.36	1.67	0.55
N válido (por lista)	17						

Fuente: Elaboración propia.

El cuadro 8 muestra que el número de salas de cine oscila entre 22 (La Rioja) y 468 (Cataluña), y en promedio se obtuvo 165 salas por CCAA. En promedio, se proyectaron 377 películas (títulos) por Comunidad. Asimismo, el número promedio de espectadores de las películas extranjeras es muy superior al de películas españolas.

El hecho que haya Comunidades con más habitantes que otras y, por lo tanto, tengan mayor cantidad de equipos de cine, proyecten más títulos, reciban más espectadores y obtengan una mayor recaudación podría ser un indicio de que existan datos atípicos. Para validar esta conjetura, se observó la dispersión de las variables en diagramas de cajas.

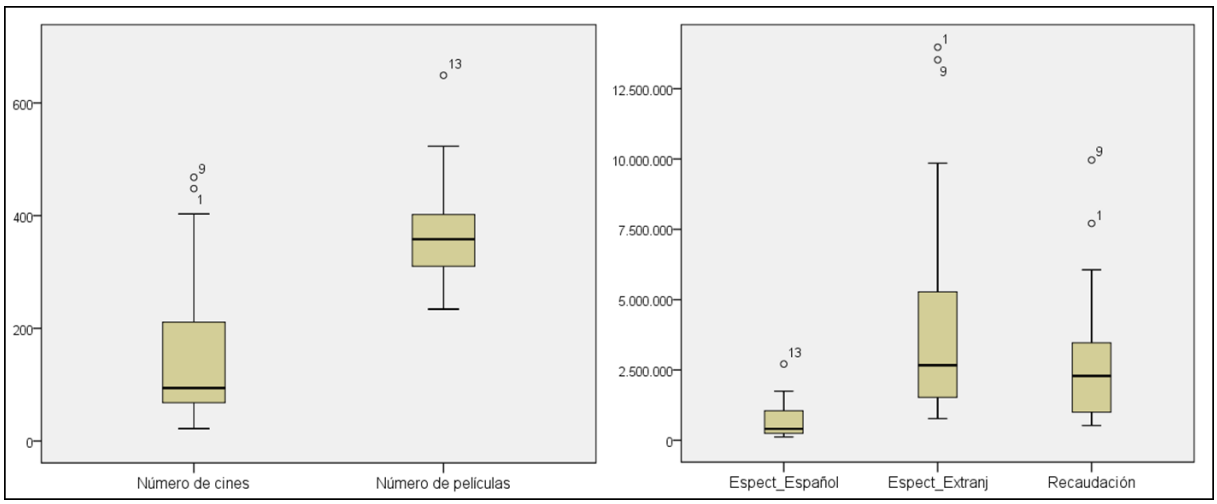

Figura 6: Diagramas de caja por variable
Fuente: Elaboración propia**.**

En la figura 6 se encontró que la variable *número de cines* presentaba dos casos atípicos, que corresponden a Andalucía (1) y Cataluña (9), que son las Comunidades con mayor número de cines. También se encontró un caso atípico en el *número de películas* estrenadas, que

corresponde a Madrid (13). Del mismo modo, respecto al *número de espectadores* y a la *recaudación*, se identificó a estas mismas tres Comunidades como casos atípicos.

El hecho de que los casos atípicos se localicen en la parte superior de la distribución indica que se trata de distribuciones asimétricas positivas, lo cual se deberá corregir antes de seguir con los siguientes pasos.

Para corregir la asimetría positiva, considerando la escalera de Tukey, se deben sustituir los datos originales por su raíz cuadrada o logaritmo, según la intensidad. En este caso, se optó por transformar con la raíz cuadrada a las variables con datos atípicos.

