UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS

FACULTAD 3

Trabajo de Diploma para optar por el título de Ingeniero en Ciencias Informáticas.

Título: "Análisis exploratorio de datos como soporte a la toma de decisiones en la formación profesional desde la producción"

Autora: Carmen Cutié Torres

Tutor: Lic. Bolívar Ernesto Medrano Broche

Lic. Lytyet Fernández Capestany

Asesor: MSc. Yunier E. Tejeda Rodríguez

La Habana, Cuba

2012

DECLARACIÓN DE AUTORÍA

Declaro ser el autor de la presente tesis y reconozco a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmo la presente a los _____ días del mes de _____del año 2012.

Autor:

_____

Carmen Cutié Torres

Tutor:

_____ _____

Lic. Bolívar Ernesto Medrano Broche

Lic. Lytyet Fernández Capestany

A mis tutores Bolivar Ernesto Medrano y Lytyet Fernández.

A Yunier Tejeda, Hugo Arnaldo Martínez.

A mi madre, Solangel.

A mi familia por el apoyo inmenso e incondicional.

A L.

A todos mis amigos...

A Odisleysi Martínez y Jorge Regalado, en especial.

Resumen

El presente trabajo es el resultado de la aplicación de métodos multivariados para la toma de decisiones. Tiene como finalidad realizar un análisis exploratorio de un conjunto de datos, que generan información provechosa para el proceso de incorporación de estudiantes a proyectos de software de la Universidad de las Ciencias Informáticas (UCI).

Se tomó una muestra de 122 estudiantes de la Facultad 3, con sus resultados docentes hasta el tercer año académico y los valores procedentes de instrumentos de diagnóstico del Centro de Innovación y Calidad de la Educación (CICE).

Para el análisis se utilizaron los métodos estadísticos: análisis de componentes principales y análisis de *cluster*. Estos permitieron reducir la dimensión de los datos y clasificar a los estudiantes en grupos similares de acuerdo a sus resultados académicos.

"**R**" fue la herramienta empleada para la implementación de un paquete de funciones que integró los métodos antes mencionados.

Se analizaron 67 variables, que fueron reducidas a 8 componentes principales, logrando explicar gran parte de la información subyacente en los datos. También se clasificó a los estudiantes en tres grupos: *Aptos, No aptos y Aptos con limitaciones*, lo cual permitirá a los expertos trazar estrategias que incidan directamente en el proceso de formación profesional.

Palabras claves: análisis multivariado, análisis de componentes principales, análisis de *cluster*.

Índice

Índice de tablas

Índice de figuras

Introducción

El desarrollo creciente de las nuevas Tecnologías de la Informática y las Comunicaciones (TIC) impone un ritmo acelerado a la forma en que el hombre adquiere y gestiona el conocimiento. Las universidades juegan un importante rol social, pues en ellas se genera y transmite gran parte de este conocimiento a través de modelos de enseñanza y aprendizaje sujetos a constante perfeccionamiento.

La formación profesional se considera el principal y más complejo proceso en una universidad. Esto se debe a la gran variabilidad en las características de los estudiantes que son el objeto a transformar, y a la variedad de condiciones que confluyen en el proceso para desarrollar esta transformación (Massy, Wilger 1995). Dicho proceso de formación, regula la forma en que se desarrollan en los alumnos las capacidades profesionales. Estas abarcan el conjunto de conocimientos, destrezas y aptitudes cuya finalidad es la realización de actividades vinculadas a una profesión.

En la Universidad de las Ciencias Informáticas (UCI) el proceso de formación profesional tiene características que lo diferencian del resto de los Centros de Educación Superior (CES) del país. Se compone de dos ciclos: el primero denominado ciclo básico que comprende los cinco primeros semestres de la carrera y un segundo denominado ciclo profesional que comprende los restantes semestres y la vinculación del estudiante a los proyectos de desarrollo de software.

La incorporación masiva de personal a la producción provoca, en ocasiones, disminución de la calidad del éxito de los proyectos que se ejecutan en la universidad. La incorrecta gestión de los recursos humanos en los equipos de desarrollo trae como consecuencia: el aumento de los problemas de comunicación entre sus miembros; la pérdida de tiempo en capacitaciones; la incorrecta distribución de las tareas; estimaciones de cronogramas de trabajo no reales; el gasto excesivo de recursos; y el aumento de los costos de producción. Estos efectos nocivos conllevan al fracaso o atraso de los proyectos, lo cual compromete el prestigio de la institución como entidad desarrolladora de software.

Se tiene referencia de los siguientes trabajos: (Medrano, Martínez, Fernández, Díaz 2010), (Martínez, Medrano, Fernández, Tejeda 2012), que abordan el tema de la selección de personal, vinculado a la producción en la UCI. Los autores realizan la clasificación de estudiantes, basada en tres grupos obtenidos al aplicar métodos de *cluster* jerárquicos a los datos procedentes de las notas finales del segundo año.

En estos artículos se brinda un procedimiento que favorece el proceso de toma de decisiones para la incorporación de estudiantes a los proyectos. No se emplea información relacionada con aspectos como la personalidad o habilidades técnicas de los alumnos, ni se establece una validación de los resultados atendiendo a estos aspectos.

Actualmente el Sistema de Gestión académica (Akademos) proporciona los datos relacionados con la trayectoria docente de los estudiantes de la UCI. Antes de comenzar el ciclo profesional, se cuenta con un total de 37 asignaturas distribuidas de la siguiente manera: primer año (13 materias), segundo año (11 materias) y tercer año (13 materias). Además, existen instrumentos de diagnóstico utilizados para evaluar la calidad de los procesos formativos.

El Centro de Innovación y Calidad de la Educación (CICE) brinda los resultados de los diagnósticos clasificados de la siguiente manera:

- Estilos decisionales (4 indicadores): Permite identificar los estilos predominantes en los estudiantes en la toma de decisiones.
- Autopercepción de Belbin (9 indicadores): Permite conocer el rol o los roles que mejor pueden desempeñar los estudiantes en un equipo de trabajo e identificar las aptitudes para el liderazgo.
- Matriz de fortalezas y debilidades (17 indicadores): Permite conocer cuáles son las principales fortalezas y debilidades presentes en los estudiantes.

En total se analizan 30 aspectos relacionados con las habilidades y capacidades.

Por un lado, se tienen los resultados docentes que caracterizan profesionalmente a los estudiantes, en un momento dado de la carrera. Mientras que por otro, se encuentran los datos del conjunto de habilidades cognitivas desarrolladas como resultado del aprendizaje durante el primer ciclo. Ambos proceden de fuentes distintas, y son agrupados y almacenados con procedimientos diferentes.

Las notas de los estudiantes y los resultados obtenidos de los instrumentos del CICE, conforman un conjunto de datos amplio (67 variables por cada estudiante). El análisis de estas observaciones puede generar información relevante para el proceso de toma de decisiones, en la selección de personal para determinado proyecto o rol. Sin embargo, la realización de una interpretación a priori de estos datos, resulta prácticamente imposible.

A partir de la **situación problemática** planteada surge el siguiente **problema a resolver**: ¿Cómo interpretar los datos que caracterizan a los estudiantes, que favorezca la toma de decisiones en el proceso de incorporación de los mismos a proyectos de desarrollo de software de la universidad?

Se propone como **objeto de estudio**: Estadística Multivariada, y como **objetivo general**: Crear una herramienta de apoyo a la toma de decisiones, en el proceso de incorporación de estudiantes a proyectos de desarrollo de software de la universidad.

En este contexto se define como **campo de acción**: El análisis de datos en procesos de toma de decisiones.

Teniendo como **idea a defender** que si se cuenta con una herramienta que facilite la interpretación de los datos que caracterizan a los estudiantes, se podrán obtener elementos

para la toma de decisiones, en la incorporación a proyectos de desarrollo de software de la universidad.

Las **tareas** a desarrollar durante la investigación son:

1. Elaboración del marco teórico de la investigación.
2. Estudio de aspectos relacionados con la gestión de recursos humanos basada en competencias, en proyectos de software.
3. Estudio del análisis multivariado y su aplicación al proceso de toma de decisiones, respecto a la incorporación de personal a proyectos de software.
4. Estudio del análisis de componentes principales y el análisis de *cluster*.
5. Estudio de herramientas utilizadas para el análisis estadístico de datos.
6. Recopilación de los datos a utilizar en el análisis.
7. Implementación de funciones que permitan aplicar métodos de análisis multivariado.
8. Análisis e interpretación de los resultados de las técnicas de análisis exploratorio de datos aplicadas.

Entre los **métodos investigativos** utilizados se destacan los que se mencionan a continuación:

- **Método Análisis histórico lógico:** Para el estudio de la evolución histórica de algunos métodos, algoritmos y herramientas que se emplean en el análisis exploratorio de datos. También para la revisión de los trabajos anteriores que abordan el tema de la investigación y la extracción de aspectos necesarios que puedan ser utilizados como punto de referencia y comparación de los resultados.
- **Método Inductivo-Deductivo:** Permite a través de un razonamiento llegar a un grupo de conocimientos particulares y generales.
- **Método Analítico - Sintético**: Para descomponer todos los aspectos teóricos que se analizan en la investigación y luego sintetizarlos en la propuesta de solución.

La investigación reflejada en el presente documento se encuentra estructurada en 3 capítulos. Estos se exponen a continuación:

Capítulo 1: Fundamentación teórica

Se analiza la importancia de los recursos humanos en proyectos de software; el aporte del análisis multivariado al proceso de selección de personal; los métodos: análisis de componentes principales y análisis de *cluster;* y las herramientas utilizadas para el análisis de datos.

Capítulo 2: Paquete de funciones para el análisis estadístico de datos.

Se desarrollan las funciones que integran el paquete para el análisis estadístico de datos, implementado en R.

Capítulo 3: Caso de Prueba.

Se realiza la prueba y validación del paquete de funciones a través de un caso de estudio.

Posibles resultados:

Al concluir la investigación se espera contar con un paquete de funciones para el análisis estadístico de datos que pueda ser utilizado como soporte a la toma de decisiones en el proceso de formación profesional de la UCI.

Capítulo 1: Fundamentación teórica

En este capítulo se exponen los elementos relacionados con la importancia de los recursos humanos en proyectos de desarrollo de software y la aplicación de técnicas de análisis multivariado a los procesos de toma de decisiones. También se describen los métodos estadísticos: análisis de componentes principales y análisis de *cluster*, así como las herramientas más utilizadas para el análisis de datos.

1.1. El factor humano en proyectos de software

La gestión eficaz de un proyecto de software se centra en las llamadas cuatro P: personal, producto, proceso y proyecto, donde el orden no resulta arbitrario (Pressman 2004). En este sentido, son varias las investigaciones donde se reconoce que los recursos humanos juegan un papel crítico en el éxito o fracaso de un proyecto de software. Sin embargo, el personal continúa siendo el factor menos formalizado en los modelos de procesos y las metodologías de desarrollo de software, las cuales se centran más en aspectos técnicos que en los aspectos humanos (Wastell 1999).

La industria de software cubana también experimenta proyectos fuera de cronograma, presupuesto y calidad. Encuestas realizadas en diversos ámbitos laborales demuestran que *la asignación de personal no adecuado, los problemas de liderazgo y los problemas entre los miembros del equipo de proyecto* se detectan como tres de las dificultades comunes asociadas con factores humanos que afectan el resultado de los proyectos de software (André 2009).

La gestión de recursos humanos (GRH) en los proyectos de software

La GRH de un proyecto desarrollo de software incluye los procesos que organizan y dirigen el equipo. Es un área de conocimiento clave dentro de la gestión de proyectos y considera la gestión por competencias como elemento esencial para alcanzar resultados exitosos en el proyecto.

En cualquier organización la GRH por competencias constituye una herramienta de apoyo para los procesos de selección, contratación, capacitación, evaluación y compensación del personal. Según estudios realizados (André 2009), resulta necesaria la incorporación del análisis de competencias individuales en los procesos de selección del personal en aras de fortalecer la interacción humana entre todos los miembros del equipo, así como el dominio y conocimiento por parte de los líderes de sus fortalezas y debilidades.

Los proyectos de desarrollo de software grandes se logran debido al esfuerzo de grupos. Las características de estos últimos juegan un papel muy importante. Entre grupos con diferentes conocimientos y diferentes niveles de experiencia, hay una diferencia del orden de 5 a 1 en la productividad. Entre grupos con similares conocimientos y niveles de

experiencia, hay una diferencia en la productividad de 2,5 a 1 (Molina 2000). Esto evidencia la necesidad de lograr equipos de proyectos con capacidades profesionales desarrolladas al mismo nivel.

De Marco y Lister exponen algunas razones por las que es necesario mantener unido a los equipos y evitar cambios repentinos en su estructura:

- Mayor productividad.
- Costes iniciales menores.
- Menor riesgo de problemas con el personal.
- Menos cambios de personal.

El proceso de selección de personal para integrar proyectos de software en la UCI, cuenta con una particularidad que lo diferencia del resto de las empresas desarrolladoras: los estudiantes constituyen el capital humano. Partiendo de esta realidad, se hace necesario el análisis de las competencias profesionales que se han desarrollado durante el primer ciclo de formación. La identificación adecuada del rol que con mayor eficiencia podrían desempeñar en un proyecto, puede contribuir a lograr mayor productividad y calidad en los resultados de trabajo. Sin embargo, el desarrollo de las competencias de los estudiantes antes de arribar a los proyectos puede ser parcial. Por tanto, se deben considerar estrategias que permitan formar completamente sus capacidades técnicas o ayudar en la adquisición de nuevas.

Dados los objetivos de esta investigación, resulta esencial tener en cuenta todo lo planteado en el epígrafe respecto a la gestión de recursos humanos en los proyecto de software, por los autores André, Molina, De Marco y Lister. Teniendo en cuenta la importancia del análisis de competencias en el proceso de selección de personal, se considera necesaria la inclusión del análisis de los resultados académicos y las encuestas “Matriz de fortalezas y debilidades”, “Estilos de decisionales” y “Autopercepción de Belbin”, para la conformación de los equipos de proyectos en la universidad.

1.2. Aplicación del análisis multivariado como soporte a la toma de decisiones en el proceso de formación

El análisis multivariado agrupa un conjunto de métodos estadísticos que analizan simultáneamente medidas múltiples de cada individuo u objeto sometido a investigación. Según Hair y Anderson todas las variables deben ser aleatorias y estar interrelacionadas de tal forma que sus diferentes efectos no puedan ser interpretados separadamente con algún sentido.

Estos autores proponen un esquema para la selección de la técnica multivariada apropiada (Fig. 1). En general existen dos tipos de técnica: las de dependencia o predicción y las de interdependencia o de identificación de estructuras.

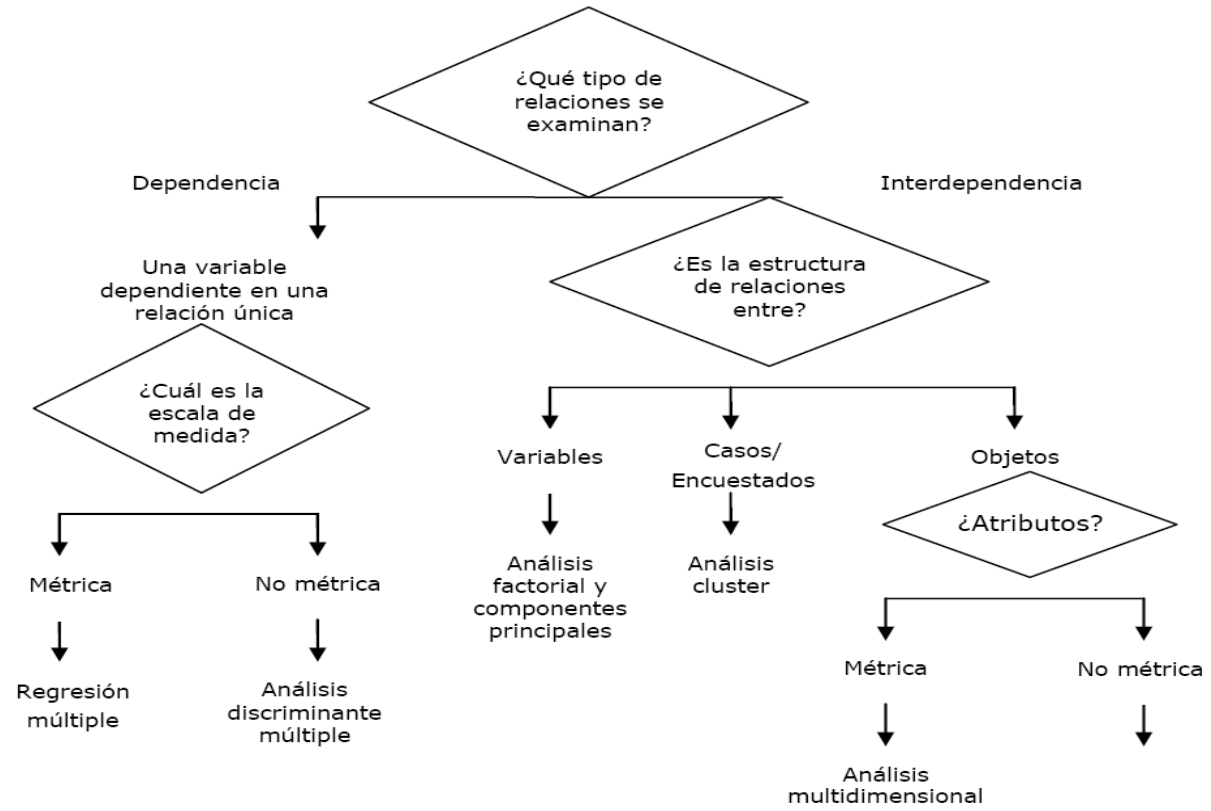


Fig. 1. Clasificación de técnicas de análisis multivariado.

El autor (Heredia 2011) aplica técnicas de estadística multivariada y de minería de datos a los procesos de formación de la Facultad de Ingeniería Industrial del Instituto Superior politécnico José Antonio Echeverría (CUJAE). Concibe un modelo orientado a aprovechar las reservas de productividad relacionadas con la mejora en las decisiones. La información empleada en el modelo se deriva de los datos que se analizan de los estudiantes.

A continuación se exponen algunos trabajos que aplican el análisis multivariado para la solución de problemas en el proceso de formación de la UCI:

En (Quintana 2011) se propone un modelo para el análisis de los datos educacionales de estudiantes de la Facultad 1, mediante técnicas minería de datos y estadística inferencial. El aporte de este estudio para la presente investigación es que combina datos de diversas fuentes y trabaja con indicadores relativos a los conocimientos, hábitos y valores.

En (Medrano et al. 2010) se propone un procedimiento donde se aplican técnicas de análisis multivariado. Los métodos aplicados son: análisis de componentes principales y análisis de *cluster*. Se emplean para generar información que es utilizada como soporte a la toma de decisiones, respecto a la selección de estudiantes para proyectos de software de la Facultad 3. En este análisis no se tuvieron en cuenta las notas del ciclo básico completo, ni la caracterización de habilidades de los estudiantes que recoge el CICE.

La documentación consultada acerca de la aplicación de técnicas de análisis multivariado se caracterizó por: establecer relaciones, leyes; investigar estructuras latentes y ensayar

maneras de organizar los datos; evaluar asociaciones entre las características; comparar los individuos y evaluar el grado de semejanza entre ellos.

Específicamente en el caso de (Quintana 2011) y (Medrano et al. 2010) hay un acercamiento a la utilización de estas técnicas para la clasificación de individuos, ya sea con fines docentes o productivos. Por tanto, se utilizarán los criterios de la estadística multivariada, unidos con los resultados obtenidos en los estudios antes mencionados.

Se debe tener en cuenta que el campo de acción y el objetivo de la presente investigación comparten similitudes con los propuestos por Medrano en su artículo. En ambas investigaciones se realiza un análisis exploratorio de datos que permite clasificar e identificar características en los alumnos que pasan a la producción. El tipo de datos (escala ordinal), el tipo variables (discretas), así como las relaciones de interdependencia que se establecen entre ellas, también son semejantes.

Se propone mejorar la calidad de la información generada en el artículo de Medrano, ya que este solo utiliza los resultados académicos hasta segundo año de la carrera, y no recoge datos relacionados con las capacidades o habilidades con que cuentan los estudiantes.

En próximos capítulos se realizará un análisis que integrará datos procedentes de las notas de los estudiantes hasta tercer año (ciclo básico) y los resultados obtenidos en encuestas del CICE (“Matriz de fortalezas y debilidades”, “Test de Autopercepción de Belbin”, “Estilos decisionales”). La interpretación de estos datos permitirá la caracterización de los estudiantes, de forma que esta información pueda ser utilizada en procesos de toma de decisión.

1.3. Tipos de datos y variables para análisis multivariado

Los datos constituyen información cruda y no son conocimiento por sí mismos. Son conjuntos de dígitos o letras empleadas para representar un hecho o caracterizar un elemento (Bouthiller, Shearer 2002). Cuando los datos se agrupan, se clasifican y se resumen para obtener una tendencia o comportamiento de un fenómeno, es decir, cuando se les agrega valor, se convierten en información (Ponjuán 2006).

Los datos pueden ser cualitativos (no métricos) o cuantitativos (métricos), y a cada uno de ellos le corresponden dos tipos de escalas. Los datos cualitativos pueden estar en una escala *nominal* o en escala *ordinal*, y los datos cuantitativos pueden expresarse en escala *de intervalo* o en escala *proporcional* (“de razón”).

Escala Nominal: Los valores en esta escala no pueden ordenarse. Ejemplo: (natación, baloncesto, gimnasia), (femenino, masculino).

