UNIVERSIDAD POLITÉCNICA DE CATALUÑA (UPC) BARCELONATECH

FACULTAD INFORMÁTICA DE BARCELONA (FIB)

MASTER EN INGENIERIA INFORMÁTICA

Trabajo de Fin de Máster

Implementación de un módulo avanzado de imputación de datos faltantes para KLASS

Autor:

Juan de los Reyes Piedra

Directora:

Karina Gibert Oliveras.

Departamento:

Estadística i Investigació Operativa.

Fecha de presentación:

25 de abril de 2019

RESUMEN

El presente trabajo describe los inconvenientes que puede representar la existencia de valores faltantes en una matriz de datos que va a ser utilizada para realizar procesos de análisis, inferencia y generación de conocimiento. La presencia de estos valores es casi segura al trabajar con bases de datos, y puede producirse por muchas causas.

El tratamiento de valores faltantes se torna un tema fundamental al momento que extraer información válida, nueva y pertinente de una matriz de datos, y un mal tratamiento de los mismos, puede no solo no ayudar a solucionar el problema, sino acabar empeorando la situación al introducir sesgo en las variables, eliminado grupos completos de la población que se está representando o introduciendo ruido en la matriz. Estos inconvenientes pueden aparecer especialmente cuando el tratamiento de faltantes se realiza sin poner mayor atención en las particularidades de la matriz de datos, factores como el tamaño, la distribución de faltantes y el tipo de datos deben ser tomados muy en cuenta al momento de realizar este proceso.

El proyecto revisa algunos de los métodos más utilizados para la imputación de datos faltantes y presenta el método ***Mixed Intelligent-Multivariate Missing Imputation (MIMMI),*** descrito en (Gibert 2013), que es un método de compromiso entre precisión y complejidad, cuya implementación en el sistema Java-KLASS es el principal objetivo de este proyecto. Java-KLASS es un sistema propuesto y diseñado en la Facultad de Informática de Barcelona (FIB), bajo la dirección de la Dra. Karina Gibert, como un software que comprende un compendio de herramientas para la realización de Minería de Datos, que anterior a este trabajo de fin de Máster, no contaba con un método adecuado para el tratamiento de faltantes. Para terminar se presenta un caso de estudio con datos reales donde se analiza las ventajas que MIMMI presenta sobre métodos más básicos con los que contaba Java-KLASS previamente, y se sustenta si la implementación mantuvo y mejoró la funcionalidad del sistema.

Palabras Clave: Imputación de valores faltantes, MIMMI, Ciencia de Datos, soporte a la decisión

Juan Francisco de los Reyes Piedra

AGRADECIMIENTOS

Quiero agradecer sinceramente a Karina Gibert, Ph.D, por brindarme todo su apoyo para la realización de este proyecto, sin todo el tiempo y esfuerzo que ha dedicado a este trabajo como tutora, este trabajo no hubiera sido posible. También es justo agradecer a todos quienes han formado parte del desarrollo de KLASS, *la tribu-KLASS*, ya que sin todos los aportes que se han venido dando a este sistema a lo largo del tiempo, sería muy difícil plantearse en este momento el desarrollo que fue realizado. Agradecer también al departamento de Salud Mental y Abuso de Sustancias de la OMS (Organización Mundial de la Salud) por facilitar los datos utilizados en este trabajo. Y para terminar quiero agradecer al Gobierno Nacional del Ecuador, ya que sin el programa de becas implementado por el Estado Ecuatoriano, no me hubiera sido posible cursar el Máster en Ingeniería Informática de la Facultad Informática de Barcelona.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1

INTRODUCCIÓN

1.1 Contexto

Al momento de realizar análisis o tratamientos de datos es común encontrarse con datos faltantes dentro de una base de datos. Sin importar la fuente de la que provengan los mismos, este fenómeno aparece frecuentemente, complicando la tarea de extraer información de los datos, ya que muchos métodos de análisis no soportan la presencia de faltantes [Gibert, 2013]; por este motivo la tarea de procesar previamente la información de forma adecuada se torna fundamental, ya que se debe asegurar que el conjunto de valores con el que se va a trabajar sea una muestra completa y significativa de la población objeto que desea representar.

De hecho el tratamiento de datos faltantes es una de las tareas de mayor envergadura en la etapa de pre-procesamiento de datos en cualquier proceso de análisis de datos o de ciencia de datos [Gibert et al., 2016]. Aún y así es raro el software estadístico o de data science que ofrece un módulo bien estructurado de imputación de datos faltantes permita afrontar este proceso de forma ágil, recayendo aún hoy en día la mayor parte de la responsabilidad y desarrollo en el propio analista de datos.

Los datos faltantes o perdidos, también conocidos por su nombre en Inglés “missing values” o simplemente “missings”, son aquellos valores que faltan, como su nombre indica, es decir que no aparecen en una matriz de datos [Dagnino, 2014]. Estos datos no se encuentran registrados por un sin número de motivos.

Existen algunos métodos para lidiar con los datos faltantes, por ejemplo la eliminación de la matriz de datos de aquellas filas que tienen valores faltantes en alguna de sus columnas (listwise deletion), opción por defecto en los softwares de análisis de datos muy frecuentemente, lo cual puede causar que la base de datos reduzca drásticamente su tamaño, o si los valores faltantes excluidos no son aleatorios, puede causar serios problemas de representatividad de los datos [Gibert et al., 2016]. Otro enfoque es la substitución de los valores faltantes por la media global de la variable, lo que permite aprovechar la información útil de la matriz, pero puede reducir la varianza de dicha variable y distorsionar la relación de la variable con las otras variables, alterando fuertemente los modelos que se deriven de la matriz de datos imputada [Gibert, 2013]. Existe otra familia de métodos más sofisticados que tienen por objetivo remplazar estos valores perdidos por estimaciones, que preserven las relaciones entre variables. Estos métodos son llamados métodos de imputación de datos o valores faltantes, y si bien son una opción mucho más correcta y estadísticamente fundada, son más complejos, requieren de experiencia y consumen un tiempo considerable para su implementación [Gibert, 2013].

Dentro de esta última familia de métodos para lidiar con datos faltantes, se encuentra el método MIMMI (Mixed Intelligent-Multivariate Missing Imputation), a medio camino entre los métodos simples como el de substitución de faltantes por la media global de la variable (rápido pero altamente impreciso) y los métodos más complejos de imputación de faltantes (más precisos pero mucho más lentos). MIMMI tiene una aplicación fácil e intuitiva, la cual toma en cuenta las relaciones globales entre las variables para la imputación, en un marco libre de hipótesis técnicas y distribucionales sobre los datos, siendo adecuada para cualquier aplicación sin hacer suposiciones estadísticas. Además, incorpora implícitamente el conocimiento específico de dominio a priori de los expertos en el proceso de imputación [Gibert, 2013] y propone un buen compromiso entre el tiempo requerido para la imputación y la precisión de las mismas.

Tomando en cuenta lo expuesto y la importancia que tiene el tratamiento adecuado de valores faltantes para realizar un correcto análisis de una base de datos, el objetivo de este trabajo es la implementación del método MIMMI para la substitución de valores faltantes como una nueva funcionalidad del software Java-KLASS. Este sistema está definido en el **capítulo 3.1** de este documento, pero para dar una corta definición del mismo, es un

sistema que realiza procesos integrales de data science e incorpora entre su funcionalidad herramientas para el tratamiento de faltantes.

Al dotar a este sistema de la funcionalidad de imputación de valores faltantes a través del método MIMMI, se amplían las opciones con las que al momento se cuenta para la substitución de valores faltantes, lo que a más de extender su funcionalidad y dotarle de una herramienta muy poderosas para lidiar con este tipo de valores, fortalece y apoya los procesos que se lleven a cabo en el sistema, convirtiendo a Java-KLASS en una herramienta aún más poderosa para en análisis de datos.

1.2 Estructura del Documento

El documento actual estará compuesto por la siguiente estructura:

- **Capítulo 1: Introducción.** Contexto motivación y objetivos.
- **Capítulo 2: Estado del Arte.** Introducción, distribución de datos faltantes y metodologías de tratamiento.
- **Capítulo 3: Desarrollo del Proyecto.** Descripción del problema, Introducción a Java-KLASS y desarrollo de la funcionalidad del Método MIMMI.
- **Capítulo 4: Caso de Estudio.** Se realiza una comparativa entre los resultados de la clasificación al utilizar uno de los métodos de sustitución de faltantes con los que contaba Java-KLASS y al utilizar el nuevo método implementado.
- **Capítulo 5: Conclusiones.** Análisis de resultados, argumentos y trabajos futuros.

1.3 Motivación

Dado que el tratamiento de los valores faltantes es un tema fundamental dentro del pre-procesamiento de los datos antes de su análisis, es importante contar con una herramienta que permita realizar esta tarea de forma correcta. Si bien en la actualidad los software especializados en el tratamiento de datos cuentan con opciones que pueden realizar este proceso, la mayoría descansan sobre una serie de hipótesis técnicas que limitan su aplicabilidad a ciertos tipos de problemas y se torna fundamental conocer exactamente cómo son los datos y cuáles son las condiciones de aplicabilidad del tipo de método que se va a utilizar, para asegurar un uso correcto del mismo. Puesto que realizar un tratamiento de forma inadecuada puede causar inconvenientes como alterar la distribución de referencia de las variables que se analizan, introducir sesgos en el análisis y la disminución

de la varianza de las variables, lo que puede tener graves consecuencias para en el resultado del análisis.

Existen pocos métodos de imputación de propósito general que permitan superar la fase de imputación de datos faltantes con agilidad en aplicaciones reales, y en [Gibert, 2013] se describe el método MIMMI que es una posible solución a este problema.

Por las ventajas propias de MIMMI que se describen en el **Capítulo 2.3.2** de este documento, es que la implementación del mismo dentro del sistema Java-KLASS nos ayudará a potenciar la funcionalidad del sistema en cuanto al tratamiento de datos faltantes. Además esta implementación nos permitirá una amplia evaluación de las prestaciones del método, actualmente objeto de implementación experimental, y poder comparar las ventajas de éste frente a otros métodos de imputación de valores faltantes, algunos más básicos, ya implementados en el sistema Java-KLASS o externos al mismo.

1.4 Objetivos

Ampliar la Funcionalidad del sistema informático Java-KLASS respecto al tratamiento de los datos faltantes, implementando la nueva funcionalidad del método de imputación MIMMI, que permita remplazar los datos faltantes tanto de variables categóricas como numéricas, logrando de esta manera mantener y potenciar la funcionalidad del sistema.

Este objetivo se concretara a través de los siguientes objetivos específicos:

- Ampliar la Funcionalidad de Java-KLASS respecto al tratamiento de los datos faltantes, en particular implementando el método MIMMI.
- Diseñar e implementar la estructura del Informe que el sistema presentará al usuario para el proceso de imputación de datos faltantes.
- Realizar pruebas del funcionamiento del método de imputación datos faltantes con datos reales, asegurando que la implementación es correcta y que queda bien integrado en la aplicación, manteniendo y potenciando las funcionalidades de JAVA-KLASS.

CAPÍTULO 2

ESTADO DEL ARTE

2.1 Introducción

En general cuando se trata de analizar datos, y principalmente en estadística, los datos que son objeto de tratamientos están representados en una matriz de datos, en la cual tradicionalmente las filas de la matriz representan individuos, y las columnas representan las variables medidas cada una en su respectiva unidad [Little et al., 2002], pudiendo ser estos valores numéricos o categóricos; un dato faltante se presenta cuando por cualquier motivo no existe un valor asociado a una celda de la matriz de datos con las cuales se pretende realizar el análisis.

Este fenómeno es muy común, sobre todo cuando se trabaja con datos obtenidos de aplicaciones reales. Un dato faltante puede ser producto de muchos factores y tener diferentes naturalezas [Gibert et al., 2016], ya que puede ser producto de fallas en la digitación, un sensor temporalmente sin conexión con el servidor, datos deliberadamente ocultos, datos no proporcionados, datos que corresponde a valores especiales, datos que se perdieron o corrompieron a lo largo del tiempo por algún error tecnológico, política de codificación de datos incorrecta en la recopilación de datos, entre otros muchos [Gibert,

2013]. Sea como fuese, estos valores faltantes deben ser detectados y luego diagnosticados [Gibert et al., 2016].

En lo que se refiere al diagnóstico, en la fase de preprocesamiento de los datos se debe intentar establecer las posibles razones que causaron los valores faltantes, y con esta información tratar de tomar las mejores decisiones para lidiar con los ellos ya que es importante identificar la naturaleza de los datos faltantes para así realizar un mejor tratamiento de los mismos.

2.2. Tipología de los Datos Faltantes

Los datos faltantes pueden clasificarse, según su naturaleza, en datos faltantes aleatorios y datos faltantes no aleatorios [Gibert et al., 2016], haciéndose fundamental distinguir el tipo de faltante de la matriz de datos con la cual vamos a trabajar, para así, realizar un tratamiento adecuado de los mismos.

A continuación se describe cada uno de estas familias de datos faltantes:

- **Los Datos Faltantes Aleatorios (incluyen faltantes completamente al azar MCAR y faltantes al azar MAR, generalmente referidos en la literatura):** Suponiendo una variable de interés, que tiene algunos valores faltantes, podemos decir que estos datos faltantes son aleatorios si no siguen ningún patrón particular. Esto recoge el escenario en el que no se identifica un perfil específico de individuo que produzca dato faltante en esa variable y por tanto se podrá después asumir que la distribución del faltante es la misma que la de la variable de interés, y que su presencia no altera la representatividad de la muestra. Son los únicos valores faltantes susceptibles de ser ignorados en circunstancias muy concretas (análisis univariante) sin perjuicio de alterar los resultados

  Datos Faltantes no Aleatorios (NMAR): O datos faltantes en forma no aleatoria; se presentan cuando la probabilidad de valores faltantes en una variable de interés, están relacionados con ciertos valores de otras variables (presentes o no en la matriz de datos), que configuran una subpoblación determinada, y por tanto, su existencia no se distribuye igual que la variable de referencia.

Cualquier política de tratamiento de datos faltantes que pase por la supresión de observaciones corre el riesgo de liquidar todo un subconjunto específico de la población objetivo en este caso, lo cual compromete muy seriamente la aplicabilidad de los resultados del análisis.

En este punto cabe recalcar que una de las causas para la aparición de valores faltantes NMAR, es la denominada faltante estructural, se trata de características que se representan a través de las variables, con las cuales algunos de los individuos de la matriz no cuentan; por ejemplo si tenemos una matriz de datos sobre especies animales de un lugar en concreto, y una de las variables corresponde al número de patas del animal, para aquellos animales como los peces que no cuentan con esta característica, el valor que se representará en la matriz será un dato faltante, causado por el propio diseño de la estructura de la matriz.

Ahora bien, se debe tener mucho cuidado a la hora de determinar el tipo de datos faltantes. Como se puede apreciar, existen 2 grandes grupos los de distribución aleatoria de los datos y los de distribución no aleatoria. El segundo caso es de especial interés, ya que si los datos faltantes de nuestra matriz de datos siguen una distribución no aleatoria, debemos necesariamente tratar de imputarlos y está totalmente contraindicado cualquier método que elimine las filas de la matriz con datos faltantes, puesto que con ello podemos eliminar todo un sub conjunto de la población objetivo de nuestro análisis, lo cual tendría efectos desastrosos en los modelos, que dejarían de representar la población de referencia, sino solo la subpoblación que no presenta datos faltantes.

2.3 Metodologías de tratamiento de datos faltantes

2.3.1 Métodos que Eliminan Filas Con Datos Faltantes

Listwise Deletion: Este método de tratamiento de datos faltantes consiste simplemente en excluir de la matriz de datos con la cual se va a realizar el análisis las filas que contienen valores faltantes en ALGUNA de sus variables, para trabajar únicamente con observaciones que disponen de información completa para todas las variables [Nakai and Weiming, 2011]. Cuando se utiliza listwise deletion se asume que las distribución de los faltantes es completamente aleatoria (MCAR), y que la submatriz resultante de la eliminación de filas

conserva las características de los datos completos, lo que en la práctica es muy improbable. Este tipo de sistema de tratamiento de valores perdidos suele ser el estándar por defecto en los paquetes estadísticos y la mayoría de sistemas de análisis de datos, data science, etc. Con lo cual es muy relevante asegurarse de que está desactivada la opción de tratamiento automático de los datos de faltantes para evitar la reducción drástica de los datos que realmente se analizan de forma transparente al usuario.

Si los valores faltantes tienen una distribución completamente aleatoria MCAR y si se puede asegurar que la distribución conjunta de los datos que permanecen en la matriz de datos luego de utilizar listwise deletion, es idéntica a la distribución conjunta real de la población, este método presenta las siguientes ventajas: su complejidad es mínima, requiere muy pocos recursos de tiempo y computación y genera una matriz completa que puede ser utilizada en las siguientes fases del análisis de datos [Soley-Bori, 2013].

Por el contrario, la desventaja de su uso es que puede excluir una fracción grande y representativa de la matriz de datos, lo que puede dar como resultado que la submatriz resultante no represente objetivamente la población que se está intentando representar, introduciendo sesgos serios en los modelos que se inducirán después con estos datos. Esto ocurre cuando los datos faltantes no son aleatorios. Cuando lo son, igualmente este método reduce el tamaño de la muestra lo que incrementará la varianza de todos los estimadores y por tanto redundará en la pérdida de precisión y de índices de calidad y bondad de ajuste de cualquier modelo que se induzca con los datos después de la imputación.

2.3.2 Métodos de Imputación de Datos Faltantes

Los métodos de imputación de datos faltantes, buscan una posible solución a falta parcial de valores en una matriz de datos, sustituyendo cada uno de los valores que no se encuentran disponibles a través de un procedimiento que utiliza información contenida en la misma muestra para asignar un valor a aquellas variables que tienen registros con valor ausentes. La razón principal para utilizar métodos de imputación de datos faltantes es para obtener un conjunto completo y consistente de datos, aprovechando toda la información útil presente en la matriz de datos, que se pierde con la eliminación de filas parcialmente incompletas del listwise deleteion. Con la matriz imputada se puedan aplicar métodos de tratamiento de datos, como si no hubiera faltantes, y así poder llegar a obtener información relevante de la matriz de datos [Otero, 2011].

Existen muchos métodos de imputación, pero en general se clasifican en 2, Imputación Simple e Imputación Múltiple. A continuación describimos estos dos y describimos brevemente algunos de los métodos más comunes de cada uno.

Imputación Simple

La imputación simple remplaza los valores faltantes en la matriz de datos por un único valor, obtenido de la misma matriz, construyendo así una matriz completa de datos que puede ser utilizada para continuar con el proceso de análisis.

Luego de realizar este proceso se trata la matriz de datos como una muestra completa, lo que puede causar que se subestime los errores estándar de las variables que presentaban valores faltantes originalmente.

A continuación se describen algunos de los métodos de imputación simple más difundidos:

Imputación de datos faltantes estructurales: Existen un tipo de valores faltantes llamados faltantes estructurales, para los cuales se debe realizar el proceso de imputación de este tipo de valores específico del dominio. La aparición de este tipo de faltantes se da cuando existe una o más variables que no corresponden a las características de todos los individuos, por ejemplo utilizando el ejemplo sobre la matriz de datos de especies animales de un lugar en concreto, la variable forma del pico representa una característica que ciertas especies como los mamíferos no cuentan, por lo cual es preciso imputar estos valores con el número cero, para que no se introduzca ruido en la matriz.

Imputación de datos Faltantes por la Media Global de la variable: Este método de sustitución de datos utiliza los valores que no son valores faltantes de una variable que contiene algunos de ellos, para obtener la media de la misma, y así imputar los mismos por la media de las observaciones útiles. La imputación de datos faltantes utilizando promedios es una vieja práctica común en diversas disciplinas, a pesar de que por sus limitaciones teóricas no se considera un procedimiento apropiado. En su aplicación se asume que los datos faltantes siguen un patrón MCAR, y ha sido ampliamente documentado que su aplicación afecta la distribución de probabilidad de la variable

imputada, sesga la media, subestima la varianza, no toma en cuenta las relaciones globales entre las variables para la imputación atenuando la correlación con el resto de variables de la matriz, entre otras inconvenientes. [Soley-Bori, 2013].

A pesar de lo todo, este método es muy utilizado como estimador de datos faltantes, quizás por su sencillez o por la creencia errónea que la media de una variable es un buen valor para substituir faltantes en una muestra con una distribución normal, pero en realidad si bien permite de forma rápida y sencilla contar con una matriz de datos completa para el análisis, y permite incluir en el subsiguiente análisis todas las filas de la matriz original, su uso no es recomendado, por ser el menos preciso de todos los métodos de imputación.

Maximum Likelihood Estimation (MLE, método de la máxima verosimilitud): Este método, también conocido en castellano como método de máxima verosimilitud, fue propuesto hace ya muchos años por Fisher (1890-1962), y aunque ha tenido un sin número de modificaciones y correcciones, es en la actualidad que gracias a los avances en la computación, se ha empezado a utilizar de forma generalizada, debido a que el método requiere la resolución de problemas numéricos de una magnitud considerable, que sin las ventajas que presentan los ordenadores, estos se convierten en una barrera difícil de superar para la correcta ejecución del mismo [Molinero, 2003].

El método de máxima verosimilitud proporciona una técnica de estimación dada una muestra finita de datos, cuya idea fundamental es tomar como estimación el valor que maximice la probabilidad conjunta de obtener la muestra observada, asumiendo que los datos faltantes siguen un patrón MAR [Graham, 2009]. Para ello utiliza la función de verosimilitud, que en rasgos generales permite comparar cuanto más verosímil es un parámetro para explicar el evento observado, de esta forma se buscará el parámetro poblacional que mejor explique los datos, y ése será el que se utilice para la imputación.

La ventaja de este método para la imputación de faltantes es que proporciona estimaciones eficientes y con buenas propiedades teóricas, siempre que cuente con un modelo estadístico sólido disponible para el problema objetivo [Gibert, 2013], y entre sus limitaciones es que puede consumir mucho tiempo y recursos, tanto de quiénes están involucrados en el proceso como en lo referente a computación, llegando en ocasiones a convertirse en un verdadero desafío.

El procedimiento para imputar los datos faltantes de una muestra se resume según [Medina and Galvan, 2007] a través de los siguientes pasos:

1. Estimar los parámetros del modelo con los datos completos con la función de máxima verosimilitud.

2. Utilizar los parámetros estimados para predecir los valores omitidos.

3. Sustituir los datos por las predicciones, y obtener nuevos valores de los parámetros maximizando la verosimilitud de la muestra completa.

4. El algoritmo se aplica hasta lograr la convergencia, la cual se obtiene cuando el valor de los parámetros no cambia entre dos iteraciones sucesivas.

Método Expectation-Maximization (EM): El método de EM, está basado en factorizar de forma iterativa la función de máxima verosimilitud, permitiendo obtener estimaciones máximo verosímiles a pesar de no contar con datos completos en una matriz determinada [Otero, 2011].

Este algoritmo se basa en un paso de expectativa y un paso de maximización, que se repiten varias veces hasta que se obtengan estimaciones de máxima verosimilitud, en cada paso o iteración se calculan los valores esperados para la información faltante a partir de los valores observados y las estimaciones actuales, para posteriormente reemplazar la información faltante con los valores esperados obtenidos [Soley-Bori].

Para la aplicación correcta de este algoritmo la distribución de los valores faltantes de la matriz de datos ha de ser aleatoria MAR, y si bien este método está bien fundado y proporciona una buena opción para la substitución de faltantes, es importante recalcar que requiere que la matriz de datos sea de un tamaño considerable, a más de un gran conocimiento y experiencia por parte de quienes están encargados de ponerlo en práctica, para asegurar que se realice satisfactoriamente, ya que no es un método sencillo de implementar de forma correcta.

Imputación a través de Métodos Predictivos Clásicos: Una forma bastante precisa de imputar los datos faltantes en una cierta variable es construir un modelo predictivo para la misma en función de las demás, ya sea vía regresión múltiple o ANCOVA (si se está imputando una variable numérica) o una regresión multinomial, para una variable categórica, o cualquier otro modelo de ciencia de datos más sofisticado. El principal

problema de este enfoque, por lo que es muy raramente utilizado en aplicaciones reales, es que el tiempo requerido para encontrar estos modelos predictivos para la imputación es largo y muchas veces consume más tiempo que el disponible para la totalidad del análisis. Por otro lado, estos modelos solo se pueden construir si todas las variables explicativas están completas, es decir, que no contienen valores faltantes, lo cual es difícil que se cumpla en aplicaciones reales. Si bien pueden ser modelos de imputación muy precisos, el costo computacional y de tiempo de experto suele ser inasumible en aplicaciones reales

Imputación por el método del kNN: Es una de las más antiguas propuestas de propósito general. Se basa en la idea de seleccionar los casos de la BD más cercanos a la fila que se quiere imputar y copiar los valores de las variables faltantes (o promediar los de los 2 o 3 vecinos más cercanos). Requiere de la definición de una distancia entre filas de la matriz de datos, que se defina sobre las varialbes que no son faltantes. En general existen implementaciones solo para variables numéricas y presenta el mismo problema que el anterior, puesto que la distancia no se puede calcular con datos faltantes

Interpolación para variables temporales: Cuando se trata de variables temporales se suele realizar una imputación por interpolación, utilizando únicamente los datos anteriores y posteriores de la variable en cuestión y proyectando las tendencias crecientes o decrecientes sobre el valor faltante de modo más o menos sofisticado.

Método Mixed Intelligent-Multivariate Missing Imputation (MIMMI): Es un método introducido por Gibert en [Gibert, 2013] que se base en realizar clustering previo con un subconjunto de pocas variables relevantes y predictivas que contengan un rango bajo de missings, he imputar los valores perdidos de las variables con las medias condicionales de los clusters descubiertos [Gibert, Sanchez 2016].

El método MIMMI, que "***se basa en realizar un clustering previo con un subconjunto de variables completas e imputar los valores perdidos de las variables restantes con las medias condicionales de los grupos descubiertos."*** [Gibert, Sanchez 2016] es una opción válida para el óptimo tratamiento de valores faltantes dentro de una matriz de datos. Este método no es complejo y proporciona una matriz completa de datos para trabajar en poco tiempo, las clases determinan grupos de objetos con características comunes y relaciones entre múltiples variables, que son tomadas en cuenta al momento de decidir el valor de la imputación. Al tomar en cuenta las medias condicionales de cada grupo para realizar la imputación, este método no implica reducciones drásticas de las varianzas de las

estimaciones después de la imputación, ya que para cada variable se utilizan diferentes valores de imputación [Gibert, 2013]. Además está apoyado por el conocimiento de expertos en el tema, quienes serán los encargados de seleccionar las variables que permitirán realizar el clustering previo.

Es un método que proporciona un buen balance entre precisión / exactitud y complejidad / tiempo, siendo bastante intuitivo y simple de aplicar. Tiene en cuenta las relaciones globales entre las variables para la imputación, en un marco libre de supuestos, siendo adecuada para cualquier aplicación sin hacer suposiciones estadísticas. [Gibert et al., 2016].

El método propiamente dicho consiste en que dada una matriz de datos $X = \{X_1 \ldots X_K\}$, conformada por un número n individuos {1…n}, que se ubicaran en las filas, y que están descritos a través de K variables, numéricas o no, ubicadas en las columnas, cuyas celdas (x_{ik}) contienen el valor tomado para la variable X_k, (k = 1: K) por el individuo i, y siendo $*_k$ la tasa de valores perdidos para la variable X_k, se construye una nueva matriz X', de forma que los valores de la matriz que son faltantes sean imputados realizando los siguientes pasos descritos en [Gibert, 2013]:

1. Establecer una δ como la tasa permitida de datos faltantes en el paso 1 del proceso de imputación (el δ debe ser muy pequeño).

2. Construir una matriz X_*, seleccionado un subconjunto de variables relevantes para el experto, se espera que el número de variables seleccionadas sea pequeño. Dichas variables no deben contener más de δ% de valores faltantes

3. Utilizando la imputación inteligente para X*, un conjunto de expertos proporcionará valores de imputación de consenso para los datos faltantes en X*, de acuerdo con sus conocimientos previos en el dominio de destino. Al ser δ muy pequeño, los expertos deberán analizar manualmente un conjunto muy reducido de casos por lo que se puede realizar en una breve reunión de dichos expertos y obtener una versión completa de X* que se denotará como X*' con cierta agilidad. El resultado de este paso es un conjunto imputado de datos X*' con las mismas variables que X_* y sin valores faltantes. X*' es consistente con el conocimiento previo del experto sobre el dominio de aplicación.

4. Realizar clustering multivariado utilizando X*', con el objetivo de agrupar en clases a los individuos de la muestra. Debido a que no se conoce el número existente de clusters previamente, es altamente recomendado utilizar clustering jerárquico.

5. Para cada variable en X\X*', calcule las medias condicionales de las variables dadas las clases resultantes en el paso anterior.

6. Construir la matriz X' de datos imputados sustituyendo los datos faltantes de las variables restantes (aquéllas en X \ X*') por la media condicional dada la clase, calculada en el paso anterior.

Luego de completar los pasos descritos, contaremos con un set de datos completo en tiempo reducido, pero cabe recalcar que para ello el paso de la imputación inteligente debe involucrar un conjunto pequeño de datos faltantes. Otro punto a destacar es que al realizar la imputación utilizando clases que agrupan individuos con características comunes, diferentes a las de otra clase se logra que no existan reducciones drásticas de las variancias de los estimadores tras la imputación, ya que para cada variable se utilizan diferentes valores de imputación.

En este proyecto se desarrolla la introducción del método MIMMI en Java-KLASS como una nueva opción para el tratamiento de datos faltantes que amplié la funcionalidad del sistema y aporte alternativas viables a las limitaciones de las opciones disponibles al inicio de este proyecto.

La intervención se desarrollará de forma que todas las opciones disponibles en KLASS en materia de clustering estarán disponibles para el usuario de MIMMI. Este punto es de gran importancia para el éxito de la imputación, ya que se desea como en cualquier proceso de clustering, se anhela agrupar a los individuos con características comunes entre sí, y que se diferencien del resto de grupos [Gordon 1987], ya que mientras mayor sea la similitud de los valores agrupados, y mejor representen el comportamiento de la población que se desea representar, mejor será la imputación de datos faltantes, y mejores resultados se obtendrán de la matriz en los análisis posteriores.

Imputación Múltiple

La imputación múltiple es un enfoque general del problema de los datos faltantes cuyo objetivo es permitir la incertidumbre sobre los datos faltantes creando varios conjuntos de datos imputados plausibles diferentes y combinar adecuadamente los resultados obtenidos de cada uno de ellos. La imputación múltiple tiende a minimizar el sesgo o la pérdida de potencia estadística causada por la pérdida de datos con una distribución aleatoria, a cambio de requerir fuertes hipótesis distribucionales de partida sobre las

variables a imputar, y mecanismos específicos de consenso para la integración de los resultados de todas las réplicas de imputación en un único resultado final.

Método Multiple Imputation by Chained Equations (MICE): O Imputación Múltiple por Ecuaciones Encadenadas, en castellano; es un proceso iterativo que resuelve el tratamiento de valores faltantes dentro de una matriz de datos, imputando los valores en función de un modelo de regresión construido con las variables restantes [Azur et al. 2011] . Al igual que casi todos los métodos de imputación de valores faltantes realiza su implementación y bajo la suposición de que los datos faltantes son Missing At Random (MAR).

Este método construye un número m de matrices de datos imputadas, que da lugar a m estimaciones con sus respectivas varianzas, entonces se combinan las estimaciones realizadas, ejecutando una operación de consenso de las estimaciones para dar como resultado una matriz de datos completa.

El MICE según [Zhou et al., 2014] puede llevarse a cabo mediante una serie ordenada e iterativa de siete pasos, que se describen a continuación:

1. Examinar los patrones de datos faltantes e identificar las variables que van a ser imputadas Decidir la secuencia de la imputación.

2. Inicializar los valores faltantes con un método de imputación simple como sustituir por la media.

3. Construir el modelo de imputación para la primera variable que se ha decidido tratar, se utiliza un modelo de regresión en el que la variable dependiente es la variable que se está imputando, y, las variables independientes, las demás que hay en la base de datos original.

4. Reemplazar los valores faltantes para la variable por las predicciones obtenidas por el método del ajuste de la media predictiva. Para los siguientes modelos de imputación que requieran los valores de la variable que actualmente se está imputando, se toman sus valores observados y los imputados, considerando dicha variable como completa.

5. Repetir los pasos 3 y 4 para cada una de las restantes variables con datos faltantes.

6. Repetir los pasos del 2-5 para obtener tantos conjuntos de datos como número de imputaciones se haya elegido. Esto produce un conjunto de matrices imputadas sobre las que con posterioridad se podrá realizar el análisis deseado.

7. Analizar cada una de las matrices imputadas obtenidas en 6. En este último paso hay que realizar una operación de consenso de los modelos de forma que se entregará un único modelo resultante que representa la agrupación de los obtenidos con cada matriz imputada. En el caso de la regresión es simple, porque bastará con construir un modelo de regresión general que use como coeficientes de la recta de regresión las medias de los coeficientes obtenidos en cada matriz imputada.

Un punto a tener en consideración es la cantidad de iteraciones o ciclos que se deben llevar a cabo para una correcta imputación utilizando este método, y este número depende principalmente de tres factores el tamaño de la matriz de datos, la cantidad de valores faltantes y el sistema informático en el que se esté llevando a cabo el proceso.

El problema más serio de este método es el paso de consenso entre los modelos inducidos con cada matriz de imputación. En realidad, este paso depende del método de modelado y requiere software específico Como ya se ha dicho, la regresión es un caso simple, pero consensuar, por ejemplo, los árboles de decisión que resultan de todas las matrices imputadas no es en absoluto trivial.

Otro punto a tener muy en cuenta con este método es cuando todas las variables de la matriz de datos presentan valores faltantes, ya que para realizar el paso número 3, se ha de realizar con variables completas, es decir que no presenten faltantes, y al no haber variables que cumplan esta condición, se presenta un serio inconveniente para realizar la imputación con este método, ya que simplemente no podría realizarse.

CAPÍTULO 3

INTRODUCCIÓN A JAVA-KLASS

3.1 Descripción

El sistema KLASS, es un software desarrollado en su primera versión por Karina Gibert como parte de su tesis de licenciatura y posterior tesis doctoral, el cual fue propuesto y diseñado en la Facultad Informática de Barcelona (FIB) el cual comprende un conjunto de herramientas para gestionar y ayudar a los expertos en procesos de Minería de Datos, principalmente orientado a la clasificación automática de dominios poco estructurados [Gibert, 1991].

El sistema en su primera versión fue desarrollado en LISP y su ejecución se realizaba sobre UNIX, para luego re-escribirlo en Java, ya que presentaba algunas ventajas sobre su lenguaje de desarrollo original como eliminación de costos de licencia para desarrollar sobre LISP, la portabilidad ya que un sistema informático desarrollado en Java puede ejecutarse sobre cualquier plataforma, la posibilidad de distribuir un software ejecutable independiente de su código fuente, entre otros

Este sistema desde su aparición se ha mantenido en constante crecimiento, incorporando nuevas opciones y funcionalidades, al mismo tiempo que depurando y mejorando sus capacidades, esto en gran parte a trabajos de Máster, proyectos del grupo de investigación, tesis Doctorales, o de trabajos en varias asignaturas tanto en los grados de estadística como de ingeniería en informática en la Universidad Politécnica de Cataluña (UPC) o la Universidad Illes Balears (UIB).

3.2 Funcionalidades de Java-KLASS

En este apartado se presenta un listado de las funcionalidades que ofrece Java-KLASS hasta su más reciente versión.

- Representación de matrices de datos, variables cuantitativas, cualitativas, semánticas y manejo de metadatos.
- Selección de variables e individuos basada en criterios definidos por el usuario para generar submatrices basado en muestreo aleatorio.
- Recodificación o discretización de variables y generación de variables nuevas.
- Gestión de bases de conocimiento.
- Gestión y visualización de ontologías.
- Gestión y visualización de termómetros.
- Estadística descriptiva extensa univariante, bivariante y trivariante de los datos y de distribuciones condicionadas (CPGs)
- Visualización 3D.
- Análisis dinámico [Gibert et al., 2010a].
- Cálculo de distancias con métricas de distintas familias como: Euclídea, Métrica del valor absoluto, Minkovski, mixta de Gibert [Gibert et al., 2005], ralambondrainy, Gower, Gowda-Diday, Ichino-Yaguchi, mixta de Gibert generalizada, Chi-cuadrado, Hamming generalizado, s seda-based distance.
- Clustering automático con métodos jerárquicos clásicos, basados en reglas, en ontologías, con métodos basados en densidades como BDSCAN o OPTICS [Molla Santiago, 2014] o métodos escalables como CURE.
- Interpretación de clases vía TLP, IRBBP [Gibert et al., 2012b, Gibert et al., 2013] [Gibert and Conti, 2016, Gibert et al., 2008a] y conceptualmente CCEC
- Evaluación de bases de conocimiento (BC)
- Métodos de interoperabilidad.
- Gestión de Sistemas heterogéneos que incluye información numérica, cualitativa, semántica, BV, ontologías.

3.3 Cronología

A continuación pequeño resumen gráfico de cómo ha ido avanzando Java-KLASS en el tiempo hasta la última versión

- Feb. 1991 **KLASS v0**. Tesina Karina Gibert. “KLASS. Estudi d'un sistema d'ajuda al tractament estadstic de grans bases de dades". Clasificación de matrices de datos heterogéneas con la distancia mixta de Gibert [Gibert, 1991].
- Nov. 1994 **KLASS v1**. Tesis Karina Gibert. “L'us de la informacio simbolica en l'automatitzacio del tractament estadstic de dominis poc estructurats". Es una ampliación de KLASS v0 que incorpora la clasificación basada en reglas [Gibert, 1995].
- Jul. 1996 **KLASS v1.1**. PFC Xavier Castillejo. Incorpora a **KLASS.v1** una interfaz de ventanas independiente con un sistema que facilita el uso de KLASS desde SUN y desde PC a usuarios que desconocen Lisp y UNIX. Llamaremos **xcn.KLASS** al nLcleo Lisp de esta nueva versión y xcn.i en la interfaz C[Castillejo, 1996].
- Oct. 1997 **jj.KLASS**. PFC Juan José Márquez y Juan Carlos Martín. Incorporan a la versión KLASS.v1 nuevas opciones para el tratamiento de datos faltantes, la posibilidad de trabajar con objetos ponderados e implementan un test no paramétrico de comparación de clasificaciones [Márquez and Martín, 1997].
- Sep. 1999 **KLASS v1.2**. PFC Xavier Tubau (versión Beta). Incorpora a la versión **xcn.KLASS** el módulo de comparación de clasificaciones de **jj.KLASS**, la métrica mixta y Ralambondrainy y prepara la formulación de tres más para su posterior implementación. Llamaremos xt.KLASS al núcleo Lisp de esta nueva versión y xt.i en la interfaz C asociada [Tubau, 1999].
- 1999-2000 **KLASS + v1**. PFC Silvia Bayona. Fusión definitiva de la versión **xt.KLASS** con **jj.KLASS**. Incorpora además un módulo de análisis descriptivo de los datos, también de las clases resultantes, reorientando KLASS hacia un propósito más general y menos especializado. Llamaremos **sbh.KLASS** al núcleo Lisp de esta nueva versión y sbh.i en la interfaz C asociada [Bayona, 2000].
- 2000-2002 **KLASS + v2.** PFC Josep Oliveras. Añade a sbh.KLASS las métricas mixtas pendientes (Gower, Gowda-Diday y Ichino-Yaguchi). Llamaremos joc.KLASS a esta nueva versión.
- 2000-2003 **jr.KLASS +.** Tesis doctoral Jorge Rodas. Integra **KLASS + v.2** y Columbus, que se introduce más adelante [Rodas-Osollo, 2004].

- 2000-2003 Investigación Anna Salvador y Fernando Vázquez. Desarrollo de **CIA-DEC**, que se introduce más adelante [Gibert and Salvador, 2000] [Vazquez and Gibert, 2002].

- 2002-2003 **Java-KLASS v0**. PFC Ma. del Mar Colillas. Versión Java del módulo de análisis descriptivo e integración con CIADEC y Columbus.

- 2003-2005 **Java-KLASS v0.22**. Colaboración con Mar Colillas. Ampliación del análisis descriptivo e introducción de herramientas de gestión de datos (definición de ordenaciones en los informes, posibilidad de varias matrices de objetos en el sistema simultáneamente, cambio de matriz activa).

- 2005-2006 **Java-KLASS v1.0**. Colaboración con Mar Colillas. Incluye lectura y visualización de dendrograma aislados, así como la generación de particiones a partir de ellos.

- 2006-2007 **Java-KLASS v2.0**. PFC José Ignacio Mateos. Ampliación de Java-KLASS con un módulo de cálculo de distancias para diferentes tipos de matrices de datos, incluyendo las que combinan información cualitativa y cuantitativa, tratamiento de missing y creación de submatrices.

- 2006-2007 **Java-KLASS v3.0**. PFC Roberto Tuda. Incluye un módulo de clasificación automática por métodos jerárquicos, utilizando todas las distancias implementadas en la v2.0 y una opción para estudiar agregaciones de objetos paso a paso. Se crea la opción de poder seleccionar el directorio de trabajo predeterminado. Se le agrega la opción de añadir y guardar objetos con peso.

- 2006-2007 **Java-KLASS v4.0.** PFC Laia Riera Guerra. Introducción, gestión y evaluación de bases de conocimiento, ampliación de Java-KLASS con un módulo de trasformación de variables que permite desratización, recodificación y cálculos aritméticos con variables numéricas. Por último, esta versión incluye la definición de submatrices vía filtros lógicos sobre los objetos, la edición de metainformación de las variables de la matriz, eliminación de variables e importación de archivos en formato .dat estandar.

- 2.007 **Java-KLASS v5.0**. PFC Andreu Raya. Incluye la clasificación condicionada, la clasificación basada en reglas y funcionalidades de división de la base de Datos y de gestión de árboles de clasificación (o dendrograma) asociados a las diferentes matrices de datos.

- 2007 **Java-KLASS v6.0** Trabajo de investigación tutelada Alejandro Garcia. Clasificación basada en reglas exógenas. Intenacionalitzacion y localización de a tres idiomas (Catalan, Inglés y Castellano). Fusión de matrices.

- 2008 **Java-KLASS v6.4**. Trabajo de Master Alfonso Bosch Sansa, Patricia García Giménez, Ismael Sayyad Hernando. Boxplot-based discretization, Boxplot-based Induction rules.

- 2008 Tesis doctoral Alejandra Perez. Caracterización por condicionamientos sucesivos, metodología que induce automáticamente a conceptos asociados a las clases descubiertas.