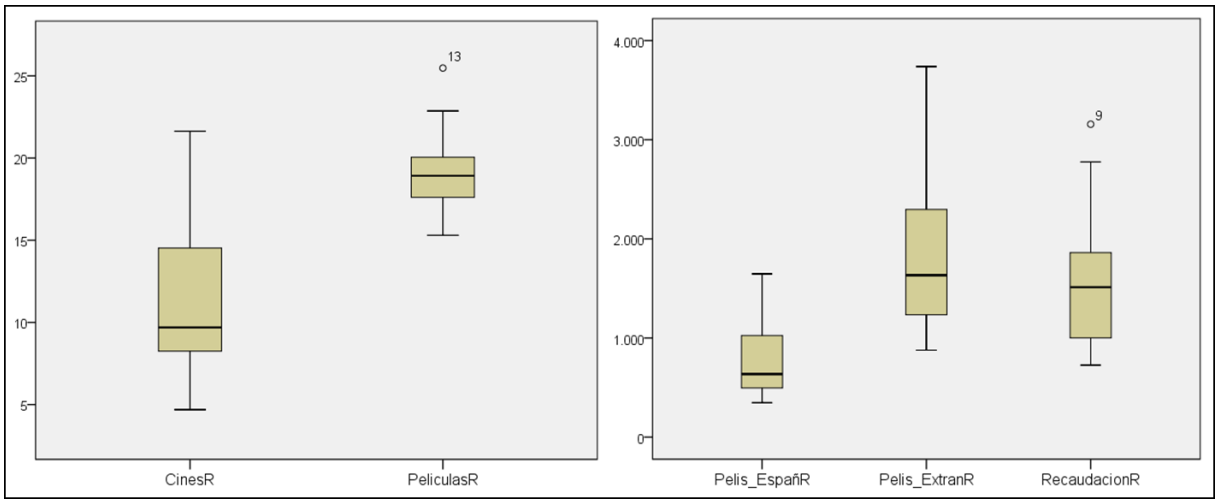

Figura 7: Diagramas de caja por variable transformada
Fuente: Elaboración propia.

Como muestra la figura 7, al transformar las variables por su raíz cuadrada, se logró reducir la asimetría de las variables y, por lo tanto, la influencia que puedan tener los datos atípicos.

Antes de proceder con el análisis clúster en sí, fue necesario comprobar hasta qué punto los datos cumplen con los supuestos mencionados en el capítulo anterior:

- **Ausencia de correlación entre las variables:**
  Se analizó la matriz de correlaciones para ver el grado de correlación entre las variables consideradas.

Cuadro 9: Correlaciones bivariadas (variables transformadas).

		Cines	Películas	Pelis_Españ	Pelis_Extran	Recaudación
Cines	Correlación de Pearson	1	.318	,942**	,801**	,913**
	Sig. (bilateral)		.214	.000	.000	.000
Películas	Correlación de Pearson	.318	1	.451	-.040	.156
	Sig. (bilateral)	.214		.069	.879	.550
Pelis_Españ	Correlación de Pearson	,942**	.451	1	,616**	,795**
	Sig. (bilateral)	.000	.069		.008	.000
Pelis_Extran	Correlación de Pearson	,801**	-.040	,616**	1	,960**
	Sig. (bilateral)	.000	.879	.008		.000
Recaudación	Correlación de Pearson	,913**	.156	,795**	,960**	1
	Sig. (bilateral)	.000	.550	.000	.000	

**. La correlación es significativa en el nivel 0,01 (2 colas).

Fuente: Elaboración propia.

Conceptualmente, la variable *recaudación* estaría fuertemente correlacionada con el *número de espectadores*. En el cuadro 9, se evidenció que existe una elevada correlación de la *recaudación* con *el número de salas de cine* y de *espectadores*. Para reducir el efecto de esta alta correlación, y para una mejor interpretación de los resultados, se calculó el *Gasto medio por espectador* (recaudación divida entre el número de espectadores) para reemplazar esta variable.

Cuadro 10: Correlaciones bivariadas – Gasto medio.

		Cines	Películas	Pelis_Españ	Pelis_Extran	Gasto_medio
Gasto_medio	Correlación de Pearson	.081	,623**	.286	-.212	1
	Sig. (bilateral)	.758	.008	.266	.414	

**. La correlación es significativa en el nivel 0,01 (2 colas).

Fuente: Elaboración propia.

El cuadro 10 muestra la correlación (considerando variables transformadas) del *Gasto medio* contra el resto de variables. En él se observa que, se logró reducir la alta correlación que tenía la *recaudación* con el *número de salas de cine* y de *espectadores*.

- **Número de variables no muy elevado.**

  En este caso, se dispone de cinco variables, el cual es un número manejable y pertinente para el ejemplo. Además, debido a que no se presentó algún problema de correlación, no hubo necesidad de excluir alguna variable.

- **Las variables no deben estar medidas en escalas diferentes.**
  Las variables de estudio están medidas en unidades diferentes. Para solucionar la diferencia de métricas de las variables, estas fueron estandarizadas o tipificadas (ver Anexo 2).

Finalmente, se tiene que las Comunidades Autónomas españolas serán clasificadas considerando las siguientes variables (transformadas y estandarizadas):

- Número de cines (ZCinesR).
- Número de películas proyectadas (ZPeliculasR).
- Número de espectadores de películas españolas (ZPelis_EspañR).
- Número de espectadores de películas extranjeras (ZPelis_ExtranR).
- Gasto promedio por espectador (ZGasto_medioR).

