“ELABORACIÓN DEL ÍNDICE DE CALIDAD DEL CAFÉ A PARTIR DE LAS CARACTERÍSTICAS ORGANOLÉPTICAS Y AGRONÓMICAS EMPLEANDO ANÁLISIS FACTORIAL BAYESIANO”

CLAUDIA ALEJANDRA SALAMANCA ROMERO

UNIVERSIDAD DEL VALLE

FACULTAD DE INGENIERIA

PROGRAMA DE ESTADISTICA

SANTIAGO DE CALI

2012

“ELABORACIÓN DEL ÍNDICE DE CALIDAD DEL CAFÉ A PARTIR DE LAS CARACTERÍSTICAS ORGANOLÉPTICAS Y AGRONÓMICAS EMPLEANDO ANÁLISIS FACTORIAL BAYESIANO”

CLAUDIA ALEJANDRA SALAMANCA ROMERO

Proyecto De Grado

Director:

Álvaro Flórez

UNIVERSIDAD DEL VALLE

FACULTAD DE INGENIERIA

PROGRAMA DE ESTADISTICA

SANTIAGO DE CALI

2012

Nota de Aceptación:

_____

ALVARO JOSE FLORES
Director
Universidad del Valle

_____

Jurado

_____

Jurado

Dedicatoria:

A Adriana Romero, mi querida madre, a quien le debo todo, a mi esposo Frédéric Quintana, por su apoyo incondicional y a mi hija Andrea Gaëlle para que le sirva de ejemplo en su formación y la motive a construir un futuro prometedor.

AGRADECIMIENTOS

A la Universidad del Valle por haberme formado moral e intelectualmente durante todo la carrera.

A los profesores Dr. Marc Saez y Dra. Isabel Villaescusa de la Universitat de Girona - Catalunya, por toda su ayuda incondicional en la elaboración y desarrollo de este trabajo.

Al Dr. Héctor Fabio Ospina, a la Dra. Eugenia Balanta, y al Catador Manuel Peña del Comité de Cafeteros del Valle del Cauca por su valiosa colaboración.

A la empresa Federación Nacional De Cafeteros De Colombia, por la ayuda prestada al suministrar la información pertinente para el desarrollo de esta tesis de grado.

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE GRAFICAS

INTRODUCCIÓN

La zona cafetera de Colombia tiene aproximadamente 3'050.141 Ha, de las cuales el departamento del Valle del Cauca aporta 75.600 Ha distribuidas en los 39 de sus 42 municipios (FNC; 2008,2009).Por su importancia económica y social, el cultivo del café es catalogado como uno de los principales productos agrícolas del país. Para el Ministerio de Agricultura, el café es la principal fuente de demanda y desarrollo en la región interandina, donde se concentra la mayor parte de la población rural y en la cual se genera cerca de 500.000 empleos directos y alrededor de 5.000 empleos en la parte agroindustrial (Espinal, et al. 2005). A parte de los aspectos económicos y sociales, está la combinación particular de diversos factores ambientales correspondientes a la agroclimatología del cultivo (suelos, las variedades, el clima, la temperatura, la lluvia) y las prácticas de manejo en su fase de cosecha y poscosecha que finalmente resultan en producir un café de excelente calidad. Por este motivo, el Comité de Cafeteros del Valle del Cauca quiere a través de este proyecto, seleccionar e identificar dentro de su departamento, aquellas zonas cafeteras que se caracterizan por un tener una buena calidad, necesario en un futuro para los procesos de certificación y comercialización.

Para ello cuenta con alrededor de 900 pruebas de taza que abarcan aproximadamente un año completo (2009-2010), con las cuales se busca relacionar las características organolépticas del café con las características agroclimáticas. La base de datos aportada, posee una gran cantidad de información, pero con la limitante de presentar valores faltantes (missing values), para resolver este problema, se escogió una de tres técnicas de imputación de datos (listwise o case deletion - CD, Hot deck - HD e imputación múltiple –MI) según sus bondades. Después se halló el índice de calidad usando un modelo factorial bayesiano, esto porque debido a la naturaleza de los datos y a una posible y compleja parametrización, las técnicas bayesianas permiten abordar los problemas de estimación y de bondad de ajuste, mediante la inferencia en la distribución a posterior.

1. DEFINICION DEL PROBLEMA

1.1 ANTECEDENTES

Existe muy poca información acerca de estudios previos donde se haya generado un índice de calidad de Café. Cros (1979) junto con otros investigadores evaluaron la calidad del café, mediante la determinación de su fracción volátil, empleado un método para el análisis de trazas de componentes volátiles orgánicos, contenidos en el aire y los líquidos biológicos, que consiste en una preconcentración de la fracción volátil en un polímero absorbente, posteriormente, efectúan un análisis de los perfiles cromatográficos de las fracciones volátiles de un café verde sano y de un café verde maloliente de procedencia, de este modo demostraron que es posible caracterizar analíticamente un lote de café a través del estudio únicamente la fracción volátil del grano como índice de calidad del Café tostado, sin embargo, no dan información estadística de las técnicas empleadas para llegar a obtener el índice de calidad.

Inmersa en esta apreciación de calidad, el Comité departamental de Cafeteros del Valle del Cauca ha buscado identificar y fortalecer los sistemas de producción a través del programa de Cafés Especiales en el departamento, bajo criterios de sostenibilidad que faciliten calidad para satisfacer la demanda existente en el mercado nacional y mundial.(Benítez, et al. l2003; Tobasura, 2006)

1.2 FORMULACIÓN DEL PROBLEMA

El problema central que se ha considerado por parte del comité departamental cafeteros del Valle del Cauca, es que se adolece de una herramienta que permita seleccionar e identificar aquellas zonas o áreas cafeteras que se caracterizan por tener un café con buena calidad, necesarias en un futuro para los procesos de certificación y comercialización.

La solución de esta problemática, podría dar vía a los directivos, profesionales de apoyo técnico y a los caficultores a aprovechar al máximo las áreas identificadas, con el fin de contribuir en un futuro a que los cafés especiales colombianos obtengan cualquiera de los sellos o certificaciones y códigos de conducta que utiliza la FNC en el café que comercializa.

Ahora, el Comité de Cafeteros del Valle del Cauca cuenta con alrededor de 900 pruebas de taza, de un año completo (2009-2010) con las cuales se busca relacionar las características organolépticas del café con las características del cultivo, recolección y beneficio, con el fin de generar un índice de calidad y seleccionar, clasificar y zonificar los diferentes sabores de café.

Esta base de datos presenta valores faltantes, por lo que se utilizaron tres técnicas de imputación de datos de las cuales se escogió solo una y con ella se generará una base de datos completa para realizar el análisis factorial bayesiano del cual se obtuvo el índice de calidad del café.

2. JUSTIFICACION

El Comité Departamental de Cafeteros del Valle del cauca - FNC, se encuentra ejecutando el programa de cafés especiales como una de las estrategias para dar valor agregado al café suave colombiano y que surgió de las nuevas tendencias del mercado en donde el valor del café se ha alejado del cafetal y se ha acercado más a la taza, lo cual obligó a repensar la estrategia de posicionamiento del café de Colombia. Con este fin, la FNC busca entonces, generar un mayor valor para el café de Colombia mejorando los ingresos de los caficultores mediante el programa de cafés especiales y promoviendo la calidad del café como uno de los principales determinantes para competir en el mercado bajo diferentes certificaciones con los sellos de cafés sostenibles como Rainforest Alliance-certified, Utz Certified, Fairtrade y los cafés certificados como orgánicos.(Muñoz, 2007;FNC, 2006)

Desde este punto de vista entonces, se está trabajando en mejorar la capacidad de los diferentes sistemas de producción del café, permitiendo así establecer cualidades y atributos comunes y no tan comunes que hacen efecto en la calidad, así como también los procesos de beneficio y almacenamiento que en la mayoría de los casos, tienen una incidencia negativa muy marcada en la producción, calidad e imagen del café, es decir pues, que si se hacen adecuadamente todas las labores y los procesos, *se obtiene café con inocuidad*, *buena calidad física, buena calidad sensorial y una composición química natural*. Por eso, el Comité de Cafeteros se ha visto la necesidad de plantear este proyecto motivado ante todo, establecer un índice de calidad a partir de la afinidad o relación entre las características organolépticas y las características del cultivo, recolección y beneficio, que sin lugar a duda definen en un alto grado la calidad global del café . (2006, FNC; 2009, Ruiz)

Como la base de datos obtenida presenta valores faltantes y puestos que los métodos estadísticos tradicionales se aplican a conjuntos de datos completos, utilizó una técnica de imputación de datos para generar un nuevo conjunto de

datos completos. En la literatura se pueden encontrar gran cantidad de técnicas para resolver el problema de datos faltantes, pero hasta ahora no hay un consenso sobre cuál de ellas es la mejor. Amón (2010), menciona que trabajar con datos faltantes o aplicar procedimientos de imputación inapropiados, puede introducir sesgos, reducir el poder explicativo de los métodos estadísticos, restar eficiencia a la fase de inferencia y puede incluso invalidar las conclusiones del estudio, además de sobredimensionar el poder explicativo de los modelos, generar estimadores sesgados que distorsionan las relaciones de causalidad entre las variables, subestimación en la varianza y alterar el valor de los coeficientes de Correlación. Por esta razón se compararon tres técnicas de imputación Listwise ó Case Delection CD, Hot Deck- HD o Imputación Múltiple – MI y se escogió la más conveniente según sus bondades.

Como el objetivo principal del estudio es hallar un índice de calidad del café y pese al problema de valores faltantes, a la naturaleza de las variables (ordinales y continuas) y sumado a esto que existe una compleja parametrización para describir los sujetos y las características organolépticas y agronómicas, se empleó un análisis factorial bayesiano, escogido porque gracias a la Inferencia bayesiana en el factor de modelos analíticos, se afrontan problemas aplicados a la estimación de los parámetros y a la bondad de ajuste que dependan de parámetros desconocidos por medio del estudio de la distribución posteriori. Este tipo de análisis es una de las técnicas más potentes para estimar un conjunto de muestras que constituyen el estimador bayesiano que se aproxime a la distribución posteriori, porque genera estructuras factoriales utilizando métodos de simulación basados en cadenas de Markov (MCMC) que se han relevado como especialmente eficaces para este tipo de problemas. Revuelta J. (2001),Gallizo y Salvador (1997, 2003); Gallizo y otros (2000, 2002,2004).

3. OBJETIVOS

3.1 OBJETIVO GENERAL

Crear el índice de calidad de café a partir de las características organolépticas y agronómicas empleando análisis factorial bayesiano.

3.2 OBJETIVOS ESPECIFICOS

- Obtener un archivo de datos completo y consistente a través de la evaluación y selección de una de tres técnicas de imputación de datos (Listwise, hotdeck e imputación múltiple).

- Estimar el índice de calidad a través del Análisis Factorial bayesiano mediante MCMC (Cadenas de Markov Monte Carlo).

- Calcular las cargas factoriales y el porcentaje de varianza explicada, para cuantificar la contribución al factor común (índice de calidad) de las características organolépticas y agronómicas.

- Presentar la zonificación geográfica del índice de calidad y las áreas con mayor probabilidad de este índice.

4. MARCO DE REFERENCIA

4.1 ASPECTOS ESTADISTICOS

4.1.1 DATOS FALTANTES (MISSING VALUES)

En el análisis de datos es muy común encontrar valores faltantes, así como las observaciones atípicas, esta situación es una limitante para analizarlos mediante métodos estadísticos tradicionales que han sido diseñados para conjuntos de datos completos, ya que afecta sustancialmente la fase de inferencia estadística. Esta falta de datos en los registros obtenidos está asociada a diversas causas. Al desconocimiento de la información solicitada, a la fatiga del informante o al rechazo de participar de la persona que diligencia el formato.

Según se cita en Medina y Galván (2007), los valores faltantes forman parte de un conjunto de observaciones con características especiales que incluyen a los datos agrupados, agregados, redondeados censurados o truncados. En estas y otras situaciones se debe diferenciar estos registros para que puedan ser analizados adecuadamente, ya que el uso del análisis normal sin un conocimiento especifico, alteraría las medidas de dispersión y de tendencia central (Donner, 1982;, Little y Rubín ,1987, Roth et al, 1996). Aunque existen diferentes tratamientos para los valores faltantes no existe ninguno método que pueda ser recomendado a cabalidad, ya que esto depende de diferentes aspectos como el porcentaje de omisión en la variable, que si es alto, la imputación atentaría contra la confiabilidad estadística del estudio y posteriormente invalidaría las conclusiones del estudio. (Medina y Galván, 2007)

4.1.2 MECANISMO DE PÉRDIDA DE LOS DATOS FALTANTES

Han surgido muchas metodologías para el tratamiento de la información con datos faltantes dando resultados aceptables. Según Valer C. et al, (2004),

estos datos faltantes en las variables se simulan conforme el proceso en el que se produce la pérdida de la información en un experimento y según las características de la probabilidad de respuesta. Por ello Rubín en 1976, hace una rigurosa definición de los distintos supuestos que se podrían hacer acerca del mecanismo de pérdida de los datos faltantes. Si se supone que hay datos faltantes sobre una determinada variable de Y, se pueden decir que hay tres clases, pueden estar perdidos completamente al azar- MCAR, datos perdidos al azar MAR ó no perdidos al azar- MNAR (Allison, 2002) .Todo depende de la información incompleta en la variable, esta se puede presentar de manera aleatoria, o sea, estar ligada a valores correspondientes a otra variable relacionada con la que presenta pérdidas o en categorías de la propia variable, es decir, que los valores no observados sean diferentes a los observados.

La notación para los mecanismos de pérdida es la siguiente, Y (variables dependientes) es una matriz rectangular $(n\,x\,k)$, donde n es el número de casos y k es el número de variables, entonces Yij seria la $j-esima$ variable en el $i-esimo$ caso de la matriz de datos Y, ahora, si hay datos faltantes se puede descomponer la matriz Y en, $(Yobs, Ymis, R)$, donde $Yobs$ contiene los valores observados de Y, $Ymis$ contiene los valores faltantes de Y, y R está definida como la matriz indicador de respuesta que define las posiciones de valores observados y de los no observados (datos faltantes) en la matriz de datos Y y X son variables independientes (todas observadas):

$$R_{ij}\begin{cases} 0 \;\; si\; y_{ij} \;\; es \;\; no\; observado \\ 1 \quad si\; y_{ij} \;\; es \;\; observado \end{cases}$$

Cualquier mecanismo de pérdida de datos puede definirse mediante la especificación de la distribución condicional del indicador de respuesta dado los datos completos $P(R|y, X)$.

***Datos perdidos completamente al azar - MCAR*:** Ocurre cuando la probabilidad de que los valores faltantes en Y no está relacionada con el valor de Y misma o con los valores de otras variables en el conjunto de datos, cuando se cumple esta suposición, el conjunto de individuos con datos completos puede considerarse como una submuestra aleatoria simple del

conjunto original de observaciones, dicho de otra forma, si la probabilidad de los datos faltantes es independiente de los datos observados, no observados y de las variables independientes, se dice que el proceso es MCAR, es decir que la ausencia de la información no está originada por ninguna variable presente en la matriz de datos y R es independiente de y_{obs} y y_{mis}:

$$P(R|y,X) = P(R|X) \forall Y$$

***Datos perdidos al azar – MAR*:** se dice que los datos en Y son faltantes de forma aleatoria, si la probabilidad de los datos faltantes en Y no está relacionado con el valor de Y después de que se han controlado otras variables en el análisis. Es decir, que los datos perdidos son dependientes o están relacionados únicamente con los datos observados, por lo que su ausencia está asociada a variables presentes en la matriz de datos y R es independiente de y_{mis}:

$$P(R|y,X) = P(R \mid X, y_{obs}) \forall Y_{mis}$$

***Datos no perdidos al azar - MNAR*:** Ocurre cuando la probabilidad de los datos perdidos en una variable Y depende de los valores de dicha variable una vez que se han controlado el resto de las variables, es decir que la probabilidad de respuesta no es independiente de las variables no observadas completamente y posiblemente tampoco de las observadas:

$P(R|y)$ Depende de y_{mis}

El mecanismo de pérdida (proceso de no respuesta), según Allison (2002) y Heckman's (1976) para MCAR y MAR, es ignorable, esto básicamente significa que no se necesita modelar el mecanismo de de datos faltantes, como parte del proceso de estimación y para MNAR es "no ignorable ". Tanto MCAR y MAR permiten valores perdidos que dependen de la variable X. Es imposible comprobar si la condición MAR se cumple, ya que dado que no se conocen los valores de los datos que faltan, no se puede comparar los valores de los casos con y sin perder los datos para ver si difieren sistemáticamente en una variable determinada.

4.1.3 MÉTODOS PARA EL MANEJO DE DATOS FALTANTES

Existen métodos convencionales utilizados para el manejo de datos faltantes, dentro de los cuales podemos mencionar dos, *Listwise* y *PairWise,* los algoritmos de imputación se describen a continuación:

ListWise (Case Deletion) o eliminación por listas

Es una de las formas más sencillas de obtener un conjunto completo de datos y asume que los datos faltantes siguen un patrón **MCAR.** En este método se elimina sólo los casos con valores perdidos considerándose así el método más fácil y sencillo de hacer frente a los datos faltantes y se basa en la eliminación de los casos incompletos (registros con datos faltantes en una variable) del conjunto de datos. Una ventaja en el uso de eliminación por lista es que todos los análisis se calculan con el mismo conjunto de casos. Pero este método presenta dos grandes inconvenientes. En primer lugar, es la pérdida substancial de la información en bases de datos, lo que altera el tamaño de la muestra y el número de variables. (Carter, 2006; Magnani , 2004).El segundo problema es que supone que la observación de los valores que faltan no es importante y puede ser ignorado, o sea que la falta de respuesta se genera de manera aleatoria asumiendo que la submuestra de los datos eliminados tiene las misma distribución que los datos completos lo que en situaciones prácticas no se cumplen y los valores estimados serán sesgados de los parámetros poblaciones lo que podría invalidar las conclusiones. (Carter, 2006; Medina y Galván, 2007)

PairWise o Eliminación por pareja

Este método incluye todos los datos disponibles y a diferencia del método Listwise, la eliminación por parejas o Pairwise sólo elimina los valores específicos que falta en el análisis (no todo el caso o sujeto). Es decir, todos los datos disponibles se incluyen, contrario al Listwise donde se elimina los casos (sujetos) que tienen valores perdidos en cualquiera de las variables objeto de

análisis. Este método asume un patrón de datos perdidos **MCAR** en los datos omitidos, y hace uso de toda la información disponible sin efectuar ningún tipo de corrección en los factores de expansión. Las observaciones que no tienen datos se eliminan, y se hacen los cálculos con diferentes tamaños de muestra lo que limita comparación de resultados en este caso, el registro con los valores que faltan sólo se utiliza en los análisis que no requieren de la variable que falta. (Medina y Galván, 2007; Marwala, 2009). Su principal ventaja con respecto al Listwise es que preserva una mayor cantidad de información, aunque a cambio hace complicada la interpretación de los coeficientes de correlación, ya que cada coeficiente tendrá un tamaño muestral diferente. (Varela, et al. 1998)

4.1.3.1 MÉTODO DE IMPUTACIÓN SIMPLE

Los métodos de imputación simple radican en asignar un valor por cada valor faltante, tomando como referencia el valor de la propia variable o de otras variables, generando así, una base de datos completa, la técnica de imputación simple usada en esta aplicación es la **Hot-Deck**.

