Extensión del algoritmo *ClustImpute* para variables cualitativas y mixtas: una aplicación al capítulo de cultivos de Bogotá D.C del III Censo Nacional Agropecuario.

William Camilo Rojas Pulido

Director: Mario José Pacheco López

Trabajo de grado presentado como requisito para optar por el título de Estadístico

Noviembre de 2023

1. Introducción

En el actual contexto, marcado por cambios demográficos, económicos y ambientales en las últimas décadas, así como la posición del país frente a las políticas de desarrollo rural y mejora de la calidad de vida en áreas rurales, se hace necesario obtener información censal abarcando aspectos sociales, económicos y ambientales. La realización de un censo agropecuario en Colombia se destaca como una herramienta fundamental para comprender las múltiples variables que impactan en este sector de la economía. En un país con una economía donde el sector agropecuario desempeña un papel relevante, la recopilación de datos precisos sobre la actividad rural es esencial para múltiples aspectos. Este censo ofrece una mirada detallada al sector agropecuario colombiano, proporcionando datos que permitan identificar los aspectos en lo que se deban enfocarse las políticas y los esfuerzos de las diferentes entidades que velan por este sector. Un censo agropecuario proporciona una radiografía detallada de la realidad del campo colombiano. El censo ofrece información acerca de la extensión de tierras cultivadas, las semillas utilizadas, la finalidad de los cultivos, y otros datos que inciden en la toma de decisiones tanto a nivel nacional como regional.

Con esta premisa se llevo a cabo el III Censo Nacional Agropecuario durante el año 2014, este ejercicio proporciono información estadística, georreferenciada o de ubicación satelital y actualizada del sector agropecuario del país. Tuvo una cobertura operativa del 98.9%, cubriendo

los 1.101 municipios del país, el archipiélago de San Andrés, Providencia y Santa Catalina, 32 departamentos, 20 áreas no municipalizadas, 773 resguardos indígenas, 181 tierras de comunidades negras y 56 parques nacionales naturales. En este trabajo, el enfoque estará centrado en el aspecto de los cultivos en la ciudad de Bogotá. El análisis y la comprensión de los resultados permitieron comprender aspectos fundamentales en este sector, involucrando una serie de elementos que contribuyeron a una caracterización completa y adecuada.

Los datos faltantes, como los que se encuentran en esta encuesta (4% de registros con datos faltantes en al menos una variable), son un desafío inherente en el análisis de datos. Para enfrentar el problema de datos faltantes, se han desarrollado enfoques estadísticos diversos y eficaces, como la imputación de valores faltantes, que permiten tratar esta problemática de manera adecuada. A través de métodos como la imputación por la media, imputación basada en vecinos más cercanos, imputación por regresión y la imputación múltiple, se busca abordar los vacíos en los datos y así mantener la integridad de los análisis resultantes (Arteaga, 2009). Además, técnicas avanzadas como el algoritmo MICE (Multiple Imputation by Chained Equations), introducido por van Buuren (2011), han demostrado su eficacia al generar múltiples imputaciones basadas en modelos condicionales específicos para cada variable con datos faltantes. Otros métodos encontrados de la literatura son los métodos de aumento de datos, de donantes, bayesianos, entre otros.

Dentro del contexto de análisis multivariado y en general de análisis de datos, el agrupamiento emerge como una herramienta esencial. Esta técnica tiene como propósito fundamental organizar conjuntos de datos en grupos o clústeres con similitudes. El proceso se basa en la medición de similitudes o distancias entre observaciones, y su objetivo primordial es dividir los datos en subconjuntos coherentes y homogéneos que permitan una comprensión más clara de las características presentes en los datos (Jain et al., 1999). Esta metodología abarca diferentes enfoques, como el agrupamiento jerárquico y no jerárquico, el agrupamiento basado en densidades, entre otros. Dentro de los algoritmos de agrupamiento jerárquico encontramos el algoritmo de k-medias, el cual es un método no supervisado de agrupamiento y que servirá de base para el presente trabajo. Para el caso de datos faltantes existen algunas alternativas de tratamiento dentro de los métodos de agrupamiento, en particular encontramos el algoritmo *ClustImpute* introducido por Pfaffel (2020), el cual combina una técnica de imputación con el algoritmo de k-medias para variables numéricas.

Este trabajo propone utilizar el III Censo Nacional Agropecuario enfocado en el capítulo de Cultivos en Bogotá, para aplicar una extensión del algoritmo *ClustImpute* el cual aprovechará las variables tanto cuantitativas como cualitativas presentes en el censo, y así dar una alternativa a los métodos convencionales de imputación en el análisis de datos de características similares, para un análisis posterior adecuado y fiable.

2. Tratamiento de datos faltantes

Según Little et al. (2012), los datos faltantes corresponden a valores no disponibles que tendrían relevancia en el análisis si estuvieran presentes. Es esencial comprender la razón detrás de su ausencia y las implicaciones resultantes, para lo cual se necesita entender el mecanismo de los datos faltantes para explicar cómo se relacionan con los datos observados.

El mecanismo de datos faltantes se puede clasificar según Arteaga (2009) en tres tipos: datos faltantes completamente aleatorios (MCAR), en los que la usencia no depende de los valores de los datos, observados y faltantes; datos faltantes aleatorios (MAR), donde la ausencia depende solo de los datos observados; y datos faltantes no aleatorios (NMAR), donde la probabilidad que falte un elemento depende del valor no observado.

En el proceso de tratamiento de estos datos, diversos enfoques son utilizados. Estos enfoques incluyen la exclusión de casos con datos faltantes (casos completos), uso de subconjuntos de datos observados que permiten estimaciones en variables con datos faltantes (casos disponibles), métodos de imputación, en los que encontramos métodos como imputación por media incondicional, media condicional, regresión estocástica, hot-deck, múltiples imputaciones, entre otros, para reemplazar valores faltantes, con valores plausibles.

2.1 Algoritmo MICE

Un enfoque común para tratar valores faltantes es el método de imputación multivariada por ecuaciones encadenadas (MICE). También conocido como "imputación múltiple secuencial de regresión", MICE genera múltiples predicciones para cada valor perdido. Sin embargo, su efectividad depende de la calidad de los valores observados relacionados con el valor faltante. Azur (2011). Dado que se asume comúnmente un mecanismo MAR

Algoritmo:

1. Crear valores provisionales usando por ejemplo, la media de los datos observados para cada variable.
2. Estimar un modelo por ejemplo de regresión para cada variable con datos faltantes, utilizando los datos observados. Los valores generados en el paso anterior se reemplazan con los valores estimados.
3. Replicar este proceso para todas las variables con valores faltantes, iterando a lo largo de ciclos.
4. Repetir los pasos 2-3 en ciclos múltiples para actualizar las imputaciones.

2.2 Reglas de Rubin

Dado el parámetro de interés θ y su correspondiente estimador $\hat{\theta}$ junto con varianza estimada $\hat{V}(\hat{\theta})$, se generan m imputaciones que resultan en estimaciones $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \dots, \hat{\theta}^{(m)}$ y varianzas estimadas $\hat{V}(\hat{\theta})^{(1)}, \hat{V}(\hat{\theta})^{(2)}, \dots, \hat{V}\left(\hat{\theta}\right)^{(m)}$, se logrará obtener una estimación

conjunta de θ y su varianza de acuerdo con las reglas propuestas por Rubin, ver van Buuren (2018), de la siguiente manera:

$$\bar{\theta} = \frac{1}{m}\sum_{j=1}^{m}\hat{\theta}^{(j)}$$

$$\hat{V}(\bar{\theta}) = \frac{1}{m}\sum_{j=1}^{m}\hat{V}\left(\bar{\theta}^{(j)}\right) + \left(1+\frac{1}{m}\right)B$$

Donde:

$$B = 1 + \frac{1}{m}\sum_{j=1}^{m}(\hat{\theta}^{(j)} - \bar{\theta})(\hat{\theta}^{(j)} - \bar{\theta})^{T}$$

Como medidas de severidad producto del proceso de imputación, se puede calcular:

- Proporción de la varianza debido a datos faltantes:

$$\lambda = \frac{\left(1+\frac{1}{m}\right)B}{\hat{V}(\bar{\theta})}$$

- Incremento relativo de la varianza:

$$r = \frac{\left(1+\frac{1}{m}\right)B}{\bar{V}(\theta)}$$

- Fracción de información faltante sobre θ:

$$\gamma = (1+r)^{-1}\left[r + \frac{2}{v+3}\right]$$

con $v = \frac{n(n-1)}{(n+2)}(1-\lambda)$ grados de libertad.