- 2008 Tesis doctoral Gustavo Rodríguez. Clasificación basada en reglas para estados que permite análisis de sistemas dinámicos.

- 2008: **Java-KLASS v7.0**.: TRT Alejandro Garca Rudolph. Fusion de matrices y gestion de variables activas.

- 2009: **Java-KLASS v8**.: Tesis de m ster de Ester Lozano. Criterios Best Local Concept and no close world Assumption del CCECS. PT Alejandro Garca Rudolph. Clasicacion basada en reglas para estados.

- 2010: **Java-KLASS V8.1**.: Práctica SISPD. Narcis Maragall. Boxplot Based Induction Rules

- 2012: **Java-KLASS v8.6**.: Práctica SISPD. Pau. metodología CCECS.

- 2012: **Java-KLASS v9**.: Práctica SISPD. Marco Villegas. Criterios CCECS.

- 2013 **Java-KLASS v10**.: Práctica SISPD. Emili Boronat. Trac Light Panel.

- 2014: **Java-KLASS v11**.: Proyecto nal de Carrera Ingeniera Informática FIB. Sheila Molla. DBSCAN, OPTICS, 3D Visualization.

- 2014: **Java-KLASS v12**.: Práctica SISPD. Jonathan Moreno. Optimizacion de expresiones logicas.

- 2015 **Java-KLASSv15**.: Practicas IKPDI + SISPD Sergio Santamaría y Daniel Gibert y otros practicas Gestión de ONTOLOGIAS, distancias semánticas. Clasificación basada en ontologías.

- 2016 **Java-KLASSv16**.: TFG Valerio Di Matteo (U. La Sapienza, Roma, Italia). Muestreo y Escalabilidad: Generación de variables aleatorias, extracción de muestras aleatorias sobre la matriz de datos, k-Nearest Neighbour, CURE.

- Jun 2016 **Java-KLASSv17.:** TM David Canudes + practicas IKPD des2015: Gestión termómetros + automatización de TLPs.

- Nov 2016 **Java-KLASSv18**.: prácticas IKPD: Implementación de TLPs anotados. Primeras infraestructuras para gestionar variables multivaluadas (desarrollo y concatenaciones)

- Mar 2018 **Java-KLASSv18.2**.: TM Luis Daniel Pérez Tamayo: Gestión de variables multivaluadas y consolidación trabajo anterior.
- May 2018 **Java-KLASSv18.3**.: TM Carlos Luis Jordán y TM Johnny Avila: termómetros cualitativos y conexión con semáforos, y reorganización de todos los métodos de inducción de conceptos.
- Abr 2019 **Java-KLASSv20**.: TM Juan de los Reyes: Implementación del método de imputación de datos faltantes MIMMI.

A continuación se muestra un esquema grafico de la cronología descrita previamente (figura No.1 y figura No. 2)

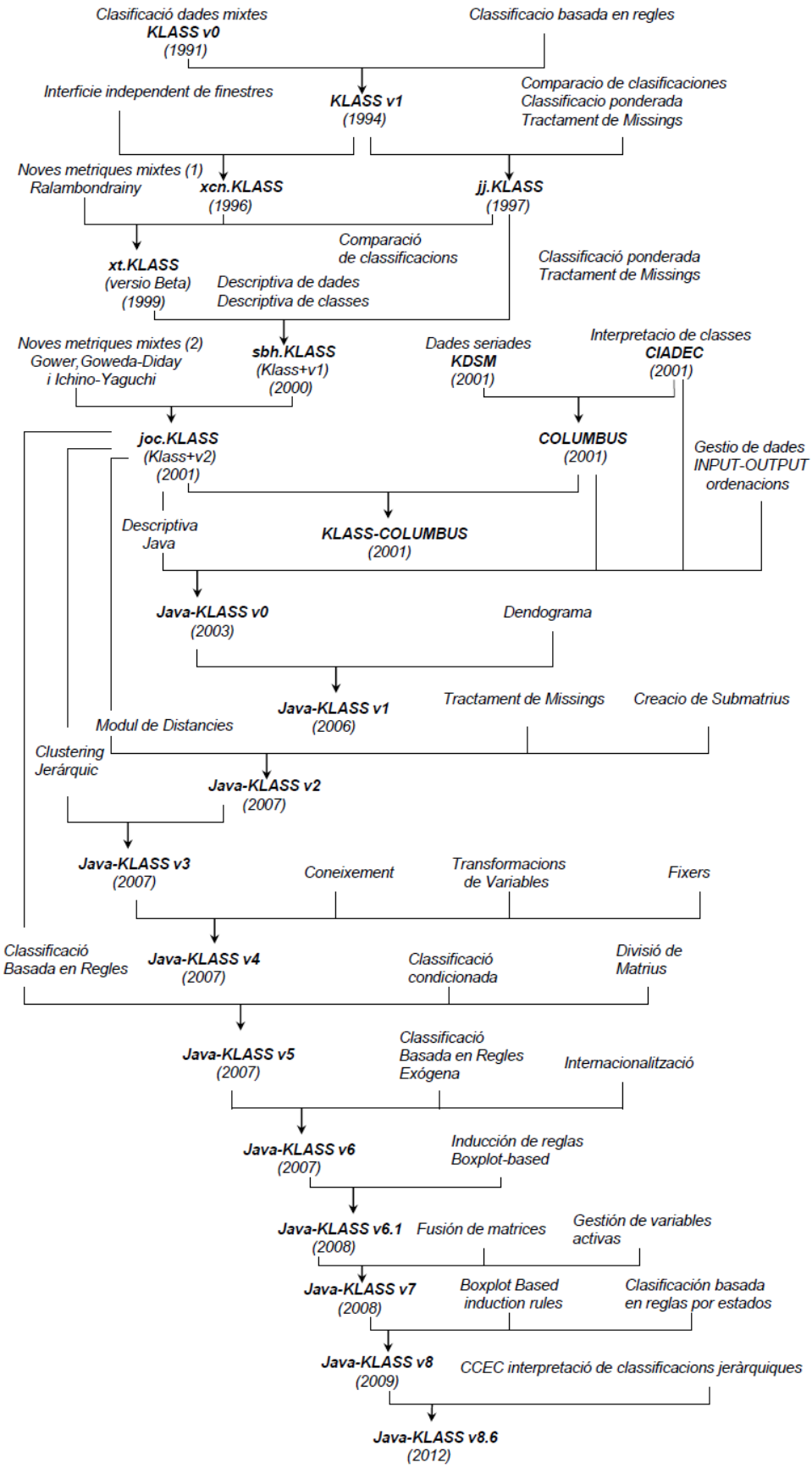


Figura No.1: Cronología de Java-KLASS parte 1

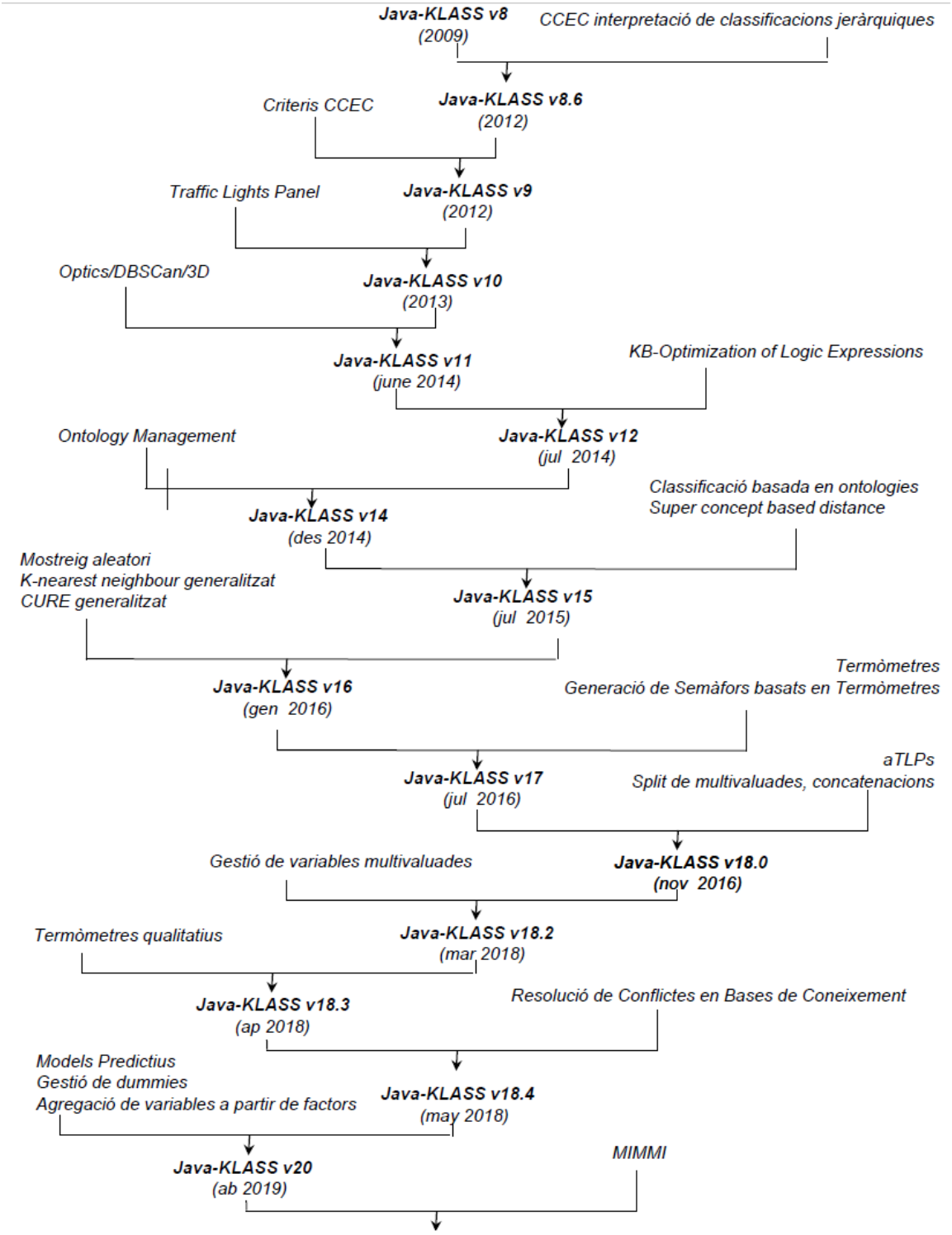


Figura No.2: Cronología de Java-KLASS parte 2

3.4 Java-KLASS y el Tratamiento de Valores Faltantes

Al momento cuenta con dos métodos de tratamiento de faltantes, el primero es la imputación de valores faltantes de las variables numéricas por cero, y el otro método es la imputación de valores faltantes en variables numéricas por la media global de la variable. Aunque son mecanismos útiles en situaciones concretas y tiene pleno sentido su inclusión en un sistema como Java-KLASS, ninguno de los dos métodos son realmente adecuados para el tratamiento de faltantes en el caso general como ya vimos en el **Capítulo 2**. Además de todo esto no se cuenta con un método que resuelva el problema para el caso de valores faltantes para variables categóricas, aunque esta no representa una limitación crítica puesto que es fácil imputarlos datos faltantes en una variable cualitativa por una modalidad "UNKNOWN" operación para la que KLASS da actualmente soporte.

3.4.1 Tratamiento de Valores Faltantes Variables Numéricas

Previo al desarrollo de este trabajo, el sistema Java-KLASS, que se encontraba en su versión 19, contaba con 2 métodos de tratamiento de faltantes para variables numéricas:

La substitución de valores faltantes numéricos por el número cero: Esta opción si bien es sumamente sencilla y permite continuar de una forma rápida con las siguientes fases en el análisis de datos con el software, solucionando el problema que presentan la mayoría de los métodos de clasificación al no poder trabajar con datos faltantes, es sin duda poco fundamentada en principios estadísticos, ya que introduce valores sin considerar absolutamente ninguna característica de la variable, por lo que además de introducir sesgo en la misma, puede alterar relaciones existentes y dificultar la obtención de resultados validos en el análisis posterior de datos que han sido sometidos a este método, por ello sin lugar a dudas es poco recomendada al momento de tratar datos faltantes.

La Substitución de los Valores Faltantes por la Media Global de la Variable: Este método como pudimos apreciar al describirlo en el capítulo anterior a pesar de no demandar mayores recursos computacionales, de poder realizarse de forma rápida y sencilla y de ser ampliamente difundido, no es un método que sea de gran eficacia al momento de remplazar datos faltantes con eficiencia, ya que no se considera un procedimiento apropiado. En su aplicación se asume que los datos faltantes siguen un

patrón MCAR, y ha sido ampliamente documentado que su aplicación afecta la distribución de probabilidad de la variable imputada, atenúa la correlación con el resto de las variables y subestima la varianza.

3.4.2 Tratamiento de Valores Faltantes Variables Categóricas

El tratamiento de datos faltantes para variables categóricas es un punto que no estaba atendido de ninguna forma por el sistema Java-KLASS, puesto que no contaba con método explícito alguno para hacer frente a esta circunstancia, que al igual que para los datos numéricos se presenta frecuentemente en las matrices de datos, y los cuales son de igual importancia y relevancia a la hora de caracterizar una población objetivo, y no pueden ser dejadas de la do al momento de ser analizadas. Hasta el momento se podía utilizar la funcionalidad del Recode para recodificar los valores cualitativos faltantes por una nueva modalidad “NS/NC” para poder hacer frente a esta situación.

CAPÍTULO 4

DESARROLLO DEL PROYECTO

4.1 Descripción del Problema

Como se ha descrito la aparición de datos faltantes es un inconveniente que aparece en casi todos las matrices de datos al momento de realizar análisis de los mismo, sobre todo cuando se trabaja con valores recogidos del mundo real, causando un sin número de inconvenientes al momento de querer extraer información válida y relevante de dicha matriz, por lo que toma especial interés contar con un método adecuado de imputación.

Este problema de la aparición de datos faltantes afecta también al software Java-KLASS**,** por lo que dotar de un método adecuado de tratamiento de faltantes a este sistema se vuelve fundamental a la hora de poder extraer de él toda su funcionalidad y coadyuvar en que los resultados de los análisis realizados con el mismo se realicen de la mejor manera.

Se ha visto en el **capítulo 2** existen muchos métodos de tratamiento de faltantes, entre los que hemos descrito algunos de los más difundidos y utilizados, no todos son pertinentes ni adecuados para tratar valores faltantes en todos los casos.

Lo que se busca es implementar este método en el sistema Java-KLASS, para que el mismo pueda imputar datos de una matriz que se haya cargado en el software, en decir que dada

una matriz de datos **X**, la cual puede contener variables numéricas, variables categóricas o una mezcla de los dos tipos de variables, representadas por **k**, la cual presenta $*_k$ tasa de valores perdidos para la variable $\mathbf{X}_k$. Se deben imputar los datos faltantes para llegar a una matriz sin valores faltantes, la cual sea una buena representación del conjunto que se está intentando representar, que no afecte o que afecte en el menor grado posibles las relaciones entre variables y a los resultados de los análisis que se vayan a realizar con estos datos posterior a la imputación.

4.2 Especificaciones

El proyecto se centra en la introducción de un nuevo módulo en KLASS que implemente el método MIMMI para el caso general sobre cualquier matriz de datos que se pueda cargar en el sistema.

Esto debido a que es un método fácil de usar, fácil de comprender, incluye conocimiento del experto, balanceado entre dificultad, tiempo, esfuerzo, etc. y es un método de propósito general con un buen compromiso entre la precisión y el costo computacional

Para ello será necesario tener en cuenta los siguientes aspectos:

1) Toda la intervención garantizará en todo momento que todas las demás funcionalidades de KLASS se preservan, y funcionan con normalidad para la matriz de datos imputada resultante.
Para la fase 1 del proceso MIMMI Se ha de permitir seleccionar al usuario de entre las variables de la matriz a tratar, aquéllas que considera son predictivas relevantes, y con ellas se construirá una matriz $\mathbf{X}_{*}$**'**.

2) Debido a que Java-KLASS presenta ventajas al momento de clasificar, permitiendo realizar la misma con variables que presentan valores faltantes, lo cual es viable utilizando la distancia de Gower, se pueden relajar las condiciones de la fase 1 del método descrito en (Gibert, 2013) y permitir que en X*' existan datos faltantes.

3) En caso de no querer utilizar la distancia de Gower en la clasificación, deberá configurar la aplicación para que muestre solo las variables que no tienen valores faltantes.

4) Si las variables que el usuario quiere utilizar para X* para realizar la clasificación auxiliar, es responsabilidad del usuario realizar la imputación de las variables.

5) Bajo estas premisas, no es necesario fijar un valor de δ.

6) Por ello, en este proyecto no se incluirán herramientas para realizar la imputación inteligente basada en expertos, puesto que si se han elegido variables con faltantes para conformar X*', es posible realizar el proceso basado en la distancia de Gower para la clasificación.

7) La fase 1 de MIMMI, consistente en existe métodos que soportan dicha elección, por lo tanto el conjunto X*', será automáticamente el conjunto de datos X', con el que se trabajara en el siguiente paso.

8) La clasificación auxiliar multivariada de la matriz X*', descansará en el módulo de clustering de Java-KLASS, ya existente y de notable robustez, el cual cuenta con varios el métodos, criterios y métricas de clasificación, con el objetivo de agrupar en clases a los valores de X*'.

9) En este proyecto, y de acuerdo con la definición de MIMMI dada en (Gibert 2013) se asume que la fase 1 se resuelve con un método de clustering de la familia jerárquica. Debido a que no se conoce a priori el número apropiado de clusters, se ha de dotar al sistema de una forma de visualización adecuada del proceso de clasificación de la fase 1 que facilite al usuario elegir el número de clases resultantes de la clasificación a la vista del dendrograma.

Este paso conectará automáticamente con las siguientes fases del algoritmo MIMMI, se garantizará que si en el futuro se amplía la funcionalidad de KLASS en materia de métodos de clustering, MIMMI incluirá estas ampliaciones de forma automática, sin necesidad de ajustar su código. Para ello se descansará en los paneles propios del módulo de clustering, el cual descansa a su vez en los paneles del módulo de cálculo de distancias. Una vez definidas las clases se ha de realizar el cálculo de las medias condicionadas a las clases por cada variable numérica en X\X*', e imputarlos valores faltantes por los valores calculados. Se implementarán las clases y métodos necesarios para cubrir esta parte

10) En el caso de los faltantes para variables categóricas se optará por una política que incluya las ventajas de MIMMI y que permita también obtener un valor representativo por cada clase, para realizar la imputación de este tipo de propiedades.

11) Se ofrecerán al usuario opciones en cuanto a los niveles de transparencia del sistema, permitiendo la visualización de los valores imputados o no, según preferencias del usuario en cada momento.

12) Una vez se haya realizado la imputación se debe mostrar un informe que permita reflejar los por menores de la ejecución. Se diseñará cuidadosamente la estructura de dicho informe.

13) En todos los pasos donde se requieran funcionalidades que Java-KLASS ya contempla, se reutilizarán los métodos correspondientes incluyendo en el código a desarrollar llamadas a los mismos convenientemente parametrizadas.

14) En los casos que se requiera, se introducirán objetos y clases Java nuevos siguiendo la filosofía general del sistema.

15) Se respetará la arquitectura de fachada que presenta Java-Klass.

16) Se ampliará el manual de usuario con las nuevas funcionalidades

4.3 Trabajo Realizado

4.3.1 Desarrollo Funcionalidad Método MIMMI

Implementación de MIMMI en Java-KLASS: Ahora bien, para dotar de la funcionalidad necesaria al software Java-KLASS del método de imputación de faltantes MIMMI, se va a realizar la implementación desarrollando la funcionalidades necesarias para su implementación, pero además, y como ya se ha dicho, se ha de aprovechar y reutilizar todas los procedimientos y características con las que ya cuenta Java-KLASS, y así lograr implementar MIMMI en el sistema manteniendo y potenciando su funcionalidad actual.

En esta sección se describen las acciones que se desarrollan para la imputación de datos faltantes a través de MIMMI, en el apartado siguiente se describirá la interface y el proceso que el usuario debe seguir para realizar la imputación, por temas didácticos nos apoyaremos de los seis pasos que se describen en [Gibert, 2013] como los que se deben seguir para realizar la imputación con MIMMI, y que fueron descritos en el capítulo 2 de este documento. Ahora bien, ya que el método se está implementando en un sistema que tiene mucho tiempo en constante desarrollo y depuración, y que por lo tanto es robusto y cuenta con mucha funcionalidad y conocimiento ya implementado.

Paso 1. En este paso se va a permitir al usuario del sistema seleccionar entre realizar la selección de las variables relevantes que conforman la submatriz X_*, de entre variables que no contienen datos faltantes o si desea incluir también variables que tienen datos faltantes. Esto ya que como aclaramos previamente, la robustez de Java-KLASS nos permite realizar la clasificación para el clustering a través de métricas que soportan la presencia de faltantes como es el caso del criterio Ward tomando la distancia de Gower. Para el caso de permitir datos faltantes en X* solamente la distancia de Gower será aceptada para la clasificación de la Fase1.

Por este motivo es que se ha desarrollado funcionalidad que permite al usuario elegir si desea visualizar solo variables completas o todas las variables de la matriz al momento de seleccionar las mismas para la clasificación auxiliar, así el usuario estará consciente de que va a realizar una clasificación con variables que presentan faltantes o no, además en caso de seleccionar variables que presenten valores faltantes se implementaron alertas que en todo momento avisarán al experto, que se trata de variables incompletas y de cuántos valores faltantes presentan exactamente, para que se cuente siempre con toda la información y se pueda tomar decisiones de manera informada.

Otra de las ventajas con las que cuenta Java-KLASS, la cual es muy importante y que es aprovechada para la implementación de este métodos, es que permite realizar la clasificación utilizando distancias para variables numéricas, mixtas, categóricas y semánticas, lo que se traduce en poder realizar la clasificación de la Fase1 sin importar si se cuenta solo con variables numéricas, solo categóricas o una mezcla de las dos; lo que es una gran ventaja, ya que el experto cuenta con una gran libertad a la hora de seleccionar las variables que conformarán X*, con ello vienen implícitas las ventajas de contar con una clasificación que incluye sin importar el tipo, cualquier variable que se considere relevante para realizar la clasificación.

Paso 2. No se desarrolla dentro del sistema los procesos que pueden llevar a realizar la imputación inteligente con los expertos, puesto que si se han elegido variables con valores faltantes para conformar la submatriz, existen métodos que soportan dicha elección, y en cualquier caso este es un proceso basado en el experto que se puede afrontar con anterioridad a cargar los datos en Java-KLASS. De ser el caso que solo se elijan variables completas para dicha submatriz, es obvio que tampoco será necesaria la imputación inteligente, por lo tanto el conjunto X_*', será automáticamente el conjunto de datos X', con el que se trabajara en el siguiente paso.

Paso 3. Se realizará la clasificación de la fase 1 multivariada utilizando X*, se elegirá por parte del usuario de entre varios el método, criterio y métrica de clasificación, la más adecuada para los datos que se están tratando, con el objetivo de realizar una clasificación y agrupamiento apropiado para conseguir el mejor resultado en la imputación.

Se presenta el resultado de la clasificación de la Fase 1 en forma de un dendrograma, conectando con dando la funcionalidad necesaria (ya existente en Java-Klass) para que el usuario realice el corte del dendrograma en un nivel que considere apropiado para realizar el clustering, y así continuar con los pasos siguientes que deben realizarse para completar el método MIMMI.

Aquí cabe recalcar que para la ejecución del método se recomienda tomar la tercer o cuarto mejor corte, con más grupos y más homogéneo que el agrupamiento óptimo [Gibert, 2013].

Paso 4. Siguiendo la definición original de MIMMI de [Gibert, 2013] para cada variable en X\X*', se calculará la media condicionales de las variables de clase, según selecciones el usuario. La media es un buen representante de la tendencia central de una clase cuando las variables tienen una distribución normal, o cuando menos, una distribución muy simétrica. Sin embargo, en el caso que exista mucha asimetría en los datos será mejor la utilización de la mediana, que se utiliza para representar la tendencia central en este tipo de distribuciones. Por esta razón, el proyecto aporta una contribución original a la implementación de MIMMI y se amplía un poco la visión de (Gibert 2013), permitiendo al usuario elegir entre la media o la mediana como representante de clase

En este punto cabe recalcar, que en casos en los que un cluster especifico tenga para una variable solo datos faltantes, lo que imposibilita el cálculo de la media o mediana de esa variable para esa clase, se tomará el valor de la media o mediana global, esto ya que este valor puede representar la tendencia general de la variable.

Ahora bien, llegados a este punto es importante definir cómo se va a tratar con las variables cualitativas, ya que hasta ahora solo hemos hablado de cómo imputar los valores faltantes en variables numéricas. Para el caso de las cualitativas, se va a remplazar los valores faltantes también agrupando los individuos por clase, pero en lugar de imputar los valores por la media o mediana condicional del grupo, que no existe en variables cualitativas, se va a imputar con una nueva modalidad de la variable cualitativa compuesto por el nombre de la variable, el nombre de la clase y el sufijo ".UNKNOWN" común para todas las variables, así de esta forma también existe un valor distinto de imputación para cada clase, manteniendo la variabilidad de las columnas. Es frecuente codificar el valor

faltante de una variable categórica como una columna más y esto resulta más robusto que utilizar la moda condicionada a la clase como estimador de la tendencia central de la variable en cada clase.

Paso 6. Construir la matriz X de datos imputados sustituyendo los datos faltantes de las variables por los valores obtenidos en el paso anterior.

Aquí cabe recalcar que se ha de proporcionar al usuario la posibilidad de incluir en la matriz original la variable de clase con la que se realizó la clasificación auxiliar descrita en el punto 3, esta variable se posiciona luego de las columnas de la matriz de datos original.

También se permitirá en caso que el usuario así lo requiera, conservar una copia de las variables que fueron imputadas, que conservaran los datos originales de la misma, previo al proceso de imputación. Estas variables se han de crear como nuevas columnas en la matriz de datos con el nombre de la variable original más el subfijo ".AUX" y se posicionan de forma posterior a las columnas de la matriz original.

Paso 7 Elaborar un informe completo de lo que se ha realizado, para que el usuario pueda analizar el resultado del proceso de imputación.

Desarrollo de la Implementación de MIMMI en Java-KLASS

Para poner en práctica todo lo descrito en el apartado anterior, se debe realizar el desarrollo informático que permita implementar toda la funcionalidad definida. Este desarrollo a de asegurar que el nuevo método sirva para imputar correctamente valores faltantes en matrices de datos cargadas en el sistema y que todas las opciones con la que actualmente cuenta Java-KLASS, sigan trabajando de forma óptima, sin presentar inconvenientes o novedades.

Para empezar el desarrollo en Java-KLASS, se ha de tener en cuanta primero lo siguiente:

- El sistema cuenta con un patrón de arquitectura que separa claramente la interfaz gráfica de la lógica del negocio. La interfaz gráfica identificada en los paquetes "UI", ha de manejar solo la interacción con el usuario y pequeñas validaciones que permitan un correcto ingreso y obtención de datos. Por otro lado en el core del sistema "NUCLI", se encuentran las clases que han de realizar la obtención y transformación de los datos, ningún proceso de la logia de negocio puede realizarse fuera de esta área.

- El sistema como ya lo hemos visto en el **Capítulo No. 3**, tiene ya larga data en funcionamiento, y se ha ido mejorando y corrigiendo a lo largo del tiempo, llegando a ser una herramienta muy robusta, por ello durante el desarrollo se debe respetar los protocolos de desarrollo, y tratar de que el código quede los más claro, auto describible y reutilizable posible. Esto ya que como mucha gente trabaja en colaboración para que Java-KLASS continúe su crecimiento y mantenga funcionando de forma correcta, si no se cumplen con estos parámetros se entorpece la labor de otros colaboradores.

Ahora bien, como se ha mencionado esta implementación trata de reutilizar las opciones con las que cuenta Java-KLASS, adaptándolas o perfeccionando su funcionamiento para mantener en todas las opciones una sola implementación de la funcionalidad, esto ya que en caso de que la funcionalidad deba ser actualizada, bastará actualizar un solo código para actualizar todas las partes de la aplicación donde esta funcionalidad se utilice cumpliendo así también los paradigmas de programación orientada a objetos y clean code.

Se realizaron cambios e implementaciones tanto en el core del sistema "nucli" como en la interfaz gráfica "ui" de la aplicación, a continuación una muy breve descripción del desarrollo realizado en cada clase:

Nucli

- **GestorMatriu (Actualización):** Implementación de funcionalidad para el manejo de las estadísticas por clase de cada, Gestion de opciones y parámetros "Reporte MIMMI", Imputación de datos faltantes con MIMMI, Actualización estadísticas variables categóricas y numéricas, funcionalidad creación de variables auxiliares para copia de columnas imputadas, entre otras muchas que permiten todo el proceso de imputación de faltantes.
- **GeneradorText (Actualización):** Se actualizo funcionalidad para permitir gestión y visualización Informe MIMMI.
- **GeneradorTextMIMMI(Creación):** Elaboración dinámica de código LaTex, para la generación del informe MIMMI (texto, tablas y gráficos), en base a los parámetros seleccionados por el usuario previamente.
- **StatisticsByClass(Creación):** Calculo de las estadísticas por variable y clase (descriptiva por clase).
- **StatisticsByClassMIMMI(Creación):** Calculo de estadísticas por variable y clase de las clasificación auxiliar para imputación de datos faltantes.

UI (Funcionalidad Ampliada a detalle en la siguiente sección: ***Descripción de la Interfaz***)

- **DlgOpcMiss (Actualización):** Creación del menú para MIMMI y visualización de opciones
- **DlgOpcMIMMI (Creación):** Pantalla de parametrización de opciones para la imputación.
- **PanelMIMMI (Creación):** Pantalla principal del método MIMMI
- **PanelClasifica(Actualización):** Se actualiza para el soporte de parámetros y métricas propias de MIMMI.
- **DlgOpcDescriptivaMIMMI(Creación):** Opciones de parametrización para construcción del "Informe MIMMI".
- **DlgOpcDendo(Actualización):** Cambios requeridos para soporte de parámetros requeridos en caso de imputación por el método MIMMI.

En el siguiente diagrama de clases se caracteriza lo correspondiente a la implementación realizada:

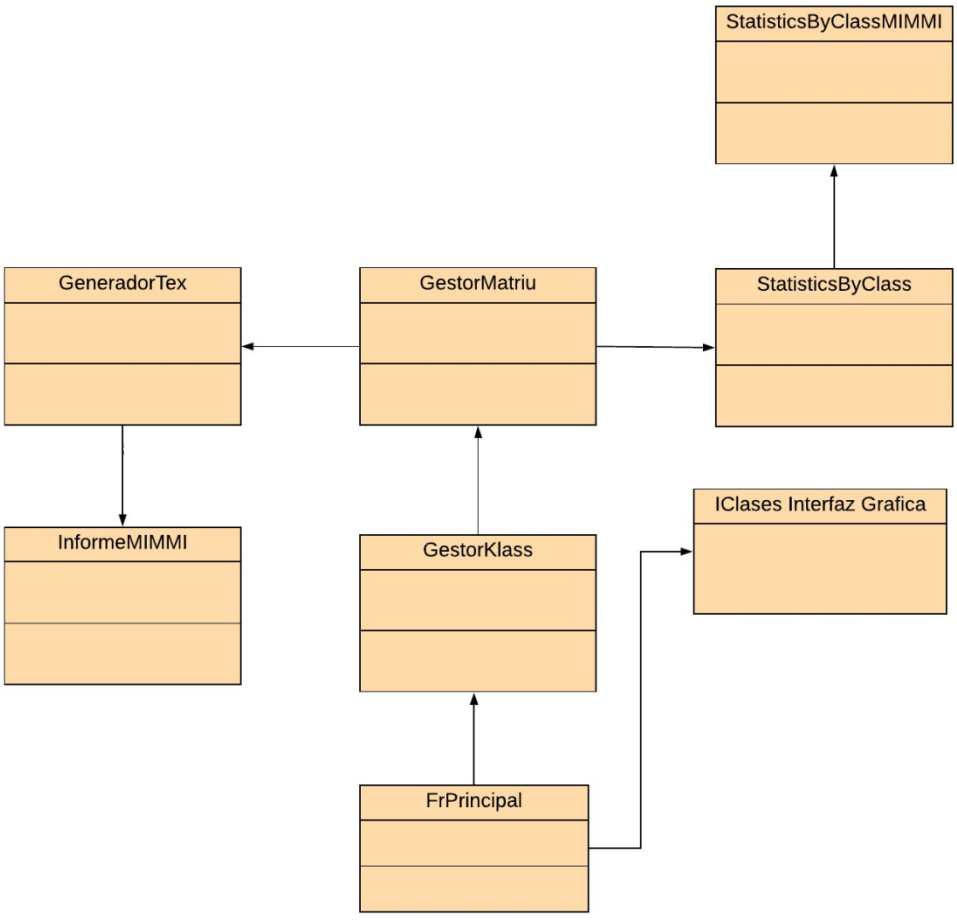


Figura No.3: Diagrama de clases

Descripción de la Interfaz

Menú Tractament de mancants: El primer paso para implementar el método MIMMI (descrito en Gibert 2013) en el sistema Java-KLASS, es agregarlo en el menú de tratamiento de faltantes de la aplicación, en la cual ya contábamos con las opciones de substituir las variables por el valor de cero o por la mediana de la variable, a continuación se muestra la pantalla (Figura No.4) de opciones de tratamiento de faltantes una vez incluido MIMMI:

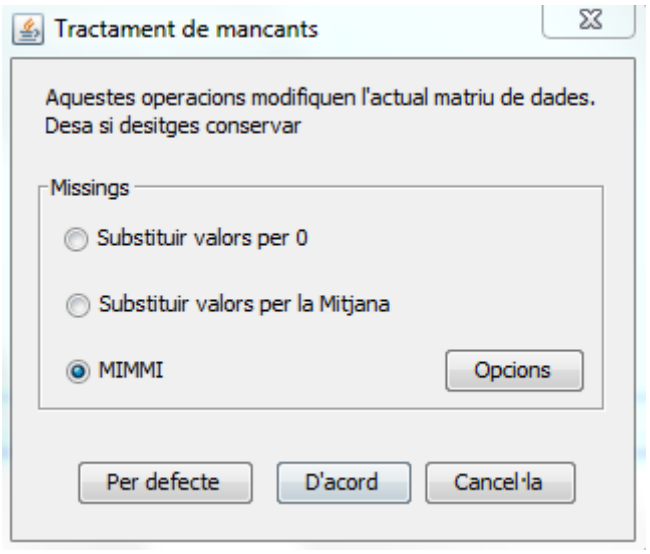


Figura No. 4: Menú Tratamiento de Faltantes Java-KLASS

Como se puede apreciar, junto a la opción de MIMMI en el menú de tratamiento de faltantes se presenta un botón de opciones, estas opciones permiten al usuario configurar las condiciones en que se ejecutará el método MIMMI, que se deben conocer antes de ingresar a la pantalla de ejecución del método propiamente dicha, como son (Figura No.5):

- **Ocultar les variables incompletes: (**Valor por defecto: NO**)**
  - **Si:** Solo se muestran las variables de la matriz de datos a tratar que no tienen valores faltantes en el listado de variables candidatas para realizar la clasificación auxiliar de la Fase1
  - **No:** Se muestran todas las variables de la matriz de datos a tratar, tengan o no missings.

- **Conservar còpia de las variables originals:** (Valor por defecto: SI)
    - **Si:** Se incluye en la matriz original una copia de las variables que fueron imputadas, que conservan sus valores originales en forma de nuevas variables.
    - **No:** Posterior a la imputación, no se conserva copia de las variables cuyos valores faltantes fueron imputados.

- **Valor d'imputació:** (Valor por defecto: Media)
    - **Media:** Se utilizará el valor de la media calculada por cada clase resultante del proceso de clasificación auxiliar para la imputación de los valores faltantes de las variable numéricas este valor es recomendado frente a variables de distribución normal o bastante simétrica
    - **Mediana:** Se utilizará el valor de la mediana calculada por cada clase resultante del proceso de clasificación auxiliar para la imputación de los valores faltantes de las variable numéricas, este valor es recomendado frente a variables de distribución asimétrica

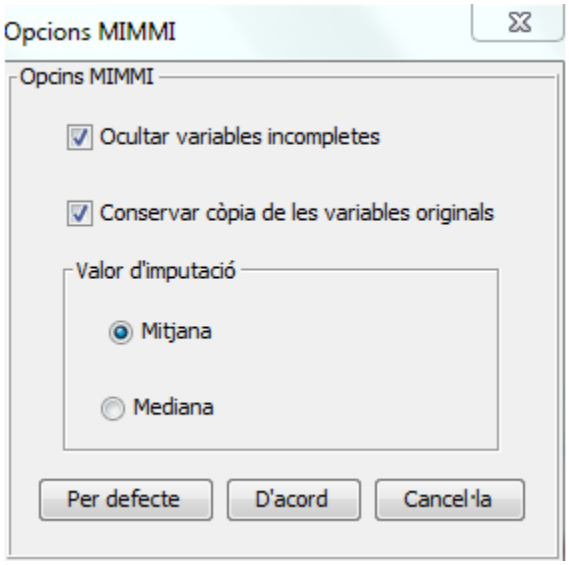


Figura No. 5: Menú opciones método MIMMI

- **Botones:**
    - **Per defecte:** Si se pulsa este botón, las opciones toman sus valores por defecto.
    - **D'acord:** Al pulsar este el botón se guardan los valores el usuario ha seleccionado y se regresa a la pantalla "Menú Tractament de mancants".
    - **Cancel.la:** Esta opción permite abandonar el menú de "opcions MIMMI" sin guardar los cambios realizados en los parámetros de este menú.

Tras configurar los parámetros de ejecución de MIMMI, aparece la pantalla **MIMMI** automáticamente, la cual se describe a continuación

Panel MIMMI

Este panel que permite completar la configuración del método MIMMI y realizar la imputación. A continuación (figura No. 6) se muestra un ejemplo de la pantalla tal como se presenta, previo a la selección por parte del usuario de las variables para realizar la clasificación de la Fase 1 del método.

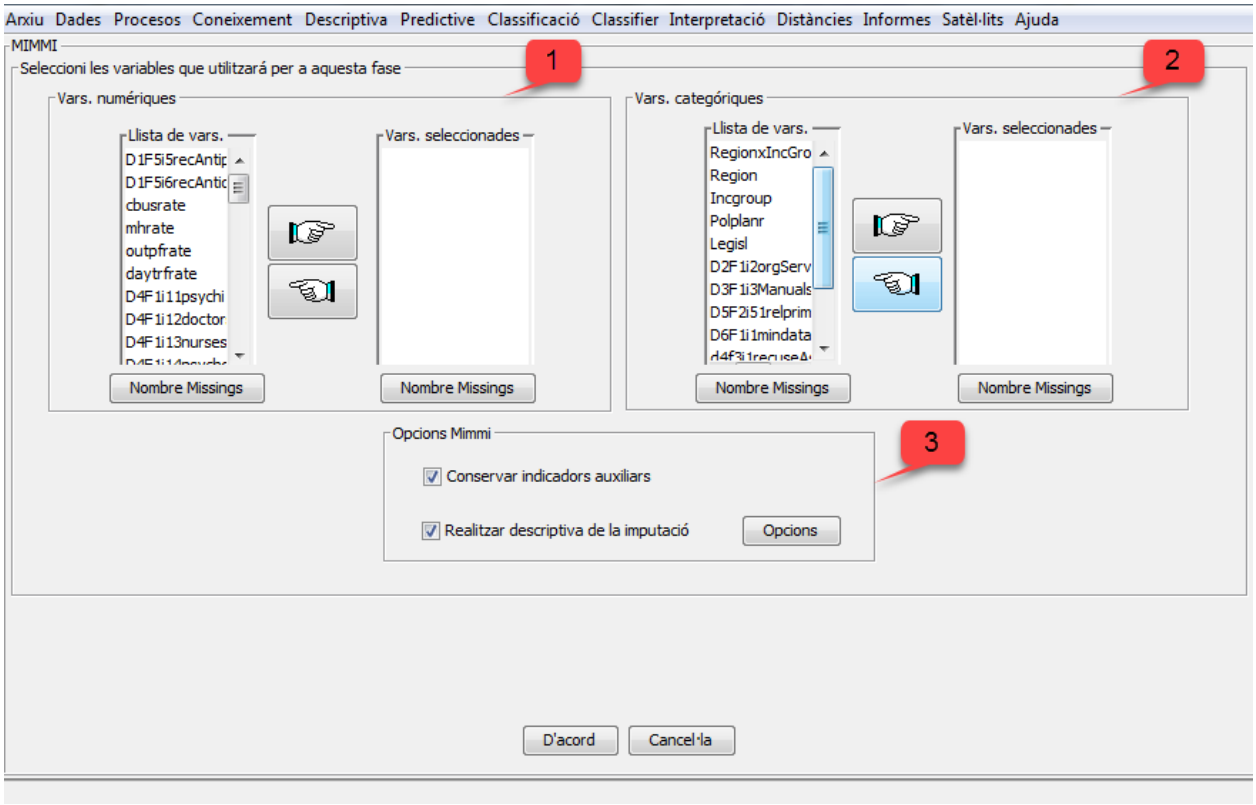


Figura No. 6: Pantalla método MIMMI Java-KLASS

El panel contiene dos secciones principales:

Sección 1 Selección de variables: El usuario puede seleccionar de aquí las variables que configurarán la matriz X* para la clasificación de la fase 1 de MIMMI. A su vez, se distinguen dos subsecciones en este apartado

- **Variables numériques (etiqueta 1 de la figura No. 6):** Variables numéricas elegibles para realizar el clustering de la Fase 1 de MIMMI (figura 6).
  Si en la pantalla anterior el usuario ha determinado "Ocultar variables incompletas" se mostraran solo variables sin valores faltantes. En caso contrario, se mostraran todas las variables numéricas de la matriz que se está tratando. La lista de la derecha muestra las variables que ya han sido seleccionadas por el usuario, por ello al iniciar la lista se presenta vacía. De acuerdo con el funcionamiento de todos los paneles de selección de variables de Java-KLASS, para seleccionar una variable debemos dar clic en la misma en la lista de la izquierda y pasarla utilizando el botón con la flecha hacia la lista derecha. Por el contrario si deseamos excluir una variable de la selección seleccionaremos la variable y utilizaremos el botón "Flecha hacia la izquierda".