4.3 Paso 2: Elección de la medida de asociación o proximidad.

Utilizando la distancia euclidiana al cuadrado sobre las variables de análisis se obtuvo la siguiente matriz de distancias, proximidad o similaridad entre las 17 Comunidades Autónomas, calculando 136 medidas de proximidad.

Cuadro 11: Matriz de proximidad.

	Andalucía	Aragón	Asturias	Baleares	Canarias	Cantabria	Cast. Mancha	Cast. León	Cataluña	Valencia	Extrema.	Galicia	Madrid	Murcia	Navarra	País Vasco	La Rioja
Andalucía	0.00	12.99	20.54	25.08	11.35	27.48	5.07	14.09	8.32	2.92	17.01	7.32	30.73	13.93	23.75	10.08	27.46
Aragón	12.99	0.00	1.97	6.44	2.39	2.95	2.32	2.08	14.72	7.72	3.58	1.53	25.24	0.95	3.48	1.97	3.16
Asturias	20.54	1.97	0.00	3.00	2.27	1.73	6.97	4.61	21.70	11.91	3.15	3.58	25.79	0.77	0.53	4.06	1.26
Baleares	25.08	6.44	3.00	0.00	6.38	5.17	12.37	13.79	18.50	12.29	11.42	8.42	16.12	5.08	1.31	4.49	6.53
Canarias	11.35	2.39	2.27	6.38	0.00	6.98	3.77	3.98	18.02	6.12	2.00	0.95	27.49	0.67	3.55	3.86	5.60
Cantabria	27.48	2.95	1.73	5.17	6.98	0.00	10.07	5.92	24.62	18.13	6.77	7.53	29.79	3.58	2.36	6.08	0.64
Cast. Mancha	5.07	2.32	6.97	12.37	3.77	10.07	0.00	3.20	9.96	3.61	5.80	1.02	24.17	3.54	9.50	2.85	10.04
Cast. León	14.09	2.08	4.61	13.79	3.98	5.92	3.20	0.00	22.52	12.26	1.76	2.53	37.28	2.51	8.00	6.87	4.38
Cataluña	8.32	14.72	21.70	18.50	18.02	24.62	9.96	22.52	0.00	5.06	27.95	13.17	14.34	18.08	22.20	7.02	28.83
Valencia	2.92	7.72	11.91	12.29	6.12	18.13	3.61	12.26	5.06	0.00	13.12	4.09	17.33	7.96	12.87	3.78	18.98
Extremadura	17.01	3.58	3.15	11.42	2.00	6.77	5.80	1.76	27.95	13.12	0.00	2.83	38.01	1.68	5.71	8.57	4.01
Galicia	7.32	1.53	3.58	8.42	0.95	7.53	1.02	2.53	13.17	4.09	2.83	0.00	25.12	1.10	5.60	2.46	6.76
Madrid	30.73	25.24	25.79	16.12	27.49	29.79	24.17	37.28	14.34	17.33	38.01	25.12	0.00	26.23	22.52	15.67	34.38
Murcia	13.93	0.95	0.77	5.08	0.67	3.58	3.54	2.51	18.08	7.96	1.68	1.10	26.23	0.00	2.10	2.95	2.73
Navarra	23.75	3.48	0.53	1.31	3.55	2.36	9.50	8.00	22.20	12.87	5.71	5.60	22.52	2.10	0.00	4.58	2.41
País Vasco	10.08	1.97	4.06	4.49	3.86	6.08	2.85	6.87	7.02	3.78	8.57	2.46	15.67	2.95	4.58	0.00	7.70
La Rioja	27.46	3.16	1.26	6.53	5.60	0.64	10.04	4.38	28.83	18.98	4.01	6.76	34.38	2.73	2.41	7.70	0.00

Fuente: Elaboración propia.

La matriz de proximidad mostró que existe una gran similitud en la actividad de los cines de Asturias y Navarra (distancia 0.53). También se encontró que Extremadura y Madrid son las Comunidades más diferentes (distancia 38.01), en cuanto a su actividad cinematográfica.

Se tuvo en cuenta esta información para realizar un seguimiento del proceso de agrupamiento en los siguiente pasos, donde se observó que Asturias (3) y Navarra (15) fueron las primeras CCAA en agruparse.

4.4 Paso 3: Elección del criterio de agrupamiento (Algoritmo de Ward)

La matriz de pertenencia o historial de agrupamiento muestra el proceso de fusión en cada etapa, según el algoritmo de Ward. En este caso, la matriz partió de 17 conglomerados, formados por cada una de las Comunidades, hasta agruparlas en solo un clúster.