3. Agrupamiento

El agrupamiento, es una técnica en estadística que implica la clasificación de un conjunto de datos en subgrupos homogéneos, donde las observaciones dentro de cada subgrupo son similares entre sí en comparación con las observaciones en otros subgrupos. Este enfoque permite descubrir patrones y estructuras inherentes en los datos sin la necesidad

de etiquetas previas. (Hastie, Tibshirani, & Friedman, 2009).

3.1 Algoritmo k-medias

El algoritmo k-medias es una técnica de agrupamiento utilizada para datos numéricos continuos. Desarrollado por Stuart P. Lloyd en 1982. Su objetivo es dividir un conjunto de datos en k grupos o clústeres, de modo que los datos dentro de cada clúster sean similares entre sí y diferentes de los de otros clústeres. El proceso se basa en minimizar la distancia entre los datos y el centroide del clúster al que están asignados (Lloyd, 1982).

Pasos del Algoritmo k-medias:

1. Inicialización: selecciona K centroides iniciales. Estos centroides pueden ser puntos aleatorios del conjunto de datos o seleccionados de alguna otra manera.
2. Asignación a Clústeres: asigna cada dato al clúster cuyo centroide sea el más cercano en términos de distancia euclidiana u otra métrica de distancia. La distancia euclidiana entre un dato X y un centroide C se calcula como:

$$d(X, C) = \sqrt{\sum_{i=1}^{n} (x_i - c_i)^2}$$

Donde x_i es la coordenada i-ésima del dato X, y c_i es la coordenada i -ésima del centroide C (Lloyd, 1982).

3. Actualización de Centroides: calcula el nuevo centroide de cada clúster tomando la media de todas las coordenadas de los datos asignados a ese clúster. Para el j -ésimo atributo, la coordenada del nuevo centroide C_j se calcula como:

$$c_j = \frac{1}{|S|} \sum_{x \in S} x_j$$

Donde S es el conjunto de datos asignados al clúster y x_j es el valor del atributo j en el dato x.

4. Iteración: repite los pasos 2 y 3 hasta que los centroides no cambien significativamente o se cumpla un criterio de detención, como un número máximo de iteraciones.

El algoritmo k-medias busca converger hacia una solución en la que los datos estén agrupados de manera coherente y los centroides representen el "centro" de cada clúster. Sin embargo, es importante destacar que k-medias es sensible a la inicialización de los centroides y puede converger a mínimos locales. Por lo tanto, es común ejecutar el algoritmo varias veces con diferentes inicializaciones y elegir la mejor solución.

3.2 Algoritmo k-modas

El algoritmo k-modas es una técnica de agrupamiento utilizada para datos categóricos, donde las variables son discretas o etiquetas en lugar de valores numéricos. A diferencia de k-medias, que se aplica a datos numéricos continuos, k-modas se basa en la noción de similitud entre objetos categóricos para formar agrupaciones con características similares (Huang, 1997; Huang, 1998).

Este emplea la distancia de *Hamming* entre dos objetos categóricos x e y la cual se calcula como la suma de las diferencias entre los valores de los atributos x_i e y_i:

$$d_1(X,Y) = \sum_{j=1}^{m} \delta(x_j, y_j)$$

Donde $d_1(X,Y)$ es la distancia *Hamming* entre los objetos categóricos X e Y, x_j e y_j son los valores de los atributos j de los objetos X e Y, respectivamente, y $\delta(x_j, y_j)$ es una función delta que vale 1 si x_j e y_j son diferentes, y 0 si son iguales (Huang, 1998).

Para calcular los modos y centroides en k-modas, se selecciona el valor más frecuente de cada atributo para un grupo específico. Dado un grupo con N objetos categóricos, el modo de cada atributo se determina como el valor más frecuente dentro del grupo, el cual es el valor de x_i que minimiza:

$$\sum_{i=1}^{n} d_1(X_i, Q)$$

X_i es el valor del atributo i en el grupo, y $d_1(X_i, Q)$ es una función delta que vale 1 si X_i es igual al valor Q, y 0 si son diferentes (Huang, 1998).

La actualización de los centroides en k-modas se realiza iterativamente, tomando el objeto más cercano al centroide en términos de distancia *Hamming*. En cada iteración, se calcula el nuevo centroide de un grupo seleccionando el objeto más cercano al centroide en términos de distancia *Hamming*.

$$C_k^t = \arg min_{x \in S_k^t} \left(\sum_{y \in S_k^t} d_1(X,Y) \right)$$

Donde C_k^t es el nuevo centroide (modo) del grupo k en la iteración t, S_k^t es el conjunto de objetos asignados al grupo k en la iteración t, y $d_1(X,Y)$ es la distancia Hamming entre los objetos categóricos X e Y (Cao, Liang, & Bai, 2009).

Pasos del Algoritmo k-modas:

1. Inicialización: selecciona K objetos aleatoriamente como centroides iniciales.
2. Asignación a Grupos: asigna cada objeto del conjunto X al grupo cuyo centroide sea el más cercano en términos de distancia Hamming.
3. Actualización de Centroides (Modos): calcula el nuevo centroide (modo) de cada grupo seleccionando el valor más frecuente de cada atributo dentro del grupo.
4. Iteración: repetir los pasos 2 y 3 hasta que se alcance la convergencia o se cumpla un criterio de detención.

3.3 Algoritmo k-prototipos

El algoritmo k-prototipos es una técnica de agrupamiento utilizada para datos mixtos, que combinan variables categóricas y numéricas en un mismo conjunto de datos. A diferencia de k-medias y k-modas, que se aplican a datos completamente numéricos o categóricos, k-prototipos aborda la agrupación de datos con atributos de diferentes tipos (Huang, 1997; Huang, 1998; Z. Cao, Liang, & Bai, 2009).

La distancia empleada en k-prototipos combina la distancia Euclidiana para atributos numéricos y la distancia *Hamming* (o una medida similar) para atributos categóricos. La distancia entre dos objetos X e Y se calcula como la raíz cuadrada de la suma ponderada de las distancias cuadráticas de los atributos numéricos y las distancias *Hamming* de los atributos categóricos:

$$d_{kp}(X,Y) = \sqrt{\sum_{j=1}^{m} w_j \cdot d_j^2(X,Y) + \sum_{j=m+1}^{m} \delta(x_j, y_j)}$$

Donde:

- $d_{kp}(X,Y)$ es la distancia K-prototipos entre los objetos X e Y.

- w_j es un peso asignado al atributo j (mayor para atributos numéricos, menor para categóricos).

- $d_j(X,Y)$ es la distancia Euclidiana entre los valores numéricos x_j e y_j.

- $\delta(x_j, y_j)$ es una función delta que vale 1 si los atributos categóricos x_j y y_j son diferentes, y 0 si son iguales.

La actualización de los centroides se realiza de manera similar a k-medias para atributos numéricos y a k-modas para atributos categóricos:

- Para atributos numéricos, se calcula la media de los valores numéricos dentro del grupo.

- Para atributos categóricos, se selecciona el valor más frecuente dentro del grupo.

Pasos del Algoritmo k-prototipos:

1. Inicialización: selecciona k objetos aleatoriamente como centroides iniciales (prototipos).
2. Asignación de Objetos a Grupos: asigna cada objeto del conjunto X al grupo cuyo prototipo (centroide) sea el más cercano en términos de distancia K-prototipos.
3. Actualización de Centroides: calcula los nuevos prototipos para atributos numéricos y categóricos dentro de cada grupo.
4. Iteración: Repetir los pasos 2 y 3 hasta que se alcance la convergencia o se cumpla un criterio de detención.

3.4 Algoritmo *ClustImpute*

El algoritmo *clustImpute* es una técnica de imputación de datos faltantes que combina el proceso de agrupamiento con la imputación de valores faltantes. A través de la agrupación, *clustImpute* busca identificar patrones similares entre los datos y luego utiliza estos patrones para imputar los valores faltantes de manera coherente (Li, 2019; Zhao et al., 2015).

En *clustImpute*, el proceso de imputación se realiza mediante el análisis de grupos de datos similares. Los pasos clave del algoritmo son los siguientes:

1. Agrupamiento: utiliza el algoritmo de agrupamiento de k-medias para agrupar los datos en grupos basados en la similitud entre los atributos. Cada grupo representa un patrón de datos similares.
2. Identificación de Patrones: una vez que los grupos se han formado, se analizan los patrones de datos dentro de cada grupo. Se busca entender cómo se distribuyen los valores de los atributos y cuál es la relación entre ellos.
3. Imputación de Valores Faltantes: cuando se encuentra un valor faltante en un atributo de un objeto en el conjunto de datos, *clustImpute* busca objetos similares dentro del mismo grupo y utiliza sus valores para imputar el valor faltante. Esta imputación se realiza de manera coherente con el patrón de datos del grupo.

