Universidad Miguel Hernández

Facultad de Ciencias Sociales y Jurídicas de Elche

Grado en Estadistica Empresarial

Trabajo Fin de Grado

Curso 2017/2018

ANALISIS CLUSTER APLICADO AL AMBITO AGRONOMO PARA ESTUDIAR EL COMPORTAMIENTO DE NUEVOS TIPOS DE COMPOST

Tutor: Xavier Barber
Julio Alberto Ibáñez Perea

Índice

1. Introducción **3**
1.1. Extracción de la información 4

2. Objetivos **5**

3. Material y métodos **6**

4. Análisis estadístico **8**
4.1. Análisis clúster 8
4.1.1. Etapas básicas para el análisis Cluster 8
4.1.2. Clasificación de tipos de análisis Clúster 9
4.1.3. Métodos clúster jerárquicos 11
4.1.4. Metodos clúster no jerarquicos 12
4.2. Medidas de distancias y similaridad 15
4.2.1. Definiciones de distancias y proximidad 15
4.2.2. Medidas de distancias 16
4.2.3. Medidas de similaridad o proximidad 20

5. Resultados **22**
5.1. Análisis descriptivo 23
5.2. Análisis comparativo 28
5.2.1. Análisis clúster para variables 28
5.2.2. Análisis clúster para las observaciones 35

6. Conclusiones **54**

Índice de figuras

1. Barreno (instrumento para la extracción de la muestra de raices.) 4
2. Ejemplo de dendograma 12
3. Estructura de los datos 22
4. Cuadro de mínimos, máximos y medianas 23
5. Gráfico de correlaciones 24
6. Gráfico para la longitud de raíz (suma) por planta 25
7. Gráfico para el diámetro promedio de la suma de raíces por planta 25
8. Gráfico para las bifurcaciones por planta, indica el grado de ramificación 25
9. Gráfico para concentración hidrosoluble del elemento Na (sodio) 26
10. Gráfico para concentración hidrosoluble del elemento K (potasio) 26
11. Gráfico para concentración hidrosoluble del elemento P (fósforo) 26
12. Datos tipificados 29
13. Dendograma para las variables 30
14. Dendograma para las variables agrupadas en 8 clusters 31
15. Dendograma para las variables agrupadas en 7 clusters 31
16. Dendograma para las variables agrupadas en 6 clusters tras análisis jerárquico 32
17. Composición de los clúster para las variables tras análisis K-MEANS 33
18. Estructura de los datos una vez ajustados 35
19. Dendograma para las variables tras el análisis jerárquico 36
20. Dendograma para las variables agrupadas en 5 clusters tras el análisis jerárquico 37
21. Gráfico de K-MEANS para todas las observaciones agrupadas en 5 clusters 38
22. Gráfico de K-MEANS para todas las observaciones agrupadas en 3 clusters 39
23. Composición de los clusters según el tipo de planta 39
24. Composición de los clusters según el porcentaje aplicado de compost 40
25. Tabla de medias y desviaciones tipicas para todas las observaciones 41
26. Composición de los clusters para todas las variables 44
27. Gráfico K-MEANS para las observaciones de lechuga, brócoli y escarola agrupadas en 5 clusters 45
28. Gráfico K-MEANS para las observaciones de lechuga, brócoli y escarola agrupadas en 4 clusters 46
29. Gráfico K-MEANS para las observaciones de lechuga, brócoli y escarola agrupadas en 3 clusters 47
30. Tabla de medias y desviaciones típicas para las observaciones de lechuga, brócoli y escarola 48
31. Composición de clusters para las observaciones de lechuga, brócoli y escarola 49
32. Gráfico K-MEANS para las observaciones del pimiento 51
33. Tabla de medias y desviaciones típicas para las observaciones del pimiento 52
34. Composición de clusters para las observaciones del pimiento . . . 52

1. Introducción

En el presente trabajo se va a llevar a cabo un estudio estadístico basado en el análisis tanto del comportamiento de las raíces de diferentes tipos de cultivos, como la eficacia de diferentes sustratos elaborados a partir de nuevos tipos de compost de diversos orígenes. Con este análisis se pretende llegar a conclusiones de que tipo o tipos de compost han mostrado mejores resultados para las variables utilizadas para el escaneo de las raíces, es decir, que tipo o tipos de compost muestran una mayor calidad en el cultivo basándonos en las características radiculares de las plantas. Con esto se quiere estudiar si existe un medio de cultivo que pueda sustituir al medio tradicional más común hasta hoy utilizado, el cual es la turba. Juntando toda esta información del análisis, veremos qué tipo o tipos de compost pueden comportarse de una manera similar y que repercusión tiene esto en las propias raíces y en su composición química de los diferentes cultivos.

Como ya se ha introducido, las variables necesarias para observar la eficiencia de los diferentes sustratos estudiados como posibles sustititos son variables que provienen del comportamieto (rendimiento y características morfológicas) de las raíces. Por lo que la mayoría de información necesaria para el análisis queda oculta en el sistema radicular, cuya evaluación es difícil especialmente respecto al efecto positivo o negativo del sustrato sobre las plantas. Para la recogida de esta información entra en juego el uso de sistemas basados en instrumentación óptica para el posterior análisis de la imagen mediante programas especializados, que más adelante se detallara este procedimiento con más detalle.

Las raíces son la parte principal de las plantas, ya que es donde se desempeñan las funciones esenciales para el correcto crecimiento y sobrevivencia de estas. En concreto, estas funciones esenciales son: el anclaje en el sustrato y soporte estructural, la absorción del agua, minerales, compuestos orgánicos y otras sustancias del suelo, la generación de presión de raíz capaz de reparar la cavitación, el almacenamiento de agua, minerales, carbohidratos, etc. Para cada tipo de estas funciones se ha de generar una variable de estudio y poder cuantificar cada aspecto para su posterior análisis.

Para la recogida de información hay que destacar que, desde el punto de vista del diagnóstico, es importante contar con información obtenida de forma periódica y sistemática sobre la abundancia y la calidad de las raíces. Para obtener esta información con las mediciones de estas características sobre los sistemas radiculares de las plantas se ha utilizado un sistema de análisis computacional llamado WinRhizo, el cual es un paquete informático diseñado para el análisis de la abundancia y la distribución de las raíces en el suelo, extraídas por métodos destructivos de muestreo sistemático y lavadas antes de su procesamiento. Este es un sistema computacional de imágenes, diseñado para realizar mediciones de las raíces de las plantas en términos de su morfología, topología, arquitectura, color, presencia de relaciones simbióticas, ramificación, densidad y estado sanitario, entre otros. Efectivamente estas son las características que se necesitan en este estudio para poder evaluar la calidad de raíces, y de este modo poder concluir sobre qué tipo o tipos de compost proporciona mejores cualidades a las plantas en relación con las características de estas.

1.1. Extracción de la información

Ahora se va a exponer el procedimiento de muestreo llevado a cabo en los diferentes cultivos para la recogida de información. Para esta recogida de información se va a utilizar un muestreo sistemático, el cual es un componente del diseño experimental implementado a menudo en los estudios de las raíces, y su complejidad depende en gran medida del tipo de suelo o sustrato, y de la edad y el tamaño del sistema radicular bajo estudio.

Como ya se ha introducido, la información necesaria proviene de las raíces sumergidas en la tierra, por lo que las muestras de raíces se obtendrán con una herramienta llamada barreno (Figura 1). Se recomiendo realizar muestreos secuenciales a lo largo del perfil del suelo, que reflejen la extensión y profundidad del sistema radicular de las plantas. Estas muestras de suelo extraídas con el barreno se mantienen frescas y en la oscuridad hasta su procesado en el lugar de análisis. Esto es necesario para no incurrir en errores de muestreo y poder llegar a conclusiones erróneas por modificación de las características de las raíces en la recogida de información. Una vez en las instalaciones, las muestras pueden ser lavadas en agua y dejadas en remojo para facilitar el desprendimiento de terrones y otras adherencias. Es conveniente establecer un tiempo de lavado y recolección de las raíces estándar para cada muestra, de esta forma se logra una mayor uniformidad en el procedimiento de estas.

Una vez recogida todas las muestras necesarias con todos los nuevos tipos de compost y para todas las plantas analizadas (en el punto siguiente se especificarán que tipo de compost y plantas han sido analizadas) entra en juego el programa informático anteriormente mencionado WinRhizo. Este paquete informático produce un resultado de parámetros morfo-fisiológicos de las raíces, el cual es esencial para poder diagnosticar su estado funcional y calidad de las mismas. Concretamente, este programa devuelve los siguientes parámetros: número, diámetro, longitud, superficie, volumen, densidad (cm/cm^3) y orden topológico de las raíces. Además del peso fresco y seco de las raíces y otras observaciones como el color, firmeza, flexibilidad, consistencia, porosidad, deciduosidad y longevidad.

Figura 1: Barreno (instrumento para la extracción de la muestra de raices.)

2. Objetivos

En este punto se van a exponer los diferentes objetivos del presente estudio. Por un lado, existe un objetivo principal, el cual está ligado al estudio del comportamiento de las raíces dependiendo el tipo de compost utilizado y así poder observar posibles sustitutos a métodos tradicionales utilizados. Y, por otro lado, tenemos un objetivo secundario más enfocado al análisis estadístico, el cual se basa en analizar que técnicas del análisis utilizado son más adecuadas al tipo de datos que obtenemos.

El objetivo principal de este estudio es conocer que tipo o tipos de compost han mostrado mejores resultados para las variables utilizadas y analizadas en el escaneo de raíces llevado a cabo. Para dar conclusiones sobre que compost proporciona mejores resultados, habrá que realizar un estudio comparativo entre las características morfológicas de las raíces de diferentes cultivos hortícolas (brócoli, escarola, lechuga y pimiento) germinados y propagados en medios alternativos al tradicional, frente a los desarrollados en medios tradicionales (turba). Este estudio comparativo se basará a nivel de estructura de raíz, como longitudes, diámetros y superficies obtenidos mediante escaneo radicular como método de estimación de la calidad de dichos medios. De este modo se podrá llegar a conclusiones sobre qué tipo de compost proporciona una mayor calidad y así poder ser un posible sustituto a los medios de cultivo tradicionales.

Además, se quiere analizar que patrón de agrupación muestran estos tipos de compost respecto a la composición fisicoquímica del mismo. De aquí nace el objetivo secundario de este estudio, el cual consiste en estudiar qué tipo de análisis estadístico utilizado será más conveniente para el tipo de datos que se recogen. Este tipo de análisis de agrupación se denomina análisis clúster (el cual será explicado con más detalle en puntos siguientes). Para llegar al fin de este objetivo se realizarán diferentes procedimientos existentes para desarrollar un análisis clúster, y de esta forma estudiar qué tipo de procedimiento llevado a cabo se ajusta mejor a los datos que se recogen como muestra. De esta forma se podrá llegar a conclusiones más acertadas para solventar nuestro objetivo principal.

3. Material y métodos

En este punto se va a especificar el material que se ha utilizado para poder llevar a cabo el estudio, y como en cualquier estudio estadístico, el material utilizado es la información recogida y de la que se dispone para llevar a cabo el estudio. También se especificará el método o métodos desarrollados para poder llegar a conclusiones claras y fiables.

Con el fin de poder extraer agrupaciones lógicas de las propiedades de los sustratos con relación a la morfología de las raíces de las plantas, se ha llevado a cabo una recogida de información en el proceso de germinación-propagación de las plantas hasta nivel comercial de trasplante. Esta información proviene de 49 métodos de cultivo distintos, ya que se ha analizado el proceso de germinación-propagación para la turba y 16 compost diferentes con su grado creciente de participación del compost en cada uno de ellos: 25, 50, 75 %. Estos compost analizados provienen de diferentes orígenes y procesos:

1. Compost ganaderos primarios provenientes de estiércoles:
   PL
   AP
   CIG1
2. Compost ganaderos secundarios provenientes de digeridos:
   Z1
   Z3
   Z4
   Z6
   Zmix
   C60
   C61
   C62
3. Compost obtenidos mediante un tratamiento de desalinización mediante lavado con agua en una proporción 1:1 y posterior agitación durante 30 minutos a partir de sus homónimos sin lavar:
   Zmix-Lav
   PL-Lav
   AP-Lav
   C61-Lav
   CIG1-Lav

Estos son todos los tipos de compost analizados a la hora de estudiar la agrupación en relación con la calidad que proporciona cada uno de ellos observando cómo afectan a la morfología de las raíces y plantas. Para hacer el estudio más amplio se han analizado las raíces de diferentes plantas: brócoli, escarola, lechuga y pimiento. Por lo que con cada combinación de planta, compost y porcentaje

aplicado se empieza el estudio con una base de datos de 264 observaciones, ya que para cada combinación se han realizado tres observaciones en momentos en el tiempo diferentes.

Como ya se ha introducido anteriormente, para obtener la información morfología radicular se ha utilizado el programa anteriormente expuesto WinRhizo. A través de este análisis se ha obtenido las características que serán nuestras variables de estudio, las cuales estan definidas en el Anexo 1.

Para llevar a cabo el análisis estadístico por el cual llegamos a conclusiones de agrupaciones óptimas para analizar la calidad de los diferentes sustratos se ha utilizado el programa estadístico R, el cual proporciona una serie de paquetes específicos para llevar a cabo, por un lado, un primer análisis descriptivo y, por otro lado, los diferentes análisis cluster que nos darán la información necesaria para poder llegar solventar nuestros objetivos.

4. Análisis estadístico

En este apartado se va a exponer toda la explicación y desarrollo teórico sobre el método estadístico utilizado, método clúster. De esta forma se entenderá mejor cómo se comporta este método a la hora de agrupar las observaciones de las que se disponen.

4.1. Análisis clúster

Este análisis clúster se ha desarrollado durante el presente siglo, el detonante para crear una investigación profunda sobre el mismo fue la publicación del libro Principios de Taxonomía Numérica, publicado en 1963 por dos biólogos (Sokal y Sneath). A partir de ese momento se crea un considerable desarrollo de dicho método. Este fuerte desarrollo se debe principalmente a dos factores importantes. Por un lado, el desarrollo computacional, y por otro, la importancia que tiene la clasificación como procedimiento científico.

El análisis clúster es una técnica estadística multivariante, que como su propio nombre indica, son técnicas que analizan estudios que contengan tres o más variables influyentes. Este método clúster también es llamado habitualmente como análisis de conglomerado o análisis de agrupamiento. Esta técnica busca agrupar un conjunto de entidades (observaciones y/o variables) en grupos con una relativa homogeneidad dentro de ellos, pero a la vez se busca que estos grupos sean lo más distantes (heterogéneos) entre ellos. Estos grupos o categorías son denominados clusters. Este análisis, como cualquier análisis estadístico, comienza con un conjunto de datos donde existen una serie de variables que explican las características de las observaciones que se obtengan. A partir de este conjunto de datos se busca reducir dimensiones formando diferentes categorías (clusters) donde incluir a cada una de ellas, un determinado número de observaciones (o variables) que tengan una semejanza en relación a su comportamiento. En este análisis, el investigador no conoce apenas información sobre la estructura de las categorías a formar, este motivo es el que lo diferencia del resto de métodos multivariantes. Tampoco se conoce el número de categorías en las que se va a dividir el conjunto de observaciones, por estas razones se pretende encontrar un conjunto de categorías a las que ir asignando las diferentes observaciones por algún criterio de homogeneidad. Por lo que es imprescindible definir una medida de similitud o divergencia para una clasificación lo más homogénea posible de los individuos (observaciones) dentro las diferentes categorías y lo más heterogénea entre ellas. Estas mediadas de similitud o distancia se definirán en el capítulo siguiente.

4.1.1. Etapas básicas para el análisis Cluster

En este punto se desarrollarán los pasos básicos que se deben seguir para un correcto análisis clúster.

Lo primero que se debe hacer para llevar a cabo correctamente el análisis clúster será realizar una correcta elección de las variables que van a describir a los diferentes individuos (observaciones). Una de las cuestiones importantes a

tener en cuenta sobre la elección de variables es ver si realmente son relevantes para el tipo de clasificación a la que se quiere llegar. Ya que, como se ha dicho anteriormente, el analista no tiene ninguna información sobre los posibles grupos a formar. Por lo que es conveniente saber de antemano que tipo de clasificación se quiere obtener como resultado e intentar recoger la información acorde a este.