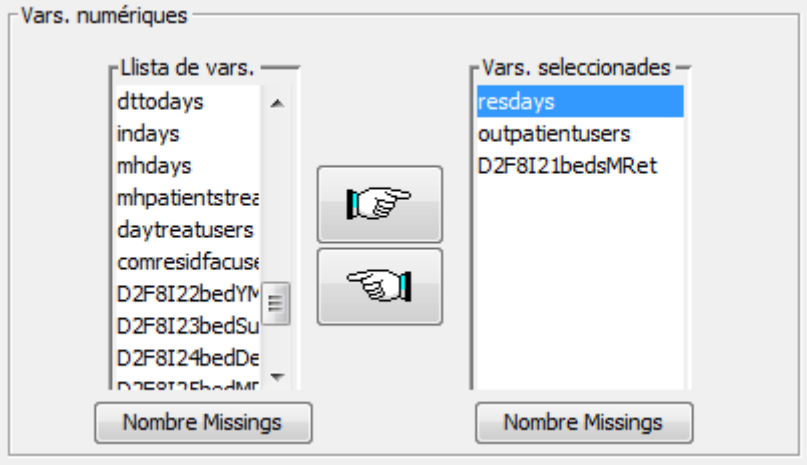


Figura No. 7: Variables numéricas

- **Botón Nombre Missings:** al pincharlo, permite ver el número de faltantes que tiene una variable de la lista que se encuentre sobre el botón. Para ello la variable ha de estar seleccionada y el resultado se muestra a través de un pop-up como el de la figura No. 8:

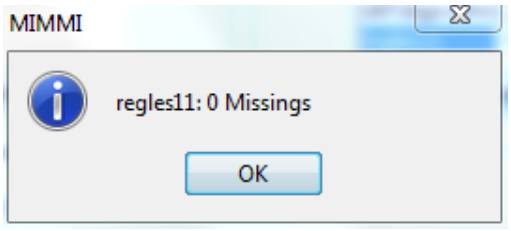


Figura No. 8: Mensaje número de missings

Si se seleccionan variables que contienen faltantes aparecerá un mensaje (pop-up) informándonos que la variable tiene valores perdidos, en número de los mismos y preguntándonos si estamos seguros de seleccionar la variable (figura No. 9).

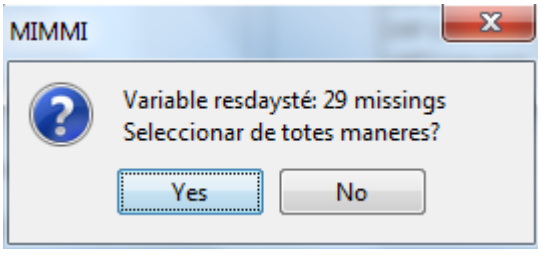


Figura No. 9: Mensaje confirmación inclusión variable con missings

- **Variables Categòriques (etiqueta 2 figura No. 6):** Variables numéricas elegibles para realizar el clustering de la Fase 1 de MIMMI (figura No. 10).

  Puedan ser incluidas, al igual que las numéricas dependiendo de la selección que haya realizado el usuario en las opciones de MIMMI, se han de presentar solo variables completas o todas las variables categóricas.

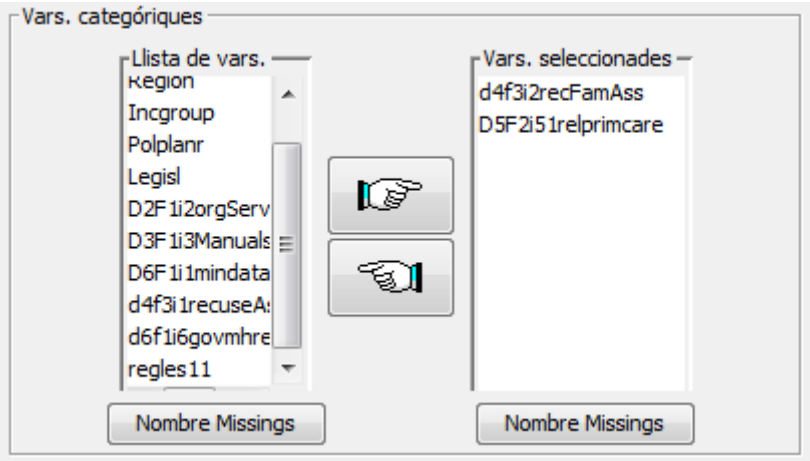


Figura No. 10: Variables categóricas

El funcionamiento de esta sección es idéntico al de la sección número 1, salvo que en esta se trata de variables categóricas, por lo que no nos extendemos en más detalles.

Sección 2 Opciones MIMMI (etiqueta 3 de la figura No. 6): Aquí se puede seleccionar las siguientes opciones (figura No. 11):

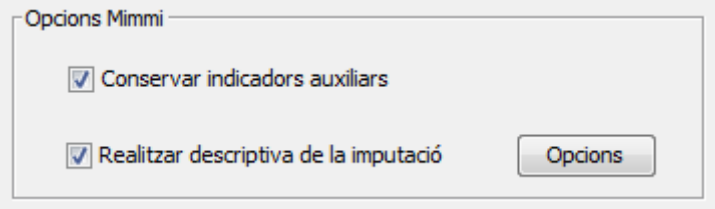


Figura No. 11: Opciones MIMMI

- **Conservar indicadors auxiliars (**figura No. 10**):** (defecto: Si)
  - Si: conserva la variable de clase que se obtiene del clustering de la Fase1 en la matriz de datos a tratar
  - No: la variable de clase se utilizará internamente para el proceso de imputación de faltantes, pero no será visible para el usuario

- **Realitzar descriptiva de la imputació** (figura No. 11): (defecto SI)
  - **Si:** genera el Informe MIMMI, con los datos correspondientes a la ejecución de la imputación de los valores faltantes, el cual se almacena en la carpeta "resultats", cuya ubicación se ha definido por el usuario en la configuración de Java-KLASS con el nombre de la matriz de datos más el subfijo "MIMMI".
  - **No:** Al finalizar el proceso de imputación no se crea ni visualiza el informe MIMMI correspondiente a la ejecución de imputación de los valores.
  - **Opcions** (figura No 12): al pulsar este botón aparece una ventana pop-up, en la cual se puede parametrizar el número máximo de variables numéricas y categóricas que se van a mostrar en las tablas de la descriptiva comparativa del Informe MIMMI.

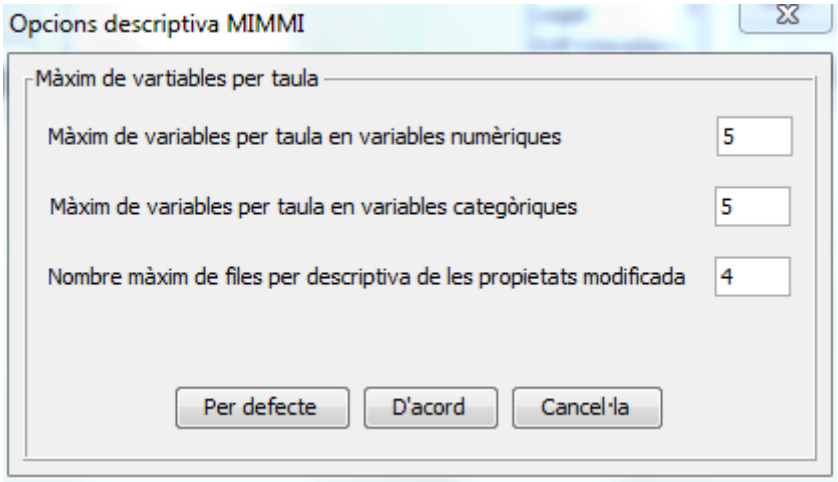


Figura No. 12: Opciones descriptiva MIMMI

Una vez se ha seleccionado las variables con las que se desea realizar la clasificación, si se pulsa el botón "D'acord" de la pantalla principal de MIMMI, y si se cumplen las validaciones de por lo menos haber agregado una variable para la clasificación, aparece la pantalla de opciones para la clasificación. Al igual que la pantalla principal de MIMMI, esta pantalla se ha dividido en tres secciones para facilitar su explicación, y es la siguiente:

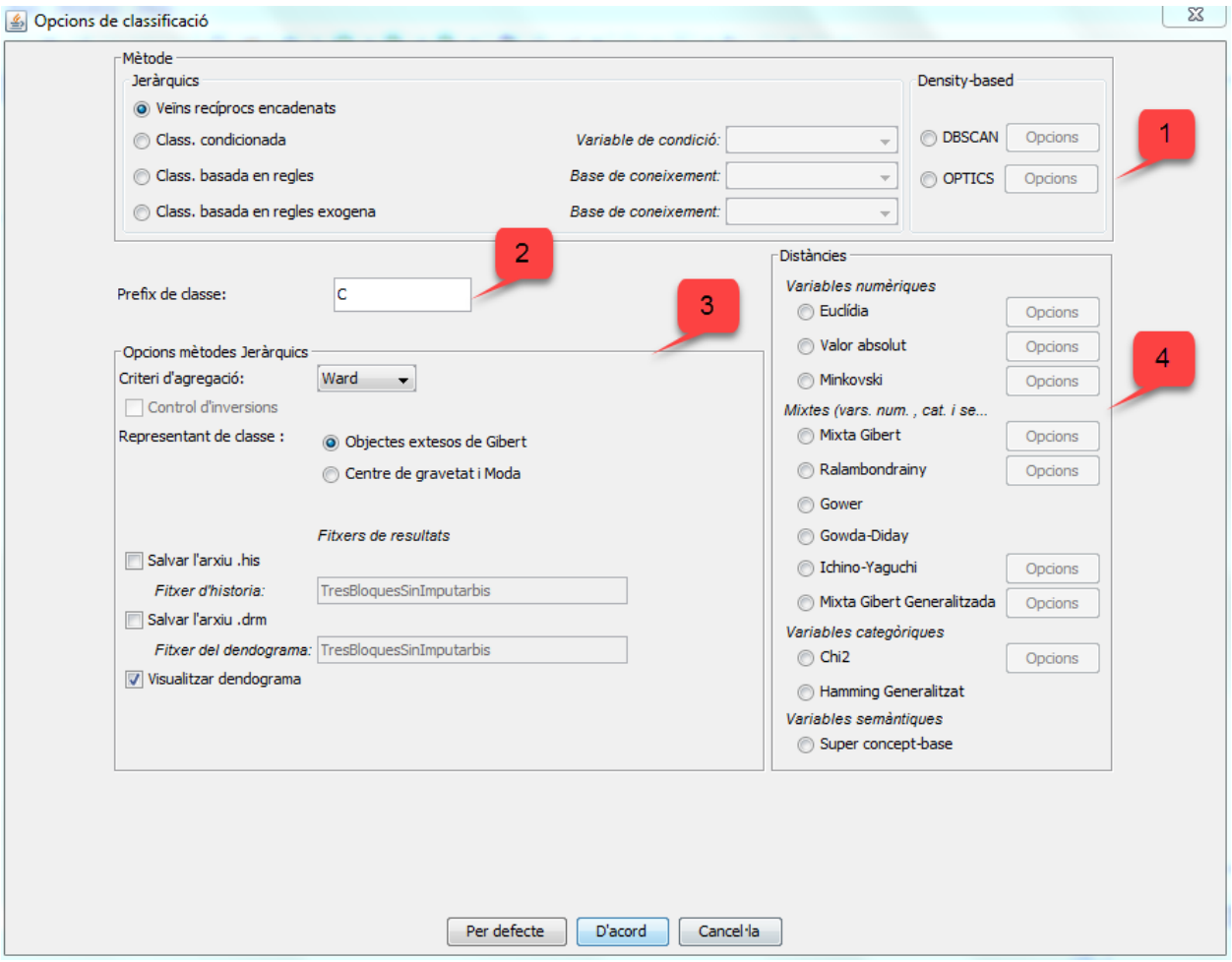


Figura No. 13: Opciones Clasificación

Sección 1 Método (etiqueta 1 de la figura No. 13): En esta sección se podrá seleccionar entre algunos métodos para realizar la clasificación con las variables seleccionadas en la pantalla anterior, cabe recalcar que depende mucho de los datos con los que contemos, para seleccionar el método que se utilice para la clasificación. Principalmente se dividen en Jerárquicos y basados en densidad

Jeràrquics: Como su nombre indica, construyen una jerarquía de agrupamientos, uniendo o dividiendo los grupos de acuerdo a una cierta función de similaridad/disimilaridad entre los grupos, que dan como resultado un “arbol” de clusters llamado dendrograma [Pacual, 2010]. Este tipo de método de aglomerativos y divisivos, Se consideran agrupamientos aglomerativos cuando comienzan con grupos unitarios y recursivamente une dos a más clusters con características comunes, por otro lado se consideran agrupamientos divisivos, cuando comienzan con un solo cluster, en el que se encuentran todos los individuos, y

recursivamente va dividiendo los individuos en grupos según sus características. Entre las ventajas de los algoritmos de agrupamiento jerárquicos se puede mencionar la flexibilidad con respecto al nivel de granularidad, son fáciles de manejar y son aplicables a cualquier tipo de atributo [Pacual, 2010].

Los métodos jerárquicos con los que cuenta Java-KLASS son:

- Veins recíprocs encadenats
- Clasificación condicionada
- Clasificación basada en reglas
- Clasificación basada en reglas exógena

Density-based: Estos algoritmos se basan en detectar áreas en las que existe alta densidad de puntos y que están separadas por áreas vacías o con escasa presencia de los mismos. Estos puntos son representaciones en un área espacial de los individuos que están caracterizados en la matriz de datos [Campello et al., 2013].

Los métodos basados en densidad con los que cuenta JavaKLASS son:

- DBSCAN
- OPTICS

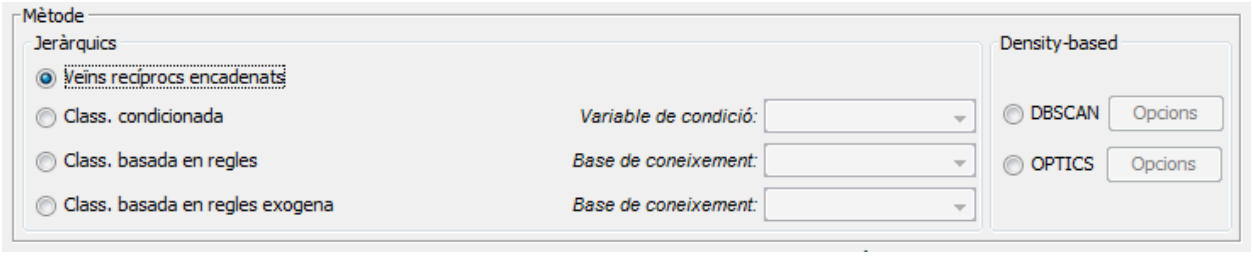


Figura No. 14: Métodos de Clasificación

Sección 2 Prefix de clase (etiqueta 2 de la figura No. 13): Este campo nos permite indicar que prefijo tendrán las clases que se obtengan de la clasificación, podemos colocar un valor que nos ayude a distinguir los resultados de la clarificación que estamos realizando. El valor por defecto es “C”.

Figura No. 15: Prefix de classe

Sección 3 Opcions mètodes Jeràrquics (etiqueta 3 de la figura No. 13): En caso de seleccionar un método jerárquico para la clasificación, se puede parametrizar el criterio de agregación, la representación de clase, entre otras opciones.

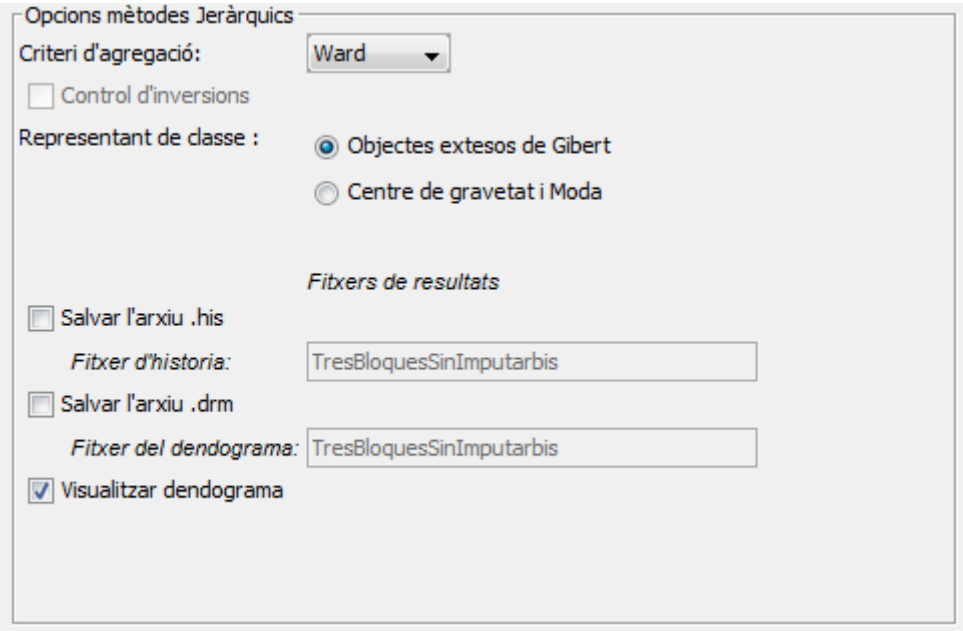


Figura No. 16: Opcions mètodes Jeràrquics

Distàncies (etiqueta 4 de la figura No. 13): En el ámbito de la estadística y el análisis de datos, las distancias son medidas simétricas no negativas que nos permiten cuantificar que tan símiles son los individuos comparados, a medida que la distancia sea más grande, los individuos serán menos símiles, y a medida que la distancia sea menor su similitud crecerá.

Las distancias clasificadas por el tipo de variables en las que se pueden utilizar, con los que cuenta Java-KLASS para determinar la similitud entre individuos son las que se ven en el panel, y se describen en el manual de KLASS (Manual d'usuari Java-KLASS 2018):

D'Acord: el sistema en base a todas las selecciones realiza la clasificación de la fase 1 del Método MIMMI; y dado que el método recomienda utilizar clustering jerárquico, el resultado se muestra en la pantalla de forma automática.

Figure 1: CAJ. Arbre general de classificació

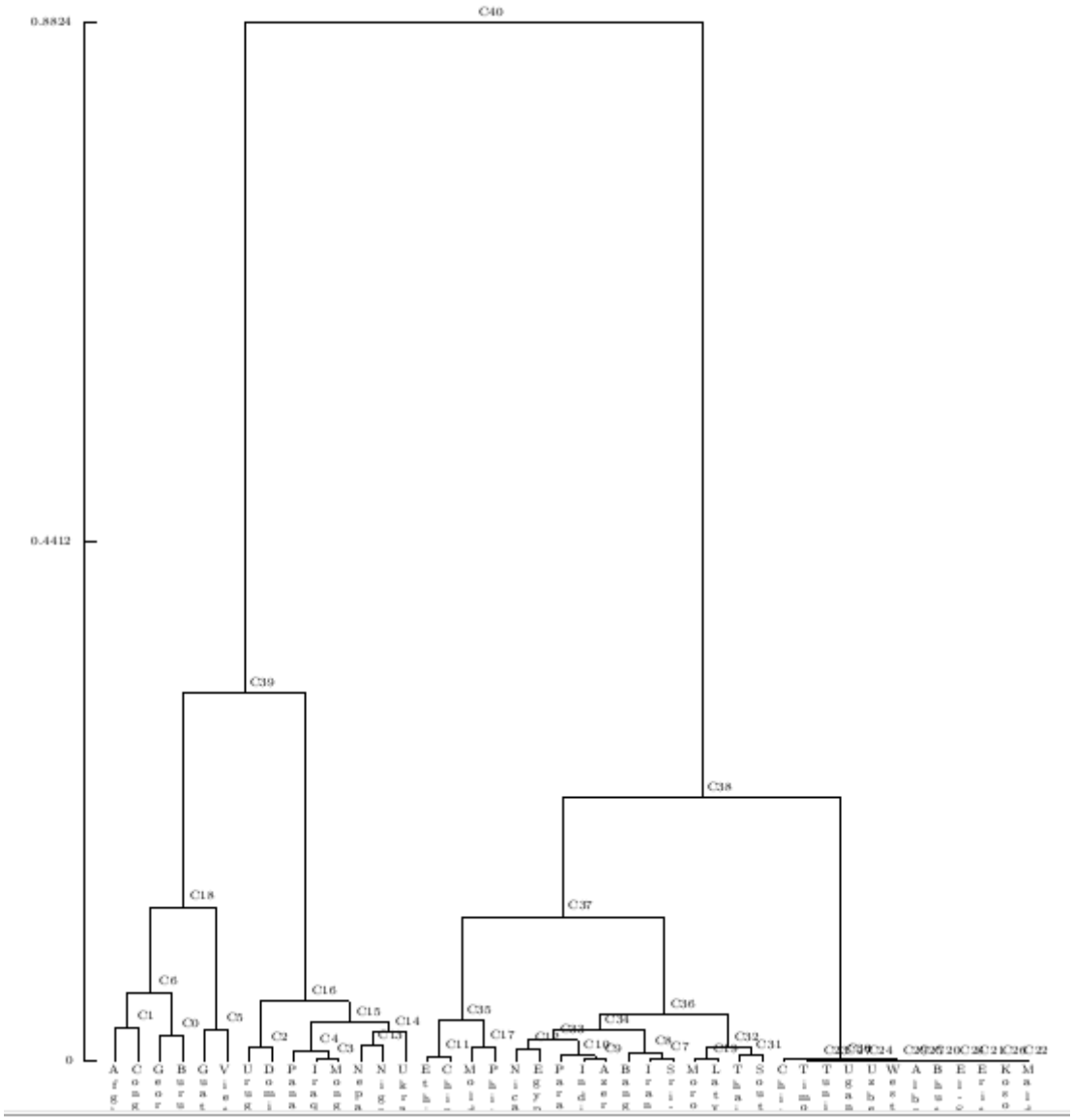


Figura No. 17: Dendrograma

El experto deberá analizar el dendrograma que se ha generado como resultado de la clasificación, y en base a su conocimiento realizar la tala del mismo, para ello posterior a la apreciación del árbol, se visualizará de forma automática la pantalla pop-up ***Opciones de***

classificació (figura No. 18), que permite realizar la tala de acuerdo con el comportamiento habitual de Java-KLASS para los dendrogramas.

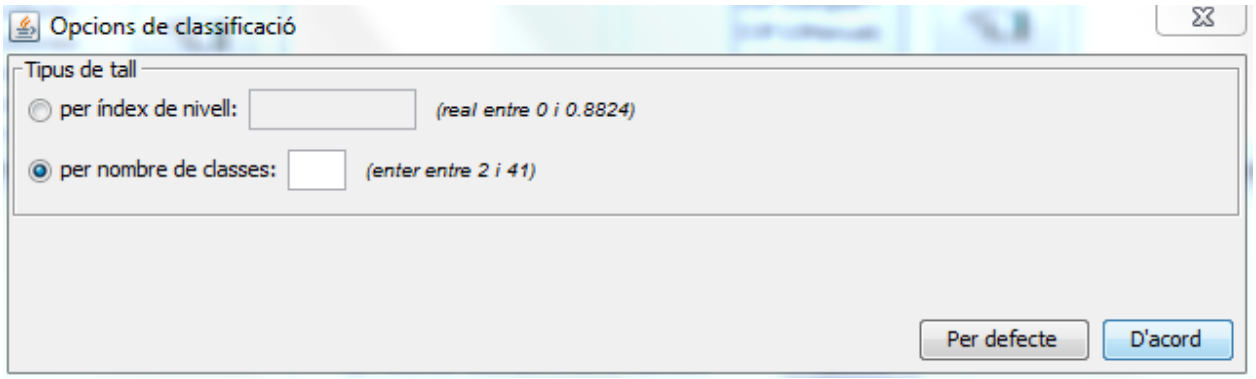


Figura No. 18: Opciones de clasificación

D'acord: realiza el corte del árbol y obtiene una variable de clase. Se crean los clusters, se calculan medias o medianas condicionadas a las clases para las variables numéricas que no han intervenido en la Fase1 (según configuración especificada por el usuario), se etiquetan los valores faltantes de las categóricas que no han intervenido en la Fase 1 según las clases. Posterior al cálculo de la media o mediana por clase, se utiliza estos valores para sustituir los valores faltantes por los valores de imputación calculados y se genera una matriz de datos imputada.

Si "*Conservar indicadors auxiliars*" estaba activado, se incluirá luego de las columnas de las variables de la matriz de datos, una nueva columna con los valores de la variable de clase para cada individuo de la matriz.

Si "*Conservar còpia de las variables originals"* estaba activado, las variables seleccionadas originalment para la fase 1, incluyendo sus valores faltantes, permanecerán en la matriz de datos como columnas nuevas en la parte final de la matriz, presentadas luego de la columna de variable de clase, y sufijadas por AUX. En su posición original se presentarán debidamente imputadas con las medias/medianas condicionadas a las clases.

Per defecte: Al pulsar este botón se coloca el valor por defecto para el número de clases, que servirá para realizar la tala del árbol.

Informe MIMMI

El informe se ordena siguiendo los siguientes parámetros de diseñó para que se pueda evidenciar todas las particularidades del proceso de imputación de datos faltantes que se ha realizado con Java-KLASS utilizando el método MIMMI. Este informe está dividido en dos fases, la Fase 1 corresponde a la clasificación auxiliar y la Fase 2 que corresponde al proceso de imputación de datos faltantes.

La Fase 1 del informe empieza listando las variables que se utilizaron para realizar la clasificación auxiliar, ya que es importante conocer exactamente las variables que nos han permitido realizar la clasificación auxiliar. Además de conocer el listado de variables es importante que se pueda visualizar las características de cada una de estas variables, por ello a continuación del listado se muestra una descriptiva de las variables utilizadas, esta descriptiva varía entre las variables de tipo numérico y las de tipo categórico, puesto que sus características son distintas y los métodos que consideramos adecuados para visualizar sus propiedades más importantes difieren.

Para el caso de las variables numéricas se consideró que mostrar el histograma, el boxplot y la tabla de estadísticos sumarios, permitirán visualizar la distribución de los datos de cada variable de forma diáfana y comprensible.

En el caso de las variables categóricas se muestra el diagrama de barras y la tabla de frecuencias de cada variable, esto al igual que en caso de las variables numéricas permiten evidenciar las características de la variable de forma clara.

Una vez se conoce las variables con las que se realizó la clasificación auxiliar y sus características, hemos de conocer los parámetros seleccionados por el usuario para realizar dicha clasificación, ya que esto es el otro dato fundamental a la hora de realizar este método de imputación. Se ha de presentar una tabla con todas las particularidades con las que se realizó esta clasificación, no solo para tener presente las características de la clasificación que nos permitirán entender y conocer algo más del proceso de clasificación sino también estos datos nos deben permitir en caso de ser necesario replicar dicha clasificación de forma exacta.

Para terminar la fase de clasificación auxiliar se muestra el resultado de la misma a través de una descriptiva de la variable de clase resultante del proceso de clasificación auxiliar. Al

igual que con otras variables categóricas se mostrará el diagrama de barras y la tabla de frecuencias de esta variable resultante, con ello será más fácil visualizar el resultado del proceso.

La segunda fase, ha de mostrar el resultado de la imputación de tal manera que quien lo visualice tenga una clara idea del proceso realizado y del resultado del mismo, para ello se debe mostrar los valores con los cuales se han imputado los valores faltantes por cada variable y por cada clase de la clasificación auxiliar. Además de quedar claro los valores que se utilizaron para realizar la imputación es fundamental saber cuántos valores faltantes fueron imputados por cada valor de sustitución, esto debe mostrarse tanto para variables numéricas como para categóricas.

Ahora bien, para finalizar el informe se muestra una descriptiva comparada de las variables imputadas antes y después del proceso. En esta comparativa debe ser fácil observar las características de las variables antes y después de la imputación; para ello se muestra al igual que en descriptivas anteriores de este informe los histogramas, boxplots y estadísticas sumarias para variables numéricas y tabla de frecuencias y diagrama de barras para las variables categóricas, de forma que se visualice el antes y el después uno junto al otro y facilite la comparación.

Este informe se presentará en caso de seleccionar "*Realitzar descriptiva de la imputació"*. A continuación se presenta el detalle más técnico sobre la estructura de este informe

La estructura del informe está dividida en 2 partes:

Parte 1: Classificació Auxiliar: hace referencia a la clasificación auxiliar de la Fase 1 del método. En este apartado, se presentan los siguientes elementos:

1) **Variables seleccionades per la fase de Clasificació:** En esta sección se listan las variables que se utilizaron pata la clasificación auxiliar. A manera de ejemplo se presenta la primera página del informe (figura 19) que muestra también esta sección

Informe MIMMI

Fase 1: Classificació Auxiliar

Variables seleccionades per la fase de Classificació auxiliar

- D1F5i5recAntipshic
- D1F5i6recAntidepr

Anàlisi descriptiva de les variables anteriors

Variable D1F5i5recAntipshic

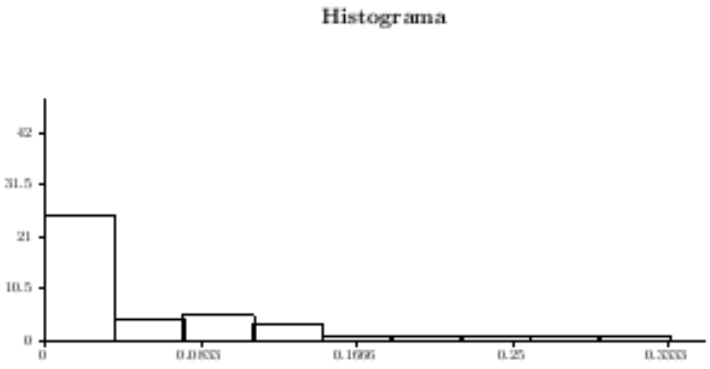


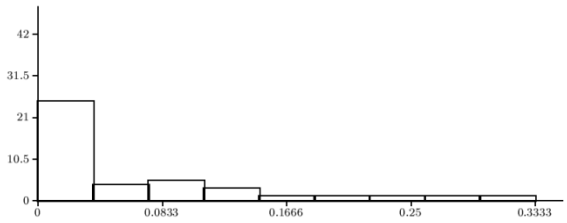
Figura No. 19: Métodos de Clasificación

2) **Anàlisi descriptiva de les variables anteriors:** En esta sección se muestra la descriptiva de las variables que se utilizan para la clasificación auxiliar. La descriptiva varía entre variables numéricas y categóricas. Para el caso de las variables categóricas se muestra la tabla de frecuencias y el diagrama de barras de cada una de ellas (ver figura No. 20), y para las variables numéricas se muestra el histograma, boxplot y la estadística sumeria de cada variables (ver figura No. 21).

Variables Categóricas

Modalitats	Freq. absol.	Freq. acum.	Freq. relat.	Freq. rel. acum.
LOW	13	13	0.3095	0.3095
LOWER	24	37	0.5714	0.881
UPPER	5	42	0.119	1
dades mancants	0	N = 42	0	

Tabla No. 1: Tabla de frecuencias



Variables Numéricas

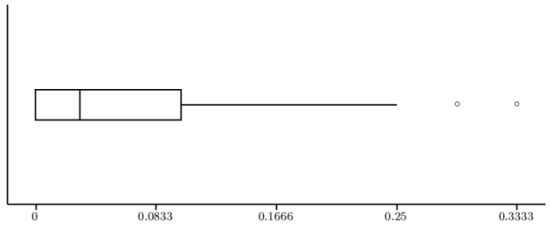


Figura No. 20: Descriptiva de una variable categórica: Tabla de distribución de frecuencias y diagrama de barras

Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	0.0505
Mediana	0.032
Primer quartil (Q1)	0
Tercer quartil (Q3)	0.086
Mínim	0
Màxim	0.179
Quasi-desviació típica	0.0528
Coeficient de variació	1.0336

Figura No. 21: Descriptiva de una variable numérica: Histograma, boxplot y Estadísticos Sumarios

Paràmetres de la Clasificació Auxiliar: En esta tabla se muestra los parámetros seleccionados por el usuario para realizar el clustering de la fase 1 con las variables

seleccionadas. La Tabla No. 1 muestra todos los datos relevantes y que podrían de ser el caso ayudar a reproducir esta clasificación auxiliar si se desea.

PARÀMETRE	VALOR
Mètode	Vens recíprocs encadenados
Criteri	Ward
Representant de classe	Objectes extesos de Gibert
Prefix de classe	C
Mètrica	Mixta de Gibert
Categoria	Alfa i beta automatics
Alfa	0.016042188
Beta	0.9839578
Nombre de classes resultants	4

Tabla No. 1: Parámetros de la Clasificación Auxiliar

Descriptiva Univariant de la variable de classe: Se muestra una descriptiva de la clase resultado de la clasificación auxiliar de la fase 1, al ser esta una variable categórica, se describe de igual manera que las variables categóricas que se utilizaron para la clasificación, es decir se presenta la tabla de frecuencias y el diagrama de barras de la misma (figura No. 22)

Modalitats	Freq. absol.	Freq. acum.	Freq. relat.	Freq. rel. acum.
C18	6	6	0.1429	0.1429
C30	12	18	0.2857	0.4286
C37	16	34	0.381	0.8095
C16	8	42	0.1905	1
dades mancants	0	N = 42	0	

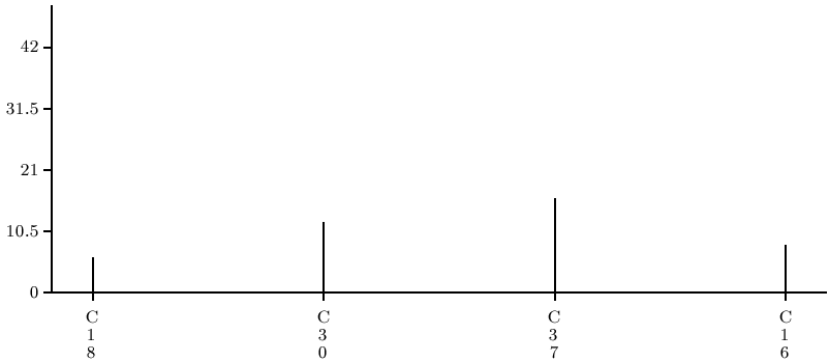


Figura No. 22: Diagrama de barras – Variable de clase

Parte2: Imputación: Detalla el proceso de imputación.

Valors d'imputació per las variables numèriques: En esta sección se muestra en tablas los valores de la media o mediana obtenidos por cada clase, en cada variable que presenta valores faltantes. Con estos valores será con los cuales se va a realizar la sustitución de los datos faltantes. Cabe resaltar que como título se resalta que tipos de medida de tendencia central se utilizó, si la media o la mediana.

Media

idClase	chtreaprev	D4F1I16octher	D4F1I17othW	D2F2I2treatOut
C37	0.0048	0.0892	1.0216	452.277
C36	0.0065	0.139	5.5542	667.1341
C34	0.0037	0.0657	9.254	1220.38
C35	0.0225	0.722	9.2852	1551.6525

Tabla No. 2. : Valors de imputació per las variables numèriques

Valors d'imputació per las variables qualitatives: Como se ha descrito previamente, las variables cualitativas que son imputadas con este método, remplazan los valores faltantes por el nombre de la variables, el nombre de la clase y un subfijo, a continuación un ejemplo:

idClase	Region
C37	RegionC37.UNKNOWN
C36	RegionC36.UNKNOWN
C34	RegionC34.UNKNOWN
C35	RegionC35.UNKNOWN

Tabla No.3. : Valors de imputació per las variables qualitatives

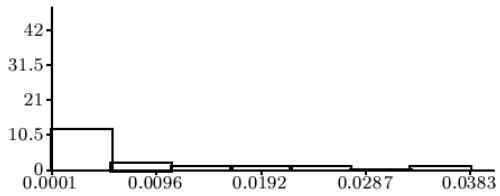
Nombre missings remplazados per clase: esta sección muestra el número de imputaciones que se realizaron por variable y por clase.

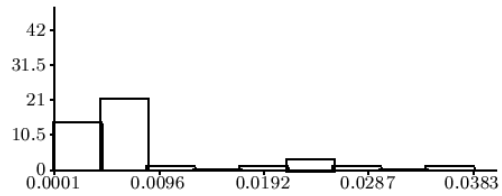
idClase	chtreaprev	d4f3i1recuseAss	d4f3i2recFamAss	D4F1I16octher
C37	4	3	2	0
C36	14	4	5	1
C34	3	1	1	0
C35	3	1	0	0

Tabla No. 4: Nombre missings remplazados per clase

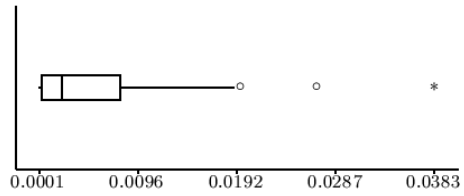
Descriptiva de las Propietats Modificades: En esta sección se realiza una descriptiva de las variables que fueron imputadas, con la particularidad que se presentan para su comparación el resultado de la descriptiva de la propiedad antes y después de ser imputadas. De esta forma se logra que se visualice de forma diáfana la diferencia en la distribución de los datos antes y después de la sustitución de valores faltantes. Al igual que en descriptivas anteriores, se visualiza para variables numéricas el histograma, boxplot y las estadísticas sumarias para cada propiedad y para las propiedades cualitativas, se visualiza el tabla de frecuencias y el diagrama de barras. A continuación se muestra un ejemplo.

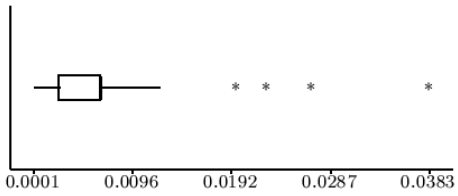
ANTES	DESPUES
chtreaprev	





ANTES	DESPUES
chtreaprev	





ANTES	DESPUES
chtreaprev	

Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	24
Nombre d'observacions útils	18
Mitjana	0.0071
Mediana	0.0023
Primer quartil (Q1)	0.0005
Tercer quartil (Q3)	0.0079
Mínim	0.0001
Màxim	0.0383
Quasi-desviació típica	0.0108
Coeficient de variació	1.4817

Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	0.0075
Mediana	0.0065
Primer quartil (Q1)	0.0026
Tercer quartil (Q3)	0.0065
Mínim	0.0001
Màxim	0.0383
Quasi-desviació típica	0.0082
Coeficient de variació	1.0746

Figura No 23: Descriptiva de las Propietats Modificades - Histograma, boxplot y estadistics sumaris

ANTES	DESPUES
d4f3i1recuseAss	

Taula de freqüències

Modalitats	Freq. absol.	Freq. acum.	Freq. relat.	Freq. rel. acum.
Present	14	14	0.4242	0.4242
Absent	19	33	0.5758	1
dades mancants	9	N = 42	0.2143	

Taula de freqüències

Modalitats	Freq. absol.	Freq. acum.	Freq. relat.	Freq. rel. acum.
Present	14	14	0.3333	0.3333
Absent	19	33	0.4524	0.7857
d4f3i1recuseAssC37.UNKNOWN	4	37	0.0952	0.881
d4f3i1recuseAssC30.UNKNOWN	4	41	0.0952	0.9762
d4f3i1recuseAssC16.UNKNOWN	1	42	0.0238	1
dades mancants	0	N = 42	0	

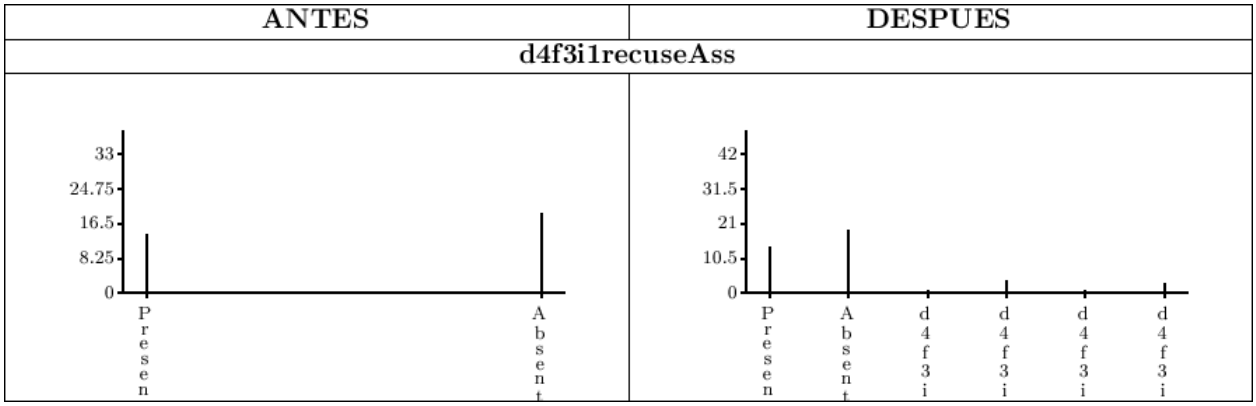


Figura No. 24: Descriptiva de las Propietats Modificades - Taula de freqüencies y diagrama de Barras

CAPÍTULO 5

CASO DE ESTUDIO

5.1 Introducción

A continuación presentaremos el caso de estudio que utilizaremos como referencia en este proyecto

Se utilizarán los datos brindados por la OMS (Organización Mundial de la Salud), que es el ente rector y coordinador del sistema de salud dentro de las Naciones Unidas. Esta matriz proviene de un estudio realizado en países de bajos o medios recursos o conocidos también como LAMIC (Low and middle income countries), que tiene como objetivo verificar si la depresión de madres jóvenes en estos países, es una causa que contribuye en la tasa de mortalidad de los neonatos y corresponde a una colaboración que la Dra. K. Gibert, directora de este proyecto mantuvo con el departamento de Salud Mental y Abuso de Sustancias de la OMS y para el cuál se diseñó el método MIMMI, que se implementó puntualmente en R para aquella ocasión. La sección 5.3 describe el caso de estudio.

5.2 Experimentación

Se intentará evaluar el impacto de utilizar MIMMI o no para imputar los datos faltantes de una matriz, con respecto al método básico de imputación por la media global de las variables. Con este comparativo lo que se intenta es evidenciar en la práctica las ventajas que presenta el método MIMMI al momento de realizar la imputación de valores faltantes en una matriz de datos reales y corroborar el aporte que se realiza al implementar este método en el sistema Java-KLASS.