4. Métodos de Validación

Los métodos de validación de clusters son herramientas esenciales en el campo de la minería de datos y el análisis de clustering. Estos métodos se utilizan para evaluar la calidad de los clusters generados por algoritmos de agrupamiento (Milligan & Cooper, 1985). El objetivo principal de la validación es determinar si los grupos identificados son apropiados y si el algoritmo de clustering utilizado es el adecuado para los datos en cuestión.

Existen dos categorías principales de métodos de validación de agrupamiento: medidas internas y medidas externas.

- **Medidas de validación internas:** estas medidas evalúan la calidad de los grupos sin hacer referencia a ninguna información externa. Evalúan la cohesión dentro de los grupos y la separación entre los mismos. Algunas medidas internas comunes incluyen el Índice de Silueta, el Índice de McClain-Rao y el Índice Davies-Bouldin.
- **Medidas de validación externas:** estas medidas comparan los grupos generados por un algoritmo con un conjunto de datos de referencia o la verdad fundamental. Evalúan qué tan bien se ajustan los grupos a la estructura de datos conocida (Milligan & Cooper, 1985). Ejemplos de medidas externas incluyen el Índice Rand, la Puntuación de Información Mutua (AMI) y el Índice Fowlkes-Mallows.

4.1 Indice Silueta

El Índice de Silueta es una medida interna de validación de clusters que evalúa la cohesión y la separación de los clusters (Rousseeuw, 1987). Se calcula para cada punto de datos en función de la distancia entre ese punto y los puntos de su mismo conglomerado, así como la distancia promedio a los puntos en el grupo más cercano que no contiene al punto. El Índice de Silueta proporciona valores entre -1 y 1.

$$Silueta = \frac{1}{n}\sum_{i=1}^{n}\frac{b(i)-a(i)}{\max\left(a(i),b(i)\right)}$$

a(i) es la disimilitud media del objeto *i*-esimo con respecto a todos los demás objetos de mismo grupo. *b(i)=min(d(i,C)),* donde *d(i,C)* es la disimilitud media del objeto *i*-esimo con respecto a todos los demás conglomerados, excepto el mismo grupo. El valor máximo del índice se utiliza para indicar el numero óptimo de grupos.

4.2 Indice McClain-Rao

El Índice de McClain-Rao es otra medida interna de validación de clusters que se basa en la densidad y la separación de los clusters (McClain & Rao, 1975). Se calcula como la relación entre la dispersión intra-cluster y la dispersión inter-cluster. Un índice de McClain-Rao más alto indica que los grupos son más densos y mejor separados.

$$McClain = \frac{\bar{S}_w}{\bar{S}_b}$$

$\bar{S}_w$ e la suma de las distancias dentro del grupo dividida por el número de distancias dentro del grupo y $\bar{S}_b$es la suma de las distancias de los grupos dividida por el número de distancias entre grupos. El valor mínimo del índice se utiliza para indicar el número óptimo de conglomerados.

4.3 Distancia Media entre Medias (ADM)

La medida de la Distancia Promedio entre Medias (ADM, por sus siglas en inglés) calcula la distancia promedio entre los centros de los grupos para las observaciones ubicadas en el mismo grupo mediante el agrupamiento basado en el conjunto de datos completo y el agrupamiento basado en los datos con una sola columna eliminada. Se define de la siguiente manera:

$$ADM = \frac{1}{MN}\sum_{i=1}^{N}\sum_{i=1}^{M} dist(\bar{x}_{C^{i,l}}, \bar{x}_{C^{i,0}})$$

Donde $\bar{x}_{C^{i,0}}$ es la media de las observaciones en el grupo que contiene la observación i cuando el agrupamiento se basa en el conjunto de datos completo, y $\bar{x}_{C^{i,l}}$ se define de manera similar. También tiene un valor entre cero e infinito, y nuevamente se prefieren valores más pequeños.

5. Datos

Cuadro 1: Descripción de Variables

Variable	Descripción
TIPO_REG	Región
PAIS	País
P_DEPTO	Departamento
P_MUNIC	Municipio
UC_UO	Unidad de observación
ENCUESTA	Numero de encuesta
COD_VEREDA	Código de vereda
P_S6P45B	Numero de lote

Cuadro 1: Descripción de Variables

P_S6P45A	Cultivo registrado estaba presente o era pasado
P_S6P46	Cultivo o plantación forestal que tiene el lote
P_S6P47A	Mes de siembra
P_S6P47B	Año de siembra
P_S6P48	Indica si el cultivo se realiza de forma individual (solo) o en asociación (asociado).
P_S6P50	Tipo de semilla utilizada
P_S6P51_SP1	Cultivo bajo cubierta
P_S6P51_SP2	Cultivo a cielo abierto
P_S6P51_SP3	Cultivo en hidroponía
P_S6P53	Finalidad de la plantación
P_S6P57A	Cantidad obtenida
P_S6P59_UNIF	Rendimiento tonelada/hectáreas
P_S6P60	Fenómeno natural que afecto el cultivo
AREA_SEMBRADA	Área sembrada
AREA_COSECHADA	Área cosechada

El III Censo Nacional Agropecuario del año 2014, proporciono información estadística, georreferenciada o de ubicación satelital y actualizada del sector agropecuario del país. En cuanto al aspecto de los cultivos, se incluyen datos que abarcan la ubicación de estos, su presencia, la época de siembra, el tipo de semilla utilizada y la finalidad de cada cultivo, entre otros aspectos relevantes, los cuales pueden observar en el Cuadro 1.

Para acceder a los datos lo puede hacer desde el siguiente enlace https://microdatos.dane.gov.co/index.php/catalog/513/get-microdata

Un conjunto de variables de ubicación geográfica será excluido del análisis, ya que su inclusión no aportaría información en un análisis de grupos, dado que nos concentramos únicamente en una región del país. Además, algunas variables serán consolidadas en una única variable, ya que forman parte de una misma pregunta pero fueron formuladas por separado, generando así numerosos valores faltantes en estas variables.

Finalmente, se emplearán los datos del año 2013, que son los registros más recientes en el censo. Los registros de años anteriores involucraban las mismas unidades de observación, y la inclusión de sujetos repetidos carecía de aportes significativos tanto para la imputación como para el análisis subsiguiente.

6. Metodología

Se propone un método de agrupación tipo k-prototipo para variables mixtas basada en el algoritmo ClustImpute e implementado en el software R versión 4.2.3.

Para la implementación de la función, se realizó una revisión de los códigos fuente de las funciones kmodes de la librería klaR, kproto de la librería clustMixType y ClustImpute de la librería ClustImpute. La función ClustImpute fue la base sobre la cual se realizaró modificaciones en varios pasos clave.

Uno de los primeros pasos que se modificó es la asignación de pesos a las observaciones de las variables. Este peso influye en cómo se tratan los valores faltantes en el algoritmo de k-medias utilizado por ClustImpute. Conforme avanza el proceso de agrupamiento, el peso de las observaciones faltantes disminuye gradualmente hacia cero. Sin embargo, dado que se trabajó con variables categóricas, este paso se omitió.

Los pasos por modificar fueron los relacionados con el proceso de agrupamiento en sí. La función original utiliza el método k-medias, pero se exploró alternativas adecuadas para el tipo de datos y análisis requeridos. También se ajustó el paso que asigna grupos a los datos de entrada, ya que se buscó adaptar el proceso de asignación de clústeres a las modificaciones realizadas en el proceso de agrupamiento.

La función *ClustImputeMix* cuenta con los siguientes argumentos:

- X: base de datos con las variables a imputar.
- nr_cluster: número de grupos que se utilizarán en el proceso de agrupación.
- nr_iter: número de iteraciones que se llevarán a cabo en el proceso de agrupación. Por defecto, son 10 iteraciones.
- seed_nr: número entero que se utiliza como semilla para que los resultados sean reproducibles. Puede ser NULL.

Para iniciar con el proceso se realiza la carga de los paquetes necesarios: dplyr y clusMixType. En primer lugar, la función toma como entrada el conjunto de datos *X*, si *X* no es un dataframe, lo convierte en uno para asegurar la consistencia en el procesamiento de datos. Luego, se crea un indicador de datos faltantes llamado *mis_ind* para rastrear las ubicaciones de los valores faltantes en *X*. La función realiza una verificación inicial para determinar si todos los valores en *X* son faltantes. Si esta condición se cumple, se muestra un mensaje de error indicando que todos los valores son NA.

Luego de completar las etapas anteriores, comienza la imputación de los valores faltantes mediante la sustitución de estos por valores aleatorios obtenidos de las observaciones que ya están disponibles en el conjunto de datos. Este proceso se lleva a cabo individualmente para cada variable en el conjunto de datos *X*. En primer lugar, se identifican las variables que no tienen datos faltantes, y luego se elige al azar una de estas observaciones sin faltantes para reemplazar los valores ausentes en esa columna en particular.