En el segundo paso se llevará a cabo la elección del sistema de asociación entre individuos, es decir, una medida que permita medir la proximidad de los individuos. Generalmente esta medida de proximidad viene dada en términos de distancias, aunque también se pueden utilizar medidas de similitud. En el siguiente punto se desarrollarán con más detalle las medidas más utilizadas.

El tercer paso consistirá en seleccionar la técnica clúster adecuada para el estudio. Debido a la existencia de numerosos y diversos métodos, es importante elegir un método acorde con los datos que tenemos para realizar el estudio. Esta elección dependerá de la naturaleza de los datos y de los objetivos a los que se quiera llegar. En la práctica, es conveniente no ceñirse solamente a un solo método, sino realizar el estudio con varias posibilidades y de este modo poder contrastar los resultados y poder llegar a conclusiones más fiables.

Y, por último, habrá que validar e interpretar los resultados. Esta etapa podría ser da las más importantes, ya que será donde se expongan las conclusiones definitivas del análisis. Existen diferentes métodos para validar los resultados dependiendo de los métodos llegados a cabo (jerárquico o no jerárquico). Para validar los métodos jerárquicos se generan dos cuestiones importantes: por un lado, en qué medida representa la estructura obtenida las similitudes entre los individuos del análisis y, por otro lado, cuál sería el número ideal de clusters.

Para validar los métodos no jerárquicos, las cuestiones planteadas anteriormente no tienen tanto sentido, ya que por ejemplo en estos métodos el analista ya sabe de ante mano de cuantos clusters se compone el estudio. Por lo que la validación de estos métodos se basa en el análisis de la homogeneidad de los grupos que se han ido formando durante el desarrollo del método. Para ello algunos autores proponen utilizar múltiples análisis de la varianza (ANOVA) sobre cada variable dentro de cada clúster. Pero esta validación no debe ser considerada como definitiva, y habrá que reforzarla con otra técnica comúnmente utilizada, que consiste en tomar varias submuestras de la original e ir repitiendo el estudio sobre cada una de las submuestras.

4.1.2. Clasificación de tipos de análisis Clúster

Este método tiene diferentes variantes dependiendo del procedimiento a seguir. Aquí vamos a exponer solamente los diferentes tipos de análisis clúster que se pueden llevar a cabo, ya que en capítulos siguientes serán desarrollados con mayor detalle. De forma genérica existen dos categorías de métodos: métodos jerárquicos y no jerárquicos.

Los métodos jerárquicos tienen la principal función de agrupar diferentes clusters para formar un nuevo o bien separar uno ya existente para crear otros dos, estas agrupaciones o divisiones se basan en minimizar alguna función de

distancia o maximizar alguna medida de similitud. La característica principal de estos métodos es que cuando una unidad es asignada a un clúster es irrevocable, es decir, una vez que un individuo sea asignado a un determinado grupo no podrá ser asignado a cualquier otro. Estos métodos jerárquicos se pueden dividir en aglomerativos y disociativos. Los métodos aglomerativos se comienza el análisis con tantos clusters como individuos existan en el estudio. A partir de ahí se van agrupando los diferentes objetos, y esas agrupaciones iniciales se van fusionando de acuerdo a sus similitudes de forma ascendente, hasta que al final del proceso todas las observaciones pertenecen al mismo conglomerado. Dentro de este método jerárquico aglomerativo existen los siguientes métodos:

- Método de aglomeramiento simple.
- Método de aglomeramiento completo.
- Método del promedio entre grupos.
- Método del centroide.
- Método de la mediana.
- Método de Ward.

Los métodos jerárquicos disociativos, también llamados divisivos, llevan a cabo el proceso contrario al anterior. Comienza el estudio con un solo grupo que engloba a todas las observaciones, y a partir de ahí va realizando divisiones en grupos cada vez más pequeños, para acabar con tantos grupos como individuos existan en el estudio. Dentro de los métodos disociativos destacan los siguientes (además de los anteriores que siguen siendo válidos):

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

En el análisis clúster con métodos no jerárquicos, que también acogen el nombre de partitivos o de optimización, hay que elegir a priori el número de grupos en los que se quiere dividir la muestra. Por lo que tiene como objetivo realizar una sola partición de las observaciones en un determinado número de grupos. Es posible que esta razón sea la principal diferencia respecto a los métodos jerárquicos. La asignación de los individuos a los respectivos grupos se basa en maximizar la homogeneidad dentro de los grupos y la heterogeneidad entre los grupos. Pedret en 1986 agrupa estos métodos partitivos en cuatro familias:

- Métodos de reasignación:
  - Método de K-Medias.
  - El Quick-Cluster análisis.
  - Método de Forgy.
  - Método de las nubes dinámicas.
- Métodos de búsqueda de densidad:
  - Análisis modal de Wishart.
  - Método Taxmap.
  - Método de Fortin.

- Métodos directos:

  Método de Block-Clustering.

- Métodos de reducción de dimensiones:

  Análisis factorial tipo Q.

4.1.3. Métodos clúster jerárquicos

En este punto se explicarán con más detalle los métodos de clúster jerárquicos. Como ya se ha dicho anteriormente, son métodos que buscan tanto agrupar como dividir individuos de forma iterativa, buscando minimizar alguna medida de distancia, o bien, maximizar alguna medida de similitud. Existen dos tipos de métodos jerárquicos, que también se diferenciaran en este capítulo más detalladamente: aglomerativos y divisivos.

Estos métodos tienen una particularidad, y es que al ir formando grupos tanto ascendentemente (aglomerativos) como descendientemente (divisivos) estos quedan ya jerárquicamente agrupados para el resto del estudio, es decir, si por ejemplo un individuo se agrupa a un grupo formado anteriormente, ese grupo no podrá desprenderse de ningún individuo que ya le haya sido asignado. Estos métodos jerárquicos tienen una forma gráfica muy buena para ver el proceso que se ha ido llevando a cabo, este gráfico se denomina *dendograma*. En este gráfico se puede observar el procedimiento de unión o división seguido, es decir, los grupos que se han ido uniendo o dividiendo, en el momento que lo han hecho y también el valor de la medida de asociación en cada momento de unión o división, ver figura 2.

4.1.3.1. Métodos jerárquicos aglomerativos

Estos métodos aglomerativos también reciben el nombre de métodos ascendentes, ya que el análisis parte con tantos grupos como individuos existan en el estudio, y a partir de estos grupos se van formando grupos de forma ascendente. Al final del análisis se llega a tener todos los casos agrupados en un mismo grupo. Esta agrupación puede estar llevada a cabo a través de diferentes estrategias de agrupación. Estas diferentes estrategias son las siguientes:

- Estrategia de la distancia mínima o similitud máxima.
- Estrategia de la distancia, o similitud, promedio no ponderada.
- Estrategia de la distancia, o similitud, promedio ponderada.
- Métodos basados en el centroide.
- Método Ward.

Todas estas estrategias se utilizan para los métodos jerárquicos aglomerativos y proporcionan diversos criterios para determinar que grupos se deben unir.

4.1.3.2. Métodos jerárquicos divisivos

En la práctica, este método divisivo no es muy utilizado a la hora de realizar análisis clusters, ya que el aglomerativo es el más común. Este método jerárquico divisivo toma el camino contrario a los aglomerativos, ya que el estudio parte de un solo grupo donde se engloban a todos los casos del análisis. A partir de este grupo se van realizando sucesivas divisiones formando grupos cada vez más pequeños. Al final del análisis, el algoritmo agrupa a todos los casos en un mismo grupo.

En cuanto a los tipos de estrategias que se siguen en este método divisivo podemos acoger las mencionadas en el método aglomerativo, ya que seguimos buscando separar casos que maximicen su distancia (o minimicen su similitud). Y además de dichas estrategias se pueden adoptar las siguientes:

- Estrategias monotéticas, las cuales dividen los datos sobre la base de un solo atributo y suelen emplearse cuando los datos son de tipo binario.
- Estrategias politéticas, cuyas divisiones se basan en los valores tomados por todas las variables.

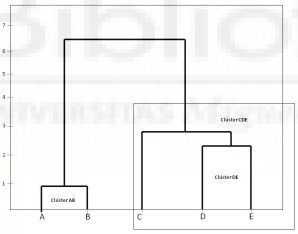

Figura 2: Ejemplo de dendograma

4.1.4. Metodos clúster no jerarquicos

Estos métodos no jerárquicos están diseñados para agrupar observaciones (individuos), pero de todas formas en la práctica se puede utilizar para agrupar variables en algún punto concreto si la composición de los datos lo permite, aunque lo ideal para este método es que existan muchas más observaciones que variables a la hora de realizar el análisis. Para comenzar el procedimiento de estos métodos es necesario establecer un número concreto de K clusters a priori. Una vez que el analista ha determinado el número de grupos (clusters) en el que quiere agrupar los individuos, el procedimiento consiste en iterar intercambiando a los individuos entre los diferentes grupos para llegar a la *mejor* composición posible. Los diferentes métodos existentes se diferencian en como poder llegar a la *mejor partición posible*, es decir, el camino que siguen para obtener las agrupaciones óptimas.

A diferencia de los métodos jerárquicos, estos no trabajan sobre la matriz de distancias, es decir, el análisis se realiza directamente sobre la matriz original de datos. Este aspecto hace que los métodos no jerárquicos sean más idóneos en la práctica a la hora de analizar gran número de observaciones y/o variables, ya que no es necesario guardar la matriz de distancias, por lo que no se precisa de tanta memoria en el computador.

Los diferentes tipos de métodos de agrupación no jerárquicos son los siguientes:

- Métodos de reasignación:

  Método de K-Medias.

  El Quick-Cluster análisis.

  Método de Forgy.

  Método de las nubes dinámicas.

- Métodos de búsqueda de densidad:

  Análisis modal de Wishart.

  Método Taxmap.

  Método de Fortin.

- Métodos directos:

  Método de Block-Clustering.

- Métodos de reducción de dimensiones:

  Análisis factorial tipo Q.

De todos los métodos presentados anteriormente, el método más utilizado y común que hoy en día se utiliza es el método de K-MEANS.

A continuación, se especificará a modo de resumen, el procedimiento que siguen estos métodos no jerárquicos:

1. Determinar los centroides iniciales de los K grupos. Este primer paso se denomina más comúnmente como *elección de puntos de semilla*, para el cual existen diferentes procesos. Los más comunes son los siguientes:

   - Elegir puntos de semilla las primeras K observaciones (individuos) de la matriz de datos. Este procedimiento es el más simple.
   - Enumerar los individuos de 1 a m y elegir los siguientes:

     $$[\tfrac{m}{k}], [\tfrac{2m}{k}], \ldots, [\tfrac{(k-1)m}{k}] \text{ y } m$$

     donde los corchetes ($[x]$) representan la parte entera de x. Con este sistema se evita obtener alguna secuencia no aleatoria.
   - Etiquetar de igual forma que el método anterior los casos de 1 a m y elegir k números aleatorios diferentes (McRae, 1971).

2. Formación de los k grupos. Los procedimientos más comunes para generar unas particiones iniciales son los siguientes:

- Tras la elección de los puntos de semilla, cada observación es asignada al clúster con el punto de semilla más próximo a dicha observación, de este modo los puntos de semilla siguen permaneciendo estacionarios durante el proceso de asignación. Esto nos reporta un conjunto de clusters independientes de la secuencia con la que los individuos han sido introducidos.
- Partiendo de un conjunto de clusters unitarios, siendo estos los puntos de semilla iniciales, se va asignando cada observación al clúter con el controide más próximo. Tras ser asignado, se vuelve a recalcular el centroide del clúster, por lo que los clusters pueden ir moviendo y modificando su centroide inicial, esto lleva a que la distancia entre un individuo y un centroide puede ir variando durante el proceso.

3. Iterar recalculando los centroides y formando grupos hasta la estabilidad.

4.1.4.1. Método no jerárquico K-MEANS

Aquí se va a introducir el método más común para llevar a cabo un análisis clúster no jerárquico. Este método es muy adecuado a la hora de analizar el agrupamiento de los casos a estudiar cuando se quiere agrupar de forma iterativa.

Este método fue diseñado por McQueen en el año 1972. Este método consiste básicamente en asignar cada uno de los casos a su correspondiente clúster (de los K que ya se han fijado previamente) minimizando la distancia con el centroide asignado a cada clúster. Este método está incluido dentro de la gama de métodos denominados de reasignación, ya que una observación puede ser asignada a un determinado clúster en un paso concreto y posteriormente, en un siguiente paso, puede ser reasignado a un otro clúster diferente. Además, tiene una ventaja significante, y es que los centroides de cada clúster se van recalculando tras cada asignación y no al final del ciclo. Por esta razón siempre se va a estar iterando, no solo en buscar centroides cercanos a cada observación, sino que además, se iterará buscando el "mejorçentroide para cada clúster. Los pasos que sigue el algoritmo propuesto por McQueen son los siguientes:

- En primer lugar se toman las primeras K observaciones como clusters individualizados.
- Posteriormente, ir asignando cada caso (individuos) restante a un determinado clúster con el centroide más cercano. Después de realizar una asignación se vuelve a calcular el centroide del clúster creado, más próximo a los casos que se han agrupado.
- Una vez que ya están todos los casos englobados en su respectivo clúster, se toma cada centrodie como punto de semilla fijo. Teniendo en cuenta esos puntos de semilla, se lleva a cabo una última iteración ajustando los datos al punto de semilla (fijado en la última etapa) que más cercano se encuentre.

De forma analítica, el método de K-MEANS busca dividir los puntos $X_1, ..., X_n$ en K grupos. Para ello se buscan los centroides de los clusters que minimicen:

$$W_n = \frac{1}{n} \sum_{i=1}^{n} \min_{1 \leq j \leq k} \parallel X_i - a_j \parallel^2$$

donde $\parallel \vdots \parallel$ denota la medida de distancia euclidea. a_j denota el valor del centro de cada clúster. Por lo que el método consiste en asignar cada X_i al clúster cuyo centro a_j este más cercano. Por lo tanto, cada centro a_j adquiere un conjunto de puntos para formar un clúster (C_j). El valor medio de los puntos en C_j tiene que ser igual a a_j, si no coinciden podemos remplazar este centroide a_j por la media del clúster, y, por lo tanto, se podrá minimizar en mayor medida W_n. De este modo se podrá reasignar algunos X_i a nuevos centroides (clusters).

4.2. Medidas de distancias y similaridad

Aquí se desarrollarán las medidas numéricas de similitudes y distancias más usadas a la hora de agrupar individuos para formar grupos homogéneos. A partir de una matriz de datos con tamaño mxn, aplicando una determinada medida de distancia/similitud, se obtendrá una matriz simétrica de distancias/similitudes D ($d_{ij} = d_{ji}$). Cada componente de esta matriz representará el valor resultante de la medida elegida para los correspondientes casos i y j. Por ejemplo, el componente de la matriz $d_{12} = d_{21}$ será la distancia que exista entre el caso 1 y el caso 2. Como ya se ha dicho anteriormente, es necesario elegir una medida de asociación apropiada para cada problema concreto.

Cuando se decide trabajar con distancias como medida de asociación, los individuos que se encuentren en el mismo grupo tendrán minimizada la distancia entre ellos. Y, por consiguiente, cuando se trabaje con medidas de similitud los individuos del mismo grupo tendrán maximizada esta medida. Las matrices de distancias (D) creadas a partir de la misma matriz de datos originales, pueden variar de forma considerable dependiendo de la medida de distancia/similitud utilizada. Pero al igual que estas matrices pueden variar, también existen medidas de distancia relacionadas entre sí, lo que proporcionará resultados similares. Esto quiere decir que, si realizando diferentes análisis se llegan a resultados similares, no siempre se puede afirmar que se haya encontrado la verdadera estructura de los datos, ya que estos resultados similares pueden corresponder a la relación entre las medidas de distancia utilizadas. Para medir la semejanza entre las observaciones es necesario utilizar diferentes procedimientos (medidas), los cuales dependen del tipo de variable al que se les aplica. De manera general, encontramos diferenciadas dos tipos de medidas:

- Medidas de distancia o disimilaridad.
- Medidas de proximidad o similaridad

4.2.1. Definiciones de distancias y proximidad

Antes de explicar los tipos de medidas de asociación que son usualmente utilizadas, se introducirán las definiciones métricas de distancias y similaridad.