El esquema de la experimentación es el siguiente:

1) Carga de los datos originales sin pre-tratar y descriptiva inicial
2) Imputación de valores faltantes por el método de la media global
3) Clustering de los datos imputados por este método y análisis del resultado
4) Imputación de los datos faltantes de la matriz original con el método MIMMII,
    a. análisis de la creación de la matriz de clasificación auxiliar,
    b. análisis del informe final del método
5) Clasificación de los datos bajo los mismos parámetros con los que se realizó el cluster del punto 3 y análisis de los resultados
6) Comparativa justificada

5.3 Descripción datos de la OMS

Como matriz de datos para la experimentación se va a utilizar la proporcionada por "WHO-AIMS v2.2" la cual contiene datos recopilados en 42 países. WHO-AIMS es un instrumento diseñado específicamente por la OMS para evaluar el estado de los sistemas de salud mental en países LAMIC y está compuesto por 22 facetas, 155 ítems y 256 variables, tomando en cuenta los siguientes dominios:

- El marco político y legislativo del país
- Los servicios de salud mental.
- La salud mental en la atención primaria.
- Los recursos humanos.
- La información pública y los vínculos con otros sectores.
- La monitorización de los sistemas de salud mental y la investigación en ese ámbito.

A continuación se clasificarán estos datos utilizando dos preprocesados distintos para la imputación de valores faltantes para después comparar la calidad de los perfiles identificados.

5.4 Clasificación Utilizando Método de imputación de Media Global

5.4.1 Datos originales

El primer paso que vamos a realizar es cargar la matriz de datos descrita en el apartado anterior de este capítulo, en Java-KLASS. Esta matriz cuenta con 256 variables.

Los expertos de la OMS proporcionan una selección de las 19 variables relevantes que se van a tener en cuenta para la clasificación. Son las siguientes:

Variable	**Significado**	**Tipo de Variable**
Incgroup	Nivel de ingreso del país.	Cualitativa
Totprofmh	Número total de profesionales dedicados a la salud mental en el país por cada 100 000 habitantes.	Cuantitativa
Usmhexperca	Gasto en salud mental per cápita en USD.	Cuantitativa
Treatpre	Parte de la población diagnosticada y atendida por cada 100 000 habitantes.	Cuantitativa
Capratiosch	Cobertura del tratamiento de la esquizofrenia.	Cuantitativa
d2f11i1closepsybeds	Camas psiquiátricas ubicadas en o cerca de la ciudad más grande (proporción per cápita).	Cuantitativa
d1f5i2exmhos	Gasto en hospitales mentales (%).	Cuantitativa
d2f6i71mhrec10y	Proporción de pacientes que permanecen en hospitales psiquiátricos durante 10 años o más.	Cuantitativa
Comcarewor	Proporción de usuarios tratados en hospitales mentales.	Cuantitativa
lundpararectrail	Ratio entre consultas externas y días en que el paciente está hospitalizado, indica si el sistema de salud da prioridad a mantener al paciente o ingresarlo lo más pronto posible.	Cuantitativa

D3f1i3Manuals	Disponibilidad de manuales de tratamiento y evaluación en atención primaria.	Cualitativa
Legisl	Presencia de legislación.	Cualitativa – Binaria
Polplanr	Presencia de un plan de salud mental.	Cualitativa – Binaria
d6f1i6govmhrep	Informe sobre salud mental publicado por el departamento de salud del gobierno.	Cualitativa
Cbusrate	Unidades de pacientes internados basadas en la comunidad por 100 000 habitantes.	Cuantitativa
Outpfrate	Instalaciones ambulatorias por 100 000 habitantes.	Cuantitativa
Daytrfrate	Instalaciones de centros de día por 100 000 habitantes.	Cuantitativa
(D5F2i51)relprimcare	Relación de colaboración formal con el departamento de atención primaria.	Cualitativa
Region	Región a la que pertenece el país.	Cualitativa

Tabla No. 5: Variables utilizadas para el caso de estudio (WHO-AIMS v2.2)

Una vez cargadas la variables y seleccionadas las pertinentes para nuestro análisis, procedemos a obtener a través de Java-KLASS, de la descriptiva de las variables seleccionadas, esto para poder visualizar las características de las mismas, entre las elementos que nos interesa conocer está el identificar que variables tienen valores faltantes y el número de ellos por variable. Además esta descripción previa nos permitirá realizar una comparación de las características de las variables que tienen faltantes antes y después de la imputación por el método de imputación por media global de la variable en el siguiente apartado **5.4.2 Imputación de Valores Faltantes Utilizando Método de Media Global,** si se desea ahondar más en la descriptiva realizada se puede consultar la misma en el **Anexo 1** de este documento**.**

En el siguiente resumen (tabla No. 5) se muestra la cantidad de variables con valores faltantes, cuántos? valores son y qué porcentaje representan del total de cada variable?

Variable	Numero de valores faltantes	Porcentaje faltantes sobre el total
Incgroup	9	21,43
Totprofmh	8	19,05
Usmhexperca	8	19,05
Treatpre	8	19,05
Capratiosch	9	21,43
d2f11i1closepsybeds	7	16,67
d1f5i2exmhos	9	21,43
d2f6i71mhrec10y	9	21,43
Comcarewor	9	21,43
Lundpararectrail	9	21,43
Region	8	19,05

Tabla No. 6: Resumen valores faltantes

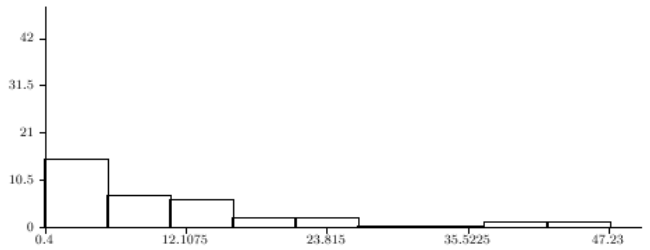
5.4.2 Imputación de Valores Faltantes Utilizando Método de Media Global

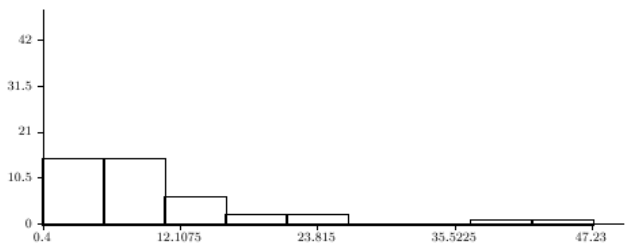
El sistema Java-KLASS realiza esta operación de imputación de forma rápida y sin consumir muchos recursos ni de tiempo ni de computación. Este proceso no requiere de parámetro alguno por parte del usuario o de expertos en el tema, es simplemente seleccionar la opción en el menú de tratamiento de faltantes y el sistema realiza el proceso.

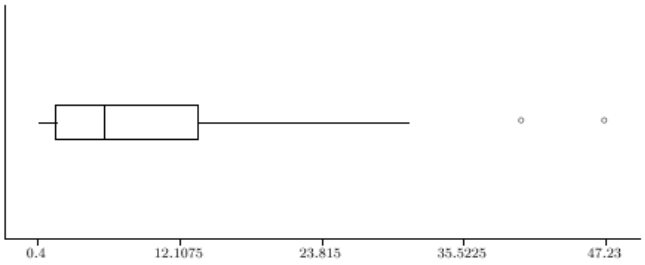
Ahora en este caso contamos con una variable categórica que tiene faltantes, como ya se ha dicho, el método de imputación por media global, no permite imputar variables de tipo categórico, por ello se ha de utilizar otra de las ventajas Java-KLASS, y se ha de recodificar los valores faltantes imputándolos de cierta manera con una etiqueta que represente el valor faltante y que puede ser “UNKNOWN”, para todos los valores faltantes en la variables categóricas, esto presenta desde luego inconvenientes, ya que no toma en cuenta ninguna de las particularidades de la variable para hacer la “imputación”, introduciendo ruido en las variables.

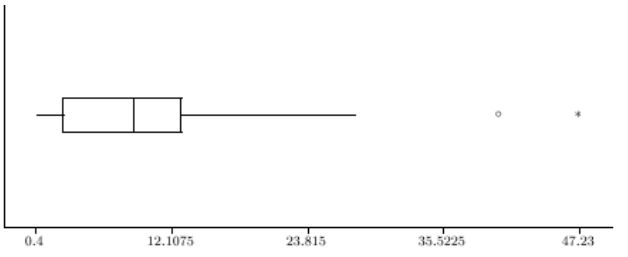
Una vez realizada la imputación, la distribución de las variables ha cambiado, por lo que es interesante hacer una comparativa de las variables antes y después de realizar la imputación , para ello a continuación se muestra la descriptiva de las variables antes y después de realizada la operación.

Antes **Después**







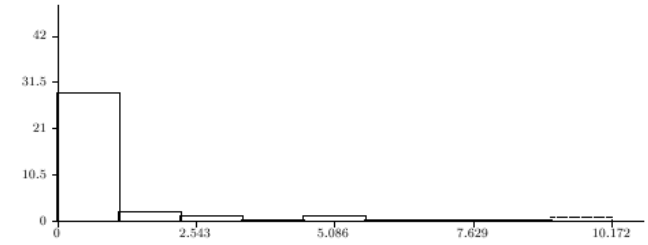


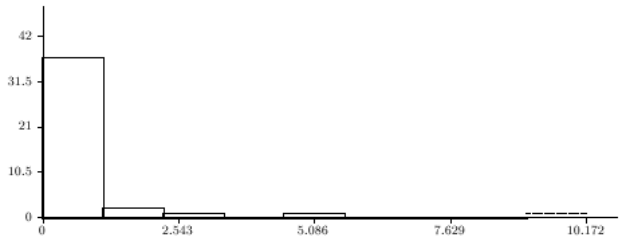
Nombre d'objectes	42
Nombre de dades mancants	8
Nombre d'observacions útils	34
Mitjana	9.7289
Mediana	5.93
Primer quartil (Q1)	1.93
Tercer quartil (Q3)	13.5507
Mínim	0.4
Màxim	47.23
Quasi-desviació típica	10.8552
Coeficient de variació	1.0992

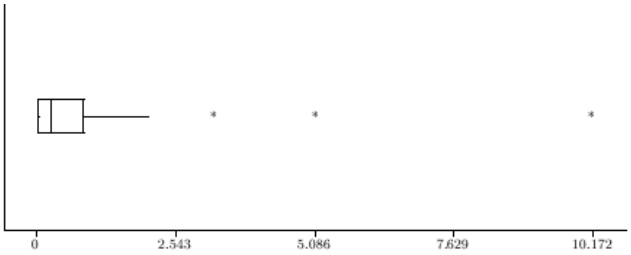
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	9.7289
Mediana	8.8445
Primer quartil (Q1)	2.76
Tercer quartil (Q3)	12.84
Mínim	0.4
Màxim	47.23
Quasi-desviació típica	9.7388
Coeficient de variació	0.989

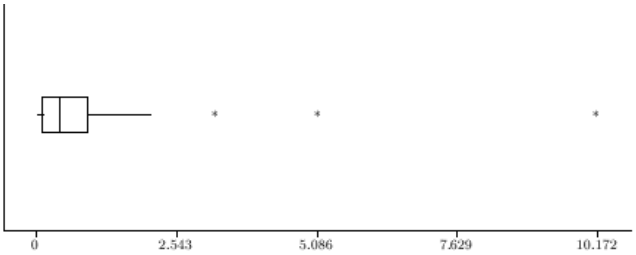
Figura No. 25: Descriptiva variable **Totprofmh** antes y después imputación media global: Histograma, boxplot y estadística sumaria

Antes **Después**







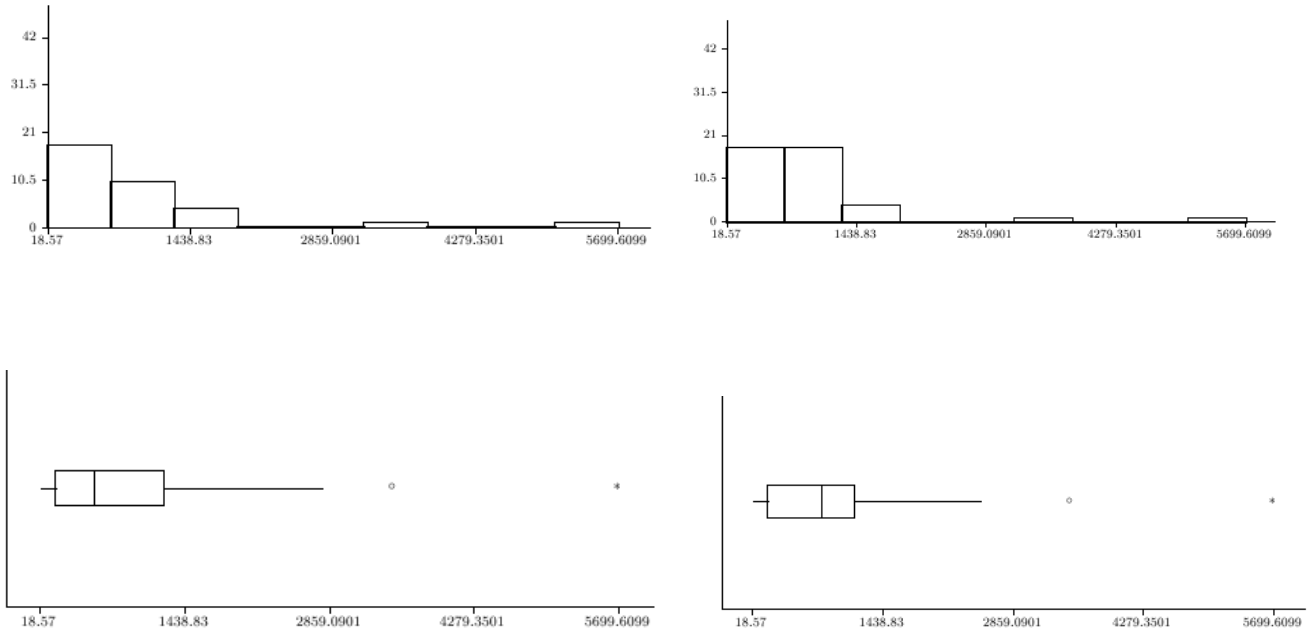


Nombre d'objectes	42
Nombre de dades mancants	8
Nombre d'observacions útils	34
Mitjana	0.9031
Mediana	0.2751
Primer quartil (Q1)	0.041
Tercer quartil (Q3)	0.8426
Mínim	0
Màxim	10.172
Quasi-desviació típica	1.9355
Coeficient de variació	2.1114

Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	0.9031
Mediana	0.424
Primer quartil (Q1)	0.1261
Tercer quartil (Q3)	0.9031
Mínim	0
Màxim	10.172
Quasi-desviació típica	1.7364
Coeficient de variació	1.8997

Figura No. 26: Descriptiva variable **Usmhexperca** antes y después imputación media global: Histograma, boxplot y estadística sumaria

Antes **Después**

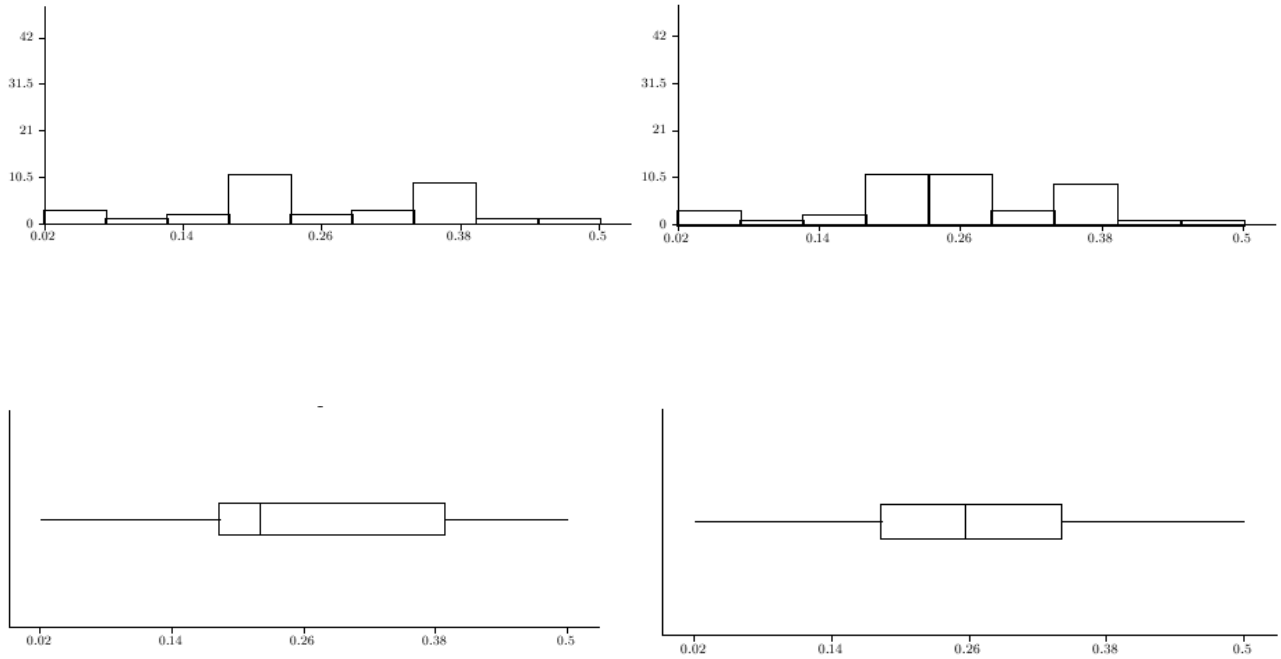


Nombre d'objectes	42
Nombre de dades mancants	8
Nombre d'observacions útils	34
Mitjana	865.43
Mediana	553.9938
Primer quartil (Q1)	172.77
Tercer quartil (Q3)	1219.4614
Mínim	18.57
Màxim	5699.6099
Quasi-desviació típica	1118.3401
Coeficient de variació	1.2731

Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	865.43
Mediana	776.37
Primer quartil (Q1)	192.22
Tercer quartil (Q3)	1120.13
Mínim	18.57
Màxim	5699.6099
Quasi-desviació típica	1003.3187
Coeficient de variació	1.1454

Figura No. 27: Descriptiva variable **Treatpre** antes y después imputación media global: Histograma, boxplot y estadística sumaria

Antes **Después**

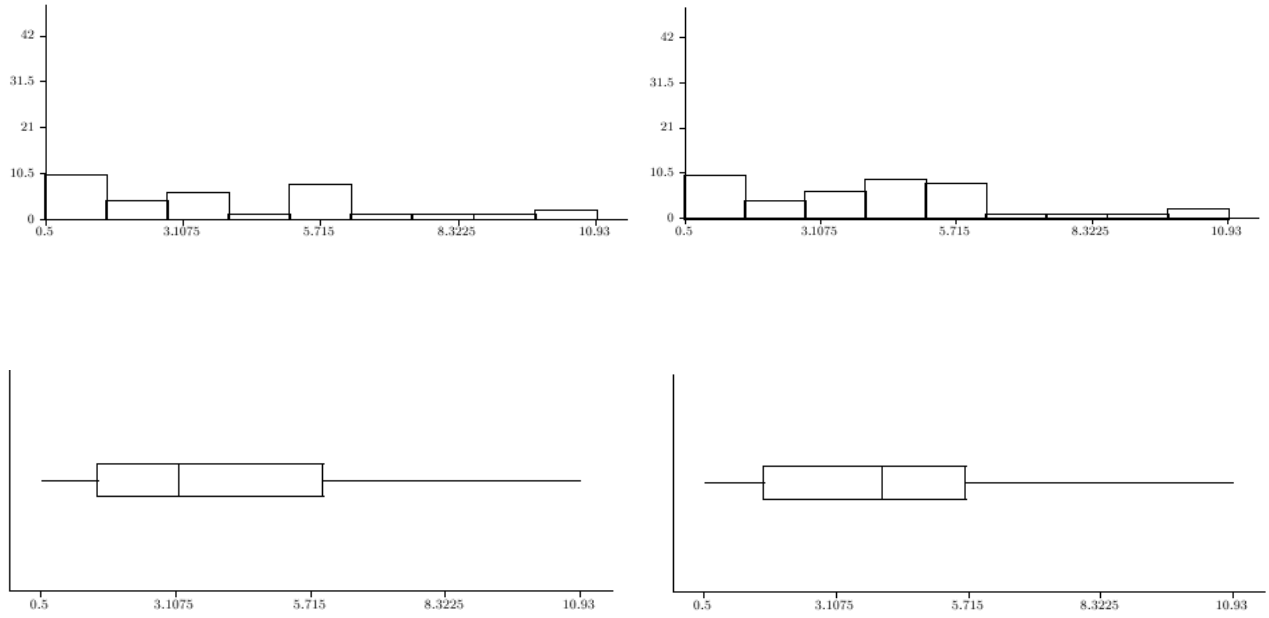


Nombre d'objectes	42
Nombre de dades mancants	9
Nombre d'observacions útils	33
Mitjana	0.2569
Mediana	0.22
Primer quartil (Q1)	0.1833
Tercer quartil (Q3)	0.3875
Mínim	0.02
Màxim	0.5
Quasi-desviació típica	0.1236
Coeficient de variació	0.4737

Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	0.2569
Mediana	0.2569
Primer quartil (Q1)	0.1833
Tercer quartil (Q3)	0.34
Mínim	0.02
Màxim	0.5
Quasi-desviació típica	0.1092
Coeficient de variació	0.4199

Figura No. 28: Descriptiva variable **Capratiosch** antes y después imputación media global: Histograma, boxplot y estadística sumaria

Antes **Después**

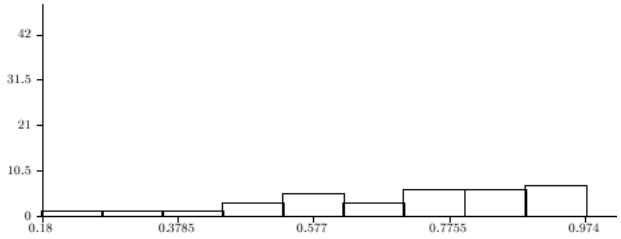


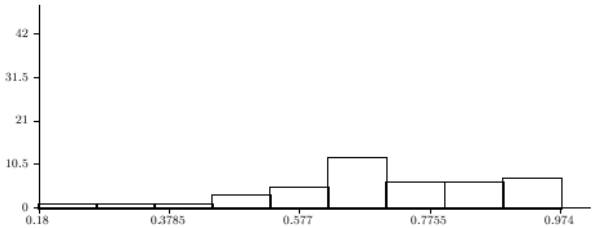
Nombre d'objectes	42
Nombre de dades mancants	8
Nombre d'observacions útils	34
Mitjana	4.0169
Mediana	3.155
Primer quartil (Q1)	1.59
Tercer quartil (Q3)	5.932
Mínim	0.5
Màxim	10.93
Quasi-desviació típica	2.8456
Coeficient de variació	0.6979

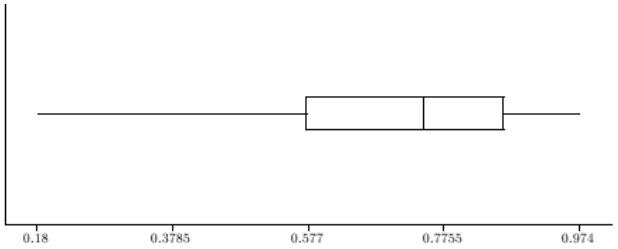
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	4.0169
Mediana	4.0169
Primer quartil (Q1)	1.68
Tercer quartil (Q3)	5.64
Mínim	0.5
Màxim	10.93
Quasi-desviació típica	2.553
Coeficient de variació	0.6279

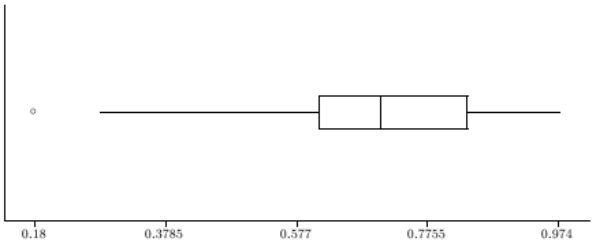
Figura No. 29: Descriptiva variable **d2f11i1closepsybeds** antes y después imputación media global: Histograma, boxplot y estadística sumaria

Antes **Después**







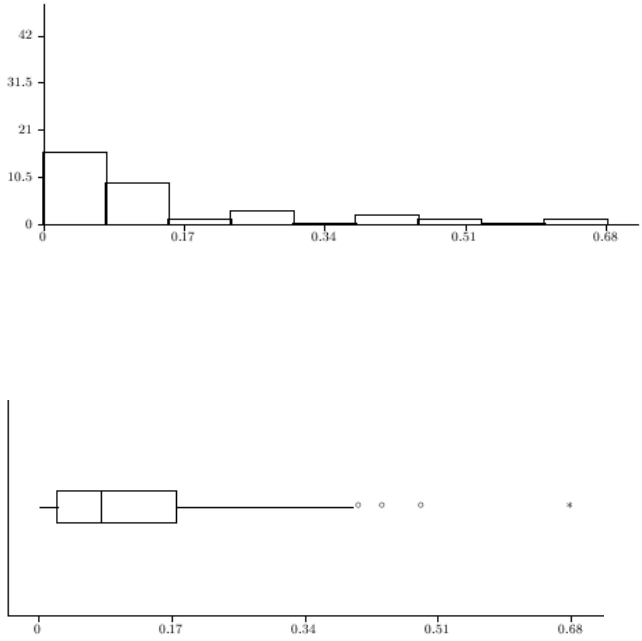


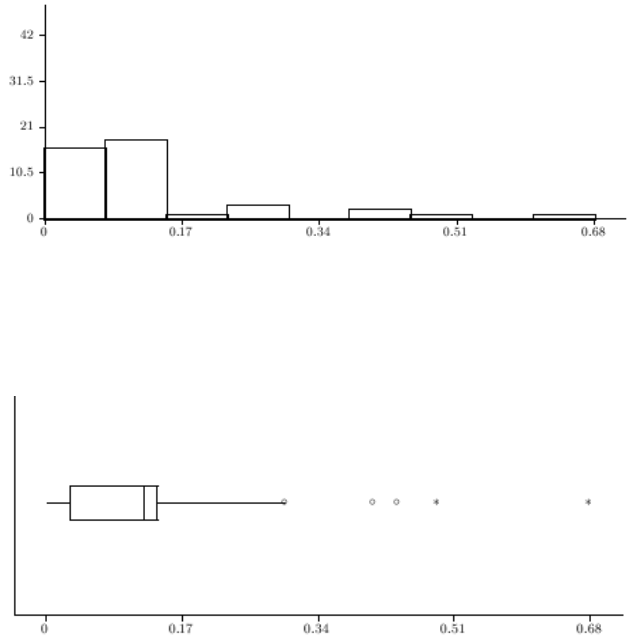
Nombre d'objectes	42
Nombre de dades mancants	9
Nombre d'observacions útils	33
Mitjana	0.705
Mediana	0.7463
Primer quartil (Q1)	0.5744
Tercer quartil (Q3)	0.8615
Mínim	0.18
Màxim	0.974
Quasi-desviació típica	0.1913
Coeficient de variació	0.2672

Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	0.705
Mediana	0.705
Primer quartil (Q1)	0.611
Tercer quartil (Q3)	0.833
Mínim	0.18
Màxim	0.974
Quasi-desviació típica	0.169
Coeficient de variació	0.2368

Figura No. 30: Descriptiva variable **d1f5i2exmhos** antes y después imputación media global: Histograma, boxplot y estadística sumaria

Antes **Después**



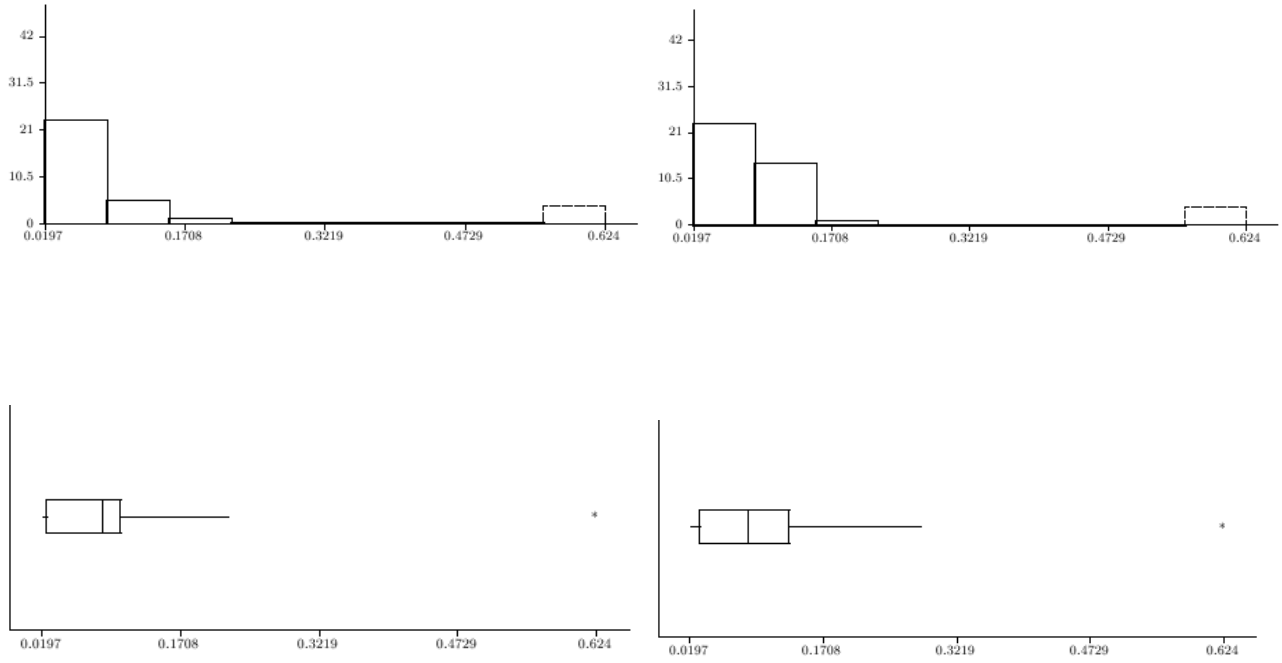


Nombre d'objectes	42
Nombre de dades mancants	9
Nombre d'observacions útils	33
Mitjana	0.1354
Mediana	0.08
Primer quartil (Q1)	0.0237
Tercer quartil (Q3)	0.1746
Mínim	0
Màxim	0.68
Quasi-desviació típica	0.1653
Coeficient de variació	1.2017

Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	0.1354
Mediana	0.1227
Primer quartil (Q1)	0.03
Tercer quartil (Q3)	0.1367
Mínim	0
Màxim	0.68
Quasi-desviació típica	0.146
Coeficient de variació	1.0652

Figura No. 31: Descriptiva variable **d2f6i71mhrec10y** antes y después imputación media global: Histograma, boxplot y estadística sumaria

Antes **Después**

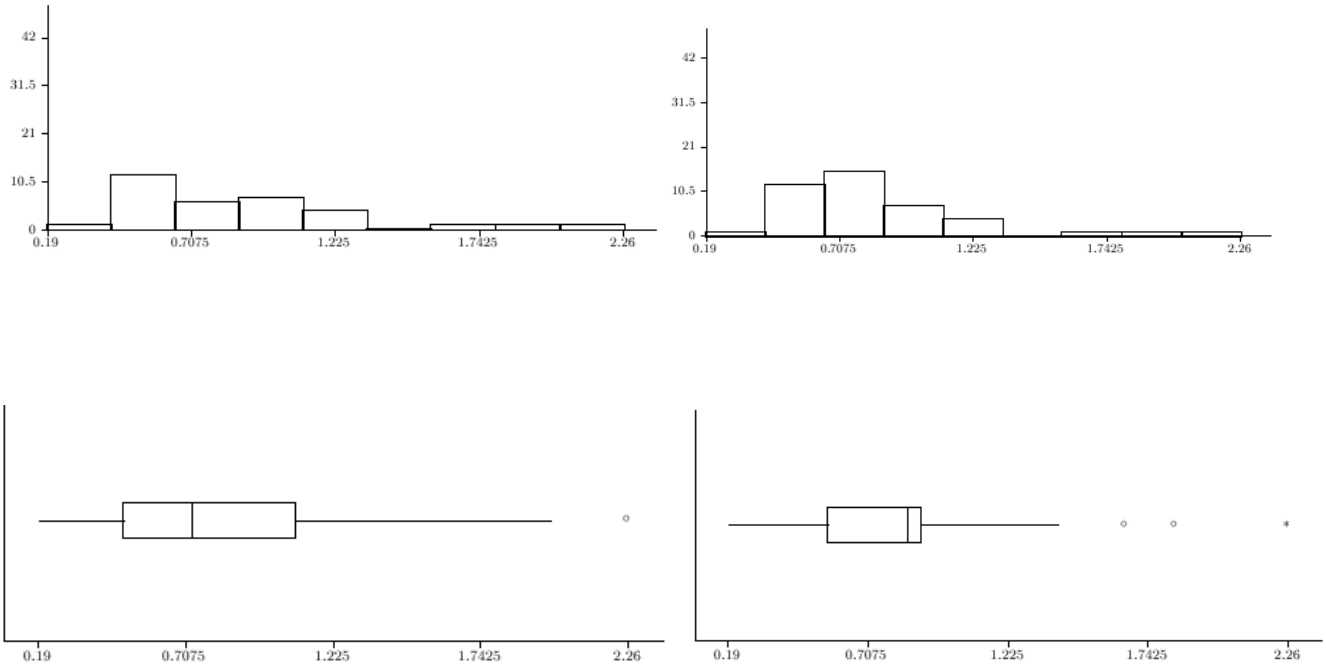


Nombre d'objectes	42
Nombre de dades mancants	9
Nombre d'observacions útils	33
Mitjana	0.1313
Mediana	0.0856
Primer quartil (Q1)	0.0245
Tercer quartil (Q3)	0.1036
Mínim	0.0197
Màxim	0.624
Quasi-desviació típica	0.1905
Coeficient de variació	1.4285

Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	0.1313
Mediana	0.0856
Primer quartil (Q1)	0.0314
Tercer quartil (Q3)	0.1313
Mínim	0.0197
Màxim	0.624
Quasi-desviació típica	0.1683
Coeficient de variació	1.2662

Figura No. 32: Descriptiva variable **Comcarewor** antes y después imputación media global: Histograma, boxplot y estadística sumaria

Antes **Después**

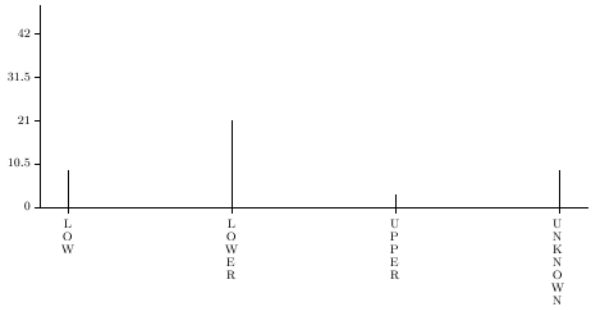


Nombre d'objectes	42
Nombre de dades mancants	9
Nombre d'observacions útils	33
Mitjana	0.8523
Mediana	0.73
Primer quartil (Q1)	0.49
Tercer quartil (Q3)	1.09
Mínim	0.19
Màxim	2.26
Quasi-desviació típica	0.44
Coeficient de variació	0.5083

Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	0.8523
Mediana	0.8523
Primer quartil (Q1)	0.56
Tercer quartil (Q3)	0.9
Mínim	0.19
Màxim	2.26
Quasi-desviació típica	0.3887
Coeficient de variació	0.4506

Figura No. 33: Descriptiva variable **lundpararectrail** antes y después imputación media global: Histograma, boxplot y estadística sumaria

Antes **Después**

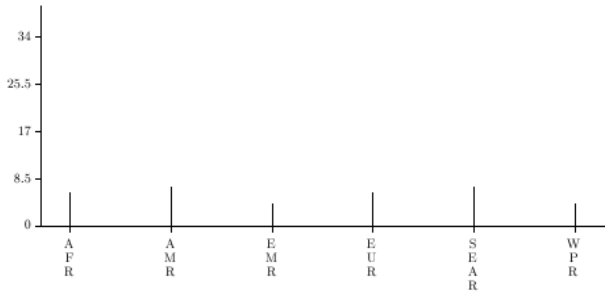


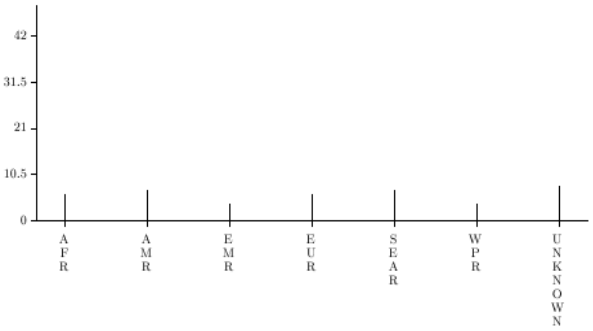
Modalitats	Freq. absol.	Freq. acum.	Freq. relat.	Freq. rel. acum.
LOW	9	9	0.2727	0.2727
LOWER	21	30	0.6364	0.9091
UPPER	3	33	0.0909	1
dades mancants	9	N = 42	0.2143	

Modalitats	Freq. absol.	Freq. acum.	Freq. relat.	Freq. rel. acum.
LOW	9	9	0.2143	0.2143
LOWER	21	30	0.5	0.7143
UPPER	3	33	0.0714	0.7857
UNKNOWN	9	42	0.2143	1
dades mancants	0	N = 42	0	

Figura No. 34: Descriptiva variable **Incgroup** antes y después imputación media global: Taula de frecuencias y diagrama de Barras

Antes **Después**





Modalitats	Freq. absol.	Freq. acum.	Freq. relat.	Freq. rel. acum.
AFR	6	6	0.1818	0.1818
AMR	8	14	0.2424	0.4242
EMR	6	20	0.1818	0.6061
EUR	6	26	0.1818	0.7879
SEAR	5	31	0.1515	0.9394
WPR	2	33	0.0606	1
dades mancants	9	N = 42	0.2143	

Modalitats	Freq. absol.	Freq. acum.	Freq. relat.	Freq. rel. acum.
AFR	6	6	0.1429	0.1429
AMR	7	13	0.1667	0.3095
EMR	4	17	0.0952	0.4048
EUR	6	23	0.1429	0.5476
SEAR	7	30	0.1667	0.7143
WPR	4	34	0.0952	0.8095
UNKNOWN	8	42	0.1905	1
dades mancants	0	N = 42	0	

Figura No. 35: Descriptiva variable **Region** antes y después imputación media global: Taula de frecuencias y diagrama de Barras

Como se puede observar, la imputación altera la distribución de los valores de las variables, en el caso de las variables categóricas, se crea una nueva modalidad, con la que se imputan todos los valores faltantes de una misma variable, sin tomar en cuenta ninguna particularidad ni relación de los datos. Tomando en cuenta que el porcentaje de valores faltantes en las variables es cercano al 20%, y que todos ellos toman el mismo valor, estaríamos hablando de una variación considerable en los datos, que no tendría correspondencia con de la población objetivo que se está representando.

Algo muy parecido sucede con las variables numéricas, ya que todos los valores faltantes de una variable se imputan por un único valor, en este caso la media, sin tener consideración de aspectos como la distribución de los faltantes, el tamaño de la matriz, la relación de la variable con otras variables. Este particular es fácil de apreciar, sobre todo en el histograma, ya que se nota claramente cómo crece la tendencia hacia el valor de la media, introduciendo sesgo en la variable y alterando la distribución de la misma, si bien ya habíamos hablado de este tema, contar con un ejemplo tan claro como este nos ayuda a visualizar la teoría antes descrita.

Aunque lo más grave al realizar la imputación por media global, es la reducción de la desviación típica y del coeficiente de variación que se puede observar en las tablas de estadística sumaria de la comparativa anterior, aunque para mejor comprensión se presenta la tabla No. 7 de resumen a continuación:

Variable	Desviación típica		Coeficiente de variación	
	Antes de imputar	Después de imputar	Antes de imputar	Después de imputar
Totprofmh	10,8552	9,7388	1,0992	0,989
Usmhexperca	1,9355	1,7364	2,1114	1,8997
Treatpre	1118,3401	1003,3187	1,2731	1,1454
Capratiosch	0,1236	0,1092	0,4737	0,4199
d2f11i1closepsybeds	2,8456	2,553	0,6979	0,6279
d1f5i2exmhos	0,1913	0,169	0,2672	0,2368
d2f6i71mhrec10y	0,1653	0,146	1,2017	1,0652
Comcarewor	0,1905	0,1683	1,4285	1,2662
lundpararectrail	0,44	0,3887	0,5083	0,4506

Tabla No. 7: Resumen cambios en la desviación típica y en el coeficiente de variación de las variables imputadas con el método de media global

Todo esto lo que nos indica es que la varianza de las variables disminuyó con respecto de la original, como ya se había visto en el capítulo **2.3.2 Métodos de Imputación de Datos Faltantes,** este inconveniente es una de las desventajas propias de este método, pero ahora podemos observarlo claramente en nuestro caso práctico y verificar que es un inconveniente que ocurre con la imputación de variables utilizando el valor de la media global, lo que se traduce en la el modelo pierde representatividad de la población objetivo, alterando el modelo y con ello los resultados que se pueden obtener del mimo.

5.3.2 Clasificación Utilizando Matriz Resultante

Una vez se ha realizado la imputación y contamos con una matriz e datos "completa" que nos permite continuar en el proceso extracción de conocimiento de una fuente de datos, procedemos a realizar la clasificación mediante el sistema Java-KLASS, para ello utilizamos uno de los muchos métodos de clustering disponibles en java-KLASS y descritos en el **Capítulo 4.3.1.**

Al no tener ninguna hipótesis previa sobre el número de clusters que se quieren construir, seleccionaremos uno de los métodos de clustering jerárquico, que construyen una jerarquía de agrupamientos, uniendo si es ascendente o dividiendo si es divisivo los grupos de acuerdo a una cierta función que busca similitud o disimilitud entre los grupos. Java-KLASS solo implementa opciones de cluster jerárquico ascendente. En este caso utilizaremos el algoritmo básico de los vecinos recíprocos encadenados para analizar estrictamente el impacto del método de imputación en la clasificación, aunque JavaKLASS aporta otros métodos ascendentes jerárquicos híbridos que incluyen elementos de IA como la clasificación basada en reglas que inyecta conocimiento específico de dominio a priori al proceso jerárquico.

El criterio de agregación que se utilizará es el Ward por estar relacionado con la cantidad de información contenida en los datos, y presentar buenas propiedades en aplicaciones realce, con tendencia a generar grupos bastante interpretables por parte del experto

Dado que las variables seleccionadas para la construcción del cluster son numéricas y categóricas, hemos de utilizar una distancia mixta, que nos permita realizar la clasificación con una mezcla de variables numéricas y categóricas, específicamente hemos seleccionado la distancia Mixta Gibert [Gibert, 97], ya que nos permite un tratamiento homogéneo de diferentes tipos de variables, manteniéndolas en su forma original, y está ampliamente documentado su buen comportamiento [Gibert and Conti, 2016] y [Gibert et al., 2010a].