Cuadro 12: Matriz de pertenencia.

Etapa	Clúster combinado		Coeficientes	Primera aparición del clúster de etapa		Etapa siguiente
	Clúster 1	Clúster 2		Clúster 1	Clúster 2	
1	3	15	.265	0	0	7
2	6	17	.583	0	0	11
3	5	14	.916	0	0	10
4	7	12	1.427	0	0	9
5	8	11	2.306	0	0	10
6	2	16	3.289	0	0	9
7	3	4	4.638	1	0	11
8	1	10	6.097	0	0	13
9	2	7	7.638	6	4	12
10	5	8	9.571	3	5	12
11	3	6	12.629	7	2	14
12	2	5	16.409	9	10	14
13	1	9	20.382	8	0	15
14	2	3	31.371	12	11	16
15	1	13	45.613	13	0	16
16	1	2	80.000	15	14	0

Fuente: Elaboración propia.

De acuerdo al cuadro 12, en la primera etapa se formó un clúster con la unión de Asturias (3) y Navarra (15), tal como lo indicó la matriz de proximidades. El coeficiente 0.265 indica la distancia acumulada, o el incremento de la varianza al usar el método de Ward, del clúster formado. Este nuevo grupo se unió a otra Comunidad o clúster en la etapa siete.

En la segunda etapa se unieron las Comunidades de Cantabria (6) y La Rioja (17), a una distancia acumulada de 0.583 (0.265+0.318).

En la séptima etapa se juntaron la Comunidad de Baleares (4) con el clúster constituido por Asturias (3) y Navarra (15) en la etapa uno, formando un nuevo clúster (Baleares, Asturias y Navarra) a una distancia acumulada de 4.638.

4.5 Paso 4: Selección de los clústeres

El dendograma muestra a cada Comunidad y la distancia a la que fueron agrupadas en diferentes etapas, representando gráficamente el proceso observado en la matriz de pertenencia.

La distancia que muestra el dendograma que se obtiene al usar el software SPSS, no necesariamente guardará relación con las distancias acumuladas obtenidas en la matriz de pertenencia, puesto que el SPSS re-escala las distancias de 0 a 25.

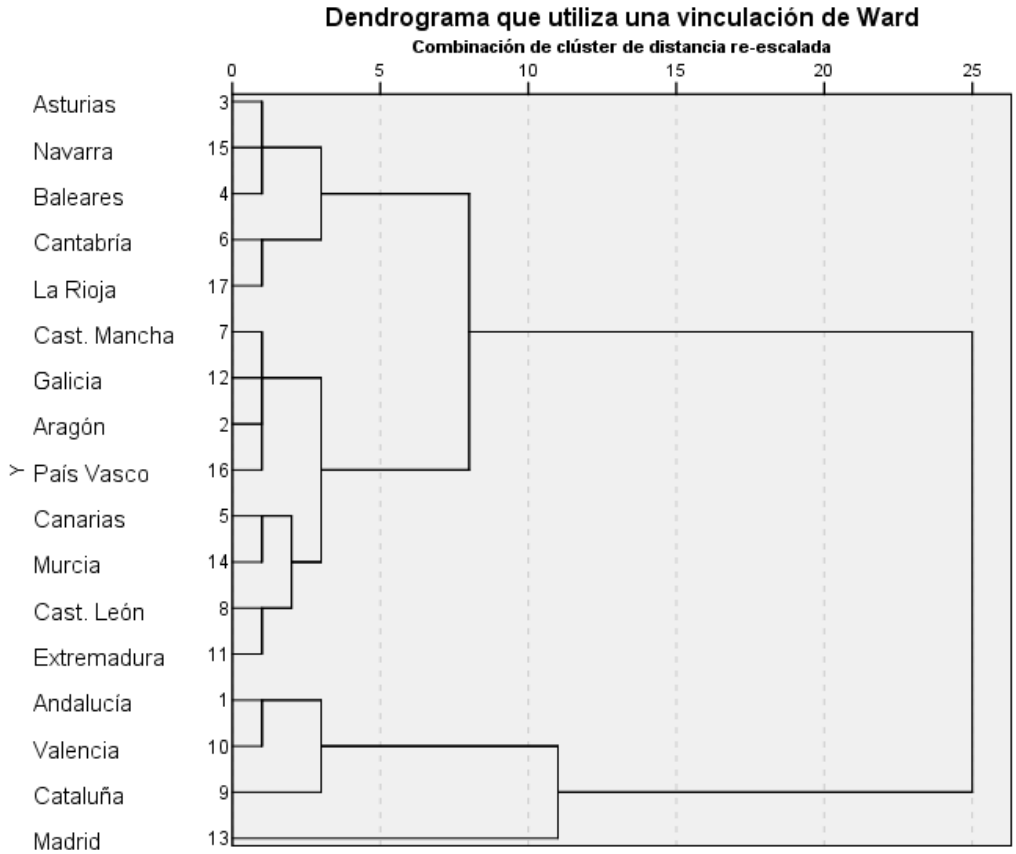

Figura 8: Dendograma usando el algoritmo de Ward
Fuente: Elaboración propia.