Hot-Deck

Es un procedimiento imputación de datos no paramétrico a través de una distribución no condicionada, que consiste en ubicar registros completos (donantes) e incompletos (receptores) e identificar las características comunes entre donantes y receptores, consecutivamente se hace la imputación entre observaciones con características comunes, la selección de los donantes se realiza en forma aleatoria, así la suposición se basa en que dentro de cada grupo de clasificación los valores faltantes siguen la misma distribución como aquellos que no son faltantes. La ventaja de este método es evitar que se introduzcan sesgos en el estimador de la varianza preservando las distribuciones conjuntas y marginales, cuando los datos faltantes **siguen un patrón MCAR** produce estimaciones consistentes (Ávila, 2002).

Existen otros métodos de imputación simple conocidos como **Medias no condicionadas** que consiste en imputar cada valor faltante por el promedio de la variable, bajo este procedimiento se asume que los datos faltantes siguen un patrón MCAR, **Medias condicionadas para datos agrupados** que es una variación de las medias no condicionadas, aquí se forman categorías a partir de las covariables correlacionadas con la variable de interés, después se imputan los datos faltantes con observaciones provenientes de una submuestra que comparte características comunes, generando así, tantos promedios como categorías hayan. Asume que los datos faltantes siguen un patrón MCAR, por **Variables ficticias** que consiste en crear una variable indicadora para identificar las observaciones con datos faltante, posteriormente se realizan las estimaciones mediante el modelo de regresión lineal como en Listwise, asignando "0" si no se conoce y "1" de otra manera, por **Regresión** que consiste en eliminar las observaciones con datos incompletos y ajustar una ecuación de regresión para predecir los valores $\hat{y}$ que se usaran para sustituir los valores de Y que faltan, de modo que valor es construido como una media condicionada de las covariables $X's$ y los datos faltantes siguen un patrón MCAR y por **máxima verosimilitud -**MV que asume que los datos faltantes siguen un patrón MAR y se demuestra que la distribución marginal de los registros observados está asociada a una función de verosimilitud para un parámetro desconocido θ, bajo el supuesto de que el modelo es adecuado para el conjunto de datos completo, el procedimiento para estimar los parámetros del modelo utiliza una muestra con datos faltantes y estima los parámetros del modelo con los datos completos empleando la función de máxima verosimilitud, luego utiliza los parámetros estimados para predecir los valores omitidos y posteriormente sustituye los datos por las predicciones obteniendo los nuevos valores de los parámetros maximizando la verosimilitud de la muestra completa hasta lograr que el valor de los parámetros no cambie entre dos iteraciones sucesivas.(Useche y Mesa, 2006; Schafer y Graham, 2002; Acock y Demo, 1994; Durrent, 2005; Shafer y Graham,2002; Little y Rubín ,1987)

4.1.3.2 MÉTODO DE IMPUTACION MULTIPLE

Rubín en 1987, introdujo el método de imputación múltiple en donde cada dato faltante debía ser remplazado a partir de $m > 1$ simulaciones a través de un proceso estocástico, aquí se seleccionan posibles valores para la información faltante y se utilizan dichos valores para obtener un conjuntos de datos completo. (Alfaro y Fuenzalida, 2009).

Van (2007) menciona que cuando se presentan variables categóricas y continuas**,** el proceso de imputación múltiple **por ecuaciones encadenas – MICE, proporciona mejores resultados**, en todos los casos la validez del proceso de imputación se basa en el supuesto de que el patrón de distribución de los valores faltantes es aleatorio MCAR o MAR (Mediavilla, 2010). El sustento teórico de la imputación múltiple, se encuentra en la estadística bayesiana, la cual utiliza la información de la muestra para realizar inferencia respecto de los parámetros. Para obtener la inferencia deseada se debe realizar sus estimaciones y combinar los resultados a través de las distintas bases de datos. Para lograr este objetivo, Rubín (1986) propone reglas simples que permiten combinar los resultados obtenidos en las distintas bases imputadas, de modo que sea posible ajustar los errores estándar de los estimadores para considerar la incertidumbre generada de las imputaciones. El método de imputación múltiple no soluciona el problema de la información faltante, sino que lo acomoda desde una perspectiva estadística, es decir que se podrá contar con información completa, pero se deberá manejar múltiples bases de datos donde cada una de ellas tiene un valor posible para la observación faltante. El investigador entonces deberá desarrollar su análisis en cada una de las $\boldsymbol{m}$ bases de datos completas y luego combinar los resultados a fin de obtener las conclusiones finales de su investigación. Según Alfaro y Fuenzalida (2009), mencionan que el número optimo de bases de datos ($\boldsymbol{m}$) depende del porcentaje de información faltante, pero según Allison (2001), en la practica la elección comúnmente utilizada es $m = 5$, sin embargo, StataCorp (2009) establece que $m = 20$ es un numero razonable de imputaciones ya que minimiza de forma empírica el rango de incertidumbre de la estimación del intervalo de confianza.

Descripción De La Técnica De Imputación Múltiple - MI

La imputación múltiple es un método de Monte Carlo descrito por Rubín (1987) con el contexto de emplearla en la falta de respuesta en las encuestas por muestreo, la técnica es muy general y puede ser fácilmente utilizada y que conducen a una única instrucción de inferencia sobre los parámetros de interés, el procedimiento de imputación múltiple se describe a continuación.

Q es la cantidad escalar de interés en el estudio para ser estimada, por ejemplo, la media población de Y, coeficiente de regresión, la correlación, etc. Con m imputaciones, se puede calcular m diferentes versiones de $\hat{Q}$ y U, sea:

$$\hat{Q}^t = \hat{Q}(Y_{obs}, Y^t_{mis}) \text{ y } U^t = U(Y_{obs}, Y^t_{mis})$$

Donde $\hat{Q}^t$ es la estimación estadística de Q y U^t su varianza estimada utilizando el t-ésimo conjunto de datos imputados $t = 1,2,\ldots,m$.

La estimación puntual de las imputaciones para Q es simplemente el promedio de las estimaciones puntuales de los datos completos:

$$\bar{Q} = \frac{1}{m}\sum_{t=1}^{m}\hat{Q}^t \quad (1)$$

Para obtener el error estándar de $\bar{Q}$ uno debe calcular la variación dentro de las imputaciones y entre las imputaciones:

- *La variación dentro de la imputación*, que es el promedio de las estimaciones de la varianza de los datos completos:

$$\bar{U} = \frac{1}{m}\sum_{t=1}^{m}U^t \quad (2)$$

- *La variación entre las imputaciones* que es la varianza de las estimaciones puntuales de los datos completos,

$$B = \frac{1}{m-1}\sum_{t=1}^{m}\left(\hat{Q}^t - \bar{Q}\right)^2 \quad (3)$$

La varianza total es definida como:

$$T = \bar{U} + (1 + m^{-1})B \quad (4)$$

El intervalo estimado y el nivel de significancia para el escalar Q son formados usando una distribución de referencia t y las inferencias son basadas en la aproximación:

$$T^{-\frac{1}{2}}(Q - \bar{Q}) \sim t_v \quad (5)$$

Donde los grados de libertad son dados por:

$$v = (m-1)\left[1 + \frac{\bar{U}}{(1+m^{-1})B}\right]^2 \quad (6)$$

Así, el intervalo de estimación $100(1-\alpha)\%$ para Q es:

$$\bar{Q} \pm t_{v,1-\alpha/2}\sqrt{T} \quad (7)$$

y un p-valor para probar la hipótesis nula $\bar{Q} = Q'$ en contra de una alternativa bilateral es:

$2P(t_v \geq |\bar{Q} - Q'| T^{-\frac{1}{2}})$ ó equivalentemente, $P\left[F_{1,v} \geq T^{-\frac{1}{2}}(\bar{Q} - Q')^2\right]$

Ahora, cuando se presentan datos faltantes se tiene en cuenta lo siguiente:
Nótese que los grados de libertad de la ecuación (6) Dependen no sólo de m, sino también de la relación:

$$r = \frac{(1+m^{-1})B}{\bar{U}} \quad (8)$$

Rubín llama a r el *incremento relativo de la varianza debido a la falta de información* o más precisamente la varianza condicional dado $B_\infty = B$, donde $\bar{U}$ representa la estimación total de la varianza cuando no hay información faltante en Q (por ejemplo cuando $B = 0$). Cuando m es grande y/o r es pequeño, los grados de libertad son grandes y se aproxima a una normal. Sin información faltante, la parte posterior llegar a ser normal con media $\bar{Q}$ y varianza $\bar{U}$, para lo cual la información es $\bar{U}^{-1}$.

Resultando en:

$$\hat{\lambda} = (\bar{U}^{-1}(v+1)(v+3)^{-1}T^{-1})\,\bar{U}$$
$$= \frac{r + 2/(v+3)}{r+1}$$

Que es una estimación de la fracción de información faltante sobre Q. En las aplicaciones, el cálculo de r y $\hat{\lambda}$ es muy recomendable, ya que son interesantes y útiles en el diagnóstico para evaluar la forma en que los datos faltantes contribuyen a la incertidumbre inferencial de Q.

Finalmente, Rubín y Schenker (1986) demuestran que este método conduce a inferencias que están bien calibrados desde un punto de vista frecuentista y que las estimaciones del intervalo de las imputaciones múltiples cubren el parámetro real al menos el 95% del tiempo, incluso cuando se tiene un m tan pequeño como 2. La razón fundamental para esto en primer lugar, es que la imputación múltiple se basa en la simulación para resolver problema del aspecto de los datos perdidos. Como con cualquier método de simulación, una forma efectiva podría eliminar el error Monte Carlo eligiendo un m lo suficientemente grande, pero, con la imputación múltiple la ganancia resultante de la eficiencia típicamente sería poco importante debido a que el error de Monte Carlo es una parte relativamente pequeña de la incertidumbre inferencial general y la otra razón, son las reglas para combinar los m análisis de datos completos para el conteo de forma explícitamente del error Monte Carlo. Ahora, cuando hay muestras grandes como se demuestra en Rubín 1976, los procedimientos de imputación conducen a inferencias que son válidas porque la inferencia en los datos completos es basada en la aleatorización cuando hay ausencia de falta de respuesta y además, porque el método de imputación es correcto. Ahora, se puede calcular eficiencia relativa cuando se utiliza un número finito de imputaciones adecuadas m, en lugar de un número infinito.

$$RE = (1 + \lambda / m)^{-1/2}$$

Donde el escalar λ, es la fracción de la información faltante y la eficiencia relativa- RE de un punto estimado es dada unidades de errores estándar. Cuando $\lambda = 0{,}2$, por ejemplo, una estimación basada en $m = 3$ imputaciones tenderá a tener un error estándar sólo en un $\sqrt{1 + (0.2 / 3)} = 1.033$ veces más grande que el estimado con $m = \infty$ ó con $\lambda = 0.5$, una estimación basada en

$m = 5$ imputaciones tienden a tener un error estándar de $\sqrt{1 + (0.5 / 5)} = 1.049\,1$ veces más grande que el estimado con $m = \infty$. Por ello las imputaciones con m pequeño son tal validas colo las obtenidas con un m grande.

Imputación Múltiple por Ecuaciones Encadenadas – MICE

Según la publicación de Melissa, et al (2011), y Stef y Karim (2011), las ecuaciones encadenados se utilizan para imputar valores faltantes cuando las variables pueden ser de diferentes tipos y los patrones de valores perdidos son arbitrarios. Para hacer frente a los problemas planteados por la complejidad real de los datos, la especificación se produce al nivel variable. Hipotéticamente los datos completos Y son muestras aleatorias parcialmente observadas de la P-Variable de la distribución multivariada $P(Y_{obs}|\theta)$, se supondrá entonces que la distribución multivariante de Y está completamente especificada por θ, explícita o implícitamente. Las ecuaciones encadenadas proponen para obtener una distribución posterior de θ, mediante el muestreo de forma iterativa a partir de distribuciones condicionales de la forma:

$$P(Y_1|Y_{-1}, \theta_1), P(Y_2|Y_{-2}, \theta_2), \ldots\ldots P(Y_p|Y_{-p}, \theta_p)$$

Donde $Y = (Y_1, Y_2, \ldots, Y_p)$ es un conjunto de p variables aleatorias, $Y_j\ con\ (j = 1,2, \ldots, p)$ es una de p variables incompletas y Y_{-j}=$(Y_1, Y_2, \ldots Y_{j-1}, Y_{j+1}, \ldots Y_p)$ es la coleccion de $p - 1$ variables en Y exceptoY_j, los parámetros $\theta_1, \theta_2, \theta_3 \ldots, \theta_p$ son específicos de las densidades condicionales respectivas y no necesariamente son el producto de una factorización de la "verdadera" distribución conjunta $P(Y_{obs}|\theta)$.

A partir de una simple extracción de la distribución marginal observada, la $t - esima$ iteración de la cadena de las ecuaciones se realiza mediante el muestreador de gibbs:

$$\theta_1^{*t} \sim P\left(\theta_1 \left| Y_1^{(obs)}, Y_2^{(t-1)}, \ldots\ldots, Y_p^{(t-1)}\right.\right)$$

$$Y_1^{*t} \sim P\left(Y_1 \left| Y_1^{(obs)}, Y_2^{(t-1)}, \dots\dots, Y_p^{(t-1)}, \theta_1^{*t}\right.\right)$$
$$\vdots$$
$$\theta_p^{*t} \sim P\left(\theta_p \left| Y_p^{(obs)}, Y_1^{(t)}, Y_2^{(t)}, \dots\dots, Y_{p-1}^{(t)}\right.\right)$$
$$Y_p^{*t} \sim P\left(Y_p \left| Y_p^{(obs)}, Y_1^{(t)}, Y_2^{(t)}, \dots\dots, Y_p^{(t)}, \theta_p^{*t}\right.\right)$$

Donde $Y_j^{(t)} = (Y_j^{(obs)}, Y_j^{*(t)})$ es la $j - esima$ variable imputada para la iteración t , se puede observar que las imputaciones anteriores $Y_j^{*(t-1)}$ sólo entran $Y_j^{*(t)}$ a través de su relación con otras variables, y no directamente. El nombre de ***ecuaciones encadenadas*** se refiere al hecho de que el muestreador de Gibbs se puede implementar fácilmente como una concatenación de los procedimientos univariantes para completar los datos faltantes.

MICE es entonces, una técnica de imputación múltiple (Raghunathan et al, 2001), que opera bajo el supuesto de que, dadas las variables utilizadas en el procedimiento de imputación, los datos faltantes están ausentes de forma aleatoria (MAR), lo que significa que la probabilidad de que un valor no esté presente sólo depende de los valores observados y no de los valores no observados. En otras palabras, después de controlar todos los datos disponibles (es decir, las variables incluidas en el modelo de imputación) "cualquier valor faltante restante es completamente al azar" y cuando los datos no son perdidos MAR podría dar lugar a estimaciones sesgadas. Schafer y Graham , 2002,Graham, 2009,Stuart et al, 2009).

Pasos MICE

En el proceso de ecuación en cadena el problema es extraer las imputación de $P(Y)$, la distribución incondicional multivariado de Y. Sea t un contador de iteración. Se puede repetir la siguiente secuencia de iteraciones utilizando el muestreador de Gibbs:

Para Y_1: extraer imputaciones de Y_1^{t+1} de $P(Y_1 | Y_2^t, Y_3^t, \dots, Y_p^t)$

Para Y_2: : extraer imputaciones de Y_2^{t+1} de $P(Y_2 | Y_1^{t+1}, Y_3^{t}, \dots, Y_p^{t})$

Para Y_j: extraer imputaciones de Y_j^{t+1} de $P(Y_j | Y_1^{t+1}, Y_2^{t+1}, \dots, Y_{j-1}^{t})$

Con las anteriores ecuaciones se puede deducir los siguientes pasos:

1) Se realiza una imputación simple, como la imputación de la media, para cada valor faltante en el conjunto de datos. Estas medias imputadas pueden ser consideradas como "marcadores de posición".

2) El "marcador de posición" imputación de la media para una variable (Y_j) es un conjunto de regreso para los valores faltantes.

3) Los valores observados de la variable "Y_j" en el paso 2 son regresores en las otras variables en el modelo de imputación, que pueden o no pueden consistir en todas las variables del conjunto de datos. En otras palabras, "Y_j" es la variable dependiente en un modelo de regresión y todas las otras variables son variables independientes en dicho modelo. Estos modelos de regresión operan bajo los mismos supuestos que uno hace cuando se realizan los modelos lineales de regresión logística, o Poison, es decir, fuera del contexto de imputación de datos faltantes.

4) Los valores faltantes para " Y_j " se reemplazan con las predicciones (imputaciones) del modelo de regresión. Cuando se utiliza " Y_j " posteriormente como una variable independiente en los modelos de regresión de las variables, tanto los valores observados como los imputados son utilizados.

5) Los pasos 2-4 se repiten para cada variable con datos faltantes. El ciclaje a través de cada una de las variables constituye una iteración o "ciclo". Al final de un ciclo de todos los valores faltante se han sustituido con las predicciones de las regresiones que reflejan las relaciones observadas en los datos.

6) los pasos 2-4 se repite para un número de ciclos, con las imputaciones siendo actualizadas y alimentadas en cada ciclo.

Al final de estos ciclos las imputaciones finales se conservan, dando como resultando un conjunto de datos completo. La idea es que al final de los ciclos la distribución de los parámetros que rigen las imputaciones haya convergido en el sentido de convertirse en estable. Esto, porque de cierta manera, evita la dependencia del orden en el que las variables se imputan. Una vez que el número designado de ciclos se ha completado, el proceso de imputación completo se repite para generar múltiples conjuntos de datos imputados. Los datos observados, evidentemente, serán los mismos a través de los conjuntos de datos imputados; sólo los valores que originalmente habían sido faltantes serán diferentes.

Ahora, una correcta especificación del modelo, puede mejorar aún más el modelo de imputación y la creación de imputaciones válidas, para ello, se deben tener en identificar las condiciones bajo las cuales una variable debe o no debe ser imputados. Esto podría ocurrir si hay falta de diseño en los datos o si hay cuestiones jerárquicas (También llamado salto de patrones). Después de la especificación del modelo sigue la evaluación del procedimiento de imputación, y ocurre una vez que el modelo de imputación se ha especificado y ha creado las imputaciones iniciales, aquí se revisa el modelo y se perfecciona si es necesario. Esto puede hacerse a través del Resumen de estadísticas SUMMARY que proporcionan información sobre los valores observados, los valores imputados donde se pueden comparan las diferencias de medias y desviaciones estándar, muy útil para la identificación de las variables problemáticas y las variables que hay que modificar antes de su uso en los análisis. Otras comparaciones gráficas y numéricas entre los datos observados e imputados que ofrecen representaciones visuales de la medida en que los valores imputados se diferencian de los valores observados y que son muy útiles en el diagnóstico de problemas (Abayomi et al., 2008), son los histogramas, los diagramas de cuantil-cuantil y los gráficos de densidad.

Finalmente, según Stuart et al. (2009) es importante reconocer que variables presentan diferencias muy marcadas entre los valores observados y los imputados, pero esta magnitud no necesariamente significa que las

imputaciones no son exactas, sino que de hecho puede ser razonables debido a que puede ser indicativo del sesgo de las imputaciones que se están tratando de resolver.