Los parámetros α y β, para la distancia Mixta Gibert, se tomaron de la opción de cálculo automático de Java-KLASS. Con estos parámetros se procedió a realizar la clasificación que dio como resultado el siguiente dendrograma (Figura No. 27).

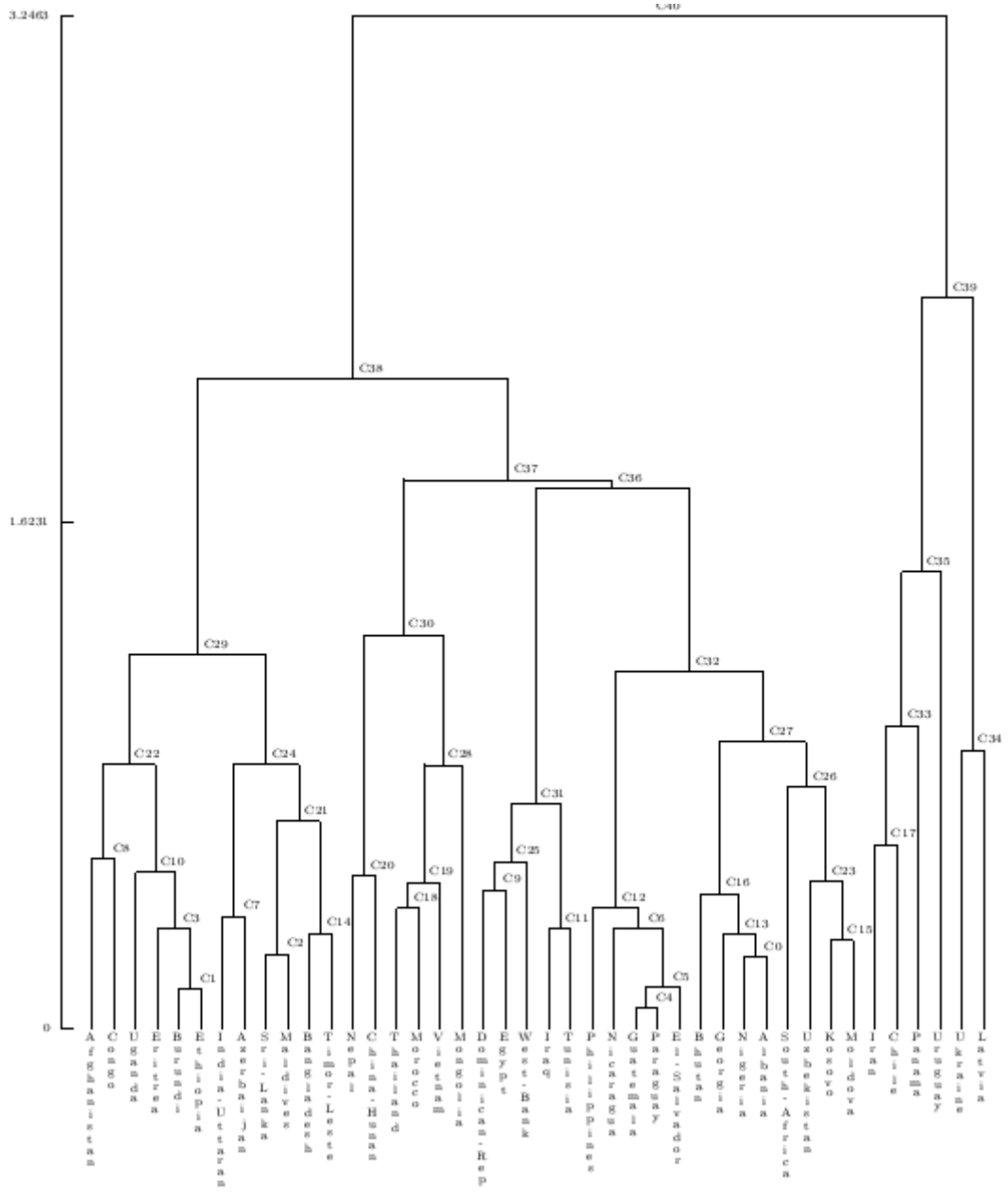


Figura No. 36: Dendrograma de resultado de la clasificación posterior a imputación con media global.

A la vista del dendrograma se utiliza el heurístico de cortar el árbol por una zona donde exista una discontinuidad en los índices de nivel de las clases, lo cual es equivalente a optimizar el criterio de Calinski-Harabasz sobre los cortes del árbol. Según este criterio, el mejor corte sería en 2 clases, que se desprecia por resultar poco informativo, práctica habitual en las aplicaciones reales de clustering. El segundo mejor corte es en 4 clases y el tercer mejor corte en 7. De acuerdo a los intereses de los expertos se realizó el corte en 7 clases resultantes. Con la ayuda de la funcionalidad de Java-KLASS, visualizamos a continuación el resultado de la clasificación, empezamos por el listado de objetos por clase (tabla No.), luego las Distribuciones condicionadas por clase (figuras No. 37 y 38) y por ultimo un mapa con la distribución de los países por clase.

Classe	Objectes
C29	Afghanistan, Azerbaijan, Bangladesh, Burundi, Congo, Eritrea, Ethiopia, India-Uttaranc, Maldives, Sri-Lanka, Timor-Leste, Uganda
C32	Albania, Bhutan, El-Salvador, Georgia, Guatemala, Kosovo, Moldova, Nicaragua, Nigeria, Paraguay, Philippines, South-Africa, Uzbekistan
C33	Chile, Iran, Panama
C30	China-Hunan, Mongolia, Morocco, Nepal, Thailand, Vietnam
C31	Dominican-Repu, Egypt, Iraq, Tunisia, West-Bank
C34	Latvia, Ukraine
Uruguay	Uruguay

Tabla No. 8: Descripción extensional de clases resultantes de la clasificación

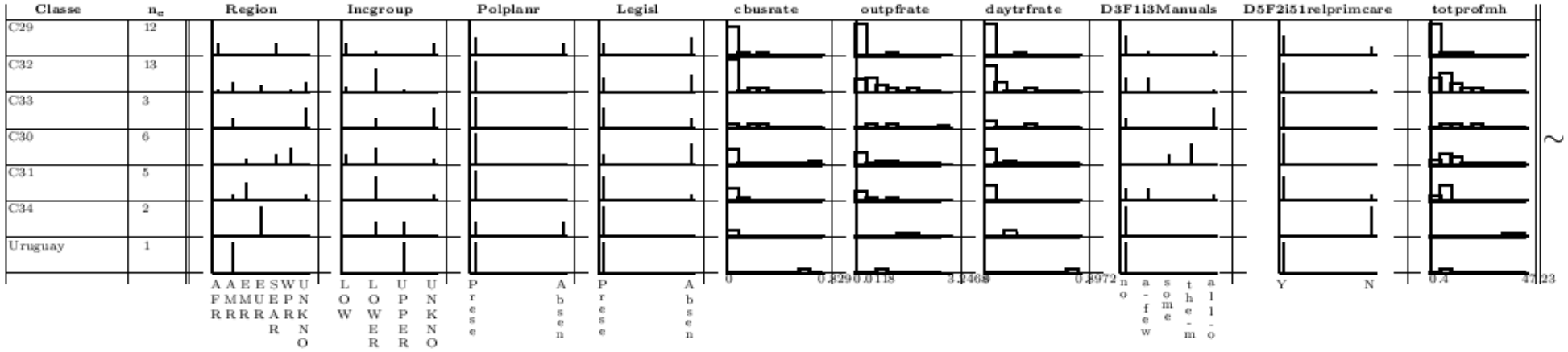


Figura No. 37: Distribuciones condicionadas por clase parte 1.

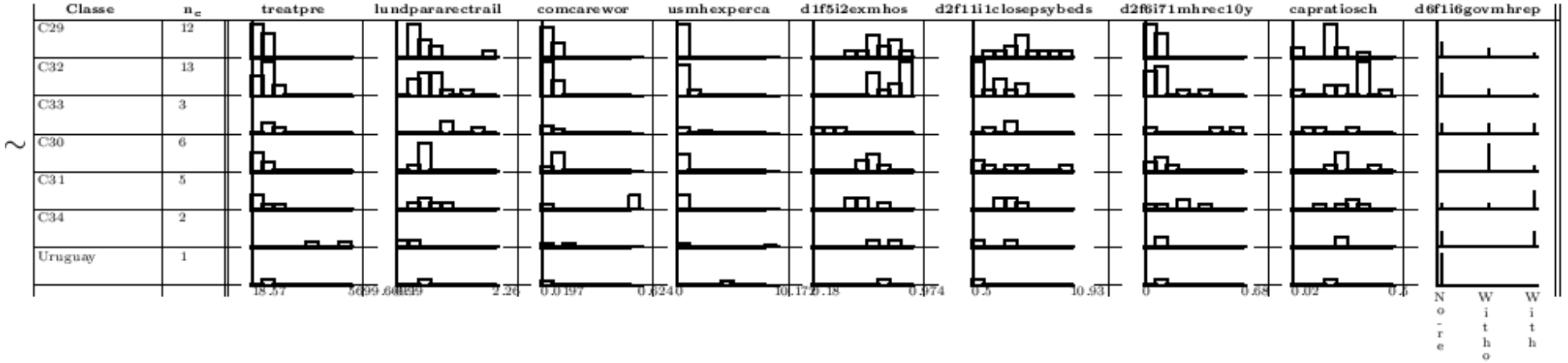


Figura No. 38: Distribuciones condicionadas por clase parte 2.

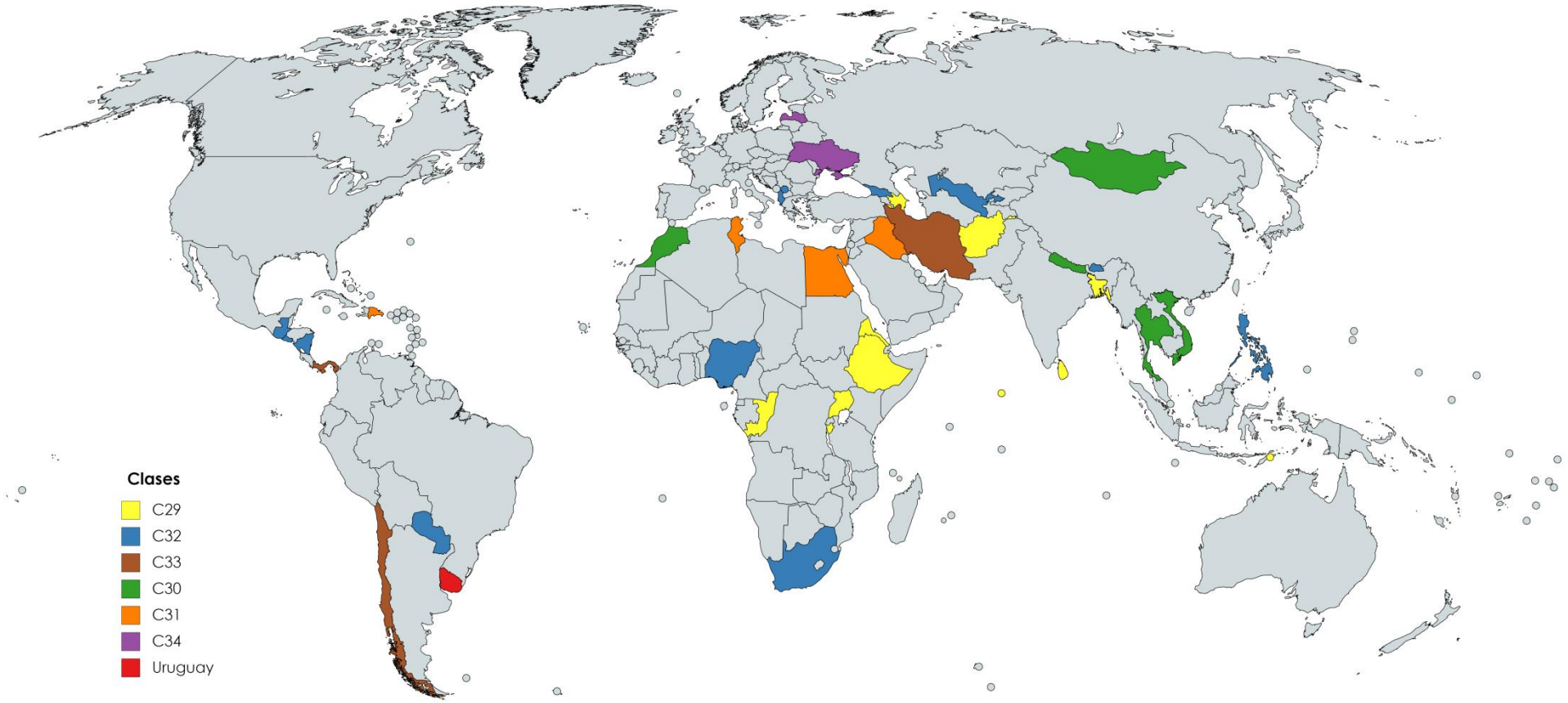


Figura No. 39: Representación de la clasificación de los países por clase.

5.4 Clasificación Utilizando MIMMI

5.4.1 Selección de la Submatriz para Clasificación Auxiliar de la FASE1

En [Gibert, 2010], con ayuda de expertos se identifica un conjunto de 16 variables relevantes para utilizar como explicativas en el proceso de imputación de datos cuando se quiere utilizar el MIMMI (tabla No. 9).

Variable	Significado	Tipo de Variable
polplanr	Presencia de política o plan de salud mental en el país	Cualitativa
Legisl	Presencia de legislación en salud mental en el país	Cualitativa
d1f5i5rec(antipsych)	Asequibilidad de la medicina antipsicótica.	Cuantitativa
d1f5i6rec(antidepr)	Asequibilidad a antidepresivos.	Cuantitativa
D2f1i2(orgServices)	Organización de servicios .	Cualitativa
Cbusrate	Unidades de pacientes internados basadas en la comunidad por 100 000 habitantes.	Cuantitativa
Mhrate	Hospitales psiquiátricos por 100 000 habitantes.	Cuantitativa
Outpfrate	Instalaciones ambulatorias por 100 000 habitantes.	Cuantitativa
Daytrfrate	Instalaciones de centros de día por 100 000.	Cuantitativa
d4f1i11(psychi)	Psiquiatras por 100 000 habitantes.	Cuantitativa
D4F1I12(doctors)	Otros doctores por 100 000 habitantes.	Cuantitativa
D4F1I13(nurses)	Enfermeras por 100 000 habitantes.	Cuantitativa
D4F1I14(psycho)	Psicólogos por 100 000 habitantes.	Cuantitativa
D4F1I15(socWorK)	Trabajadores sociales por 100 000 habitantes.	Cuantitativa
d3f1i3(manuals)	Disponibilidad de manuales de tratamiento y evaluación.	Cualitativa
Relprimcare	Relación de colaboración formal con el departamento de atención primaria.	Cualitativa

Tabla No. 9: Descripción de las variables utilizadas para la clasificación auxiliar

Al realizar un análisis descriptivo de estas variables, se verificó que ninguna tenía valores faltantes, por lo que no se requiere realizar imputación inteligente o utilizar forzosamente la distancia de Gower para la clasificación auxiliar.

5.4.2 FASE 1: Clasificación Auxiliar

Utilizando clustering jerárquico, en la funcionalidad ya descrita de MIMMI para realizar la clasificación auxiliar, realizamos la misma con el criterio de Ward y la distancia Mixta Gibert tomando los valores de α y β, de la opción automática que presenta Java-KLASS.

La clasificación que dio como resultado el siguiente dendrograma (Figura No. 28)

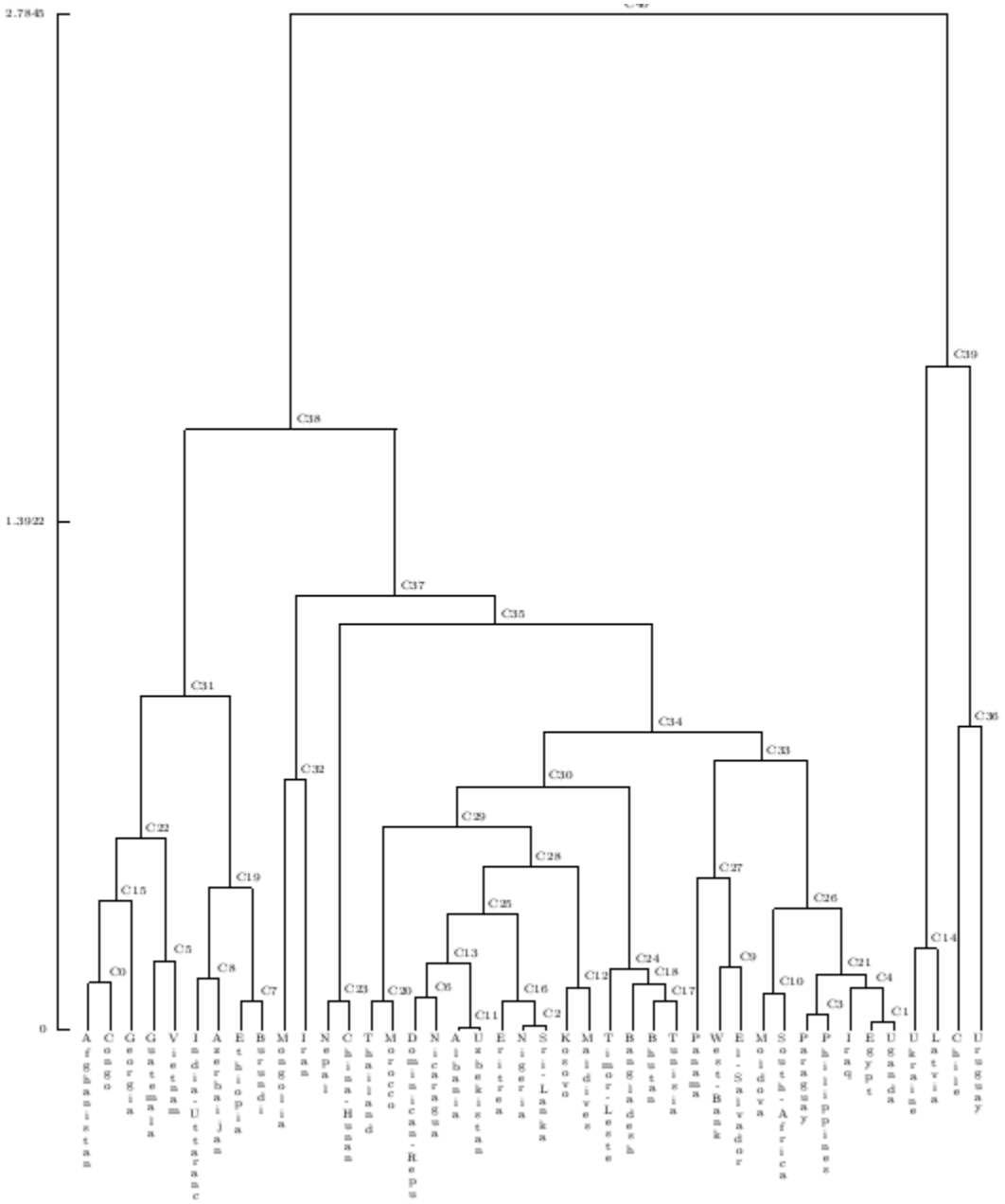


Figura No. 40: Dendrograma de resultado de la clasificación auxiliar.

En este caso se utiliza un criterio parecido al anterior para cortar el árbol, pero se recomienda cortar en algunas clases más para disponer de más varianza en las medias condicionadas que se utilizarán en la siguiente fase de MIMMI y por tanto, se decide realizar el corte del árbol en 7 clusters, de ahí el sistema pasa automáticamente a la siguiente fase.

Imputación de datos faltantes por la media condicionada a clase

Una vez el sistema realiza la agrupación, comienza el cálculo de la media condicionada por cluster y la imputación de los datos faltantes por el valor de la media correspondiente a su clase, este proceso se realiza internamente y solo se realiza un aviso de la ejecución con éxito del proceso.

5.4.3 Análisis Reporte Final de Imputación de Datos Faltantes con MIMMI

Si se ha seleccionado la opción de mostrar descriptiva de la imputación a continuación del proceso de imputación se construye el informe MIMMI, como ya se describió en el **Capítulo 3** de este documento. En este caso obviamente se seleccionó la opción para realizar la descriptiva que nos va a permitir visualizar mejor el proceso de imputación que se efectuó, es importante aclarar que en el **Anexo 2,** se presenta el informe MIMMI completo, en caso de requerir ampliar alguna información.

A continuación en la tabla No. 10, se muestran los valores de la media de las variables numéricas, para cada una de las clases resultantes de la clasificación auxiliar, estos serán los valores que servirán para imputar los faltantes de la variable.

MITJANA

idClase	totprofmh	treatpre	lundpararectrail	comcarewor
C22	5.1	1115.5867	1.1425	0.0975
C34	7.1795	520.1282	0.817	0.1583
C19	5.6004	657.065	0.66	0.0275
C36	24.15	1221.6001	1.16	0.0245
C23	13.5507	283.61	0.67	0.1313
C32	13.97	1120.13	1.84	0.1206
C14	43.81	4595.1797	0.375	0.1424

idClase	usmhexperca	d1f5i2exmhos	d2f11i1closepsybeds	d2f6i71mhrec10y
C22	0.2683	0.7167	3.282	0.057
C34	0.687	0.724	4.1169	0.1732
C19	0.1213	0.8747	5.804	0.0486
C36	2.5634	0.525	0.85	0.49
C23	0.0393	0.6787	6.265	0.0684
C32	0.1281	0.408	1.59	0.02
C14	5.5864	0.804	1.33	0.1134

idClase	capratiosch
C22	0.2245
C34	0.2656
C19	0.2383
C36	0.2475
C23	0.44
C32	0.21
C14	0.2569

Tabla No. 10: Valores de imputación variable numérica

También podemos apreciar en la tabla No. 10, los valores de imputación por cada variable y clase para las variables categóricas, en este caso como ya se había mencionado, obviamente no se utiliza el valor de la media o media, sino un valor que también toma en cuenta la clasificación auxiliar previamente realizada. Este valor está conformado por el nombre de la variable, el valor de la clase correspondiente al valor faltante y la constante "UNKNOWN".

idClase	Region	Incgroup
C22	RegionC22.UNKNOWN	IncgroupC22.UNKNOWN
C34	RegionC34.UNKNOWN	IncgroupC34.UNKNOWN
C19	RegionC19.UNKNOWN	IncgroupC19.UNKNOWN
C36	RegionC36.UNKNOWN	IncgroupC36.UNKNOWN
C23	RegionC23.UNKNOWN	IncgroupC23.UNKNOWN
C32	RegionC32.UNKNOWN	IncgroupC32.UNKNOWN
C14	RegionC14.UNKNOWN	IncgroupC14.UNKNOWN

Tabla No. 11: Valores de imputación variables categóricas

Además de conocer los valores por los que fueron realizadas las imputaciones, también es de interés conocer el número de imputaciones realizadas por clase y variable (tabla No. 10), y al observar la tabla se puede ver como los valores de imputación se dispersan en las variables según la clase a la que pertenezca el valor faltante, al contrario de lo que sucede con el método de imputación por la variable global, en el cual todos los valores de imputación son iguales, lo que como ya se ha dicho perjudica la varianza de la propiedad.

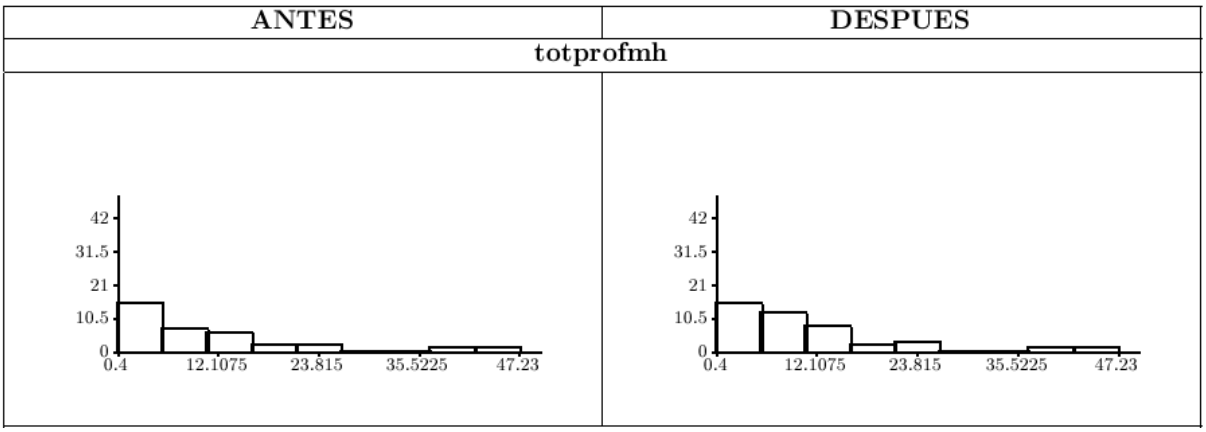
idClase	Region	Incgroup	totprofmh	treatpre
C22	2	2	0	2
C34	4	4	5	3
C19	0	1	0	2
C36	1	1	1	0
C23	0	0	1	0
C32	1	1	1	1
C14	0	0	0	0

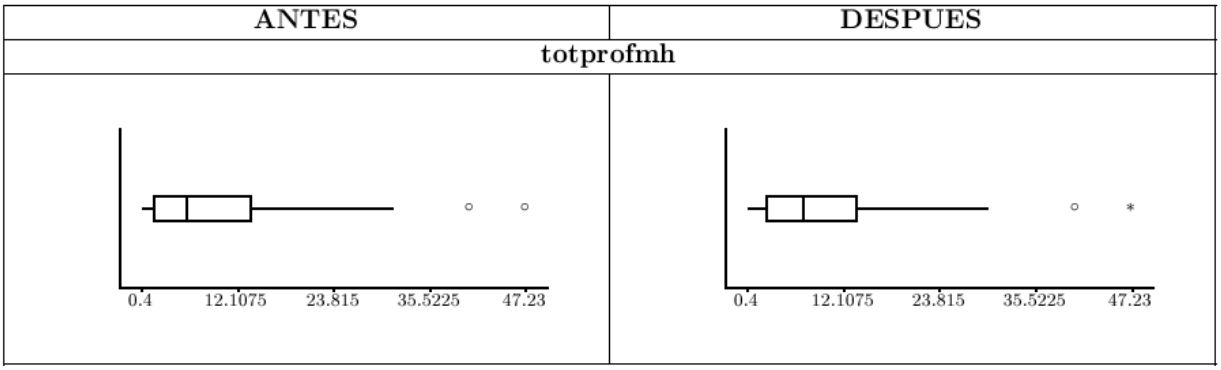
idClase	lundpararectrail	comcarewor	usmhexperca	d1f5i2exmhos
C22	1	3	0	0
C34	4	3	7	7
C19	1	1	1	1
C36	1	0	0	0
C23	1	2	0	0
C32	1	0	0	0
C14	0	0	0	1

idClase	d2f11i1closepsybeds	d2f6i71mhrec10y	capratiosch
C22	0	0	0
C34	4	7	6
C19	1	0	0
C36	1	1	0
C23	0	0	1
C32	1	1	0
C14	1	0	2

Tabla No. 12: Número de Valores de imputación por clase y variable

Continuamos con la descriptiva de las variables antes y después de la imputación, dado que el "Reporte MIMMI" se incorpora a este documento como **Anexo no. 2,** no se muestra aquí la descriptiva comparativa de todas las variables, ya que podría volver engorrosa la lectura del documento, solo se incluye a manera de ejemplo algunas de las comparativas, lo que debe tratar de notarse en este punto en la diferencia de la distribución de los datos faltantes en las variables, al contrario del método anterior, MIMMI realiza la imputación distribuyendo los valores entre las clases generadas de la clasificación auxiliar, lo que se refleja en una menor disminución de la varianza y una distribución más apegada a la realidad de los valores faltantes.





ANTES

totprofmh

Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	8
Nombre d'observacions útils	34
Mitjana	9.7289
Mediana	5.93
Primer quartil (Q1)	1.93
Tercer quartil (Q3)	13.5507
Mínim	0.4
Màxim	47.23
Quasi-desviació típica	10.8552
Coeficient de variació	1.0992

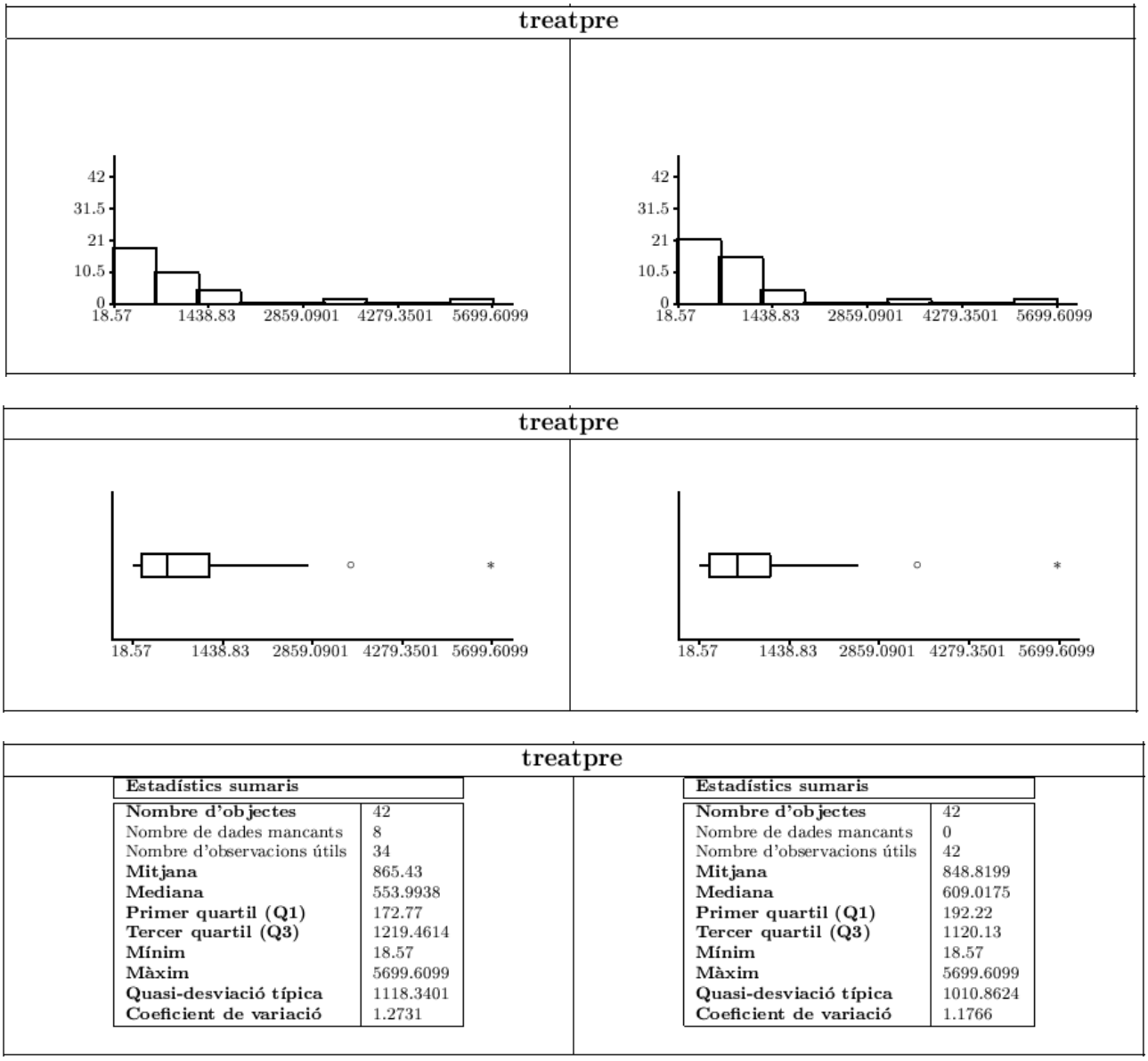
DESPUES

totprofmh

Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	9.9607
Mediana	7.1795
Primer quartil (Q1)	2.76
Tercer quartil (Q3)	13.5507
Mínim	0.4
Màxim	47.23
Quasi-desviació típica	10.0722
Coeficient de variació	0.9991

Figura No. 41: Descriptiva variable **totprofmh** antes y después imputación media global: Histograma, boxplot y estadísticas sumarias

Antes **Después**

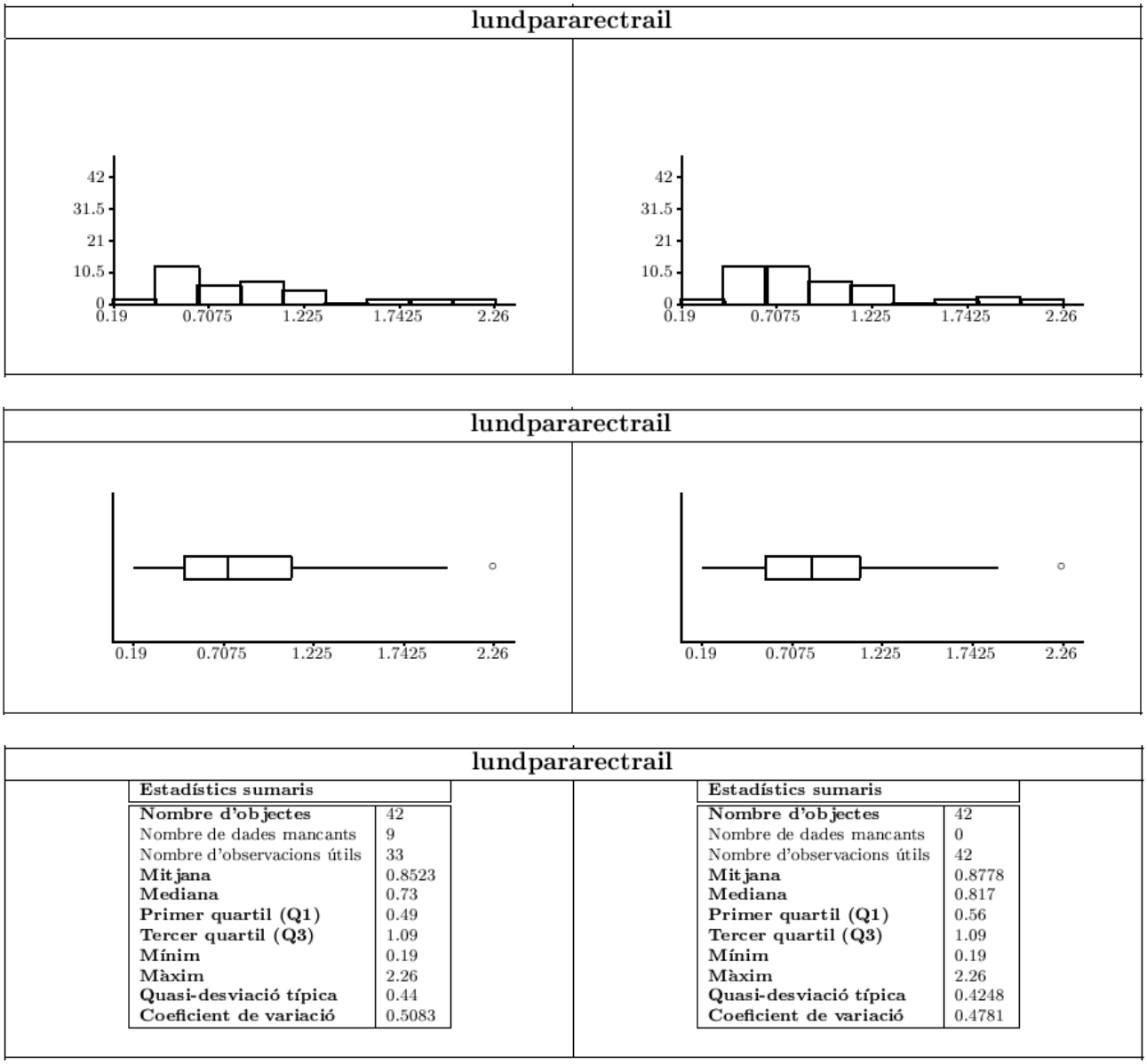


Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	8
Nombre d'observacions útils	34
Mitjana	865.43
Mediana	553.9938
Primer quartil (Q1)	172.77
Tercer quartil (Q3)	1219.4614
Mínim	18.57
Màxim	5699.6099
Quasi-desviació típica	1118.3401
Coeficient de variació	1.2731

Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	848.8199
Mediana	609.0175
Primer quartil (Q1)	192.22
Tercer quartil (Q3)	1120.13
Mínim	18.57
Màxim	5699.6099
Quasi-desviació típica	1010.8624
Coeficient de variació	1.1766

Figura No. 42: Descriptiva variable **treatpre** antes y después imputación media global: Histograma, boxplot y estadística sumaria

Antes **Después**

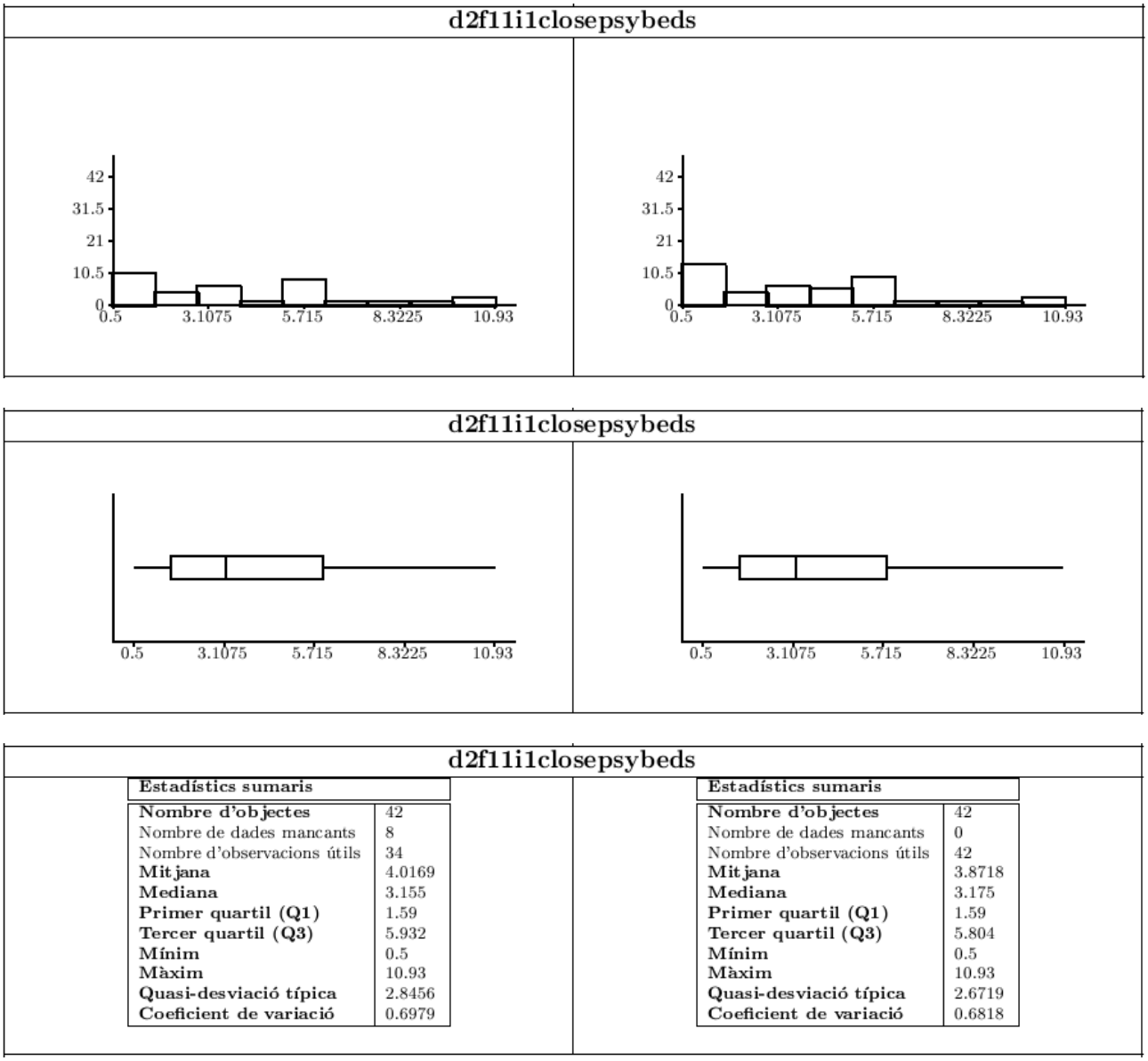


Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	9
Nombre d'observacions útils	33
Mitjana	0.8523
Mediana	0.73
Primer quartil (Q1)	0.49
Tercer quartil (Q3)	1.09
Mínim	0.19
Màxim	2.26
Quasi-desviació típica	0.44
Coeficient de variació	0.5083

Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	0.8778
Mediana	0.817
Primer quartil (Q1)	0.56
Tercer quartil (Q3)	1.09
Mínim	0.19
Màxim	2.26
Quasi-desviació típica	0.4248
Coeficient de variació	0.4781

Figura No. 43: Descriptiva variable **lundpararectrail** antes y después imputación media global: Histograma, boxplot y estadística sumaria

Antes **Después**



Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	8
Nombre d'observacions útils	34
Mitjana	4.0169
Mediana	3.155
Primer quartil (Q1)	1.59
Tercer quartil (Q3)	5.932
Mínim	0.5
Màxim	10.93
Quasi-desviació típica	2.8456
Coeficient de variació	0.6979

Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	3.8718
Mediana	3.175
Primer quartil (Q1)	1.59
Tercer quartil (Q3)	5.804
Mínim	0.5
Màxim	10.93
Quasi-desviació típica	2.6719
Coeficient de variació	0.6818

Figura No. 44: Descriptiva variable **d2f11ilclosepsybeds** antes y después imputación media global: Histograma, boxplot y estadística sumaria

Antes **Después**

Region

Taula de freqüències

Modalitats	Freq. absol.	Freq. acum.	Freq. relat.	Freq. rel. acum.
AFR	6	6	0.1765	0.1765
AMR	7	13	0.2059	0.3824
EMR	4	17	0.1176	0.5
EUR	6	23	0.1765	0.6765
SEAR	7	30	0.2059	0.8824
WPR	4	34	0.1176	1
dades mancants	8	N = 42	0.1905	