Escala Ordinal: Los valores en esta escala pueden ordenarse, *pero no hay intervalo de magnitud definida e igual entre cualquier par de niveles consecutivos de la variable*. Ejemplo: (poco, suficiente, bastante, mucho), (amor, agrado, indiferencia, desagrado, odio).

Los datos en escala nominal u ordinal sólo pueden representarse mediante variables *discretas* porque el conjunto soporte (conjunto de todos los posibles valores que toma una variable) resulta finito o numerable.

Escala de Intervalo: Tiene orden e iguales diferencias entre dos puntos consecutivos en cualquier lugar de la escala, pero el punto cero es arbitrario. Ejemplo: (-10º C, 0º C, 10º C, 20º C, 30º C), (2006, 2007, 2008).

Escala proporcional: Tiene orden, posee intervalos iguales entre niveles consecutivos de la variable, y tiene un cero absoluto. Ejemplo: (ratio) (0 cm., 10 cm., 20 cm.), (0ºK, 1ºK).

Las escalas de intervalos y las proporcionales corresponden a datos de tipo cuantitativo y por tanto admiten cálculos aritméticos.

Las variables cuantitativas pueden ser *discretas* o *continuas*; generalmente los valores de las primeras se obtienen como resultado de un *conteo* (cantidad de alumnos, cantidad de células), mientras que los valores de las segundas se obtienen mediante una *medición*.

En ocasiones para facilitar el análisis estadístico de conjuntos de datos se recurre a la transformación de los mismos aplicando métodos de **discretización** y **numerización**. El primero se utiliza cuando se necesita convertir un dato numérico en ordinal. Puede realizarse cuando:

- El error en la precisión de la medición realizada sobre un conjunto de datos es grande.
- Existen ciertos umbrales significativos.
- Se desea hacer comparables datos con medidas diferentes.
- Se desea hacer un análisis conjunto con un grupo de variables cualitativas mediante una técnica que solo admita variables discretas.

El segundo se utiliza cuando se desea aplicar procedimientos estadísticos que no admiten valores cualitativos. En este caso se forman variables a partir de datos que establecen clases. Se aplica a datos nominales y datos ordinales:

Datos nominales: Creación de variables numéricas dicotómicas.

Datos ordinales: Si una variable ordinal se encuentra en estado Likert [1] se le otorga una numeración ¨1 a 1¨. Esta conversión es muy utilizada cuando se trabaja con encuestas.

En este epígrafe se definen el tipo de datos y variables a utilizar en el análisis. En el caso de las notas de los estudiantes, la escala ordinal es la establecida para el sistema de calificaciones académicas utilizado en la Educación Superior cubana. En el caso de las encuestas las respuestas se encuentran clasificadas en escala ordinal, según el tipo de variables que puede ser discreta o continua.

[1] También denominado **método de evaluaciones sumarias.** Sus valores cumplen las condiciones de una variable proporcional (valores espaciados y un cero absoluto).

Teniendo en cuenta que las variables definidas son discretas y están en escala ordinal, se decide transformarlas a numéricas. Para eliminar las diferencias de escala entre ellas, se aplicarán transformaciones de estandarización.

1.4. Análisis de componentes principales

El análisis de componentes principales es una técnica estadística de síntesis de la información, o reducción de la dimensión de los datos a través de la condensación del número de variables. Es decir, ante una muestra de datos con muchas variables, el objetivo será reducirlas a un menor número perdiendo la menor cantidad de información posible (Rodríguez 2009) .

Este método solo se puede realizar si existen variables correlacionadas. Es útil para combinar la información brindada por un gran número de mediciones en unos pocos indicadores (componentes principales) que los resuman. Los indicadores formados son independientes entre sí y podrán ser determinadas a través de las siguientes condiciones:

- Los que tengan valores propios mayores que 1, porque son capaces de explicar al menos una variable.
- El conjunto que represente al menos un 60% de la varianza total, si es una aplicación de Ciencias Sociales, o hasta el 90%, si es una aplicación de Ciencias Naturales.

Para la aplicación del análisis de componentes principales se considera que el tamaño de la muestra no debe ser inferior a 50 observaciones, y preferiblemente debería ser 100 ó más grande.

La muestra de datos recolectada en esta investigación es de 122 observaciones (estudiantes de quinto año de la Facultad 3, centro CEIGE). Las variables que serán analizadas son: 37 pertenecientes a los resultados académicos y 30 a los resultados de encuestas del CICE, para un total de 67.

1.4.1. Matriz de componentes principales

Sean:

n: Cantidad de variables

m: Cantidad de datos de cada variable

$\mathbf{M}_{mxn}$: Matriz de los datos (simétrica)

$\mathbf{C}_{nxn}$: Matriz de correlaciones de las variables

Se debe tener en cuenta que la matriz de correlaciones de los datos ($\mathbf{C}_{nxn}$) tiene como máximo n pares de autovectores ($\mathbf{x}_{1xn}$) y autovalores (λ) linealmente independientes,

$\mathbf{C}\ \mathbf{x} = \lambda\ \mathbf{x}$;

$\mathbf{x}_{1xn}$: autovector

$\mathbf{X}_{nxn}$: matriz de autovectores

a partir de la matriz de autovectores $\mathbf{X_{nxn}}$ y de la matriz de los datos ($\mathbf{M}_{mxn}$), se obtiene $\mathbf{S}_{mxn}$, la matriz de las *componentes principales*:

M X = **S**,

$\mathbf{S}_{mxn}$: matriz de las *componentes principales,*

donde los vectores componentes se organizan de tal forma que $\mathbf{X_1}$ es el vector formado por un conjunto de constantes que hacen que $\mathbf{S_1}$, *primera componente principal*, tenga la máxima varianza que puede tener una combinación lineal de los vectores de los datos.

1.5. Análisis de *cluster*

El análisis de conglomerados o *cluster* consiste en la agrupación de objetos o individuos de una población en conjuntos homogéneos. Las unidades pertenecientes a cada uno de los grupos que se formen serán lo más parecidas entre sí, aunque muy diferentes respecto a las de otros grupos. Por tanto, si la clasificación es correcta, los objetos dentro de los conglomerados estarán muy próximos cuando se representen gráficamente, y los diferentes grupos estarán muy alejados (Hair, Anderson 1999).

Esta técnica de análisis exploratorio permite descubrir asociaciones y estructuras en los datos que no son evidentes a priori, pero pueden ser útiles una vez que se han encontrado. Consta de varias etapas de trabajo:

1. Preparación de los datos.
2. Elección de las variables.
3. Elección de la medida de distancia.
4. Elección y aplicación de la técnica de *cluster*.
5. Validación de los resultados.
6. Interpretación de los resultados.

1.5.1. Preparación de los datos

Los datos contenidos en la muestra deben encontrarse en óptimas condiciones. Para ello sería conveniente realizar una serie de transformaciones para evitar inconsistencias o ruido. También una correcta selección e integración de los mismos en caso de que provengan de distintas fuentes (bases de datos, archivos, bibliotecas, etc.).

1.5.2. Elección de las variables

El punto de partida para la clasificación de individuos de una muestra radica en la adecuada selección de las variables. Esto garantiza que los valores que tomen las mismas representen de forma acertada las características de cada individuo que lo diferencian del resto.

1.5.3. Elección de la medida de distancia

Para medir lo similares o disimilares que son los individuos existen diversos índices de similaridad y de disimilaridad. A continuación se explican alguna de estas medidas teniendo en cuenta los criterios de (Anderberg 1973),(Xui, Wunsch 2009).

- ✓ **Medidas de distancia:**

Las funciones de distancia son diferentes en dependencia de la escala de medición de los datos.

Distancias para variables numéricas:

Sea la tabla de medidas:

	1		*k*	*…*	*m*	*…*	*p*
1							
…							
i			x_{ik}		x_{im}		
…							
j			x_{jk}		x_{jm}		
…							
n							
			$x_{.k}$		$x_{.m}$		

Tabla 1. Tabla de medidas.

Distancia Euclídea: Dependiente de la escala y la magnitud.

$$d_{ij}^2 = \sum_{k=1}^{p} \left(X_{ij} + X_{jk}\right)^2$$

Propiedades

d(i,j) ≥ 0

d(i,i) = 0

d(i,j) = d(j,i)

d(i,j) ≤ d(i,k) + d(k,j) (Desigualdad *triangular*)

Distancia de Manhattan: Disminuye el efecto de los puntos atípicos.

$$d_{ij} = \sum_{k=1}^{p} \left|X_{ij} + X_{jk}\right|$$

Distancia Euclídea ponderada: Se utiliza para resaltar la importancia de ciertas variables.

$$d_{ij}^2 = \sum_{k=1}^{p} w_k \left(X_{ij} + X_{jk}\right)^2$$

$w_{k>0}$ (Variable ponderada cuya sumatoria debe ser 1)

Distancia de Minkowski: Es una generalización de la distancia Euclídea y Manhattan.

$$d_{ij} = (\sum_{i=1}^{p} (x_{i-} X_j)^q)^{1/q}$$

Para q=1 Manhattan; q=2 Euclídea

Distancia para variables ordinales:

El cálculo de la distancia se precede de la siguiente transformación:

a) Los datos de la variable son ordenados y reemplazados por su número de orden.

$r_{if} \in \{f1; \ldots; Mfg\}$, donde Mf es el índice del valor más alto de la variable.

b) Se forma el siguiente cociente, para obtener un resultado entre 0 y 1, dándole el mismo peso a todas las variables.

$$Z_{if} = \frac{r_{if} - 1}{M_{f-} 1}$$

c) Se aplica a los valores de $\boldsymbol{Z_{if}}$(z-score) cualquiera de las anteriores medidas de distancia para variables numéricas.

Distancia cuando se requiere eliminar el efecto de las diferencias de escala o magnitud:

Distancia euclidiana reducida o Distancia de Pearson:

$$d_{ij}^2 = \sum_{k=1}^{p} \left(X_{ij} - X_{jk}\right)^2 / S_k^2$$

S^2_k: varianza de la variable k.

Distancia cuando se requiere considerar la correlación entre variables:

Distancia de Mahalanobis:

$$d_{(x,y)} = \sqrt{\sqrt{(x-y)^T} S^{-1} (x-y)}$$

S: matriz de covarianza.

Distancia para variables nominales:

Distancia X^2

$$d_{ij}^2 = \sum_{k=1}^{p} (N/n_{.k}) \left((n_{ik}/n_{i.}) - \left(n_{jk}/n_j\right)\right)^2$$

1.5.4. Elección de la técnica de *cluster*

Los métodos de *cluster* pueden ser clasificados en:

a) Jerárquico.
b) No jerárquico o Particional.
c) Basados en densidad.
d) Basados en modelos (*cluster* conceptual).
e) Basados en teoría de grafos.
f) Basados en búsqueda combinatoria.
g) Basados en técnicas Fuzzy.

A continuación se realizará una descripción más detallada del análisis de *cluster* jerárquico y no jerárquico, ya que son los que se emplearán para llegar a la solución.

Cluster jerárquico

En los métodos jerárquicos los elementos de la muestra no se particionan en *clusters* de una vez, sino que requieren de particiones iterativas y jerárquicas de acuerdo al tipo de método seleccionado (aglomerativo o divisivo). Son empleados para datos numéricos.

En la clasificación jerárquica **aglomerativa** se parte de cada individuo como un grupo individual y paulatinamente se va agrupando con otros, de acuerdo a la similitud existente entre ellos (Fig. 2).

Según (Anderberg 1973) algunos de los métodos más conocidos son:

- Enlace simple (Vecino más cercano): Los grupos se unen considerando la distancia mínima que existe entre ellos.
- Enlace completo (Vecino más lejano): Los grupos se unen considerando la distancia máxima que existe entre ellos.
- Enlace Promedio: Los grupos se unen considerando la distancia media que existe entre ellos.
- Método de Ward: En cada iteración se unen los grupos (o elementos) que dan lugar a un menor incremento de la SCE (Suma de cuadrados error), es decir, la suma de cuadrados de la distancia intra-conglomerados:

$$SCE = \sum_{j=1}^{k} \left(\sum_{i=1}^{n_j} X_{ij}^2 - \frac{1}{n_j} \left(\sum_{i=1}^{n_j} X_{ij} \right)^2 \right)$$

Algoritmos basados en estos métodos:

- BIRCH (*Balanced Iterative Reducing and Clustering using Hierarchies*).
- CURE (*Clustering Using Representatives*).
- ROCK (*Robust Clustering using links*).

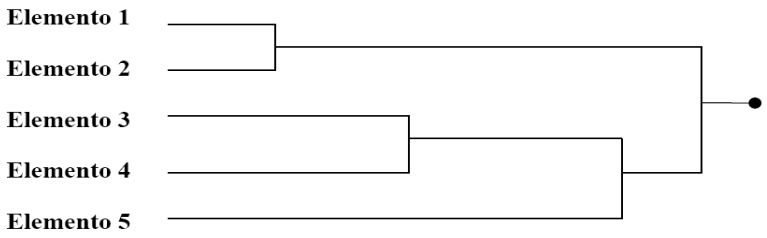


Fig. 2. Representación de un *cluster* jerárquico aglomerativo.

En la clasificación jerárquica **divisiva** se parte de la agrupación de todos los individuos en un solo grupo. A través de iteraciones sucesivas este se va particionado en conglomerados independientes, pero con cierto grado de homogeneidad (Fig. 3).

Algoritmos basados en este método:

- DIANA (*Divisive Analysis*)
- MONA (*Monothetic Analysis*)

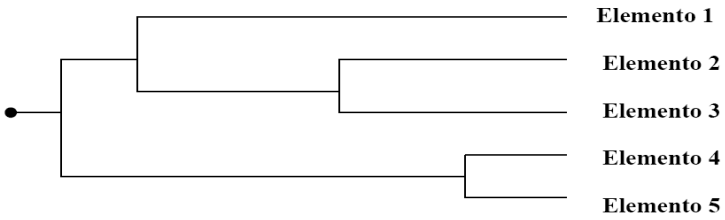


Fig. 3. Representación de un *cluster* jerárquico divisivo.

Los métodos jerárquicos permiten visualizar los resultados en un árbol de clasificación o dendrograma. Los conglomerados se representan mediante trazos verticales y las etapas de la fusión mediante trazos horizontales. Por ejemplo, en la Fig. 4 se muestran los distintos conglomerados que se forman y las distancias que hay entre ellos. Se puede comprobar que la escala estandarizada del dendrograma refleja con efectividad la proporcionalidad existente entre las distancias originales.

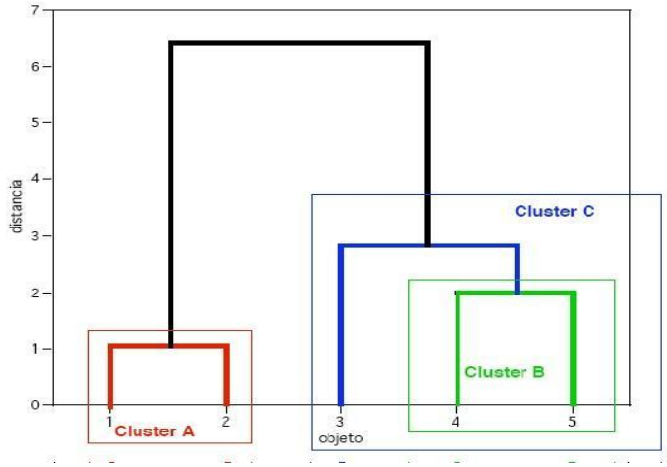


Fig. 4. Árbol de clasificación o dendrograma.

Cluster no jerárquico

Los métodos de análisis de *cluster* no jerárquicos asumen un conocimiento a priori del número de grupos en que debe ser dividido el conjunto de datos.

El método que habitualmente se usa es el K-Medias y permite asignar a cada observación el *cluster* que se encuentra más próximo en términos del centroide (media). En general, la distancia empleada es la Euclídea.

Pasos del método K-Medias:

1. Se predefinen **K** *clusters* iniciales y se determinan sus centroides.
2. Se calcula las distancias de cada punto a los centroides de los *clusters* y se reasignan a los grupos que estén más próximos. Luego se vuelven a recalcular los centroides de los **K** *clusters*.
3. Se repite el paso anterior hasta que no se produzca ninguna reasignación, es decir, hasta que los elementos se estabilicen en algún grupo.

En la práctica, se observan la mayor parte de reasignaciones en las primeras iteraciones.

Algoritmos basados en métodos no jerárquicos:

- K-Mediana
- CLARA *(Clustering Large Applications)*
- CLARANS (*Clustering Large Applications based on Randomized Search*)
- SOM (*Self-organizing Maps*)
- PAM (*Partition around medoids*)

1.5.5. Validación de los resultados

Las índices o criterios de evaluación son de vital importancia para proporcionar a los usuarios un grado de confianza en el resultado de la agrupación. Tienen dos propósitos fundamentales: determinar el número de conglomerados, o encontrar la mejor partición correspondiente.

Según Gordon (Gordon, Hayashi, N. Ohsumi 1998) existen tres categorías de criterios de prueba:

- Índices externos.
- Índices internos.
- Índices relativos.

Índices externos

Si P es una partición especificada de un conjunto de datos X con N puntos y es independiente de la estructura de *cluster* C, resultante de un algoritmo *cluster*, entonces la evaluación de C por criterio externo se obtiene comparando C con P. Entre los principales

índices que se aplican a estos casos está el índice de Rand, Jaccard, Fowlkes y Mallows, Γ de Hubert y Arabie.

Un valor de cualquiera de estos algoritmos cercano a 1 indica un buen agrupamiento (Xui, Wunsch 2009).

Índices internos

Estadísticas basadas en las sumas de cuadrados entre *clusters* y dentro de *clusters*. El número de conglomerados K es aquel que maximiza o minimiza uno de estos índices. Entre los principales están el índice de Dunn, el Índice de Davies-Bouldin.

- ✓ Índice de Dunn (1974): La idea es identificar los *clusters* que están bien compactos y bien separados de los demás. Dada una partición de *clusters* donde c_i representa el i-ésimo *cluster* de la partición, se define el índice de Dunn por:

$$D = \min_{1\leq i\leq n}\left\{\min_{1\leq j\neq n}\left\{\frac{d(c_i,c_j)}{\max_{1\leq k\leq n}(d(c_k))}\right\}\right\},$$

donde $d(c_i\ ,\ c_j)$ es la distancia entre los *clusters* c_i, y c_j y $d(c_k)$representa la distancia intracluster del *cluster* c_k.

- ✓ Índice de Davies-Bouldin (Das, Abraham, Konar 2009):

$$DB(k) = \frac{1}{k}\sum_{i=1}^{k} R_{i,rt}\ \ ,$$

donde $R_{i,rt} = max_{j\in k, j\neq 1} = \left\{\frac{S_{i,r}+S_{j,r}}{d_{ij,t}}\right\}$

El menor valor de *DB* (k) indica una partición óptima.

Índices relativos

Criterios externos e internos requieren de pruebas estadísticas que pueden convertirse en intensos cálculos computacionales. El criterio relativo elimina tales requerimientos y se concentra en la comparación de resultados de *cluster* generados por diferentes algoritmos o por el mismo, pero con diferentes parámetros de entrada.

Para algoritmos de *cluster* jerárquico, un punto de corte sobre un dendrograma puede mostrar el nivel adecuado al que se forman conjuntos con un elevado nivel de homogeneidad.

En el caso de los algoritmos no jerárquicos, K (número de *clusters* que se desea obtener) debe ser previamente especificado. En algunos casos, K puede ser estimado en concordancia con la experiencia del usuario, o de acuerdo a la información brindada a priori por otras aplicaciones.

Coeficiente cofenético

El coeficiente de correlación cofenético puede ser usado para medir cuán bien la estructura jerárquica del dendrograma representa a las verdaderas distancias. Se define como la correlación entre las $n(n - 1)/2$ pares de disimilaridades y las distancias cofenéticas del dendrograma.

1.6. Herramientas para el análisis de datos

Se hace necesario utilizar una herramienta que facilite la aplicación de técnicas de análisis multivariado como soporte a la toma de decisión en la UCI. A continuación se analizan las que se consideraron más potenciales de acuerdo a sus características:

- **WEKA**

WEKA, (*Waikato Environment for Knowledge Analysis)* es un entorno de trabajo desarrollado por la universidad de Waikato en Australia, para el análisis de datos. Se distribuye como software libre desarrollado en Java y está constituido por una serie de paquetes de código abierto que permiten aplicar, analizar y evaluar las técnicas más relevantes de análisis de datos, principalmente las provenientes del aprendizaje automático.

La función **Cluster** del **Explorer** permite aplicar algoritmos de agrupamiento de instancias como: **K-medias, EM, COBWEB, Sib** (*Sequential information bottleneck*)**, Farthest-rst** (variante mejorada de K-means)**, MakeDensityBasedClusted, Optics.**

Desventajas:

- Soporta un único formato específico conocido como ARFF (Attribute-Relation File Format), o cualquiera de estos tres soportes: fichero de texto, acceso a una base de datos y acceso a través de internet sobre una dirección URL de un servidor web.
- No permite mostrar resultados en 3 dimensiones.

- **RapidMiner**

RapidMiner (inicialmente YALE) es una herramienta de análisis y minería de datos, desarrollada en la plataforma Java, y distribuida bajo licencia AGLP. Su entorno gráfico permite el uso de más de 400 operadores que llevan a cabo las operaciones relacionadas con el procesamiento, análisis, limpieza, filtrado, e imputación de datos, detección de valores atípicos, conversiones de datos, operaciones con matrices, entre otras. Importa datos desde una gran cantidad de formatos (*.cvs,*.xls,*.xml,*.ARFF,*.XRFF, SPSS, Stata, Dbase, Bibtex, URL.).