4.1.3.3 DISCUSIÓN DE LOS MÉTODOS DE IMPUTACION

Para seleccionar un buen método de imputación se debe tener claro que algunas técnicas podrían dar mejores aproximaciones a los valores reales de la investigación, por eso, en la tabla 1, muestra las ventajas y desventajas de los métodos de imputación (CD, HD y MI) seleccionados para este estudio. Sin embargo, no hay reglas concretas, pero esto dependerá del tipo y tamaño del conjunto de datos, de los objetivos de la investigación, características específicas de la población, características generales de cada variable, etc. Goicoechea (2002), indica los criterios que se deben tener en cuenta para seleccionar una técnica de imputación adecuada y se describen a continuación:

1) Determinar el tipo de variable a imputar: Si es continua, tomar en cuenta el Intervalo para la cual se define, y si es cualitativa, tanto nominal como ordinal, las categorías de las variables.

2) Seleccionar los parámetros que se desean estimar: Si se desea conocer sólo valores agregados como la media y el total, se pueden aplicar métodos sencillos como imputación con la media o moda, sin embargo, puede haber subestimación de la varianza. En caso de que se requiera mantener la distribución de frecuencia de la variable y las asociaciones entre las distintas variables, se deben emplear métodos más elaborados aplicando imputación de todas las variables faltantes del registro.

3) calcular las tasas de no respuesta y exactitud necesaria: Cuando el porcentaje de no respuesta es alto en una base de datos, se considera que no hay confiabilidad en los resultados que se obtengan con el análisis de esta base.

4) Información auxiliar disponible: información auxiliar disponible, ya que con ella podemos deducir información de los valores ausentes de una

variable o hallar grupos homogéneos respecto a una variable auxiliar que se encuentre altamente correlacionada con la variable a imputar, y de esta manera encontrar un donante adecuado que sea similar al registro receptor.

Para evaluar el método de imputación el mismo autor sugiere los siguientes pasos:

1) Una vez que se dispone de un archivo con datos faltantes, se recopila y valida toda la información auxiliar disponible que pueda ser de ayuda para la imputación.

2) Se estudia el patrón de pérdida de respuesta. Posteriormente se observa si hay un gran número de registros que simultáneamente tienen no respuesta en un conjunto de variables.

3) Se seleccionan varios métodos de imputación posibles y se contrastan los resultados.

4) Se calculan las varianzas para los distintos métodos de imputación seleccionados con el objetivo de obtener estimaciones con el mínimo sesgo y la mejor precisión.

5) concluye a partir de los resultados obtenidos.

Cabe anotar que la validez del método se sustenta en la forma como se obtienen las imputaciones, ya que los estimadores finales se calculan como un promedio de los valores sustituidos lo cual garantiza que se mantenga la variabilidad en los valores imputados.

Tabla 1. Ventajas y desventajas de los métodos de imputación CD, HD y MI, según autores como Otero, (2011); Goicochea (2002) – Amón (2010)

Listwise: Eliminación de registros con valores faltantes.		**Hot - deck** : Es un procedimientos de Duplicación. Cuando falta información en un registro se duplica un valor ya existente en la muestra para reemplazarlo.		**Múltiple**: se basa en la imputación de más de un valor para cada valor ausente. MI consiste en generar m>1 valores aleatorios para cada valor perdido por no respuesta de manera que se dispone de m conjuntos de datos completos	
Ventajas	**Desventajas**	**Ventajas**	**Desventajas**	**Ventajas**	**Desventajas**
Facilidad de su implementación y posibilidad de comparar los estadísticos univariante	Perdida de información, asume que la submuestra de datos excluidos tiene las mismas características que los datos completos	Los procedimientos conducen a una post-estratificación sencilla. No presentan problemas a la hora de encajar conjuntos de datos No se necesitan supuestos fuertes para estimar los valores individuales de las respuestas que faltan. Conserva la distribución de la Variable.	Cuando hay muchos datos faltantes se reporta muchas veces mismo valor duplicado. Distorsiona la relación con el resto de las variables Carece de un mecanismo de probabilidad Requiere tomar decisiones subjetivas que afectan a la calidad de los datos, lo que imposibilita calcular su confianza Proporciona estimaciones sesgadas de la media	Incrementa la eficiencia de los estimadores ya que minimiza los errores estándares Obtiene inferencias validas simplemente mediante la combinación de las inferencias conseguidas con las bases de datos completas. Estudia directamente la sensibilidad de las inferencias usando los métodos de las bases de datos completas.	Requiere un mayor esfuerzo y tiempo para crear las bases de datos imputadas. No produce una única respuesta, el investigador deberá manejar múltiples bases de datos donde cada una de ellas tiene un valor posible para la observación faltante.

4.1.3.4 ANALISIS FACTORIAL

Es una técnica estadística multivariada que tiene como principal objetivo sintetizar y explicar las interrelaciones observadas entre un conjunto de variables observadas mediante un pequeño número de variables latentes no observadas (no medidas). Estas variables aleatorias inobservables, llamadas ***factores comunes***, se emplean para explicar todas las covarianzas o correlaciones y cualquier porción de la varianza que es inexplicada por dichos factores se asigna a términos de errores residuales llamados factores ***únicos o específicos***. Estos factores no deben estar correlacionados entre sí y es posible que la búsqueda de ellos se confunda con la búsqueda de las componentes principales de las variables, sin embargo, el análisis de componentes principales - ACP se construyen para explicar las varianzas entre las variables, mientras que, los factores se construyen para explicar las covarianzas o correlaciones entre las variables, también el ACP es una herramienta descriptiva o de tipo exploratoria, mientras que el análisis factorial- AF exige un modelo estadístico formal para la generación de datos, y por lo tanto acude a las pruebas de hipótesis y a la inferencia. El AF puede ser un análisis **exploratorio** cuando no se conocen a priori el número de factores y es en la aplicación empírica donde se determina este número y puede ser de tipo **confirmatorio cuando** los factores están fijados a priori, utilizándose contrastes de hipótesis para su corroboración. (Pérez y Santín, 2007; Kline ,1998).

Los supuestos que fundamentan la técnica implican, primero que factores pueden usarse para explicar fenómenos complejos y segundo, que las correlaciones observadas entre las variables son el resultado del hecho que tales variables comparten los mismos factores. Por lo tanto, caben dos enfoques en el análisis factorial, el primero, cuando se analiza toda la varianza, la común o compartida y la no común (Varianza específica de cada variable + Varianza debida a errores de medición), en este caso el método más empleado, es el de análisis de **Componentes Principales – ACP.** En el segundo enfoque, se puede analizar solamente la varianza común o compartida, en este caso se sustituyen los unos de la diagonal de la matriz correlaciones por las estimaciones de la varianza que cada

ítem tiene en común con los demás y que se denominan comunalidades. Para estimar las comunalidades no hay un cálculo definido y para ello existen diversos procedimientos como las correlaciones múltiples o coeficientes de fiabilidad. A este procedimiento se le denominan **análisis de factores comunes – AFC**. Los pasos en el AF comienzan con: (Grajales 2000; Morales, 2011; Morales y Martínez, 2008).

a) El análisis de la matriz de correlación

Se computa una matriz de correlación de todas las variables con el objetivo de de identificar las variables no correlacionadas con otras. Las variables que tienen pequeñas o nulas correlaciones entre sí, son aquellas que no comparten factores en común. Por medio de la **prueba de esfericidad de Bartlett** se prueba la hipótesis de que la matriz de correlación es una matriz de identidad (*todos los valores en la diagonal son 1 y todos fuera de la diagonal son 0).* Es decir que con esta prueba, se muestra la probabilidad estadística de que la matriz de correlación tiene correlaciones significativas al menos entre algunas variables. La significancia debe ser menor a 0.05 para proceder con la técnica. Otra forma de observar el grado de correlación entre la variables es por medio de los **coeficientes de correlación parcial** que son estimados de la correlación entre los factores únicos y deben ser cercanos a cero. Cuando se cumplen las suposiciones del AF la matriz que contiene los coeficientes de correlación parciales negativos *(La matriz de correlaciones anti-imágen)* debe mostrar una proporción muy reducida de coeficientes de correlación altos, a fin de que pueda considerarse el análisis factorial. Además, existe una **medida de la adecuación de la muestra (MSA)** que a través de un índice compara las magnitudes de los coeficientes de correlación observados y las magnitudes de los coeficientes de la correlación parcial, es decir que cuantifica el grado de intercorrelación entre las variables y lo apropiado del análisis de factor. Si una variable tiene una MSA igual 1, puede ser predicha de manera perfecta sin error por las otras variables en el estudio. Otra prueba de adecuación de la muestra se denomina **Káiser-Meyer-Olkin (KMO)** en la cual su índice nos indica que cuando presenta valores pequeños (cercanos a cero), no es

recomendable usar AF, ya que las correlaciones entre pares de variables no son explicadas por otras variables. **El coeficiente R^2** o coeficiente de correlación múltiple al cuadrado entre una variable y el resto, es otro indicador de la fortaleza de la asociación lineal entre las variables y es reconocido como Comunalidad. Cuando este coeficiente es pequeño para un variable en particular, es recomendable considerar la posibilidad de eliminarla del conjunto de variables en estudio.

b) Extracción factorial

El objetivo es establecer los factores que subyacen en las variables medidas u observadas y existen varias técnicas que se emplean, como **el ACP** donde se forma una combinación lineal de las variables observadas con el propósito de seleccionar el número de factores necesarios para representar los datos y que almacena el porcentaje total de la varianza que es explicada por cada uno de ellos, aquí la primera componente principal recoge la mayor cantidad de la varianza muestral total, a la segunda componente principal le corresponde la siguiente cantidad de varianza inmediatamente inferior a la primera y no está correlacionada con la primera componente. Así sucesivamente el resto de las componentes explican proporciones menores de la varianza muestral total, cabe aclarar que la varianza total es la suma de las varianzas de cada variable. El **AFC** analiza la varianza común o compartida de todas las variables en factores, donde cada factor está compuesto por todos los ítems, pero en cada instrumento los ítems tienen un peso específico distinto según sea su relación con cada factor. La varianza de todas estas nuevas medidas es equivalente a la máxima varianza de la medida original que es posible explicar. Por ello se dice que en el AFC se reduce a la búsqueda de los pesos específicos atribuido a cada factor. Los factores con coeficiente grandes (en valores absolutos) para una variable son factores que están estrechamente relacionados con dicha variable. Esta información se puede observar en la *matriz de estructura factorial* que muestra las correlaciones simples entre las variables y los factores además de contener las correlaciones entre los factores, pero estos pesos contienen la varianza única o

compartida entre las variables y los factores más las correlaciones entre los factores, cuando estas correlaciones entre los factores tienden a ser más grandes, se hace más difícil distinguir cuáles son variables que aportan un mayor peso en cada factor.

Cuando los factores estimados no se han correlacionado entre sí, es decir, que son ortogonales, se puede calcular la proporción de la varianza de cada variable que es explicada por el modelo factorial para emitir un juicio sobre la bondad del modelo factorial al describir las variables originales. Esto porque al no estar correlacionados los factores, el total de la proporción de varianza explicada es simplemente la suma de las proporciones de las varianzas explicadas por cada factor. Finalmente, el objetivo de la fase de extracción de los factores es determinar el número de factores comunes necesarios para describir los datos y es una decisión que se basa en los eigenvalues y el porcentaje de la varianza total que aporta cada uno de los diferentes factores.

c) La fase de rotación

Con la rotación de los factores se busca transformar la matriz inicial en otra matriz que sea más fácil de interpretar para poder identificar los factores que son substancialmente significativos. Existen dos tipos de rotación la ortogonal o la oblicua. La **rotación es ortogonal** cuando los ejes de coordenadas se giran manteniendo un ángulo de 90 grados entre ellos y esto asume que los factores que son identificados no se relacionan entre sí, contrario una **rotación oblicua que** ocurre cuando los ejes que se rotan conservan entre sí un ángulo diferente a 90 grados y se asume que hay cierto grado de relación entre los factores que lleguen a conformarse. La rotación de los ejes no afecta la bondad de la solución factorial, y aunque la matriz factorial cambia así como los porcentajes atribuibles a cada factor, las comunalidades y los porcentajes de la varianza total explicada no varían, es decir pues que la rotación solo redistribuye la varianza explicada por los factores individuales. La forma de calcular la matriz factorial da lugar a los distintos *métodos de rotación ortogonales* y oblicuos de los cuáles los más utilizados son

los siguientes: **Método Varimax** (minimiza el número de variables que tienen una alta carga en un factor para fortalecer la interpretación de dichos los factores. Este método busca aumentar varianza de las cargas factoriales al cuadrado de cada factor logrando que algunas de sus cargas factoriales tiendan a aproximarse a uno, mientras que otras se acerquen a cero, logrando un sentido de pertenencia más palpable de cada variable a ese factor), **Método Quartimax** (El objetivo de este método es minimizar el factores necesarios para explicar la variable, donde cada variable número de tenga una alta correlación con un pequeño número de factores a través de la maximización de la varianza de las cargas factoriales al cuadrado de cada una de ellas en los factores, es decir, que el método logra que cada variable concentre su pertenencia en un factor determinado a través de una carga factorial alta mientras que en el resto de los factores, las cargas factoriales tiendan a ser bajas. En cuanto a la interpretación con el método se suele ser más claro en relación a la comunalidad total de cada variable), **Método Equama** (Este método es una combinación de los dos anteriores, que busca maximizar la media de los criterios anteriores a través de la simplificación de los factores y las variables. Para su interpretación es necesario agrupar las variables que tienen una alta carga respecto al mismo factor o también se puede requerir que las cargas pequeñas sean omitidas posteriormente es indispensable un esfuerzo teórico lógico por parte del investigador para encontrar y demostrar significado y sentido a los resultados), **Método Oblimin directo** (En la rotación oblicua las cargas factoriales no coinciden con las correlaciones entre el factor y la variable (matriz de configuración), ya que los factores están correlacionados entre sí.

Es por ello que cuando las correlaciones son muy altas en la matriz de correlaciones entre los factores es conveniente emplear las rotaciones oblicuas, como este método que minimiza la suma productos de los cuadrados de los coeficientes de la matriz de configuración) y el **Método Promax** (Este método consiste en alterar los resultados obtenidos de una rotación ortogonal hasta poder crear una solución con cargas factoriales que sean lo más próximas a la estructura ideal que se obtiene elevando las cargas factoriales obtenidas a una

potencia entre 2 y 4 en una rotación ortogonal, esto porque entre mayor sea la potencia más oblicua será la solución obtenida).

d) Cálculo de puntuaciones factoriales

Después de rotar los factores se calcula la *matriz de puntuaciones factoriales* con el fin de conocer, a) Los sujetos más extremos o atípicos a través de la representación gráfica de las puntuaciones factoriales para cada par de ejes factoriales. b) La ubicación de ciertos grupos en la muestra. c) En qué factores sobresalen o no algunos los sujetos y d) la interpretación, análisis o explicación de la información obtenida, y de los valores que toman los factores en cada observación. Existen muchos métodos para estimar la *matriz de puntuaciones factoriales* pero los factores estimados deben cumplir algunas propiedades que son deseables, como que cada factor estimado tenga una alta correlación con el verdadero factor, que tengan una nula correlación con los demás factores verdaderos, que estén incorrelacionados o mutuamente ortogonales si son ortogonales y que sean estimadores insesgados de los verdaderos factores. No obstante, el problema de la estimación es complejo factores comunes y en la mayoría de los escenarios, no hay una solución exacta ni única.

Los métodos de estimación de las puntaciones factoriales más empleados son el **método de regresión** que estima la *matriz de puntuaciones* mediante el método de los mínimos cuadrados y da lugar a puntuaciones con una máxima correlación con las puntuaciones teóricas. Pero el estimador no es insesgado, ni unívoco y, en el caso de que los factores sean ortogonales, puede dar lugar a puntuaciones correladas, el **método de Barlett que** utiliza el método de los mínimos cuadrados generalizados para estimar dicha matriz que genera puntuaciones correladas con las puntuaciones teóricas, insesgadas y unívocas, pero cuando los factores son ortogonales, se da lugar a puntuaciones correladas y el **método de Anderson-Rubín** que estima la *matriz de puntuaciones factoriales* mediante el método de los mínimos cuadrados generalizados imponiendo la condición adicional en donde

las puntuaciones ortogonales obtenidas están correladas con las puntuaciones teóricas, pero el estimador no es insesgado ni es unívoco.

4.1.3.5 ANALISIS FACTORIAL BAYESIANO

El análisis Bayesiano combina las distribuciones a priori de los parámetros con la probabilidad de los datos para formar distribuciones a posteriori de los parámetros estimados. Las a priori puede ser difusas (no informativa) o informativas, donde la información debe provenir de estudios previos. La parte posterior brinda una estimación en la forma de una media, la mediana o la moda de la distribución posterior. El enfoque Bayesiano también puede utilizar las cadenas de Markov con algoritmo Montecarlo - MCMC para extraer y obtener una distribución marginal posterior de cada cantidad estimada. La idea detrás del uso de las MCMC es que la distribución condicional de un conjunto de parámetros pueda ser usada para extraer aleatoriamente valores de parámetros, finalmente el resultando es una aproximación de la distribución conjunta de todos los parámetros.

Hay cuatro puntos clave que motivan a realizar un análisis Bayesiano y son los siguientes: Se puede aprender sobre las estimaciones de parámetros y modelos de ajuste que se ilustra mediante estimaciones de los parámetros que no tienen una distribución normal. El mejor rendimiento se puede obtener de muestras pequeñas y en las muestras grandes teóricamente no es necesario. Los análisis pueden ser menos exigentes computacionalmente y se obtienen estimaciones altamente confiables. Los nuevos tipos de modelos pueden ser analizados, ya que se pueden estudiar modelos con un gran número de parámetros ó donde la máxima verosimilitud no proporcione una aproximación natural.

Este tipo de modelo de análisis factorial es usado cuando se quiere identificar el número y naturaleza de los factores fundamentales responsables de la covariaciones en las p variables observables. Según Quinn (2004), en muchos casos, algunos las puntuaciones observadas de las variables son continuas, mientras que en otras son ordinales, cuando se presenta estas situaciones, los investigadores a menudo ven dentro de sus opciones de análisis de datos las

siguientes apreciaciones: **(a)** Hacer un tratamiento de las variables ordinales como continúas y usar un modelo de análisis factorial normal con teoría estándar, **(b)** Discretizar las variables continuas y usar un modelo de ítem respuesta**, (c)** Descartar las variables ordinales (continuas) observadas y trabajar sólo con variables continuas (ordinales), **(d)** Renunciar a una estrategia de medición basada en el modelo completo. Pero cada uno de estos movimientos resulta en consecuencias negativas (posiblemente graves) para la correcta inferencia. Cuando se hace caso omiso de la naturaleza ordinal de algunas de las variables observadas se puede dar lugar a estimaciones falsamente precisas y posiblemente sesgadas. Con la discretización de las variables continuas, se deshace la información y se disminuye en la precisión de las estimaciones. Descartar algunas variables de manera similar reduce la precisión de las estimaciones y puede hacen que sea imposible de descubrir la estructura compleja y multidimensional de los datos observados y por último, los métodos no basados en los modelos de medición no puede dar cuenta de los errores de medición y no permitirá a los investigadores evaluar la incertidumbre en las medidas resultantes.

Ante la situación anterior, Quinn (2004), se enfoca en estos problemas mediante la formulación de un modelo de medición que sea apropiado para respuestas multivariantes que tienen a lgunas componentes continuas, algunas ordinales o para combinaciones continuas y ordinales, permitiendo una sencilla interpretación de los parámetros del modelo. El autor toma un enfoque bayesiano para este modelo formulado y usa cadenas de Markov con algoritmo Monte Carlo (MCMC) para ajustar el modelo, lo que tiene el beneficio adicional de que las cantidades de interés a posterior (tal como, la probabilidad de que la unidad i tenga un mayor valor de constructo latente con relación a la unidad j) sean fáciles de calcular.

a) Modelo de análisis factorial para datos mixtos

Como en el análisis factorial estándar y en la teoría de ítem de respuesta, el objetivo del modelo es de captar los patrones de asociación entre las variables

observadas a través de un modelo relativamente parsimonioso, este tipo de modelo puede interpretarse también como variable latentes en el caso de que los patrones de asociación observados surjan de una o unas variables inobservadas (es decir, latentes). Visto desde esta perspectiva, las variables de respuesta observadas son indicadores imperfectos de la(s) variable(s).