Taula de freqüències

Modalitats	Freq. absol.	Freq. acum.	Freq. relat.	Freq. rel. acum.
AFR	6	6	0.1429	0.1429
AMR	7	13	0.1667	0.3095
EMR	4	17	0.0952	0.4048
EUR	6	23	0.1429	0.5476
SEAR	7	30	0.1667	0.7143
WPR	4	34	0.0952	0.8095
RegionC22.UNKNOWN	2	36	0.0476	0.8571
RegionC34.UNKNOWN	4	40	0.0952	0.9524
RegionC36.UNKNOWN	1	41	0.0238	0.9762
RegionC32.UNKNOWN	1	42	0.0238	1
dades mancants	0	N = 42	0	

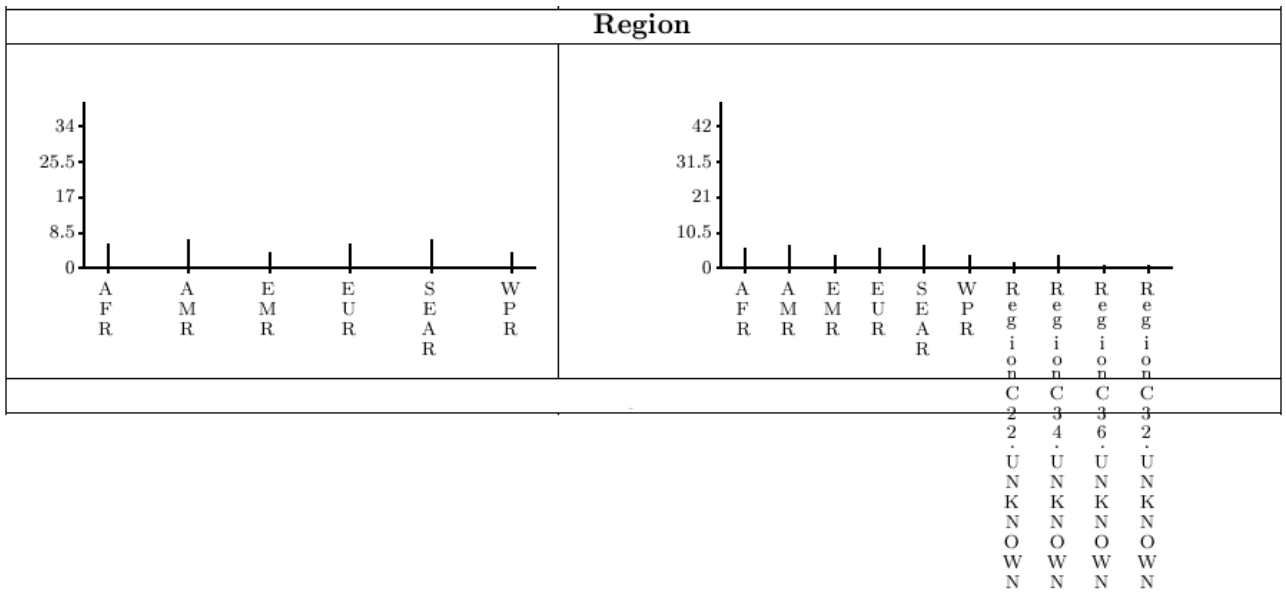


Figura No. 45: Descriptiva variable **Region** antes y después imputación media global: Taula de frecuencias y diagrama de Barras

Antes **Después**

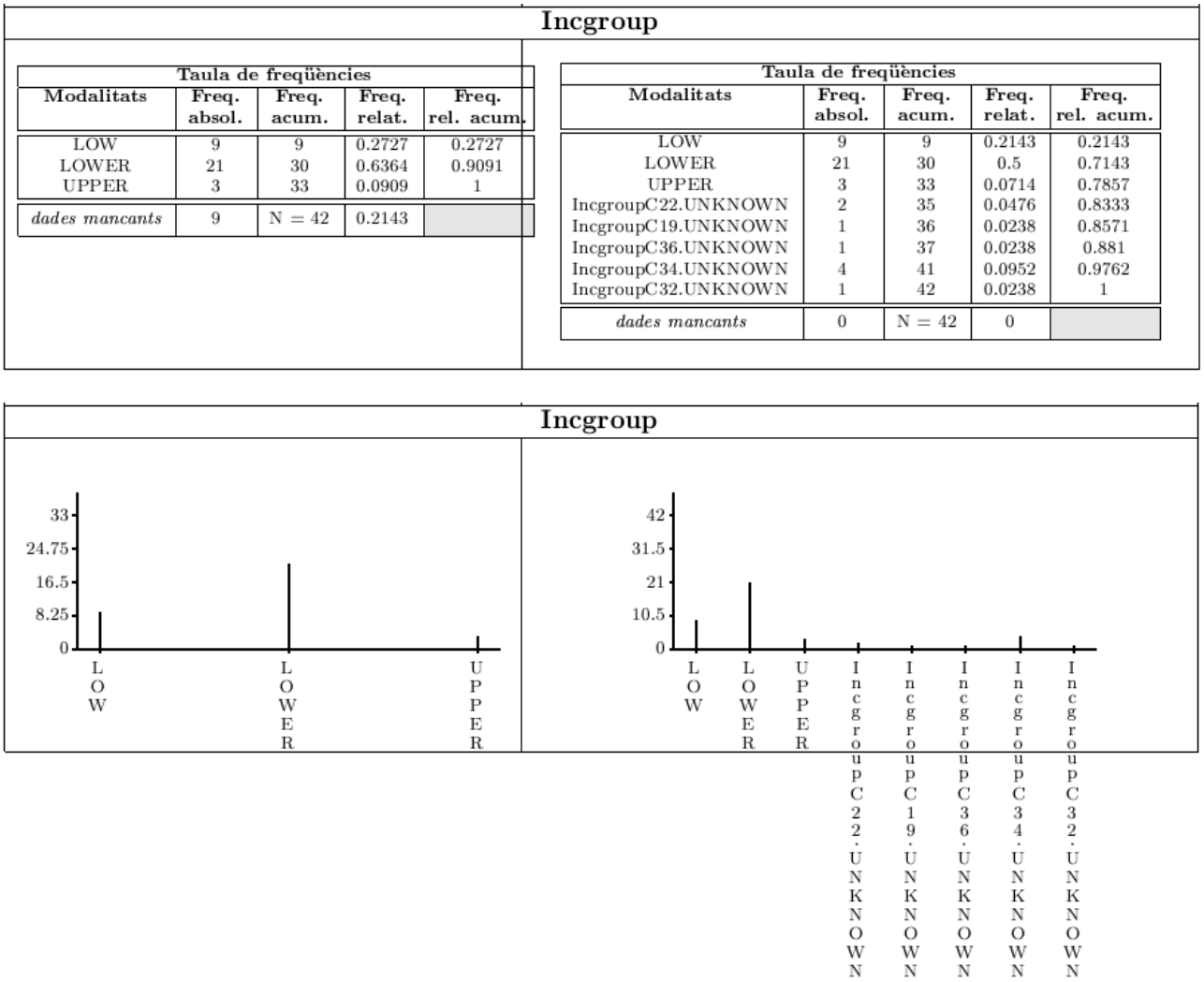


Figura No. 46: Descriptiva variable **Incgroup** antes y después imputación media global: Taula de frecuencias y diagrama de Barras

5.4.5 Clasificación con los datos resultantes luego de utilizar MIMMI

Una vez se ha completado el proceso de imputación, es posible continuar con el análisis de los datos de nuestra matriz, en este caso al igual que se realiza con el otro método, se procede a realizar la clasificación utilizando el mismo método, criterio y métrica de capítulo **5.3.2 Clasificación Utilizando Matriz Resultante,** es decir clustering jerárquico, con el criterio de selección de Ward y la distancia Mixta Gibert con alfa y beta automaticos

Una vez realizado el proceso, el sistema nos muestra el siguiente dendrograma (figura No. 31) de la clasificación.

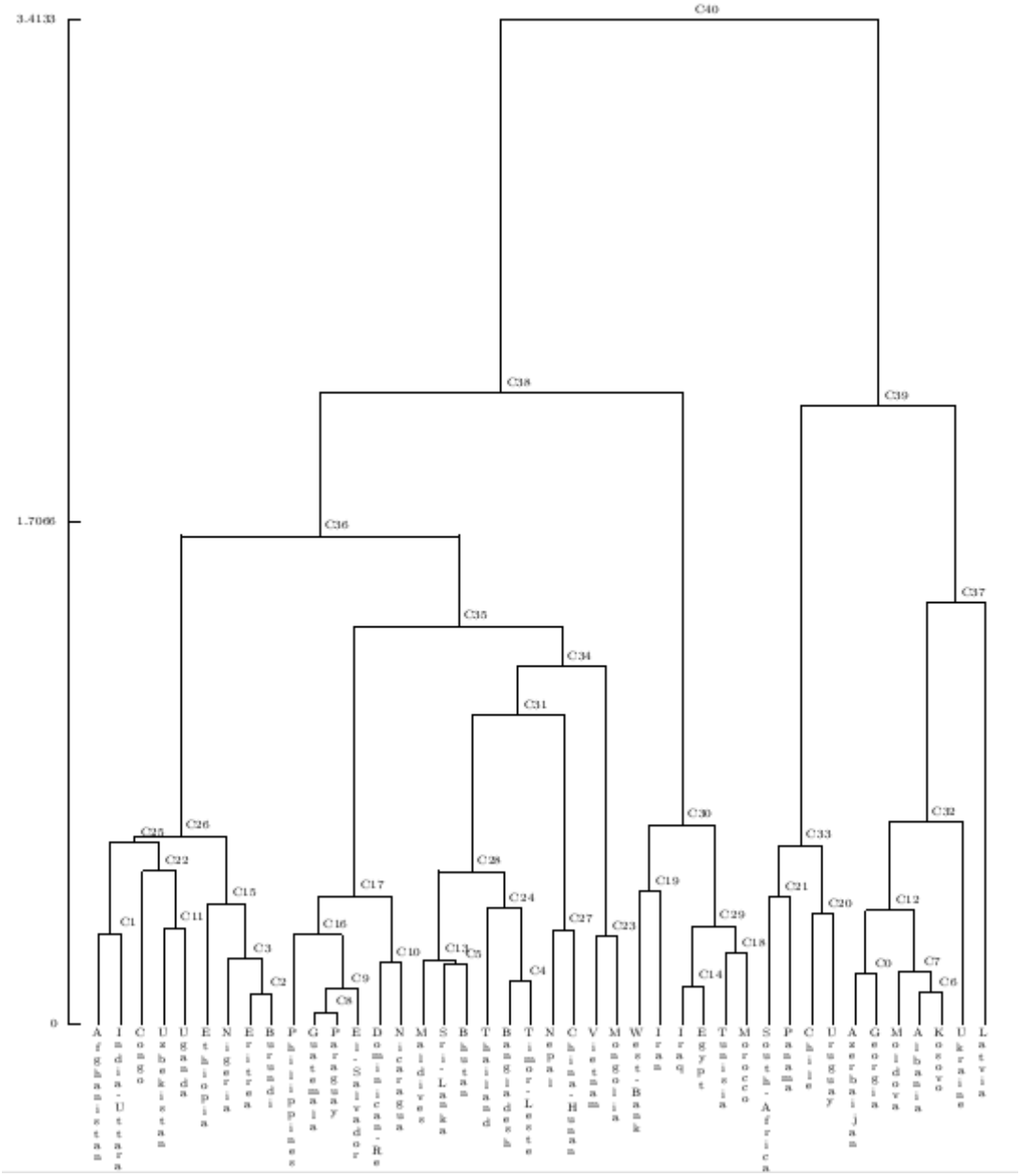


Figura No. 47: Dendrograma de resultado de la clasificación posterior a la imputación con MIMMI.

Ahora apoyándonos nuevamente en la funcionalidad de Java-KLASS, visualizamos al igual que con el método anterior el resultado de la clasificación, empezamos por el listado de objetos por clase (tabla No. 13), luego las Distribuciones condicionadas por clase (figuras No. 48 y 49) y por ultimo un mapa con la distribución de los países por clase.

Classe	Objectes
C26	Afghanistan, Burundi, Congo, Eritrea, Ethiopia, India-Uttaranc, Nigeria, Uganda, Uzbekistan
C32	Albania, Azerbaijan, Georgia, Kosovo, Moldova, Ukraine
C28	Bangladesh, Bhutan, Maldives, Sri-Lanka, Thailand, Timor-Leste
C33	Chile, Panama, South-Africa, Uruguay
C27	China-Hunan, Nepal
C17	Dominican-Repu, El-Salvador, Guatemala, Nicaragua, Paraguay, Philippines
C30	Egypt, Iran, Iraq, Morocco, Tunisia, West-Bank
Latvia	Latvia
C23	Mongolia, Vietnam

Tabla No. 13: Descripción extensional de clases resultantes de la clasificación

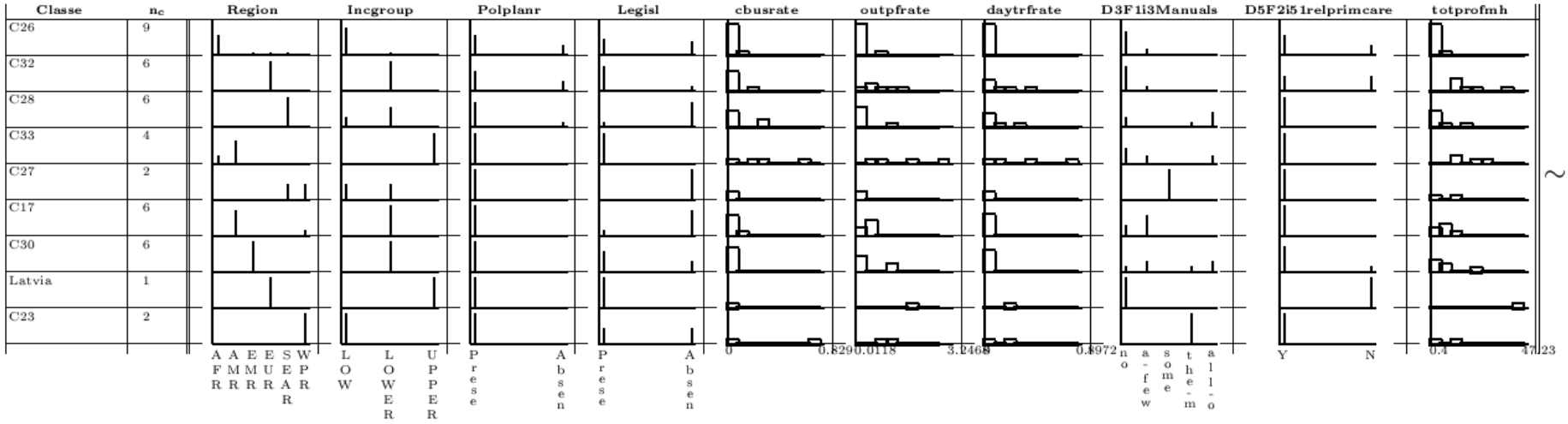


Figura No. 48 Distribuciones condicionadas por clase parte 1.

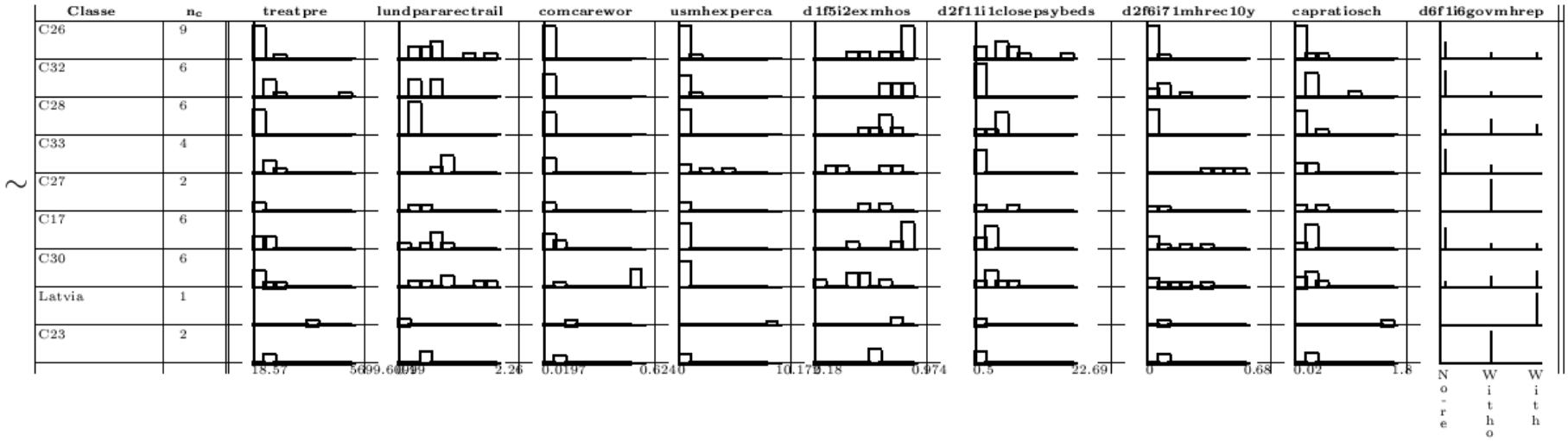


Figura No. 49: Distribuciones condicionadas por clase parte 2.

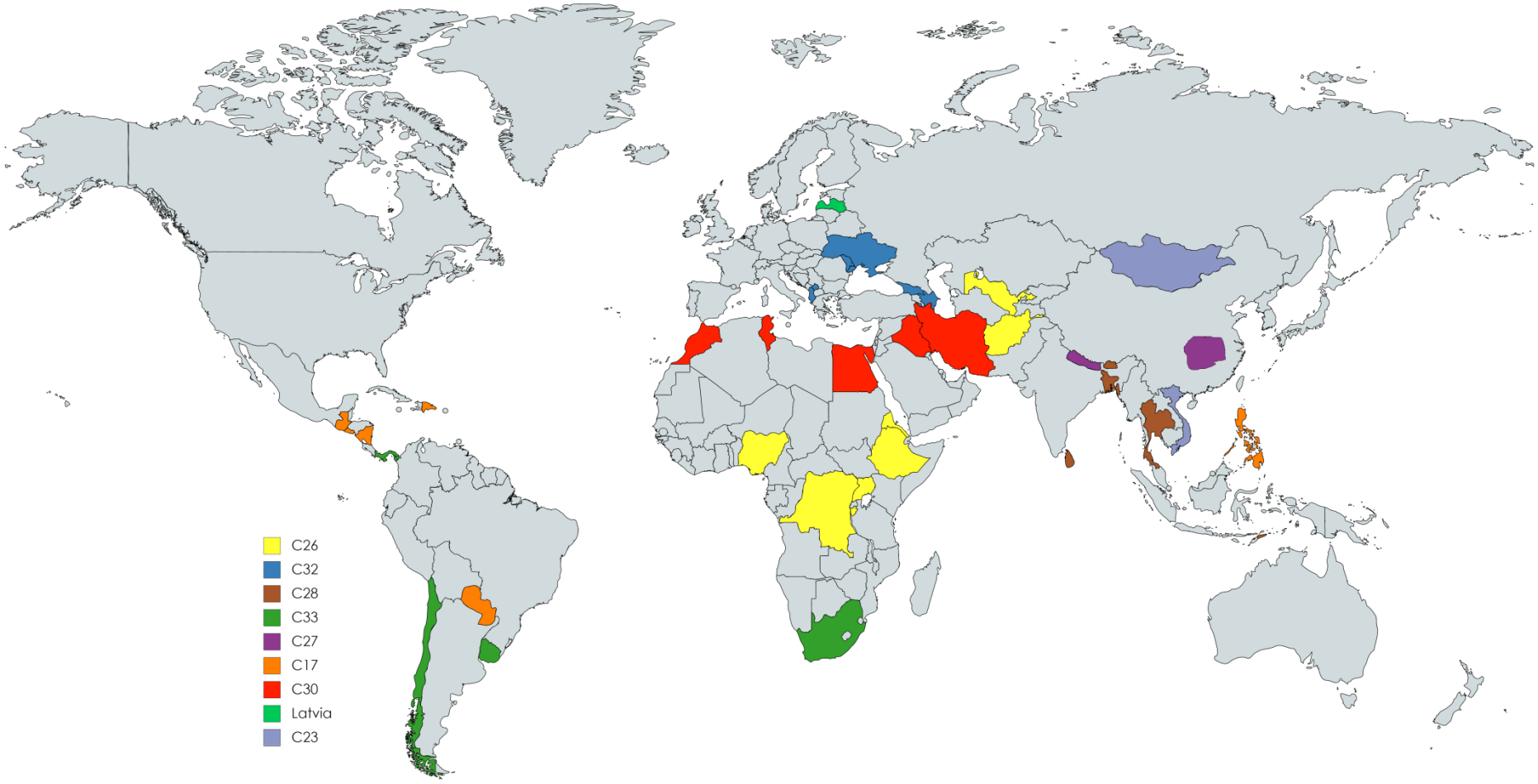


Figura No. 50: Representación de la clasificación de los países por clase

5.5 Comparativa de Resultados entre las clasificaciones Realizadas

Se puede apreciar en el apartado **5.4.2 Imputación de Valores Faltantes Utilizando Método de Media Global,** de este documento, las variables que fueron sometidas a la imputación por el método de media global, presentaron reducción en su desviación típica, lo que significa que dicha variable luego de la imputación refleja un comportamiento alterado del real.

La imputación con MIMMI, como se puede apreciar en las comparativas de las descriptivas de antes y después de la imputación, o en el Informe MIMMI que se encuentra como **Anexo No. 2** de este documento, distribuye los datos faltantes según la clase a la que pertenece el faltante, manteniendo el comportamiento y la varianza de la variable, dando como resultado una imputación más cercana al comportamiento original de la misma.

Una vez realizadas las imputaciones y posteriores clasificaciones podemos apreciar claramente que la clasificación realizada luego de la imputación por el método de media global, da como resultado un grupo de clases que no parece tener lógica, siendo clases más heterogéneas y que se alejan de lo que los expertos en [Gibert et al., 2010a] aceptaron como una clasificación correcta. Esto se puede apreciar ayudándose del recurso del mapa, en el que se señalaron los países por clase identificados por colores.

Por otro lado, la clasificación realizada luego de realizar la imputación con MIMMI, es mucho más lógica, se aprecia que las clases son más homogéneas e incluso al observarlas en el mapa en que se caracterizaron, tienen un sentido natural de asociación, el cual obviamente está apoyado por los datos de la matriz. En realidad, la clasificación obtenida después de imputar con el método MIMMI proporciona clusters que tienen una continuidad territorial muy interesante, puesto que referentes culturales de contexto comunes a los países agrupados en las mismas clases son perfectamente identificables en el mapa de clusters de la figura 50.

En efecto, es claramente observable que

- C26 corresponde al cuerno de África, con los países más pobres y por ello no incluye Sudáfrica
- C32 agrupa todos los países derivados de la desmembración de la antigua URSS, patrón ya reconocido en [Gibert et al., 2016b] y caracterizado por una aproximación muy orientada a la reclusión del paciente mental en hospitales frenopáticos y poco

popular en la actualidad. De hecho esta clase es la que menor valor del parámetro Lund presenta después del singleton representado por Latvia.

- C28 contiene los países asiáticos más pequeños, también identificados en [Gibert et al., 2016b].
- C27 Nepal y China-Hunan (región de China), contiene los países asiáticos más grandes, separados del grupo C23 por el valor de la población diagnosticada y atendida por cada 100 000 habitantes (treatpre), para este grupo este valor es considerablemente más pequeño.
- C23 Vietnam y Mongolia, agrupa el resto de países asiáticos grandes, con una tasa de población diagnosticada y atendida por cada 100 000 habitantes (treatpre), mayor al de C27.
- C33 Cono sur de América y Sudáfrica que son los países más desarrollados entre los LAMIC, con el mayor número de profesionales en salud mental después de Latvia y con más orientación a la inclusión social de los pacientes mentales (Lund)
- C17 reúne todos los países del Magreb, con más orientación a la inclusión social de los pacientes mentales (Lund)
- C30 Países de América Central y Paraguay (de corte muy diferenciado respecto a Chile o Argentina), con más orientación a la inclusión social de los pacientes mentales (Lund) aunque en esta clase se mantiene la mayor proporción de pacientes tratados en hospitales mentales todavía (comecarewor). Se está estudiando si esta alta proporción con el alto Lund implica que la población registra mayores niveles de severidad de los trastornos mentales en estos países
- Latvia, que es el más desarrollado de todos los países de esta base de datos y que de hecho poco después de la recogida de datos saltó a la franja de países desarrollados y ya no se incluye en la lista de países LAMIC. Este país se comporta como un outlier en esta base de datos debido a este hecho.

CAPÍTULO 6

CONCLUSIONES

Una vez realizado este Trabajo de Fin de Máster que tuvo como objetivo implementar un módulo avanzado de imputación de valores faltantes en el sistema Java-KLAS, debido a que el mismo contaba con opciones de imputación de valores faltantes útiles en situaciones muy restrictivas y no disponía hasta la fecha de un método de propósito general. Como hemos visto a lo largo de este trabajo es fundamental al momento de intentar extraer información de una matriz de datos.

De este proyecto se pueden obtener las siguientes conclusiones:

- El tratamiento de faltantes es una tarea fundamental al momento de pre procesar información para extraer conocimiento válido de un matriz de datos que presentan esta particularidad, y teniendo en cuenta que es un fenómeno muy común se puede decir que es un paso obligado en todos los proyectos de extracción de conocimiento basados en Minería de Datos y otros procesos de análisis de datos y ciencia de datos

- Métodos como Listwise Deletion, que eliminan filas que poseen faltantes pueden reducir la cantidad de datos válidos de la matriz y sobre todo excluir una subpoblación específica del análisis, además de introducir sesgos que pueden generar peores análisis, y el hecho que sea casi un estándar en los paquetes informáticos que se utilizan para el tratamiento de datos, no los hace más válidos.

- Que si bien existen métodos de imputación bien fundamentados, basados en hipótesis estadísticas fuertes y que proporcionan técnicas apropiadas para imputar valores faltantes de una matriz de datos como MLE y EM, éstos requieren tiempo, experiencia y conocimientos técnicos considerables para una correcta implementación, lo cual complica su implementación y utilización en la práctica, y el hecho de que partan de ciertos supuestos de partida restringe su uso a aquéllas bases de datos donde se pueda verificar el cumplimiento de dichos supuestos, lo cual raramente se da en la realidad.

- En el caso de MICE, que cuenta también con fundamentos estadísticos sólidos y permiten realizar de forma adecuada la imputación de faltantes, presenta una grave restricción, ya que si todas las variables de la matriz tienen valores faltantes, no se puede aplicar este método, y si bien este escenario tal vez no sea tan común, sí es probable que ocurra, lo que dejaría al método sin capacidad de realizar su propósito. Además, requiere de software especial para consensuar los resultados del análisis de las diferentes matrices resultantes de la imputación por MICE, lo cual sólo está disponible en la mayoría de las implementaciones para métodos de regresión y parecidos. Si se opta por consensuar la propia matriz de datos, se acaban obteniendo estimaciones de la media global en las casillas de la matriz de datos, lo cual reduce la aportación de MICE de forma considerable.

- En el caso de KNN, basado en distancias, se dispone solamente de implementaciones basadas en la distancia euclídea, y si bien es un método de propósito general inspirado en case based-reasoning y basado en la metáfora del razonamiento por analogía, en sus implementaciones disponibles se limita a casos donde sólo hay variables numéricas, y adolece de la misma limitación que MICE, no siendo viable si todas ellas presentan valores faltantes.

- Que el método Mixed Intelligent-Multivariate Missing Imputation es una muy buena opción para la imputación de faltantes, ya que incorpora desde las fases iniciales de la imputación el conocimiento de expertos (que seleccionan las variables a considerar en la fase 1 de acuerdo con su relevancia), no altera de forma drástica la variabilidad de las variables como otros métodos, permite tratar los faltantes tanto en variables categóricas como numéricas, de una forma adecuada, implica un nivel relativamente bajo de complejidad y uso de recursos. Este método representa una opción de propósito general con un buen compromiso entre el tiempo que requiere,

y su costo computacional asociado y la precisión de las imputaciones generadas, que ha mostrado buenos resultados en algunas aplicaciones reales previas [Gibert et al., 2016b], MIMMI es un método prácticamente automático, donde la única decisión que recae en el experto es con qué variables relevantes se puede realizar la fase1 y si se utiliza media o mediana como estimador de imputación. Las recomendaciones del sistema ayudan al usuario a determinar este segundo parámetro

- Al implantar el método MIMMI en Java-KLASS, este sistema mantuvo y mejoró su funcionalidad, además de ampliar sus capacidad para realizar tareas relacionadas a la Minería de Datos, ya que al contar con un método apropiado para el tratamiento de faltantes como los es Mixed Intelligent-Multivariate Missing Imputation, que puede imputar valores faltantes tanto en variables cualitativas como cuantitativas, de forma eficiente, manteniendo la varianza y otras características propias de cada variable, se amplía el universo de bases de datos con las que se puede trabajar con este software directamente, sin necesidad de realizar la imputación de faltantes previamente con otras herramientas, robusteciendo aún más sus capacidades. Además, con esta intervención, JavaKLASS se convierte en el primer software que incorpora una implentación de MIMMI

- El informe MIMMI que se ha diseñado en este trabajo y desarrollado para mostrar los detalles de la imputación por MIMMI en las aplicaciones, permite conocer todas las características y particularidades con las que se efectuó la imputación: la descripción de las variables utilizadas en la fase1, parámetros y métricas con las que se realizó la clasificación auxiliar; la descriptiva completa de las variables antes y después de ser imputadas; los valores por los que fueron imputados los faltantes de cada variable en cada clase y el número de imputaciones realizadas por variable y por clase. Por lo que se puede afirmar que el informe permite conocer todos los pormenores de la imputación de faltantes, y nos ayuda a comprender y visualizar cómo se llevó a cabo el mismo, permitiendo a partir de dicho informe reproducir de ser necesario el método para obtener los mismos resultados.

Por último, se ha realizado una comparativa sobre un caso real de datos de la OMS, la cual permitió evidenciar las ventajas de usar MIMMI respecto a los procedimientos anteriores con los que contaba el sistema Java-KLASS. El resultado fue contundente, mientras la clasificación realizada con MIMMI, se apegaba muchísimo a la realizada y validada por

expertos en [Gibert et al., 2010a], mostrando robustez, facilidad de uso y sobre todo precisión y efectividad, el resultado por el método anterior se alejó mucho de lo esperado dando como resultado clases menos heterogéneas, lo que se pude apreciar comprando el mapa en el que se representaron los países por clase de las dos clasificaciones. Cabe decir que la clasificación que se ha realizado en este proyecto no ha tenido en cuenta el conocimiento a priori que los expertos aportaron para inyectar al proceso de clasificación en [Gibert et al., 2010a] a través del algoritmo de la clasificación basada en reglas con la finalidad de poder estudiar estrictamente el impacto de cambiar el método de imputación de datos. Esto explica que las clases resultantes no sean idénticas a las obtenidas en [Gibert et al., 2010a], pero la esencia de la clasificación se mantiene usando la implementación de MIMMI realizada con un método de clasificación de vecinos recíprocos encadenados, a pesar de no haber inyectado conocimiento a priori del experto en el proceso.

Como trabajo futuro se apuntan algunos aspectos de continuidad de la línea de investigación muy directamente relacionados con el trabajo que se ha desarrollado en este proyecto

- Actualmente para obtener el número de clusters resultantes de la clasificación auxiliar, el usuario decide este número después de visualizar y analizar con su conocimiento el dendrograma que se le presenta como resultado de la clasificación. Este proceso puede mejorarse, ya que existen métodos que permiten calcular el número óptimo de clusters en una clasificación, una de ellas es el Índice de Calinski & Harabasz [Calinski and Harabasz, 1974], que permita identificar qué tan compactos son los clusters y qué tan alejados entre ellos están, para así poder calcular cual es número exacto de clusters de una clasificación [Bonaccorso, 2017]. Sin embargo, para el caso general de clasificar con variables numéricas y categóricas simultaneamente, Calinski Harabasz no sería directamente aplicable. Actualmente se está trabajando en el diseño de un índice generalizado
- Al momento la implementación realizada de MIMMI en Java-KLASS, solo permite una única elección entre media o mediana para todas las variables que se van a imputar. Ahora bien, si nos enfrentamos a una matriz de datos que tiene por un lado variables que presentan una distribución simétrica y por otro otras que presentan una distribución muy asimétrica, esto nos podría llegar a representar un problema, ya que si utilizamos la media para aquellas que son muy asimétricas, el valor no sería el más adecuado, y así mismo para una variable simétrica sería preferible utilizar el valor de la media que es el estimador estándar, y responde a un álgebra bastante más manejable e independiente de la ordenación de las observaciones, siempre costosa desde el punto de vista computacional. Por ello se debería poder determinar que variables se desean imputar con la media y cuales

con la mediana al momento de realizar la configuración de la imputación, y así asegurar que el valor que se utiliza para la imputación adecuado para cada variable.

- En un orden más genérico, sería interesante estudiar cómo se comportaría KNN respecto a MIMMI, pero esto solo sería viable en un entorno limitado a variables numéricas donde no todas tuvieran valores faltantes. Para poder realizar una comparativa general, sería necesario implementar una versión de KNN que aceptara distancias parametrizadas, entre las que se pudieran incluir distancias mixtas para datos heterogéneos resolviendo así la primera limitación y preferiblemente aquéllas que admitan valores faltantes como podría ser la distancia de Gower, con tal de resolver la segunda limitación. Java-KLASS representa un excelente contexto para realizar este paso puesto que dispone de un módulo de distancias de amplio espectro ya implementadas para realizar esta generalización del KNN

BIBLIOGRAFÍA

[Azur et al. 2011] Azur M., Stuart E., Frangakis C. and Leaf P. (2011) Multiple imputation by chained equations: What is it and how does it work?, International Journal of Methods in Psychiatric Research.

[Bayona, 2000] Bayona, S. (2000). Descriptiva de dades y de classes. PFC Facultad d'Informática, UPC.

[Bonaccorso, 2017] Bonaccorso G. (2017). Assessing clustering optimality with instability index, urlhttps://www.bonaccorso.eu/2017/08/03/assessing-clustering-optimality-instability-index/ Accedido 12-04-2019.

[Calinski and Harabasz, 1974] Calinski, T. and Harabasz, J. (1974) A dendrite method for cluster analysis. Communications in Statistics - Simulation and Computation, 3(1):1—27, Taylor and Francis

[Campello et al., 2013] Campello R, Moulavi D., Sander J. (2013) Density-Based Clustering Based on Hierarchical Density Estimates.

[Castillejo, 1996] Castillejo, X. (1996). Un entorn de treball per a klass. PFC Facultat d'Informatica, UPC.

[Dagnino, 2014] Dagnino, J. (2014). Datos Faltantes (MISSING VALUES) urlhttp://revistachilenadeanestesia.cl/datos-faltantes-missing-values/ Accedido 15-01-2019.

[Gibert, 1991] Gibert, K. (1991). Klass. estudi d'un sistema d'ajuda al tractament estadístic de grans bases de dades. Master's thesis, Master's thesis, UPC.

[Gibert, 1995] Gibert, K. (1995). L'us de la informació simbólica en l'automatització del tractament estadístic de dominis poc estructurats.

[Gibert and Salvador, 2000] Gibert, K. and Salvador, A. (2000). Aproximación difusa a la identificación de situaciones características en el tratamiento de aguas residuales. In X Congreso Español sobre tecnologías y lógica fuzzy.

[Gibert et al., 2010a] Gibert, K., García-Alonso, C., and Salvador-Carulla, L. (2010a). Integrating clinicians, knowledge and data: expert-based cooperative analysis in healthcare decision support. Health research policy and systems.

[Gibert et al., 2010b] Gibert, K., Martín, J., and . Salvador-Carulla, L. (2010b). Who-aims: General clustering. Technical part.

[Gibert, 2013] Gibert, K. (2013). Mixed intelligent-multivariate missing imputation. International Journal of Computer Mathematics, 91(1):85-96.

[Gibert et al., 2016a] Gibert, K., Sánchez-Marríee, M., and Izquierdo, J. (2016a). A survey on pre-processing techniques: Relevant issues in the context of environmental data mining. AI Communications, 29(6):627-663.

[Gibert et al., 2016b] Gibert K., Salvador Carulla L., Morris J., Lora A., Saxena S. (2016b) The data mining approach as the starting point for Mental Health policy-making in low and middle income countries at World Health Organization.

[Gibert and Conti, 2016] Gibert, K., Conti, D. (2016). On the understanding of profiles by means of post-processing techniques: an application to financial assets. Environmental Engineering and Management. International Journal of Computer Mathematics

[Little et al., 2002] Roderick J. A. Little , Donald B. Rubin (2002). Statistical Analysis with Missing Data, 2nd Edition.

[Medina and Galvan, 2007] Medina F., Galvan M., (2007) Imputación de Datos: Teoría y Práctica.

[Márquez and Martín, 1997] Márquez, J. and Martín, J. (1997). La clasificación automática en las ciencias de la salud. PFC. Facultat de Matematiques i Estadística, UPC.

[Molinero, 2003] Molinero, L. (2003) ¿Qué es el método de estimación de máxima verosimilitud y cómo se interpreta? urlhttps://www.seh-lelha.org/que-es-el-metodo-de-estimacion-de-maxima-verosimilitud-y-como-se-interpreta/ Accedido 20-02-2019.

[Nakai and Weiming, 2011] Nakai M., Weiming K. (2011). Review of the Methods for Handling Missing Data in Longitudinal Data Analysis. Journal of Math. Analysis, Vol. 5, 2011, no. 1, 1 – 13.

[Otero, 2011] Otero, D. (2011) Imputación de datos faltantes en un Sistema de Información sobre Conductas de Riesgo.

[Pacual, 2010] Pacual G. (2010) Algoritmos de Agrupamiento basados en densidad y Validación de clusters.

[Soley-Bori, 2013] Soley-Bori, M. (2013). Dealing with missing data: key assumptions and methods for applied analysis.

[Rodas-Osollo, 2004] Rodas-Osollo, J. (2004). Knowledge discovery in repeated and very short serial measures with a blocking factor. AI Communications, 17(3):175-178.

[Tubau, 1999] Tubau, X. (1999). Sobre el comportament de les metriques mixtes en algorismes de clustering.

[Vázquez and Gibert, 2002] Vázquez, F. and Gibert, K. (2002). Robustness of class prediction depending on references partition in-ill-structured domains. 8th. In Iberoamerican Conference on Artificial Intelligence. Sevilla, España.

[Zhou et al., 2014] Zhou X., Zhou C., Lui D., Ding X. (2014) Applied Missing Data Analysis in the Health Sciences. Hoboken, New Jersey: John Wiley & Sons.