Luego de completar los pasos previos, se llevan a cabo dos procedimientos en la preparación de los datos. En primer lugar, se procede a la conversión de las variables que son de naturaleza categórica en un formato denominado *factor*. Esto es necesario para que el modelo pueda trabajar de manera adecuada con estas variables. En segundo lugar, se crea una variable adicional llamada *org_is_false*, que sirve como indicador de los datos faltantes originales. En esta variable, se marca como *TRUE* aquellos lugares donde se encontraban valores faltantes en *X* y *NA* los datos que originalmente no eran faltantes.

Después se lleva a cabo la agrupación de datos utilizando la función *kproto* en múltiples iteraciones. El proceso comienza con la inicialización de los clústeres en la primera iteración y continúa hasta el valor dado en el argumento *nr_iter*. El número de clústeres se define previamente como *nr_cluster*. Durante cada iteración, se verifica si hay filas duplicadas en los centroides calculados para los clústeres. Si se encuentran duplicados, se emite un mensaje de advertencia.

Luego, se asignan los clústeres a las observaciones en función de la distancia a los centroides con la ayuda de la función *predict*. Posteriormente, se realiza la imputación de los valores faltantes basándose en los clústeres actuales. Para cada clúster y variable, se seleccionan aleatoriamente muestras de observaciones sin datos faltantes en ese clúster y se utilizan para reemplazar los valores faltantes en la misma variable. Este proceso se repite a lo largo de las iteraciones definidas por *nr_iter*.

Finalmente, se obtienen los centroides finales después de todas las iteraciones y se devuelven junto con otros resultados relevantes, como los clústeres asignados a las observaciones y la matriz de centroides. Estos resultados se estructuran en un objeto y se devuelven como el resultado de la función *ClustImputeMix*.

6.1 Algoritmo ClustImputeMix

ClustImputeMix es un algoritmo que combina la agrupación y la imputación de datos faltantes. Su enfoque se centra en encontrar patrones similares en los datos a través de la agrupación y luego utiliza esos patrones para rellenar de manera coherente los valores faltantes. Esta técnica permite analizar conjuntos de datos que contienen una combinación de variables numéricas y categóricas, al mismo tiempo que maneja la presencia de valores faltantes. Esto facilita la realización de un análisis completo de los datos en una etapa posterior.

Lo que hace que *ClustImputeMix* con k-prototipos sea único es que combina la imputación y la agrupación en un solo paso. En lugar de abordar estos dos problemas por separado, este algoritmo los aborda de manera conjunta, lo que lo convierte en una herramienta muy útil para analizar conjuntos de datos complejos.

En *clustImputeMix*, el proceso de imputación se realiza mediante el análisis de grupos de datos similares. Los pasos clave del algoritmo son los siguientes:

1. Imputación Aleatoria de Datos Faltantes: reemplaza los valores faltantes con nuevos valores tomados de observaciones existentes de manera aleatoria.
2. Transformación de Variables Categóricas: convierte las variables categóricas en factores para su posterior uso en el algoritmo.
3. Agrupación k-prototipos en Múltiples Iteraciones: realiza la agrupación k-prototipos en un bucle que se ejecuta un número de veces especificado (*nr_iter*).
4. Inicialización k-prototipos: selecciona k objetos aleatoriamente como centroides iniciales (prototipos).
5. Asignación de Objetos a Grupos: asigna cada objeto del conjunto X al grupo cuyo prototipo (centroide) sea el más cercano en términos de distancia K-prototipos.
6. Actualización de Centroides: calcula los nuevos prototipos para atributos numéricos y categóricos dentro de cada grupo.
7. Iteración: repetir los pasos 2 y 3 hasta que se alcance la convergencia o se cumpla un criterio de detención.
8. Asignación de Observaciones a Clústeres: asigna cada observación a uno de los clústeres basándose en el modelo k-prototipos actual.
9. Imputación de Datos Faltantes Basada en Clústeres: imputa los valores faltantes en función de los clústeres actuales. Para cada clúster y variable, toma muestras de observaciones sin faltantes dentro del mismo clúster y las utiliza para imputar los valores faltantes.
10. Iteración: repetir los pasos 3 a 10 hasta alcanzar convergencia o que se cumpla el criterio *nr_iter*.
11. En la primera iteración, inicializa el modelo k-prototipos con una semilla aleatoria. En las iteraciones siguientes, actualiza el modelo con una nueva semilla aleatoria y compara con el modelo anterior.

7. Ejemplo de Aplicación

Se realizo una aplicación al capítulo de cultivos del III Censo Nacional Agropecuario, se buscó imputar la información en las variables que presentan valores faltantes. Dado que estos valores faltantes impiden realizar un análisis de agrupamiento.

Para iniciar el análisis de la información, fue necesario descargar los datos recopilados de la página de microdatos del DANE y revisar el diccionario de datos correspondiente. Esto permitió realizar una primera revisión de las variables del censo y conocer qué tipo de información contenía, después de la revisión de los datos, se llevó a cabo un tratamiento de la base de datos. Se incluyeron las modificaciones de los nombres de las columnas, la conversión de variables cualitativas con valores numéricos a su representación adecuada a formato cualitativo y la exclusión de las variables que no serán empleadas en el análisis. Las variables excluidas, a saber, TIPO_REG, PAIS, P_DEPTO, P_MUNIC, UC_UO, ENCUESTA, COD_VEREDA y P_S6P45B, fueron excluidas del análisis de grupos debido el poco aporte que darían el análisis posterior.

Se llevaron a cabo otros ajustes en la base de datos original, el primero fue la creación de una variable a partir de las variables P_S6P51_SP1, P_S6P51_SP2, P_S6P51_SP3, debido a que contenían muchos faltantes, debido a que su valor dependía de una respuesta anterior. El segundo ajuste fue cambiar los valores de la variable P_S6P47A, el cual indicaba el mes de siembra, se cambió a trimestre de siembra según el mes que se indicara, para mejorar la agrupación posterior.

Luego de disponer de esta base de datos, se empleó la función propuesta para llevar a cabo la imputación de los valores faltantes. En la función, se estableció que se realizara la imputación con el número adecuado de bases y con la cantidad de iteraciones necesarias para garantizar una convergencia satisfactoria en el proceso de imputación.

Después de obtener los resultados de la función, se llevó a cabo un análisis multivariado de las bases imputadas con el propósito de aprovechar los datos que se obtuvieron y observar cómo estas variables ser relacionan entre ellas. Para este análisis se utilizó el paquete FactoMineR, y para la visualización de los resultados, se utilizará la función fviz_pca_var del paquete factoextra en R.

Además de las funciones y paquetes mencionados anteriormente, también se aprovechó el paquete ggplot2 para la visualización de datos. Además, se emplearon funciones como select y mutate con el propósito de transformar la base de datos y prepararla para su posterior análisis.

La siguiente tabla presenta la proporción de datos faltantes del III Censo Nacional Agropecuario enfocado en Cultivos en Bogotá.

7.1 Descripción de los faltantes

Cuadro 2: Proporción de Datos Faltantes de la Base Original

Variable	% Datos Faltantes
Región	0
País	0
Departamento	0
Municipio	0
Unidad de Observación	0
ENCUESTA	0
Vereda	0
No Lote	0
Presencia Cultivo	0
Cultivo o Forestal	0
Mes de Siembra	4,07
Año de Siembra	0
Cultivo esta	0
Tipo de Semilla	1,48
Bajo Cubierta	98,11
Cielo Abierto	3,59
En hidroponía	99,79
Finalidad de la Plantación	95,65
Cantidad Obtenida	0
Rendimiento Ton/Ha	45,21
FN Afecto	5,55
Área Sembrada	0
Área Cosechada	0
Cultivo se Encuentra	1,48

Patrón de datos faltantes

La Figura 1 ofrece una visualización de la cantidad de datos faltantes para cada variable. En este contexto, se han excluido las variables con porcentajes altos de datos faltantes y aquellas que introducen dominios adicionales. El grafico funciona de la siguiente manera, en el lado derecho, se presenta el recuento de variables con datos faltantes en cada instancia, mientras que en la parte inferior se muestra el número de observaciones que contienen datos faltantes en la variable resaltada en la parte superior del gráfico.