RapidMiner cuenta con algoritmos para el análisis de conglomerados como: **K-Medias, K-Medias (Kernel), K-Mediana, DBSCAN, EM, SV** (*Support Vector Machine Clustering*), **Aglomerativo** (Enlace Simple, Enlace Completo., Promedio entre grupos), **Top Down Clustering**, **Flatten Clustering**.

Es extensible a otros programas y dispone de un módulo de integración con R. También puede importar algoritmos desde WEKA.

Desventajas:

- No permite mostrar resultados en 3 dimensiones.
- No permite la programación de funciones específicas.

➢ **"R"**

R es un lenguaje interpretado (scripting) y un conjunto de librerías creado por John Chambers en los Laboratorios Bell. Provee al usuario gran variedad de técnicas estadísticas (modelos lineales y no lineales, test estadísticos, análisis de series de tiempo, *cluster*, etc.) y gráficas. Puede ser ejecutado y compilado en plataformas como UNIX, Windows y MacOS y se distribuye sin costo y bajo licencia GPL. Permite la construcción de paquetes que favorecen la reutilización y combinación de los componentes creados por otros usuarios y los predefinidos en el kernel base-R.

Esta herramienta cuenta con una serie de algoritmos de análisis de *cluster* como: **AGNES, MONA y DIANA**, de tipo jerárquico, y **PAM**, **CLARA**, **SOM y FANNY**, no jerárquico. También permite la combinación de diferentes distancias y métodos.

Existen paquetes que permiten conectar R a distintos SGBD, por ejemplo: RPgSQL (para PostgreSQL), ROracle (para Oracle), RMySQL (para MySQL) y RODBC (para cualquier origen de datos ODBC).

También se cuenta con el lenguaje de programación procedural **pl/R** que permite utilizar las funciones y servicios de PostrgeSQL, integradas a la potencia de cálculo y salidas gráfica del paquete estadístico R.

Selección de la herramienta

En este epígrafe se realizó el estudio de algunas herramientas utilizadas para el análisis estadístico de datos. Se detectó que WEKA presenta algunos inconvenientes con respecto a la conectividad a los SGBD, al soporte de formatos de datos diferentes, a la programación de funciones y a la generación de gráficos en 3 dimensiones.

RapidMiner, a pesar de contar con un entorno de desarrollo más amigable y soportar mayor cantidad de formatos, no permite la programación de funciones específicas, sino su importación desde otras herramientas.

Sin embargo, R presenta un grupo de características que lo convierten en la mejor elección. Estas se describen a continuación:

- Multiplataforma.
- Alta conectividad con SGBD.
- Capacidad de manipular y modificar datos y funciones estadísticas.

- Capacidad de crear nuevas funciones estadísticas.
- Gráficos de mejor calidad y precisión.
- Capacidad de crear paquetes de funciones con propósitos bien definidos.
- Es similar a los software Matlab y Octave. Su sintaxis recuerda a los lenguajes C/C++.

Teniendo en cuenta lo antes expuesto se decide seleccionar a **R** como herramienta para el análisis estadístico de datos, ya que permitirá la creación de un paquete de funciones que integrarán las técnicas de análisis de componentes principales y análisis de *cluster*.

1.7. Conclusiones del capítulo

Se concluye que a partir de la importancia que tiene la selección de personal basada en competencias profesionales, se hizo necesario determinar como técnica de análisis: el análisis multivariado.

Teniendo en cuenta la naturaleza de los datos que se utilizarán en el presente trabajo, se escogieron como métodos estadísticos: el análisis de componentes principales y el análisis de *cluster*, los cuales se implementarán en un paquete de funciones usando como lenguaje "**R**".

Capítulo 2: Paquete de funciones para el análisis estadístico de datos

El objetivo de este capítulo es desarrollar las funciones que conforman el paquete para el análisis estadístico de datos desarrollado en R, basadas en el procedimiento estadístico planteado en (Medrano et al. 2010 y Martínez et al. 2012).

2.1. Procedimiento para el análisis de los datos

En este epígrafe se realiza un análisis del procedimiento estadístico empleado por los autores Medrano et al. 2010 y Martínez et al. 2012, para el tratamiento de los datos pertenecientes a estudiantes de la UCI.

El objetivo principal de este procedimiento es establecer una secuencia lógica para la aplicación de las técnicas de análisis multivariado que permitan la generación de información. Esta será usada como soporte a la toma de decisiones, en el proceso de selección de personal que integra un proyecto de desarrollo de software (Fig. 5).

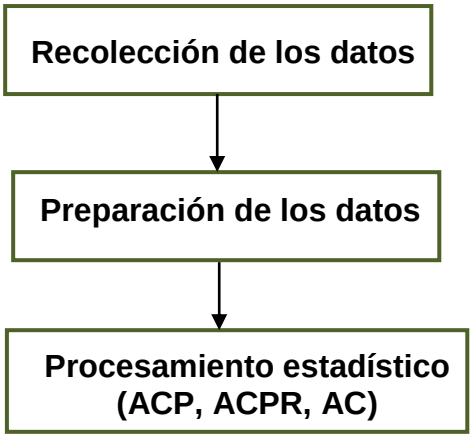


Fig. 5. Procedimiento de Medrano y otros autores.

A continuación se describen los pasos del procedimiento:

1. **Recolección de datos:**

En este caso se utilizan los datos correspondientes a las notas finales de los estudiantes hasta segundo año. Como el principal objetivo es contar con información sobre los estudiantes que van a ingresar a los proyectos de desarrollo de software, se sugiere que las asignaturas seleccionadas sean las que aporten más a la formación técnica del estudiante. Se deben eliminar del análisis las calificaciones de la asignatura Educación Física, por no tributar directamente a la formación de competencias técnicas.

2. **Preparación de los datos:**

Para procesar los datos se debe tener en cuenta el tipo de datos en que vienen expresadas las variables (las notas del sistema educacional cubano se expresan en una escala ordinal,

sin intervalo de magnitud definida e igual entre cualquier par de niveles consecutivos de la variable (Rodríguez 2009), por tanto, se deben transformar a numérico).

3. Procesamiento estadístico:

Para el procesamiento estadístico de los datos se propone el uso de los siguientes métodos:

Reducción

- ✓ **Análisis de componentes principales (ACP**):

Permite la exploración y reducción de la dimensión de los datos. Ayuda a los investigadores a adquirir cierta percepción respecto a un conjunto de datos, y a encontrar la verdadera dimensión de estos. A través del ACP se puede seleccionar un número menor de variables que son combinaciones lineales de las anteriores; permite sintetizar grandes cantidades de información y una mejor comprensión de la misma. Determina una forma de resumir los resultados académicos de un alumno, con la mínima pérdida de información, así como las relaciones hipotéticas entre variables y estructuras latentes.

- ✓ **Análisis de Componentes principales Robusto (ACPR):**

Al realizar el ACP se debe tener mucho cuidado con los valores atípicos, ya que estos son capaces de incrementar artificialmente la varianza, haciendo que las componentes principales sean atraídas por ellos. El ACPR determina las direcciones estimando la varianza por una manera robusta en lugar de la varianza clásica.

Los valores atípicos se dividen en dos:

- Puntos de palanca (*leverage*): Valores alejados de la nube de puntos en el plano vectorial, con una distancia ortogonal pequeña (buenos) o grande (malos).
- Ortogonales: Valores que se encuentran dentro de la nube de puntos pero proyectados en dirección perpendicular al eje de coordenadas.

Después de ser localizados, pueden ser retirados de la muestra de análisis. Este paso no siempre es recomendable, ya que en ocasiones pueden afectar el resultado final.

Clasificación, Agrupación

- ✓ **Análisis de *cluster* (AC):**

Procedimiento orientado al descubrimiento de asociaciones en los datos y taxonomías de carácter homogéneo. La inclusión de este método en el procedimiento tiene como objetivo la formación de tres grupos de estudiantes, cuya existencia se asume basada en la experiencia de expertos en la labor docente y productiva en la universidad:

- Aptos: Podrán integrarse directamente a un proyecto de desarrollo y comenzar a desempeñarse en un rol afín.
- Aptos con limitaciones: Presentan limitaciones técnicas y por tanto deberán recibir capacitación básica para poder desempeñarse en un proyecto o tendrán responsabilidades de menor importancia.

- No aptos: No pueden formar parte inmediatamente de un proyecto de producción. Por lo que se trazarán estrategias para elevar su nivel, en aras de una futura incorporación.

2.2. Construcción de un paquete de funciones en R

En esta sección aparecen descritas las funciones que conforman el paquete desarrollado en R como soporte a la toma de decisiones en el proceso de incorporación de estudiantes a proyectos de software. Esta herramienta está basada en el procedimiento estadístico descrito en el epígrafe anterior (Fig. 5).

Descripción de las funciones del paquete

Las funciones del paquete que se propone como solución, pueden ser utilizadas de forma independiente. Algunas dependen de paquetes incluidos en el Kernel R-base y otras pueden ser descargadas de Internet.

En la siguiente tabla se muestra un resumen de las funciones:

Función	Descripción
mcorr	Correlación entre variables. Multicolinealidad.
pca	Análisis de componentes principales clásico.
pca	Análisis de componentes principales combinado con análisis de *cluster*.
pcacomb	Combina el ACP con AC.
outdetect	Detección de elementos atípicos.
clusterh	Análisis de *cluster* jerárquico.
clusternh	Análisis de *cluster* no jerárquico.
clustermb	Análisis de *cluster* basado en modelos.

Tabla 2. Resumen de las funciones del paquete.

- **Función mcorr()**

Una condición necesaria para la aplicación del método ACP es la existencia de relaciones lineales entre las variables predictoras en el modelo lineal, lo que indica que parte de la información en una o más variables es redundante. Bajo condiciones de colinealidad resulta imposible distinguir los valores individuales de cada variable predictora, debido a que la fuerza de correlación entre ellas produce relaciones lineales de similar magnitud entre los coeficientes (Jackson 1991).

La función **mcorr()** realiza el análisis de correlación. Recibe como parámetros de entrada el rango de columnas que será analizado, la matriz de datos y las etiquetas de las variables.

Entre los primeros pasos del algoritmo se encuentra la obtención de la matriz de correlación con sus valores propios. Luego se establece el número de condición (nc) a través de las siguientes líneas de código:

```
VP <- eigen(cor(D))$value
nc <- (VP[1]/VP[a[2]])
```

donde el número de condición equivale al mayor de los índices de condición y es igual a la división del mayor valor propio por el menor valor propio de la matriz.

En la Fig. 6 se muestra el funcionamiento de **mcorr()** a través de un diagrama de flujo, donde X_{ij}: matriz de datos original y MC_{ij}: matriz de correlación**.** Esta función va a devolver como resultado final el tipo de correlación que existe entre cada variable (fuerte, ligera, débil), y si existe multicolinealidad en la muestra de datos.

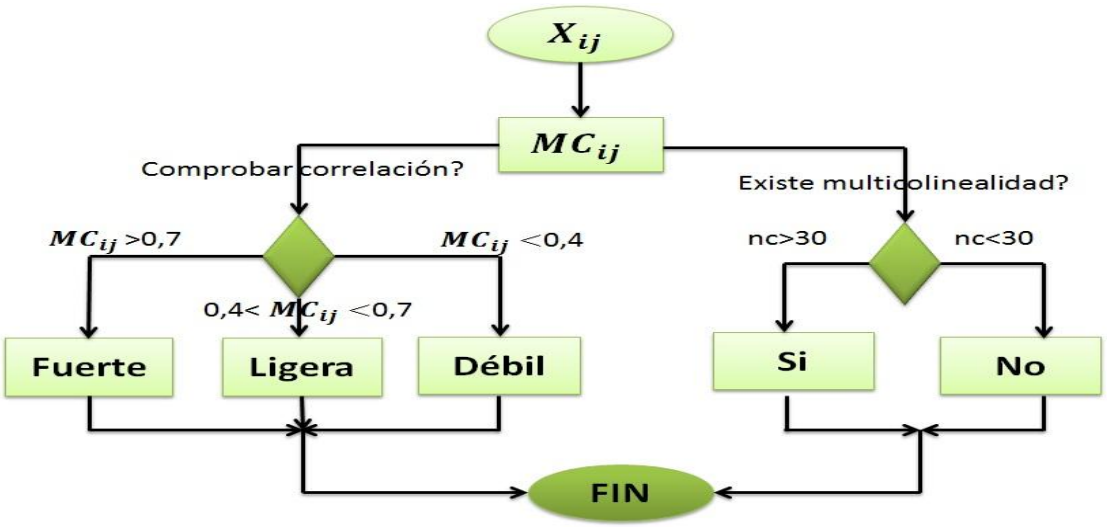


Fig. 6. Función mcorr().

- **Función pca()**

La función **pca()** implementa el ACP clásico. Recibe como parámetros de entrada el rango de columnas de la matriz de datos a analizar, la proporción de varianza acumulada, y las etiquetas de cada variable.

Después de realizar los procedimientos de estandarización o centrado de la matriz, se obtiene la matriz de carga de las componentes principales, la de *scores*, además de la matriz diagonal que contiene las desviaciones estándares de las componentes. Se calculan las proporciones de varianza y las proporciones de varianza acumulada de cada variable respecto a los componentes principales.

El siguiente código ejemplifica la forma en que se implementó este último paso en R:

```
propvar <- D^2/sum(D^2)
propacum <- numeric(length(D));propacum[1]<-propvar[1]
for (i in 2:length(D))(propacum[i] <- propvar[i] + propacum[i-1])
```

donde D es la matriz que contiene las desviaciones estándares.

Se selecciona la cantidad de componentes principales que se desean obtener de acuerdo al valor de varianza acumulada entrado como parámetro. Luego se determinan las variables que son más significativas con respecto a las componentes principales, a través del siguiente criterio:

```
for (j in 1:length(Pccp[1,]))
{for (i in 1:length(Pccp[,1]))
 {if ((abs(Pccp[i,j])) < max(abs(Pccp[,j]))/2)
  {Pccp[i,j] <- 0
  }}}
```

El algoritmo devuelve dos tablas: la primera con un resumen de la proporción de varianza y la varianza acumulada de cada componente principal y la segunda con los valores más significativos de carga de las variables con respecto a cada componente. También se obtiene el gráfico de segmentación, y los de carga de las 3 primeras componentes con sus respectivos gráficos de dispersión de individuos.

En la Fig. 7 se aprecia el funcionamiento de este algoritmo, donde MC_{ij}: matriz de carga de las componentes principales, MS_{ij}: matriz de scores, MD_{ij}: matriz diagonal, MR_{ij}: matriz resumen, CP: componentes principales, GD: gráfico de dispersión de individuos.

Fig. 7. Función pca().

- **Función pcacomb()**

Esta función combina las funciones **pca()** y **clusterh().** Devuelve los gráficos correspondientes a **pca(),** pero con la diferencia de que en el gráfico de dispersión de individuos se reflejan los conglomerados formados. También devuelve los vectores de

pertenencia de cada grupo etiquetados con los valores de la encuesta “Matriz de fortalezas y debilidades”.

- **Función outdetect()**

Esta función depende de la librería Chemometrics y recibe como entrada el rango de columnas de la matriz de datos, la proporción de varianza acumulada, y las etiquetas de las variables.

Su objetivo fundamental es detectar los valores atípicos con el cálculo de las distancias de puntaje (score) y las ortogonales. También generar los gráficos de diagnóstico mediante la función predefinida en R:

```
pcaDiagplot(Dmatriz, pca, a=CCP
```

Después de tipificar los valores atípicos se procede a la clasificación de los mismos en P.L.G (bueno), P.L.G (malo), O.O (ortogonal), según las condiciones que se establecen.

En la Fig. 8 se puede apreciar el funcionamiento de este algoritmo, donde ACP: análisis de componentes principales, SD: distancia de scores, VCSD: valor crítico de las distancias de score, OD: distancia ortogonal, VCOD: valor crítico de las distancias ortogonales.

La salida que produce esta función es una lista con los identificadores de todos los individuos de la muestra, el tipo de atípico, en caso de haberlo, y los gráficos de diagnóstico.

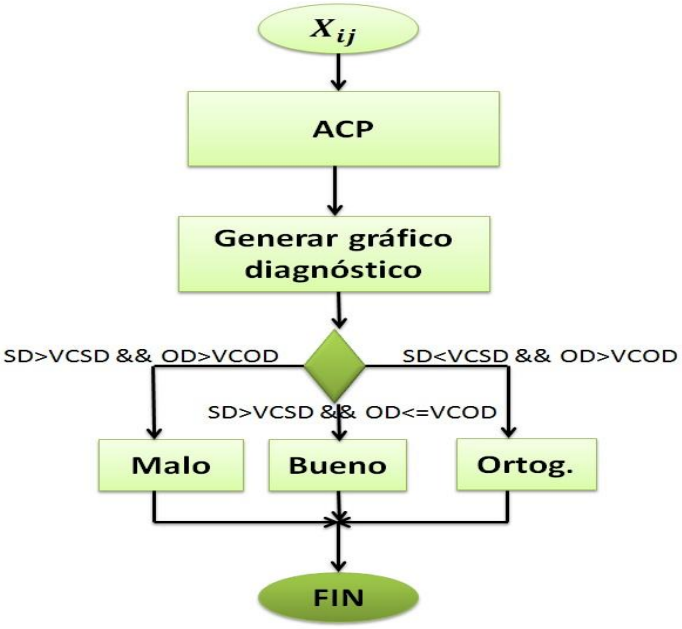


Fig. 8. Función outdetect().

- **Función clusterh()**

Esta función implementa los siguientes métodos de agrupamiento jerárquico (aglomerativo):

- Enlace Simple.
- Enlace Completo.

- Enlace Promedio.
- Enlace Promedio Mediana.
- Enlace Centroide.
- Enlace de Ward.
- Enlace de MacQuitty.

Las distancias que se emplean son:

- Distancia Euclídea
- Distancia Manhattan.
- Distancia del Máximo.
- Distancia de Canberra.
- Distancia de Minkoswki.

Devuelve los dendrogramas de todos los métodos, sus coeficientes cofenéticos y vectores de pertenencia. También permite etiquetar los elementos de cada conglomerado con variables de relevancia para la toma de decisiones en los procesos de gestión de recursos humanos (Test "Matriz de fortalezas y debilidades" y otros).

- **Función clusternh()**

Esta función implementa el método no jerárquico K-medias. La distancia utilizada es la Euclídea. Devuelve una tabla con el resumen de la cantidad de elementos por grupo.

- **Función clustermb()**

Esta función implementa un método de *cluster* basado en modelos. Toma elementos del algoritmo EM y de Criterios de Información Bayesiana. Devuelve una tabla con el resumen de la cantidad de elementos por grupo.

2.3. Creación de paquetes en R

Los paquetes de R permiten compartir las innovaciones e información generada en este marco de trabajo con numerosos usuarios. Son extensiones, conjuntos de funciones y datos desarrollados de forma compacta y documentados con más rigor que la mayoría de las aplicaciones estándares. Debido a que R que se distribuye bajo licencias de software libre, los usuarios pueden contribuir tanto a las mejoras del código como a su socialización a través del CRAN.

Estructura de paquetes y archivos

Para crear paquetes en R se deben contar con un conjunto mínimo de utilidades: Rtools, Perl y el compilador de la ayuda HTML de Microsoft o cualquier editor de texto como LaTeX o Tinn-R.

Después de cargar en la consola de R las funciones con las que se quiere conformar el paquete y de especificar su ubicación en el ordenador, se utiliza la función:

```
package.skeleton(¨nombre_paquete¨)
```

Entonces se creará una estructura de carpetas y archivos en el lugar indicado con anterioridad.

Documentación de paquetes

La documentación de los paquetes es uno de los pasos más importantes de este proceso. Si la documentación es confusa o inexistente, entonces el paquete quedará completamente inutilizable. Por tanto, se recomienda prestar especial atención al llenado de las planillas donde se especifican las características del producto.

Para cada función del paquete, se debe crear un archivo de documentación (Rd) que contiene los puntos obligatorios 'name', 'alias', 'title','description', 'usage' en la cabecera y 'keyword' en el pié, además de los puntos opcionales como 'arguments', 'value', 'details', 'references', 'seealso', 'examples' en el cuerpo del archivo.

Revisión y construcción de paquetes

Antes de construir definitivamente el paquete se recomienda comprobar que sus funciones estén en perfecto estado usando: `R CMD check`. Luego se construye la ayuda del sistema en Html ejecutando las siguientes sentencias en la consola:

```
R CMD Rdconv –t=html –o=mifun.html mifun.Rd
```

Por último, se emplea la función `R CMD build miPaquete` que creará un archivo "tar". Este se puede someter al CRAN o distribuir entre los usuarios.

2.4. Conclusiones del capítulo

En este capítulo se implementaron la siguientes funciones: mcorr(), outdetec(), pca(), pcacomb(), clusterh(), clusternh(), clustermb(), las cuales conforman el paquete para el análisis estadístico de datos.

A partir de dichas funciones se podrán obtener los gráficos combinados de análisis de componentes principales y análisis de *cluster*, que incluyen los resultados de estos dos métodos. Además la función pcacomb() o clusterh() permitirá realizar el etiquetado de conglomerados con los resultados de las encuestas del CICE.

Capítulo 3: Caso de Estudio

En este capítulo se muestran los resultados de la aplicación de las técnicas de análisis de componentes principales y análisis de *cluster* a un caso de estudio, y de esta forma se valida el paquete de funciones para el análisis estadístico de datos desarrollado en R.