En primer lugar, se analizará la definición de distancia. Las medidas de distancia son las más utilizadas para llevar a cabo el análisis clúster. En realidad,

son medidas de diferencia, donde los valores más elevados quieren decir que existe una similitud menor y por lo tanto más diferentes serán los objetos y menor será la probabilidad de que los métodos de clasificación los incluyan en el mismo clúster.

A continuación se expone la definición de distancia métrica:

Definición 1. Sea U un conjunto finito o infinito de elementos. Una función *d:* $U \times U \to \mathbb{R}$ se llama una distancia métrica si $\forall$ *x,y,z* $\in U$ se tiene lo siguiente:

1. *d(x,y)* ≥ 0
2. *d(x,y)* $= 0 \Leftrightarrow x = y$
3. *d(x,y)* $=$ *d(y,x)*
4. *d(x,z)* $\leq$ *d(x,y)* $+$ *d(y,z)*

Una vez definida a distancia métrica, se expondrá la definición de similaridad métrica:

Definición 2. Sea U un conjunto finito o infinito de elementos.Una función *s:* $U \times U \to \mathbb{R}$ se llama similaridad métrica si cumple las siguientes propiedades: $\forall$ *x,y,z* $\in U$

1. *s(x,y)* $\leq s_0$
2. *s(x,x)* $= s_0$
3. *s(x,y)* $=$ *s(y,x)*
4. *s(x,y)* $= s_0 \Longrightarrow x = y$
5. $\mid$ *s(x,y)* $+$ *s(y,z)* $\mid$ *s(x,z)* $\geq$ *s(x,y)s(y,z)*

donde s_0 es un número real finito arbitrario. Una vez definidas las medidas de distancias y similitudes pasaremos a introducir las más usadas en la práctica.

4.2.2. Medidas de distancias

Antes de introducir las diferentes medidas consideramos de forma general la siguiente matriz de datos. Compuesta por m individuos y n variables: $X_1, ..., X_n$. Por lo que tendremos como resultado una matriz $m \times n$ de la siguiente forma:

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mj} & \cdots & x_{mn} \end{pmatrix}$$

Dada esta matriz, la i-ésima fila de la matriz X contiene los valores del conjunto de variables para el i-ésimo individuo. Y la j-ésima columna representa los

valores de la j-ésima variable para el conjunto de individuos de la muestra.

Dado dos individuos escogidos de la muestra i y j, es decir, dos filas de nuestros datos (matriz X):

$$x_i = \left(\; x_{i1}, \ldots, x_{in} \; \right)$$

$$x_j = \left(\; x_{j1}, \ldots, x_{jn} \; \right)$$

4.2.2.1. Distancias para variables cuantitativas

Distancia euclídea

Esta distancia euclídea entre dos individuos i y j viene dada por:

$$d_E(x_i, x_j) = \sqrt{\sum_{l=1}^{n} (x_{il} - x_{jl})^2} = \sqrt{(x_i - x_j)'(x_i - x_j)}$$

Esta medida de distancia tiene una serie de inconvenientes, el más destacado es que es una medida no invariante frente a cambios de escala de las variables, es decir, es sensible a las unidades de medida de las variables. Esto quiere decir que las diferencias entre variables que contengan valores altos contribuirán en mayor medida al estudio que las diferencias entre variables con valores bajos. Por consiguiente, los cambios de escala llevarán a cambios en la distancia medida entre los individuos. La solución más usual en estos casos es tipificar los valores de las variables.

Otro inconveniente que tiene la distancia euclídea está relacionado con la propia naturaleza de las variables. Ya que esta medida presupone que las variables no están correlacionadas, por lo que, si existen variables que si lo estén, esta medida inflará la distancia entre los individuos, debido a que los valores individuales de algunas variables podrían explicarse por las diferencias entre otras (al existir correlación entre ellas). Para solucionas este problema se pondera la contribución de cada par de variables con pesos inversamente proporcionales a las correlaciones. Esta solución lleva a la utilización de la medida de distancia llamada Mahalanobis que se explica a continuación.

Distancia de Mahalanobis

La distancia de Mahalanobis entre los individuos i y j viene definida por:

$$d_M(x_i, x_j) = \sqrt{(x_i - x_j)' S^{-1} (x_i - x_j)}$$

donde S es una matriz de varianzas-covarianzas de las variables de la matriz de datos X. Por lo que S^{-1} será su inversa. Matricialmente podemos expresar dicha matriz de la siguiente forma:

$$S = \frac{1}{m} \tilde{X}' \tilde{X} \text{ con } \tilde{X} = (\tilde{x}_{ij}) \; ; \; \tilde{x}_{ij} = x_{ij} - \bar{x}_j \; i = 1, ..., m \; ; \; j = 1, ..., n$$

Esta medida de distancia presenta dos ventajas importantes respecto a la distancia euclídea. Por un lado, se presenta invariante frente a los cambios de

escala (transformaciones lineales), y, por lo tanto, no depende de las unidades de medida.

Por otro lado, al utilizar la matriz de varianzas-covarianzas (S) tiene en cuenta las correlaciones entre las variables, por lo que corrige el efecto de la redundancia. Es decir, no aumentará la medida de distancia por incorporar mayor número de variables observadas, solamente aumentará cuando se incorporen nuevas variables que no sean redundantes con respecto al resto de variables ya observadas.

A pesar de estas ventajas, esta medida presenta una desventaja la cual está relacionada con el cálculo de la matriz de varianzas-covarianzas, ya que está basado en todos los individuos de forma conjunta, por lo que no trata de manera separada los objetos de cada clúster, que sería lo ideal para el estudio.

Distancia de Minkowski

Esta distancia viene definida de la siguiente forma:

$$d_p(x_i, x_j) = (\sum_{l=1}^{n} |(x_{il} - x_{jl})|^p)^{\frac{1}{p}} \text{ , para } p > 0$$

Esta medida de distancia presenta los mimos inconvenientes que la distancia euclídea. Ya que tampoco se muestra invariante a cambios en la escala existentes en las variables. Esta medida da lugar a dos casos particulares:

1. Distancia ciudad o de Manhattan, para $p = 1$:

$$d_1(x_i, x_j) = \sum_{l=1}^{n} |(x_{il} - x_{jl})|$$

2. Distancia dominante o del máximo, para $p = \infty$:

$$d_\infty(x_i, x_j) = \max_{l=1,...,n} |(x_{il} - x_{jl})|$$

4.2.2.2. Distancias para variables binarias

En este apartado se expondrán los coeficientes más usuales a la hora de medir distancias entre datos binarios, por lo que son variables compuestas por mediciones que solo pueden tomar los valores de 1 o 0. Antes de introducir dichos coeficientes, sean $X_1, ..., X_n$ variables binarias con posibles valores 0,1:

- a = número de variables con respuesta 1 (atributos presentes) en ambos individuos (i,j).
- b = número de variables con respuesta 0 (atributos ausentes) en el individuo i y con respuesta 1 en el individuo j.
- c = número de variables con respuesta 1 en el individuo i y con respuesta 0 en el individuo j.
- d = número de respuesta 0 en ambos individuos.

Notese que $a+b+c+d = p$ Una vez introducida la notación necesaria para exponer los diferentes coeficientes, pasaremos a mostrar los más utilizados en la práctica:

1. Distancia Euclidea.

$$d_E(x_i, x_j) = \sqrt{b + c}$$

2. Distancia Euclidea al cuadrado.

$$d_E^2(x_i, x_j) = b + c$$

3. Distancia de Tamaño.

$$d_T(x_i, x_j) = \frac{(b - c)^2}{(a + b + c)^2}$$

4. Varianza.

$$d_V(x_i, x_j) = \frac{(b + c)}{4(p)}$$

5. Lance y Wiliams.

$$d_L(x_i, x_j) = \frac{(b + c)}{2a + b + c}$$

4.2.2.3. Distancias para variables categóricas

En este apartado se introducirá el estadístico Chi-cuadrado χ^2 como método para medir las diferencias (distancias) que pueden existir entre dos variables categóricas. Este estadístico χ^2 suele ser usado, por un lado, para probar la independencia o determinar la asociación entre variables categóricas. O, por otro lado, para poder determinar si un modelo estadístico ajusta los datos de manera adecuada. Para llevar a cabo un análisis clúster, interesa utilizar dicho estadístico con la primera aplicación mencionada, ya que nos ayudará a poder extraer una medida de distancia entre variables categóricas.

Para poder llevar a cabo este análisis, y poder medir la asociación existente entre dos variables a través del estadístico χ^2, se utilizan las llamadas *tablas de contingencias*. Estas tablas recogen el total de valores observados en las diferentes composiciones de las categorías. Antes de introducir la composición de estas tablas categóricas notemos que, de forma general:

- o_{ij} = valor (frecuencia) observado en la posición *i,j*.
- e_{ij} = valor (frecuencia) esperado bajo la hipótesis de independencia.

Además, p y q son el número de categorías que pueden tomar las variables estudiadas. Por lo que una tabla de contingencia tendría la siguiente forma:

Variable A/Var B	1	...	j	...	q	
1	n_{11}	...	n_{1j}	...	n_{1q}	n_1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
i	n_{i1}	...	n_{ij}	...	n_{iq}	n_i
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
p	n_{p1}	...	n_{pj}	...	n_{pq}	n_p
	$n_{,1}$	...	$n_{.j}$	...	$n_{.q}$	$n_{..}$

Con toda esta notación se puede definir el estadístico χ^2 como

$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

4.2.3. Medidas de similaridad o proximidad

4.2.3.1. Proximidad para variables cuantitativas

Correlación entre individuos

De manera formal se puede utilizar el coeficiente de correlación entre dos vectores como una medida de similitud (distancia) entre individuos. El cual se presenta de la siguiente forma:

$$r_{ij} = \frac{\sum_{l=1}^{n}((x_{il} - x_i)(x_{jl} - x_j))}{s_i s_j}$$

donde se ha definido:

- $\bar{x}_h = \frac{1}{n} \sum_{l=1}^{n} x_{hl}$ $h = i, j$ como la media de cada individuo.
- $s_h = \sqrt{\sum_{l=1}^{n}(x_{hl} - \bar{x}_h)^2}$ $h = i, j$ como la desviación cuadrática de cada individuo.

4.2.3.2. Proximidad para variables binarias

Coeficiente de concordancia simple

Este coeficiente tiene en cuenta las semejanzas entre la presencia del atributo en ambos casos (a) y la ausencia de este en ambos casos (d). Este coeficiente de concordancia simple queda de la siguiente forma:

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

Este coeficiente tomará valores entre 0 y 1. Cuanto más se acerque a cero menos similitud existirá entre las variables medidas. Por el contrario, contra más se acerque a 1, más similitud representarán las variables medidas.

Coeficiente de Jacard

Este coeficiente presenta la siguiente forma:

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

Este coeficiente no tiene en cuenta cuando en ambos valores existe ausencia de atributo (d). En cambio, el coeficiente de concordancia simple si, esto es debido a que en algunas situaciones las variables de tipo binarias no son simétricas, por lo que puede llegar a ser más informativo la presencia de atributo (1) que la ausencia (0). En estos casos es más conveniente utilizar coeficientes del tipo Jacard.

5. Resultados

En este apartado se van a exponer los resultados obtenidos tras el análisis estadístico de este estudio basado en analizar el comportamiento de diferentes tipos de compost y así, llegar a conclusiones sobre que tipo de compost presenta mejores prestaciones a los diferentes cultivos. Para ello se prestará mayor atención a variables más significativas a la hora de comercializar el compost, como por ejemplo son, las variables relacionadas con el NPK del sustrato. Este componente (NPK) es la composición que existe en el sustrato de sodio, fósforo y potasio.

Para llevar a cabo este estudio se utilizará el programa estadístico R. En primer lugar, se expondrá un análisis descriptivo de las variables utilizadas y de los parámetros de los sustratos. De este modo se podrá empezar a tener una leve idea de la estructura de los datos. Y, en segundo lugar, se expondrá el análisis clúster llevado a cabo analizando que tipo de este será más adecuado para los datos de los que se disponen. Una vez que se haya expuesto todo el análisis, estaremos en disposición de extraer las conclusiones oportunas.

Antes de empezar el análisis estadístico, se va a exponer la estructura de la base de datos de la que se dispone. Como ya se ha dicho anteriormente, las columnas representas variables que explican el comportamiento del cultivo. Y como filas, tres observaciones por cada compost, porcentaje aplicado y cultivo. A continuación, se muestra una tabla con algunas observaciones:

	tratamientos_planta	Observaciones	Length.cm.	ProjArea.cm2.	AvgDiam.mm.	Tips
1	turba 0%/lechu	O1	499	19	0.38	416
2	turba 0%/lechu	O2	599	20	0.34	416
3	turba 0%/lechu	O3	431	16	0.37	449
9	Z1-25%/lechu	O1	522	20	0.38	524
10	Z1-25%/lechu	O2	511	20	0.40	518
11	Z1-25%/lechu	O3	574	23	0.41	478

Figura 3: Estructura de los datos

5.1. Análisis descriptivo

En primer lugar, vamos a analizar el comportamiento de las variables con un análisis descriptivo de las mismas. Para empezar, se expone una tabla con los valores mínimos, máximos y medianas de las características de los diferentes sustratos utilizados.

Parámetro	Min	Max	Mediana
pH	5,81	8,18	6,66
Conductividad eléctrica (dS/m)	0,15	5,42	1,69
Materia orgánica total (%)	48,3	93,4	68,1
Densidad aparente (g/cm^3)	0,09	0,52	0,23
Espacio poroso total (%)	70,1	94,2	88,7
Capacidad de aireación (%)	19,0	67,0	43,7
Contracción (%)	0,1	35,2	14,2
Agua fácilmente disponible (%)	1,6	69,5	14,0
Capacidad retención agua (mL/L sustrato)	337	671	528

Figura 4: Cuadro de mínimos, máximos y medianas

Atendiendo a este primer descriptivo, se observa que ninguna variable expuesta presenta un comportamiento anormal, excepto la variable relacionada con el agua fácilmente disponible. Esta variable presenta un valor máximo muy elevado en comparación a su valor mínimo y mediana, ya que el resto de las variables tienen la mediana acorde con sus valores mínimos y máximos. Por lo que habrá que atender a esta variable con especial atención a la hora de desarrollar el análisis.

A continuación, se va a mostrar un gráfico donde se desarrollan las correlaciones cruzadas entre dos grupos diferentes de variables utilizadas. Por un lado, se tienen las características relacionadas con los sustratos y, por otro lado, las características morfológicas de las plantas. De este modo se podrá observar ver que aspectos del sustrato afecta tanto positivamente como negativamente a las características de las plantas.

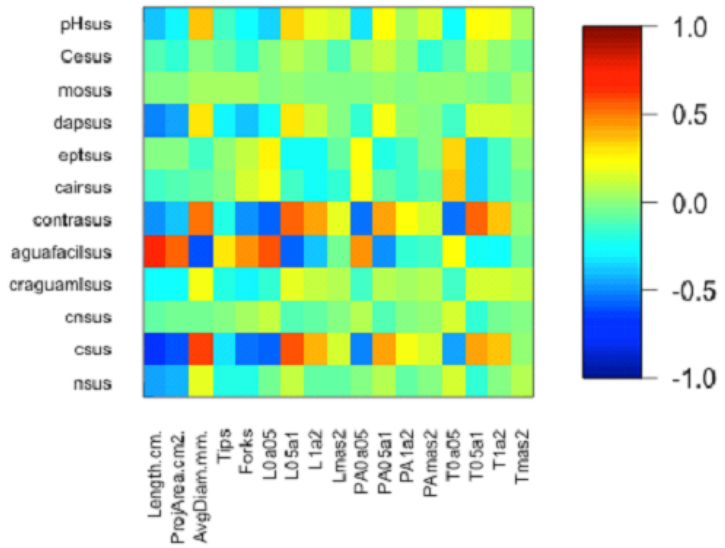

Figura 5: Gráfico de correlaciones

Atendiendo a este gráfico vemos que efectivamente existen correlaciones (tanto positivas, como negativas) para los parámetros estudiados. Entre otras, se destacan las correlaciones negativas existentes de la longitud radicular (Length) y del área proyectada (ProjArea) frente al pH (pHsus), densidad (dapsus), contracción (contrasus), contenido en carbono (csus) y nitrógeno (nsus). Entre estas correlaciones destaca la de ambas características de las plantas contra el contenido en carbono. Al tratarse de una correlación negativa, se podría realizar una primera aproximación afirmando que un alto contenido en carbono en el sustrato hace que la longitud y el área de las raíces sean menor.