X Es una matriz Nxj de respuestas observadas y cada variable observada puede ser ordinal o continua. Si la variable $j-esima$ es ordinal tendrá $\mathrm{C_j} > 1$ categorías. Se asume que los valores de los elementos de X son determinados por un matriz $\boldsymbol{X}^*$ de dimensión Nxj de variables latentes y una colección de puntos de corte γ entonces:

$$x_{ij} = \begin{cases} \mathrm{X_{ij}^*}, & si\ la\ Variable\ j\ es\ continua \\ c, & si\ \mathrm{X_{ij}^*}\ \epsilon\ \gamma_{j(c-1)}; \gamma_{j(c)}\ y\ la\ Variable\ j\ es\ ordinal \end{cases}$$

Donde:

$\mathrm{j} = 1, \ldots, \mathrm{J}$: Son los índices de las variables de respuesta

$\mathrm{i} = 1, \ldots, \mathrm{N}$: Son los índices de las observaciones

c : Puede tomar valores de $\{1, 2, \ldots, \mathrm{C_j}\}$

$$X_{NxJ}$$

Para identificar el modelo, se hace supuesto estándar de que $\gamma_{\mathrm{j0}} \equiv -\infty$; $\gamma_{\mathrm{j1}} \equiv 0$ y $\gamma_{\mathrm{jcj}} \equiv \infty$ para todo j. Los puntos de corte restantes son parámetros libres a estimar.

Los patrones de asociación entre las variables observadas en **X** son modeladas a través del factor analítico del modelo para la latente **X***:

$$\mathrm{x_i^*} = \Lambda\, \phi_\mathrm{i} + \varepsilon_\mathrm{i} \qquad \mathrm{i} = 1,2, \ldots \ldots, \mathrm{N}$$

Donde:

$\mathrm{x_i^*}$: Es el vector específico J de respuestas latentes para la i-esima observación.

Λ : Es una matriz j x k de factor de cargas

ϕ_i: Es un vector específico k de puntuaciones de los factores para la i-esima observación.

ε_i : Es vector J de perturbaciones que se $\sim \text{iid}\, N(0, \psi)$, ψ es diagonal.

Como se usan MCMC para ajustar el modelo, este se cambia ligeramente para acomodar la distribución condicional completa no *estándar* de los elementos libres de ψ. También será conveniente en algunos puntos de abajo escribir el sistema de N ecuaciones que se suministro en la ecuación anterior ($X_i^* = \Lambda \phi_i + \varepsilon_i$), como:

$$\boldsymbol{X}^* = \Phi\Lambda^t + E$$

Donde:

Φ : Es una matriz Nxk con i filas iguales a ϕ_i

E: Es una matriz Nxj con i filas iguales para ε_i

El primero elementos de ϕ_i es un set igual a 1 para todo i, esto asegura que los elementos en la primera columna de Λ funcionen como un ítems negativo para los parámetros de dificultan de las variables de respuesta ordinal. Los elementos de la primera columna de Λ que corresponden a las respuestas continuas representan la media de estas variables continuas. Los elementos de la primera columna de Λ que corresponden a las variables continuas que han sido normalizados para tener media 0 y deberán limitarse a 0. La estandarización de las variables continuas no es necesaria en este contexto, pero ayuda en la interpretación de los elementos de Λ que corresponden a las variables continúas y que pueden ser interpretados como factor de cargas.

b) Especificación a priori del modelo completo

Para identificar el modelo, algunos elementos de Λ puede ser restringidos a constantes, otros se verán restringidos a tomar sólo valores positivos (negativos), esto eliminaría la llamada invariancia de rotacional. En general, para identificar el k-1 factor del modelo se debe permutar las filas de Λ de manera que:

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_{1,1} & \lambda_{1,2} & 0 & & \dots & 0 \\ \lambda_{2,1} & \lambda_{2,2} & \lambda_{2,3} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & & & 0 \\ \lambda_{K,1} & \dots & & & & \lambda_{K,K} \\ \vdots & & & & & \vdots \\ \lambda_{J,1} & \dots & & & & \lambda_{J,K} \end{bmatrix}.$$

Al menos un elemento en cada columna de Λ debe estar restringido a tomar sólo valores positivos o negativos y se asume una densidad normal a priori recortada por debajo (o por encima) de 0, el resto de elementos de Λ siguen distribuciones normales independientes. Para ambos elementos normales atípicos a prior de Λ_{jk} recortados y no recortados, se asume que el parámetro media antes del recorte es $\mathrm{l}_{0\mathrm{jk}}$ y el parámetro precisión (varianza inversa) es $\mathrm{L}_{0\mathrm{jk}}$. Todos los elementos de Λ son asumidos como a priori independientes. Como es habitual en la literatura de la teoría de ítem de respuesta, los elementos de la diagonal de $\Psi = (\mathrm{diag}\{\Psi_1^2, \Psi_2^2, \dots., \Psi_{\mathrm{p}}^2\})$ que corresponden a las variables de respuesta ordinal están obligadas a ser igual a 1 por motivos de identificación y los elementos Ψ que corresponden a variables de respuesta continuas son dadas a priori como una distribución gamma inversa independiente, más específicamente si la variable j es continua $\Psi_{\mathrm{jj}} \sim \mathrm{I}\Gamma\left(\mathrm{a}_{0\mathrm{j}}/2, \mathrm{ab}_{0\mathrm{j}}/2\right)$.

c) Modelo de ajuste e inferencia

Con el software R para el análisis de datos y gráficos, en el paquete MCMCpack ese puede encontrar las funciones para el diagnostico y resúmenes estadísticos en MCMC, el modelo que sigue el paquete se escribe en términos de la latente **X***, es útil no sólo para construir una intuición, sino también para el ajustar del modelo. Basándose en las ideas del autor, se elige **X*** para tratar como datos latentes y se trabajar con la densidad posteriori.

$$p\left(\mathrm{X}^*, \gamma, \Lambda, \phi, (\Psi|\mathrm{X}^*)\right) \propto p\left((\Psi|\mathrm{X}^*), \phi, \Psi\right) p(\gamma) p(\Lambda) p(\Phi) p(\Psi)$$

$$\propto \left\{ \prod_{i=1}^{N} \prod_{j=1}^{J} \left\{ \|(\mathrm{X}_{ij} = \mathrm{X}_{ij}^{*}) \|(\mathrm{X}_j \ \ es\ continua) \right.\right.$$

$$\left. + \sum_{c=1}^{c_j} \|(\mathrm{X}_{ij} = c)\| \left\| \left(\mathrm{X}_{ij}^{*} \in (\gamma_{j(c-1)}, \gamma_{j(c)})\right) \right\| (\mathrm{X}_j \ \ es\ ordinal) \right\}$$

$$\left. \times \ p_N\left((\mathrm{X}_{\mathrm{i}}^{*} | \Lambda \phi_i), \Psi\right) \right\} p(\Lambda) p(\Phi) p(\Psi)$$

Donde:

$\|(a)$: Es la función del indicador, es igual a 1 si (a) es verdadero ó 0 en caso contrario.

$pN(\mathrm{z}|\mu, \Sigma)$: Es una densidad normal multivariada con media μ y matriz de varianzas-covarianzas Σ , evaluada en **z**.

$p(\Lambda)p(\Phi)p(\Psi)$: Son las densidades a priori de Λ, Φ Ψ respectivamente.

El a prior para los puntos γ es una constante para todos los valores de γ.

La ventaja de trabajar con la parte a posteriori aumentada en la ecuación anterior, es que es relativamente fácil de obtener distribuciones condicionales completas para la mayoría de los parámetros del modelo a partir de esta densidad posterior. Esto admite una muy simple cadena de Markov con algoritmo Monte Carlo (MCMC) que se utilizará para el ajuste del modelo, además, el algoritmo utilizado aquí funciona mediante un muestreador de las distribuciones condicionales completas de $\mathrm{X}^{*}, \Lambda, \phi$ y Ψ y γ llamado Metrópolis-Hastings

El modelo presentado anteriormente proporciona un medio de principios para seguir adelante con el modelo basado en la medición de este tipo de situaciones donde hay conceptos latentes que tienen tanto variables continuas como ordinales y es fácil de usar e interpretar. También es posible extender el modelo presentado anteriormente en varias direcciones: 1) permite medir las correlaciones del error de forma bastante simple. Es requerido un Ψ a priori diferente y la simulación de Ψ determina otros parámetros del modelo que también cambiarán, pero el resto del

algoritmo MCMC se mantiene intacto. 2) es posible permitir una gama más amplia de tipos de respuesta. Cuenta con Distribución de Poissón, variables censuradas y / o variables amputadas, distribuciones con variables continuas exponenciales y / o gamma son todas posible de acomodar en este marco general de modelización y 3) es posible extender el modelo para permitir la dependencia temporal, espacial o espacio-temporal. Esto es especialmente importante para aplicaciones de comparación y relación.

d) Diagnóstico convergencia de MCMC

El diagnostico de convergencia en una cadena MCMC se usa para saber si la se ha mezclado bien, para ello se utiliza un diagnostico estándar donde se evalúa si la muestra de la cadena proviene aproximadamente una distribución estacionaria. Las pruebas de convergencia más comunes son las propuestas por Gelman y Rubín (1992), Raftery y Lewis (1992) y el de Heidelberger and Welch (1983) que son en la actualidad, los más populares, por que se apoyan en los programas informáticos para su cálculo:

- El método de **Gelman y Rubín** se desarrolla en dos pasos, el primero se lleva a cabo antes del muestreo, para obtener una estimación mas dispersa de la distribución de destino y generar de ella, los puntos de partida para el número deseado de cadenas y el segundo Paso, se lleva a cabo para cada cantidad escalar de interés después de ejecutar las cadenas del muestreo de Gibbs para el número deseado de iteraciones y trata de utilizar las últimas 2n iteraciones para volver a calcular la distribución de destino de la cantidad escalar como un conservador de la distribución t de Student .

- El método de **Raftery y Lewis**, está propuesto para detectar la convergencia de la distribución estacionaria y para proporcionar una forma de delimitar las estimaciones varianza de los cuantiles de las funciones de los parámetros. primero se ejecuta una sola cadena del muestreo Gibbs para el número

mínimo de iteraciones que se necesitan para obtener la precisión deseada de la estimación si las muestras son independientes.

- El método de **Heidelberger y Welch,** es empleado para la detección de no estacionalidad, en la salida de la simulación con un enfoque de análisis espectral para la estimación de la varianza de la media de la muestra, a través de un procedimiento general para la generación de un intervalo de confianza previamente especificado para la media cuando la simulación no ha empezado en esta distribución estacionaria y procedimiento debe aplicarse a una sola cadena. Aunque su enfoque es anterior al muestro de Gibbs, es aplicable a la salida del muestreo Gibbs y otros algoritmos MCMC.

4.2 ASPECTOS AGRONOMICOS

4.2.1 CALIDAD DEL CAFÉ

La calidad del café depende de numerosos factores como la especie vegetal que se utilice (Robusta o Arábica), de la variedad de café sembrada, de los factores genéticos, depende obviamente del árbol y el entorno en que crece y el manejo agronómico del cultivo.

4.2.1.1 Componentes De La Calidad Del Café

La calidad del café de Colombia es el resultado de muchos procesos y operaciones realizados por todas las personas de la cadena del café que realizan de una manera optima, las labores de producción y los procesos. Cuando se obtiene un café de buena calidad, este debe presentar al mismo tiempo inocuidad, buena calidad física, buena calidad sensorial y una composición química natural.

4.2.1.2 La inocuidad del café

Esta característica garantiza que tanto los frutos y los granos como la bebida, no deben contener sustancias de origen químico o microbiano en cantidades que puedan causar daño a la salud como los residuos de plaguicidas, combustibles, mohos, humos y micotoxinas que además de originar defectos en el grano inciden en el aroma y sabor.

4.2.1.3 La calidad física

Cuando se presenta una buena calidad física, se puede observar mediante la apariencia homogénea de los granos de café, presencia de un olor característico a café, un color amarillo del pergamino y verde oliva del café almendra, tamaño según las especificaciones de mercado y un contenido de humedad entre el 10 y el 12%.

Por el contrario cuando los granos presentan defectos como guayabas, negros, severamente dañados por la broca, mohoso, vinagres y, los granos no despulpados o muy decolorados no son aceptables, ya que una vez son procesados, producen en la bebida sabores como fermento, *stinker*, fenólico, terroso, mohoso, acre, reposo o contaminado no muy deseables para el consumidor.

4.2.1.4 La calidad sensorial del café

Permite inferir las condiciones bajo las cuales se cuido el café, desde su cultivo hasta la obtención de la bebida. Un café de buena calidad sensorial puede presentar un balance en las características sensoriales y es imprescindible que no presente aromas ni sabores extraños que muestren un deterioro del producto o una contaminación, como los defectos acre, vinagre, fermento, stinker, reposo, fenólico, terroso, químico y ahumado entre otros.

Las características organolépticas o sensoriales del café se sienten en las papilas gustativas, y al activar los terminales del sentido del tacto se perciben la astringencia, el cuerpo y si la bebida está caliente o fría, ellas son: **La fragancia** (Esta es la característica es el aroma del café tostado y molido, es la primera cualidad que el catador percibe en la bebida del café al oler la muestra. Su intensidad, cualidad y tipo indican la calidad del café, frescura, condiciones de cultivo, beneficio y procesos para la obtención del producto**,** un buen café presenta un aroma, dulce, con notas herbales, frutales o a especies), **la acidez** (Es el sabor y aroma característico de ciertos ácidos como el acético, o de frutas cítricas como el limón o la naranja. Esta es una cualidad propia y muy positiva que presenta la especie Coffea Arábica L. cuando se ha beneficiado por la vía húmeda y la característica más apreciada en el café colombiano. La acidez resulta indeseable cuando se califica como agria, picante, acre, astringente o ausente, derivada de inadecuadas prácticas de cosecha y en el beneficio del café**, el amargor** (Es una característica normal del café debida a su composición química, la intensidad depende del grado de torrefacción y de la preparación de la bebida), **el dulzor** (Es una cualidad propia del café Arábica debida a su composición química y suavidad El sabor dulce que hace segregar una saliva espesa y viscosa, **el cuerpo de la bebida** (Las calificaciones de cuerpo muy alto, lleno, sucio o ligero, son indeseables en los cafés arábicos. Una buena bebida de café presenta cuerpo completo, moderado y balanceado) y **la impresión global** (Se refiere a la calificación general y clasificación de una bebida de café según su calidad. Debido a la impresión global, se acepta o rechaza la calidad de un café. Está relacionada con todas las propiedades percibidas con el sentido del olfato (aromas) y gusto (acidez, dulzor, cuerpo, amargo).

4.2.2 LA CATACIÓN

Es el método usado para conocer el aroma, el sabor y la sanidad del café, se denomina también evaluación sensorial de la calidad del café y prueba de taza. Por medio de esta técnica se pueden identificar los defectos presentes en la

bebida, medir la intensidad de una característica sensorial como la acidez y el dulzor, y también para calificar el sabor, el aroma y la calidad global del producto. Los métodos de la evaluación sensorial del café se usan básicamente en la industria de los alimentos para determinar la calidad de las materias primas y del producto, en el control de los procesos de fabricación y además para los estudios de preferencias en los consumidores.

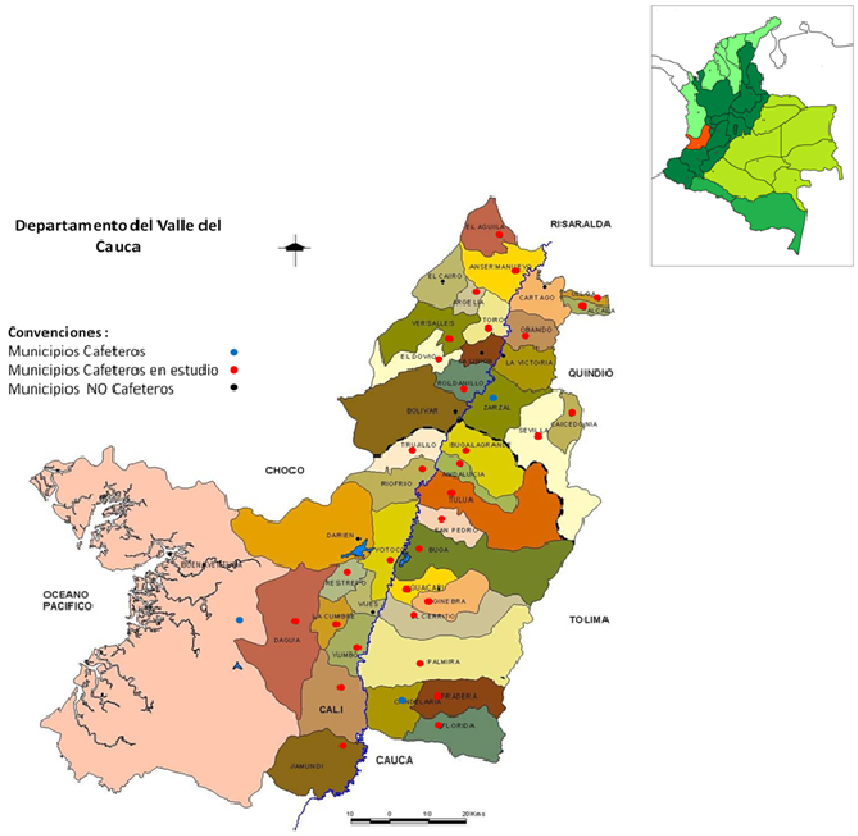
La aplicación de los métodos sensoriales puede resultar bastante compleja y costosa. Para los estudios sensoriales del café se requiere primero que todo de personal especializado y también se necesitan instalaciones adecuadas, métodos de análisis estandarizados, diseños experimentales, formularios para el registro de la información en forma escrita o digital, análisis estadísticos y experiencia para la interpretación de los resultados, de esta forma se reducen los errores inherentes a los juicios humanos. Las pruebas sensoriales del café permiten determinar la influencia de los diversos factores y condiciones de procesamiento en las características de calidad del producto, y por lo tanto, conocer los aspectos que conforman la calidad de una buena taza de café.

5. DISEÑO METODOLOGICO

5.1 UBICACIÓN GEOGRÁFICA DEL ESTUDIO

La base de datos fue proporcionada por el comité de cafeteros del Valle del Cauca y fue elaborada con 900 muestras enviadas por los caficultores de 32 de 39 municipios cafeteros del departamento del Valle del Cauca, durante el año 2009-2010, las cuales llegan identificadas con los datos de la finca y el proceso hecho al café.

Grafica 1. Mapa del Valle del Cauca con sus municipios cafeteros incluidos en el estudio



Fuente: IGAC-CVC-FEDECAFE

5.2 VARIABLES DEL ESTUDIO

Las Variable principales definidas por los investigadores (Coordinador Catación y la Coordinadora del Programa de Cafés Especiales) son las siguientes:

Análisis Sensorial

A continuación se muestran las variables organolépticas del café obtenías del formato de Análisis de Café, sobre las cuales desarrollo el estudio.