3.1. Descripción del caso de estudio

El caso de estudio está compuesto por 122 estudiantes del centro CEIGE (Facultad 3) que actualmente cursan el quinto año de la carrera de Ingeniería Informática. De ellos se conocen las notas correspondientes a las asignaturas del primer ciclo de formación profesional (Tabla 3).

Teniendo en cuenta que uno de los objetivos de esta investigación es la clasificación de estudiantes para su posterior inserción en proyectos de software, se deben priorizar aquellas asignaturas que inciden de forma relevante en la formación técnica o en la adquisición de habilidades. Es por esto que se decide retirar de la muestra la asignatura de Educación Física.

Primer Año	Segundo Año	Tercer Año
Álgebra Lineal **(AL)**	Física 1 **(F1)**	Administración de Empresas **(AE)**
Filosofía y Sociedad **(FS)**	**Física 2 (F2)**	Contabilidad y Finanzas **(CF)**
Idioma Inglés 1**(IE1)**	Idioma Inglés 3 (**IE3)**	Gráficos por Computadora **(GC)**
Idioma Inglés 2 **(IE2)**	Idioma Inglés 4 (**IE4)**	Ingeniería de Software 1 **(IS1)**
Introducción a la Programación **(IP)**	Matemática 3 **(M3)**	Ingeniería de Software 2 **(IS2)**
Matemática 1 **(M1)**	Matemática 4 **(M4)**	Programación 3 **(P3)**
Matemática 2 **(M2)**	Máquinas Computadoras 1 **(MC1)**	Programación 4 **(P4)**
Matemática Discreta **(MD)**	Máquinas Computadoras 2 **(MC2)**	Probabilidades y Estadística **(PE)**
Programación 1 **(P1)**	Programación 2 **(P2)**	Práctica Profesional 3 **(PP3)**
Panorama Histórico Cultural Cubano **(PHCC)**	Práctica Profesional 2 **(PP2)**	Problemas Sociales de la Ciencia y las Tecnologías **(PSCT)**
Panorama Histórico Cultural Universal **(PHCU)**	Sistemas de Base de Datos **(SBD)**	Sistemas Operativos **(SO)**
Práctica Profesional 1 **(PP1)**		Teleinformática 1 **(T1)**
Preparación para la Defensa **(PPD)**		Teleinformática 2 **(T2)**

Tabla 3. Distribución de asignaturas por semestre.

También se incluyen los datos provenientes de encuestas realizadas a los estudiantes por el CICE (Anexo 7). A continuación se explican las encuestas seleccionadas:

- **Matriz de fortalezas y debilidades**: Detecta las principales fortalezas y debilidades presentes en los estudiantes. Las variables recolectadas son discretas en escala ordinal: Capacidad de abstracción, análisis y síntesis (CAAS), Capacidad de aplicar los conocimientos en la práctica (CACP), Capacidad para organizar y planificar el tiempo (COPT), Responsabilidad social, compromiso ciudadano y ética de la profesión (RSCCEP), Capacidad de comunicación oral y escrita (CCOE), Capacidad de comunicación en un segundo idioma (CCSI), Capacidad de investigación (CI), Capacidad de aprender y actualizarse permanentemente (CAAP), Habilidades para buscar, procesar y analizar información procedente de diferentes fuentes (HBPAI), Capacidad crítica y autocrítica (CCA), Capacidad creativa (CC), Capacidad para identificar, plantear y resolver problemas (CIPRP), Capacidad para tomar decisiones (CTD), Capacidad de trabajo en equipo (CTE), Capacidad de motivar y conducir hacia metas comunes (CMCMC), Habilidad para trabajar en forma autónoma (HTA), Compromiso con la calidad (CCL).
- **Test "Autopercepción de Belbin":** Brinda conocimiento del rol o los roles que mejor pueden desempeñar los estudiantes en un equipo de trabajo. Identifica las aptitudes para el liderazgo. Las variables recolectadas son discretas en escala ordinal: Implementador (I), Cerebro (C), Cohesionador (CH), Impulsor (IS), Especialista (E), Investigador de recursos (IR), Coordinador (COR), Finalizador (F), Monitor-Evaluador (ME).
- **Test "Estilos decisionales":** Identifica los estilos predominantes en los estudiantes en la toma de decisiones. Las variables recolectadas son discretas en escala ordinal: Estilo directivo (ED), Estilo analítico (EA), Estilo conceptual (EC), Estilo conductual (ECD).

3.2. Análisis estadístico

En la muestra de datos se identificaron un total de 70 variables. De ellas, 3 corresponden a datos generales de los estudiantes (id, nombre y apellidos, grupo), 37 a las asignaturas que conforman el primer ciclo, y 30 a los resultados de las encuestas del CICE.

El análisis exploratorio que se realiza solamente considera las variables pertenecientes a las notas de las asignaturas. Las restantes serán empleadas para etiquetar los conglomerados que se forman.

3.3. Análisis de la matriz de correlación

Después que los datos son procesados y estandarizados se realiza el análisis de correlación entre variables. Según lo expuesto en capítulos anteriores, una de las condiciones

necesarias para la aplicación del ACP es la existencia de variables correlacionadas dentro de la muestra.

Para ello se emplea la función del paquete de análisis estadístico **mcorr()** y se obtiene como resultado la existencia de multicolinealidad entre las variables. También un resumen del comportamiento de las correlaciones: 1 correlación fuerte, 75 correlaciones ligeras y 590 correlaciones débiles.

Al analizar los coeficientes de correlación (Anexo 1), se detecta que las asignaturas IE3 e IE4 presentan valores elevados (0.72), lo cual indica que los resultados de los estudiantes en estas materias tienen comportamientos similares.

Las asignaturas de Idioma (IE1, IE2, IE3, IE4) correlaciona entre ellas de forma ligera, aunque también lo hacen con asignaturas como MC1, MC2, IP, MD. Las de Humanidades (PHCC y PHCU) correlacionan entre ellas, al igual que algunas pertenecientes a Ciencias Básicas (AL, MD, M1, M3, F1, entre otras), y la especialidad (P1, P3, P4, SBD, IS1, IS2), en un rango de 0,4 a 0,54.

En cambio, asignaturas como PSCT, PPD, PP1, PHCC, y PHCU presentan correlaciones débiles con otras variables, en un rango de 0,01 a 0,39.

De forma general, todas las variables seleccionadas tienen al menos un coeficiente de correlación con otra variable mayor que 0,30.

3.4. Detección de valores atípicos

Un *outlier* o valor atípico puede ser definido como un valor alejado del centro de la nube de puntos en el espacio vectorial. Es importante detectarlos a tiempo porque sus varianzas pueden incidir de forma negativa en los resultados de la investigación.

Con la ejecución de la función **outdetect()** se detecta que existen 6 valores atípicos en la muestras de datos. Los resultados se exponen en la siguiente tabla:

ID estudiante	Tipo de atípico
18	Ortogonal
35	Ortogonal
37	Ortogonal
47	Palanca bueno
105	Ortogonal
120	Ortogonal

Tabla 4. Clasificación de atípicos.

A continuación se muestran los gráficos de diagnóstico de componentes principales. En la Fig. 9 se advierte la presencia del atípico de tipo palanca (*leverage*) bueno, el cual se caracteriza por tener una pequeña distancia ortogonal, es decir, no tiene influencia significativa sobre el espacio de las componentes principales. En la Fig. 10 quedan

graficados los atípicos de tipo ortogonal. Se puede apreciar que sus distancias ortogonales no son significativas.

Lo antes expuesto demuestra que las observaciones que resultaron ser atípicas no serán influyentes para el ACP. No obstante, se analizan de forma independiente los estudiantes a los que corresponden las observaciones:

- El estudiante con ID número 18 presenta un índice académico de 4.1. Sus calificaciones en asignaturas como IP, P1, P2 y P3 están entre 5 y 4 puntos. No obstante, presenta nota de 3 puntos en PPD y CF (estas son materias en las que la mayoría de los estudiantes suelen alcanzar buenos resultados).
- El estudiante con ID número 35 presenta un índice académico de 3.51 y calificaciones de 3 puntos en asignaturas como IP, P1, P2, P3, P4 y SBD.
- El estudiante con ID número 37 presenta un índice académico de 3.81 y calificaciones diferentes en asignaturas complejas (4 en M1, M2, MD, 3 en MC1, MC2, 5 en F1 y F2).
- El estudiante con ID número 47 presenta un índice académico de 3.65. Obtiene calificaciones diferentes en asignaturas de la misma especialidad (4 en MC1 y MC2, 3 en IP, P1, P2, P3, P4) y calificaciones de 3 puntos en materias en las que usualmente los estudiantes suelen obtener buenos resultados (PPD y PHCC).
- El estudiante con ID número 105 presenta un índice académico de 3.91 y calificaciones diferentes en asignaturas complejas (5 en IP, 3 en P1, 4 en P2 y P3).
- El estudiante con ID número 120 presenta un índice académico de 3.97 y calificaciones diferentes en asignaturas complejas (3 en IP, P1, 4 en MC1, MC2, P2, 5 en GC).

Por lo antes expuesto, se toma la decisión de no retirar a estos estudiantes del análisis.

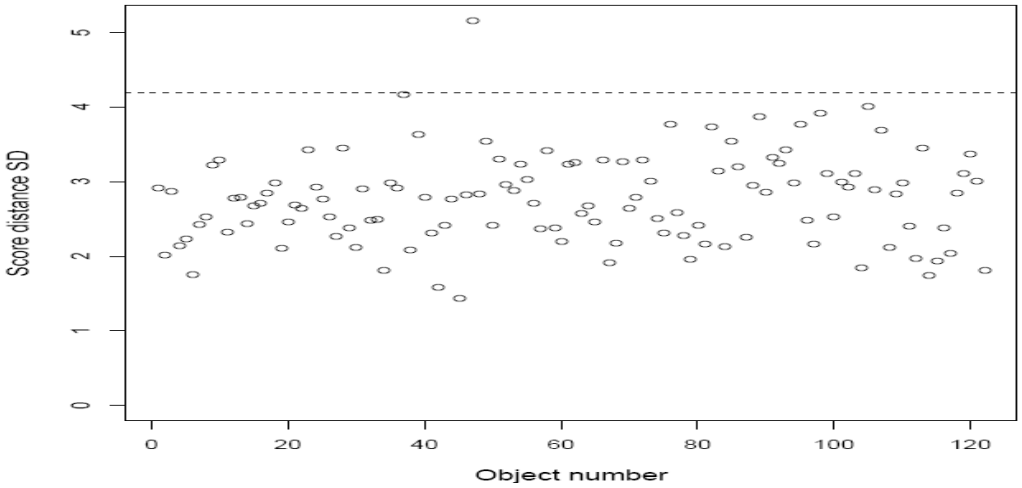


Fig. 9. Gráfico diagnóstico (a).

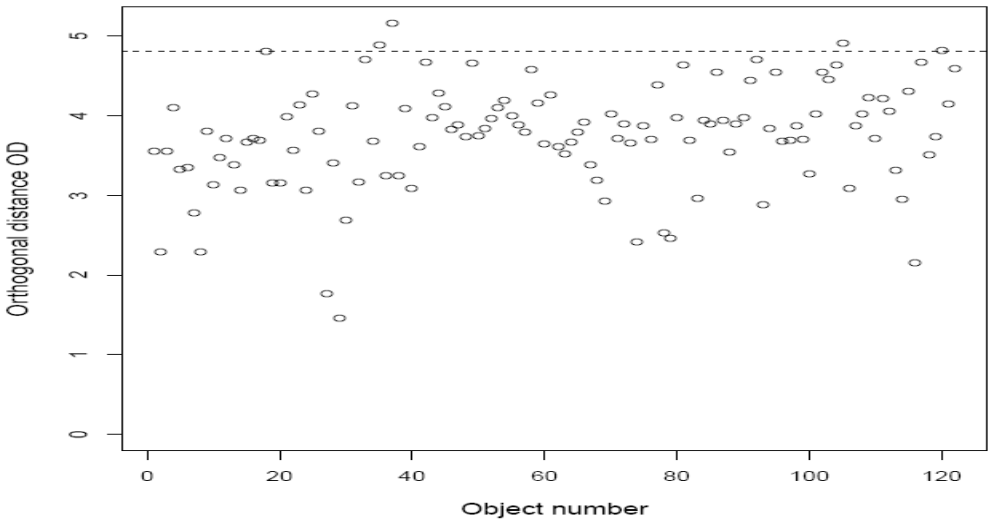


Fig. 10. Gráfico diagnóstico (b).

3.5. Análisis de componentes principales

Después de aplicar el análisis de correlación y la detección de atípicos se procede a realizar el ACP. En la Fig. 11 se muestra un resumen de los valores propios, proporción de varianza, y varianza acumulada de cada componente principal.

Se aprecia que la primera componente acumula una varianza de 0,279, lo cual se traduce en que ella sola explica casi un 30% de la variabilidad de los datos, mientras que el resto solo lo hace para un 5% como máximo.

En este caso se tomarán solo 8 componentes principales con valores propios mayores estrictos que 1, que resumen la variabilidad total de un 61% aproximadamente. Para realizar estudios de esta naturaleza (Ciencias Sociales), varios autores recomiendan trabajar con al menos un 60 % de la variabilidad acumulada.

"PCA Summary"

Comp,	[PC1]	[PC2]	[PC3]	[PC4]	[PC5]	[PC6]	[PC7]	[PC8]	[PC9]	[PC10]	[PC11]	[PC12]
VP	35.312	17.353	16.186	15.600	14.145	13.382	12.636	11.738	11.244	11.047	10.730	10.327
propvar	0.279	0.067	0.059	0.054	0.045	0.040	0.036	0.031	0.028	0.027	0.026	0.024
propacum	0.279	0.346	0.404	0.459	0.503	0.543	0.579	0.610	0.638	0.665	0.691	0.715

Comp.	[PC13]	[PC14]	[PC15]	[PC16]	[PC17]	[PC18]	[PC19]	[PC20]	[PC21]	[PC22]	[PC23]	[PC24]	[PC25]	[PC26]
VP	9.926	9.772	9.409	9.110	8.763	8.557	8.500	8.230	7.805	7.651	7.391	7.031	6.799	6.627
propvar	0.022	0.021	0.020	0.019	0.017	0.016	0.016	0.015	0.014	0.013	0.012	0.011	0.010	0.010
propacum	0.737	0.758	0.778	0.796	0.814	0.830	0.846	0.861	0.875	0.888	0.900	0.911	0.922	0.931

Comp.	[PC27]	[PC28]	[PC29]	[PC30]	[PC31]	[PC32]	[PC33]	[PC34]	[PC35]	[PC36]	[PC37]
VP	6.453	6.107	5.998	5.821	5.400	5.227	5.042	4.831	4.620	4.361	3.615
propvar	0.009	0.008	0.008	0.008	0.007	0.006	0.006	0.005	0.005	0.004	0.003
propacum	0.941	0.949	0.957	0.965	0.971	0.977	0.983	0.988	0.993	0.997	1.000

Fig. 11. Resumen de componentes principales.

A continuación se realiza el análisis gráfico de las cargas factoriales sobre los ejes de las dos primeras componentes pues son las que mayor variabilidad resumen (Fig. 12). El resto

se analizan mediante la tabla resumen del Anexo 2, donde se exponen los valores de carga de cada variable con respecto a las componentes seleccionadas.

Según el gráfico se identifica un significado para los ejes. Se puede apreciar que todas las asignaturas cargan en el mismo sentido (positivo respecto a la primera componente) y esto se debe a que todas requieren un grado de habilidades generales para su desarrollo.

Las observaciones que se ubican en esta región del gráfico de dispersión de las componentes 1 y 2 corresponden a estudiantes que por lo general tienen buenos resultados académicos.

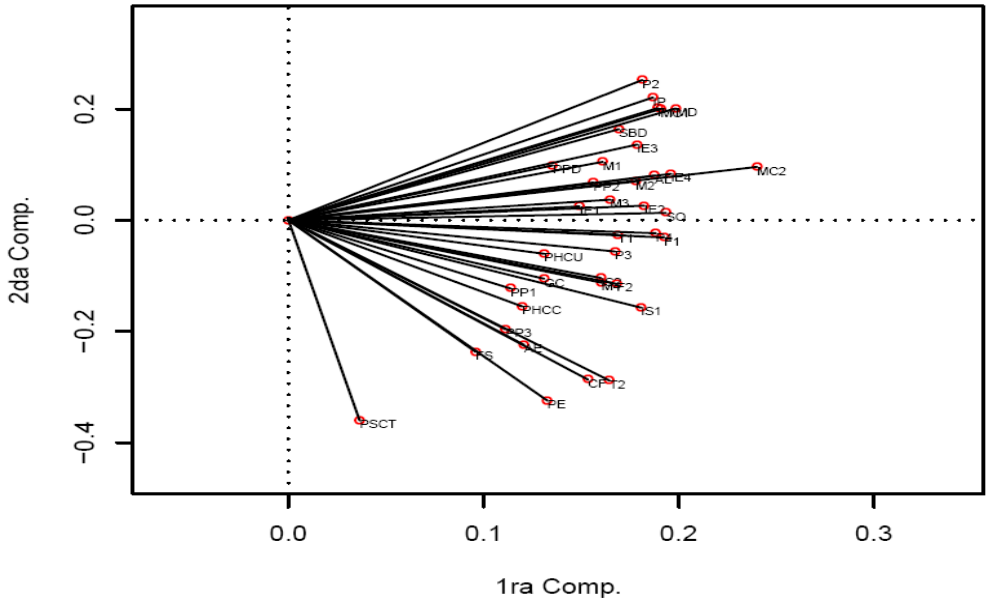


Fig. 12. Análisis de las primeras componentes principales.

Por ejemplo, al analizar la Fig. 13, se observa que el estudiante con identificador 116 está ubicado en el área donde las asignaturas cargan con fuerza sobre la primera componente y su promedio académico es de 4,89 puntos. Por tanto, la interpretación de esta gráfica de dispersión de individuos permite la conformación a priori de 3 grupos de estudiantes que pueden ser clasificados de acuerdo sus resultados académicos en aptos, no aptos y aptos con limitaciones.

Por otra parte, PSCT carga débilmente sobre la primera componente y se encuentra alejada del resto de las variables, lo cual podría significar que los resultados en esta materia tienden a mantenerse constantes en un rango de 4 a 5 puntos. En el caso de materias como MC2, MC1, MD, SO, IE4, P1, F1 los valores de carga son elevados. Esto significa que contribuyen a diferenciar a los estudiantes, ya que le introducen mayor variabilidad a los resultados.

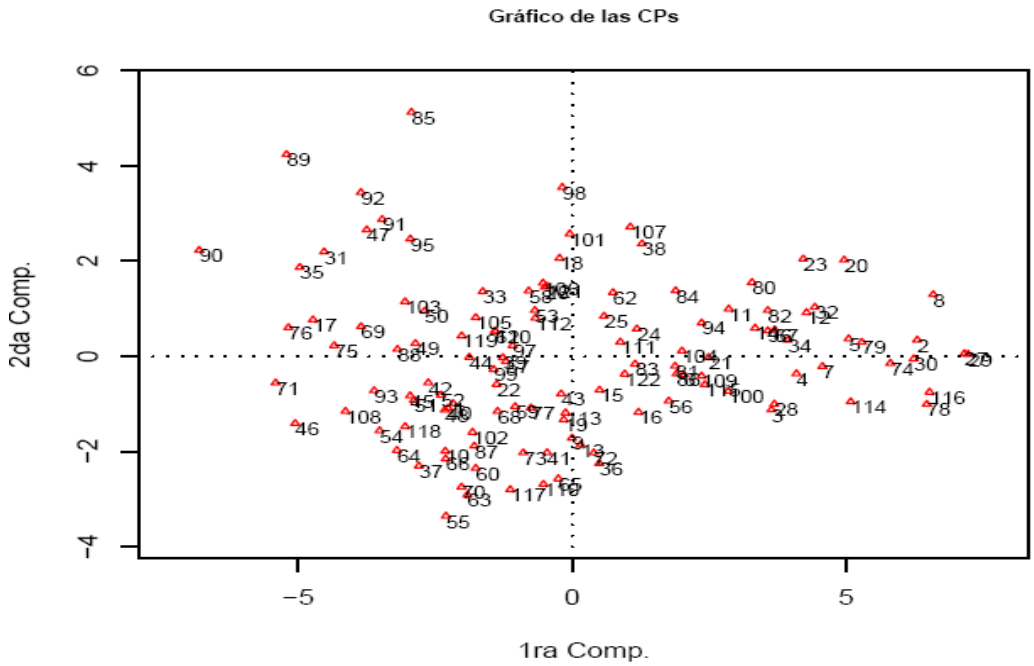


Fig. 13. Gráfico de dispersión de individuos.

3.6. Análisis de *cluster*

La Fig. 14 muestra el diagrama de segmentación que se obtiene tras aplicar el algoritmo K-Medias con la distancia Euclídea. En él se evidencia que la cantidad de conglomerados a seleccionar es 3, ya que la gráfica de codos se comienza a rectificar a partir de este punto. El empleo del diagrama de segmentación corrobora lo que plantean los autores (Medrano et al. 2010 y Martínez et al. 2012) sobre la conformación de 3 grupos de estudiantes teniendo en cuenta sus resultados académicos.