Atendiendo al factor de agua fácilmente disponible (aguafacilsus), vemos que genera una correlación positiva frente a la longitud y área de las raíces. Pero de forma opuesta, genera una correlación negativa frente a la ramificación generada en las raíces (Forks). Por estas razones se puede afirmar que la disponibilidad de agua es un parámetro clave en la morfología radicular, ya que aumentando el agua disponible aumenta la longitud y área radicular, pero a su vez genera un menor diámetro promedio (AvgDiam), por lo que la falta de agua lleva a un engrosamiento de las raíces. Por último, se observa que la salinidad (cnsus, que es la relación de carbono y nitrógeno) no muestra correlaciones significativas frente a ninguna característica morfológica de las raíces.

Una vez analizadas las correlaciones, se van a exponer una serie de gráficos con las variables más significativas a la hora de concluir que compost presenta una mayor calidad, diferenciando por tipo de compost utilizado, con esto se finalizará el análisis descriptivo.

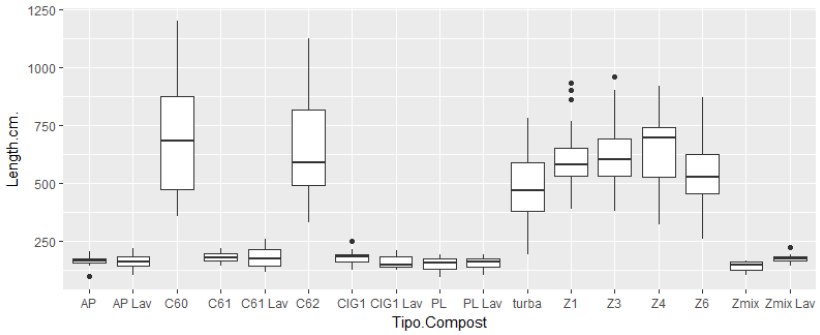

Figura 6: Gráfico para la longitud de raíz (suma) por planta

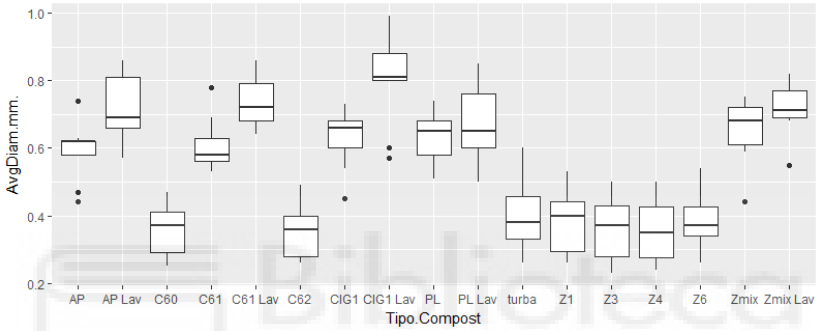

Figura 7: Gráfico para el diámetro promedio de la suma de raíces por planta

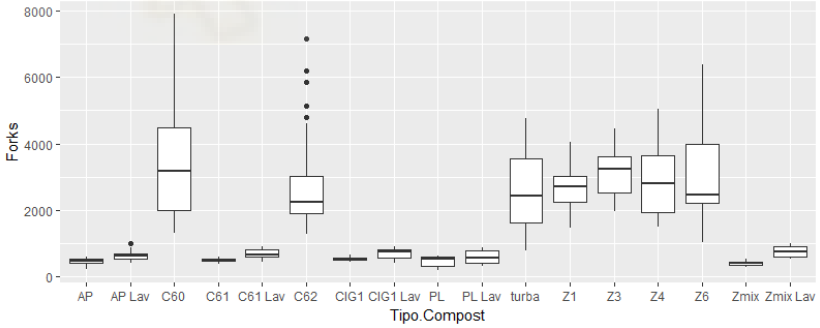

Figura 8: Gráfico para las bifurcaciones por planta, indica el grado de ramificación

.

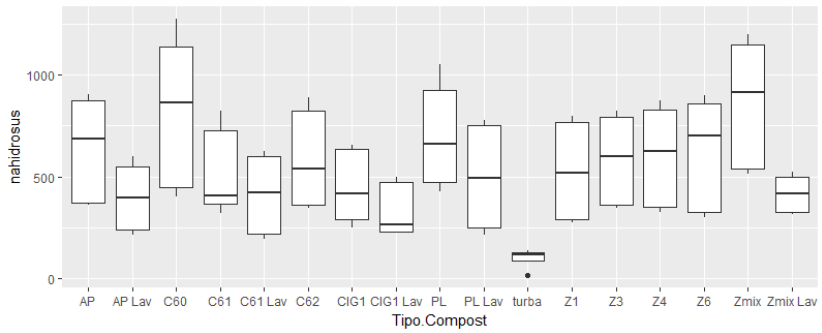

Figura 9: Gráfico para concentración hidrosoluble del elemento Na (sodio)

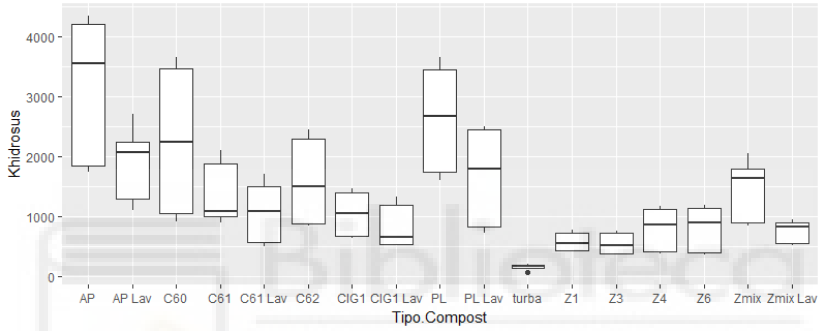

Figura 10: Gráfico para concentración hidrosoluble del elemento K (potasio)

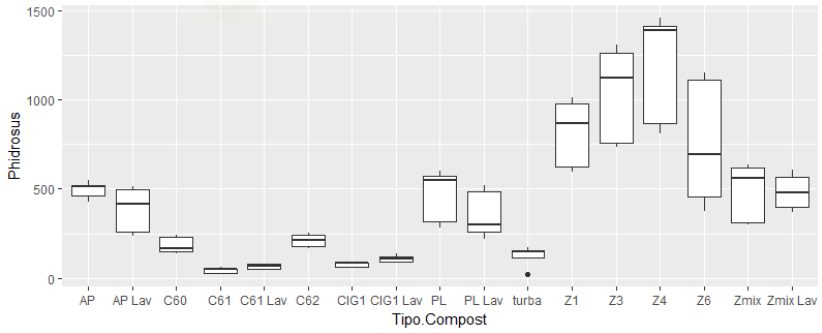

Figura 11: Gráfico para concentración hidrosoluble del elemento P (fósforo)

Como se puede observar en los gráficos expuestos, en primer lugar, destacan los compost C60 y C62 dando una mayor longitud total y ramificaciones de las raíces a las plantas. Estas dos variables están relacionadas entre sí, ya que un aumento en las ramificaciones genera una longitud total mayor en las raíces. Esto se puede observar en el comportamiento de ambos gráficos (Figura 5 y 7), ya que se comportan de la misma forma. Por lo que, en primera instancia,

se tienen estos dos tipos de compost (C60 y C62) como mayores generadores en cuanto a cantidad de raíces se refiere. En cambio, el gráfico que muestra el diámetro promedio de las raíces tiene un comportamiento opuesto a la longitud y bifurcaciones de las raíces, por lo que un menor diámetro lleva a un aumento en la longitud (y ramificaciones) de las raíces.

Atendiendo a las variables referentes al NPK del sustrato (Figura 9, Figura 10 y Figura 11), se aprecia que en conjunto no destaca ningún compost que genere valores superiores en las tres variables. Pero por el contrario sí que destaca los valores muy inferiores de la turba frente al resto de compost. Por lo que se ve evidenciado la necesidad de llevar a cabo el presente estudio, ya que se pretende encontrar un grupo de compost que puedan sustituir con más eficiencia a este método tradicional de cultivo.

5.2. Análisis comparativo

Dentro de este análisis estadístico comparativo se van a diferenciar dos puntos. En primer lugar, se va a analizar el comportamiento que tienen las variables, y así poder observar si existe alguna relación entre ellas. Y, en segundo lugar, se analizará el comportamiento de las observaciones, que es el análisis sobre el que más importancia recae, ya que será donde se extraigan las conclusiones que se necesitan para ver posibles compost sustitutos a la turba.

5.2.1. Análisis clúster para variables

En este apartado se llevará a cabo el análisis clúster correspondiente a las variables. Por lo que se podrá concluir con el comportamiento que tienen las variables a la hora de ser agrupadas. Este análisis no reportará conclusiones directas para solventar los objetivos planteados del estudio, pero si que nos ayudará a observar si existe alguna relación muy significativa entre variables. Esta relación puede llevarnos a excluir alguna variable del estudio o simplemente tener en cuanta alguna relación existente.

5.2.1.1. Análisis clúster jerárquico para variables

En primer lugar, se realizará un análisis clúster jerárquico para analizar la agrupación entre variables. Como se ha dicho en puntos anteriores, para llevar a cabo este análisis es necesario elegir la medida de distancia y estrategia de agrupación oportunas. En este estudio se han realizado varios análisis para ver cual proporciona la "mejor". agrupación. Para saber este aspecto de "mejor" agrupación, se va a analizar el llamado *coeficiente de aglomeración (división)*. Este coeficiente, proporcionado por Kaufman y Rousseeuw (1990), da una referencia sobre la calidad de los resultados. Este son las medias de medida de silueta de la cohesión (distancia) y separación de los clusters en todos los registros, $(B - A)/max(A, B)$, donde A es la distancia del caso (observación) al centro de su conglomerado y B es la distancia del caso (observación) al centro del conglomerado más cercano al que no pertenece. Este coeficiente toma valores entre -1 y 1. Contra más se acerque a 1, quiere decir que los casos pertenecientes a los cluster están muy cercanos al centro del clúster. Es decir, si el valor del coeficiente es igual a 1 implica que todos los casos están ubicados directamente en los centros de sus propios clusters. Y, por el contrario, si toma valor -1 implica, de media, que los casos están equidistantes entre el centro de su propio conglomerado y el siguiente conglomerado más cercano.

Una vez elegidas la medida de distancia y la estrategia de agrupación que proporcionan una mayor calidad de ajuste se estará en disposición de llevar a cabo el análisis, donde se realizarán los cálculos oportunos y la ilustración del dendograma para poder extraer las conclusiones necesarias. A continuación, se muestra el proceso llevado a cabo indicando las funciones utilizadas en el programa R. En primer lugar, se tipifica la matriz de datos original.

```
datos_tipificados<-scale(datos_originales[,-1])
```

Quedando los datos tipificados, sin las dos primeras columnas, de la siguiente forma (se muestra solamente las filas y columnas mostradas en la tabla anterior a modo de ilustración):

	Length.cm.	ProjArea.cm2.	AvgDiam.mm.	Tips	Forks	L0a05
1	0.1443526	0.0139169	-0.5334945	-0.4471072	-0.1171441	0.6133223
2	0.5270123	0.1328184	-0.7653341	-0.4471072	-0.1019012	0.7812121
3	-0.1158560	-0.3427877	-0.5914544	-0.4085624	-0.2365472	0.6133223
9	0.2323643	0.1328184	-0.5334945	-0.3209607	-0.1184144	0.6552948
10	0.1902717	0.1328184	-0.4175747	-0.3279688	0.0333800	0.5293774
11	0.4313474	0.4895229	-0.3596148	-0.3746898	0.0429069	0.5713499

Figura 12: Datos tipificados

Una vez que se han tipificado los datos para evitar problemas de escala a la hora de calcular la distancia, se está en disposición de calcular la matriz de distancias con la que se realizará el análisis clúster. Para este cálculo y llevar a cabo el análisis clúster de tipo jerárquico se han utilizado las siguientes funciones:

```
matriz_de_distancia<-dist(datos, method = "medida_de_distancia")
analisis_cluster<-hclust(distancias, method = "metodo_de_agrupación")
```

Estas dos funciones se han computado una serie de veces variando la medida de distancia y la estrategia de agrupación utilizadas hasta llegar al *coeficiente de aglomeracion* mas cercano a 1. Este coeficiente se obtiene con el siguiente comando:

```
coef.hclust(analisis_cluster)
```

El coeficiente más cercano a 1 se ha obtenido realizando el análisis con la medida de distancia *euclídea* y con el método de agrupación *Ward*. El valor resultante de este coeficiente para esta combinación ha sido de 0.9168, por lo que se puede afirmar que este análisis tiene muy buena calidad de ajuste, lo que quiere decir que cada caso (observación) está muy cercano del centro del clúster al que pertenece.

Con la siguiente función se obtiene el dendograma resultante del análisis llevado a cabo con la medida de distancia y método de agrupación anteriormente indicados:

```
plot(analisis_cluster)
```

Este dendograma es útil para poder decidir acerca del número de clusters final con el que analizar los datos. Además del dendograma nos ayudaremos de la visualización del proceso numérico del análisis clúster para poder extraer información sobre las medidas de distancia que han ido juntando los diferentes casos. Es necesario basarse en este análisis numérico, ya que decidir el número de clusters atendiendo al dendograma, suele ser una tarea subjetiva (sentido común).