Tabla 2. Variables del análisis sensorial de las pruebas de taza que son resumidas en el puntaje final:

	Descripción y Calificación
Fragancia	Olor del café en seco. 0 a 10 (10 es el punto máximo)
Aroma	Olor del café una vez se adiciona el agua. 0 a 10 (10 es el punto máximo)
Sabor	Impresión combinada de todas las sensaciones gustativas. 0 a 10 (10 es el punto máximo)
Sabor Residual	Sabor que queda en la boca después de probado el café. . 0 a 10 (10 es el punto máximo)
Acidez	Contribuye a la vivacidad del café, dulzor y al carácter de fruta fresca. . 0 a 10 (10 es el punto máximo)
Cuerpo	Sensación táctil del líquido en la boca. . 0 a 10 (10 es el punto máximo)
Uniformidad	Igualdad de cada una de las tazas de la muestra2 puntos por taza. 2 puntos por taza
Balance	Equilibrio entre acidez, cuerpo, sabor y dulzor. 0 a 10 (10 es el punto máximo)
Taza Limpia	Es la transparencia de la bebida, sin defectos. 2 puntos por taza.
Dulzor	Plenitud agradable del sabor en la boca. 2 puntos por taza.
Puntaje catador	Impresión global del café. 0 a 10 (10 es el punto máximo)
Defectos	Sabores negativos. Se restan 2 puntos por taza si es ligero y 4 si es rechazo.
Puntaje Final	Es la suma de los atributos menos los defectos. Calificación Máxima 100 puntos

Según Manuel peña, Coordinador de catación del comité de cafeteros del Valle del Cauca, se realizan 5 pruebas de taza por cada muestra de café, en donde deben

exceptuarse los datos de uniformidad, taza limpia y dulzor, ya que estos 3 datos, muestran si es o no limpia la taza y si hay o no igualdad entre las pruebas y además aclara que respecto al dulzor solo se cualifica si está presente o no, ya que no está calificado por intensidad y calidad.

Análisis Físico de la Calidad del Café

Las variables del Análisis físico de calidad de Café del café que se muestran a continuación son adquiridas del formato de Análisis de Café:

Tabla 3. Variables de análisis físico de calidad de Café pruebas de taza del estudio.

Variable	Descripción y Calificación
Humedad en Café Pergamino Seco - CPS	Cantidad de agua presente en el CPS, importante para la calidad y la comercialización. 10 % - 12%.
Aspecto del CPS	El pergamino debe presentar uniformidad en color y debe estar libre de olores extraños o de cualquier tipo de contaminación. 1.Color normal 2. Olor fresco característico a olores distintos (como a reposo, moho, tierra, vinagre, productos derivados del petróleo, etc.)
Almendra sana	Almendras sin efectos físicos. Sin presencia de granos con pulpa e impurezas.75% - 100 %
Merma en trilla	Denominación que se le da a la pérdida del peso del grado de café mediante el proceso de descascarado para obtener el café excelso o almendra. 0% - 20 %.
Almendra defectuosa	1. Cardenillo, 2. Cristalizado, 3 .decolorado (blanqueado, ámbar o mantequillo), 4. Manchado, 5. Mordido, cortado, 6. Picado por insectos (afectado por la broca), 7 Partido, 8. Malformado o deformado, 9. Inmaduro, 10. Aplastado, 11. Flotador o balsudo, 12 Flojo, 12.Negro balsudo, 14. Vano, 15. Astillado y partido, 16. Reposo, 17. Sucio, 18. Sabor fenólico, 19. Materias extrañas. 0% - 5%.
Color Almendra	Característico, cutícula ligeramente rojiza, cutícula rojiza, decolorada, decolorada veteada, dispareja y sobresecada.
Olor Almendra	Normal, agrio, cigarrillo, fermento, moho, tierra vinagre

Variables Espaciales y Agronómicas

Las variables espaciales y de agronomía del cultivo del café en campo y poscosecha son las que se muestran en el cuadro que se presenta a continuación:

Tabla 4. Variables Espaciales y Agronómicas del estudio.

Variable	Descripción	Calificación
Municipio	Municipios objeto de estudio	Cali, Alcalá, Andalucía, Ansermanuevo, Argelia, Bolívar, Buga, Bugalagrande, Caicedonia, Dagua, El Águila, El Cerrito, El Dovio ,Florida, Ginebra, Guacarí, Jamundí, La Cumbre, Obando, Palmira, Pradera, Restrepo, Riofrío, Roldanillo, San Pedro, Sevilla, Toro, Trujillo, Tulúa, Ulloa, Versalles, Yótoco y Yumbo
Variedad	variedades de café analizadas en las pruebas de taza	Castillo, Colombia, Catimor, Caturra, Típica, Tabí, Costa Rica
Altura (msnm)	Altura de siembra del cultivo sobre el nivel del mar	1000 msnm - 2000 msnm
Fertilización	Fertilización empleada en el cafetal	Convencional, Natural, Orgánica, ninguna
Sombrío	Tipo de exposición del café al sol	Sombrío, al sol
Beneficio	Tipo de beneficio empleado	Desmucilaginado, fermentación natural, Desmucilaginado y fermentación natural.
secado	Secado del café	Al sol, mecánico, Al sol y mecánico

5.3 OBTENCIÓN DE LA INFORMACIÓN

La Federación Nacional de Cafeteros de Colombia - FNC diseño 3 formatos: el formato de etiqueta para identificar la muestra, el formato para entregar el resultado al productor, y el formato de campo o de trabajo en laboratorio para registrar los resultados físicos. En el caso del análisis sensorial o de taza, se trabaja con un formato diseñado por la Asociación de Cafés Especiales de América – SCAA, quienes diseñaron también la metodología de catación con la que se realiza el trabajo en el Comité.

Este es un formato muy usado, el cual se maneja tanto en países productores como en los consumidores, que desean describir un café especial. Es decir, no es

un formato para catar todo el café, solo en los casos en que se quiera calificar un café especial por taza y su escala es de 6 a 10 para cada atributo y hasta 100 en puntaje final.

5.4 ANÁLISIS ESTADÍSTICO

Ante la presencia de la información faltante en algunas variables en la base de datos en estudio, se recurrieron a técnicas de imputación de datos que básicamente radican en completar la base de datos bajo cierto supuestos, ya sea eliminado los casos con datos faltantes ó reemplazando dichos datos por otros, tan similares a ellos como sea posible.

Para ello se realizaron tres técnica de imputación de datos, la Listwise - CD, Hot-Deck –HD e Imputación Múltiple- MI, esta última fue la técnica de referencia y se escogió debido a que según Rubín (1987) y otros autores toma en cuenta una cantidad considerable de valores perdidos y no informados (NA/NI) en algunas de las variable y además aquellas que presentan una elevada variabilidad de los valores imputados que según Schafer (1997), es conveniente tratar con esta técnica.

Cada una de las bases de datos completas obtenida de los diferentes métodos de imputación (Listwise o case deletion CD, Imputación Múltiple - MI, Hot Deck -HD), se analizó mediante métodos estadísticos tradicionales (valores de la media (Me), mediana (Med), desviación estándar (SD), coeficiente de asimetría (As),curtosis (K) y el contraste de normalidad de Kolmogorov-Smirnov-Lilliefors (KSL)) con el objetivo de evaluar las técnicas de imputación a través de la simulación y poder seleccionar la más adecuada. El conjunto de datos generado y los datos iniciales fueron analizados mediante el Análisis Factorial bayesiano propuesto por Quinn (2004), que soporta su teoría en la utilización de las cadenas de Markov con algoritmo Monte Carlo (MCMC) para estimar el índice de calidad objetivo central de este estudio. Posteriormente, se mostraran las zonas geográficas que presentan una mayor probabilidad de este índice de calidad. Se calcularon las

cargas factoriales y el porcentaje de varianza explicada, obtenidas para cuantificar la contribución del factor común (índice de calidad) a las características organolépticas y agronómicas. Además se verificó la convergencia de las MCMC mediante la prueba de Heidelberger and Welch (1983).

Todo el procesamiento de la información se hizo en el software R, a través de diversos paquetes estadísticos entre ellos los más destacados:

- STATS : Listwise o case deletion CD
- VIM :Hot Deck -HD
- MICE :Imputación Múltiple – MI
- MCMCpack: Análisis factorial bayesiano

Finalmente se analizaron y discutieron los resultados obtenidos a partir del análisis realizado.

6. RESULTADOS Y DISCUSION

6.1 ESTADÍSTICAS DESCRIPTIVAS DE LAS VARIABLES

A continuación se muestran las estadísticas descriptivas para las variables del estudio tanto cuantitativas como cualitativas que hicieron parte del estudio.

Tabla 5. Variables de los datos originales (sin imputación)

Variables	Etiquetas	N		Media	Mediana	Mínimo	Máximo
		Válidos	Faltantes				
Variedad	varied	692	299	10,06	12	1	15
Sombrío	sombr	649	342	9,30	5	1	27
Fertilización	fertiliz	753	238	2,93	|	1	3
ALTURA (Msnm)	msnm	529	462	1566,40	1580	1000	2000
Beneficio	benefic	748	243	2,13	2	1	4
Tipo De Secado	tsecado	761	230	3,83	4	1	5
Certificación	certific	991	0	6,68	6	1	9
Aspecto Almendra	aspect	988	3	35,88	38	1	46
Humedad Cps (%)	humcpsp	979	12	0,12	0,1176	0,075	0,3
Merma (%)	mermap	646	345	0,18	0,1856	0,1476	0,2196
Olor Almendra	coloralm	735	256	3,06	3	1	8
Color Almendra	oloralm	735	256	6,90	7	1	7
Almendra Sana (%)	almendefecp	646	345	0,77	0,7762	0,4704	0,8324
Almendra Defectuosa %	almsanap	646	345	0,05	0,0386	0,0016	0,3144
Clases De Defectos	Defec	991	0	7,82	9	1	15
Puntaje final	puntfin	991	0	2,31	2	1	4

Para las variables sin imputación que se muestran en la tabla 4. Las variables como altura, merma, almendra sana, almendra defectuosa y sombrío, mostraron el número más alto de valores perdidos en contraste con variables como certificación, aspecto de la almendra, humedad del CPS que presentaron los valores más bajos.

Con este conjunto de variables se realizaron 3 técnicas de imputación de datos Listwise o case deletion CD, Imputación Múltiple - MI, Hot Deck -HD. En la tabla 6 se presentan los estadísticos para las variables en estudio, en ellas están los valores de la media (Me), mediana (Med), desviación estándar (SD), coeficiente de asimetría (As), curtosis (K) y el contraste de normalidad de Kolmogorov-Smirnov-

Lilliefors (KSL) bajo las diferentes técnicas de imputación, y los cuales permitieron evaluar e las técnicas de imputación a través de la simulación.

Tabla 6. Variables imputadas bajo las diferentes técnicas

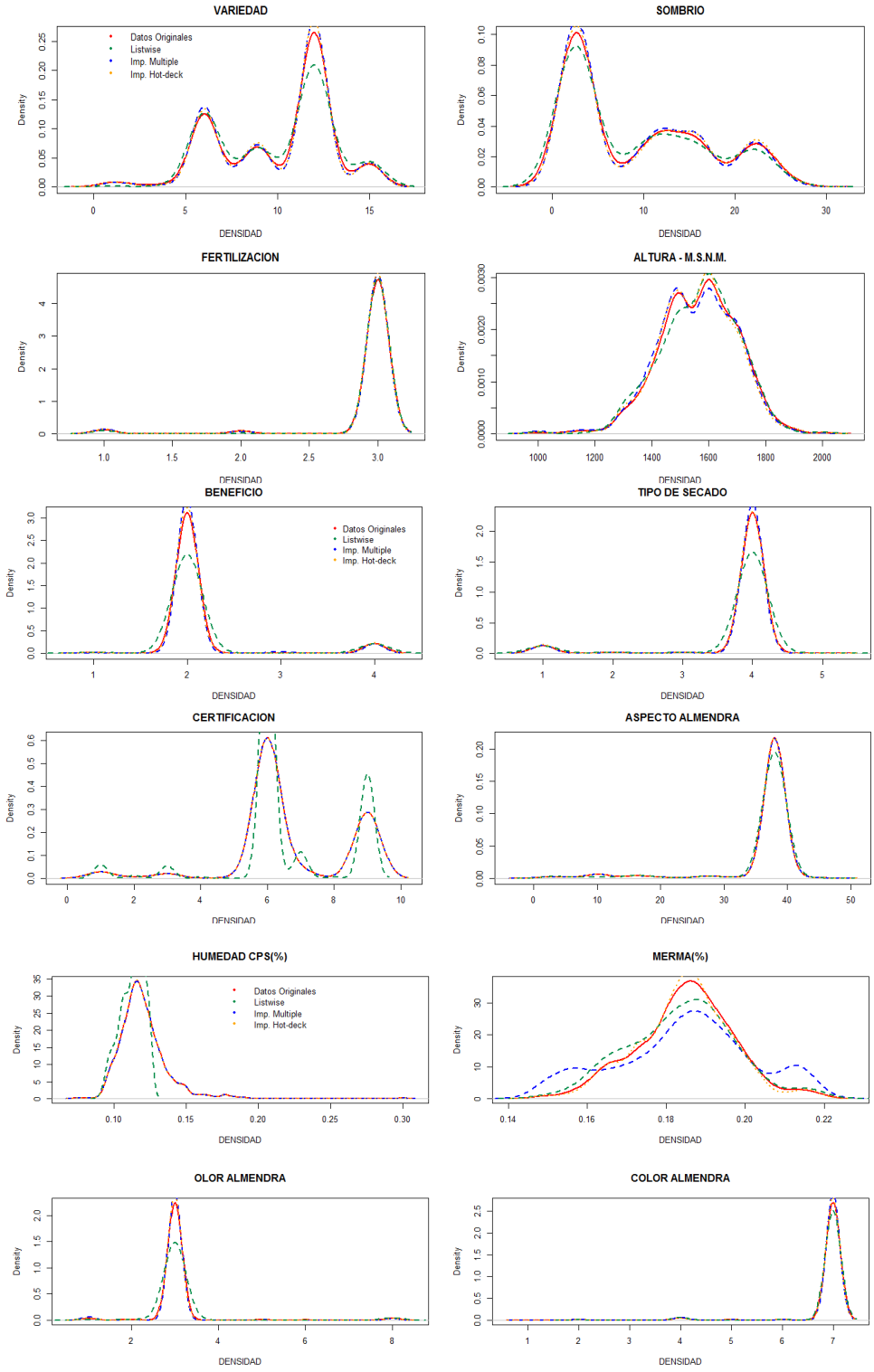
Variables	Método Imputación	N Válidos	Me	Med	SD	As	k	KSL (=0,05)	
								D	p- valor
Variedad	Imp. Mult	991	10,06	12	3,074	-0,570	-0,500	0,309	0,00
	Hot-Deck	991	10,05	12	3,072	-0,590	-0,416	0,317	0,00
	Listwise	306	10,07	12	3,038	-0,263	-1,078	0,284	0,00
Sombrío	Imp. Mult	991	9,30	5	7,512	0,605	-1,044	0,262	0,00
	Hot-Deck	991	9,54	5	7,694	0,563	-1,129	0,257	0,00
	Listwise	306	8,93	5	7,449	0,670	-0.967	0,267	0,00
Fertilización	Imp. Mult	991	2,93	3	0,346	-5,038	24,338	0,082	0,00
	Hot-Deck	991	2,93	3	0,341	-5,074	24,795	0,537	0,00
	Listwise	306	2,96	3	0,283	-6.604	42,238	0,537	0,00
ALTURA (Msnm)	Imp. Mult	991	1566	1580	136,113	-0,384	0,681	0,533	0,00
	Hot-Deck	991	1561	1570	132,621	-0,283	0,738	0,082	0,00
	Listwise	306	1567	1580	134,542	-0,332	0,513	0,086	0.022
Beneficio	Imp. Mult	991	2,13	2	0,474	3,528	11,164	0,536	0,00
	Hot-Deck	991	2,13	2	0,505	3,284	9,629	0,536	0,00
	Listwise	306	2,16	2	0,572	2,718	6,212	0,524	0,00
Tipo De Secado	Imp. Mult	991	3,83	4	0,671	-3,848	13,179	0,322	0,00
	Hot-Deck	991	3,82	4	0,705	-3,624	11,394	0,537	0,00
	Listwise	306	3,78	4	0,769	-3,187	8,500	0,530	0,00
Certificación	Imp. Mult	991	6,68	6	1,779	-0,588	1,411	0,513	0,00
	Hot-Deck	991	6,68	6	1,779	-0,588	1,411	0,322	0,00
	Listwise	306	6.52	6	1,795	-0,651	1,535	0,314	0,00
Aspecto Almendra	Imp. Mult	991	35,88	38	7,256	-3,268	9,602	0.567	0,00
	Hot-Deck	991	35,88	38	7,198	-3,295	9,809	0,514	0,00
	Listwise	306	36,24	38	6,543	-3,742	13,316	0,524	0,00
Humedad Cps	Imp. Mult	991	0,12	0,1176	0,016	2,327	16,772	0,118	0,00
	Hot-Deck	991	0,12	0,12	0,016	2,357	17,055	0,121	0,00
	Listwise	306	0,11	0,11	0,008	-0,496	-0,615	0,086	0.022
Merma	Imp. Mult	991	0,18	0,1856	0,017	-0,177	-0,512	0,057	0.003
	Hot-Deck	991	0,18	0,19	0,012	-0,077	0,326	0,047	0.027
	Listwise	306	0,18	0,18	0,013	0,012	-0,093	0,045	0.0047
Olor Almendra	Imp. Mult	991	3,06	3	0,675	4,613	37,002	0,494	0,00
	Hot-Deck	991	3,06	3	0,711	5,136	35,499	0,512	0,00
	Listwise	306	3,13	3	0,885	4,613	22, 95	0,523	0,00
Color Almendra	Imp. Mult	991	6,90	7	0,587	-6,223	41,130	0,535	0,00
	Hot-Deck	991	6,90	7	0,602	-6,272	41,488	0,536	0,00
	Listwise	306	6,91	7	0,535	0,418	40,055	0,535	0,00

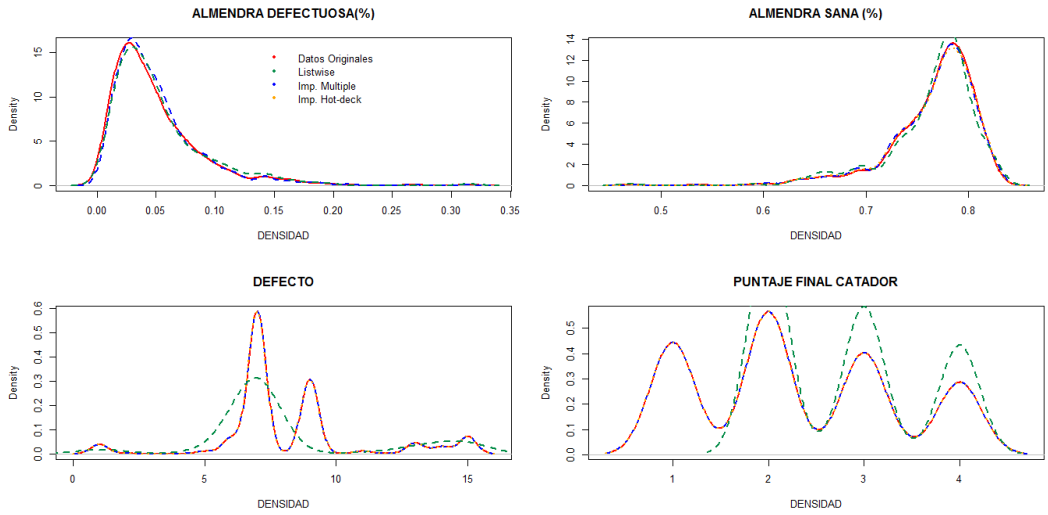
Almendra Sana	Imp. Mult	991	0,77	0,7762	0,043	-1,647	4,496	0,1190	0,00
	Hot-Deck	991	0,77	0,78	0,034	-1.649	4,419	0,125	0,00
	Listwise	306	0,76	0,78	0,045	-1,921	6,478	0,151	0,00
Almendra Defectuosa	Imp. Mult	991	0,05	0,0386	0,036	1,973	6,380	,131	0,00
	Hot-Deck	991	0,05	0,04	0.036	1,834	5,711	0,128	0,00
	Listwise	306	0,05	0,04	0,039	2,062	7,043	0,159	0,00
Clases De Defectos	Imp. Mult	991	7,82	9	2,783	0.664	1,626	0,264	0,00
	Hot-Deck	991	7,82	9	2,783	-0.664	1,626	0,213	0,00
	Listwise	306	8,03	9	3,197	0,913	0,787	0,429	0,00
Puntaje final	Imp. Mult	991	2,31	2	17,143	0,255	-0,529	0,288	0,00
	Hot-Deck	991	2,31	2	17,143	0.255	-0,529	0,288	0,00
	Listwise	306	2,78	2	17,061	0,418	-0,905	0,188	0,00

Una de las hipótesis básicas suele ser evaluar la normalidad de las variables y como se observa a través de las estadísticas de forma como lo son coeficientes de asimetría y curtosis, muestran que ninguna de las variables imputadas en este estudio sigue una distribución normal, lo que se corrobora con la prueba KSL que con un nivel de significancia de 0.05, en donde el P-valor de la prueba (la probabilidad de obtener un resultado al menos tan extremo como el que realmente se ha obtenido), fue inferior en todas las variables por lo que se rechaza la hipótesis de normalidad.