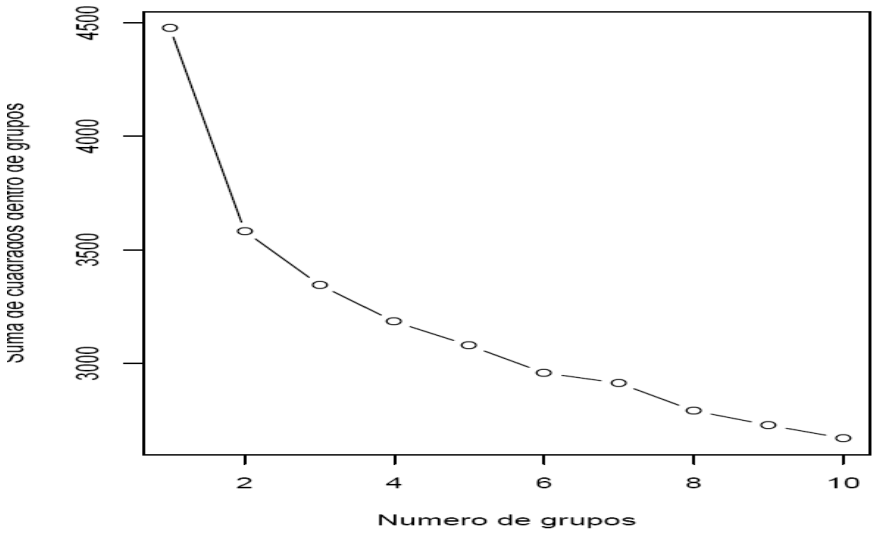


Fig. 14. Gráfico de segmentación.

Para formar los tres conglomerados de estudiantes se utilizaron varias combinaciones de distancias y enlaces. A continuación se exponen las más significativas:

Combinación Distancia-Enlace	Coeficiente Cofenético
Euclídea-Completo	0.5447257
Euclídea-Ward	0.4567209
Manhattan-Ward	0.477653
Máximo-MacQuity	0.2308923
Máximo-Ward	0.2511533
Canberra-Completo	0.4232287
Canberra-Promedio	0.548735
Canberra-Ward	0.4726828
Canberra-MacQuitty	0.4690213
Minkowski-Ward	0.4561093

Tabla 5. Combinaciones de distancia y enlace.

La mejor combinación resultó ser la del **enlace de Ward con la distancia Euclídea,** ya que sus *clusters* son más compactos y están mejor organizados. En el Anexo 6 se pueden observar las gráficas obtenidas con otras combinaciones que también se consideran efectivas.

Se realiza el AC con la combinación seleccionada. El resumen de los resultados de este método se muestra en la siguiente tabla:

Análisis de cluster		
Cluster	**Nombre del cluster**	**No. Observac.**
1	Aptos con limitaciones	65
2	Aptos	33
3	No aptos	24

Tabla 6. Resumen del análisis de *cluster* .

En la Fig. 15 se muestran los conglomerados que se obtuvieron al ejecutar la función **pcacomb()** del paquete para el análisis estadístico de datos. Esta función combina el ACP con el AC y permite obtener la información de los grupos en el gráfico de dispersión de individuos (los gráficos restantes se pueden apreciar en los Anexos 3 y 4).

Se observa que el rectángulo rojo se forma en la parte positiva de la primera componente y representa a los estudiantes clasificados como “Aptos”. Estos presentan habilidades generales en todas las asignaturas y promedios en el intervalo de [4,35; 4,97].

El rectángulo azul se ubica cercano al cero del eje de coordenadas y ligeramente retirado hacia la parte negativa de la primera componente. En este conjunto se agrupan los estudiantes “Aptos con limitaciones” con promedios en el intervalo de [3,75; 4,34]. Los

mejores resultados de aquellos que están ubicados en el área positiva del gráfico, por lo general se presentan en asignaturas como IP, P1, P3.

Por último, el rectángulo verde se ubica en la parte negativa respecto a la primera componente y agrupa a los estudiantes “No aptos”. Estos tienen promedios en el intervalo de [3,32; 3,74] y presentan mayores dificultades en el proceso de formación.

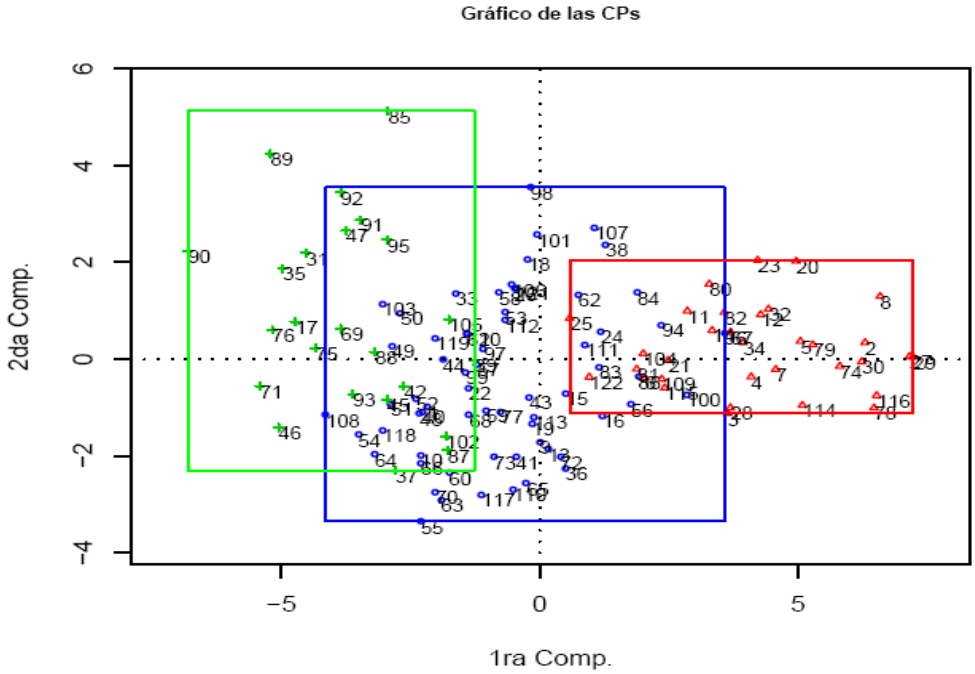


Fig. 15. Combinación de ACP con AC.

3.7. Etiquetado de los conglomerados

Una vez formados los *clusters*, se procede a determinar el comportamiento de los estudiantes en relación con los valores que ofrecen las herramientas de diagnóstico del CICE. Esta combinación de información puede ser empleada por expertos como soporte para la toma de decisiones en la gestión de recursos humanos, en cualquier centro productivo de la UCI.

En el Anexo 8 se pueden observar las salidas del software que contemplan el proceso de etiquetado de cada miembro del conglomerado con los respectivos valores de la Matriz de fortalezas y debilidades.

A continuación se resumen los resultados obtenidos:

Análisis de la Matriz de fortalezas y debilidades						
	Grupo 1		Grupo 2		Grupo 3	
	Media	Mediana	Media	Mediana	Media	Mediana
CAAS	3.615385	4	3.878788	4	3.125	3
CACP	3.953846	4	4.272727	4	3.875	4
RSCCEP	3.615385	4	3.757576	4	3.583333	4
COPT	4.184615	4	4.454545	5	4.125	4
CCOE	3.969231	4	4	4	3.666667	4
CCSI	2.676923	3	2.848485	3	2.958333	3
CI	3.738462	4	3.636364	4	3.708333	4
CAAP	3.676923	4	3.969697	4	3.708333	4
HBPAI	3.738462	4	3.757576	4	3.458333	3.5
CCA	4.107692	4	4.090909	4	3.875	4
CC	3.953846	4	3.757576	4	3.75	4
CIPRP	3.815385	4	4.090909	4	3.75	4
CTD	4	4	3.969697	4	3.916667	4
CTE	4.061538	4	4.30303	5	3.958333	4
CMCMC	3.830769	4	3.818182	4	3.791667	4
HTA	3.8	4	3.969697	4	3.75	4
CCL	3.8	4	3.969697	4	3.75	4

Tabla 7. Resumen del proceso de etiquetado.

De esta información se puede interpretar que los estudiantes que integran los grupos 1 y 2 poseen capacidades o habilidades superiores a los del grupo 3.

Los que conforman el grupo 2 podrán incorporarse a proyectos de inmediato, mientras que aquellos que pertenecen al grupo 1 deberán recibir una capacitación básica o tener responsabilidades de menor importancia dentro del equipo. En cambio, para los estudiantes que se ubican en el grupo 3 se deberán trazar estrategias de formación bien definidas, ya que no cuentan con las habilidades necesarias para desempeñarse con eficiencia en la producción.

Los resultados obtenidos confirman la clasificación de los grupos en: Aptos, Aptos con limitaciones y no Aptos, por tanto validan el funcionamiento de la herramienta propuesta.

3.8. Conclusiones del capítulo

En este capítulo se validó el paquete de funciones para el análisis estadístico de datos mediante un caso de estudio.

La aplicación de técnicas de análisis de componentes principales y análisis de *cluster,* implementadas como soporte a la toma de decisiones en los procesos de incorporación de estudiantes a proyectos productivos, arrojó las siguientes conclusiones:

Desde el punto de vista estadístico:

- ✓ El análisis de correlaciones determinó que existe multicolinealidad en la muestra de datos, y que las variables se encuentran correlacionadas de forma fuerte, ligera y débil. Este resultado demostró que los métodos escogidos para el análisis de datos se podían aplicar.
- ✓ Se logró la reducción de la matriz de datos que contenía 37 variables en 8 componentes, que describen la mayor parte de la información contenida en los datos.
- ✓ Se logró condensar el 60% de la variabilidad de los datos, condición necesaria para la realización de estudios de Ciencias Sociales.
- ✓ A partir de la gráfica de segmentación del algoritmo K-Medias se confirmó que la cantidad de grupos a formar era 3.

Desde el punto de vista del análisis:

- ✓ Las asignaturas IE3 e IE4 correlacionan con fuerza, mientras que las de Ciencias Básicas, y la especialidad lo hacen (al menos con una de ellas) ligeramente. En cambio la correlación de PSCT, PPD, PP1, PHCC, y PHCU con otras materias es débil.
- ✓ Se determinó que las asignaturas de MC2, MC1, MD, SO, IE4, P1, F1 contribuyen a diferenciar a los estudiantes ya que sus valores de carga son elevados y aportan la mayor cantidad de varianza a la muestra de datos.
- ✓ Los estudiantes que se ubican en la región positiva con respecto a la primera componente, en el gráfico de dispersión, presentan habilidades generales y buenos resultados académicos.
- ✓ El análisis de *cluster* se realizó con la combinación del enlace de Ward y la distancia Euclídea. Se obtuvieron tres grupos (Aptos, Aptos con limitaciones y No aptos) que clasifican a los estudiantes teniendo en cuenta sus resultados académicos.
- ✓ Los grupos fueron etiquetados con los resultados de la encuesta: Matriz de fortalezas y debilidades. Con la interpretación de este procedimiento se validó la formación de los conglomerados resultantes.

Conclusiones generales

Con la realización del presente trabajo se arriba a las siguientes conclusiones generales:

- ✓ Se implementó un paquete de funciones en R que proporciona salidas con un nivel de calidad aceptable. Permite combinar las técnicas de análisis de componentes principales y análisis de *cluster*, así como generar gráficos que posibilitan una mejor interpretación de los datos.
- ✓ Con la aplicación del análisis de componentes principales se disminuyó la dimensión de los datos del caso de estudio. Inicialmente se tenían 37 variables (asignaturas) que fueron reducidas a 8 componentes principales, logrando explicar el 60% de la variabilidad de los datos.
- ✓ Con la aplicación de técnicas de análisis de *cluster* se agrupó a los estudiantes con comportamientos similares en sus resultados docentes, lo cual permite identificar a aquellos con mayores habilidades y los que necesitan algún tipo de capacitación, antes o después de ingresar a un proyecto.
- ✓ Se realizó el etiquetado de los estudiantes con valores procedentes de la encuesta "Matriz de fortalezas y debilidades". Los resultados obtenidos permitieron la validación de la clasificación de los grupos formados con el método análisis de *cluster*.
- ✓ A partir de los métodos antes mencionados y la combinación de los mismos, se logró conformar tres grupos de estudiantes clasificados en: *Aptos, Aptos con limitaciones y No aptos*. Dicha clasificación permitirá trazar estrategias de formación que incidan directamente en el proceso de incorporación de estudiantes a proyectos de desarrollo de software en la UCI.

Recomendaciones

- ✓ Continuar con el desarrollo del paquete de funciones para el análisis de datos e incluir técnicas de agrupamiento no jerárquicas.
- ✓ Extender el uso del paquete a otros centros de la UCI, de forma que pueda ser utilizado por los directivos como soporte a la toma de decisiones en procesos de formación de estudiantes desde la producción.

Bibliografía

ANDERBERG, M.R, 1973. *Cluster Analysis for Applications*. New York: Academic Press.

ANDRÉ, Margarita Ampuero, 2009. *Un modelo para la asignación de recursos humanos a equipos de proyectos de software*. Tesis de Doctorado. La Habana: Instituto Superior Politécnico "José Antonio Echeverría" Facultad de Ingeniería Informática.

ANON.[no date]. The R Project for Statistical Computing. In: [online]. [Accessed 23 February 2012]. Available from: http://www.r-project.org/.

BOUTHILLER, F. and SHEARER, K., 2002. Understanding knowledge management and information management: the need for an empirical perspective. In: *Information research*. 2002. Vol. 8, no. 1.

DAS, Swagatam, ABRAHAM, Ajith and KONAR, Amit, 2009. *Metaheuristic Clustering*. Berlin: Springer-Verlag Berlin Heidelberg. ISBN 978-3-540-92172-1.

FEBLES, A., 2004. *Un modelo de referencia para la gestión de configuración en la pequeña y mediana empresa de software.* Tesis de Doctorado en Ciencias Técnicas. La Habana: Instituto Superior "José Antonio Echeverría."

GORDON, A., HAYASHI, C. and N. OHSUMI, 1998. Cluster validation. In: *Data Science, Classifi cation, and Related Methods. S.l.: NY : Springer - Verlag. pp. 22–39.*

HAIR, J. and ANDERSON, R*.,* 1999. *Análisis multivariante*. 5. Madrid: Prentice-Hall.

HEREDIA, Ing. Jobany Rico José, 2011. *Análisis de datos en apoyo a la productividad en el proceso de formación de ingenieros*. Tesis de Maestría. La Habana: Facultad de Ingeniería Industrial. Instituto Superior Pedagógico José Antonio Echeverría.

JACKSON, 1991. *A User Guide to Principal Components*. New York: Wiley.

LAKHANPAL, B., 1993. Understanding the Factors Influencing the Performance of Software Development Groups: An Exploratory Group-Level Analysis. In: *Information and Software Technology*. 1993.

LUAN, J., 2002. Data Mining and Knowledge Managment in Higher Education-Potentiak Aplications. In: *Anual Forum for Association for Institutional Reserch*. Toronto. 2002.

MARTÍNEZ, Hugo, MEDRANO, Bolivar E., FERNÁNDEZ, Lytyet and TEJEDA, Yunier, 2012. Análisis multivariado de datos como soporte a la decisión en la selección de estudiantes en proyectos de software. In: *Revista de Ingeniería Industrial*. 2012. Vol. XXXIII.

MASSY, W.F. and WILGER, A.K., 1995. Improving Productivity: What Faculty Think About It and It´s Efect on Quality. In: *Change: The Magazine of higher Learning*. 1995. Vol. 4, no. 27, pp. 10–20.

MEDRANO, Bolivar E., MARTÍNEZ, Hugo, FERNÁNDEZ, Lytyet and DÍAZ, Maybel, 2010. Técnicas de Análisis Multivariado como soporte a la toma de decisiones. In: *Simposio de Ingeniería Industrial*. La Habana: s.n. 2010.

MOLINA, M., 2000. *Gestión de Recursos Humanos en Proyectos Informáticos*. Ciudad Real: Universidad de Castilla-La Mancha.

PARDO, Campo Elías and CAMPO, Pedro César Del, 2007. Combinación de métodos factoriales y de análisis de conglomerados en R: el paquete FactoClass. In: *Revista Colombiana de Estadística*. 2007. Vol. 30, no. 2, pp. 231–245.

PONJUÁN, G., 2006. *Gestión de la información en las organizaciones. Principios, conceptos y aplicaciones.* La Habana: Félix Varela.

PRESSMAN, R. S, 2004. *Software Engineering: A Practitioner's Approach*. S.l.: McGraw-Hill Science.

QUINTANA, Lic. Hilario, 2011. *Modelo para el análisis de datos de procesos de formación del profesional.* Tesis de Maestría. La Habana: Universidad de las Ciencias Informáticas.

RODRÍGUEZ, Dra. Ing. Aida G. Hernández, 2009. *Análisis multivariado*. Tesis de Maestría. La Habana: Facultad de Ingeniería Industrial. Instituto Superior Politécnico José Antonio Echeverría.

SANTANA, Oscar Fernández, 1991. El análisis cluster: Aplicación, interpretación y validación. In: *Papers: Revista de Sociologia*. 1991. no. 37, pp. 65–76.

THOMPSON, Laura A., 2009. *R (and S-PLUS) Manual to Accompany Agresti's Categorical Data Analysis (2002) 2nd edition*. S.l.: s.n.

TUSELL, F., 2004. *Lectura, manipulación y análisis de datos en R*. 2004. S.l.: s.n.

URIARTE, Ramón-Díaz, 2003. Introducción al uso y programación del sistema estadístico R. In: Unidad de Bioinformática Centro Nacional de Investigaciones Oncológicas (CNIO). 2003.

WASTELL, D., 1999. *The Human Dimension of the Software Process*. Software Process: Principles, Methodology and Technology. S.l.: Springer-Verlag.

XUI, Rui and WUNSCH, Donald C., 2009. *Clustering*. Piscataway: IEEE Press Editorial Board. ISBN 978-0-470-27680-8.

Glosario de términos

A

Análisis exploratorio de datos: Tratamiento estadístico al que se someten las muestras recogidas durante un proceso de investigación científica.

AC: Análisis de conglomerados.

ACP: Análisis de componentes principales.

AGNES: Agglomerative Nesting.

AL: Álgebra Lineal.

AE: Administración de Empresas.

B

BIRCH: Balanced Iterative Reducing and Clustering using Hierarchies.

C

Cluster: Palabra en inglés que significa ¨conglomerado¨ o agrupación.

Conglomerado: Unión compacta de elementos.

Competencia laboral: Características subyacentes en la persona, asociada a la experiencia, que como tendencia están causalmente relacionadas con actuaciones exitosas en un puesto de trabajo contextualizado en una determinada cultura organizacional.

CES: Centro de Enseñanza Superior.

CEIGE: Centro de gobierno electrónico.

GRH: Gestión de recursos humanos.

CURE: Clustering Using Representatives.

CLARA: Clustering Large Applications.

CLARANS: Clustering Large Applications based on Randomized Search

COBWEB: Algoritmo de *cluster* jerárquico para modelos conceptuales.

CRAN: Comprehensive R Archive Network.

CF: Contabilidad y Finanzas.

CAAS: Capacidad de abstracción, análisis y síntesis.

CACP: Capacidad de aplicar los conocimientos en la práctica.

COPT: Capacidad para organizar y planificar el tiempo.

CCOE: Capacidad de comunicación oral y escrita.

CCSI: Capacidad de comunicación en un segundo idioma.

CI: Capacidad de investigación.

CAAP: Capacidad de aprender y actualizarse permanentemente.

CCA: Capacidad crítica y autocrítica.

CC: Capacidad creativa.

CIPRP: Capacidad para identificar, plantear y resolver problemas.

CTD: Capacidad para tomar decisiones.

CTE: Capacidad de trabajo en equipo.

CMCMC: Capacidad de motivar y conducir hacia metas comunes.

CCL: Compromiso con la calidad.

C: Cerebro.

CH: Cohesionador.

COR: Coordinador.

D

Discretización: Transformación de datos cualitativos a numéricos.

Dendrograma: Representación gráfica o diagrama de datos en forma de árbol.

DIANA: Divisive Analysis.

E

EM: Expectation- Maximation.

EF1: Educación Física 1.

EF2: Educación Física 2.

IE1: Idioma Inglés 1.

IE2: Idioma Inglés 2.

EF3: Educación Física 3.

EF4: Educación Física 4.

EP1: Economía Política 1.

EP2: Economía Política 2.

EF5: Educación Física 5.

EF6: Educación Física 6.

E: Especialista.

F

FANNY: Fuzzy Analysis.

F1: Física 1.

F2: Física 2.

F: Finalizador.

FS: Filosofía y Sociedad.

G

GC: Gráficos Computadoras.

H

Heterogeneidad: Mezcla de partes de diversa naturaleza en un todo.

HBPAI: Habilidades para buscar, procesar y analizar información procedente de diferentes fuentes.

HTA: Habilidad para trabajar en forma autónoma.

I

ISO: International Organization for Standardization (Organización internacional de normalización).

IP: Introducción a la Programación.

IS1: Ingeniería de Software 1.

IS2: Ingeniería de Software 2.

IE3: Idioma Extranjero 3.

IE4: Idioma Extranjero 4.

I: Implementador.

IM: Impulsor.

IR: Investigador de Recursos.

M

MONA: Monothetic Analysis.

ME: Monitor Evaluador.

Matriz simétrica: Matriz cuadrada (*m*X*n*; n=m), igual a su traspuesta.

Matriz de correlación: Matriz que muestra la correlación o interdependencia existente entre variables.

M3: Matemática 3.

M4: Matemática 4.

MC1: Máquina Computadoras 1.

MC2: Máquina Computadoras 2.

M1: Matemática 1.

M2: Matemática 2.

MD: Matemática Discreta.

N

Numeración: Transformación de variables numéricas en ordinales**.**

O

Outlier: Observación que está numéricamente distante del resto de los datos de una muestra.

P

PMBOK: Project Management Body of Knowledge (Proyecto del Órgano de Gestión del Conocimiento). Estándar elaborado por el Instituto de Gestión de Proyectos.