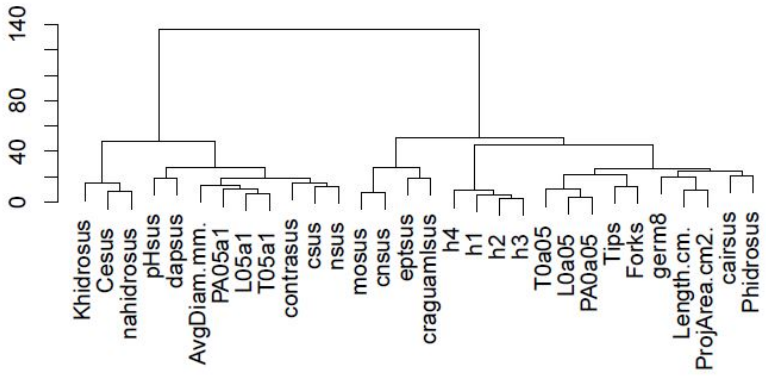

Figura 13: Dendograma para las variables

Una vez en este punto, es necesario discutir el número de clusters óptimo para explicar correctamente la agrupación llevada a cabo por el análisis de las variables y analizar la distancia existente entre los diferentes casos (variables). Dependiendo de la distancia en la que se quiera cortar el dendograma se obtendrán un número diferentes de clusters, ya que, si cortáramos por ejemplo en una distancia de 80, solamente se obtendrían dos grupos, caso que es totalmente inadecuado, ya que no explicaría ningún resultado útil para concluir. Para cortar el dendograma y obtener los grupos más significativos, nos basaremos en la composición y naturaleza de las variables. A primera vista se observa que pueden existir diferentes combinaciones en cuanto al número de clusters posibles. En primer lugar, vamos a cortar en 8 grupos para ver cómo quedaría la composición de los clusters. Una vez que se decide el número de clusters, se pueden dibujar en el gráfico original para tener una visualización más clara de esta composición y así poder ver de qué manera agrupa los casos el método llevado a cabo. Esto se puede llevar a cabo mediante la siguiente función:

```
rect.hclust(analisis_cluster, k=nuemro_de_cluster,
            border=çolor_rectangulos")
```

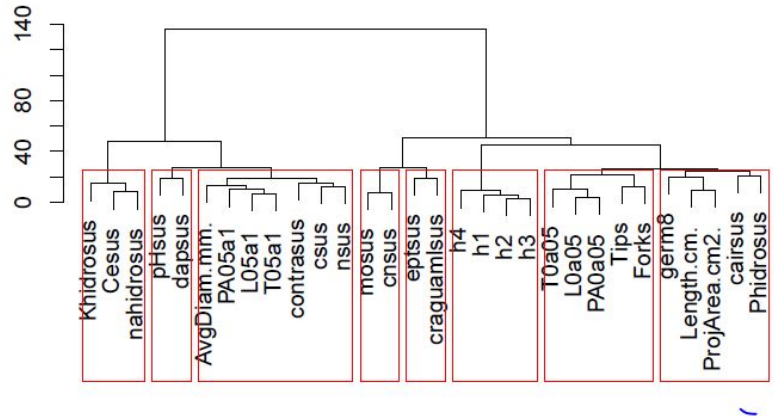

Figura 14: Dendograma para las variables agrupadas en 8 clusters

Como se puede observar, existen particiones que podrían ser un mismo clúster por su cercanía, pero cortando en 8 clusters el método los separa, como es el caso de los clusters número 4 y 5, ya que, basándonos en la naturaleza de las variables, las cuatro variables que están comprendidas en dichos clusters tienen una relación entre sí. Estas cuatro variables explican características de la composición del sustrato. Por lo que sería conveniente realizar un análisis con 7 clusters y ver si junta dichos clusters en un único grupo sin modificar la composición del resto. El dendograma resultante con 7 clusters es el siguiente:

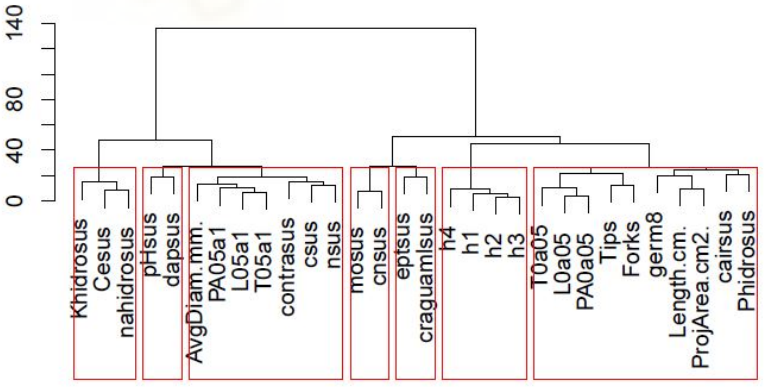

Figura 15: Dendograma para las variables agrupadas en 7 clusters

En este caso junta un clúster que no era el que se pretendía juntar, por lo que nos dice que las variables comprendidas en el clúster unido (ahora, clúster número 7) están más cerca entre ellas que las cuatro variables de los otros dos clusters (4 y 5). Con esta información ya se puede afirmar que, por ejemplo, la variable *Length* está más cerca de *Forks* y *Tips*, que las variables del clúster 4 (*mosus* y *cnsus*) a las del clúster 5 (*eptsus* y *craguamlsus*). Esto confirma que existe una relación mayor entre la suma de las longitudes, los ápices y bifurcaciones de raíces por planta, que entre la relación C/N y la capacidad de retención de agua útil del sustrato.

A continuación, se va a realizar un dendograma dividiendo las variables en 6 clusters para ver si une en un mismo grupo las variables objetivo desde el principio, ya que son 4 variables que miden características del sustrato, por lo que nos interesaría ver si aún existen variables o clusters más relacionados entre sí que estas variables del sustrato.

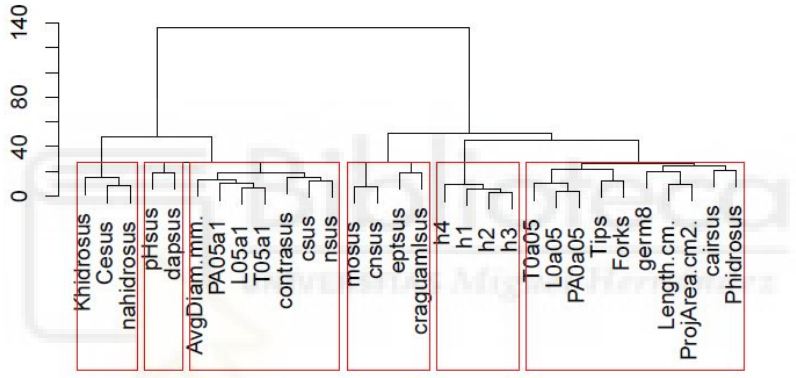

Figura 16: Dendograma para las variables agrupadas en 6 clusters tras análisis jerárquico

Como se puede apreciar, con 6 clusters ya une estas características del sustrato en un solo grupo. Comparando con las demás variables que explican características del sustrato, aparece que el clúster número 1 también está compuesto por características de este. Este clúster esta unificado desde el principio, por lo que las tres variables que lo conforman tienen una relación más cercana entre sí, que el clúster creado pasando de 7 a 6 clusters (que finalmente sería el clúster numero 4). Este comportamiento de agrupación afirma que, existe una relación mayor entre la concentración hidrosoluble elemento potasio (K), sodio (Na) y la conductividad eléctrica (salinidad); que la existente entre la materia orgánica y el espacio poroso total del sustrato.

Y, por el contrario, existen dos variables del sustrato (clúster 2) que son independientes del resto durante todo el análisis. Esto quiere decir que el pH del sustrato está altamente relacionado con la densidad aparente del mismo (*pHsus* y *dapsus*), y a la vez son variables que no tienen una relación muy elevada con el resto de las variables referentes al sustrato.

5.2.1.2. Análisis clúster no jerárquico para las variables (K-MEANS)

En este apartado se va a realizar el mismo estudio para un análisis no jerárquico, en concreto se llevará a cabo el método de K-MEANS. En este método sí es necesario prefijar el número de clúster en los que queremos agrupar las observaciones, por esta razón se ha llevado a cabo con anterioridad un análisis jerárquico, ya que este da una orientación para decidir el número de clusters con el que empezar en análisis no jerárquico. Además, para llevar a cabo este método se utiliza la matriz de datos originales, no es necesario calcular la matriz de distancias. La función que se ha utilizado para llevar a cabo este estudio es la siguiente:

```
analisis_KMEANS<-kmeans(datos_originales, numero_de_clusters)
```

Basándose en el análisis jerárquico, se va a llevar a cabo el análisis fijando 6 clusters inicialmente y así poder ver que composición desarrolla el método de K-MEANS. A continuación se muestra una tabla con la asignación de cada variable a su respectivo clúster:

Cluster.1	Cluster.2	Cluster.3	Cluster.4	Cluster.5	Cluster.6
cairsus	Length.cm.	AvgDiam.mm.	Cesus	ProjArea.cm2.	pHsus
Phidrosus	Tips	L05a1	nahidrosus	germ8	dapsus
	Forks	PA05a1	Khidrosus	h1	
	L0a05	T05a1		h2	
	PA0a05	contrasus		h3	
	T0a05	craguamlsus		h4	
		csus		mosus	
		nsus		eptsus	
				cnsus	

Figura 17: Composición de los clúster para las variables tras análisis K-MEANS

En este caso, al tener más variables que observaciones (al transponer la matriz para estudiar las variables originales) no se puede desarrollar el gráfico donde se ve ilustrado los diferentes clusters con los casos que forman cada uno.

Como podemos ver en la tabla, los clusters creados por el método de K-MEANS no todos contienen las mismas variables que los formados por el método jerárquico. En primer lugar nos fijaremos en los grupos que contienen las mismas variables. Por ejemplo, existe un clúster que se forma de igual manera en ambos métodos, estos clusters contienen las variables *pHsus* y *dapsus*. Estas dos variables explican características del sustrato, en concreto el pH que contiene y la densidad del sustrato respectivamente. Atendiendo a los resultados, y justificando los resultados del análisis clúster realizado anteriormente, podemos concluir que estas dos variables muestran un comportamiento muy similar, por lo que se deduce que a partir de la medición del pH existente en el sustrato se pueden extraer indicios de la densidad existente en el mismo. De igual modo ocurre con el clúster número 4 compuesto por las variables: *Cesus, nahidrosus* y *Khidrosus*.

Por otro lado, vemos que existen clusters que han variado las variables con respecto al último análisis jerárquico que se ha realizado. Esto quiere decir que al recalcular los centroides de los clusters que se van formando (procedimiento llevado a cabo por el método K-MEANS), estas variables pueden estar más cerca de un centroide calculado posteriormente (en iteraciones posteriores al inicio) que de las variables que forman los respectivos clusters en análisis jerárquico.

Una vez analizadas las relaciones de las variables y observar que no existe ningún comportamiento extraño, se va a dar paso al análisis de las observaciones, que será donde se extraerán las conclusiones oportunas sobre el comportamiento de los sustratos.

5.2.2. Análisis clúster para las observaciones

En este apartado se estudiará el comportamiento de las observaciones a la hora de ser agrupadas. Como ya se ha dicho anteriormente, estas observaciones son combinaciones de diferentes tipos de sustrato, diferentes porcentajes aplicados al compost y diferentes plantas estudiadas. Con esto se quiere extraer conclusiones sobre cómo afecta a cada variable (o grupos de variables) el tipo de sustrato y porcentaje que se le aplica a su correspondiente grupo.

Como se ha podido observar en la primera tabla donde se mostraba la composición de los datos, tenemos tres observaciones por cada combinación. Una forma de afrontar este problema, cuando existe más de una observación para el mismo parámetro, es transformar la base de datos para tener solamente una observación para cada combinación. Para ello se ha llevado a cabo un reajuste de los datos quedando cada observación como variable, por lo que resultarán tres subvariables por cada variable original. A continuación, se mostrará una tabla resultante tras este reajuste:

tratamientos_planta	Length.cm.O1	Length.cm.O2	Length.cm.O3	ProjArea.cm2.O1	ProjArea.cm2.O2
C60-25% /brocoli	927	1065	899	27	30
C60-25% /lechu	655	662	714	24	26
C60-50% /brocoli	897	1199	799	27	36
C60-50% /lechu	495	684	448	18	25
C60-75% /brocoli	437	451	614	11	13
C60-75% /lechu	536	388	414	17	16

tratamientos_planta	ProjArea.cm2.O3	AvgDiam.mm.O1	AvgDiam.mm.O2	AvgDiam.mm.O3	Tips.O1
C60-25% /brocoli	23	0.29	0.28	0.28	3473
C60-25% /lechu	29	0.37	0.40	0.40	516
C60-50% /brocoli	23	0.29	0.30	0.29	1267
C60-50% /lechu	18	0.37	0.36	0.39	360
C60-75% /brocoli	16	0.25	0.28	0.25	3188
C60-75% /lechu	14	0.40	0.41	0.36	345

Figura 18: Estructura de los datos una vez ajustados

Vemos que las columnas tienen tres observaciones (O1, O2, y O3) para cada variable. De este modo, teniendo solamente una observación para cada combinación de tipo de compost, porcentaje aplicado y planta, tenemos 88 observaciones, es decir, 88 combinaciones diferentes. Por lo tanto, ya se está en disposición de realizar un análisis clúster correctamente. Para ello seguiremos el mismo proceso que para analizar las variables. En primer lugar, se analizarán realizando un análisis clúster jerárquico, y en segundo lugar, se llevará a cabo un estudio a través del método no jerárquico K-MEANS.

5.2.2.1. Análisis clúster jerárquico para las observaciones

En este apartado se mostrará el proceso llevado a cabo para analizar el comportamiento de los casos con el método jerárquico. Para ello se utilizarán las

mismas funciones de R que se han utilizado para analizar las variables.

En este apartado también se ha utilizado el *coeficiente de aglomeración* para elegir la medida de distancia y el método de agrupación más adecuado. En este caso, al igual que para las variables, la medida de distancia utilizada es euclídea, y la estrategia de agrupación será Ward. Esta combinación proporciona un *coeficiente de aglomeración* del 0.9346, por lo que se puede afirmar que se trata de un ajuste con una calidad bastante elevada.

A continuación, se muestra el proceso llevado a cabo para el análisis de clúster jerárquico:

```
datos_tipicicados<-scale(datos_originales)
matriz_de_distancias<-dist(datos_tipificados, method = "medida_de_distancia")
analisis_cluster<- hclust(matriz_de_distancia, method = "metodo_de_agrupacion")
dendograma<-plot(analisis_cluster)
```

Utilizando la medida de distancia y la estrategia de agrupación anteriormente mencionadas obtenemos el siguiente dendograma:

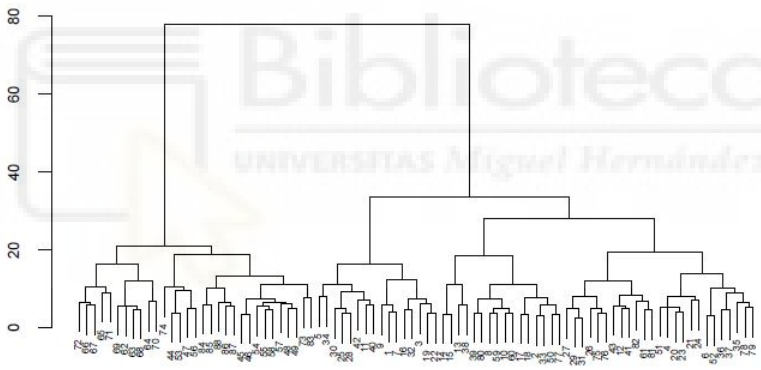

Figura 19: Dendograma para las variables tras el análisis jerárquico

En el presente dendograma se recogen todas las observaciones existentes en la base de datos, que son un total de 88. Por esta razón, visualmente no se puede exponer el nombre de cada una de ellas en el gráfico, por lo que están numeradas. Llegados a este punto hay que discutir el número de clúster más apropiados para poder explicar correctamente el comportamiento de las observaciones. Analizando el dendograma, a primera vista se puede observar que se generan 5 grandes grupos que diferencian las observaciones de forma homogénea entre ellas. A continuación, se va a ejecutar la función que agrupa las observaciones en diferentes grupos según el método y se verá si son los grupos que a primera vista parecen evidentes.

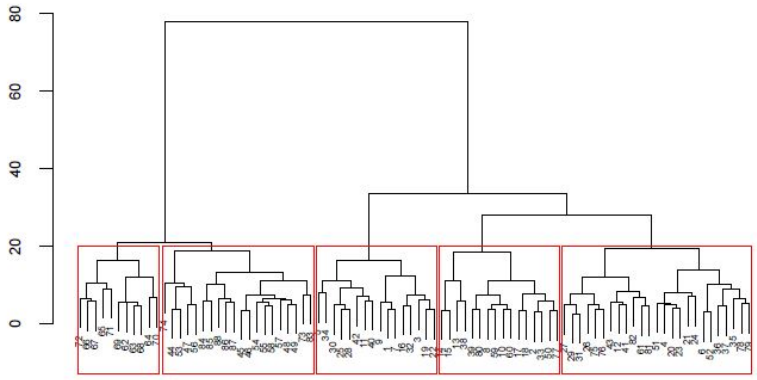

Figura 20: Dendograma para las variables agrupadas en 5 clusters tras el análisis jerárquico

Tras aplicar la función que añade los rectángulos rojos al dendograma, vemos que los grupos creados por R son los inicialmente más evidentes. Por esta razón no se realizarán más dendogramas cambiando el número de clusters, ya que los 5 grupos creados explican de manera coherente el comportamiento de las observaciones. Y, además, es un numero de clusters correcto para poder empezar el análisis no jerárquico K-MEANS en el siguiente punto. Aquí será donde se podrá ver con mas claridad los grupos gracias a una serie de gráficos y tablas que R genera para el análisis K-MEANS.

5.2.2.2. Análisis clúster no jerárquico para las observaciones (K-MEANS)

En este punto se desarrollará un análisis clúster no jerárquico utilizando el método de las K-MEANS para analizar el comportamiento de las observaciones. Como se ha hecho anteriormente, se mostrará tanto el proceso llevado a cabo como las funciones utilizadas. En este punto se profundizará más en el análisis, ya que es donde podremos extraer conclusiones más exactas a la hora de concluir sobre el comportamiento de las observaciones a la hora de ser agrupadas y así poder decidir sobre que observación o grupo de observaciones proporciona mejores cualidades en el sustrato. En este punto se desarrollarán una serie de tablas donde se podrá extraer información muy valiosa para poder extraer conclusiones. Estas tablas recogen para cada clúster los valores medios y desviaciones típicas de cada variable. Esto hace que facilite la observación de clusters muy diferentes al resto, y por lo tanto comportamientos de grupos de observaciones (clusters) diferentes.