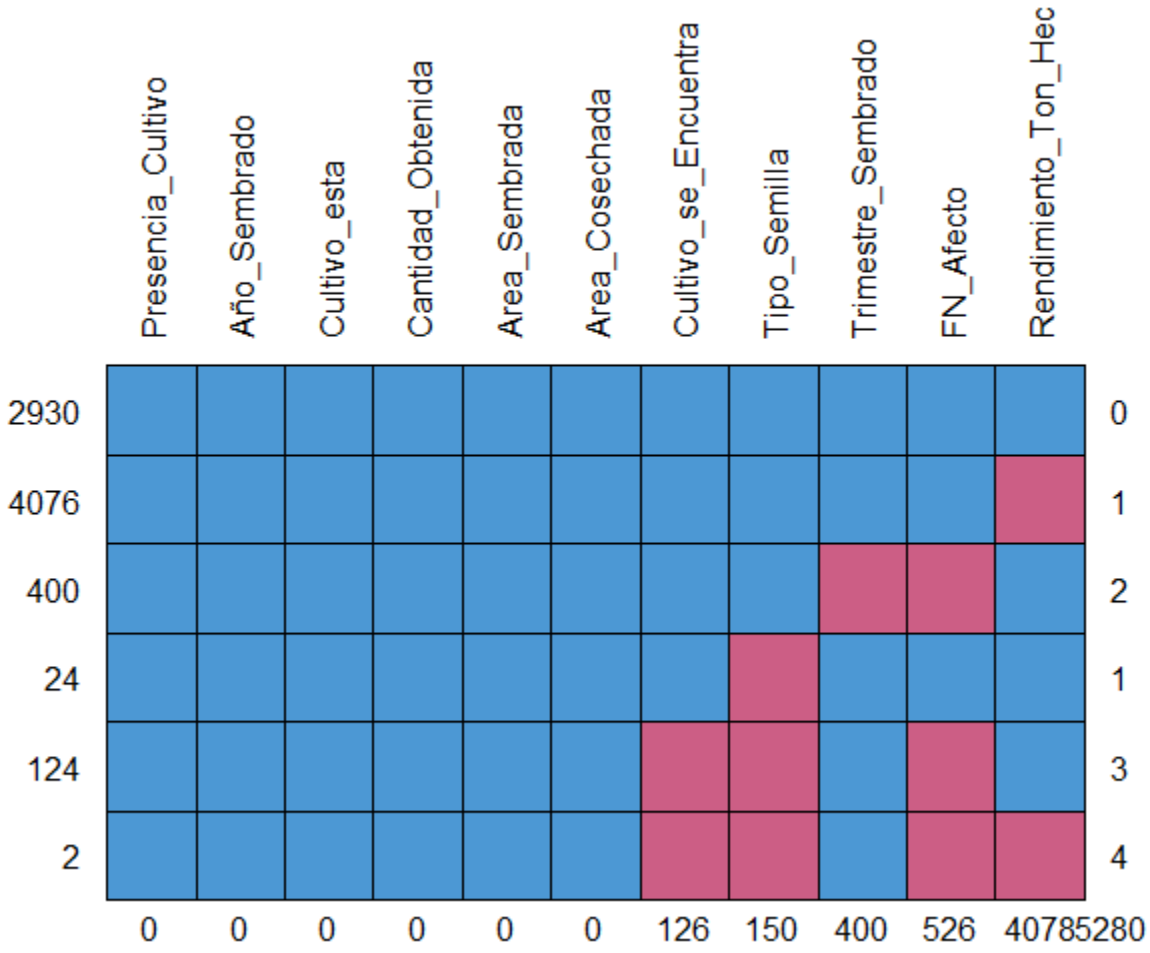

Figura 1: Patrón de datos faltantes

7.2 Medidas descriptivas después de imputación

Las siguientes tablas de frecuencias presentan un resumen de las variables categóricas sin datos faltantes, junto con una comparación entre las frecuencias de las variables con datos faltantes después de aplicar dos métodos de imputación.

Cuadro 3: Variables sin faltantes

Cultivo esta	n	Presencia de Cultivo	n
Asociado	5003	Pasado	6691
Solo	2553	Presente	865

Cuadro 4: Variables con faltantes

Variable	Categoría	ClustImputeMix	MICE
Cultivo se Encuentra	Bajo Cubierta	87	88
	Cielo Abierto	7450	7449
	Hidroponía	19	19
Tipo de Semilla	Certificada	201	203
	No Certificada	6954	6952
	Tradicional	401	401
Trimestre Sembrado	Trimestre 1	3800	3711
	Trimestre 2	1308	1595
	Trimestre 3	769	709
	Trimestre 4	1679	1544
Fenómeno Natural que Afecto	Enfermedad	6	4
	Lluvia	2	98
	Helada	17	24
	Lluvia a Destiempo	3	43
	No fue Afectado	7506	7192
	Plaga	16	18
	Sequia	6	77

En el cuadro 4 se pueden evidenciar cambios en las frecuencias de las categorías imputadas, lo cual afecta la agrupación que se realizó las cuales mostraron una diferencia considerable en los resultados.

Las siguientes tablas presentan un resumen con el promedio y la desviación estándar de cada variable. Específicamente para la variable 'Rendimiento Ton/Hec', se lleva a cabo una comparación entre los diferentes algoritmos utilizados para su análisis.

Cuadro 5: Promedio y desviación de las variables numéricas

Variable	Promedio	Desviación Estándar
Cantidad Obtenida	1,917	7,318
Area Sembrada	0,6285	1,066
Area Cosechada	0,2682	0,671

La cantidad obtenida muestra una considerable variabilidad con un promedio de 1,917 y una desviación estándar de 7,318, mientras que el área sembrada presenta menos dispersión en torno a su promedio de 0,6285 con una desviación estándar de 1,066.

Cuadro 6: Promedio y desviación de la variable imputada

	ClustImputeMix		MICE	
Variable	Promedio	Desviación Estándar	Promedio	Desviación Estándar
Rendimiento Ton/Hec	6,51	5,974	7,31	5,577

En la variable de rendimiento, se destaca que el algoritmo MICE arroja un promedio superior en comparación con el otro algoritmo utilizado. Sin embargo, la desviación estándar en ambos casos muestra una similitud significativa.

7.3 Medidas de validación de los grupos

Cuadro 7: Índices de validación de grupos para distintos grupos

		Grupos				
Algoritmo	**Índice**	**2**	**3**	**4**	**5**	**6**
ClustImputeMix	Silueta	0,091	0,255	0,5232	0,263	0,4795
	McClain	0,2231	0,6773	0,2323	0,2341	0,2226
MICE	Silueta	0,9421	0,2105	0,3065	0,4215	0,4231
	McClain	0,1311	0,5235	0,4538	0,4897	0,6157

En el Cuadro 7 se exhiben los valores de los índices de silueta y McClain para diferentes números de agrupaciones. Estos índices fueron empleados para identificar el número óptimo de grupos tras aplicar métodos de imputación en la base de datos. Basándose en estos resultados, se optó por emplear cuatro grupos para llevar a cabo una caracterización adecuada de la base de datos, dado que esta cantidad de grupos exhibe valores deseables en ambos índices, a diferencia de los otros agrupamientos.

Cuadro 8: Índices de validación de grupos

Índice	**ClustImputeMix**	**MICE**
Silueta	0,5232	0,3065
McClain	0,2323	0,4538

Estos índices resaltan que la agrupación aplicada a la base imputada con el algoritmo implementado logra una asignación más efectiva de los individuos a los grupos, mostrando una diferencia notable en comparación con el algoritmo MICE.

Basándose en los resultados obtenidos, el método implementado mostro un rendimiento superior tanto en tiempo de computación como en la validación de la agrupación. Se seleccionaron 4 grupos que permitieron una caracterización adecuada de la base de datos, con validaciones favorables en medidas como la silueta (0.5232) y McClain (0.2323). En contraste, las métricas de validación del método MICE mostraron una silueta de 0.3065 y un McClain de 0.4538, evidenciando una diferencia considerable. Esto resalta la efectividad del método ClustImputeMix en comparación con el método MICE.

7.4 Caracterización de los grupos de MICE

En el grupo 1 se constata una siembra pasada en el tercer trimestre, con un cultivo asociado y el uso de semillas no certificadas en áreas al aire libre. La cantidad obtenida fue baja, con un promedio de alrededor de 0.8 toneladas, y un rendimiento moderado de aproximadamente 9.7 toneladas por hectárea en promedio. Este cultivo no se vio afectado por factores naturales y se sembró en un área promedio de aproximadamente 0.52 hectáreas, cosechando alrededor de 0.04 hectáreas.Este grupo, caracterizado por cultivos pasados en el Trimestre 3, presenta cultivos asociados y utiliza semillas no certificadas en terrenos de cielo abierto. En promedio, muestra una cantidad obtenida relativamente baja de alrededor de 0.8 toneladas, un rendimiento moderado de aproximadamente 9.7 toneladas/hectárea y áreas sembradas y cosechadas relativamente pequeñas.

En el grupo 2 se identifica una siembra pasada en el cuarto trimestre, con un cultivo asociado utilizando semillas no certificadas en áreas al aire libre. La cantidad obtenida varió considerablemente, desde valores bajos hasta altos, con un promedio de aproximadamente 3.7 toneladas, y un rendimiento por hectárea que mostró una amplia variabilidad, con un promedio cercano a 9.7 toneladas por hectárea. Este cultivo no fue afectado por factores naturales y se sembró en áreas que presentaron diferencias, con un promedio de alrededor de 0.67 hectáreas, cosechando con variaciones en un promedio de aproximadamente 0.29 hectáreas.