Como muestra la figura 8, el primer gran incremento en las distancias se produjo al formarse cuatro clústeres, alrededor de la distancia re-escalada de 5. Otro gran incremento en las distancias ocurrió cuando formaron tres clústeres, a una distancia aproximada de 10. Si se detuviera el proceso en la distancia re-escalada de 15, se considerarían dos clústeres.

Para determinar el número aconsejable de clústeres a utilizar, se construyó el gráfico de saltos de distancia.

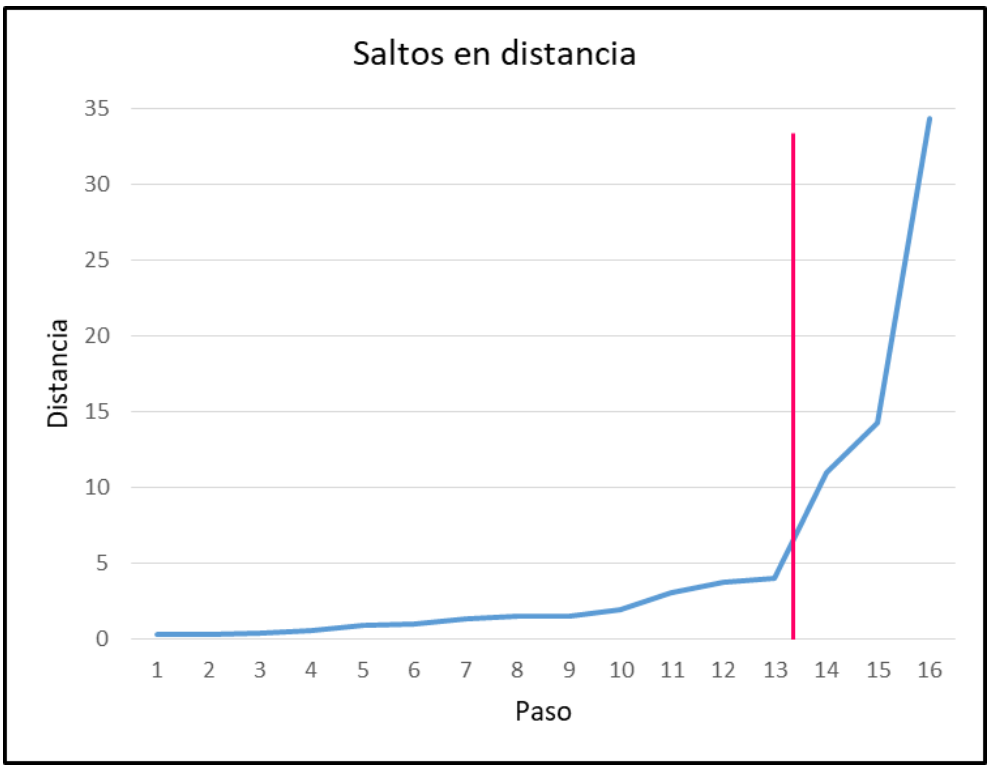

Figura 9: Saltos en distancia.
Fuente: Elaboración propia.

La figura 9 muestra un cambio significativo en los incrementos de distancia a partir del paso 13, por lo que se decidió que la solución óptima era la que presentaba cuatro clústeres.

La solución escogida consideró las siguientes agrupaciones:

- Clúster 1 = {Andalucía, Cataluña, Valencia}
- Clúster 2 = {Aragón, Canarias, Cast. Mancha, Cast. León, Extremadura, Galicia, Murcia}
- Clúster 3 = {Asturias, Baleares, Cantabria, Navarra, La Rioja}
- Clúster 4 = {Madrid}