A continuación, se mostrará el proceso llevado a cabo para el análisis clúster no jerárquico K-MEANS.

Como se ha explicado en la parte teórica, este método necesita una previa

elección de los puntos de semilla, para ello elegimos el método de elección de puntos de semilla que toma como tal las primeras k observaciones para k puntos de semilla. Esto se lleva a cabo con la función:

```
set.seed(1234)
```

Esto es necesario, ya que, sin esta función, cada vez que ejecutemos el análisis de K-MEANS este proporcionará un resultado diferente. Con esta fijación de los puntos de semilla, R siempre proporcionará el mismo resultado.

En primer lugar, se desarrollará el análisis con 5 clusters, ya que atendiendo al dendograma del análisis jerárquico parece que es lo correcto. La función que lleva a cabo el analisis es la siguiente:

```
analisis_cluster<-kmeans(datos_originales, nuemro_de_clusters)
```

Una vez realizado el análisis, este método proporciona un gráfico muy visual con el que se pueden extraer conclusiones sobre el número óptimo de clusters comparando con otros análisis llevados a cabo con diferente número de clusters. Este gráfico es el siguiente:

```
autoplot(analisis_kmeans, data = datos ,label = TRUE, label.size
= 3, frame = TRUE)
```

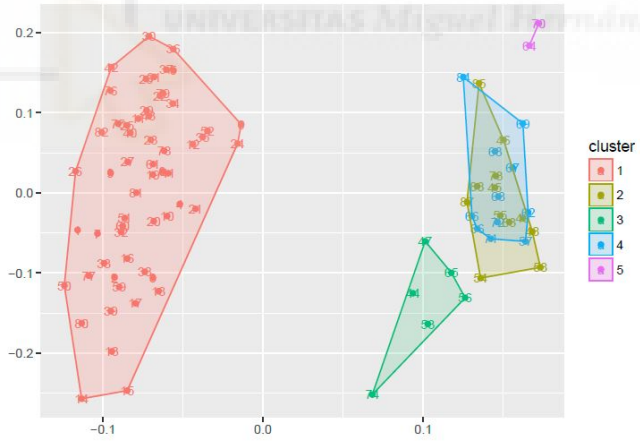

Figura 21: Gráfico de K-MEANS para todas las observaciones agrupadas en 5 clusters

Observando el gráfico resultante con el número de clusters que parecían más apropiados con el análisis clusters jerárquico (5), se puede ver que existen dos clusters superpuestos. Eso significa que las características de las observaciones que pertenecen a ambos clusters son muy parecidas. Además existe un clúster en el que solamente se incluyen dos observaciones, por lo que volveremos a realizar el análisis probando con solamente tres clusters.

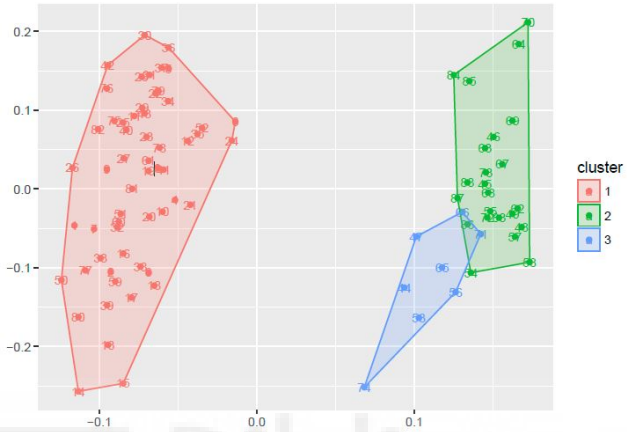

Figura 22: Gráfico de K-MEANS para todas las observaciones agrupadas en 3 clusters

Con este gráfico se hace evidente que es más coherente llevar a cabo el estudio del comportamiento de las observaciones agrupándolas en tres clusters solamente, ya que quedan mejor explicadas (los clusters generados son más heterogéneos entre ellos).

A continuación, se mostrarán las tablas con la composición de cada clúster, para ver si en un principio los clusters discriminan teniendo en cuenta algún aspecto de las observaciones (por tipo de planta o por porcentaje de compost utilizado).

brocoli	escarola	lechu	pimiento
19	19	19	0
0	0	0	23
0	0	0	8

Figura 23: Composición de los clusters según el tipo de planta

Estudiando esta tabla se observa que en el clúster número uno se incluyen las siguientes plantas: brócoli, escarola y lechuga, discriminando en su totalidad

al pimiento, el cual tiene sus observaciones divididas entre el clusters número dos y tres. Por esta razón, se puede afirmar que el comportamiento del pimiento es diferente al resto de plantas. Esta conclusión lleva a realizar el estudio por separado, es decir, por un lado, realizar un análisis clúster para las tres plantas del clúster 1 y por otro con el pimiento, ya que al estudiar conjuntamente las plantas que se comportan de diferente manera puede llevar a resultados confusos. De esta manera se podrían confundir variables intrínsecas de cada planta como diferencias significativas en los clusters. Por ejemplo, podríamos detectar como variable con diferencias significativas respecto al resto, la altura de la planta, y esta llevarnos a confusión, ya que quizás las plantas de brócoli, escarola y lechuga tengan una medida similar, y las plantas de pimiento tengan unas medidas naturales muy diferentes. Por esta razón se ve necesario realizar un análisis por separado. Terminaremos el análisis de forma conjunta para ver el proceso seguido y en el apartado siguiente se realizará el análisis correspondiente.

A continuación, se muestra una tabla similar a la anterior pero discriminando por porcentaje de compost utilizado.

0	25	50	75
3	18	18	18
0	4	9	10
1	6	1	0

Figura 24: Composición de los clusters según el porcentaje aplicado de compost

Atendiendo a la tabla que discrimina por porcentaje de compost utilizado, vemos que a primera vista no sigue un patrón por el cual discrimine con una coherencia detectable. Esto quiere decir que el porcentaje de compost no influye a la hora de agrupar las observaciones. Por esta razón es necesario profundizar en el estudio del comportamiento de los tipos de compost.

Para profundizar en el comportamiento de las observaciones (tipos de compost), a continuación, se realizarán unas tablas compuestas por las medias y desviaciones típicas de cada variable dependiendo al clúster en el que estén agrupados. De este modo, fijándonos en las observaciones que comprenda dichos clústers podremos extraer conclusiones sobre el comportamiento de estos.

Para llevar a cabo dicho proceso, es necesario unificar las variables, ya que teníamos tres subvariables por cada variable original. Para ello se realizará una media de los valores de las tres subvariables para obtener un valor por cada variable original.

La tabla resultante es la siguiente:

	Cluster-1	Cluster-2	Cluster-3	p-value	C1 vs C2	C1 vs C3	C2 vs C3
Length.cm.	622.433 ± 161.245	159.072 ± 22.803	181.875 ± 22.130	0.00	0.00	0.00	0.02
ProjArea.cm2.	22.848 ± 7.171	10.843 ± 2.211	13.746 ± 3.217	0.00	0.00	0.00	0.03
AvgDiam.mm.	0.364 ± 0.074	0.663 ± 0.063	0.694 ± 0.147	0.00	0.00	0.00	0.57
Tips	1032.520 ± 976.044	356.087 ± 61.960	406.229 ± 118.045	0.00	0.00	0.20	0.21
Forks	3040.404 ± 1239.481	521.275 ± 146.135	746.333 ± 165.962	0.00	0.00	0.00	0.01
L0a05	83.333 ± 7.169	37.225 ± 6.893	40.492 ± 14.421	0.00	0.00	0.00	0.73
L05a1	13.095 ± 5.315	50.977 ± 5.354	35.704 ± 5.948	0.00	0.00	0.00	0.00
PA0a05	57.754 ± 10.303	19.038 ± 5.838	17.504 ± 13.617	0.00	0.00	0.00	0.46
PA05a1	26.665 ± 6.638	57.906 ± 7.686	35.000 ± 18.581	0.00	0.00	0.03	0.00
T0a05	97.971 ± 1.756	87.600 ± 2.661	91.125 ± 2.015	0.00	0.00	0.00	0.00
T05a1	1.859 ± 1.534	11.378 ± 2.418	7.921 ± 2.056	0.00	0.00	0.00	0.00
germ8	94.466 ± 5.095	82.174 ± 13.319	85.792 ± 7.502	0.00	0.00	0.00	0.78
h1	4.245 ± 1.627	0.652 ± 0.129	0.838 ± 0.077	0.00	0.00	0.00	0.00
h2	6.917 ± 3.071	1.033 ± 0.163	1.104 ± 0.167	0.00	0.00	0.00	0.34
h3	8.775 ± 3.697	2.220 ± 0.559	2.862 ± 0.280	0.00	0.00	0.00	0.00
h4	9.373 ± 3.981	3.958 ± 1.025	4.713 ± 0.516	0.00	0.00	0.02	0.03
pHsus	6.450 ± 0.314	6.990 ± 0.624	6.922 ± 0.568	0.00	0.00	0.01	0.95
Cesus	2.133 ± 1.104	2.627 ± 1.063	1.023 ± 0.488	0.00	0.06	0.00	0.00
mosus	67.207 ± 10.162	63.078 ± 6.600	75.838 ± 9.527	0.01	0.10	0.03	0.00
dapsus	0.198 ± 0.057	0.259 ± 0.085	0.219 ± 0.062	0.01	0.00	0.52	0.34
eptsus	87.481 ± 2.925	84.196 ± 4.865	85.688 ± 3.857	0.01	0.00	0.27	0.43
cairsus	38.846 ± 7.907	33.292 ± 4.506	33.375 ± 6.156	0.00	0.00	0.10	0.43
contrasus	10.589 ± 6.626	22.951 ± 4.758	24.879 ± 2.930	0.00	0.00	0.00	0.20
craguamlsus	486.333 ± 85.878	516.435 ± 64.545	530.333 ± 42.944	0.11	0.08	0.14	0.67
cnsus	17.111 ± 6.674	13.897 ± 2.641	23.350 ± 9.918	0.00	0.04	0.02	0.00
csus	22.608 ± 4.731	38.401 ± 4.522	46.775 ± 1.856	0.00	0.00	0.00	0.00
nsus	1.422 ± 0.442	2.844 ± 0.440	2.220 ± 0.563	0.00	0.00	0.00	0.01
nahidrosus	603.246 ± 271.085	600.754 ± 232.704	235.167 ± 106.113	0.00	0.97	0.00	0.00
Khidrosus	1049.737 ± 830.583	1797.797 ± 988.219	824.333 ± 522.532	0.00	0.00	0.82	0.01
Phidrosus	679.789 ± 453.008	336.913 ± 214.514	154.292 ± 147.688	0.00	0.00	0.00	0.03

Figura 25: Tabla de medias y desviaciones tipicas para todas las observaciones

La tabla anterior da una visualización muy concreta sobre el comportamiento de las observaciones en cada grupo. Esto nos ayuda a poder extraer conclusiones más exactas. Como podemos ver en la tabla, existen variables con valores medios muy diferentes dependiendo al clúster en los que se encuentra. Por ejemplo, la primera variable que encontramos con diferencias significativas es *Length* (Altura), ya que en el clúster 1 toma un valor medio de 622.43 y en el resto de clusters están muy por debajo de esa media (159 y 181 para el clúster 2 y 3 respectivamente). Esto, en condiciones normales de análisis (ya se ha explicado anteriormente el problema que se puede tener al realizar un mismo estudio con tipos de plantas diferentes que pueden tener aspectos intrínsecos muy diferenciados), se trataría de una variable discriminatoria a tener en cuenta. Cuando tenemos una de estas variables con diferencias significativas, es necesario ver que observaciones pertenecen a ese clúster en concreto (lo cual aparece en la siguiente tabla) para poder concluir que las correspondientes observaciones presentan una mayor o menor medida de una variable concreta.

Cluster.1	Cluster.2	Cluster.3
C60-25% /brocoli	AP 50%/pimiento	AP 25%/pimiento
C60-25% /lechu	AP 75%/pimiento	AP Lav 25%/pimiento
C60-50% /brocoli	AP Lav 50%/pimiento	C61 25%/pimiento
C60-50% /lechu	AP Lav 75%/pimiento	C61 Lav 25%/pimiento
C60-75% /brocoli	C61 50%/pimiento	CIG1 Lav 25%/pimiento
C60-75% /lechu	C61 75%/pimiento	CIG1 Lav 50%/pimiento
C62-25% /brocoli	C61 Lav 50%/pimiento	PL Lav 25%/pimiento
C62-25% /lechu	C61 Lav 75%/pimiento	turba 0%/pimiento
C62-50% /brocoli	CIG1 25%/pimiento	
C62-50% /lechu	CIG1 50%/pimiento	
C62-75% /brocoli	CIG1 75%/pimiento	
C62-75% /lechu	CIG1 Lav 75%/pimiento	
turba 0%/brocoli	PL 25%/pimiento	
turba 0%/escarola	PL 50%/pimiento	
turba 0%/lechu	PL 75%/pimiento	
Z1-25%/brocoli	PL Lav 50%/pimiento	
Z1-25%/escarola	PL Lav 75%/pimiento	
Z1-25%/lechu	Zmix 25%/pimiento	
Z1-50%/brocoli	Zmix 50%/pimiento	
Z1-50%/escarola	Zmix 75%/pimiento	
Z1-50%/lechu	Zmix Lav 25%/pimiento	
Z1-75%/brocoli	Zmix Lav 50%/pimiento	
Z1-75%/escarola	Zmix Lav 75%/pimiento	
Z1-75%/lechu		
Z3-25% /brocoli		
Z3-25% /escarola		
Z3-25% /lechu		
Z3-50% /brocoli		
Z3-50% /lechu		
Z3-75% /brocoli		
Z3-75% /lechu		
Z4-25% /brocoli		
Z4-25% /lechu		
Z4-50% /brocoli		

Z4-50% /lechu
Z4-75% /brocoli
Z4-75% /lechu
Z6-25% /brocoli
Z6-25% /lechu
Z6-50% /brocoli
Z6-50% /lechu
Z6-75% /brocoli
Z6-75% /lechu
C60-25% /escarola
C60-50% /escarola
C60-75% /escarola
C62-25% /escarola
C62-50% /escarola
C62-75% /escarola
Z3-50% /escarola
Z3-75% /escarola
Z4-25% /escarola
Z4-50% /escarola
Z4-75% /escarola
Z6-25% /escarola
Z6-50% /escarola
Z6-75% /escarola

Figura 26: Composición de los clusters para todas las variables

Como podemos ver en esta tabla, los clusters 2 y 3 están formados solamente por el tipo de planta pimiento, otra razón que confirma para realizar un estudio por separado y así poder extraer conclusiones más acertadas sobre el comportamiento del compost y porcentaje de este en el sustrato. Además, se observa que los tipos de compost utilizados para el pimiento son diferentes a los usados en el brócoli, escarola y lechuga.

Una vez expuesto el proceso que se va a llevar a cabo para analizar las observaciones con el análisis no jerárquico K-MEANS, se va a exponer el estudio separando en dos bases de datos. Las observaciones de brócoli, escarola y lechuga, por un lado, y, por otro lado, las observaciones del pimiento. En estos análisis solo se va a exponer el análisis no jerárquico K-MEANS, ya que es el más apropiado y el método que mas facilidades proporciona a la hora de poder concluir con más certeza.

5.2.2.2.1. Análisis de las observaciones para los cultivos de lechuga, brócoli y escarola

En este apartado se realizará un análisis clúster (no jerárquico) solamente para las observaciones de lechuga, brócoli y escarola. Los tipos de compost utilizados en los cultivos, son los mimos para estos tres tipos de plantas, por lo que se podrá concluir con más certeza sobre el comportamiento de los mismos (turba con 0 %; C60 con 25 %, 50 % y 75 %; C62 con 25 %, 50 % y 75 %; Z1 con

25 %, 50 % y 75 %; Z3 con 25 %, 50 % y 75 %; Z4 con 25 %, 50 % y 75 %; Z6 con 25 %, 50 % y 75 %).