Con cultivos previos en el Trimestre 4, este grupo también muestra cultivos asociados y utiliza semillas no certificadas en terrenos de cielo abierto. Exhibe una variabilidad considerable en la cantidad obtenida, con valores que van desde bajos hasta altos, promediando alrededor de 3.7 toneladas. El rendimiento también varía ampliamente, con un promedio cercano a 9.7 toneladas/hectárea. Las áreas sembradas y cosechadas presentan variaciones considerables.

En el grupo 3 se constata una siembra pasada durante el cuarto trimestre, con un cultivo solitario utilizando semillas no certificadas en áreas al aire libre. La cantidad obtenida fue baja, alrededor de 2.3 toneladas, con un rendimiento promedio de aproximadamente 9.6 toneladas por hectárea. Este cultivo no se vio afectado por factores naturales y se sembró en un área promedio de alrededor de 0.82 hectáreas, cosechando en un área promedio de aproximadamente 0.36 hectáreas.

Caracterizado por cultivos pasados en el Trimestre 4 y cultivos en solitario, este grupo utiliza semillas no certificadas en terrenos de cielo abierto. Muestra una cantidad obtenida relativamente baja de alrededor de 2.3 toneladas en promedio, con un rendimiento promedio de aproximadamente 9.6 toneladas/hectárea. Las áreas sembradas y cosechadas son más amplias en comparación con el Grupo 1.

En el grupo 4 se identifica una siembra pasada en el cuarto trimestre, con un cultivo solitario en áreas al aire libre utilizando semillas no certificadas. La cantidad obtenida fue alta, con un promedio de alrededor de 5.3 toneladas por hectárea, y un rendimiento también alto, aproximadamente de 11.7 toneladas por hectárea en promedio. Este cultivo no se vio afectado por factores naturales y se sembró en un área mayor, con un promedio de aproximadamente 0.77 hectáreas, cosechando en un área significativamente mayor, con un promedio cercano a 0.38 hectáreas.

Este grupo también presenta cultivos pasados en el Trimestre 4, pero se caracteriza por cultivos en solitario y un uso de semillas no certificadas en terrenos de cielo abierto. Destaca por una cantidad obtenida alta, con un promedio cercano a 5.3 toneladas, y un rendimiento elevado de alrededor de 11.7 toneladas/hectárea. Las áreas sembradas y cosechadas son más grandes en comparación con los Grupos 1 y 3.