En la grafica 2, se presentan las distribuciones de probabilidad para las variables en estudio bajo las tres técnicas de imputación, como se puede observar en los gráficos de densidad, los datos generados por MI y por imputación HD se ajustaron bien a los datos originales lo que no sucedió con la técnica de CD.
Como se muestra gráficamente, ni las variables originales ni las imputadas muestran distribuciones normales, por lo cual se verifica que luego de realizar las imputaciones dichas distribuciones no han cambiado su distribución.

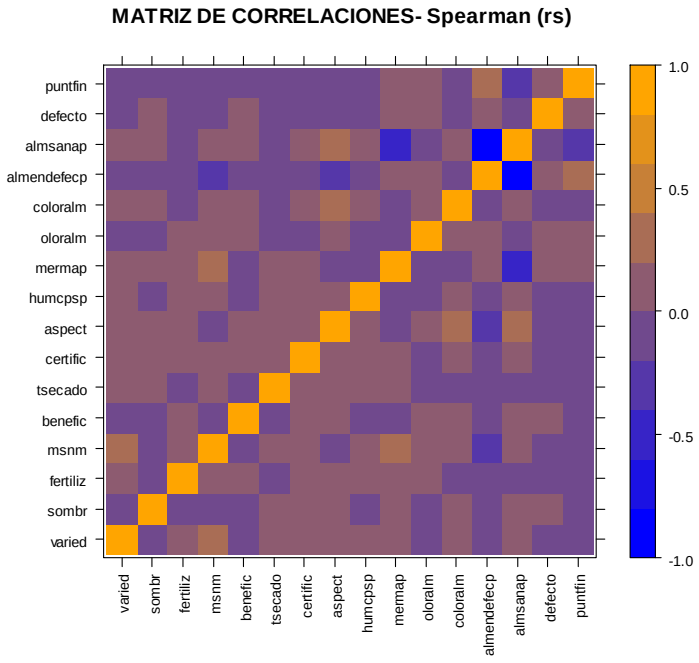
Grafica 2. Variables antes y después de los procesos de imputación.





Por último, antes de hacer análisis de los resultados bajo las diferentes técnicas, se calculó el coeficiente de correlación entre las variables lo que permite conocer las relaciones existentes entre las variables, pero debido a que no hay normalidad en las variables, no se pudo calcular coeficiente de correlación de *r Pearson,* por lo cual se utilizó el coeficiente de correlación no paramétrico de *rs Spearman*, el cual no requiere normalidad y además acepta variables de libre distribución e incluso ordinales .

Grafica 3. Relación entre las variables iniciales.



Los resultados que se presentan en la grafica 3, muestran los bajos valores del coeficiente de correlación de *rs Spearman*, lo que permite concluir que existe una relación entre las variables aunque sea muy débil, por lo que en este caso, es importante hacer un análisis de la relación entre las variables después de realizar las imputaciones mediante las diferentes técnicas.

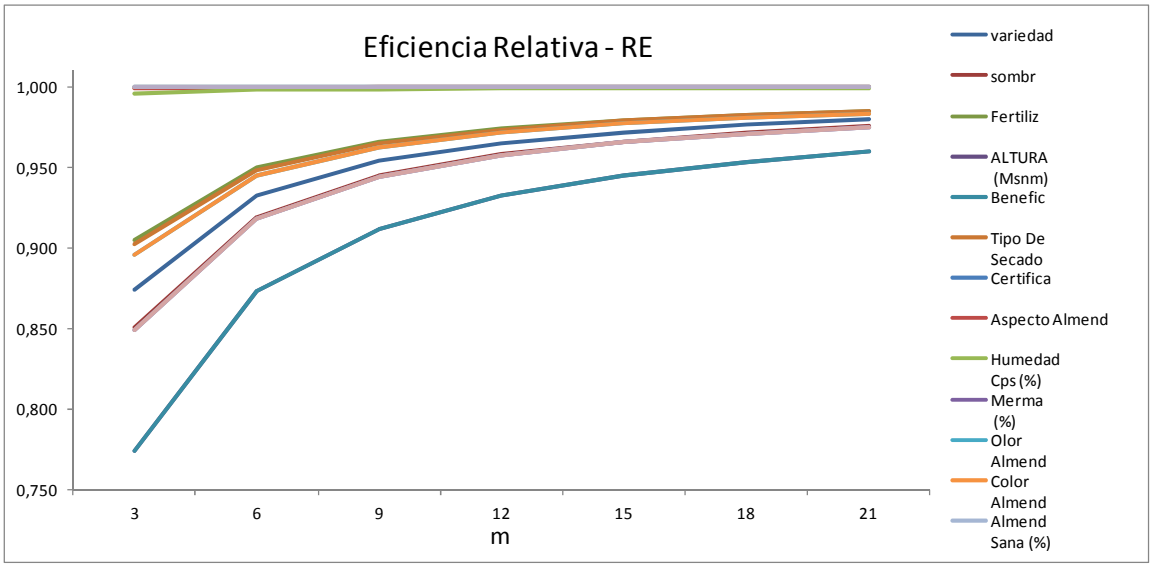
6.2 ANÁLISIS DE LOS RESULTADOS PARA VALORES FALTANTES

Según la revisión de literatura resumida en la tabla 6, los análisis de los estadísticos y los gráficos construidos, muestran que al comparar las técnicas de imputación, ninguna de ellas afectó significativamente ni la media, ni la mediana, ni siquiera omitiendo valores faltantes por la técnica de CD, una situación obvia debida que ambas son medidas de tendencia central, por lo cual son muy próximas. Con la técnica HD, en este sentido se hizo un buen trabajo eligiendo donantes que no afectaron estas medidas, debido a que se contó con un volumen de datos relativamente alto y por tanto hubo buena disponibilidad de donantes. Para la técnica de MI, sin embargo, detrás del cálculo de las estadísticas, se debe preguntar si el proceso de imputación con un número de m imputaciones genera inferencias estadísticas útiles, para lograrlo se realizó previamente el análisis de la eficiencia relativa de las estimaciones con relación al porcentaje de datos faltantes que se aprecia en la tabla 7 .Rubín (1987) ilustra que al usar entre 2 y 10 imputaciones se pierde poco eficiencia en la estimación (en relación con un número infinito de imputaciones) cuando la fracción de la información faltante es modesta y genera intervalos de confianza y pruebas de hipótesis cercanas a su cobertura nominal y niveles de significación, respectivamente. Sin embargo, Schafer (1997) muestra que cuando la fracción de la información que falta es muy grande, son necesarias más de 10 imputaciones. Con relación a lo anterior, fue ineludible realizar el análisis de la eficiencia relativa para saber el número m de imputaciones necesarias bajo el porcentaje de valores faltantes de las variables en estudio, donde se escogió un $m = 12$ imputaciones necesarias para obtener estimaciones confiables, como se muestra en la tabla 7 y la grafica 4.

Tabla 7. Eficiencia de la estimación a través de imputación múltiple según el número de imputaciones (m) y la fracción de información faltante.

	Variedad	Sombrío	Fertiliz	ALTURA (Msnm)	Benefic	Tipo De Secado	Certifica	Aspecto Almend	Humedad Cps (%)	Merma (%)	Olor Almend	Color Almend	Almend Sana (%)	Almend Defect %	Clases De Defectos	Puntaje final
m	Fraccion de la Informcion Faltante															
	0,432	0,527	0,316	0,873	0,325	0,302	0,000	0,003	0,012	0,534	0,348	0,348	0,534	0,534	0,000	0,000
3	0,874	0,851	0,905	0,775	0,902	0,908	1,000	0,999	0,996	0,849	0,896	0,896	0,849	0,849	1,000	1,000
6	0,933	0,919	0,950	0,873	0,949	0,952	1,000	0,999	0,998	0,918	0,945	0,945	0,918	0,918	1,000	1,000
9	0,954	0,945	0,966	0,912	0,965	0,968	1,000	1,000	0,999	0,944	0,963	0,963	0,944	0,944	1,000	1,000
12	0,965	0,958	0,974	0,932	0,974	0,975	1,000	1,000	0,999	0,957	0,972	0,972	0,957	0,957	1,000	1,000
15	0,972	0,966	0,979	0,945	0,979	0,980	1,000	1,000	0,999	0,966	0,977	0,977	0,966	0,966	1,000	1,000
18	0,977	0,972	0,983	0,954	0,982	0,983	1,000	1,000	0,999	0,971	0,981	0,981	0,971	0,971	1,000	1,000
21	0,980	0,976	0,985	0,960	0,985	0,986	1,000	1,000	0,999	0,975	0,984	0,984	0,975	0,975	1,000	1,000

Grafica 4. Eficiencia Relativa de la estimación a través de imputación múltiple según el número de imputaciones (m) y la fracción de información perdida ( ).

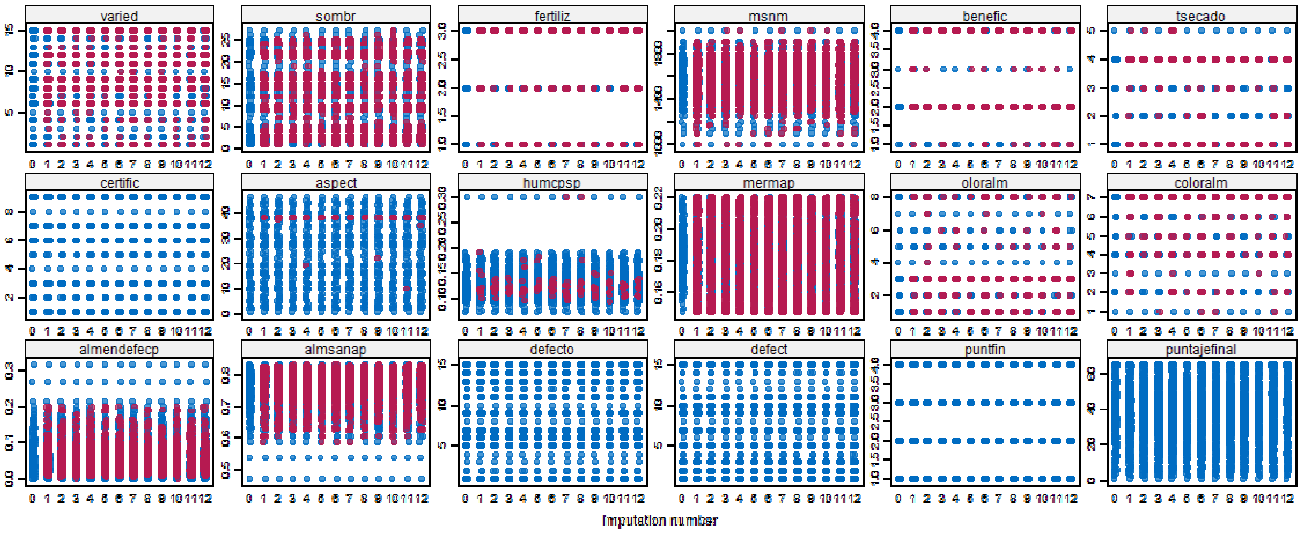


Según la tabla 7. Cuando $\lambda = 0.874$, por ejemplo, una estimación basada en $m = 3$ imputaciones, tenderá a tener un error estándar sólo en un $\sqrt{1 + \left(\frac{0,874}{3}\right)} = 1.136$ veces más grande que el estimado con $m = \infty$ ó con $\lambda = 0.012$, una

estimación basada en $m = 21$ imputaciones tienden a tener un error estándar de $\sqrt{1 + \left(\frac{0{,}012}{21}\right)} = 1.00$ veces más grande que el estimado con $m = \infty$.

Ahora, si se observa la grafica 4, se puede ver que a medida que el número de imputaciones aumenta, la eficiencia relativa también lo hace, por ello se escogió un tamaño m que tuviera una buena eficiencia sin exagerar el número de imputaciones. Con la técnica de imputación múltiple también se lograron estimar los valores que no afectaran ni la media, ni la mediana es por ello que para realizar la segunda parte del estudio y basándose en lo descrito por Rubín (1987, 1996), se consideró que este método es el más adecuado para tratar los problemas más complejos cuando hay un número considerable de datos en más de una variable, además como se requiere mantener la distribución de frecuencia de la variable y las asociaciones entre las distintas variables, se deben emplear métodos más elaborados aplicando imputación de todas las variables faltantes del registro como la MI. Goicoechea (2002).Para tener una idea más clara de lo acontecido con el proceso MI, en el grafico 5 que se muestra a continuación se representa la distribución de las variables como puntos individuales, los puntos azules son los valores observados y los puntos rojos son los valores.

Grafica 5. Variables antes y después del proceso de imputación múltiple MI.



Como se observa en el panel, para las variables en estudio como defecto y puntaje final y certificación, solo contienen valores azules, es decir que en los datos originales, estas variables tenían todos sus datos completos, sin embargo, en variables como merma, sombrío, almendra sana y altura, se puede observar que los valores más notorios fueron los imputados (color rojo), pero se puede observar que el método MI respondió bastante bien, ya que los puntos rojos siguen los puntos azules muy razonablemente, incluyendo las brechas en las distribuciones.

6.3 ANÁLISIS FACTORIAL BAYESIANO

Una vez ejecutado el proceso IM con $m = 12$ y se ha obtenido una base de datos completa, se realizó el análisis factorial bayesiano, empleando el paquete MCMCpack de R, a través de la función MCMCmixfactanal, en él se suministraron los datos y como resultado se obtuvo una cadena de Markov con algoritmo Monte Carlo – MCMC y que contiene una muestra de la densidad posterior de los datos mixtos (tanto continuos como ordinales) del modelo de análisis factorial.

El ajuste del modelo se llevo mediante lo descrito previamente y en donde las primeras 5000 exploraciones de las MCMC fueron descartadas como burn-in. Los resúmenes posteriori de los parámetros del modelo (factor de cargas (λ_2), varianzas del Error (ψ_{jj}) y factor latente de puntuaciones (ϕ_i)), fueron basados en una muestra posterior de tamaño 5000, formado mediante la ejecución de la cadena por un período adicional de 500000 exploraciones de donde se fueron almacenando cada 10 exploraciones. La cadena mixta generada, mostró un diagnóstico estándar razonablemente bueno.

```
Posteriori<- MCMCmixfactanal((~varied+sombr+fertiliz+msnm+benefic+
            tsecado+aspect+ humcpsp+mermap+oloralm+coloralm+
            almendefecp+almsanap+puntfin), factors=1,
            data=Datos.Imputados,burnin=5000,mcmc=500000,thin=10,
```

store.scores=TRUE,tune=0.25,store.lambda=TRUE)

Después, se verificó mediante la prueba de Heidelberger and Welch (1992) que la muestra de la cadena Markov generada, alcanzara la distribución estacionaria (Ver Anexo 1). Las variables se pueden observar en la función del modelo de ajuste, solo se eligió estimar un factor (Índice de calidad), el intervalo de precisión fue de 0.25, tanto para almacenar la muestra del factor de cargas (Λ), las varianzas del Error (Ψ) y el factor latente de puntuaciones (Φ) de la distribución posteriori. Básicamente lo que se pretendió, fue aventurarse en la estructura de los datos de un modelo construido con anticipación, para tratar de explicar las correlaciones entre el conjunto de variables observadas a través de un conjunto más reducido de factores latentes.

6.3.1 RESÚMENES DE LA DENSIDAD POSTERIORI

A continuación se muestran los resúmenes Λ y Ψ de la densidad posterior que proporciona información acerca de los patrones globales de la asociación entre las variables observadas del estudio y las estimaciones de las puntuaciones de los factores latentes Φ que finalmente son las que nos llevan al índice de calidad.

Factor de cargas (Λ) y Varianzas del Error (Ψ)

El factor loading o factor de cargas (Λ) permite observar las conexiones entre las variables y el factor, es decir, indican la fuerza de la dependencia de cada variable observada en cada factor.

Como se observa en la tabla 8, se muestran las variables que constituyen las propiedades organolépticas y agronómicas obtenidas en las diferentes pruebas de tasa, aquí, la estimación de la variable almendra defectuosa en el factor de cargas es de -1.00, lo que indica que presenta una asociación negativa muy alta entre esta variable y el factor, en contraste con las variables como merma y

almendra sana que presentaron una asociación positiva de 0.5529 y 0.9269 respectivamente, el resto de las variables mostraron valores muy bajos.

Tabla 8. Factor de cargas (λ_2) y del Error (ψ_{jj}

Variables	(λ_2)	(ψ_{jj})
Variedad	-0,0574	0,9975
Sombrío	-0,0625	0,9983
Fertilización	-0,0013	1,0015
ALTURA (Msnm)	-0,1794	0,9711
Beneficio	0,0253	1,0005
Tipo De Secado	0,0433	1,0008
Aspecto Almendra	-0,1019	0,9910
Humedad Cps	0,0087	1,0015
Merma	0,5529	0,6998
Olor Almendra	0,0553	0,9990
Color Almendra	0,0782	0,9953
Almendra Sana	0,9268	0,1509
Almendra Defectuosa	-1,0052	0,0007
Puntaje final	0,1312	0,9855

Con las varianzas del error se puede tener una idea de la varianza de la variable que puede ser explicada por el factor latente, es decir que para la almendra defectuosa, la varianza del error estimado se calcula en 0.0007, lo que implica que casi toda la variabilidad de la Almendra Defectuosa puede ser explicada solo por el factor latente, débilmente también se podrían tener en cuenta la merma, olor el color y la almendra sana.