P1: Programación 1.

PHCC: Panorama Histórico Cultural Cubano.

PHCU: Panorama Histórico Cultural Universal.

PP1: Práctica Profesional 1.

PPD: Preparación para la Defensa.

P2: Programación 2.

P3: Programación 3.

P4: Programación 4.

PP2: Práctica Profesional 2.

PE: Probabilidades y Estadísticas.

PP3: Práctica Profesional 3.

PSCT: Problemas Sociales de la Ciencia y las Tecnologías.

PAM: Partition around medoids.

R

ROCK: Robust Clustering using links.

SBD: Sistema de Base de Datos.

RSCCEP: Responsabilidad social, compromiso ciudadano y ética de la profesión.

Rol: Función que alguien o algo cumple. Caracteriza la particular tendencia que tiene cada persona a comportarse, contribuir y relacionarse socialmente.

S

Software libre: Solución informática con libertades para ser ejecutada, copiada, distribuida, estudiada, modificada y distribuida modificada.

SCE: Suma de cuadrados error.

SOM: Self-organizing Maps.

Sib: Sequential information bottleneck.

SV: Support Vector Machine Clustering.

SO: Sistema Operativo.

T

TIC: Tecnología de la Informática y las Comunicaciones.

T1: Teleinformática 1.

T2: Teleinformática 2.

U

URL: Localizador de recursos uniforme.

UNIX: Sistema Operativo.

W

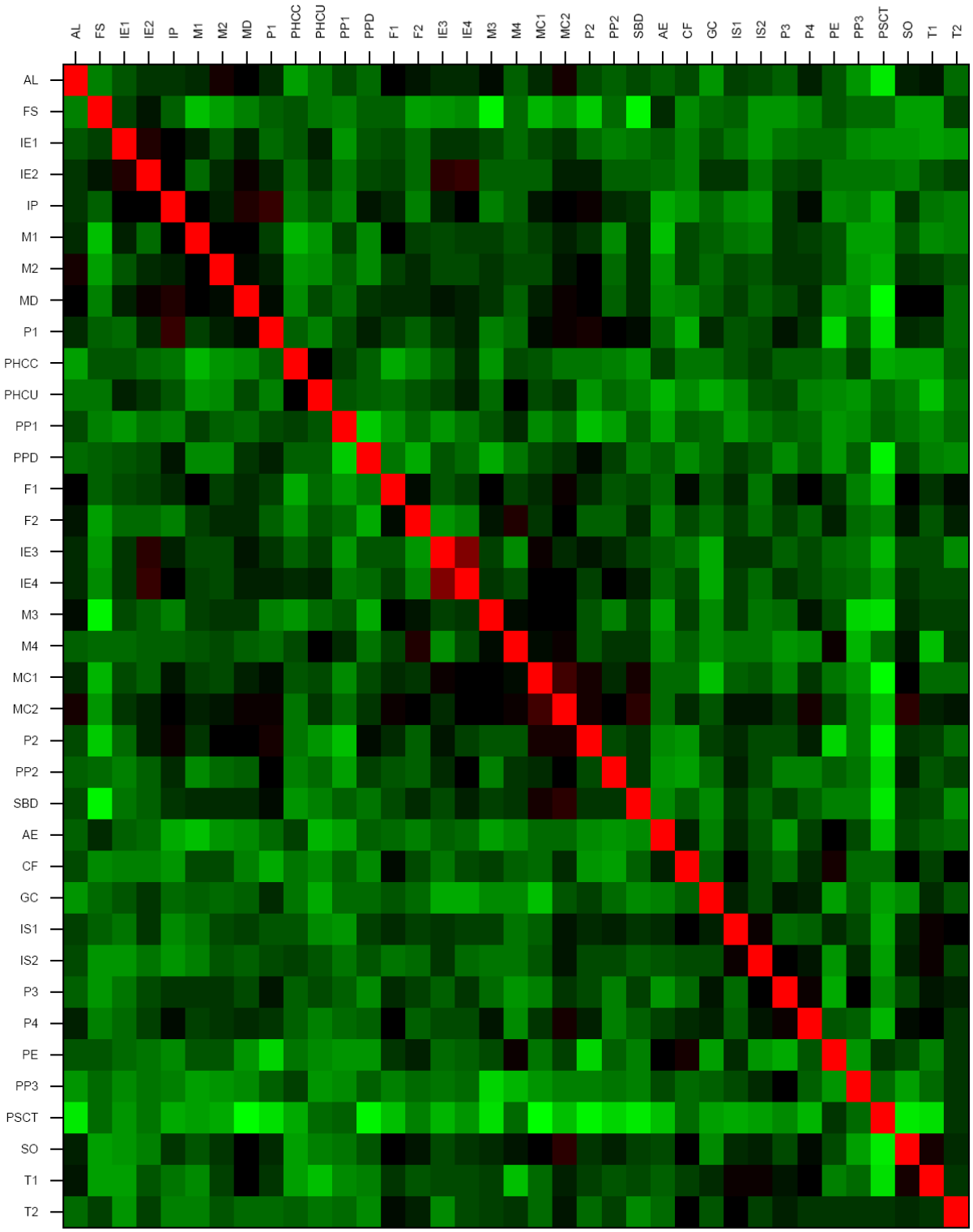
WEKA: Waikato Environment for Knowledge Analysis.

Y

Yale: Yet Another Learning Environment.

Anexos

Anexo 1: Gráfico de la matriz de correlación.

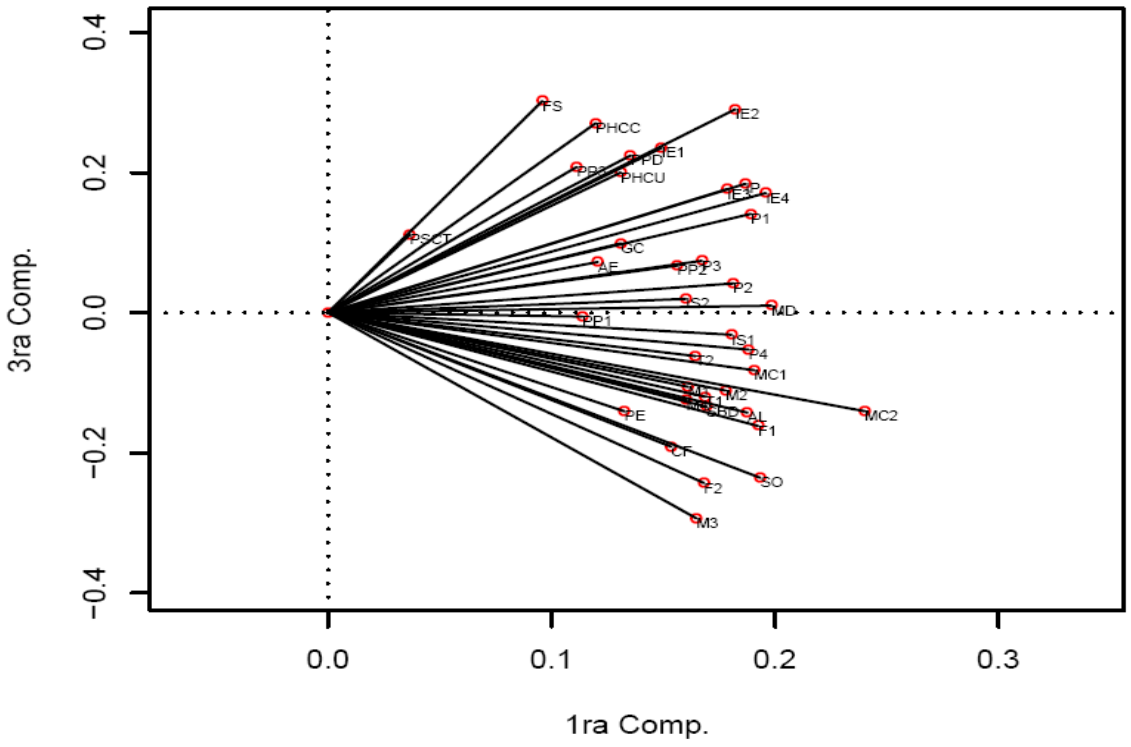


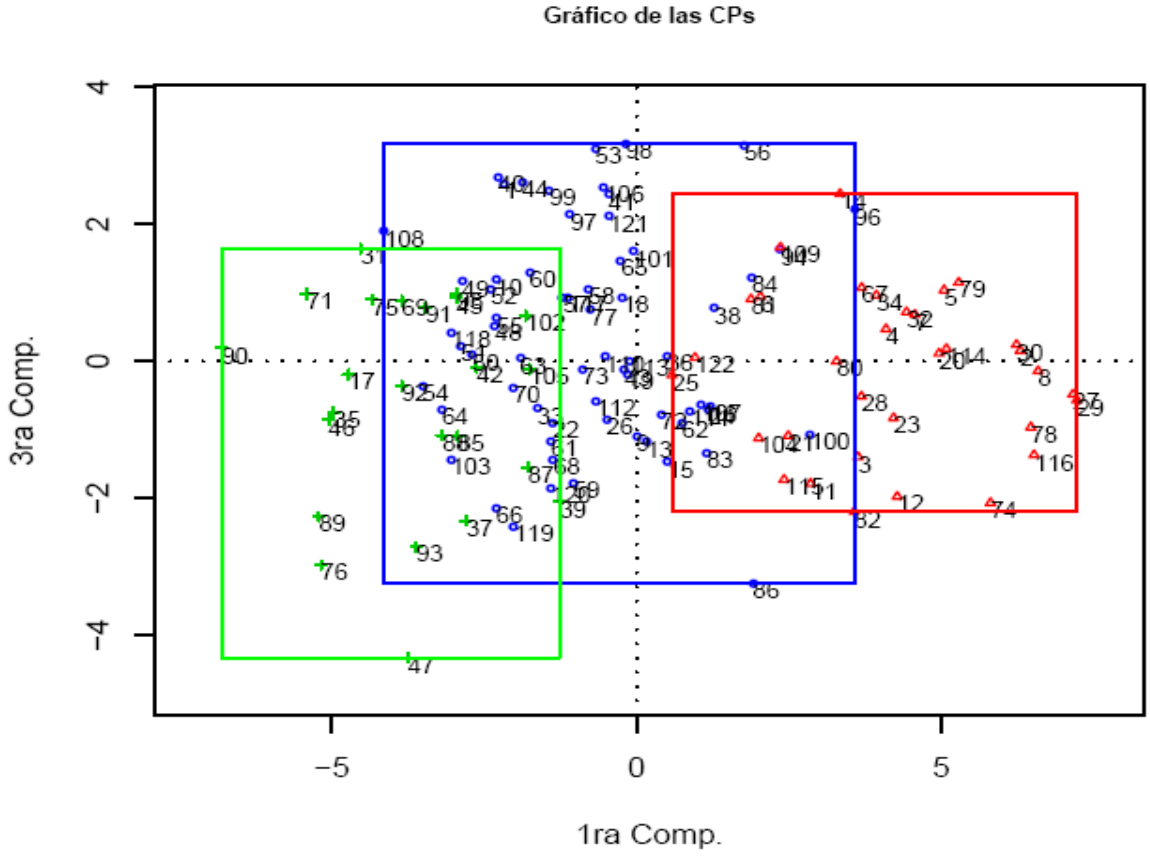
La parte roja de la grafica indica que la correlación es bien alta y positiva, la parte verde indica que es bien alta pero negativa y la parte negra indica que la correlación es cerca de cero.

Anexo2: Matriz de vectores de carga.

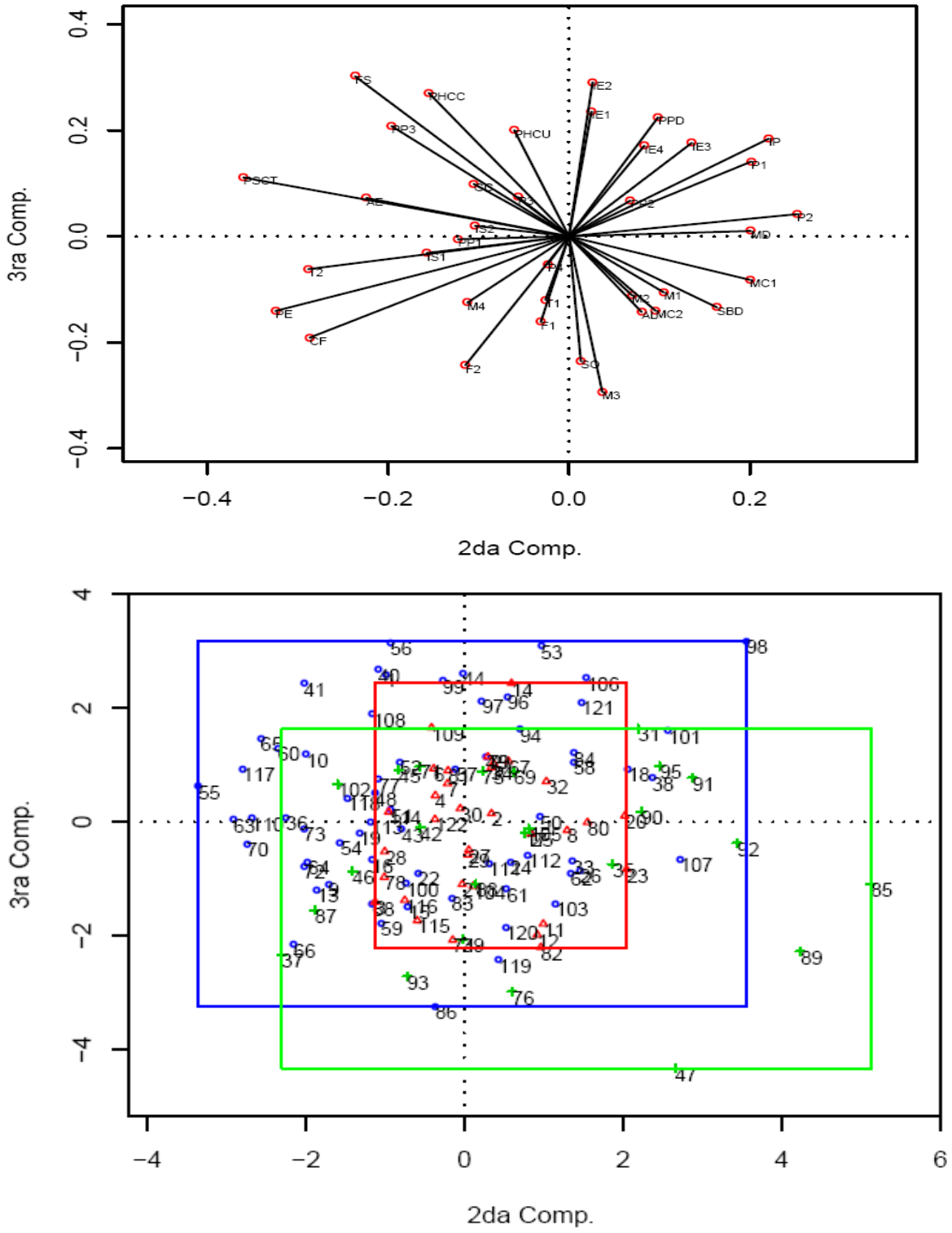
	[CP1]	[CP2]	[CP3]	[CP4]	[CP5]	[CP6]	[CP7]	[CP8]
AL	0.1877119	0.0000000	0.0000000	0.0000000	0.0000000	0.1854058	0.0000000	0.0000000
FS	0.0000000	-0.2366138	0.3018344	0.0000000	0.0000000	0.2238676	0.3084560	0.0000000
IE1	0.1491417	0.0000000	0.2350227	0.2131054	0.0000000	0.2403136	0.0000000	0.0000000
IE2	0.1823061	0.0000000	0.2903438	0.0000000	0.0000000	0.2417423	0.0000000	0.0000000
IP	0.1867218	0.2207626	0.1839447	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
M1	0.1610307	0.0000000	0.0000000	0.0000000	-0.3307582	0.0000000	0.0000000	0.0000000
M2	0.1778980	0.0000000	0.0000000	0.0000000	-0.2243710	0.0000000	0.0000000	0.0000000
MD	0.1985209	0.2011206	0.0000000	0.0000000	0.0000000	0.1403892	0.0000000	0.0000000
P1	0.1892132	0.2016258	0.0000000	0.0000000	0.0000000	0.0000000	0.2346544	0.0000000
PHCC	0.0000000	0.0000000	0.2704357	0.0000000	0.0000000	-0.2685943	0.0000000	0.3776816
PHCU	0.1309841	0.0000000	0.2000557	0.2999406	0.0000000	-0.2731125	0.0000000	0.0000000
PP1	0.0000000	0.0000000	0.0000000	0.0000000	-0.2597768	-0.2642397	0.0000000	0.2367506
PPD	0.1354089	0.0000000	0.2243407	0.0000000	0.2036330	0.0000000	0.3028234	0.0000000
F1	0.1924395	0.0000000	-0.1616305	0.0000000	0.0000000	0.1831694	0.0000000	0.0000000
F2	0.1682290	0.0000000	-0.2430235	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
IE3	0.1786768	0.0000000	0.1763541	0.0000000	0.1798774	0.0000000	-0.4352261	0.0000000
IE4	0.1959027	0.0000000	0.1708043	0.0000000	0.0000000	0.0000000	-0.3742394	-0.2233705
M3	0.1648184	0.0000000	-0.2934645	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
M4	0.1604760	0.0000000	0.0000000	0.3603296	0.0000000	-0.2332370	0.2534771	0.0000000
MC1	0.1906888	0.2000197	0.0000000	0.0000000	0.2467047	0.0000000	0.0000000	0.0000000
MC2	0.2401488	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
P2	0.1812988	0.2522277	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
PP2	0.1562361	0.0000000	0.0000000	0.0000000	0.2723198	-0.1556316	0.0000000	-0.3722513
SBD	0.1692931	0.0000000	0.0000000	0.0000000	0.0000000	-0.2081492	0.0000000	0.0000000
AE	0.1206635	-0.2242198	0.0000000	0.0000000	0.2816975	0.2449783	0.0000000	0.3086560
CF	0.1534237	-0.2866984	-0.1910647	0.0000000	0.0000000	0.1592255	0.0000000	0.0000000
GC	0.1309775	0.0000000	0.0000000	-0.2576142	-0.2686049	0.0000000	0.0000000	0.0000000
IS1	0.1805977	0.0000000	0.0000000	-0.2054604	0.2063108	0.0000000	0.0000000	0.0000000
IS2	0.1600956	0.0000000	0.0000000	-0.2817722	0.0000000	-0.2447957	0.0000000	0.0000000
P3	0.1672665	0.0000000	0.0000000	-0.2122636	-0.3140205	-0.2255124	0.0000000	0.0000000
P4	0.1882363	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.2379819
PE	0.1326778	-0.3245935	0.0000000	0.2042048	0.0000000	0.2366781	0.0000000	0.0000000
PP3	0.0000000	-0.1962419	0.2078551	-0.2197163	0.0000000	-0.1800381	0.0000000	0.0000000
PSCT	0.0000000	-0.3601434	0.0000000	0.1928556	-0.2391706	0.0000000	0.0000000	-0.3582639
SO	0.1934470	0.0000000	-0.2355890	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
T1	0.1687306	0.0000000	0.0000000	-0.3674700	0.0000000	0.0000000	0.0000000	0.0000000
T2	0.1644494	-0.2883537	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	-0.2619107

Anexo 3: Gráfico de la primera componente con la tercera.

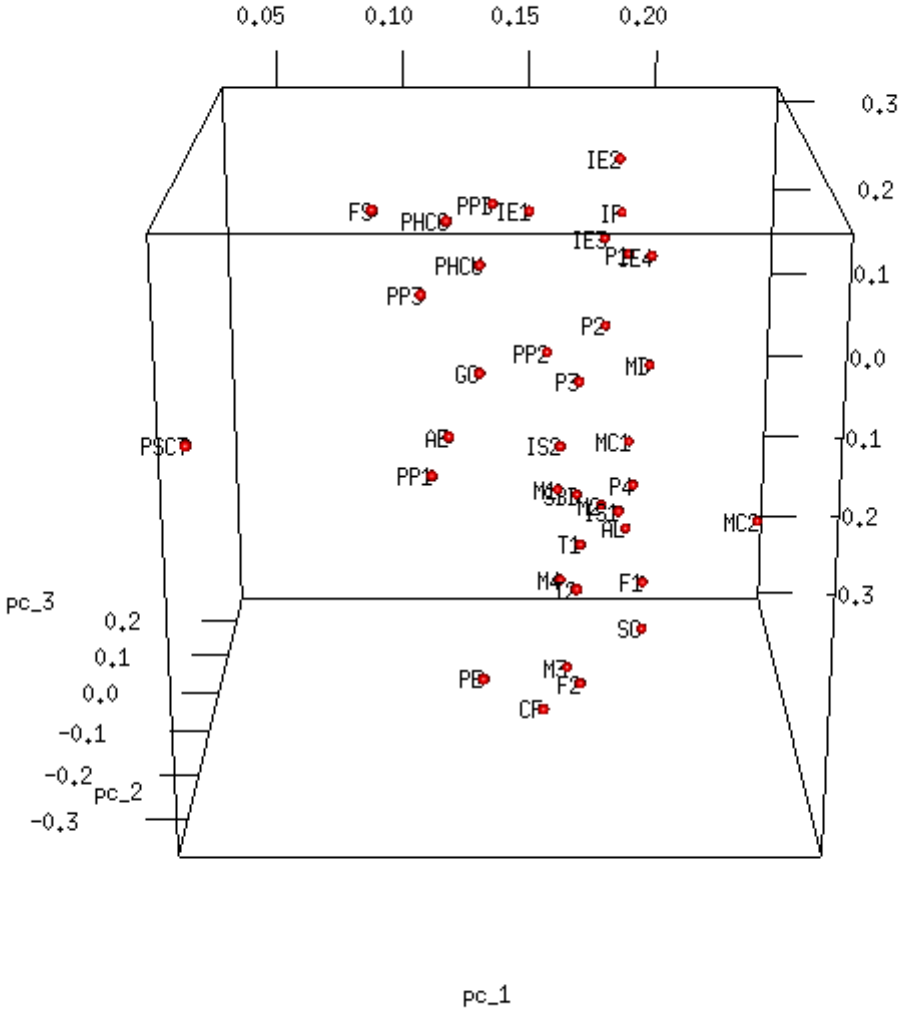




Anexo 4: Gráfico de la segunda componente con la tercera.