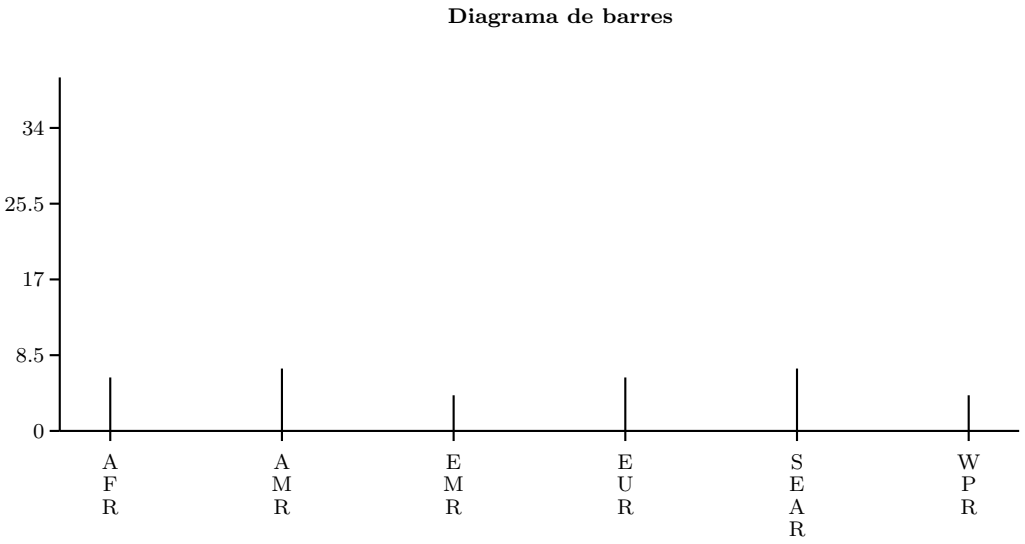
ANEXOS

Anexo 1:

Descriptiva de las variables seleccionadas de la matriz de datos “WHO-AIMS v2.2”

Anàlisi descriptiva univariant

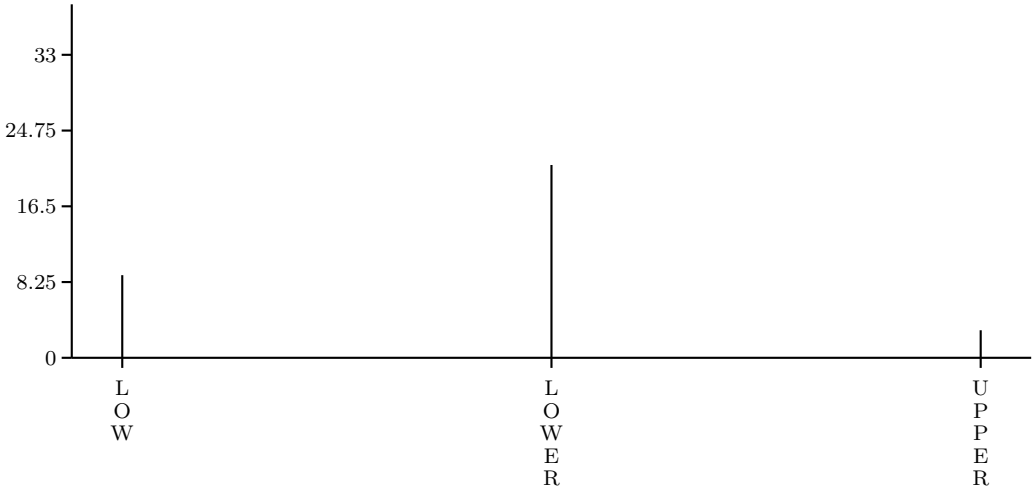
Variable Region



Taula de freqüències				
Modalitats	**Freq. absol.**	**Freq. acum.**	**Freq. relat.**	**Freq. rel. acum.**
AFR	6	6	0.1765	0.1765
AMR	7	13	0.2059	0.3824
EMR	4	17	0.1176	0.5
EUR	6	23	0.1765	0.6765
SEAR	7	30	0.2059	0.8824
WPR	4	34	0.1176	1
dades mancants	8	N = 42	0.1905	

Variable Incgroup

Diagrama de barres



Taula de freqüències				
Modalitats	**Freq. absol.**	**Freq. acum.**	**Freq. relat.**	**Freq. rel. acum.**
LOW	9	9	0.2727	0.2727
LOWER	21	30	0.6364	0.9091
UPPER	3	33	0.0909	1
dades mancants	9	N = 42	0.2143	

Variable Polplanr

Diagrama de barres

Taula de freqüències				
Modalitats	**Freq. absol.**	**Freq. acum.**	**Freq. relat.**	**Freq. rel. acum.**
Present	36	36	0.8571	0.8571
Absent	6	42	0.1429	1
dades mancants	0	N = 42	0	

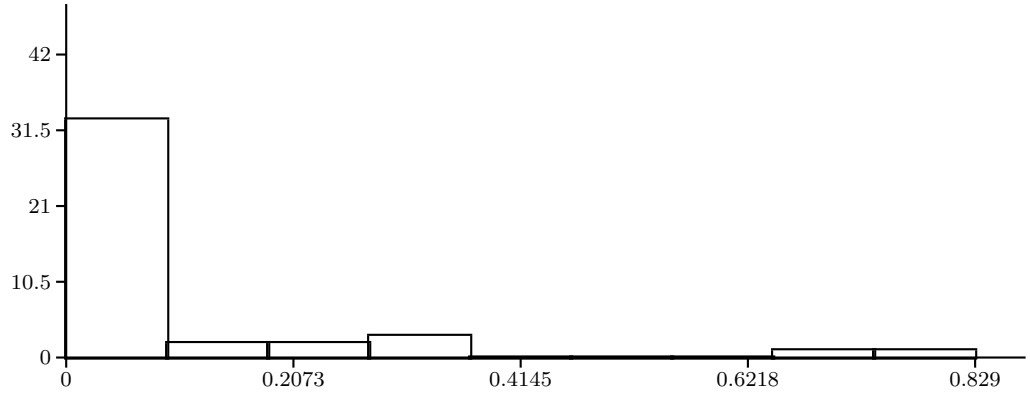
Variable Legisl

Diagrama de barres

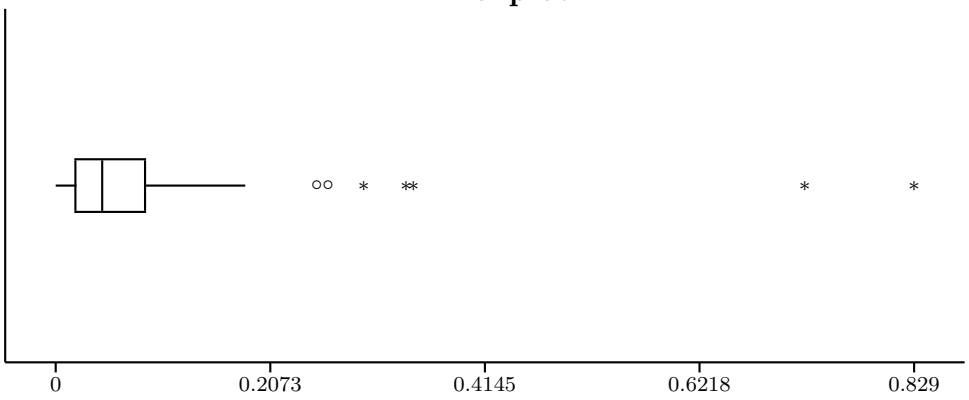
Taula de freqüències				
Modalitats	**Freq. absol.**	**Freq. acum.**	**Freq. relat.**	**Freq. rel. acum.**
Present	22	22	0.5238	0.5238
Absent	20	42	0.4762	1
dades mancants	0	N = 42	0	

Variable cbusrate

Histograma



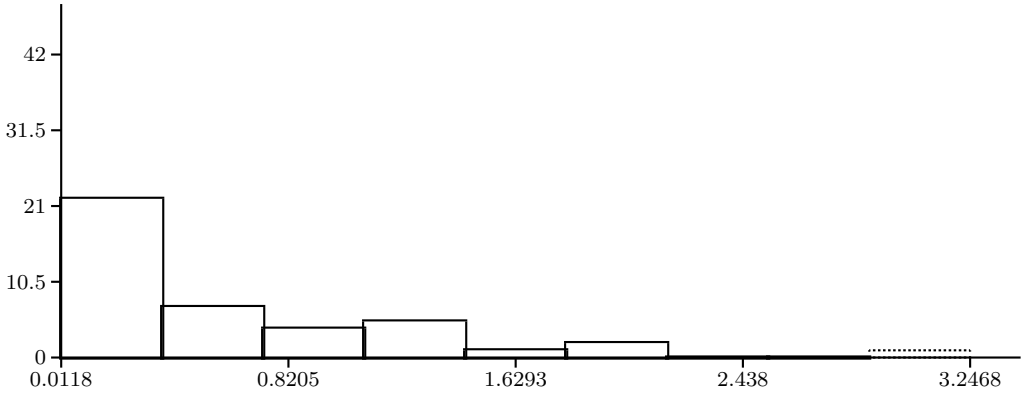
Boxplot



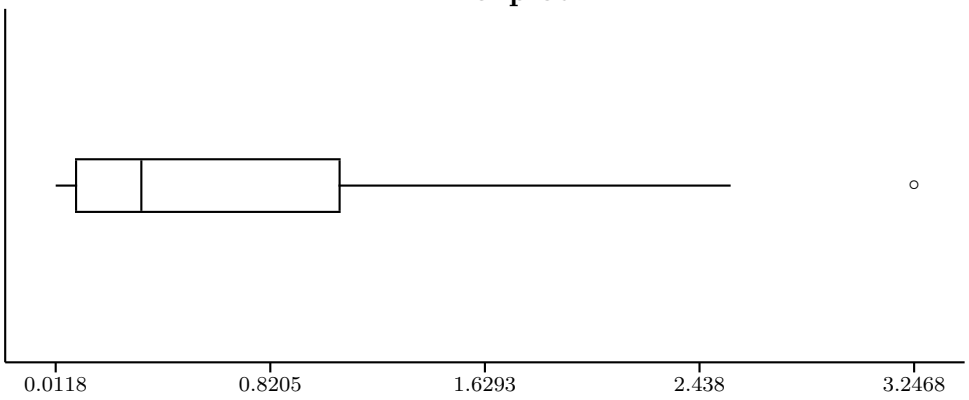
Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	0.1052
Mediana	0.0449
Primer quartil (Q1)	0.0201
Tercer quartil (Q3)	0.0853
Mínim	0
Màxim	0.829
Quasi-desviació típica	0.1772
Coeficient de variació	1.6639

Variable outpfrate

Histograma



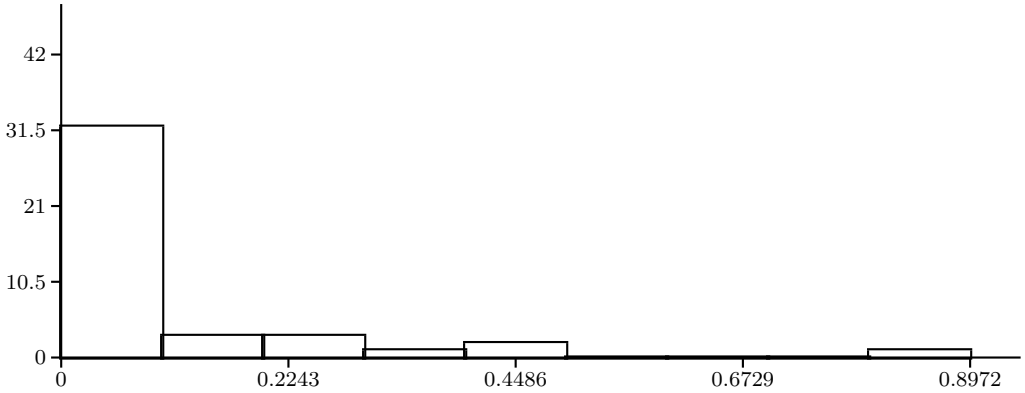
Boxplot

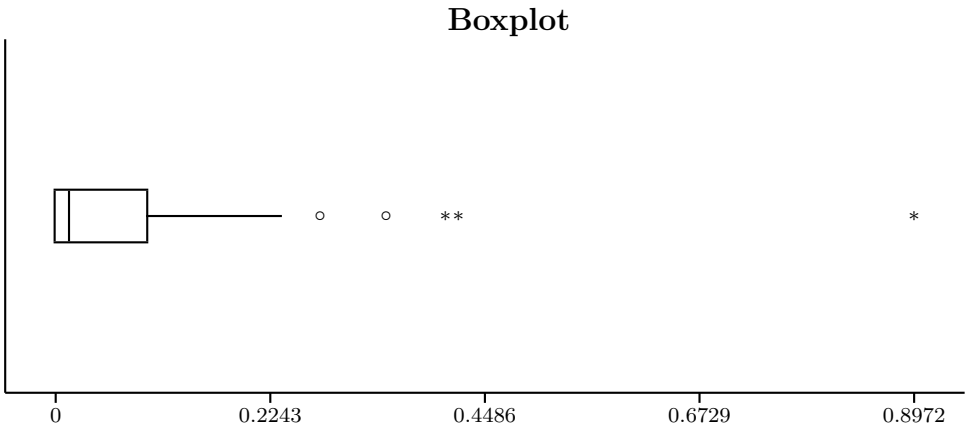


Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	0.6209
Mediana	0.3346
Primer quartil (Q1)	0.0926
Tercer quartil (Q3)	1.0777
Mínim	0.0118
Màxim	3.2468
Quasi-desviació típica	0.6992
Coeficient de variació	1.1125

Variable daytrfrate

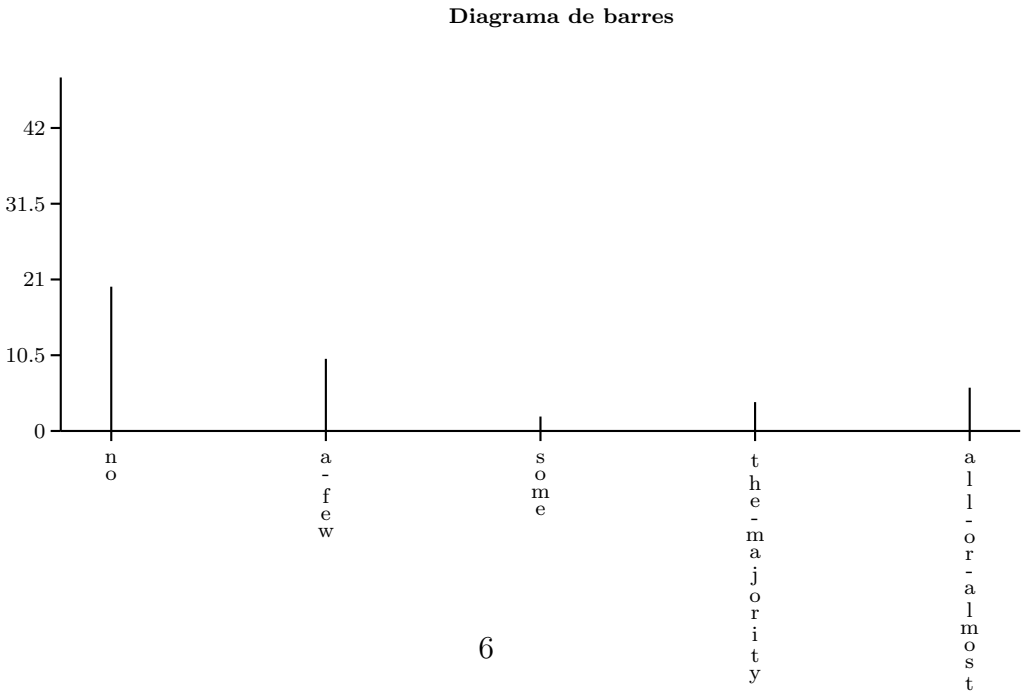
Histograma





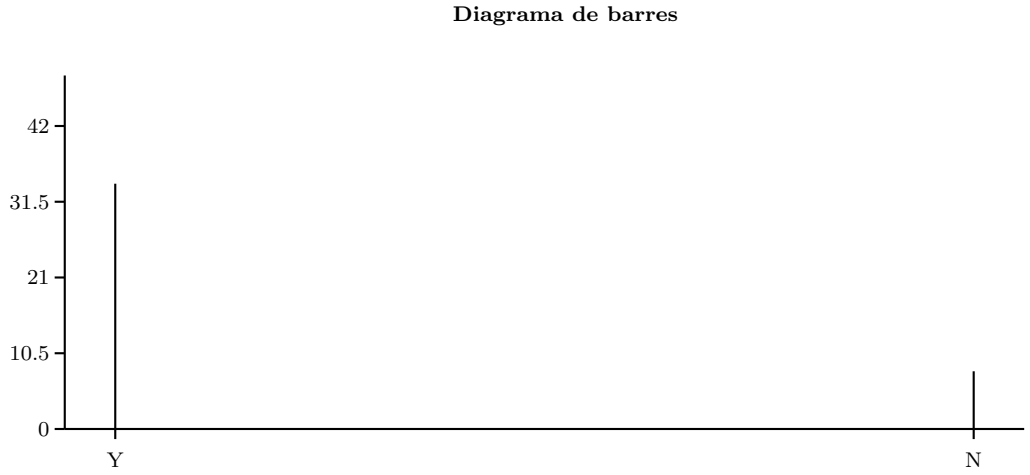
Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	0.0933
Mediana	0.0139
Primer quartil (Q1)	0
Tercer quartil (Q3)	0.0946
Mínim	0
Màxim	0.8972
Quasi-desviació típica	0.1703
Coeficient de variació	1.8026

Variable D3F1i3Manuals



Taula de freqüències				
Modalitats	**Freq. absol.**	**Freq. acum.**	**Freq. relat.**	**Freq. rel. acum.**
no	20	20	0.4762	0.4762
a-few	10	30	0.2381	0.7143
some	2	32	0.0476	0.7619
the-majority	4	36	0.0952	0.8571
all-or-almost	6	42	0.1429	1
dades mancants	0	N = 42	0	

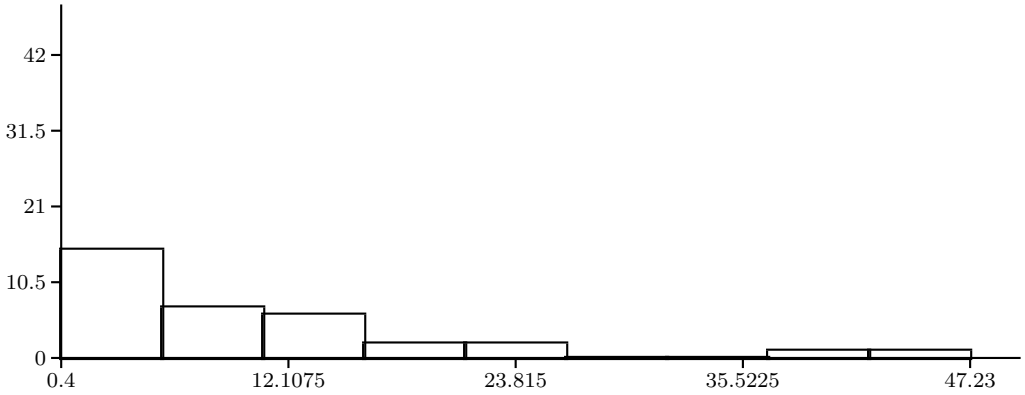
Variable D5F2i51relprimcare



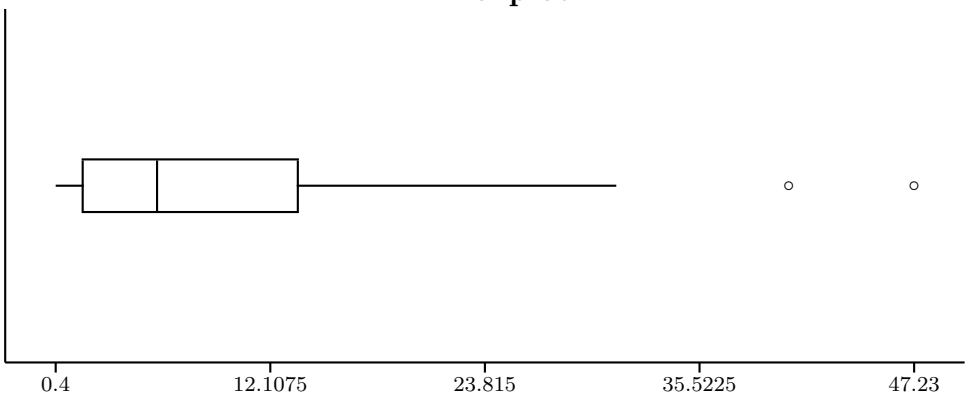
Taula de freqüències				
Modalitats	**Freq. absol.**	**Freq. acum.**	**Freq. relat.**	**Freq. rel. acum.**
Y	34	34	0.8095	0.8095
N	8	42	0.1905	1
dades mancants	0	N = 42	0	

Variable totprofmh

Histograma



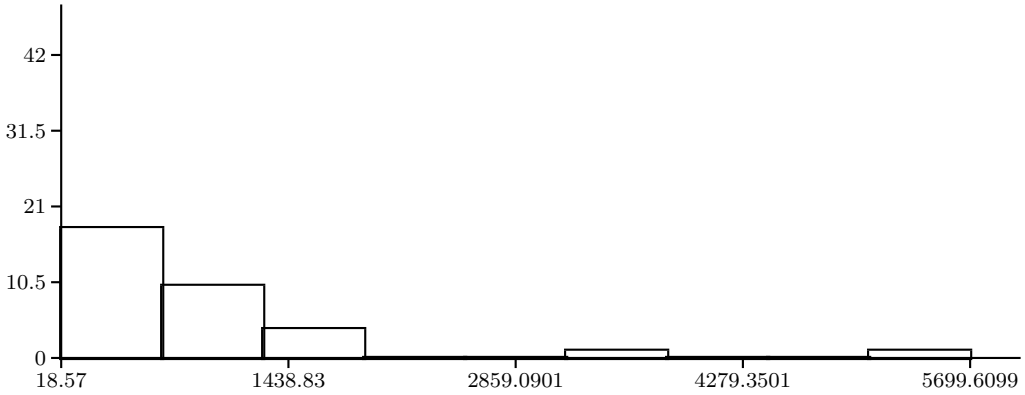
Boxplot



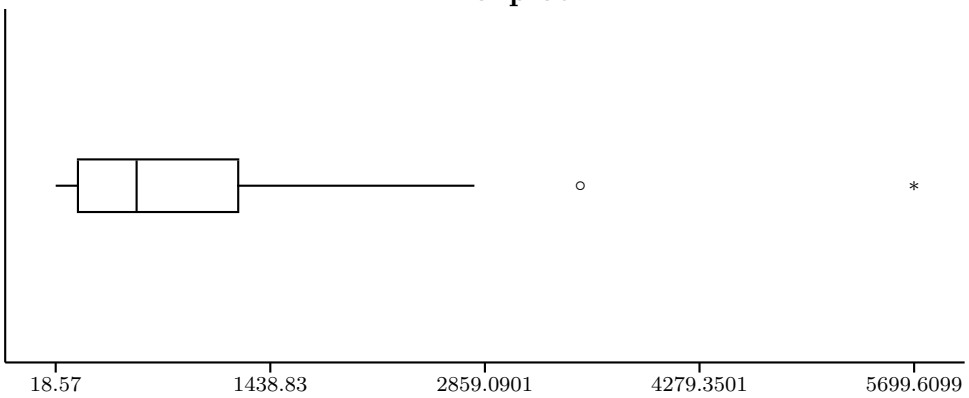
Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	8
Nombre d'observacions útils	34
Mitjana	9.7289
Mediana	5.93
Primer quartil (Q1)	1.93
Tercer quartil (Q3)	13.5507
Mínim	0.4
Màxim	47.23
Quasi-desviació típica	10.8552
Coeficient de variació	1.0992

Variable treatpre

Histograma



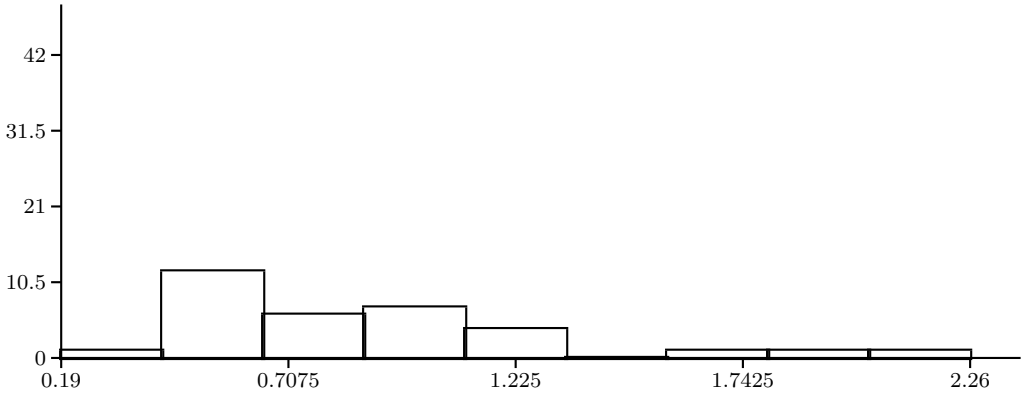
Boxplot



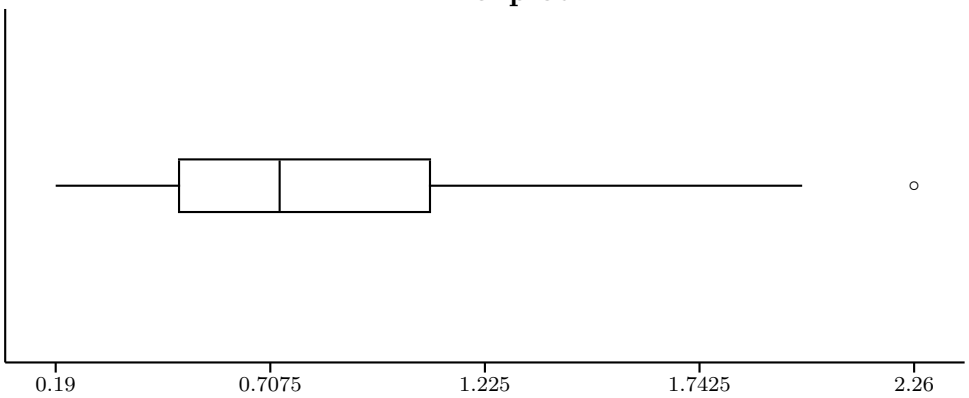
Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	8
Nombre d'observacions útils	34
Mitjana	865.43
Mediana	553.9938
Primer quartil (Q1)	172.77
Tercer quartil (Q3)	1219.4614
Mínim	18.57
Màxim	5699.6099
Quasi-desviació típica	1118.3401
Coeficient de variació	1.2731

Variable lundpararectrail

Histograma



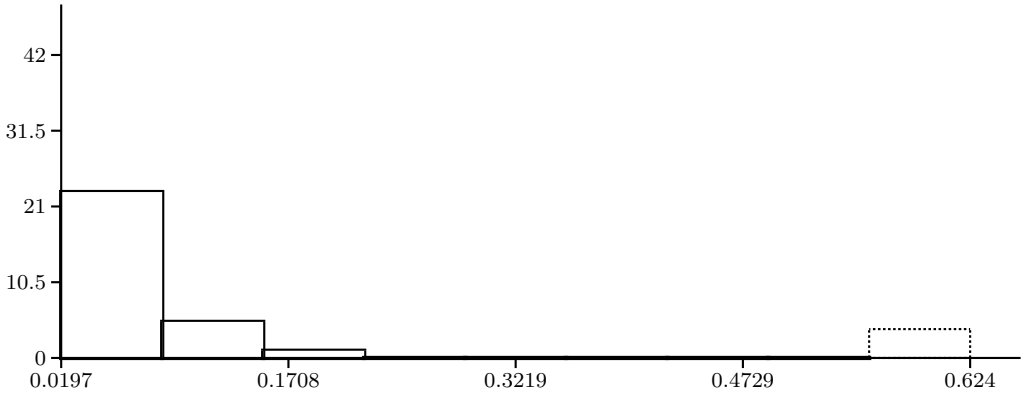
Boxplot



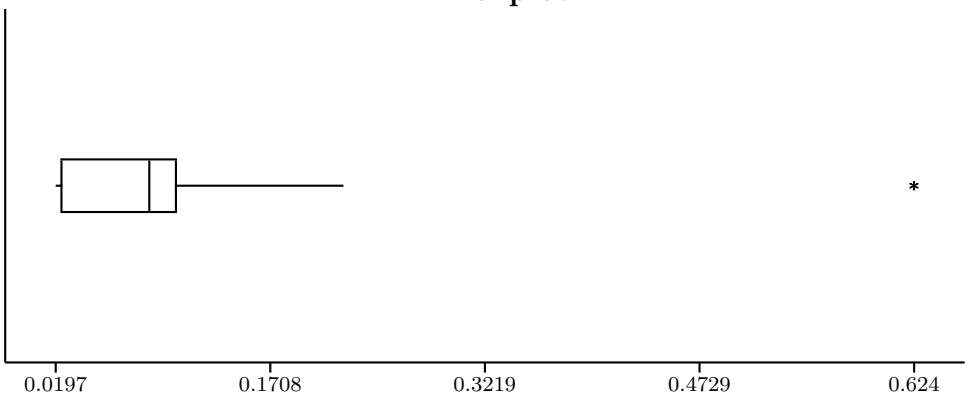
Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	9
Nombre d'observacions útils	33
Mitjana	0.8523
Mediana	0.73
Primer quartil (Q1)	0.49
Tercer quartil (Q3)	1.09
Mínim	0.19
Màxim	2.26
Quasi-desviació típica	0.44
Coeficient de variació	0.5083

Variable comcarewor

Histograma



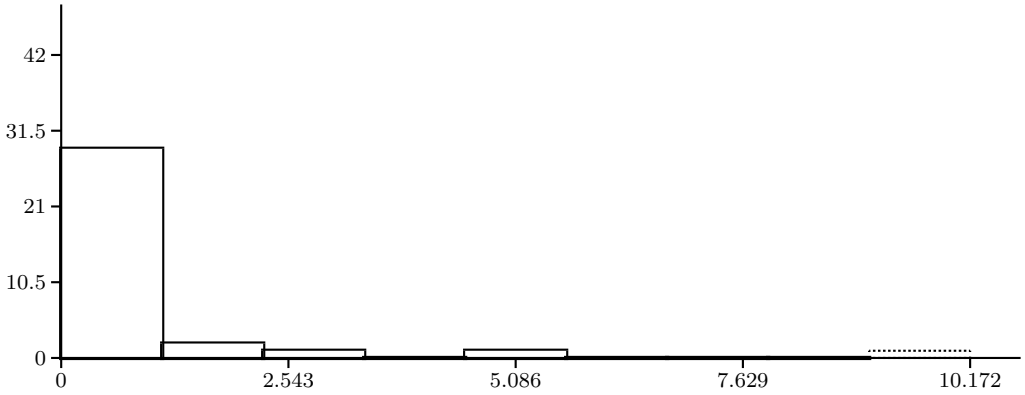
Boxplot



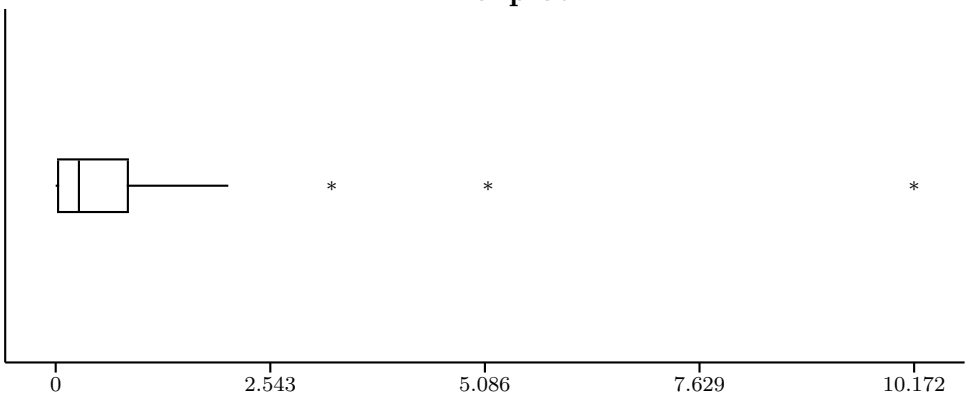
Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	9
Nombre d'observacions útils	33
Mitjana	0.1313
Mediana	0.0856
Primer quartil (Q1)	0.0245
Tercer quartil (Q3)	0.1036
Mínim	0.0197
Màxim	0.624
Quasi-desviació típica	0.1905
Coeficient de variació	1.4285

Variable usmhexperca

Histograma



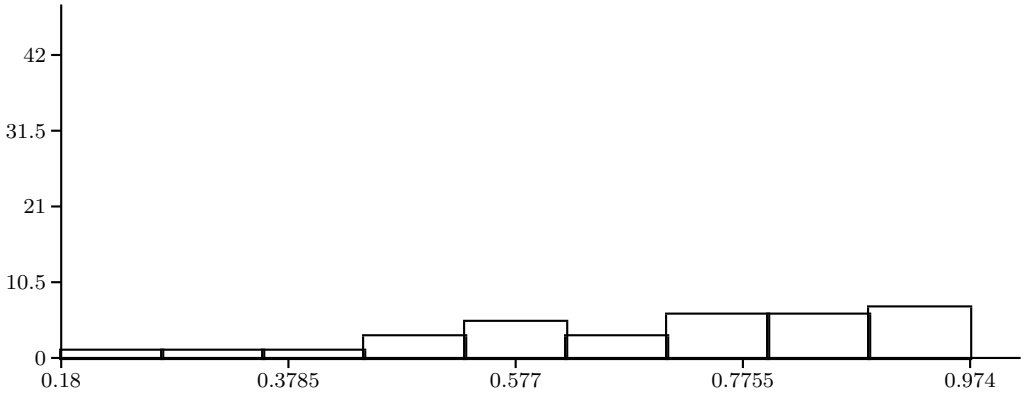
Boxplot



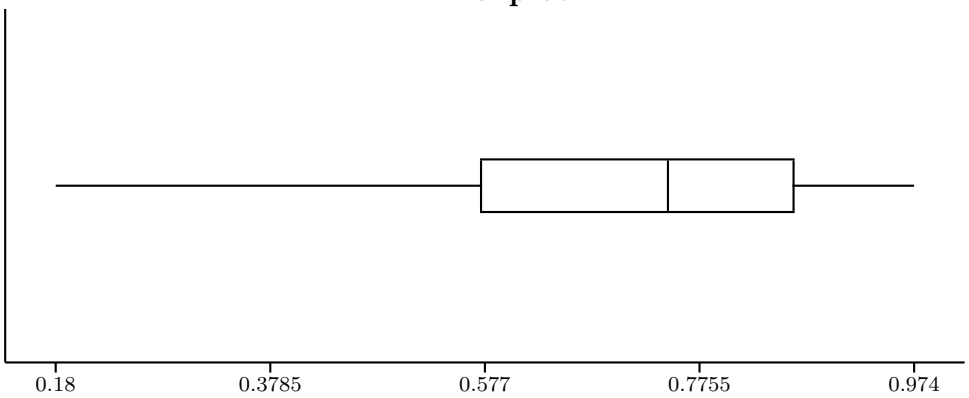
Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	8
Nombre d'observacions útils	34
Mitjana	0.9031
Mediana	0.2751
Primer quartil (Q1)	0.041
Tercer quartil (Q3)	0.8426
Mínim	0
Màxim	10.172
Quasi-desviació típica	1.9355
Coeficient de variació	2.1114

Variable d1f5i2exmhos

Histograma



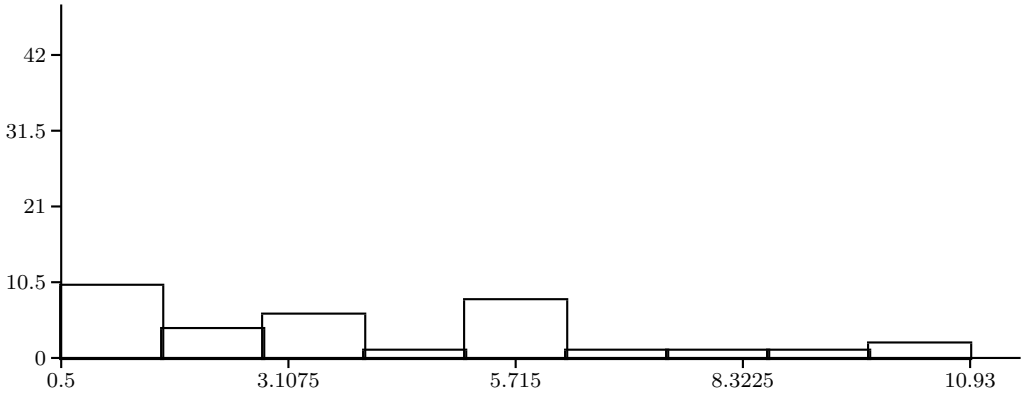
Boxplot



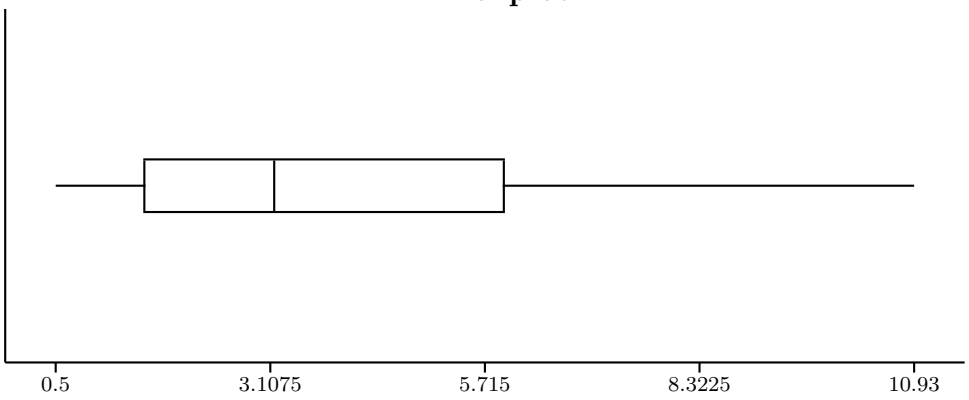
Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	9
Nombre d'observacions útils	33
Mitjana	0.705
Mediana	0.7463
Primer quartil (Q1)	0.5744
Tercer quartil (Q3)	0.8615
Mínim	0.18
Màxim	0.974
Quasi-desviació típica	0.1913
Coeficient de variació	0.2672

Variable d2f11i1closepsybeds

Histograma



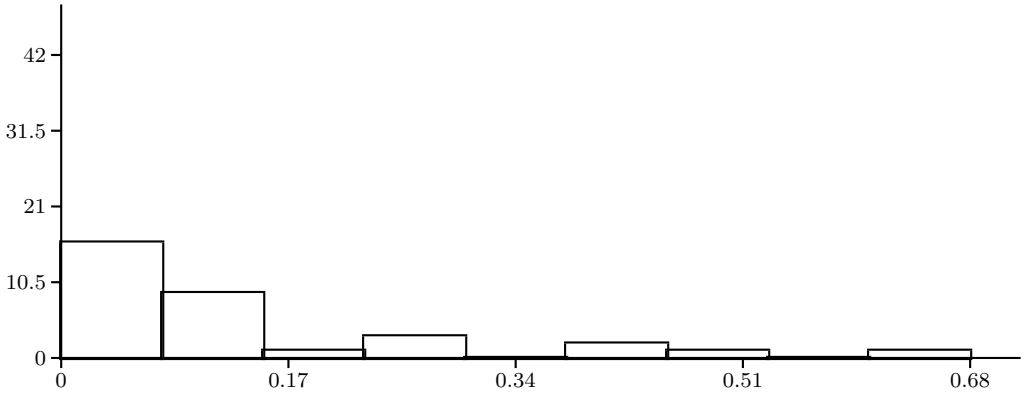
Boxplot



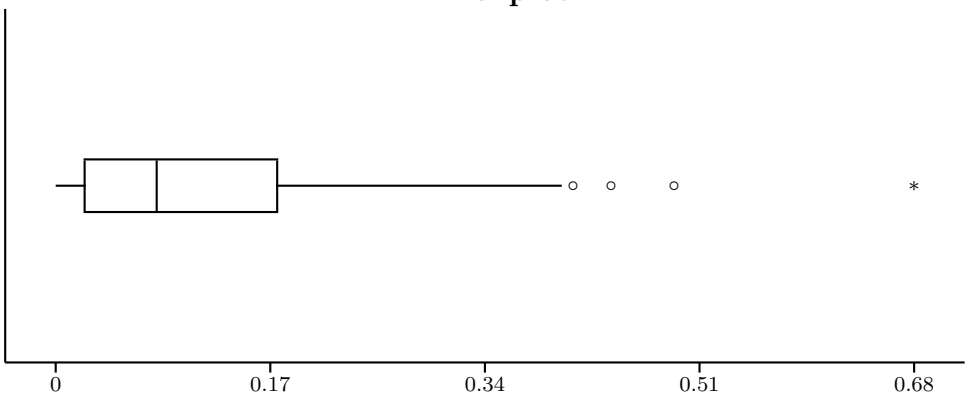
Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	8
Nombre d'observacions útils	34
Mitjana	4.0169
Mediana	3.155
Primer quartil (Q1)	1.59
Tercer quartil (Q3)	5.932
Mínim	0.5
Màxim	10.93
Quasi-desviació típica	2.8456
Coeficient de variació	0.6979

Variable d2f6i71mhrec10y

Histograma



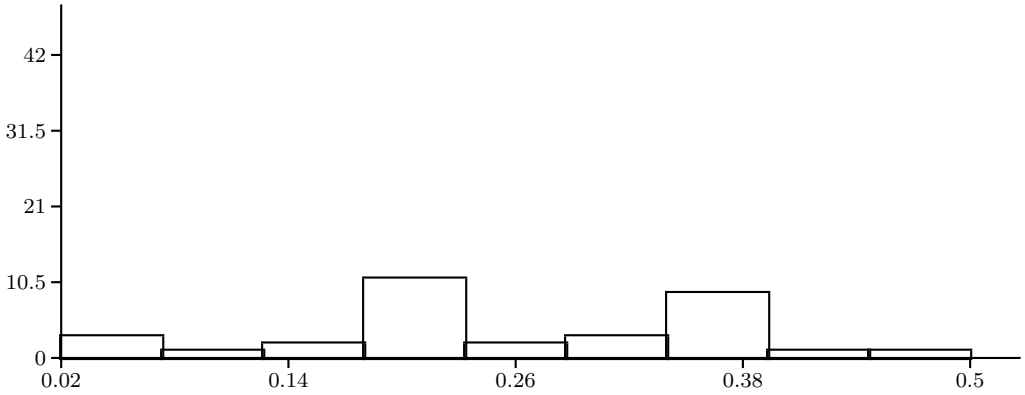
Boxplot



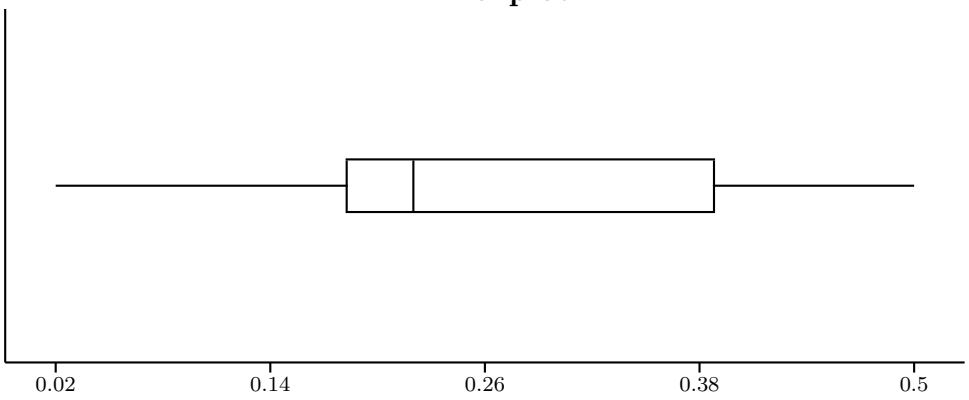
Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	9
Nombre d'observacions útils	33
Mitjana	0.1354
Mediana	0.08
Primer quartil (Q1)	0.0237
Tercer quartil (Q3)	0.1746
Mínim	0
Màxim	0.68
Quasi-desviació típica	0.1653
Coeficient de variació	1.2017

Variable capratiosch

Histograma



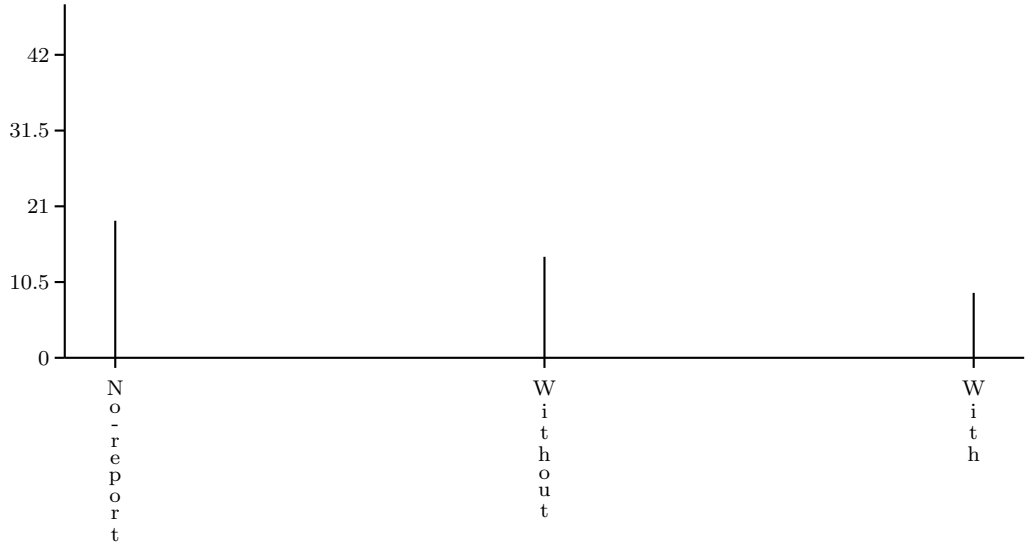
Boxplot



Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	9
Nombre d'observacions útils	33
Mitjana	0.2569
Mediana	0.22
Primer quartil (Q1)	0.1833
Tercer quartil (Q3)	0.3875
Mínim	0.02
Màxim	0.5
Quasi-desviació típica	0.1236
Coeficient de variació	0.4737