Factor latente de puntuaciones (ϕ_i)

El factor latente de puntuaciones es una variable no observada que se compone de numerosas variables (observables), ella permite explicar la variabilidad común que tienen todas ellas a través de las puntuaciones, es decir, que es una forma de consolidar numerosa información en una sola variable. Por ello, la determinación de este factor fue el interés principal de este estudio, con el se obtuvo el índice de calidad que se puede resumir en la tabla 9 para una mejor interpretación.

Tabla 9. Tabla de frecuencias para el Índice de calidad – IC (ϕ_i)

Intervalo		Media	Frecuencia Absoluta	Frecuencia relativa	Frecuencia Absoluta Acumulada	Frecuencia Relativa Acumulada	Valoración IC Café
inferior	superior						
-2	-1	-1,5	88	0,089	88	0,089	Mala calidad
-1	0	-0,5	519	0,524	607	0,613	
0	1	0,5	265	0,267	872	0,88	media
1	2	1,5	67	0,068	939	0,948	Buena
2	3	2,5	34	0,034	973	0,982	
3	4	3,5	15	0,015	988	0,997	
4	5	4,5	1	0,001	989	0,998	excelente
5	6	5,5	1	0,001	990	0,999	
6	7	6,5	1	0,001	991	1	

En la tabla 9, se puede ver que el 61,3 % de los datos presenta puntuaciones entre [-2,0] lo que se ha valorado como un índice de calidad malo, el 26.7% de las puntuaciones entre [0,1] catalogado como calidad media, el 10,2% entre [1,3] como calidad buena y solo el 1.8% de las puntuaciones presentaron una excelente calidad con valores superiores a 3.

En los gráficos 6, 7, 8,9 y 10 que se presentan a continuación se puede contrastar el índice de calidad con otras variables de interés.

- **Por municipio**

Como se ve en la gráfica 6, hay una diferencia bien marcada entre las calidades del café, con respecto a las zonas, por ejemplo la zona norte del Departamento muestra una calidad excelente, la zona Centro una calidad buena y la zona sur una calidad media. Esto es de esperarse dado a que la zona norte (El Águila, Alcalá, el Dovio, Roldanillo, Bolívar, Trujillo, Bugalagrande y Sevilla) y centro del departamento (Riofrío, Tulúa, San Pedro, Buga, Guacarí, Ginebra y Buga) posee el entorno típico de la zona cafetera, es decir, que en cuanto a los aspectos agronómicos ideales del cultivo como suelos muy fértiles, temperaturas entre los

17 y los 23 grados centígrados, precipitación atmosférica cercanas a los 2.000 milímetros anuales y altura entre los 1.200 y 1.800 metros sobre el nivel del mar muestra que se pueden producir cafés de excelente calidad.

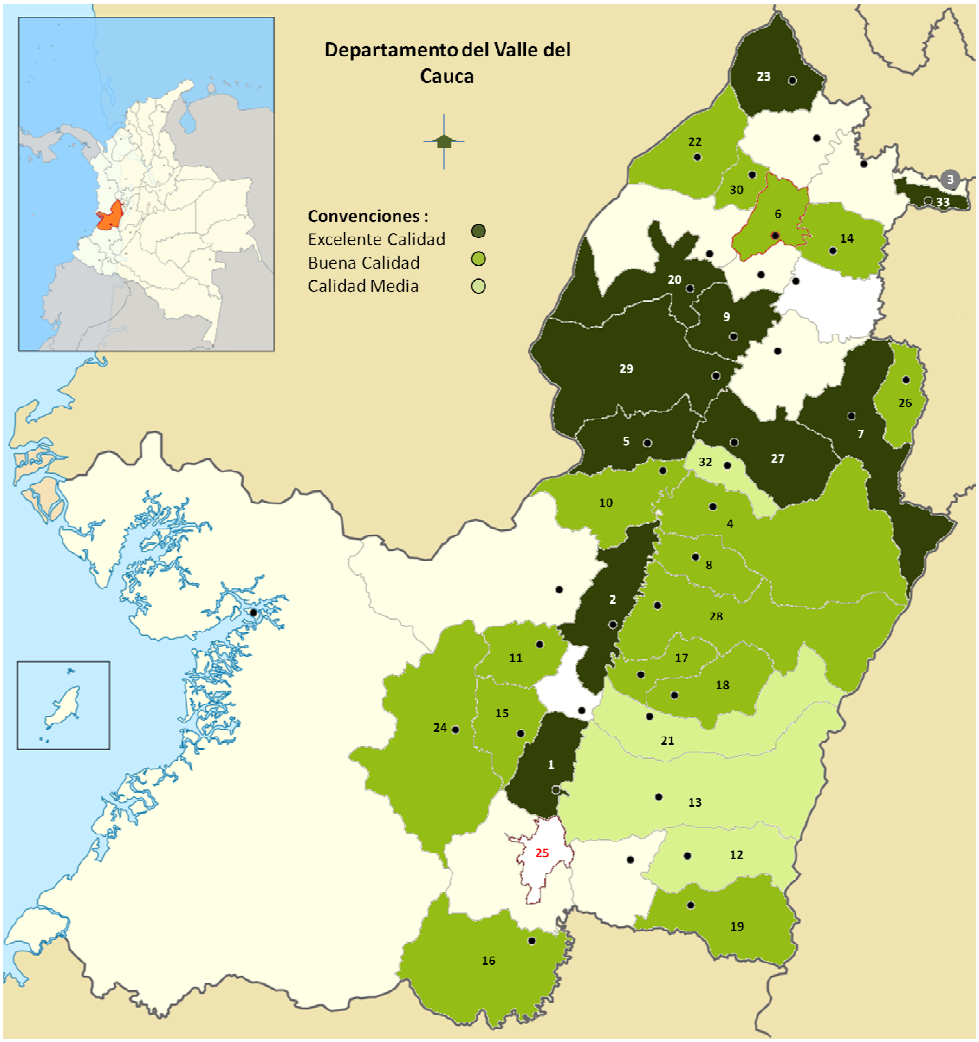
En cuando a la zona Sur del departamento, se muestra que también es posible producir un café con una calidad sobresaliente a alturas marginalmente y con niveles o frecuencia de precipitación diferentes. Esta zona pues, podría ser de especial interés, debido el alto potencial que presentan municipios como Pradera, Palmira y El Cerrito, además de municipios como Dagua, florida, la cumbre, Yótoco, Restrepo y Yumbo para producir un café con muy buena calidad.

Finalmente, cabe anotar que este mapa es una primera aproximación a la zonificación por pruebas de taza y en ningún caso es una verdad absoluta, sin embargo refleja una realidad que se puede corroborar con otros estudios realizados por Cenicafé.

Tabla 10. Municipios del Valle del Cauca que hacen parte del estudio.

#	Municipio	#	Municipio	#	Municipio	#	Municipio
1	Yumbo	10	Riofrío	19	Florida	28	Buga
2	Yótoco	11	Restrepo	20	El Dovio	29	Bolívar
3	Ulloa	12	Pradera	21	El Cerrito	30	Argelia
4	Tulúa	13	Palmira	22	El Cairo	31	Ansermanuevo
5	Trujillo	14	Obando	23	El Águila	32	Andalucía
6	Toro	15	La Cumbre	24	Dagua	33	Alcalá
7	Sevilla	16	Jamundí	25	Cali	34	No Sica
8	San Pedro	17	Guacarí	26	Caicedonia		
9	Roldanillo	18	Ginebra	27	Bugalagrande		

Grafica 6. Mapa del Departamento del Valle del Cauca que Presenta una media, buena y Excelente calidad.

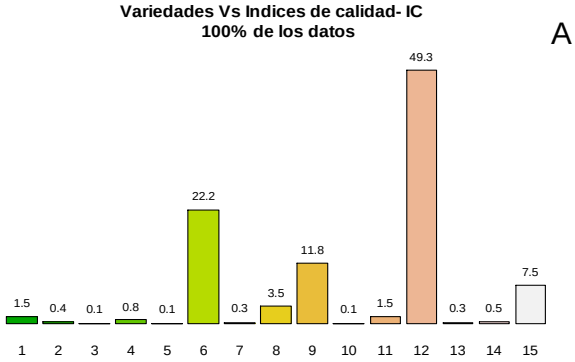


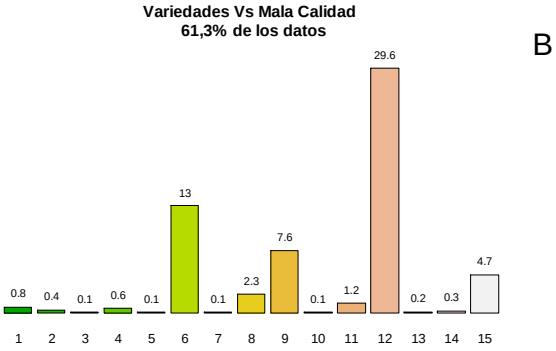
- **Por Variedad**

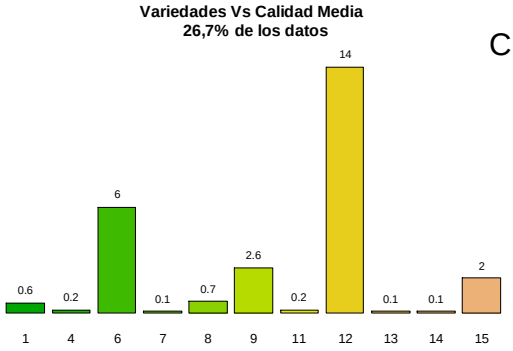
En el grafico 7A se observa que la variedad predomínate en el estudio es la Caturra con el 49.3%, seguido Colombia 22.2% y la castillo con 7.5%, además se muestra la combinación de estas variedades dentro del cultivo como la Caturra y la Colombia con un 11.8% y Caturra y típica 3.5%.

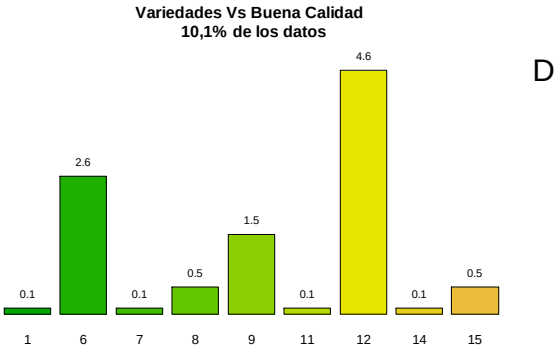
Grafica 7. Variedades de café según el Índice de calidad.

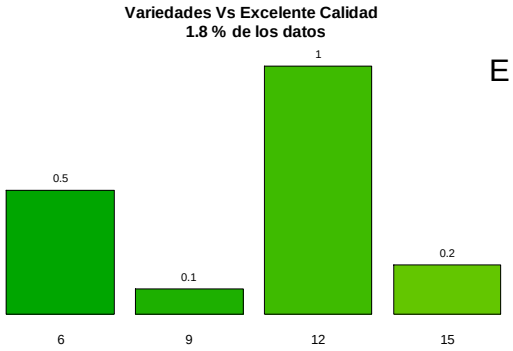
#	Variedad	#	Variedad
1	Típica	9	Caturra y Colombia
2	San Bernardo	10	Caturra y Catimor
3	Costarica	11	Caturra y Castillo
4	Colombia, Caturra y Típica	12	Caturra
5	Colombia y Típica	13	Catimor
6	Colombia	14	Castillo y Colombia
7	Caturra,Colombia y tabi	15	Castillo
8	Caturra y Típica		











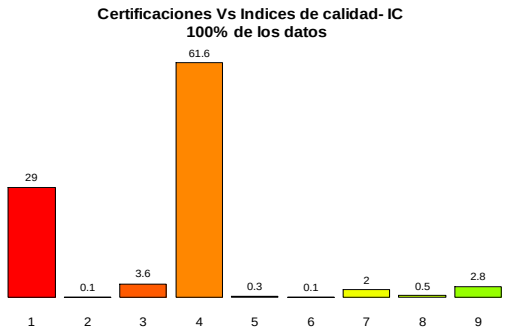
Ahora si se observa el grafico 7E, que muestra una excelente calidad según el índice hallado, se puede ver que siguen estando presentes estas cuatro variedades, así como 7B, 7C y 7D. Estos valores concuerda con estudios realizados por Cenicafé donde la variedad Colombia, caturra, típica y castillo se destacan por su calidad en taza con relación a otras variedades(Puerta, 1998 , 2000, Alvarado, 2002)

- **Por Certificación**

La certificación del café son signos que le aseguran al consumidor que el café cumple con unos estándares de calidad previamente definidos en normas legales, técnicas o reglamentos.

Grafica 8. Certificación de café según el Índice de calidad.

#	Certificación	#	Certificación
1	4C	6	Orgánico y Rainforest
2	4C y UTZ	7	Rainforest
3	FLO	8	Rainforest y 4C
4	Ninguna	9	UTZ
5	Orgánico		



Según la gráfica 8, en el estudio el 61.6% de las pruebas de taza, su café no provenía con ningún tipo de certificación, el 29% 4C(cafés procedentes de procesos productivos con buenas prácticas sociales, económicas y ambientales), el 3.6% FLO (certificación de Comercio Justo que permite recibir un precio de venta mayor y más estable donde el proceso productivo del café protege el medio ambiente), 2% Rainforest(es una certificación orientado a la protección del medio ambiente que se enfoca en el manejo sostenible de las fincas) y el 2.8% UTZ (esta certificación es para cafés de mejor calidad y producido bajo estrictas normas de calidad).

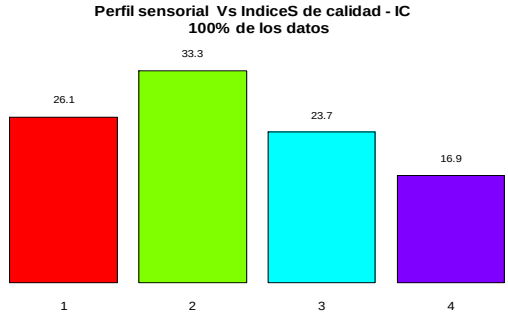
- **Por Puntaje sensorial**

La gráfica que se presenta a continuación, muestra el puntaje sensorial del café, aquí, el 33.3% presentaron muy buen perfil, el otro gran porcentaje lo contienen las muestras que no son Aptas 26.1% lo que quiere decir que presentaron una humedad muy alta y por lo tanto, afectan directamente la calidad, el 23.7% de las

pruebas de taza posee algún defecto que se mostraran más adelante y el 16.9% restante mostraron un buen puntaje sensorial.

Grafica 9. Puntaje sensorial del café según el Índice de calidad.

#	Puntaje
1	No Apto
2	Muy bueno
3	Defecto
4	Bueno



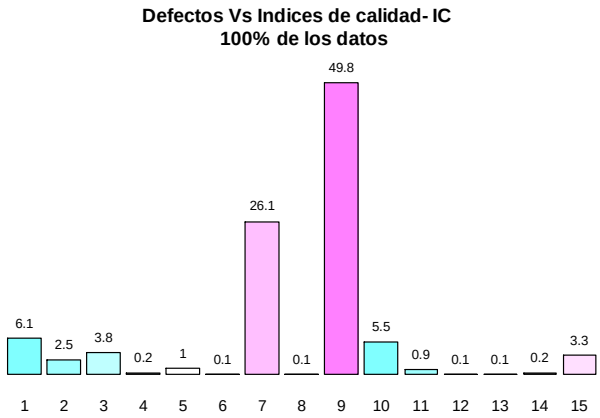
- **Por Defecto**

La gráfica 10, permite ver los defectos presentes en las pruebas de taza, es de resaltar que el 49.8% de las pruebas tuvieron una taza limpia (sin defectos), y el 26.1% como se mostró anteriormente no son aptas.

Pero los mas predominantes en las restantes muestras fueron vinagre 6.1% (causado básicamente por retrasos entre la recolección y el despulpado, fermentaciones demasiado prolongadas, uso de aguas sucias y almacenamiento húmedo del café), fermento 5.5% (se favorece el mal despulpado y por la recolección de granos sobremaduros y verdes), Sucio 3.8% (mal lavado del café), Astringente 3.3% (ocasionado por granos inmaduros), terroso 2.5%(se da en cafés que están en contacto con el suelo o agua sucia), reposo 1% (ocurre cuando los granos se almacenan durante mucho tiempo o se ha guardan en condiciones inadecuadas y en sitios muy calientes) y fenol 0. 9% (se produce por café pergamino húmedo amontonado antes de secado).

Grafica 10. Defectos presentes en las pruebas de taza contrastados con el índice de calidad.

#	Defecto	#	Defecto
1	Vinagre	9	Limpia
2	Terroso	10	Fermento
3	Sucio	11	Fenol
4	Stinker	12	Contaminado
5	Reposo	13	Cereal
6	Químico	14	Cebolla
7	No Apta	15	Astringe
8	Metálica		



Finalmente se puede decir que el índice de calidad del café hallado, guarda mucha coherencia con estudios previos realizados por Arcila et al, 2008; Puerta, 1998-1999 y 2000, para la calidad de la bebida donde se relaciona con características agronómicas del cultivo y con características organolépticas.

7. CONCLUSIONES

- El uso de la imputación de datos es una herramienta muy útil cuando se presentan datos faltantes, sin embargo, el uso de las técnicas requiere de atención, ya que su uso inadecuado podría generar datos imputados inconsistentes produciendo una base de datos no confiable y por lo tanto invalidar el estudio, por el contrario si se utiliza correctamente, se puede obtener un conjunto de datos completo, se puede reducir el sesgo debido a la ausencia de información y obtener resultados consistentes. Lastimosamente no existen criterio para decidir el porcentaje máximo de valores faltantes que deben aceptarse, pero lo que sí es cierto es que esto depende de la información que posea la base de datos y la aplicación que se vaya a realizar, debido a que habrán estudios más sensibles que otros a los valores faltantes.

- En este documento se utilizaron tres técnicas de imputación de datos Listwise o case deletion CD, Imputación Múltiple – MI y la Hot Deck – HD, de todas ellas, hasta el momento la más confiable es la MI, dada las bondades de la técnica aportada por investigaciones previas de Melissa, et al (2011), Stef y Karim (2011) Raghunathan et al, (2001), Schafer y Graham , (2002),Graham (2009),Stuart et al (2009), Allison (2001). Otero (2011), Goicochea, (2002) , Rubín y Schenker (1986) ,Van (2007) y Amón, (2010), donde menciona que la MI es una técnica que Incrementa la eficiencia de los estimadores, genera inferencias validas, permite calcular el error estándar de los estimadores, utiliza procedimientos estadísticos robustos, genera distintas opciones de datos imputados y las combina en forma adecuada, en contrate con CD donde se genera una pérdida substancial de la información en bases de datos, alterando el tamaño de la muestra y el número de variables, así como supone que la observación de los valores que faltan no es importante y puede ser ignorado por lo estimadores de los parámetros serán sesgados lo que podría invalidar las

conclusiones y con la técnica HD que tiene la limitante que cuando se presenta una gran cantidad datos faltantes se obtiene muchas veces mismo valor duplicado, lo que distorsiona la relación con el resto de las variables, carece de un mecanismo de probabilidad que imposibilita calcular su confianza y además proporciona estimaciones sesgadas de la media (Carter, 2006; Medina y Galván, 2007;Carter, 2006; Magnani , 2004).

- Al construir el Índice de calidad del Café a través de un modelo constituido por diferentes tipos de variables (continuas y ordinales) utilizando el Análisis Factorial bayesiano mediante MCMC el tema de la integración de datos empleando inferencia bayesiana fue un trabajo razonablemente bueno, esto desde la óptica de obtener resultados coherentes, es decir pues, que el factor latente de las puntuaciones proporcionó una menor dimensión que fue conveniente para resumir las variables observadas del estudio.