Cuadro 9: Tamaño de Cada Grupo

1	2	3	4
1493	1478	3485	1100

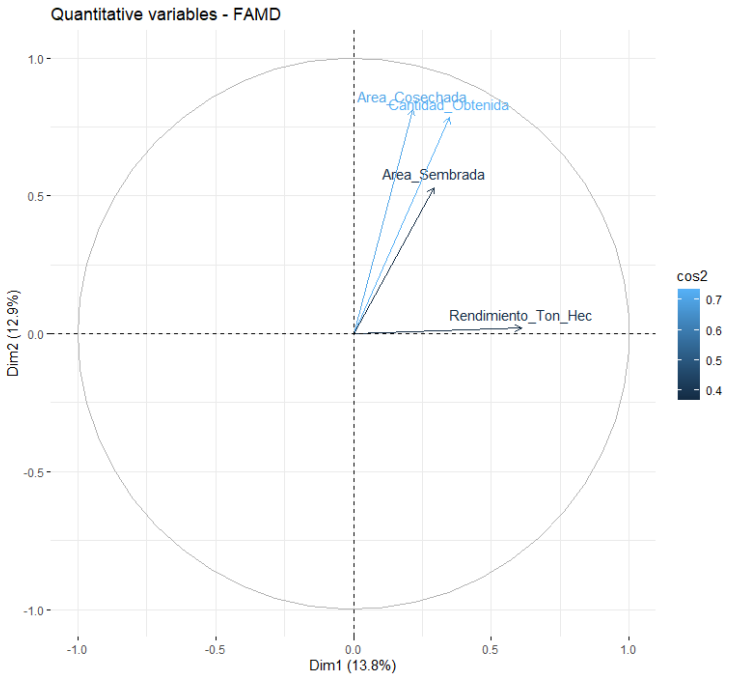

Figura 2: Grafico variables cuantitativas FAMD

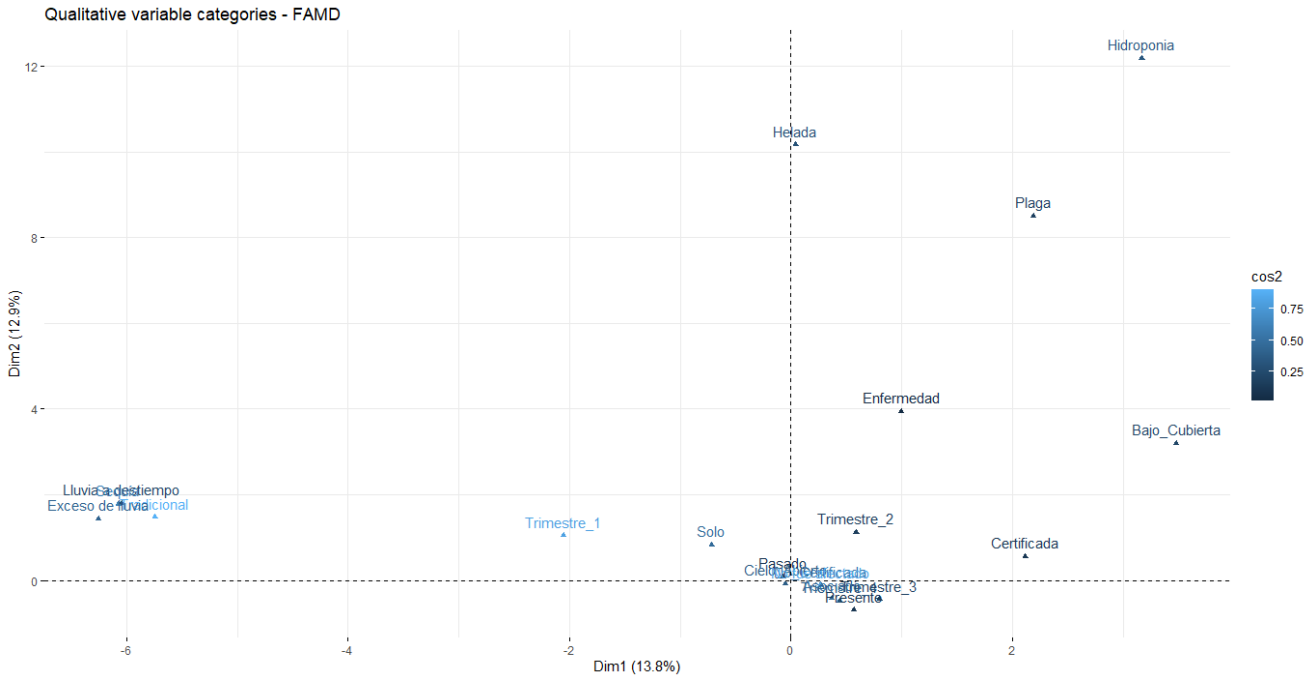

Figura 3: Grafico variables cualitativas FAMD

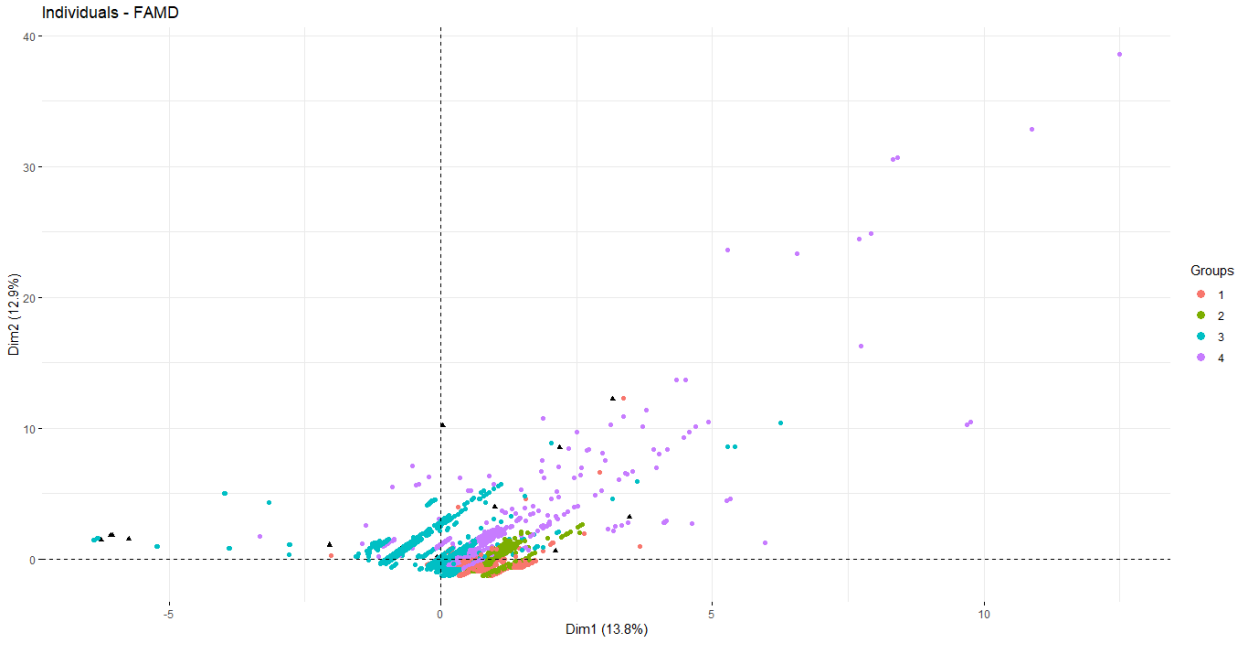

Figura 4: Grafico de individuos FAMD

Los Gráficos 5, 6 y 7 muestran cómo las variables cuantitativas y categóricas contribuyen en mayor medida a la dimensión 2, representando alrededor del 26.85% de la varianza total entre la dimensión 1 y 2. Este porcentaje registra un cambio ligero con respecto a la imputación realizada mediante el algoritmo *ClustImputeMix*. A pesar de ello, la asignación de individuos a cada grupo se deteriora, tal como evidencian los índices de validación de grupos. Además, el Cuadro 10 indica una marcada fluctuación en el tamaño de los grupos, lo que podría ser otro elemento influyente en el deterioro de la calidad de la agrupación.

7.5 Caracterización de los grupos de ClustImputeMix

En el grupo 1 se registró la siembra en el pasado en el tercer trimestre, con un cultivo asociado utilizando semillas no certificadas y desarrollado en áreas al aire libre. Se obtuvo un rendimiento promedio de alrededor de 0.65 toneladas, con un rendimiento medio de aproximadamente 1.4 toneladas por hectárea. El cultivo no se vio afectado por factores naturales, y se sembró en un área promedio de alrededor de 0.65 hectáreas, cosechándose alrededor de 0.26 hectáreas.

Este grupo representa casos en los que el cultivo fue realizado en el Trimestre 3, con cultivos asociados y utilización de semillas no certificadas en terrenos de cielo abierto. En promedio, tienen una cantidad obtenida baja, un rendimiento moderado de alrededor de 1.4 toneladas/hectárea y áreas sembradas y cosechadas relativamente pequeñas.

En el grupo 2, se constata una siembra pasada en el cuarto trimestre, con un cultivo asociado y el uso de semillas no certificadas en áreas al aire libre. Se logró una cantidad promedio de alrededor de 1.7 toneladas, con un rendimiento considerable de aproximadamente 12.2 toneladas por hectárea. Este cultivo no fue afectado por factores naturales y se sembró en un área promedio de alrededor de 0.35 hectáreas, con una superficie cosechada de aproximadamente 0.15 hectáreas.

En este grupo, los cultivos pasados se llevaron a cabo en el Trimestre 4, también con cultivos asociados y semillas no certificadas en terrenos de cielo abierto. Se observa un incremento notable en la cantidad obtenida y un rendimiento muy superior, con áreas sembradas y cosechadas en promedio más grandes que el Grupo 1.

En el grupo 3 se registra una siembra en el pasado durante el segundo trimestre, con un cultivo solitario utilizando semillas no certificadas en áreas al aire libre. Se obtuvo una cantidad promedio de alrededor de 22.6 toneladas por hectárea, con un rendimiento promedio significativo de aproximadamente 15.1 toneladas por hectárea. Este cultivo no se vio afectado por factores naturales y se sembró en un área promedio de alrededor de 1.68 hectáreas, con una superficie cosechada igual a la superficie sembrada.

Este grupo muestra una diferencia considerable, con cultivos pasados en el Trimestre 2, cultivos en solitario y semillas no certificadas en terrenos de cielo abierto. Exhiben una cantidad obtenida y un rendimiento extremadamente alto en comparación con los otros grupos, con áreas sembradas y cosechadas también notoriamente más grandes.

En el grupo 4 se evidencia una siembra pasada durante el tercer trimestre, con un cultivo solitario utilizando semillas no certificadas en áreas al aire libre. Se logró una cantidad promedio de alrededor de 0.59 toneladas por hectárea, con un rendimiento promedio de aproximadamente 1.7 toneladas por hectárea. Este cultivo no se vio afectado por factores naturales y se sembró en un área promedio de alrededor de 0.61 hectáreas, cosechando aproximadamente 0.26 hectáreas.

Similar al Grupo 1, estos individuos tienen cultivos previos en el Trimestre 3, pero con cultivos en solitario en terrenos de cielo abierto y semillas no certificadas. Muestran una cantidad obtenida similar al Grupo 1, con un rendimiento promedio comparable y áreas sembradas y cosechadas relativamente pequeñas.

Cuadro 10: Tamaño de Cada Grupo

1	2	3	4
2293	3231	273	1759

Como resultado del FAMD, en la figura 2 se obtienen las variables cuantitativas mejor representadas, la variable con menor representación en la dimensión 1 y 2 es el área sembrada. Las dimensiones 1 y 2 representan el 23,5% de la varianza total.

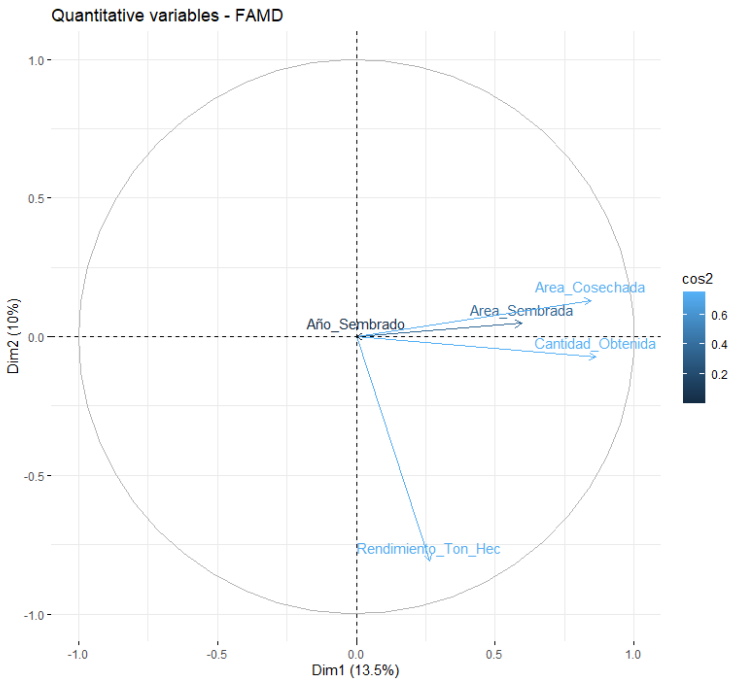

Figura 5: Grafico variables cuantitativas FAMD

En la figura 3, se observan las categorías mejor representadas, en cada dimensión, las variables cualitativas son las responsables de que las dimensiones 1 y 2 representan un porcentaje bajo de la varianza total.