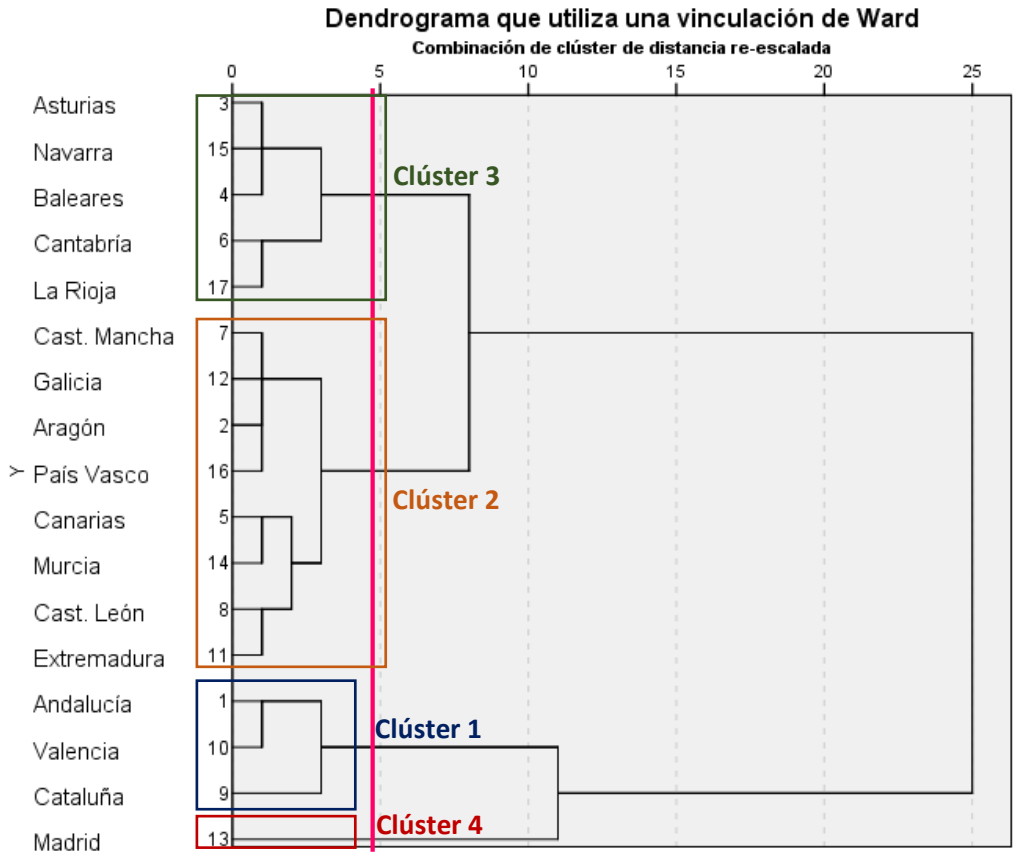

Figura 10: Clústeres seleccionados en el dendograma.
Fuente: Elaboración propia.

4.6 Paso 5: Interpretación de los resultados

Cuadro 13: Comparación de medias entre clústeres.

Método Ward		Número de cines	Número de películas	Espect_Español	Espect_Extranj	Gasto_medio
1	Media	405.33	389.00	1,463,722.00	12,451,111.00	566.58
	N	3	3	3	3	3
	Desviación estándar	91.77	53.69	248,653.53	2,264,035.97	77.46
2	Media	122.13	328.25	529,237.38	3,665,418.13	545.31
	N	8	8	8	8	8
	Desviación estándar	52.81	52.49	266,729.85	1,359,134.60	36.00
3	Media	41.60	394.20	208,875.20	1,385,856.20	606.08
	N	5	5	5	5	5
	Desviación estándar	19.53	90.05	60,733.89	484,712.69	25.24
4	Media	403.00	649.00	2,710,431.00	1,444,852.00	689.60
	N	1	1	1	1	1
	Desviación estándar					
Total	Media	164.94	377.24	728,227.76	4,414,753.59	575.43
	N	17	17	17	17	17
	Desviación estándar	149.72	97.59	700,616.83	4,161,240.76	55.25

Fuente: Elaboración propia.

Se tiene que los centroides de los clústeres son:

- Clúster 1 = {405.33; 389.00; 1,463,722.00; 12,451,111.00; 566.58}
- Clúster 2 = {122.13; 328.25; 529,237.38; 3,665,418.13; 545.31}
- Clúster 3 = { 41.60; 394.20; 208,875.20; 1,385,856.20; 606.08}
- Clúster 4 = {403.00; 649.00; 2,710,431.00; 1,444,852.00; 689.60}