- Con el cálculo de las cargas factoriales y el porcentaje de varianza explicada, se pudo observar que variables como almendra defectuosa, merma y almendra sana se destacan notablemente en el factor, así como para la almendra defectuosa, la varianza del error estimado implicó que casi toda su variabilidad puede ser explicada solo por el factor latente, débilmente también se podrían tener en cuenta la merma, olor el color y la almendra sana.

- Mediante el índice de calidad hallado se pudo zonificar la calidad del café en el departamento del valle del Cauca, dando resultados aceptables, sin embargo, cabe anotar que esto es una primera aproximación a la zonificación a través de las pruebas de taza y su relación con las propiedades agronómicas y organolépticas utilizando este tipo de análisis factorial bayesiano.

BIBLIOGRAFIA

- Abayomi K., Gelman A., LEVY, M. Diagnostics for multivariate imputations. Journal of the Royal Statistical Society: Series C (Applied Statistics), (2008) 57, 273–291, DOI: 10.1111/j.1467- 9876.2007.00613.
- ACOCK, C. A., Y D. DEMO .Family diversity and well-being, Thousand Oaks, C. A. Sage. 1994
- ALFARO RODRIGO Y FUENZALIDA MARCELO. Nota Técnica: Imputación Múltiple en Encuestas Microeconómicas. Cuadernos De Economía. Banco Central De Chile - Columbia University. Vol. 46 (Noviembre), Pp. 273-288, 2009.
- ALVARADO A., G.; PUERTA Q., G.I. La variedad Colombia y sus características de calidad física y en taza. Cenicafé. Avances Técnicos 303. 2002
- ALLISON, P. Missing Data, Quantitative Applications in the Social Sciences, A Sage University Papers Series. 2001
- AMON URIBE, IVAN. Guía Metodológica Para La Selección De Técnicas De Depuración De Datos. Universidad Nacional De Colombia. Facultad De Minas, Escuela De Sistemas. 2010
- ARCILA JAIME - FARFÁN FERNANDO - MORENO ARGEMIRO - SALAZAR LUIS FERNANDO - HINCAPIÉ EDGAR. Sistemas de producción de café en Colombia Federación Nacional De Cafeteros De Colombia – Fnc Plan estrategico 2008 - 2012. 2008. Vía internet: http://www.federaciondecafeteros.org/static/files/PLAN%20ESTRATEGICO%20FNC%202008%202012.pdf
- ÁVILA G, CARLOS. Una aplicación del procedimiento Hot Deck como método de imputación. Capítulo III. Método de imputación hot deck .Trabajo Monográfico Para optar el Título Profesional de Licenciado en Estadística. Universidad Mayor de San Marcos.2002

- BADLER, CLARA E. ALSINA, SARA M. PUIGSUBIRÁ, CRISTINA B. VITELLESCHI, MARÍA S. Tratamiento De Bases De Datos Con Información Faltante Según Análisis De Las Pérdidas Con Spss .Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística (IITAE). Novenas Jornadas "Investigaciones en la Facultad" de Ciencias Económicas y Estadística, noviembre de 2004.
- BENÍTEZ ÁLVAREZ, GILMA CRISTELA, DÍAZ OJEDA, GLADIS ADELAIDA, MUÑOZ CAMPOS, ILIANA MARCELLY. Tesis De Grado “Propuesta De Un Plan Estratégico De Comercialización Del Café Nacional Para Los Productores Del Departamento De San Miguel” Facultad De Ciencias Económicas. Universidad De Oriente. San Miguel- El Salvador. 2003.
- BENÍTEZ M Y ÁLVAREZ M. “Reconstrucción de series temporales en ciencias ambientales”, Revista Latinoamericana de Recursos Naturales, 2008. Vol. 4, núm. 3, pp. 326-335
- CAPA SANTOS HOLGER Y PACHECO TOSCANO ADRIANA. Alternativas a los problemas de la imputación clásica de de datos. Una aplicación al sistema nacional interconectado del Ecuador. Centro de Investigaciones Matemáticas Aplicadas a la Ciencia y Tecnología. 2009.
- CARTER RUFUS L. Solutions Missing Data in Structural Equation Modeling. Research & Practice in Assessment Volume 1, Issue 1 Marymount University.2006
- CROS, J.E.; GUYOT, B.; VINCENT, J.C. Profil chromatographique de la fraction volatile du café; Différence entre cafés verts sain et puant influence de la torréfaction sur le grain et la boisson.. Café Cacao Thé (Francia) 23(3):193-202. 1979..
- DARÍO REBOLLEDO IVÁN, VENTO LUIS ALBERTO. Trabajo De Grado “Propuesta De Agroindustrialización Del Proceso De Beneficio Del Café En El Municipio De La Unión (Nariño) De Acuerdo A Las Características de Calidad Esperadas Por El Cliente A Nivel Internacional”. Pontificia Universidad Javeriana. Facultad De Ingeniería . 2004

- DONNER, A. The relative effectiviness of procedures commonly used in multiple regression analysis for dealing with missing values. 1982. The American Statiscian, 36, 378-381.
- DURRANT, B. G. Imputation Methods for Handling Item-Non-response in the Social Science: A Methodological Review, ESRC National Centre for Research Methods and Southampton Statistical Science Research Institute, University of Southampton, NCRM Methods Review Papers, NCMR/002. 2005
- ESPINAL, C., MARTÍNEZ, H, ACEVEDO X. La mirada del café en Colombia. Una mirada global de su estructura y dinámica 1991 – 2005. Documento de trabajo N° 59. Ministerio de Agricultura y Desarrollo Rural. Observatorio Agrocadenas Colombia. 2005.
- FEDERACIÓN NACIONAL DE CAFETEROS DE COLOMBIA - FNC, Caficultura Colombiana - Colombia, Sinónimo de Biodiversidad. Gerencia Comercial e Inteligencia competitiva. 2009. Vía internet: http://www.federaciondecafeteros.org/static/files/El%20Caf%C3%A9%20de%20Colombia%20Contexto%20General.pdf
- FEDERACIÓN NACIONAL DE CAFETEROS DE COLOMBIA – FNC. Producción y exportaciones de café de Colombia crecieron en marzo y en lo corrido del año cafetero. Boletín producción y exportaciones.2011.
- FEDERACIÓN NACIONAL DE CAFETEROS DE COLOMBIA – FNC. Valor agregado: Del árbol a la taza. Informe del Gerente General al LXVI Congreso Nacional de Cafeteros Federación: Permanencia, Sostenibilidad y Futuro. 2006
- FEDERACIÓN NACIONAL DE CAFETEROS DE COLOMBIA - FNC. Informe Comité Departamental de Cafeteros del Valle del Cauca. 2008
- FEDERACIÓN NACIONAL DE CAFETEROS DE COLOMBIA – FNC. Nuestros Cafés Especiales. 2011. Vía internet: http://www.federaciondecafeteros.org/clientes/es/nuestra_propuesta_de_valor/portafolio_de_productos/nuestro_cafe_especial/

- FEDERACIÓN NACIONAL DE CAFETEROS DE COLOMBIA- FNC - Gerencia Comercial. Bogotá. Colombia. Compras por factor de rendimiento. Comunicación GC-251. Bogotá: FEDERACAFÉ, 1999. 3 p.
- GELFAND, A. E. AND SMITH, A. F. M. Sampling-based approaches to calculating marginal densities. *J Amer Statist Assoc*, (1990). **85**, 398-409.
- GELMAN, A. AND RUBIN, D.B., "Inference from Iterative Simulation Using Multiple Sequences (with discussion)." Statistical Science, (1992). 7,457-511.
- GEMAN AND D. GEMAN. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images".(1984) *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** (6): 721–741.
- GÓMEZ VILLEGAS, MIGUEL ÁNGEL .*¿Por qué la inferencia estadística bayesiana?* Boletín de estadística e investigación operativa : 2006 BEIO, 22 (1). pp. 6-8. ISSN 1889-3805
- GORSUCH, R. *Factor Analysis. Second Edition.* 1983.
- GRAHAM J.W. Missing data analysis: making it work in the real world. Annual Review of Psychology, (2009) 60, 549–576, DOI: 10.1146/annurev. psych.58.110405.085530
- GRAJALES T. El Análisis factorial. 2000. Extraido de: http://tgrajales.net/estfactorial.pdf
- HEIDELBERGER, P. AND WELCH, P.D. (1983) "Simulation Run Length Control in the Presence of an Initial Transient," Operations Research 31:1109-1144.
- HERNÁNDEZ J., RAMÍREZ, Mª J., RAMÍREZ, C. Introducción a la Minería de Datos. 2005 .ISBN: 8420540919.
- INSTITUTO INTERAMERICANO DE COOPERACIÓN PARA LA AGRICULTURA (IICA). Protocolo De Análisis De Calidad Del Café, Programa Cooperativo Regional Para El Desarrollo Tecnológico Y Modernización De La Caficultura Promecafe. Guatemala, 2010 ISBN 13: 978-92-9248-236-7

- KLINE, R.B. *Principles and Practice of Structural Equation Modeling.* The Guilford Press. 1998.
- LITTLE, R. J. A., y RUBIN, D. B. Statistical Analysis with missing data, New York: Wiley. (1987)
- LITTLE, R. J., Y D. Rubin . Statistical analysis with missing data, New York, Wiley. 1987.
- MAGNANI MATTEO. Techniques for Dealing with Missing Data in Knowledge Discovery Tasks..University of Bologna, department of Computer Science. Bologna - ITALY. 2004.
  MANAGEMENT .1996. Journal of Business and Psychology 11, 101-112.
- MARÍN L., S.M.; ARCILA P., J.; MONTOYA R., E.C.; OLIVEROS T., C.E. Relación entre el estado de madurez del fruto del café y las características de beneficio, rendimiento y calidad de la bebida. Cenicafé 54(4) 297-315.2003
- MARTÍNEZ E. Análisis factorial.2008. via internet: http://www.uantof.cl/facultades/csbasicas/Matematicas/academicos/emartinez/magister/factorial.pdf
- MARWALA TSHILIDZI. Computational Intelligence for Missing Data Imputation, Estimation, and Management: Knowledge Optimization TECHNIQUES. University of Witwatersrand, South Africa. 2009- ISBN 978-1-60566-336-4 (hardcover) -- ISBN 978-1-60566-337-1 (ebook).
- MEDIAVILLA MAURO. Las Becas y La Continuidad Escolar En El Nivel Secundario Post-Obligatorio.Un Análisis De Sensibilidad A Partir De La Aplicación Del Propensity Score Matchin. 2010
- MEDINA FERNANDO Y GALVÁN MARCO. Serie 54 - Imputación de datos: teoría y práctica. Estudios estadísticos y prospectivos. División de Estadística y Proyecciones Económicas. Naciones Unidas .2007. Santiago de chile. Chile.
- MELISSA J. AZUR, ELIZABETH A. STUART,CONSTANTINE FRANGAKIS & PHILIP J. LEAF. Multiple imputation by chained equations: what is it and how does it work?.International Journal of Methods in Psychiatric Research

Int. J. Methods Psychiatr. Res. 20(1): 40–49 (2011) Published online in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/mpr.329

- MORALES P. “El Análisis Factorial en la construcción e interpretación de tests, escalas y cuestionarios Universidad Pontificia Comillas, Madrid Facultad de Ciencias Humanas y Sociales . 2011.Via internet: http://www.upcomillas.es/personal/peter/investigacion/AnalisisFactorial.pdf
- MUÑOZ, ANDREA. Diseño De Una Guía De Selección Del Sello Para Cafés Sostenibles. Universidad Tecnológica De Pereira .Facultad De Ciencias Ambientales. Administración Del Medio Ambiente, Pereira .2007
- OTERO G.DEBORAH. Imputación de datos faltantes en un Sistema de Información sobre Conductas de Riesgo . Master en Tecnicas Estadisticas. Universidad da Coruña.2011
- PÉREZ C., SANTÍN D. Minería de datos. Técnicas y Herramientas .2007 ISBN: 8497324922.
- PUERTA Q., G.I. Influencia del proceso de beneficio en la calidad del café. Cenicafé 50(1):78-88. 1999.
- PUERTA Q., G. I. Influencia de los granos de café cosechados verdes, en la calidad física y organoléptica de la bebida. Cenicafé 51(2):136- 150. 2000.
- PUERTA Q., G. I. La calidad de las variedades de café Coffea arabica L. cultivadas en Colombia.Cenicafé 49 (4): 265-278. 1998.
- PUERTA Q., G. I. Calidad en taza de algunas mezclas de variedades de café de la especie. Coffea arabica L. Cenicafé 51(1): 5-19. 2000.
- QUINN K. M."Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses." Political Analysis. 2004.12: 338-353.
- RAFTERY, A.E. AND LEWIS, S, "How Many Iterations in the Gibbs Sampler?" In Bayesian Statistics 4 (eds. J.M. Bernardo, J. Berger, A.P. Dawid and A.F.M. Smith), Oxford: Oxford University Press. (1992), 763-773.
- RAGHUNATHAN T.W., LEPKOWKSI J.M., VAN HOEWYK J., SOLENBEGER P. A multivariate technique for multiply imputing missing

values using a sequence of regression models. Survey Methodology, (2001) 27, 85–95.

- REVUELTA J. Estimación y evaluación de modelos psicométricos Mediante simulaciones posteriores bayesianas. Metodología de las Ciencias del Comportamiento 3(1),1-18 AEMCCO, (2001)
- ROA M., G.; OLIVEROS T., C.E.; ÁLVAREZ G., J.; RAMÍREZ G., C.A.; SANZ U., J.R.; DÁVILA A., M.T.; ÁLVAREZ H., J.R.; ZAMBRANO F., D.A.; PUERTA Q., G.I.; RODRÍGUEZ V., N. Beneficio ecológico del café. Chinchiná, Cenicafé, 1999, 273 p.
- ROTH, P. H.; CAMPION, J. E., y JONES, S. D.The impact of four missing data techniques on validity estimates in human Resource
- Rubin, D. B. 1976. “Inference and Missing Data.” *Biometrika* 63: 581-592.
- Rubin, D.B. Multiple Imputation for Nonresponse in Surveys. New York, NY: Wiley. (1987).
- RUBIN, D.B., AND SCHENKER, N. “Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse,” *Journal of the American Statistical Association*, 1986, 81, 366-374.
- RUIZ MARÍA, UREÑA MARÍA, TORRES MARÍA Y ESPINAL FEDERICO. Los mercados del café y de los cafés especiales. Situación actual y perspectivas. Programa Mas Inversión para el Desarrollo alternativo Sostenible.MIDAS.2009
- SCHAFER, J. L. Y J. W. GRAHAM. Missing Data, Our View of the State of the Art, Psychological Methods. 2002, vol. 7, No. 2
- SCHAFER, J.L. & J.W. GRAHAM "Missing Data: Our View of the State of the Art." *Psychological Methods*, (2002). 7(2), 147-177.
- SCHAFER, J.L. *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London. 1997.
- Silva LC, Benavides A (2001) *El enfoque bayesiano*: *otra manera de inferir* Gaceta Sanitaria 15 (4): 341-346.
- STATACORP. Multiple-Imputation Reference Manual, Stata Press. 2009.

- STEF VAN BUUREN, KARIN GROOTHUIS-OUDSHOORN. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, (2011)- 45(3), 1-67.
- STUART E.A., AZUR M., FRANGAKIS C.E., LEAF P.J. Practical imputation with large datasets: a case study of the Children's Mental Health Initiative. American Journal of Epidemiology, (2009) 169, 1133–1139, DOI: 10.1093/aje/kwp026
- TOBASURA ACUÑA, ISAÍAS. La Crisis Cafetera, Una Oportunidad Para El cambio En Las Regiones Cafeteras De Colombia. Revista Agronomía ISSN: 06583076 ed:v.13 fasc.2 p.5 – 17.2006
- USECHE LELLY, MESA DULCE. Una Experimental Sur Introducción A La Imputación De Valores Perdidos. Universidad Nacional del Lago Terra Nueva Etapa, año/vol. XXII, número 031- 2006
- van Buuren, S"Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification, 2007." *Statistical Methods in Medical Research*, 16, 219–242
- VARELA MALLOU J, BRAÑA TOBÍOT, GARCÍA CARREIRA A Y RIAL BOUBETA Y VÁZQUEZ FERNÁNDEZ X. Estimación de la respuesta de los "no sabe/no contesta" en los estudios de intención de voto. Universidad de A Coruña - Universidad de Santiago .1998.REIS - *Revista Española de Investigaciones Sociológicas* 83/98 pp. 269-287
- VINCENT, J.C. Influence de la maturité des fruits sur la qualité du Café Robusta. Café, Cacao, Thé 12(3):240-249. 1968.
- ZULUAGA V., J. Los factores que determinan la calidad del café verde. In: 50 Años de Cenicafé 1938- 1988. Conferencias Conmemorativas. Chinchiná,Cenicafé, 1990. p. 167-183.

ANEXOS

(Prueba de Heidelberger and Welch para estacionariedad)

	stest	start	pvalue	htest	mean	halfwidth
Lambdavaried.2	1	1	0,2	1	-0,0695	0,001
Lambdasombr.2	1	1	0,1916	1	-0,0569	8,00E-04
Lambdafertiliz.2	1	1501	0,0604	1	0,0243	0,001
Lambdamsnm.2	1	1	0,4339	1	-0,1197	9,00E-04
Lambdabenefic.2	1	1	0,1981	1	0,0172	9,00E-04
Lambdatsecado.2	1	1	0,7793	1	0,0487	9,00E-04
Lambdaaspect.2	1	1	0,3293	1	-0,1051	0,001
Lambdahumcpsp.2	1	1	0,8285	1	-0,0119	8,00E-04
Lambdamermap.2	1	1	0,054	1	0,5666	0,0032
Lambdaoloralm.2	1	1	0,1706	1	0,0305	0,001
Lambdacoloralm.2	1	1	0,2459	1	-0,138	0,0012
Lambdaalmendefecp.2	1	1	0,0861	1	0,9239	0,0058
Lambdaalmsanap.2	1	1	0,1104	1	-1,004	0,0063
Lambdapuntfin.2	1	501	0,2059	1	0,1179	0,0012
phi.1.2	1	1	0,0616	1	-0,3623	0,0021
phi.2.2	1	1	0,8164	1	0,0268	8,00E-04
.....	.....	.....	.....	.....	.....	.....
……	.....	.....	.....	.....	.....	.....
phi.991.2	1	1	0,0511	1	1,105	0,0062
si.varied	1	1	0,9098	1	0,996	0,0011
Psi.sombr	1	1	0,7098	1	0,9989	0,0011
Psi.fertiliz	1	1	0,6857	1	1,0009	0,0013
Psi.msnm	1	1	0,4609	1	0,9888	0,0011
Psi.benefic	1	1	0,7719	1	1,0008	0,0014
Psi.tsecado	1	501	0,2265	1	0,9998	0,0013
Psi.aspect	1	1	0,9247	1	0,9903	0,0011
Psi.humcpsp	1	1	0,4286	1	1,0014	0,0014
Psi.mermap	1	1	0,0954	1	0,6838	8,00E-04
Psi.oloralm	1	1	0,6956	1	1,0012	0,0014
Psi.coloralm	1	1	0,8531	1	0,9824	0,0011
Psi.almendefecp	1	1	0,7701	1	0,154	2,00E-04
Psi.almsanap	1	1	0,7516	0	7,00E-04	1,00E-04
Psi.puntfin	1	1	0,9513	1	0,9888	0,0012