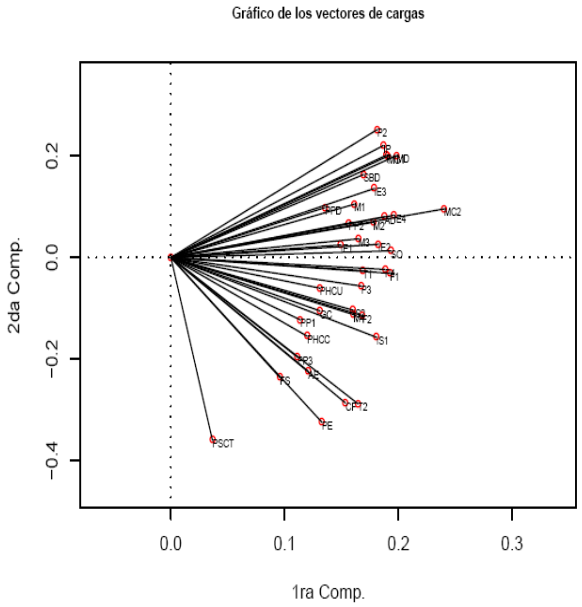


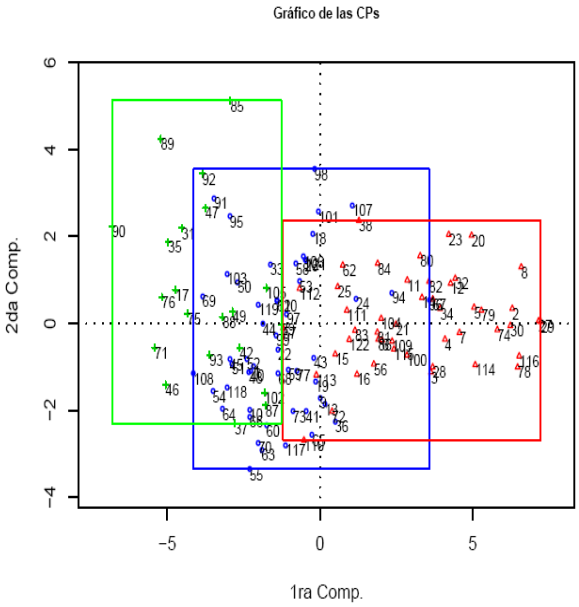
Anexo 5: Gráfico de las componentes principales en 3 dimensiones.



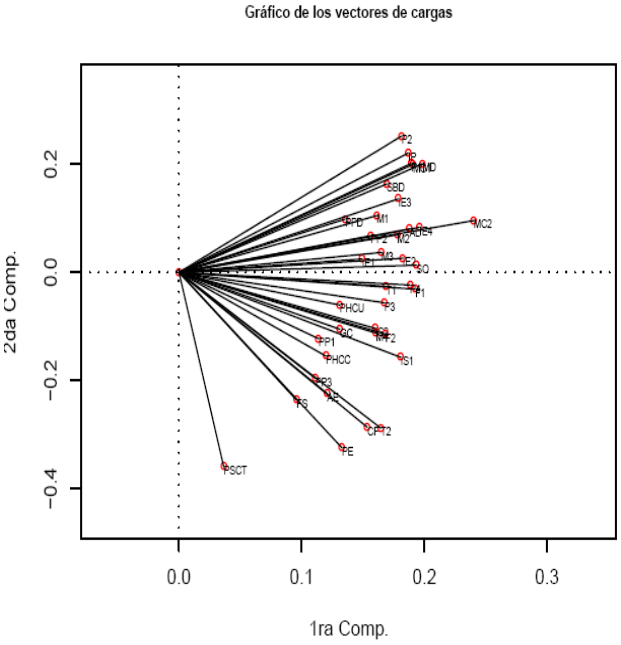
Anexo 6: Gráficos con combinaciones que obtienen buenos agrupamientos.

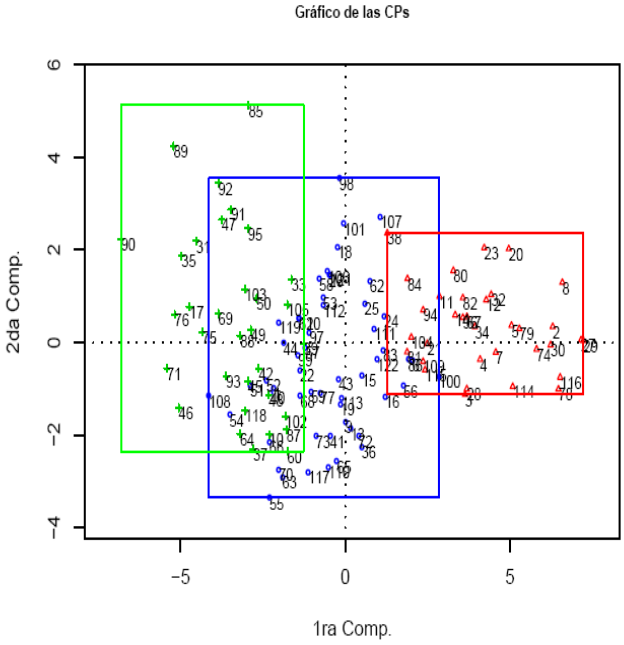
Combinación Euclídea-Completo





Combinación Minkowski-Ward





Anexo 7: Interpretación de los instrumentos para cada indicador.

Test "Autopercepción de Belbin": Conocer el rol o los roles que mejor pueden desempeñar los estudiantes en un equipo de trabajo e identificar las aptitudes para el liderazgo.	
Indicadores	**Interpretación**
Cerebro (rol mental)	Creativo, imaginativo, poco ortodoxo. Resuelve problemas difíciles.
Coordinador (rol social)	Maduro, seguro de sí mismo. Aclara las metas a alcanzar. Promueve la toma de decisiones. Delega bien.
Monitor-Evaluador (rol mental)	Serio, perspicaz y estratega. Percibe todas las opciones. Juzga con exactitud.
Implementador (rol de acción)	Disciplinado, leal, conservador y eficiente. Transforma las ideas en acciones.
Finalizador (rol de acción)	Esmerado, concienzudo, ansioso. Busca los errores y las omisiones. Realiza las tareas en el plazo establecido.
Investigador de Recursos (rol social)	Extrovertido, entusiasta, comunicativo. Busca nuevas oportunidades. Desarrolla contactos.
Impulsor (rol de acción)	Retador, dinámico, trabaja bien bajo presión. Tiene iniciativa y coraje para superar obstáculos.
Cohesionador (rol social)	Cooperador, apacible, perceptivo y diplomático. Escucha e impide los enfrentamientos.
Especialista (rol mental)	Sólo le interesa una cosa a un tiempo. Aporta cualidades y conocimientos específicos.
Test "Estilos decisionales": Identificar los estilos predominantes en los estudiantes en la toma de decisiones.	
Indicadores	**Interpretación**
Directivo	Individuos que buscan la racionalidad y evitan la ambigüedad. Toman las decisiones raudamente y se orientan en el corto plazo. Son eficientes y lógicos; su eficiencia da como producto una toma de decisiones con poca información y opciones evaluadas mínimas.
Analítico	Personas que toleran un poco más la ambigüedad que los individuos del estilo anterior. Busca más información y quiere considerar más alternativas que los del estilo directivo.
Conceptual	Individuos que buscan mucha información y alternativas de solución. Emplean la creatividad con buenos resultados. Su orientación es a largo plazo.
Conductual	Sujetos que trabajan adecuadamente con los demás. Se interesan en el logro de los compañeros y de los subordinados y aceptan las sugerencias de las demás personas, se apoyan bastante en reuniones para comunicarse. Evitan conflictos y buscan en toda circunstancia la aceptación.
Matriz de Fortalezas y Debilidades: Conocer cuáles son las principales fortalezas y debilidades presentes en los estudiantes.	
Indicadores	**Interpretación**
Fortaleza	Existencia de una capacidad o recursos personológicos que posee el individuo, en condiciones de ser aplicados para alcanzar sus objetivos y concretar planes.
Debilidad	Falta de una determinada capacidad o condición, así como pocos recursos personológicos que pueden dificultar el logro de objetivos y metas del sujeto.

Anexo 8: Etiquetado de los estudiantes de cada *cluster* con los resultados de la Matriz de fortalezas y debilidades.

Variable: *CAAS.*

```
> clusterh(4,40,DATA,41,"Single",6,1,3)
[1] "Dist. Euclidena"
[1] "Enlace de Ward"
[1] "C. Cof"
[1] 0.4567209
[1] "Grupo 1"
  E1   E9  E10  E13  E15  E16  E18  E19  E22  E24  E26  E33  E36  E38  E40  E41  E43  E44  E48  E49  E50  E51  E52  E53  E54  E55
   5    5    2    3    2    5    3    3    4    1    4    4    3    4    3    3    3    2    2    3    4    4    4    2    3    3
 E56  E57  E58  E59  E60  E61  E62  E63  E64  E65  E66  E68  E70  E72  E73  E77  E83  E84  E86  E94  E96  E97  E98  E99 E100 E101
   4    4    3    1    5    3    4    4    4    4    4    3    5    4    3    4    4    5    3    5    5    5    4    5    3    5
E103 E106 E107 E108 E110 E111 E112 E113 E117 E118 E119 E120 E121
   5    5    3    3    4    3    3    5    4    2    3    5    3
      Media Mediana C.Ind
[1,] 3.615385       4    65
[1] "Grupo 2"
  E2   E3   E4   E5   E6   E7   E8  E11  E12  E14  E20  E21  E23  E25  E27  E28  E29  E30  E32  E34  E67  E74  E78  E79  E80  E81
   5    4    4    4    5    5    3    5    3    4    2    5    4    3    4    2    5    4    3    2    1    5    3    5    5    3
 E82 E104 E109 E114 E115 E116 E122
   4    5    3    5    5    4    4
      Media Mediana C.Ind
[1,] 3.878788       4    33
[1] "Grupo 3"
 E17  E31  E35  E37  E39  E42  E45  E46  E47  E69  E71  E75  E76  E85  E87  E88  E89  E90  E91  E92  E93  E95 E102 E105
   3    3    3    3    2    4    3    2    4    2    3    4    2    3    3    3    2    3    3    3    4    5    3    5
    Media Mediana C.Ind
[1,] 3.125       3    24
```

Variable: *CACP.*

```
> clusterh(4,40,DATA,42,"Single",6,1,3)
[1] "Dist. Euclidena"
[1] "Enlace de Ward"
[1] "C. Cof"
[1] 0.4567209
[1] "Grupo 1"
  E1   E9  E10  E13  E15  E16  E18  E19  E22  E24  E26  E33  E36  E38  E40  E41  E43  E44  E48  E49  E50  E51  E52  E53  E54  E55
   5    5    3    4    3    4    3    4    4    3    5    4    4    4    4    2    3    3    4    4    4    5    4    3    4    3
 E56  E57  E58  E59  E60  E61  E62  E63  E64  E65  E66  E68  E70  E72  E73  E77  E83  E84  E86  E94  E96  E97  E98  E99 E100 E101
   5    3    5    1    4    4    4    4    4    3    4    3    5    5    3    3    4    5    4    4    5    4    4    5    4    5
E103 E106 E107 E108 E110 E111 E112 E113 E117 E118 E119 E120 E121
   5    5    2    4    5    4    4    5    4    5    4    4    4
      Media Mediana C.Ind
[1,] 3.953846       4    65
[1] "Grupo 2"
  E2   E3   E4   E5   E6   E7   E8  E11  E12  E14  E20  E21  E23  E25  E27  E28  E29  E30  E32  E34  E67  E74  E78  E79  E80  E81
   5    5    5    5    5    5    4    4    3    5    4    4    4    3    4    2    5    4    4    5    1    4    4    5    5    4
 E82 E104 E109 E114 E115 E116 E122
   5    4    5    5    5    5    4
      Media Mediana C.Ind
[1,] 4.272727       4    33
[1] "Grupo 3"
 E17  E31  E35  E37  E39  E42  E45  E46  E47  E69  E71  E75  E76  E85  E87  E88  E89  E90  E91  E92  E93  E95 E102 E105
   4    3    4    3    3    4    4    4    5    3    4    4    4    4    5    4    3    4    3    3    4    5    4    5
    Media Mediana C.Ind
[1,] 3.875       4    24
```

Variable: *COPT*

```
> clusterh(4,40,DATA,43,"Single",6,1,3)
[1] "Dist. Euclidena"
[1] "Enlace de Ward"
[1] "C. Cof"
[1] 0.4567209
[1] "Grupo 1"
  E1   E9  E10  E13  E15  E16  E18  E19  E22  E24  E26  E33  E36  E38  E40  E41  E43  E44  E48  E49  E50  E51  E52  E53  E54  E55
   3    5    4    5    4    3    3    5    4    4    5    4    4    3    5    1    4    4    3    3    5    4    4    2    5    3
 E56  E57  E58  E59  E60  E61  E62  E63  E64  E65  E66  E68  E70  E72  E73  E77  E83  E84  E86  E94  E96  E97  E98  E99 E100 E101
   2    4    4    1    4    4    4    3    4    3    5    4    4    4    3    3    5    4    4    3    3    2    3    4    3    4
E103 E106 E107 E108 E110 E111 E112 E113 E117 E118 E119 E120 E121
   3    3    2    3    5    4    3    4    5    4    4    3    2
      Media Mediana C.Ind
[1,] 3.615385       4    65
[1] "Grupo 2"
  E2   E3   E4   E5   E6   E7   E8  E11  E12  E14  E20  E21  E23  E25  E27  E28  E29  E30  E32  E34  E67  E74  E78  E79  E80  E81
   4    5    5    4    5    3    4    5    5    3    3    5    2    3    3    3    5    4    5    3    1    4    2    4    3    3
 E82 E104 E109 E114 E115 E116 E122
   3    5    4    4    3    5    4
      Media Mediana C.Ind
[1,] 3.757576       4    33
[1] "Grupo 3"
 E17  E31  E35  E37  E39  E42  E45  E46  E47  E69  E71  E75  E76  E85  E87  E88  E89  E90  E91  E92  E93  E95 E102 E105
   4    1    3    3    4    4    4    5    3    3    5    4    3    3    4    4    4    4    3    3    4    3    4    4
      Media Mediana C.Ind
[1,] 3.583333       4    24
```

Variable*: RSCCEP.*

```
> clusterh(4,40,DATA,44,"Single",6,1,3)
[1] "Dist. Euclidena"
[1] "Enlace de Ward"
[1] "C. Cof"
[1] 0.4567209
[1] "Grupo 1"
  E1   E9  E10  E13  E15  E16  E18  E19  E22  E24  E26  E33  E36  E38  E40  E41  E43  E44  E48  E49  E50  E51  E52  E53  E54  E55
   4    4    5    4    5    4    4    5    4    2    5    4    5    4    5    2    4    4    5    3    5    5    4    3    5    3
 E56  E57  E58  E59  E60  E61  E62  E63  E64  E65  E66  E68  E70  E72  E73  E77  E83  E84  E86  E94  E96  E97  E98  E99 E100 E101
   4    3    5    1    5    4    5    4    5    4    5    5    5    5    3    4    5    5    4    4    2    5    3    5    4    5
E103 E106 E107 E108 E110 E111 E112 E113 E117 E118 E119 E120 E121
   5    5    3    4    5    4    3    5    5    5    4    5    3
      Media Mediana C.Ind
[1,] 4.184615       4    65
[1] "Grupo 2"
  E2   E3   E4   E5   E6   E7   E8  E11  E12  E14  E20  E21  E23  E25  E27  E28  E29  E30  E32  E34  E67  E74  E78  E79  E80  E81
   5    5    5    5    5    4    4    5    5    5    4    5    3    3    4    3    5    4    5    5    1    5    4    5    5    5
 E82 E104 E109 E114 E115 E116 E122
   4    5    5    5    4    5    5
      Media Mediana C.Ind
[1,] 4.454545       5    33
[1] "Grupo 3"
 E17  E31  E35  E37  E39  E42  E45  E46  E47  E69  E71  E75  E76  E85  E87  E88  E89  E90  E91  E92  E93  E95 E102 E105
   4    3    5    4    4    4    4    5    4    3    5    5    4    4    5    4    4    4    3    3    4    5    4    5
    Media Mediana C.Ind
[1,] 4.125       4    24
```

Variable: *CCOE.*

```
> clusterh(4,40,DATA,45,"Single",6,1,3)
[1] "Dist. Euclidena"
[1] "Enlace de Ward"
[1] "C. Cof"
[1] 0.4567209
[1] "Grupo 1"
  E1   E9  E10  E13  E15  E16  E18  E19  E22  E24  E26  E33  E36  E38  E40  E41  E43  E44  E48  E49  E50  E51  E52  E53  E54  E55
   4    4    4    3    4    3    3    4    4    3    5    3    5    4    5    2    4    5    4    3    5    5    5    3    4    4
 E56  E57  E58  E59  E60  E61  E62  E63  E64  E65  E66  E68  E70  E72  E73  E77  E83  E84  E86  E94  E96  E97  E98  E99 E100 E101
   4    4    4    1    4    4    4    5    5    5    5    4    4    5    3    4    4    4    5    5    2    5    4    4    3    2
E103 E106 E107 E108 E110 E111 E112 E113 E117 E118 E119 E120 E121
   5    4    2    3    5    4    4    5    5    3    4    5    4
      Media Mediana C.Ind
[1,] 3.969231       4    65
[1] "Grupo 2"
  E2   E3   E4   E5   E6   E7   E8  E11  E12  E14  E20  E21  E23  E25  E27  E28  E29  E30  E32  E34  E67  E74  E78  E79  E80  E81
   2    5    5    5    5    5    5    5    5    5    3    5    4    4    4    2    3    4    4    5    3    2    5    4    4    4
 E82 E104 E109 E114 E115 E116 E122
   2    5    4    5    2    5    2
     Media Mediana C.Ind
[1,]     4       4    33
[1] "Grupo 3"
 E17  E31  E35  E37  E39  E42  E45  E46  E47  E69  E71  E75  E76  E85  E87  E88  E89  E90  E91  E92  E93  E95 E102 E105
   4    3    2    4    3    4    4    4    4    2    3    5    3    3    5    4    5    3    2    4    4    4    4    5
      Media Mediana C.Ind
[1,] 3.666667       4    24
```

Variable: *CCSI*

```
> clusterh(4,40,DATA,46,"Single",6,1,3)
[1] "Dist. Euclidena"
[1] "Enlace de Ward"
[1] "C. Cof"
[1] 0.4567209
[1] "Grupo 1"
  E1   E9  E10  E13  E15  E16  E18  E19  E22  E24  E26  E33  E36  E38  E40  E41  E43  E44  E48  E49  E50  E51  E52  E53  E54  E55
   3    2    3    1    2    1    4    3    4    1    2    1    3    3    2    4    2    3    2    2    2    3    3    2    3    3
 E56  E57  E58  E59  E60  E61  E62  E63  E64  E65  E66  E68  E70  E72  E73  E77  E83  E84  E86  E94  E96  E97  E98  E99 E100 E101
   4    3    3    1    2    1    3    3    5    3    3    2    1    4    1    3    2    5    1    3    4    2    2    4    1    4
E103 E106 E107 E108 E110 E111 E112 E113 E117 E118 E119 E120 E121
   3    3    1    3    4    4    3    3    3    3    4    2    4
      Media Mediana C.Ind
[1,] 2.676923       3    65
[1] "Grupo 2"
  E2   E3   E4   E5   E6   E7   E8  E11  E12  E14  E20  E21  E23  E25  E27  E28  E29  E30  E32  E34  E67  E74  E78  E79  E80  E81
   4    2    5    3    1    2    4    1    2    4    2    3    4    3    4    3    1    4    4    3    1    1    4    5    4    2
 E82 E104 E109 E114 E115 E116 E122
   2    2    3    3    2    4    2
      Media Mediana C.Ind
[1,] 2.848485       3    33
[1] "Grupo 3"
 E17  E31  E35  E37  E39  E42  E45  E46  E47  E69  E71  E75  E76  E85  E87  E88  E89  E90  E91  E92  E93  E95 E102 E105
   4    3    2    2    2    4    4    3    3    2    2    4    3    3    2    2    4    3    3    3    3    4    3    3
      Media Mediana C.Ind
[1,] 2.958333       3    24
```