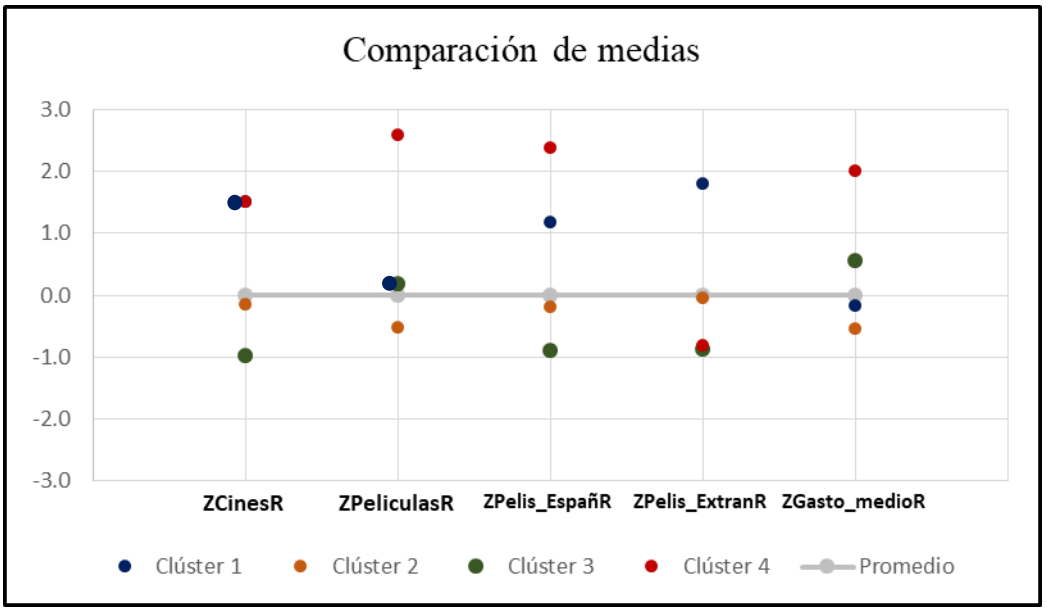

Figura 11: Comparación de medias (transformadas y estandarizadas) entre clústeres. Fuente: Elaboración propia.

Los valores medios de las variables de cada grupo (centroides) sirvieron de base para definir el perfil de los clústeres:

- El clúster 1, formado por tres CCAA, tuvo la menor tasa de preferencia por películas españolas (películas españolas/total de espectadores); sin embargo, presentó la mayor cantidad de salas de cine y de espectadores. Este clúster se caracterizó porque en éstas CCAA el mercado está algo saturado (gran cantidad de salas de cine), pero con una gran afluencia de espectadores con un gasto por debajo del promedio total, y una ligera preferencia por las películas extranjeras.
- El clúster 2 se caracterizó por tener una actividad de acuerdo al promedio del mercado; sin embargo, en estas CCAA se presentó el menor gasto promedio por espectador.
- El clúster 3 registró la menor actividad en la industria del cine. En estas CCAA existe una menor cantidad de salas de cine y un menor número de espectadores en general;

sin embargo, presenta un gasto por espectador superior al promedio y una mayor cantidad de películas de estreno. En este clúster, la actividad del cine se registró en un nicho de mercado pequeño, pero con un alto nivel de gasto.

- El clúster 4, que corresponde a Madrid, se diferenció de los demás por ser el clúster con mayor número de películas de estreno, la mayor preferencia por películas españolas (películas españolas/total de espectadores) y en la que se registró un mayor gasto promedio por espectador (690 pesetas). Se puede decir que en este clúster se generó la mayor actividad en esta industria en comparación a los otros clústeres y, además, destacó por su preferencia por el cine nacional.

VI. CONCLUSIONES

1. Antes de usar el algoritmo de Ward para agrupar casos o individuos, es importante asegurar que las variables sean relevantes en el propósito de clasificación del problema planteado; y también verificar que no exista influencia de valores extremos, que las variables no estén altamente correlacionadas y que no estén medidas en escalas diferentes. De lo contrario, habrá que seguir los pasos propuestos para corregir estos problemas en las variables y proceder con los siguientes pasos.

2. Aplicando el algoritmo de Ward se pudo detectar, en base a los saltos de distancia, cuál es el número óptimo de clústeres que se pueden formar con los datos de estudio.

3. Usando el algoritmo de Ward, luego de definir el número de clústeres a usar, se pudo encontrar grupos cuyos elementos comparten características similares dentro de cada clúster, pero que son diferentes en comparación a los otros grupos formados.

4. En el caso de las Comunidades Autónomas de España, se demostró que éstas pueden ser clasificadas en cuatro grupos, en función a la actividad de sus salas de cine, a pesar de no tener información a priori sobre la cantidad de grupos. Además, se encontró que estos cuatro clústeres tenían características diferentes entre sí, en función de las variables de estudio (número de salas, número de películas de estreno, número de espectadores y gasto medio).