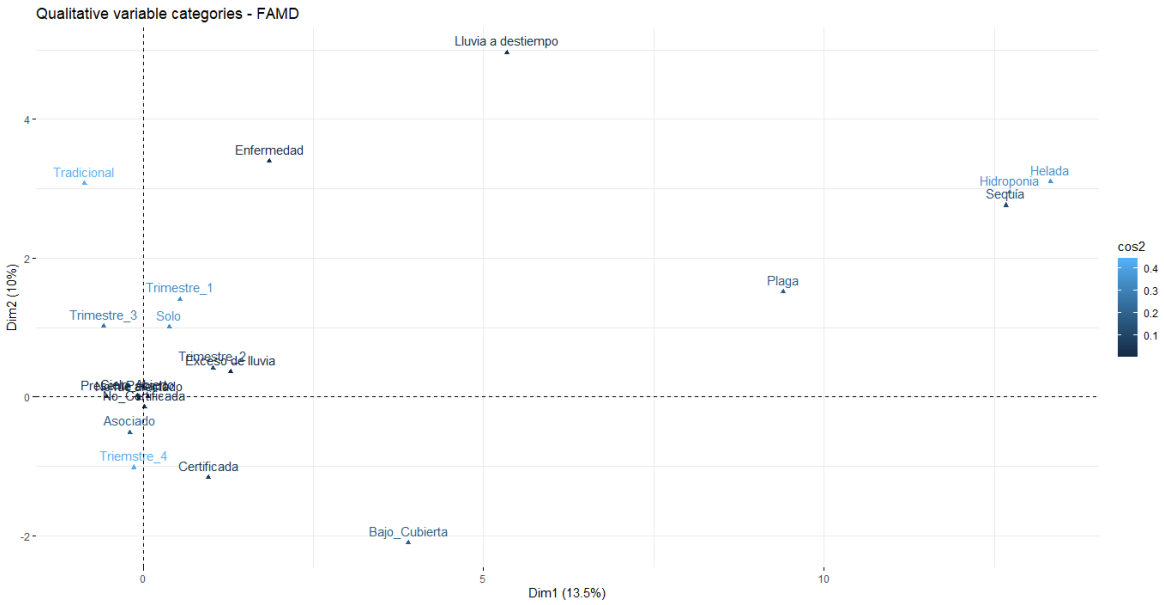

Figura 6: Grafico variables cualitativas FAMD

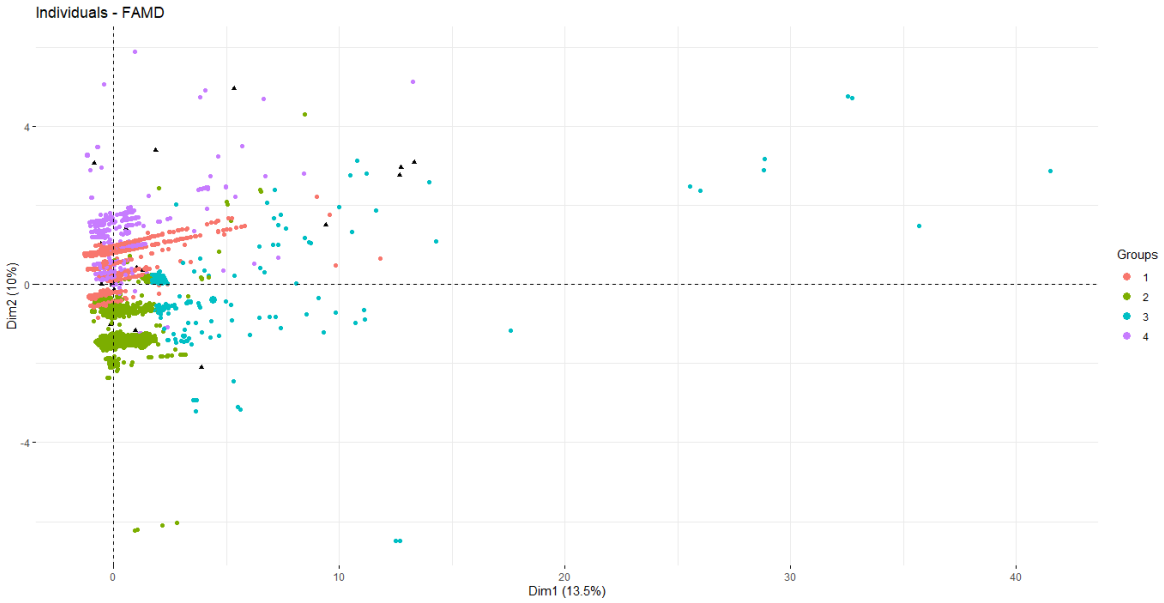

Figura 7: Grafico de individuos FAMD

En la Figura 4 se visualizan los individuos y su respectiva pertenencia a cada grupo, lo que permite observar sus características. Esta representación se apoya en los Gráficos 2 y 3, los cuales muestran las características de los individuos en cada grupo de manera más detallada.

8. Conclusiones

El objetivo principal de este trabajo fue desarrollar una función capaz de imputar valores faltantes en variables cualitativas y mixtas. Este proceso fue posible en base a la aplicación de un método de agrupación junto con un algoritmo de imputación, se implemento un algoritmo en R llamado ClustImputeMix.

La función propuesta, aplicada al capítulo de cultivos de Bogotá del III Censo Nacional Agropecuario, permite una imputación adecuada, de acuerdo con la información de cada variable haciendo uso del algoritmo propuesto que combina los métodos de agrupación k-prototipos junto con la función ClustImpute.

La base de datos presentaba valores faltantes para 7556 unidades de observación, lo que afectaba la capacidad de obtener conclusiones adecuadas sobre este sector económico en la capital. Al aplicar el algoritmo, fue posible realizar una caracterización más precisa de los individuos observados, mejorando así la comprensión y el análisis de esta área económica.

Los resultados obtenidos sugieren que la función representa una alternativa viable al MICE para la imputación de variables cualitativas y mixtas. Se espera que esta función propuesta sea de gran utilidad en futuras investigaciones, ofreciendo una herramienta útil para el manejo de datos con valores faltantes en este contexto específico.

Los resultados sugieren que el trimestre de siembra ejerce una notable influencia en la producción, destacándose el trimestre 2 como el período con el rendimiento más significativo. Además, la variabilidad observada en el tipo de cultivo y el tamaño de las áreas sembradas y cosechadas entre los distintos grupos apunta a posibles diferencias en prácticas agrícolas o condiciones de cultivo. Esta variabilidad podría reflejar estrategias específicas aplicadas por los agricultores en diferentes momentos del año, así como adaptaciones a condiciones climáticas o de suelo. Estos resultados destacan la importancia de considerar el momento de siembra y las prácticas agrícolas variadas para optimizar el rendimiento y la producción en planificaciones agrícolas.

Referencias

Alfonso, O. A. y Barrera, R. A. (2019). El ciclo mortal de los habitantes de calle en Bogotá. Revista de Economía Institucional, 21(41), julio-diciembre. https://doi.org/10.18601/01245996.v21n41.05

Arteaga, F. y Ferrer-Riquelme, A. (2009). "Missing data". En S.D. Brown, R. Tauler y B. Walczak (Eds.), Comprehensive Chemometrics. Elsevier, Oxford, pp. 285-314.

Azur, M. J. (2011). Multiple imputation by chained equations: what is it and how does it work? International Journal of Methods in Psychiatric Research, 20(1), 40-49.

Bai, L., Cao, F., & Liang, J. (2009). A new initialization method for categorical data grouping. Expert Systems with Applications, 36(3), 5992-5998.

Cao, L., & Zhao, X. (2016). A grouping-based imputation approach to missing data in a fault detection system. Neurocomputing, 173, 693-703.

Pfaffel, Oliver. (2020). CLUSTIMPUTE: AN R PACKAGE FOR K-MEANS CLUSTERING WITH BUILD-IN MISSING DATA IMPUTATION.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

Huang, Z. (1997). A fast grouping algorithm to group very large categorical data sets in data mining. Data Mining and Knowledge Discovery, 1(3), 275-288.

Huang, Z. (1998). Extensions to the k-modes algorithm for grouping large data sets with categorical values. Data Mining and Knowledge Discovery, 2(3), 283-304.

Hwang, J. T., & Lee, J. D. (2010). A grouping-based imputation method for missing data. Computational Statistics & Data Analysis, 54(12), 3095-3107.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. ACM Computing Surveys (CSUR), 31(3), 264-323.

Li, J. (2019). A novel grouping-based imputation algorithm for mixed data. Neurocomputing, 331, 322-329.

Little, R. J. A., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., ... Stern, H. (2012). The prevention and treatment of missing data in clinical trials. New England Journal of Medicine, 367(14), 1355–1360.

Lloyd, S. P. (1982). Least squares quantization in PCM. IEEE Transactions on Information Theory, 28(2), 129-137.

van Buuren, S. (2011). "mice: Multivariate imputation by chained equations in R". Journal of Statistical Software, 45, 1-67.

McClain, J. O., & Rao, V. R. (1975). Clustering and classification in marketing research. Journal of Marketing Research, 12(2), 129-134.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50(2), 159-179.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

Xiao-Hua, Z. (2019). Challenges and strategies in analysis of missing data. Biostatistics Epidemiology, 4, 15–23.

Zhao, X., Zhang, S., Wu, X., & Chen, J. (2015). Group-based missing value imputation. Information Sciences, 314, 85-101.

Zhao, Y., & Liu, H. (2014). Grouping-based multiple imputations for missing data. Information Sciences, 265, 1-12.

Zhang, J., & Yang, X. (2009). A survey on multi-objective evolutionary algorithms for the solution of the portfolio optimization problem. European Journal of Operational Research, 204(1), 28-46.

8. Anexos

- Scrip en R: función propuesta para el tratamiento de variables cualitativas y mixtas *ClustImputeMix.*
- Scrip en R: análisis de la base y aplicación de la función.
- Carta de habeas data.