Análisis no Jerárquico (K-MEANS)

Para llevar a cabo este análisis se empezará el estudio con 5 grupos, y así poder disminuir si el proceso lo requiere. El gráfico resultante con 5 clusters es el siguiente:

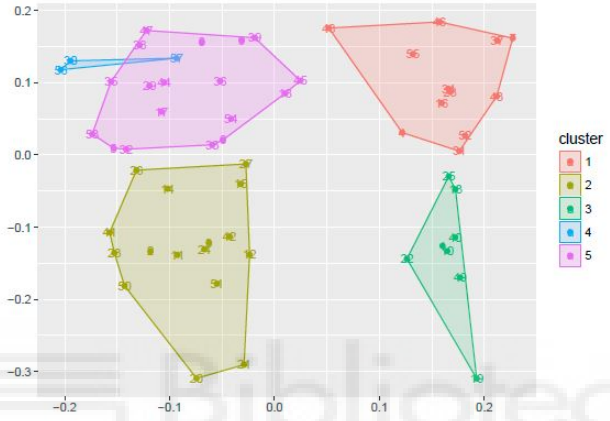

Figura 27: Gráfico K-MEANS para las observaciones de lechuga, brócoli y escarola agrupadas en 5 clusters

Como podemos observar el clúster numero 4 está compuesto por muy pocas observaciones y además esta superpuesto al clúster número 5, por estas razones se va a contrastar los resultados realizando el análisis para 4 clusters. El gráfico resultante para este análisis es el siguiente:

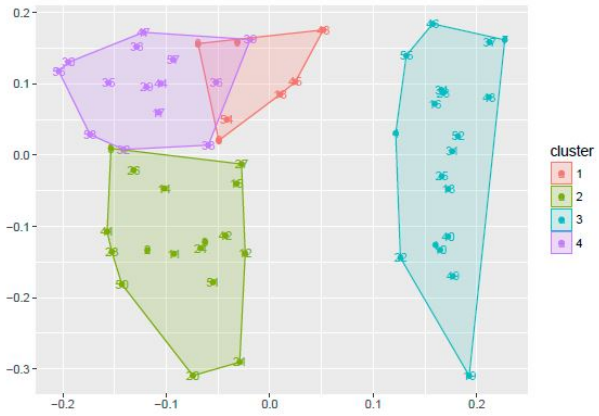

Figura 28: Gráfico K-MEANS para las observaciones de lechuga, brócoli y escarola agrupadas en 4 clusters

Podemos observar que del mismo modo, sigue existiendo un clúster (numero 1) superpuesto casi en su totalidad. Por lo que seguimos disminuyendo el número de clúster a 3. El gráfico resultante es el siguiente:

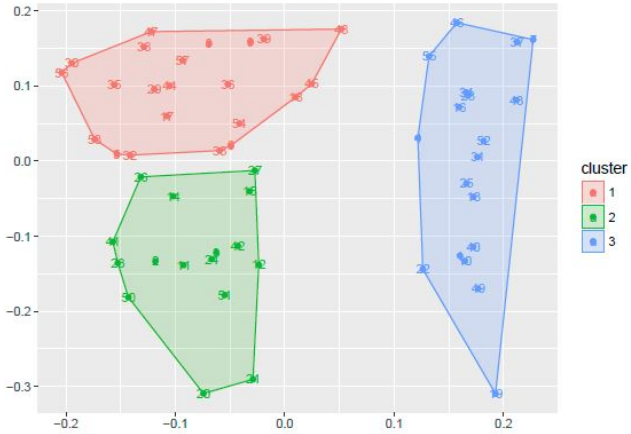

Figura 29: Gráfico K-MEANS para las observaciones de lechuga, brócoli y escarola agrupadas en 3 clusters

Con este gráfico vemos que los clusters creados están bien diferenciados entre ellos, por lo que se trata de clusters muy heterogéneos entre ellos, aspecto que es fundamental en un correcto análisis. Una vez que tenemos decidido el número de clusters en los que dividir nuestro conjunto de observaciones, estamos en disposición de estudiar conclusiones, con la ayuda de las siguientes tablas:

	Cluster-1	Cluster-2	Cluster-3	p-value	C1 vs C2	C1 vs C3	C2 vs C3
Length.cm.	568.470 ± 141.161	618.417 ± 108.820	688.298 ± 199.048	0.14	0.22	0.07	0.46
ProjArea.cm2.	24.424 ± 7.560	25.083 ± 6.112	19.140 ± 6.340	0.03	0.78	0.04	0.01
AvgDiam.mm.	0.414 ± 0.050	0.400 ± 0.038	0.275 ± 0.021	0.00	0.43	0.00	0.00
Tips	383.727 ± 107.972	441.688 ± 81.270	2281.298 ± 688.389	0.00	0.02	0.00	0.00
Forks	2621.591 ± 865.850	2615.938 ± 772.088	3882.789 ± 1511.509	0.01	0.96	0.01	0.01
L0a05	78.894 ± 5.621	79.812 ± 4.151	91.439 ± 2.055	0.00	0.82	0.00	0.00
L05a1	16.983 ± 3.474	15.317 ± 2.705	6.723 ± 1.412	0.00	0.17	0.00	0.00
PA0a05	52.212 ± 8.848	52.646 ± 5.429	68.474 ± 5.687	0.00	1.00	0.00	0.00
PA05a1	31.271 ± 2.78	30.110 ± 3.88	18.430 ± 2.73	0.00	0.19	0.00	0.00
T0a05	96.788 ± 1.453	97.188 ± 0.798	100.000 ± 0.000	0.00	0.45	0.00	0.00
T05a1	2.932 ± 1.269	2.467 ± 0.742	0.105 ± 0.059	0.00	0.25	0.00	0.00
germ8	93.934 ± 4.007	97.271 ± 1.981	92.719 ± 6.944	0.03	0.01	0.92	0.06
h1	4.758 ± 1.309	5.609 ± 1.083	2.502 ± 0.279	0.00	0.05	0.00	0.00
h2	8.636 ± 1.894	9.059 ± 1.622	3.122 ± 0.271	0.00	0.52	0.00	0.00
h3	11.154 ± 2.271	10.906 ± 1.972	4.224 ± 0.667	0.00	0.54	0.00	0.00
h4	11.976 ± 2.183	11.799 ± 1.867	4.315 ± 0.564	0.00	0.66	0.00	0.00
pHsus	6.502 ± 0.326	6.379 ± 0.297	6.450 ± 0.320	0.78	0.49	0.76	0.70
Cesus	2.718 ± 1.078	1.330 ± 0.466	2.133 ± 1.124	0.00	0.00	0.06	0.02
mosus	60.979 ± 6.213	75.771 ± 8.238	67.207 ± 10.349	0.00	0.00	0.04	0.01
dapsus	0.236 ± 0.038	0.145 ± 0.028	0.198 ± 0.058	0.00	0.00	0.03	0.01
eptsus	85.564 ± 2.155	90.117 ± 1.411	87.481 ± 2.979	0.00	0.00	0.03	0.01
cairsus	40.724 ± 7.509	36.263 ± 8.024	38.846 ± 8.052	0.16	0.06	0.41	0.30
contrasus	7.979 ± 6.558	14.179 ± 5.032	10.589 ± 6.748	0.01	0.00	0.18	0.09
craguamlsus	448.485 ± 77.383	538.375 ± 70.584	486.333 ± 87.453	0.00	0.00	0.15	0.07
cnsus	13.648 ± 2.824	21.871 ± 7.613	17.111 ± 6.797	0.00	0.00	0.07	0.02
csus	22.284 ± 5.520	23.055 ± 3.563	22.608 ± 4.818	0.67	0.38	0.70	0.63
nsus	1.651 ± 0.427	1.107 ± 0.221	1.422 ± 0.450	0.00	0.00	0.06	0.02
nahidrosus	778.091 ± 204.196	362.833 ± 134.305	603.246 ± 276.059	0.00	0.00	0.03	0.01
Khidrosus	1328.788 ± 952.668	666.042 ± 418.510	1049.737 ± 845.824	0.05	0.02	0.29	0.18
Phidrosus	863.848 ± 467.157	426.708 ± 295.336	679.789 ± 461.321	0.02	0.00	0.21	0.11

Figura 30: Tabla de medias y desviaciones típicas para las observaciones de lechuga, brócoli y escarola

.

Cluster.1	Cluster.2	Cluster.3
C60-50% /escarola	C60-25% /escarola	C60-25% /brocoli
C60-50% /lechu	C60-25% /lechu	C60-50% /brocoli
C60-75% /escarola	C62-25% /escarola	C60-75% /brocoli
C60-75% /lechu	C62-25% /lechu	C62-25% /brocoli
C62-75% /escarola	C62-50% /escarola	C62-50% /brocoli
C62-75% /lechu	C62-50% /lechu	C62-75% /brocoli
Z1-75%/escarola	turba 0%/escarola	turba 0%/brocoli
Z1-75%/lechu	turba 0%/lechu	Z1-25%/brocoli
Z3-25% /escarola	Z1-25%/escarola	Z1-50%/brocoli
Z3-25% /lechu	Z1-25%/lechu	Z1-75%/brocoli
Z3-50% /escarola	Z1-50%/escarola	Z3-25% /brocoli
Z3-50% /lechu	Z1-50%/lechu	Z3-50% /brocoli
Z3-75% /escarola	Z4-25% /escarola	Z3-75% /brocoli
Z3-75% /lechu	Z4-25% /lechu	Z4-25% /brocoli
Z4-50% /escarola	Z6-25% /escarola	Z4-50% /brocoli
Z4-50% /lechu	Z6-25% /lechu	Z4-75% /brocoli
Z4-75% /escarola		Z6-25% /brocoli
Z4-75% /lechu		Z6-50% /brocoli
Z6-50% /escarola		Z6-75% /brocoli
Z6-50% /lechu		
Z6-75% /escarola		
Z6-75% /lechu		

Figura 31: Composición de clusters para las observaciones de lechuga, brócoli y escarola

Una vez llegados a este punto, ya estamos en disposición de extraer conclusiones sobre el comportamiento de las variables de lechuga, brócoli y escarola.

Atendiendo a la tabla anterior, vemos que existen variables que presentan diferencias en los valores medios dependiendo del clúster en el que se mide, y variables que tienen aproximadamente los mismos valores en todos los clusters. En primer lugar, vemos que la variable *pHsus* presenta valores muy similares en los tres clusters, por lo que quiere decir, que el pH del sustrato no depende de la variación ni del tipo de compost, ni del porcentaje que se le aplica. Por esta razón, se puede concluir que el pH del sustrato permanece constante en todas las combinaciones (planta, tipo de compost y porcentaje) analizadas. De la misma forma ocurre con la variable *csus* (concentración total de carbono en el sustrato).

En segundo lugar, vamos a analizar las variables que muestran diferencias significativas entre los diferentes clusters. Atendiendo a las características de la planta, vemos que la variable *Tips* presenta una diferencia bastante significativa de los dos primeros clusters con respecto al tercero. Vemos que en el tercer clusters el valor medio de ápices (puntas de raíz) por planta es mucho mayor al resto. Esto nos lleva a la conclusión de que las observaciones que pertenecen al clúster número tres tienen un número de ápices mucho mayor a los casos comprendidos en los otros dos clusters. Las observaciones del clúster 3 lo forman todas las combinaciones de compost y porcentaje pertenecientes a la planta del brócoli. Esto quiere decir que el brócoli, independientemente del tipo de compost y porcentaje, siempre desarrollará un número elevado de ápices.

Atendiendo a las características del sustrato, vemos que las variables que explican la combinación de NPK presentan diferencias significativas, quedando el clusters número 2 por debajo del resto en las tres variables (*nahidrosus, Khidrosus, Phidrosus*). Por lo que las observaciones pertenecientes al clúster 2 presentan deficiencias en este aspecto. Fijándonos en la tabla de la composición de los clusters, vemos que en dicho clúster el tipo de compost Z3 no aparece en ninguna de sus maneras, por lo que en principio se puede concluir que este tipo de compost es bueno para tener un NPK elevado. Además el clúster número 1 muestra unos valores mínimamente superiores al clúster 3, lo que nos lleva a pensar que este tipo de compost funciona mejor en lechuga y escarola que en brócoli. Como podemos ver, en el clúster número dos se engloban el cultivo con turba de lechuga y escarola, por lo que se puede afirmar que las combinaciones englobadas en el clúster número 1 pueden ser buenos sustitutos a la turba para el cultivo de lechuga y escarola.

Por último, vemos que existen dos variables donde los valores medios más bajos están en el clúster número 1 y, además, los valores más elevados se encuentran en el clúster numero 2 (la contracción del sustrato: *contrasus* y la relación existente de carbono/nitrógeno: *cnsus*). Fijándonos en la tabla de la composición de los clusters, vemos que la mayoría de las combinaciones del clúster 1 están formadas con un porcentaje de compost elevado (50 % y 75 %), solamente aparece el porcentaje de 25 % para el compost Z3. Esto nos lleva a concluir que aplicando un 50 % o más de cualquier compost analizado baja el valor de las variables nombradas (*contrasus* y *cnsus*). Esta es otra de las razones por la que se puede concluir que el compost Z3 es buen sustituto a la turba, y además aplicando solamente un 25 %, ya que a la hora de ser comercializado conviene un porcentaje aplicado mínimo para poder obtener una mayor rentabilidad al producto.

5.2.2.2.2. Análisis de las observaciones para el cultivo del pimiento

En este punto se realizará el mismo estudio que anteriormente, pero en este caso nuestras observaciones esta compuestas solamente por la planta de pimiento. Para este tipo de planta se ha recogido información con los siguientes tipos de compost y porcentajes:

1. 1. Turba con 0 %.
2. 2. AP con 25 %, 50 % y 75 %.
3. 3. AP Lav con 25 %, 50 % y 75 %.
4. 4. PL con 25 %, 50 % y 75 %.
5. 5. PL Lav con 25 %, 50 % y 75 %.
6. 6. C61 con 25 %, 50 % y 75 %.
7. 7. C61 Lav con 25 %, 50 % y 75 %.
8. 8. CIG1 con 25 %, 50 % y 75 %.

9. 9. CIG1 Lav con 25 %, 50 % y 75 %.
10. 10. Zmix con 25 %, 50 % y 75 %.
11. 11. Zmix Lav con 25 %, 50 % y 75 %.

Por lo que tendremos un total de 31 observaciones, correspondientes a los tres diferentes porcentajes de cada uno de los 10 tipos de compost, más una observación de turba.