VII. REFERENCIAS BIBLIOGRÁFICAS

De La Fuente S. 2011. Análisis conglomerados. Madrid, España, Universidad Autónoma de Madrid. Libro electrónico.

Del Campo P; Pardo C. 2007. Combinación de métodos factoriales y de análisis de conglomerados en R: el paquete FactoClass. Revista Colombiana de Estadística.

Gallardo, J. s.f. Introducción al Análisis Clúster. Universidad de Granada, Granada, España. Disponible en http://www.ugr.es/~gallardo/pdf/cluster-g.pdf

Legendre P; Murtagh F. 2014. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?. Journal of Classification 31: 274-295.

Oliva C. 2015. Métodos para la segmentación de datos longitudinales. Aplicación a datos de rendimientos de cultivos en Argentina. Tesis Lic. Buenos Aires, Argentina, UBA. 72p.

Kumar V; Steinbach M; Tan PN. 2006. Introduction to Data Mining. Harlow, England, Pearson. 725p.

Pedret R; Sagnier L; Camp F. 2003. Herramientas para segmentar mercados y posicionar productos. Barcelona, España, Planeta. 329p.

ANEXOS

Anexo 1: Actividad en salas de cines por Comunidades Autónomas.

	CCAA	Nro. Cines	Nro. Películas	Nro. Espectadores		Recaudación (miles pesetas)
				Películas españolas	Películas extranjeras	
1	Andalucía	448	330	1,380,202	13,976,149	7,709,721
2	Aragón	76	310	580,526	3,513,294	2,370,874
3	Asturias	55	383	207,100	1,524,423	1,000,709
4	Baleares	68	523	280,851	2,081,987	1,496,299
5	Canarias	94	394	345,213	4,056,725	2,288,764
6	Cantabria	26	315	190,540	1,149,257	847,231
7	Cast. Mancha	211	295	1,049,698	5,319,556	3,464,668
8	Cast. León	102	234	404,716	2,406,798	1,490,303
9	Cataluña	468	402	1,743,383	13,527,492	9,963,937
10	Valencia	300	435	1,267,581	9,849,692	6,061,359
11	Extremadura	69	309	226,139	1,614,986	912,405
12	Galicia	166	341	570,921	4,465,381	2,680,531
13	Madrid	403	649	2,710,431	1,444,852	2,865,482
14	Murcia	88	358	326,445	2,669,391	1,647,870
15	Navarra	37	441	245,750	1,403,940	981,839
16	País Vasco	171	385	730,241	5,277,214	3,673,712
17	La Rioja	22	309	120,135	769,674	526,496
	Total	**2,804**	**6,413**	**12,379,872**	**75,050,810**	**49,982,201**

Fuente: Adaptado de De La Fuente 2011.

Anexo 2: Variables de estudio para el análisis clúster (estandarizadas y transformadas).

	CCAA	ZCinesR	ZPelículasR	ZPelis_EspañR	ZPelis_ExtranR	ZGasto_medioR
1	Andalucía	1.72	-0.47	1.1	2.06	-1.36
2	Aragón	-0.53	-0.70	0.0	-0.05	0.09
3	Asturias	-0.77	0.12	-0.9	-0.78	0.07
4	Baleares	-0.62	1.50	-0.7	-0.54	1.05
5	Canarias	-0.36	0.24	-0.5	0.11	-1.01
6	Cantabria	-1.19	-0.64	-0.9	-0.96	1.04
7	Cast. Mancha	0.52	-0.88	0.7	0.44	-0.56
8	Cast. León	-0.28	-1.67	-0.4	-0.42	-0.82
9	Cataluña	1.80	0.32	1.5	2.00	1.38
10	Valencia	1.02	0.66	1.0	1.38	-0.54
11	Extremadura	-0.61	-0.71	-0.8	-0.74	-1.49
12	Galicia	0.22	-0.34	-0.1	0.22	-0.78
13	Madrid	1.52	2.59	2.4	-0.81	2.01
14	Murcia	-0.41	-0.15	-0.6	-0.32	-0.45
15	Navarra	-1.01	0.72	-0.8	-0.83	0.38
16	País Vasco	0.25	0.14	0.2	0.43	0.67
17	La Rioja	-1.26	-0.71	-1.2	-1.18	0.32

Fuente: Elaboración propia

Anexo 3: Diagrama de témpanos (clústeres seleccionados).

Es otra forma de mostrar cómo quedan los grupos, dependiendo del número de clústeres a considerar. Se lee por filas y cada espacio en blanco indica que se cambia de clúster.