Variable: *CI*

```
> clusterh(4,40,DATA,47,"Single",6,1,3)
[1] "Dist. Euclidena"
[1] "Enlace de Ward"
[1] "C. Cof"
[1] 0.4567209
[1] "Grupo 1"
  E1   E9  E10  E13  E15  E16  E18  E19  E22  E24  E26  E33  E36  E38  E40  E41  E43  E44  E48  E49  E50  E51  E52  E53  E54  E55
   4    5    4    4    3    5    4    4    3    2    4    1    5    3    4    2    4    5    3    3    4    3    3    2    4    4
 E56  E57  E58  E59  E60  E61  E62  E63  E64  E65  E66  E68  E70  E72  E73  E77  E83  E84  E86  E94  E96  E97  E98  E99 E100 E101
   5    4    5    1    3    3    4    5    5    4    4    3    5    4    3    3    5    5    5    4    3    4    4    3    4    3
E103 E106 E107 E108 E110 E111 E112 E113 E117 E118 E119 E120 E121
   3    4    3    3    5    4    4    4    4    4    4    5    3
      Media Mediana C.Ind
[1,] 3.738462       4    65
[1] "Grupo 2"
  E2   E3   E4   E5   E6   E7   E8  E11  E12  E14  E20  E21  E23  E25  E27  E28  E29  E30  E32  E34  E67  E74  E78  E79  E80  E81
   4    5    4    5    5    5    4    2    5    4    3    5    3    3    3    2    3    4    5    4    1    2    3    5    4    3
 E82 E104 E109 E114 E115 E116 E122
   3    4    2    5    3    4    3
      Media Mediana C.Ind
[1,] 3.636364       4    33
[1] "Grupo 3"
 E17  E31  E35  E37  E39  E42  E45  E46  E47  E69  E71  E75  E76  E85  E87  E88  E89  E90  E91  E92  E93  E95 E102 E105
   4    2    3    3    4    4    4    5    4    4    4    4    2    4    5    3    4    4    3    3    3    5    4    4
      Media Mediana C.Ind
[1,] 3.708333       4    24
```

Variable: *CAAP*

```
> clusterh(4,40,DATA,48,"Single",6,1,3)
[1] "Dist. Euclidena"
[1] "Enlace de Ward"
[1] "C. Cof"
[1] 0.4567209
[1] "Grupo 1"
  E1   E9  E10  E13  E15  E16  E18  E19  E22  E24  E26  E33  E36  E38  E40  E41  E43  E44  E48  E49  E50  E51  E52  E53  E54  E55
   4    5    3    3    3    5    3    4    3    3    5    5    4    3    3    1    4    3    3    3    5    3    3    3    4    4
 E56  E57  E58  E59  E60  E61  E62  E63  E64  E65  E66  E68  E70  E72  E73  E77  E83  E84  E86  E94  E96  E97  E98  E99 E100 E101
   5    3    5    1    3    4    4    3    5    4    4    3    4    5    2    3    4    5    3    4    5    5    3    4    3    4
E103 E106 E107 E108 E110 E111 E112 E113 E117 E118 E119 E120 E121
   4    5    3    4    5    4    3    4    4    3    4    3    3
      Media Mediana C.Ind
[1,] 3.676923       4    65
[1] "Grupo 2"
  E2   E3   E4   E5   E6   E7   E8  E11  E12  E14  E20  E21  E23  E25  E27  E28  E29  E30  E32  E34  E67  E74  E78  E79  E80  E81
   4    5    5    4    5    4    4    3    5    4    3    5    4    3    3    3    3    4    5    5    1    3    4    4    5    4
 E82 E104 E109 E114 E115 E116 E122
   4    4    3    5    4    5    4
      Media Mediana C.Ind
[1,] 3.969697       4    33
[1] "Grupo 3"
 E17  E31  E35  E37  E39  E42  E45  E46  E47  E69  E71  E75  E76  E85  E87  E88  E89  E90  E91  E92  E93  E95 E102 E105
   4    2    3    3    3    4    4    5    4    5    4    4    3    4    4    3    4    4    2    3    3    5    4    5
      Media Mediana C.Ind
[1,] 3.708333       4    24
```

Variable: *HBPAI*

```
> clusterh(4,40,DATA,49,"Single",6,1,3)
[1] "Dist. Euclidena"
[1] "Enlace de Ward"
[1] "C. Cof"
[1] 0.4567209
[1] "Grupo 1"
  E1   E9  E10  E13  E15  E16  E18  E19  E22  E24  E26  E33  E36  E38  E40  E41  E43  E44  E48  E49  E50  E51  E52  E53  E54  E55
   4    4    3    2    3    5    4    4    3    3    5    4    4    4    4    1    3    5    2    3    4    4    3    3    4    4
 E56  E57  E58  E59  E60  E61  E62  E63  E64  E65  E66  E68  E70  E72  E73  E77  E83  E84  E86  E94  E96  E97  E98  E99 E100 E101
   5    3    5    1    4    4    4    4    5    3    4    4    4    5    3    3    4    5    3    4    4    4    3    5    4    4
E103 E106 E107 E108 E110 E111 E112 E113 E117 E118 E119 E120 E121
   5    5    3    3    5    4    4    5    4    3    4    3    2
      Media Mediana C.Ind
[1,] 3.738462       4    65
[1] "Grupo 2"
  E2   E3   E4   E5   E6   E7   E8  E11  E12  E14  E20  E21  E23  E25  E27  E28  E29  E30  E32  E34  E67  E74  E78  E79  E80  E81
   3    5    5    4    5    5    4    3    5    4    3    5    4    3    3    2    3    3    4    3    1    2    4    4    5    3
 E82 E104 E109 E114 E115 E116 E122
   4    5    4    5    4    4    3
      Media Mediana C.Ind
[1,] 3.757576       4    33
[1] "Grupo 3"
 E17  E31  E35  E37  E39  E42  E45  E46  E47  E69  E71  E75  E76  E85  E87  E88  E89  E90  E91  E92  E93  E95 E102 E105
   4    3    2    3    3    4    4    4    4    2    4    4    3    3    5    3    3    3    2    3    4    4    4    5
      Media Mediana C.Ind
[1,] 3.458333     3.5    24
```

Variable: *CCA*

```
> clusterh(4,40,DATA,50,"Single",6,1,3)
[1] "Dist. Euclidena"
[1] "Enlace de Ward"
[1] "C. Cof"
[1] 0.4567209
[1] "Grupo 1"
  E1   E9  E10  E13  E15  E16  E18  E19  E22  E24  E26  E33  E36  E38  E40  E41  E43  E44  E48  E49  E50  E51  E52  E53  E54  E55
   4    5    5    5    4    5    3    5    4    4    5    4    5    4    3    2    4    4    4    4    5    5    3    4    5    5
 E56  E57  E58  E59  E60  E61  E62  E63  E64  E65  E66  E68  E70  E72  E73  E77  E83  E84  E86  E94  E96  E97  E98  E99 E100 E101
   5    3    5    1    4    4    4    5    5    4    4    5    5    5    4    4    4    4    4    3    3    5    3    5    4    3
E103 E106 E107 E108 E110 E111 E112 E113 E117 E118 E119 E120 E121
   3    4    4    3    5    4    3    5    5    4    4    5    3
      Media Mediana C.Ind
[1,] 4.107692       4    65
[1] "Grupo 2"
  E2   E3   E4   E5   E6   E7   E8  E11  E12  E14  E20  E21  E23  E25  E27  E28  E29  E30  E32  E34  E67  E74  E78  E79  E80  E81
   3    5    5    5    5    5    4    5    5    4    4    4    3    4    4    4    2    3    4    5    1    4    5    4    4    4
 E82 E104 E109 E114 E115 E116 E122
   3    5    3    5    4    5    5
      Media Mediana C.Ind
[1,] 4.090909       4    33
[1] "Grupo 3"
 E17  E31  E35  E37  E39  E42  E45  E46  E47  E69  E71  E75  E76  E85  E87  E88  E89  E90  E91  E92  E93  E95 E102 E105
   4    2    5    4    4    4    4    5    4    5    5    4    2    4    4    4    3    3    2    3    5    4    4    5
     Media Mediana C.Ind
[1,] 3.875       4    24
```

Variable: *CC*

```
> clusterh(4,40,DATA,51,"Single",6,1,3)
[1] "Dist. Euclidena"
[1] "Enlace de Ward"
[1] "C. Cof"
[1] 0.4567209
[1] "Grupo 1"
  E1   E9  E10  E13  E15  E16  E18  E19  E22  E24  E26  E33  E36  E38  E40  E41  E43  E44  E48  E49  E50  E51  E52  E53  E54  E55
   4    4    4    3    4    3    3    3    4    3    5    3    5    3    5    2    3    5    5    4    5    5    4    2    4    4
 E56  E57  E58  E59  E60  E61  E62  E63  E64  E65  E66  E68  E70  E72  E73  E77  E83  E84  E86  E94  E96  E97  E98  E99 E100 E101
   5    4    5    1    4    4    4    4    5    3    5    3    5    5    4    4    4    5    4    3    3    5    3    5    4    3
E103 E106 E107 E108 E110 E111 E112 E113 E117 E118 E119 E120 E121
   4    4    4    3    5    4    3    5    5    5    4    5    3
      Media Mediana C.Ind
[1,] 3.953846       4    65
[1] "Grupo 2"
  E2   E3   E4   E5   E6   E7   E8  E11  E12  E14  E20  E21  E23  E25  E27  E28  E29  E30  E32  E34  E67  E74  E78  E79  E80  E81
   3    4    5    4    5    5    4    3    5    4    3    4    4    2    3    3    3    4    4    4    1    3    5    4    3    3
 E82 E104 E109 E114 E115 E116 E122
   3    4    4    5    5    4    4
      Media Mediana C.Ind
[1,] 3.757576       4    33
[1] "Grupo 3"
 E17  E31  E35  E37  E39  E42  E45  E46  E47  E69  E71  E75  E76  E85  E87  E88  E89  E90  E91  E92  E93  E95 E102 E105
   4    3    2    4    3    4    4    5    5    3    5    5    2    5    4    4    4    3    1    3    4    4    4    5
     Media Mediana C.Ind
[1,]  3.75       4    24
```

Variable: *CIPRP*

```
> clusterh(4,40,DATA,52,"Single",6,1,3)
[1] "Dist. Euclidena"
[1] "Enlace de Ward"
[1] "C. Cof"
[1] 0.4567209
[1] "Grupo 1"
  E1   E9  E10  E13  E15  E16  E18  E19  E22  E24  E26  E33  E36  E38  E40  E41  E43  E44  E48  E49  E50  E51  E52  E53  E54  E55
   4    4    4    4    5    3    4    4    4    2    5    3    4    4    3    2    4    3    4    3    4    4    4    2    5    4
 E56  E57  E58  E59  E60  E61  E62  E63  E64  E65  E66  E68  E70  E72  E73  E77  E83  E84  E86  E94  E96  E97  E98  E99 E100 E101
   3    3    5    1    5    3    3    4    5    3    4    4    5    5    3    3    4    5    5    3    4    4    3    5    4    5
E103 E106 E107 E108 E110 E111 E112 E113 E117 E118 E119 E120 E121
   5    4    4    3    5    4    3    5    4    5    4    3    2
      Media Mediana C.Ind
[1,] 3.815385       4    65
[1] "Grupo 2"
  E2   E3   E4   E5   E6   E7   E8  E11  E12  E14  E20  E21  E23  E25  E27  E28  E29  E30  E32  E34  E67  E74  E78  E79  E80  E81
   4    5    5    4    5    5    4    4    5    4    3    5    4    4    4    3    4    4    4    4    1    3    4    4    4    4
 E82 E104 E109 E114 E115 E116 E122
   4    3    5    5    5    5    4
      Media Mediana C.Ind
[1,] 4.090909       4    33
[1] "Grupo 3"
 E17  E31  E35  E37  E39  E42  E45  E46  E47  E69  E71  E75  E76  E85  E87  E88  E89  E90  E91  E92  E93  E95 E102 E105
   4    3    3    4    4    4    4    5    4    5    4    4    3    4    4    4    3    3    3    3    4    3    4    4
     Media Mediana C.Ind
[1,]  3.75       4    24
```

Variable: *CTD*

```
> clusterh(4,40,DATA,53,"Single",6,1,3)
[1] "Dist. Euclidena"
[1] "Enlace de Ward"
[1] "C. Cof"
[1] 0.4567209
[1] "Grupo 1"
  E1   E9  E10  E13  E15  E16  E18  E19  E22  E24  E26  E33  E36  E38  E40  E41  E43  E44  E48  E49  E50  E51  E52  E53  E54  E55
   4    5    4    4    5    5    3    4    4    4    5    4    5    4    2    1    4    3    5    3    5    4    4    3    5    4
 E56  E57  E58  E59  E60  E61  E62  E63  E64  E65  E66  E68  E70  E72  E73  E77  E83  E84  E86  E94  E96  E97  E98  E99 E100 E101
   3    4    5    1    5    3    4    4    5    4    4    4    4    5    2    4    5    5    5    4    5    4    3    4    3    5
E103 E106 E107 E108 E110 E111 E112 E113 E117 E118 E119 E120 E121
   5    4    4    3    5    4    3    4    5    4    4    5    3
     Media Mediana C.Ind
[1,]     4       4    65
[1] "Grupo 2"
  E2   E3   E4   E5   E6   E7   E8  E11  E12  E14  E20  E21  E23  E25  E27  E28  E29  E30  E32  E34  E67  E74  E78  E79  E80  E81
   5    5    5    5    5    4    4    4    5    4    3    5    3    2    4    3    3    4    5    5    1    3    2    4    4    3
 E82 E104 E109 E114 E115 E116 E122
   4    4    5    5    4    5    4
       Media Mediana C.Ind
[1,] 3.969697       4    33
[1] "Grupo 3"
 E17  E31  E35  E37  E39  E42  E45  E46  E47  E69  E71  E75  E76  E85  E87  E88  E89  E90  E91  E92  E93  E95 E102 E105
   3    4    4    3    3    4    4    5    4    5    4    4    3    4    5    4    4    3    2    5    4    5    4    4
       Media Mediana C.Ind
[1,] 3.916667       4    24
```

Variable: *CTE*

```
> clusterh(4,40,DATA,54,"Single",6,1,3)
[1] "Dist. Euclidena"
[1] "Enlace de Ward"
[1] "C. Cof"
[1] 0.4567209
[1] "Grupo 1"
  E1   E9  E10  E13  E15  E16  E18  E19  E22  E24  E26  E33  E36  E38  E40  E41  E43  E44  E48  E49  E50  E51  E52  E53  E54  E55
   5    5    5    5    5    5    3    5    4    4    5    4    4    4    4    1    4    4    3    3    5    3    4    3    5    4
 E56  E57  E58  E59  E60  E61  E62  E63  E64  E65  E66  E68  E70  E72  E73  E77  E83  E84  E86  E94  E96  E97  E98  E99 E100 E101
   4    3    5    1    5    3    4    3    5    4    5    3    5    5    4    3    5    5    4    4    4    3    3    5    5    4
E103 E106 E107 E108 E110 E111 E112 E113 E117 E118 E119 E120 E121
   5    4    3    3    5    4    3    5    5    4    4    5    4
       Media Mediana C.Ind
[1,] 4.061538       4    65
[1] "Grupo 2"
  E2   E3   E4   E5   E6   E7   E8  E11  E12  E14  E20  E21  E23  E25  E27  E28  E29  E30  E32  E34  E67  E74  E78  E79  E80  E81
   5    5    5    5    5    5    4    5    5    5    4    5    3    4    3    4    2    4    5    5    1    4    4    4    5    4
 E82 E104 E109 E114 E115 E116 E122
   5    5    4    5    4    5    4
      Media Mediana C.Ind
[1,] 4.30303       5    33
[1] "Grupo 3"
 E17  E31  E35  E37  E39  E42  E45  E46  E47  E69  E71  E75  E76  E85  E87  E88  E89  E90  E91  E92  E93  E95 E102 E105
   4    3    4    4    5    4    4    5    4    3    3    4    3    3    4    4    5    3    3    5    3    5    5    5
       Media Mediana C.Ind
[1,] 3.958333       4    24
```

Variable: *CMCMC*

```
> clusterh(4,40,DATA,55,"Single",6,1,3)
[1] "Dist. Euclidena"
[1] "Enlace de Ward"
[1] "C. Cof"
[1] 0.4567209
[1] "Grupo 1"
  E1   E9  E10  E13  E15  E16  E18  E19  E22  E24  E26  E33  E36  E38  E40  E41  E43  E44  E48  E49  E50  E51  E52  E53  E54  E55
   5    5    4    4    3    5    4    5    4    2    5    4    4    4    3    1    4    3    4    3    5    3    4    3    5    4
 E56  E57  E58  E59  E60  E61  E62  E63  E64  E65  E66  E68  E70  E72  E73  E77  E83  E84  E86  E94  E96  E97  E98  E99 E100 E101
   2    4    5    1    4    3    3    3    5    3    5    3    4    5    3    3    5    4    4    3    4    4    2    5    4    4
E103 E106 E107 E108 E110 E111 E112 E113 E117 E118 E119 E120 E121
   5    4    3    3    5    5    3    5    5    4    4    5    3
      Media Mediana C.Ind
[1,] 3.830769       4    65
[1] "Grupo 2"
  E2   E3   E4   E5   E6   E7   E8  E11  E12  E14  E20  E21  E23  E25  E27  E28  E29  E30  E32  E34  E67  E74  E78  E79  E80  E81
   3    5    5    4    5    5    4    5    5    5    3    5    3    3    3    2    2    4    5    4    1    4    4    3    4    4
 E82 E104 E109 E114 E115 E116 E122
   4    4    3    5    4    4    2
      Media Mediana C.Ind
[1,] 3.818182       4    33
[1] "Grupo 3"
 E17  E31  E35  E37  E39  E42  E45  E46  E47  E69  E71  E75  E76  E85  E87  E88  E89  E90  E91  E92  E93  E95 E102 E105
   4    3    4    3    4    4    3    5    4    3    4    4    3    4    4    3    4    4    3    5    3    4    4    5
      Media Mediana C.Ind
[1,] 3.791667       4    24
```

Variable: *HTA*

```
> clusterh(4,40,DATA,55,"Single",6,1,3)
[1] "Dist. Euclidena"
[1] "Enlace de Ward"
[1] "C. Cof"
[1] 0.4567209
[1] "Grupo 1"
  E1   E9  E10  E13  E15  E16  E18  E19  E22  E24  E26  E33  E36  E38  E40  E41  E43  E44  E48  E49  E50  E51  E52  E53  E54  E55
   5    5    4    4    3    5    4    5    4    2    5    4    4    4    3    1    4    3    4    3    5    3    4    3    5    4
 E56  E57  E58  E59  E60  E61  E62  E63  E64  E65  E66  E68  E70  E72  E73  E77  E83  E84  E86  E94  E96  E97  E98  E99 E100 E101
   2    4    5    1    4    3    3    3    5    3    5    3    4    5    3    3    5    4    4    3    4    4    2    5    4    4
E103 E106 E107 E108 E110 E111 E112 E113 E117 E118 E119 E120 E121
   5    4    3    3    5    5    3    5    5    4    4    5    3
      Media Mediana C.Ind
[1,] 3.830769       4    65
[1] "Grupo 2"
  E2   E3   E4   E5   E6   E7   E8  E11  E12  E14  E20  E21  E23  E25  E27  E28  E29  E30  E32  E34  E67  E74  E78  E79  E80  E81
   3    5    5    4    5    5    4    5    5    5    3    5    3    3    3    2    2    4    5    4    1    4    4    3    4    4
 E82 E104 E109 E114 E115 E116 E122
   4    4    3    5    4    4    2
      Media Mediana C.Ind
[1,] 3.818182       4    33
[1] "Grupo 3"
 E17  E31  E35  E37  E39  E42  E45  E46  E47  E69  E71  E75  E76  E85  E87  E88  E89  E90  E91  E92  E93  E95 E102 E105
   4    3    4    3    4    4    3    5    4    3    4    4    3    4    4    3    4    4    3    5    3    4    4    5
      Media Mediana C.Ind
[1,] 3.791667       4    24
```

Variable: *CCL*

```
> clusterh(4,40,DATA,57,"Single",6,1,3)
[1] "Dist. Euclidena"
[1] "Enlace de Ward"
[1] "C. Cof"
[1] 0.4567209
[1] "Grupo 1"
  E1   E9  E10  E13  E15  E16  E18  E19  E22  E24  E26  E33  E36  E38  E40  E41  E43  E44  E48  E49  E50  E51  E52  E53  E54  E55
   5    5    5    4    5    4    3    5    4    4    5    5    5    4    5    1    4    4    5    3    5    4    5    4    5    4
 E56  E57  E58  E59  E60  E61  E62  E63  E64  E65  E66  E68  E70  E72  E73  E77  E83  E84  E86  E94  E96  E97  E98  E99 E100 E101
   2    4    5    1    5    5    4    4    5    4    4    5    5    5    3    4    5    5    4    5    5    3    3    5    5    5
E103 E106 E107 E108 E110 E111 E112 E113 E117 E118 E119 E120 E121
   5    4    4    4    5    4    3    5    5    5    4    5    3
       Media Mediana C.Ind
[1,] 4.276923       5    65
[1] "Grupo 2"
  E2   E3   E4   E5   E6   E7   E8  E11  E12  E14  E20  E21  E23  E25  E27  E28  E29  E30  E32  E34  E67  E74  E78  E79  E80  E81
   4    5    5    5    5    5    4    5    5    5    3    5    4    4    4    3    5    3    4    5    1    4    5    4    5    4
 E82 E104 E109 E114 E115 E116 E122
   5    5    3    5    5    5    5
       Media Mediana C.Ind
[1,] 4.363636       5    33
[1] "Grupo 3"
 E17  E31  E35  E37  E39  E42  E45  E46  E47  E69  E71  E75  E76  E85  E87  E88  E89  E90  E91  E92  E93  E95 E102 E105
   4    4    5    4    5    4    4    5    4    4    5    4    3    4    5    4    5    5    3    5    4    5    4    5
       Media Mediana C.Ind
[1,] 4.333333       4    24
```