Análisis no Jerárquico (K-MEANS)

Del mismo modo que los anteriores análisis, se mostrará el grafico con el número de clusters más apropiado en relación a las observaciones que se tienen, y a continuación se extraerán conclusiones con la visualización de la tabla de medias y desviaciones típicas y la tabla de la composición de los clusters. El gráfico resultante con el número de clusters óptimos, en este caso son 3, es el siguiente:

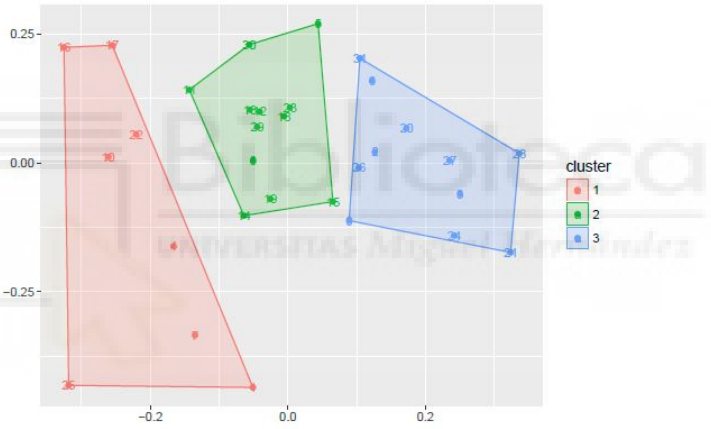

Figura 32: Gráfico K-MEANS para las observaciones del pimiento

Como podemos observar, siendo tres el número de clusters a formar, se comportan de forma heterogénea entre ellos, por lo que se puede concluir que tres clusters es un buen número en el que dividir las observaciones que tenemos para el pimiento. Como resultado de esta discriminación obtenemos la siguiente tabla de medias y desviaciones típicas:

	Cluster-1	Cluster-2	Cluster-3	p-value	C1 vs C2	C1 vs C3	C2 vs C3
Length.cm.	181.875 ± 22.130	168.472 ± 20.830	148.818 ± 21.093	0.01	0.15	0.01	0.03
ProjArea.cm2.	13.746 ± 3.217	12.092 ± 1.451	9.482 ± 2.128	0.00	0.20	0.01	0.00
AvgDiam.mm.	0.694 ± 0.147	0.686 ± 0.060	0.638 ± 0.059	0.31	0.97	0.34	0.13
Tips	406.229 ± 118.045	324.278 ± 62.761	390.788 ± 39.679	0.06	0.10	0.65	0.02
Forks	746.333 ± 165.962	599.694 ± 124.671	435.727 ± 120.229	0.00	0.08	0.00	0.01
L0a05	40.492 ± 14.421	34.075 ± 7.063	40.661 ± 4.977	0.15	0.42	0.84	0.03
L05a1	35.704 ± 5.948	51.972 ± 7.121	49.891 ± 2.190	0.00	0.00	0.00	0.44
PA0a05	17.504 ± 13.617	17.186 ± 6.119	21.058 ± 5.020	0.31	0.67	0.39	0.12
PA05a1	35.000 ± 18.581	56.325 ± 9.771	59.630 ± 4.322	0.00	0.01	0.00	0.64
T0a05	91.125 ± 2.015	86.322 ± 2.252	88.994 ± 2.430	0.00	0.00	0.04	0.01
T05a1	7.921 ± 2.056	12.447 ± 2.093	10.212 ± 2.274	0.00	0.00	0.02	0.02
germ8	85.792 ± 7.502	89.167 ± 7.403	74.545 ± 14.397	0.01	0.31	0.10	0.01
h1	0.838 ± 0.077	0.703 ± 0.126	0.597 ± 0.114	0.00	0.01	0.00	0.03
h2	1.104 ± 0.167	1.122 ± 0.134	0.936 ± 0.137	0.02	0.79	0.04	0.01
h3	2.862 ± 0.280	2.508 ± 0.569	1.906 ± 0.351	0.00	0.07	0.00	0.01
h4	4.713 ± 0.516	4.453 ± 1.095	3.418 ± 0.613	0.00	0.49	0.00	0.02
pHsus	6.922 ± 0.568	7.099 ± 0.575	6.870 ± 0.680	0.63	0.67	0.77	0.34
Cesus	1.023 ± 0.488	1.880 ± 0.568	3.442 ± 0.853	0.00	0.00	0.00	0.00
mosus	75.838 ± 9.527	66.414 ± 6.110	59.439 ± 5.182	0.00	0.02	0.00	0.01
dapsus	0.219 ± 0.062	0.272 ± 0.090	0.245 ± 0.082	0.47	0.25	0.65	0.48
eptsus	85.688 ± 3.857	83.214 ± 5.106	85.267 ± 4.580	0.42	0.26	0.84	0.31
cairsus	33.375 ± 6.156	31.916 ± 4.589	34.795 ± 4.093	0.23	0.13	0.90	0.19
contrasus	24.879 ± 2.930	22.350 ± 4.724	23.606 ± 4.934	0.37	0.18	0.39	0.64
craguamlsus	530.333 ± 42.944	521.833 ± 56.923	510.545 ± 74.350	0.88	0.85	0.59	0.90
cnsus	23.350 ± 9.918	15.417 ± 2.585	12.239 ± 1.481	0.00	0.01	0.00	0.00
csus	46.775 ± 1.856	40.656 ± 3.189	35.942 ± 4.588	0.00	0.00	0.00	0.01
nsus	2.220 ± 0.563	2.716 ± 0.425	2.983 ± 0.433	0.01	0.05	0.01	0.17
nahidrosus	235.167 ± 106.113	442.861 ± 108.016	773.000 ± 209.156	0.00	0.00	0.00	0.00
nahidrosus	235.167 ± 106.113	442.861 ± 108.016	773.000 ± 209.156	0.00	0.00	0.00	0.00
Khidrosus	824.333 ± 522.532	1251.778 ± 464.804	2393.455 ± 1078.317	0.00	0.06	0.00	0.01
Phidrosus	154.292 ± 147.688	202.028 ± 161.092	484.061 ± 164.436	0.00	0.46	0.00	0.00

Figura 33: Tabla de medias y desviaciones típicas para las observaciones del pimiento

Para complementar esta tabla y así poder obtener conclusiones de forma óptima, es necesario complementarla con la composición de los clusters:

Cluster.1	Cluster.2	Cluster.3
AP 25%/pimiento	AP Lav 50%/pimiento	AP 50%/pimiento
AP Lav 25%/pimiento	C61 50%/pimiento	AP 75%/pimiento
C61 25%/pimiento	C61 Lav 50%/pimiento	AP Lav 75%/pimiento
C61 Lav 25%/pimiento	C61 Lav 75%/pimiento	C61 75%/pimiento
CIG1 Lav 25%/pimiento	CIG1 25%/pimiento	PL 50%/pimiento
CIG1 Lav 50%/pimiento	CIG1 50%/pimiento	PL 75%/pimiento
PL Lav 25%/pimiento	CIG1 75%/pimiento	PL Lav 75%/pimiento
turba 0%/pimiento	CIG1 Lav 75%/pimiento	Zmix 25%/pimiento
	PL 25%/pimiento	Zmix 50%/pimiento
	PL Lav 50%/pimiento	Zmix 75%/pimiento
	Zmix Lav 25%/pimiento	Zmix Lav 75%/pimiento
	Zmix Lav 50%/pimiento	

Figura 34: Composición de clusters para las observaciones del pimiento

Una vez mostradas las tablas realizadas, ya podemos pasar a extraer las conclusiones oportunas. Como vemos en la tabla de medias y desviaciones, las variables que presentan diferencias en los valores medios de cada clúster, son muy similares a las que lo hacían también en el análisis de las plantas de lechuga, brócoli y escarola. Por un lado tenemos la variable *Tips*, que en este caso el valor medio más elevado pertenece al clúster número 1, aunque no muestra una diferencia muy significativa con respecto el clúster número 3. En el clúster 1 se recogen la mayoría de sustratos aplicados en un 25 %, excepto el compost CIG1 Lav. Con esto podemos decir que aplicando menos cantidad de compost, el número de ápices en las raíces aumenta para la planta del pimiento.

Fijándonos en las variables que explican características del sustrato, vemos que en la variable *Cesus* (conductividad eléctrica) existe una diferencia del clúster 3 con respecto los otros dos clusters. El valor medio de esta variable en el clúster 3 es más elevado que el resto, por lo que las observaciones (combinaciones de tipo de compost y porcentaje aplicado) pertenecientes a este grupo presentan una mayor conductividad (salinidad) en el sustrato.

Atendiendo ahora al resto de variables referentes al sustrato, vemos que existen un número de variables que no presentan diferencias entre clusters. Esto quiere decir que estas características no varían en la planta del pimiento con cambios en el tipo de compost utilizado, ni en el porcentaje que se le aplica al sustrato. Estas variables a las que nos referimos son: *dapsus, eptsus, contrasus, craguamlsus y nsus.*

Por último, analizando el NPK del sustrato, observamos que existen diferencias entre clusters en estas tres variables. De forma general, vemos que el clúster número 3 recoge los valores más elevados para estas tres variables, por esta razón las combinaciones de tipo de compost y porcentajes pertenecientes a dicho grupo son los idóneos para el cultivo del pimiento. Además, vemos que el clúster número 1, donde pertenece la turba, es el que presenta los valores más pequeños, esto hace evidente la sustitución del método tradicional de cultivo (turba) por una de las combinaciones pertenecientes al clúster 3. Además, observamos que la única combinación perteneciente al clúster 3 con un 25 % de compost aplicado es el compost *Zmix*, atendiendo a la rentabilidad de la comercialización del sustituto de la turba, podemos afirmar que esta combinación del sustrato tipo *Zmix* con un 25 % es la mejor opción para sustituir a la turba.

6. Conclusiones

En este apartado se van a exponer de forma conjunta las conclusiones y así poder dar respuesta a los objetivos marcados al principio para este estudio. A modo de recordatorio, el objetivo principal marcado es conocer que tipo o tipos de compost han mostrado mejores resultados para las variables utilizadas y analizadas en el escaneo de raíces llevado a cabo. Y, como objetivo secundario, se quiere analizar que técnicas del análisis utilizado son más adecuadas al tipo de datos que obtenemos.

En cuanto al objetivo secundario propuesto, se ve evidente que dentro del análisis clúster, a la hora de estudiar el comportamiento de los compost (observaciones) ha sido mucho más útil el análisis no jerárquico K-MEANS. Este ha proporcionado una serie de gráficos y tablas muy fáciles de interpretar a la hora de analizar el comportamiento de las observaciones. Gracias a estos gráficos y tablas se ha podido analizar de forma más certera que tipos de compost (con un concreto porcentaje aplicado al sustrato) serian los indicados para poder sustituir al medio tradicional de cultivo (turba). A demás, han sido muy visuales a la hora de estudiar la calidad del compost atendiendo a las características del NPK de cada uno de ellos, aspecto que es importante en este estudio. Dentro de este objetivo secundario nace la necesidad de concluir sobre, que medida de distancia y método de agrupación han sido las que mejores resultados han proporcionada para la estructura de datos disponible. En el presenta estudio se ha utilizado la distancia euclídea y el método de agrupación Ward.

Pasando a concluir sobre el objetivo principal, por un lado, se concluirá sobre los cultivos de brócoli, lechuga y escarola, y, por otro lado, sobre el cultivo del pimiento. Atendiendo al primer grupo de plantas, se ha llegado a la conclusión de que las observaciones que estaban agrupadas en el primer clúster eran las más indicadas para sustituir a la turba, en lo que al NPK se refiere y, además, es el único clúster donde no se engloba a la turba. Concretando un poco más, se puede observar que el compost Z3 para la lechuga y la escarola solo se engloba en dicho clúster, mientras que los demás también están englobados en el resto de clusters con sus diferentes porcentajes. Por esta razón, el compost Z3 aplicado un 25 % esta englobado en el primer clúster, esto lleva a concluir que esta combinación de observaciones es la mas indicada como posible sustituto a la turba para los cultivos de lechuga, brócoli y escarola. Ya que además de proporcionar una calidad muy buena al compost basada en el NPK, se podrá aplicar solamente un 25 % al compost, por lo que a términos de comercialización será mas eficiente.

Atendiendo al cultivo del pimiento, se obtiene como mejores observaciones las comprendidas en el tercer clúster de este estudio. En este clúster se engloba al compost Zmix en las tres combinaciones, y, además, es el único tipo englobado solamente en dicho clúster, el resto de las observaciones de este clúster tienen repartidas observaciones con diferentes porcentajes en los otros dos clusters restantes. Por lo que se puede concluir que este tipo de compost (Zmix) es el mas adecuado para sustituir al medio tradicional para el cultivo del pimiento. También se observa que, en el clúster numero 3 el único porcentaje del 25 % que aparece es para este compost. Esta razón hace concluir con mas rotundidad que esta combinación (Zmix al 25 %) es la observación mas idónea para sustituir a

la turba, ya que es una de las combinaciones que mejores características proporciona al cultivo y con un porcentaje de aplicación bajo, por lo que a la hora de comercializarlo será más fácil y eficiente.

Anexo 1

Variable	Explicación	Parámetro asociado a:
Length(cm)	Longitud de raíz (suma) por planta	Raíz
ProjArea(cm2)	Superficie radicular (ej sombra, superficie que ocupa la raíz en 2D) por planta	Raíz
AvgDiam(mm)	Diámetro promedio de la suma de raíces por planta	Raíz
Tips	Ápices (puntas de raíz) por planta	Raíz
Forks	Bifurcaciones (tenedores, cuando una raíz de divide en varias.., cuenta los puntos de bifurcación) por planta, indica el grado de ramificación	Raíz
L0a05	% de racies con longitud de 0 a 0.5	Raíz
L05a1	% de raíces con longitud de 0.5 a 1	Raíz
PA0a05	% del área proyectada de raíz para raíces con diámetro entre 0 a 0.5 por planta	Raíz
PA05a1	% del área proyectada de raíz para raíces con diámetro entre 0.5 a 1 por planta	Raíz
T0a05	% ápices (tips) en raíces con diámetro entre 0 y 0.5 mm	Raíz
T05a1	% ápices (tips) en raíces con diámetro entre 0 y 0.5 mm	Raíz
germ8	Germinación	Planta
h1	Altura planta muestreo 1	Planta
h2	Altura planta muestreo 2	Planta
h3	Altura planta muestreo 3	Planta
h4	Altura planta muestreo 4	Planta
pHsus	pH del substrato	Medio cultivo (sustrato)
Cesus	Conductividad eléctrica (salinidad)	Medio cultivo (sustrato)
mosus	Materia organica	Medio cultivo (sustrato)
dapsus	Densidad aparente	Medio cultivo (sustrato)

eptsus	Espacio poroso total	Medio cultivo (sustrato)
cairsus	Capacidad aireación	Medio cultivo (sustrato)
contrasus	Contracción	Medio cultivo (sustrato)
craguamlsus	Capacidad retención agua útil	Medio cultivo (sustrato)
cnsus	Relación Carbono/Nitrógeno	Medio cultivo (sustrato)
csus	Concentracion total elemento C carbono	Medio cultivo (sustrato)
nsus	Concentracion total elemento C nitrogeno	Medio cultivo (sustrato)
nahidrosus	Concentracion hidrosoluble elemento Na sodio	Medio cultivo (sustrato)
Khidrosus	Concentracion hidrosoluble elemento K potasio	Medio cultivo (sustrato)
Phidrosus	Concentracion hidrosoluble elemento P fosforo	Medio cultivo (sustrato)

Referencias

Julián Santos Peñas, Ángel Munñoz Alamillos, Pedro Juez Martel y Pedro Cortiñas Vázquez (2003). *Diseño de encuentas para estudios de mercado. Técnicas de muestreo y análisis multivariante*. Editorial centro de estudios Ramón Arecer, S.A. (Madrid).

Joseph F. Hair, Rolph E. Anderson, Ronald L. Tatham, William C. Black (1999). *Análisis multivariante*. Pearson educación, S.A. (Madrid).

Universidad de Granada (2012). *Introducción al Análisis Cluster. Consideraciones generales*, http://www.ugr.es/ gallardo/pdf/cluster-1.pdf

Universidad de Granada (2012). *Medidas de Asociación*, http://www.ugr.es/ gallardo/pdf/cluster-2.pdf

Universidad de Granada (2012). *Métodos Jerárquicos de Análisis Cluster.*, http://www.ugr.es/ gallardo/pdf/cluster-3.pdf

Universidad de Granada (2012). *Métodos no Jerárquicos de Análisis Cluster.*, http://www.ugr.es/ gallardo/pdf/cluster-4.pdf

Universidad de Valencia (2010). *INTRODUCCIÓN AL ANÁLISIS CLUSTER*, https://www.uv.es/ceaces/multivari/cluster/CLUSTER2.htm

Aurea Grané. Universidad Carlos III de Madrid (2017). *Distancias estadísticas y Escalado Multidimensional*, http://halweb.uc3m.es/esp/Personal/

personas/agrane/ficheros_docencia/MULTIVARIANT/slides_Coorp_reducido.pdf

Salvador Figueras, M (2001) *Análisis de conglomerados o cluster*, http://www.5campus.org/leccion/cluster

Universidad de Granada (2013) *RESUMEN ANÁLISIS CLUSTER*, http://www.ugr.es/ mvargas/2.RESUMENANLISISCLUSTER.pdf

Universidad de Granada (2013) *MÉTODOS DE ANÁLISIS MULTIVARIANTE: ANÁLISIS CLÚSTER*, http://wpd.ugr.es/ bioestad/guia-spss/practica-8/#18

Yanina Gimenez. Universidad de Buenos Aires (2010) *Clasificación no supervisada: El método de k–medias*, http://cms.dm.uba.ar/academico/carreras/licenciatura/tesis/2010/Gimenez_Yanina.pdf