Variable d6f1i6govmhrep

Diagrama de barres



Taula de freqüències				
Modalitats	**Freq. absol.**	**Freq. acum.**	**Freq. relat.**	**Freq. rel. acum.**
No-report	19	19	0.4524	0.4524
Without	14	33	0.3333	0.7857
With	9	42	0.2143	1
dades mancants	0	N = 42	0	

Anexo 2:

Informe MIMMI: Descriptiva de la imputación de la matriz de datos del Caso del Estudio

Informe MIMMI

Fase 1: Classificació Auxiliar

Variables seleccionades per la fase de Classificació auxiliar

- Polplanr
- Legisl
- D1F5i5recAntipshic
- D1F5i6recAntidepr
- cbusrate
- mhrate
- outpfrate
- daytrfrate
- D4F1i11psychi
- D4F1i12doctors
- D4F1i13nurses
- D4F1i14psycho
- D4F1i15socwork
- D3F1i3Manuals
- D5F2i51relprimcare

Anàlisi descriptiva de les variables anteriors

Variable Polplanr

Taula de freqüències				
Modalitats	**Freq. absol.**	**Freq. acum.**	**Freq. relat.**	**Freq. rel. acum.**
Present	36	36	0.8571	0.8571
Absent	6	42	0.1429	1
dades mancants	0	N = 42	0	

Diagrama de barres

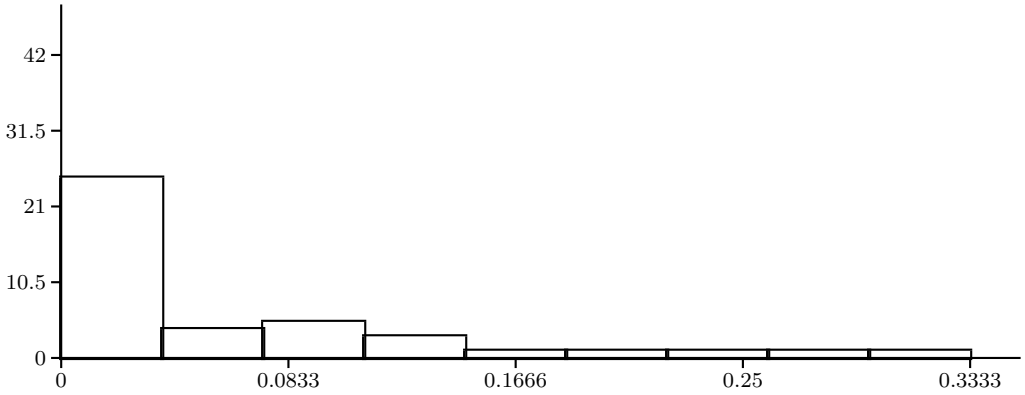
Variable Legisl

Taula de freqüències				
Modalitats	**Freq. absol.**	**Freq. acum.**	**Freq. relat.**	**Freq. rel. acum.**
Present	22	22	0.5238	0.5238
Absent	20	42	0.4762	1
dades mancants	0	N = 42	0	

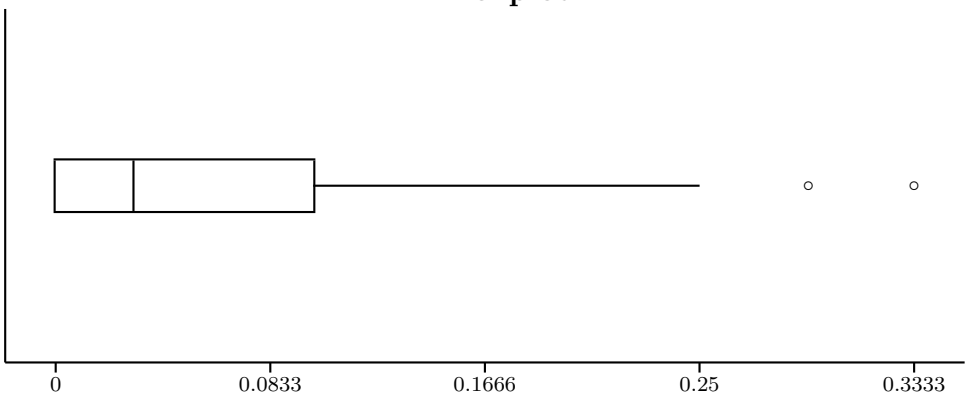
Diagrama de barres

Variable D1F5i5recAntipshic

Histograma



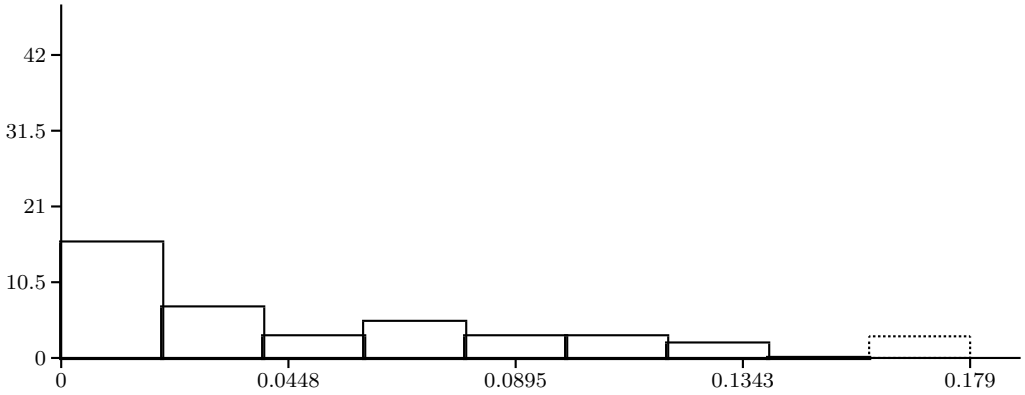
Boxplot



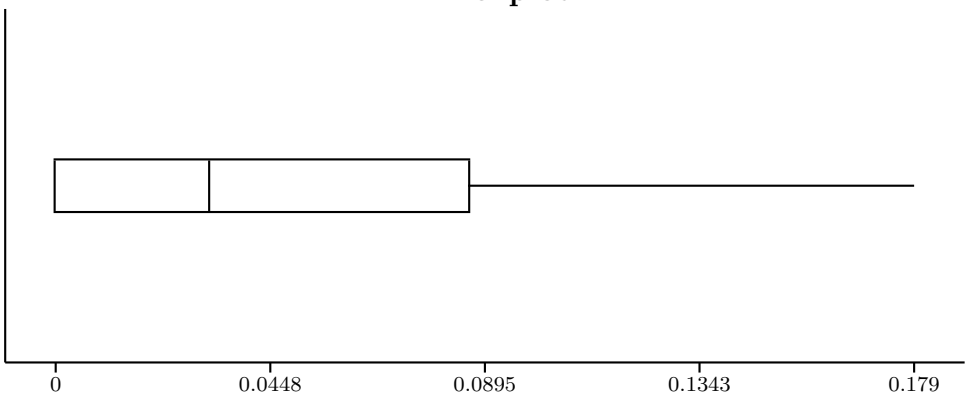
Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	0.0612
Mediana	0.0302
Primer quartil (Q1)	0
Tercer quartil (Q3)	0.1
Mínim	0
Màxim	0.3333
Quasi-desviació típica	0.0812
Coeficient de variació	1.3112

Variable D1F5i6recAntidepr

Histograma



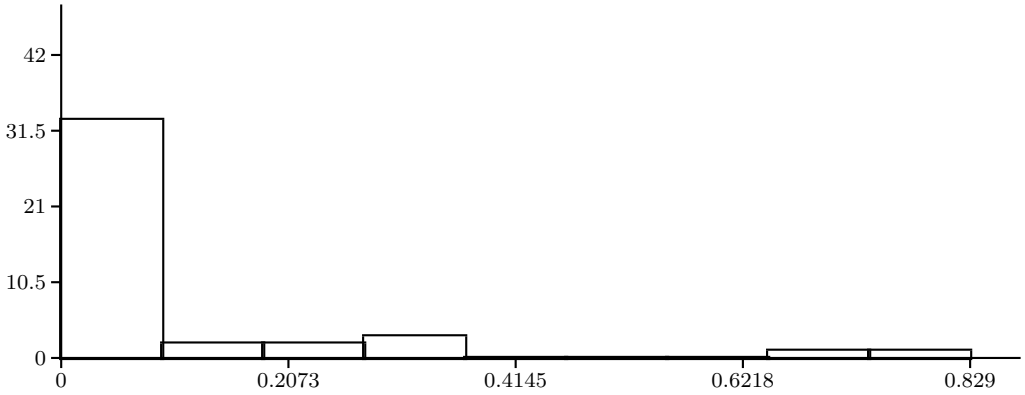
Boxplot



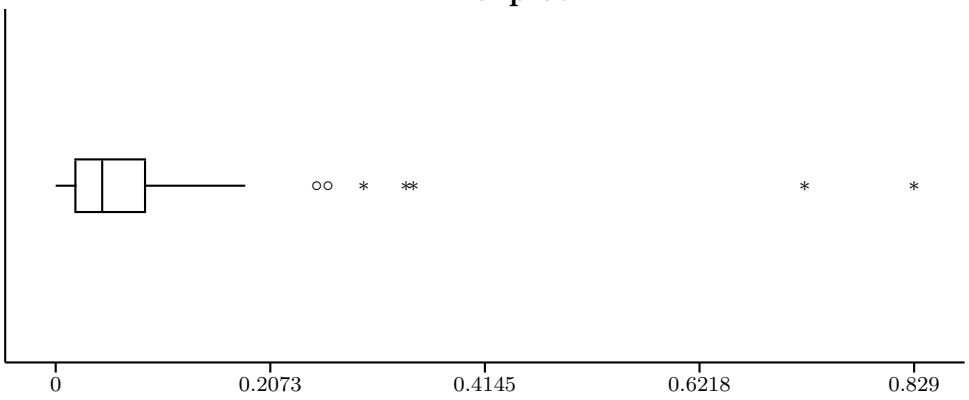
Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	0.0505
Mediana	0.032
Primer quartil (Q1)	0
Tercer quartil (Q3)	0.086
Mínim	0
Màxim	0.179
Quasi-desviació típica	0.0528
Coeficient de variació	1.0336

Variable cbusrate

Histograma



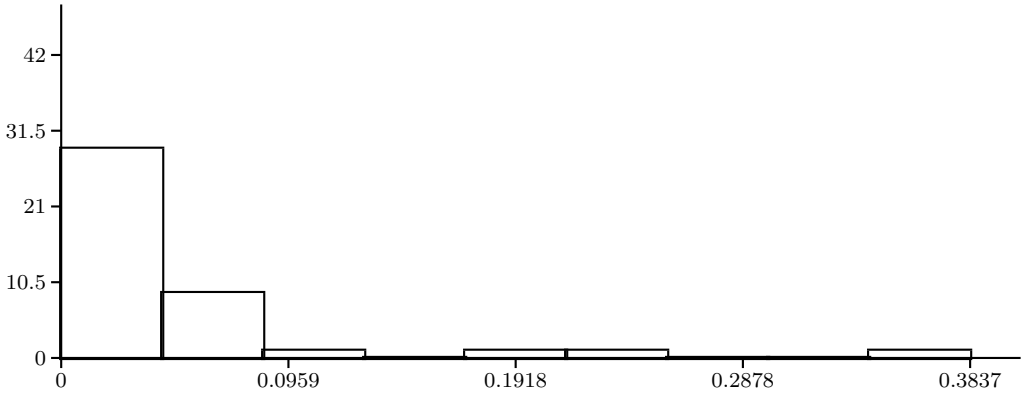
Boxplot



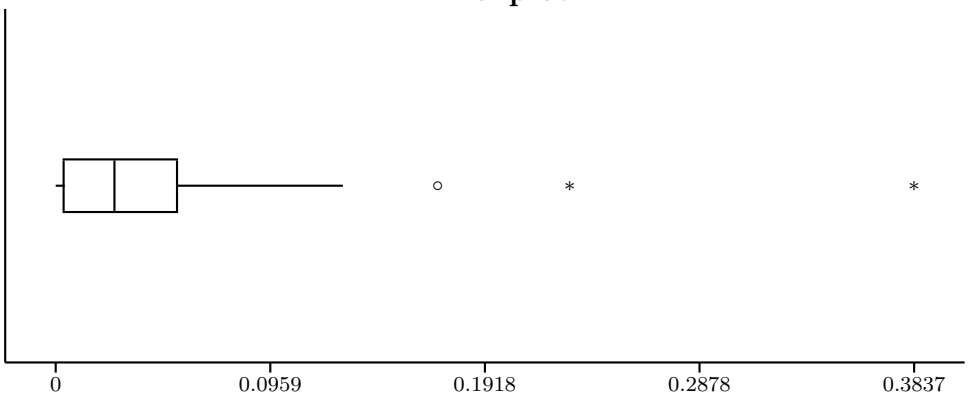
Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	0.1052
Mediana	0.0449
Primer quartil (Q1)	0.0201
Tercer quartil (Q3)	0.0853
Mínim	0
Màxim	0.829
Quasi-desviació típica	0.1772
Coeficient de variació	1.6639

Variable mhrate

Histograma



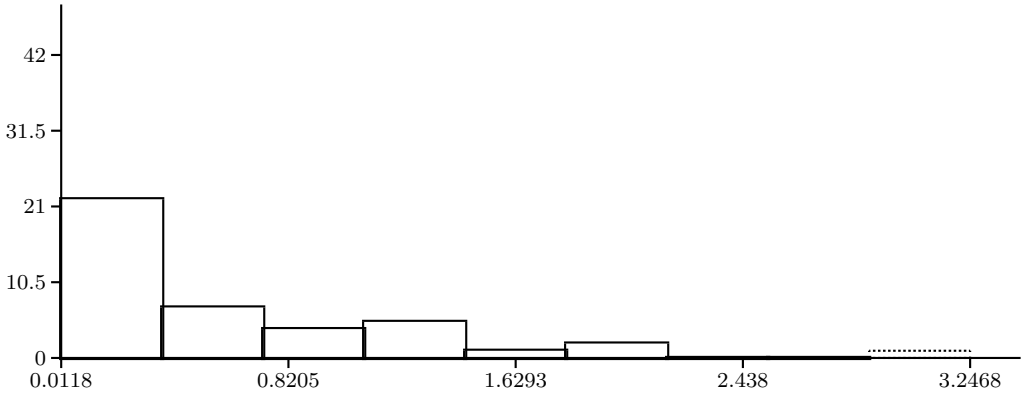
Boxplot



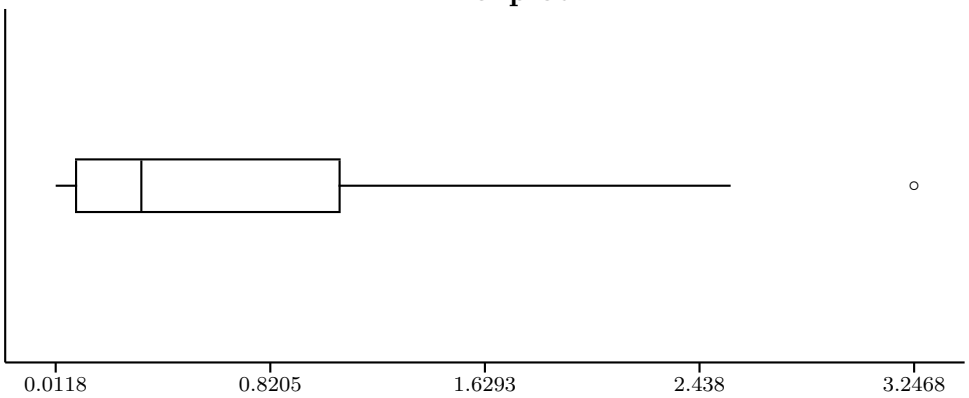
Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	0.0435
Mediana	0.0263
Primer quartil (Q1)	0.004
Tercer quartil (Q3)	0.0537
Mínim	0
Màxim	0.3837
Quasi-desviació típica	0.0704
Coeficient de variació	1.5995

Variable outpfrate

Histograma



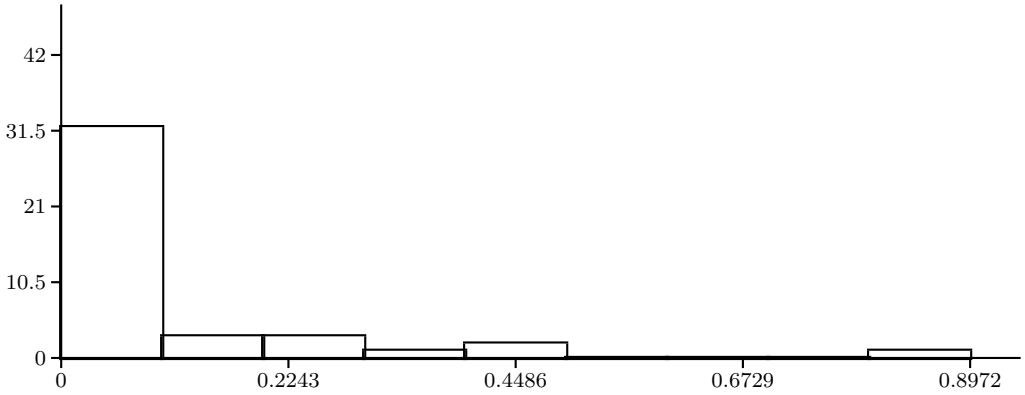
Boxplot



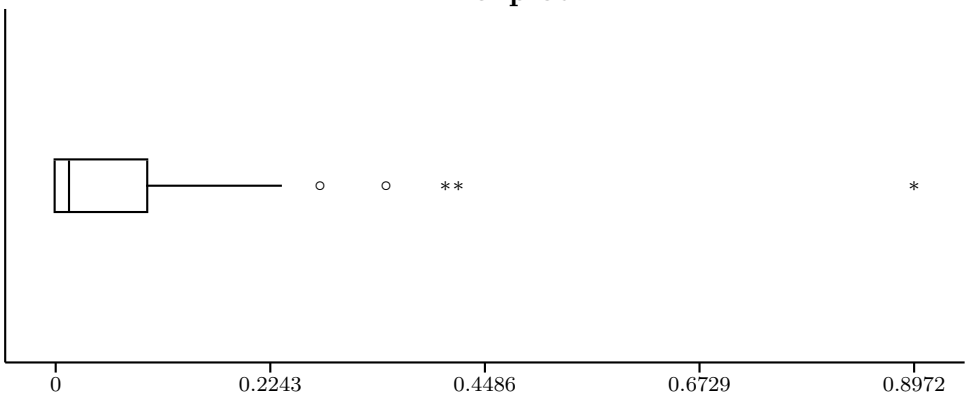
Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	0.6209
Mediana	0.3346
Primer quartil (Q1)	0.0926
Tercer quartil (Q3)	1.0777
Mínim	0.0118
Màxim	3.2468
Quasi-desviació típica	0.6992
Coeficient de variació	1.1125

Variable daytrfrate

Histograma



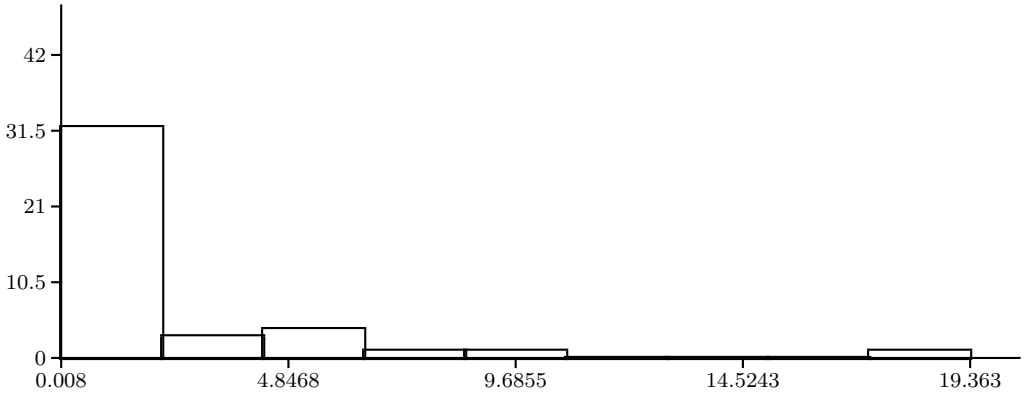
Boxplot



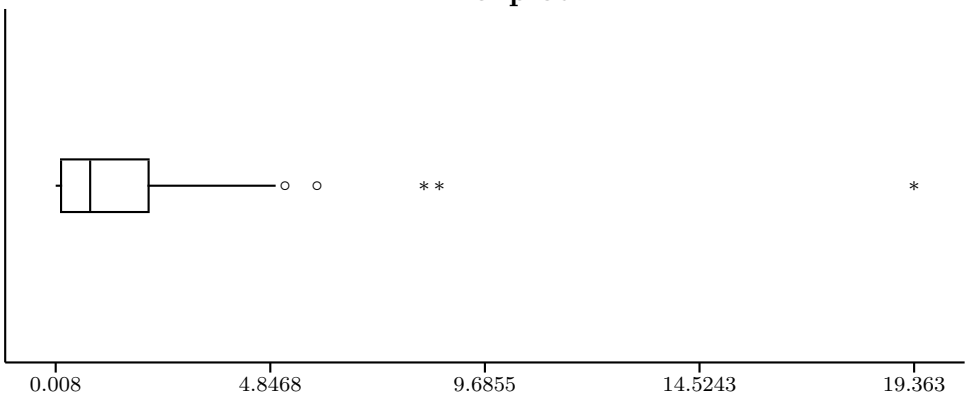
Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	0.0933
Mediana	0.0139
Primer quartil (Q1)	0
Tercer quartil (Q3)	0.0946
Mínim	0
Màxim	0.8972
Quasi-desviació típica	0.1703
Coeficient de variació	1.8026

Variable D4F1i11psychi

Histograma



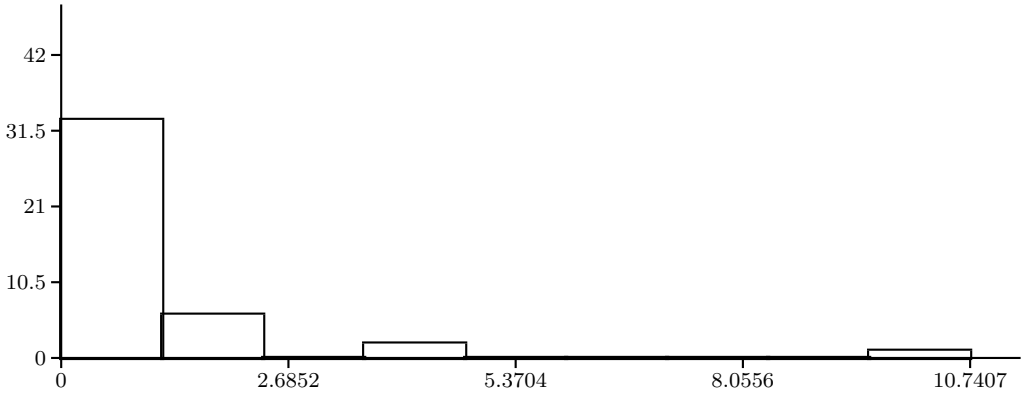
Boxplot



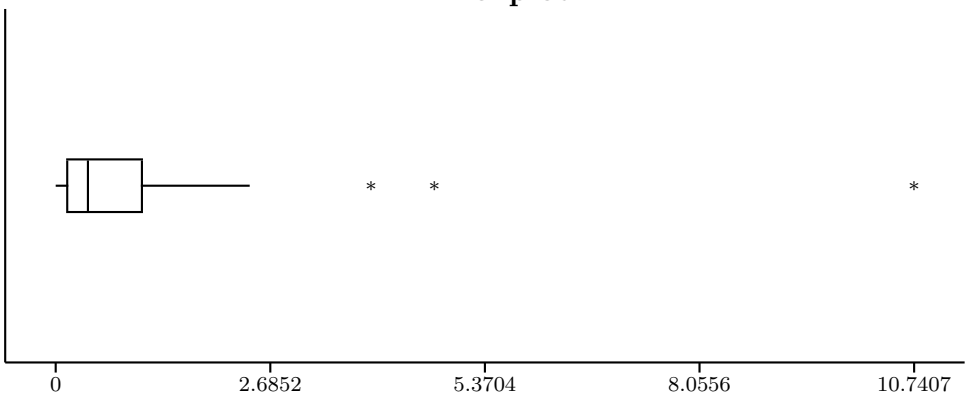
Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	2.0809
Mediana	0.7853
Primer quartil (Q1)	0.153
Tercer quartil (Q3)	2.0788
Mínim	0.008
Màxim	19.363
Quasi-desviació típica	3.5012
Coeficient de variació	1.6624

Variable D4F1i12doctors

Histograma



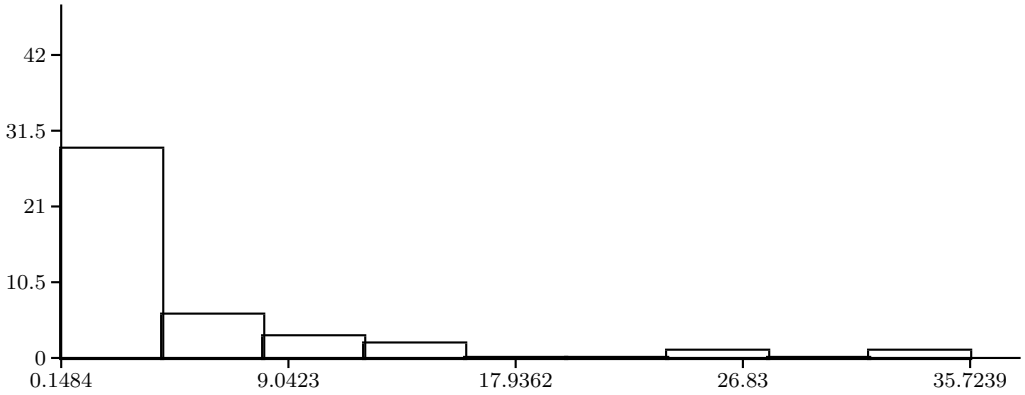
Boxplot



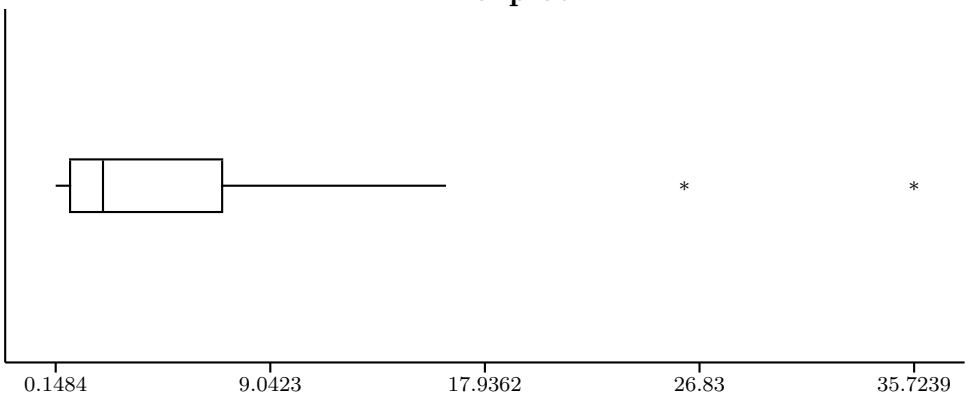
Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	0.9491
Mediana	0.4033
Primer quartil (Q1)	0.1576
Tercer quartil (Q3)	1.0655
Mínim	0
Màxim	10.7407
Quasi-desviació típica	1.828
Coeficient de variació	1.9029

Variable D4F1i13nurses

Histograma



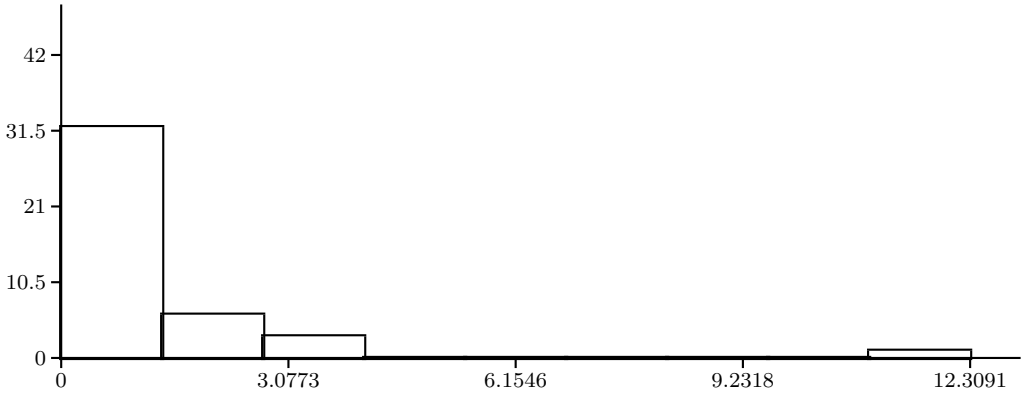
Boxplot



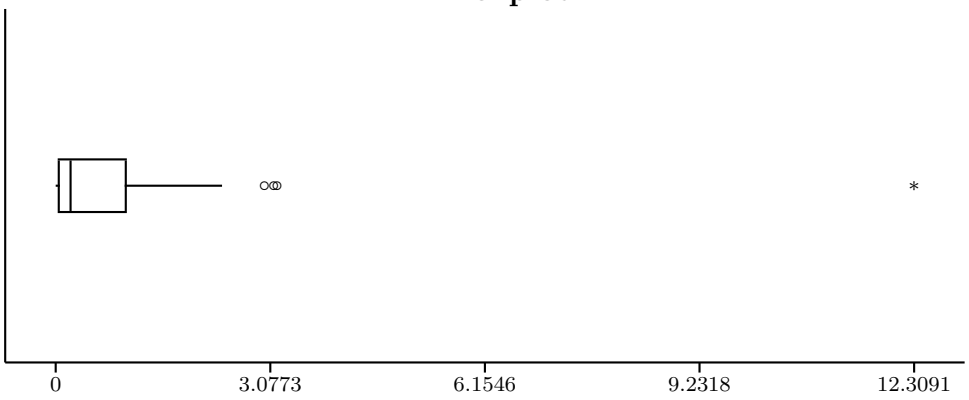
Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	4.8697
Mediana	2.1088
Primer quartil (Q1)	0.7884
Tercer quartil (Q3)	7.0027
Mínim	0.1484
Màxim	35.7239
Quasi-desviació típica	7.1164
Coeficient de variació	1.4439

Variable D4F1i14psycho

Histograma



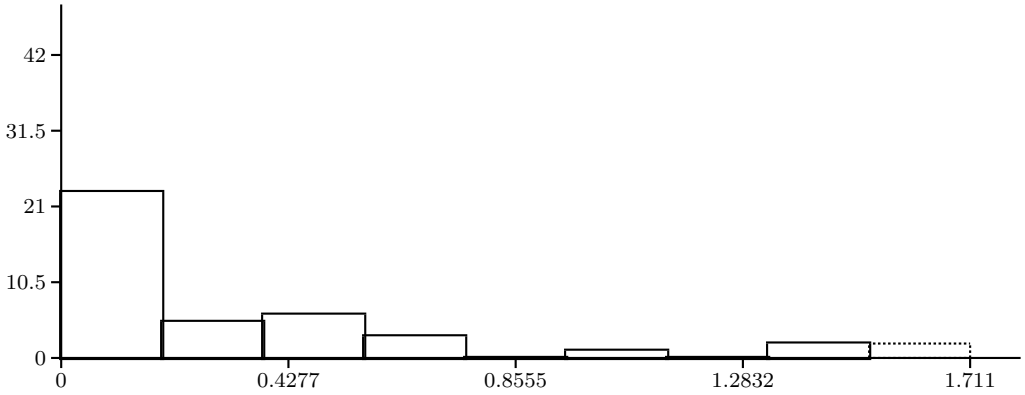
Boxplot



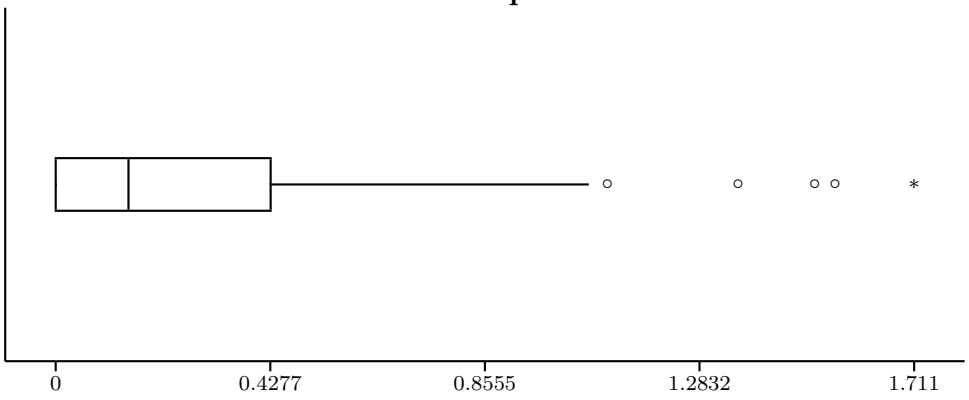
Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	0.9138
Mediana	0.2133
Primer quartil (Q1)	0.0593
Tercer quartil (Q3)	0.9897
Mínim	0
Màxim	12.3091
Quasi-desviació típica	2.0208
Coeficient de variació	2.1848

Variable D4F1i15socwork

Histograma



Boxplot

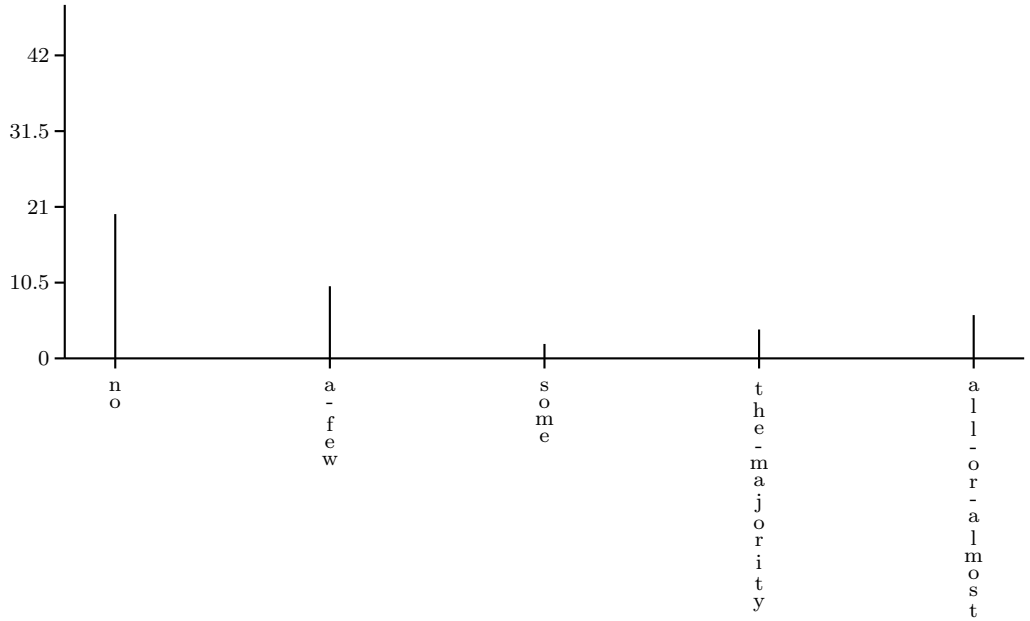


Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	0.3454
Mediana	0.1451
Primer quartil (Q1)	0.0021
Tercer quartil (Q3)	0.4263
Mínim	0
Màxim	1.711
Quasi-desviació típica	0.4663
Coeficient de variació	1.3337

Variable D3F1i3Manuals

Taula de freqüències				
Modalitats	**Freq. absol.**	**Freq. acum.**	**Freq. relat.**	**Freq. rel. acum.**
no	20	20	0.4762	0.4762
a-few	10	30	0.2381	0.7143
some	2	32	0.0476	0.7619
the-majority	4	36	0.0952	0.8571
all-or-almost	6	42	0.1429	1
dades mancants	0	N = 42	0	

Diagrama de barres



Variable D5F2i51relprimcare

Taula de freqüències				
Modalitats	**Freq. absol.**	**Freq. acum.**	**Freq. relat.**	**Freq. rel. acum.**
Y	34	34	0.8095	0.8095
N	8	42	0.1905	1
dades mancants	0	N = 42	0	

Diagrama de barres

Paràmetres de la classificació auxiliar

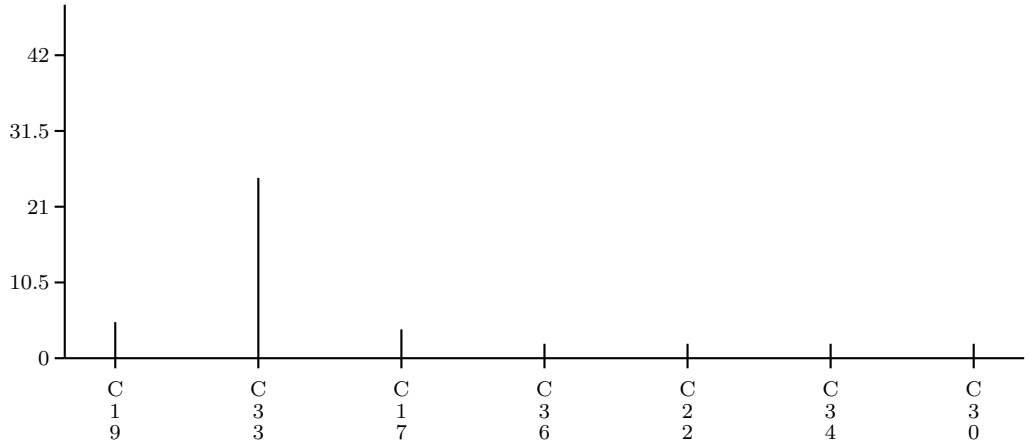
PARÀMETRE	VALOR
Mètode	Vens recíprocs encadenados
Criteri	Ward
Representant de classe	Objectes extesos de Gibert
Prefix de classe	C
Mètrica	Mixta de Gibert
Categoria	Alfa i beta automatics
Alfa	0.027377522
Beta	0.97262245
Nombre de classes resultants	7

Descriptiva univariant de la variable de clase

Variable vreWardGib0TresBloquesSinImputarbisK7

Taula de freqüències				
Modalitats	**Freq. absol.**	**Freq. acum.**	**Freq. relat.**	**Freq. rel. acum.**
C19	5	5	0.119	0.119
C33	25	30	0.5952	0.7143
C17	4	34	0.0952	0.8095
C36	2	36	0.0476	0.8571
C22	2	38	0.0476	0.9048
C34	2	40	0.0476	0.9524
C30	2	42	0.0476	1
dades mancants	0	N = 42	0	

Diagrama de barres



Fase 2: Imputació

Valors d'imputació per les variables numèriques

MITJANA

idClase	totprofmh	treatpre	lundpararectrail	comcarewor
C19	5.1	1115.5867	1.1425	0.0975
C33	7.1795	520.1282	0.817	0.1583
C17	5.6004	657.065	0.66	0.0275
C36	24.15	1221.6001	1.16	0.0245
C22	13.5507	283.61	0.67	0.1313
C34	13.97	1120.13	1.84	0.1206
C30	43.81	4595.1797	0.375	0.1424

idClase	usmhexperca	d1f5i2exmhos	d2f11i1closepsybeds	d2f6i71mhrec10y
C19	0.2683	0.7167	3.282	0.057
C33	0.687	0.724	4.1169	0.1732
C17	0.1213	0.8747	5.804	0.0486
C36	2.5634	0.525	0.85	0.49
C22	0.0393	0.6787	6.265	0.0684
C34	0.1281	0.408	1.59	0.02
C30	5.5864	0.804	1.33	0.1134

idClase	capratiosch
C19	0.2245
C33	0.2656
C17	0.2383
C36	0.2475
C22	0.44
C34	0.21
C30	0.2569

Valors d'imputació per les variables qualitatives

idClase	Region	Incgroup
C19	RegionC19.UNKNOWN	IncgroupC19.UNKNOWN
C33	RegionC33.UNKNOWN	IncgroupC33.UNKNOWN
C17	RegionC17.UNKNOWN	IncgroupC17.UNKNOWN
C36	RegionC36.UNKNOWN	IncgroupC36.UNKNOWN
C22	RegionC22.UNKNOWN	IncgroupC22.UNKNOWN
C34	RegionC34.UNKNOWN	IncgroupC34.UNKNOWN
C30	RegionC30.UNKNOWN	IncgroupC30.UNKNOWN

Nombre de missings reemplaçats per classe

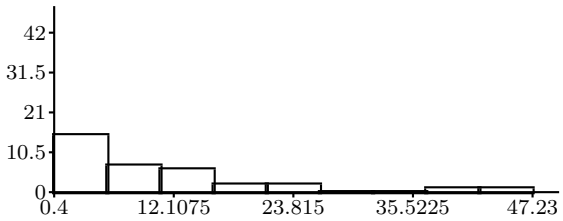
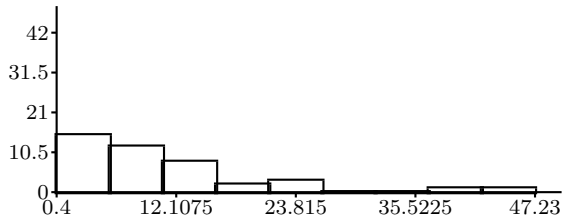
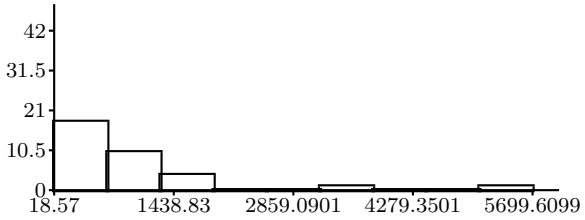
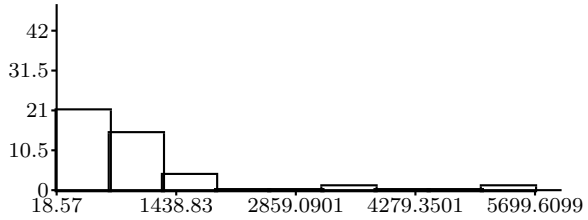
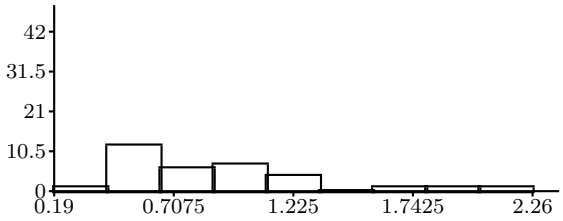
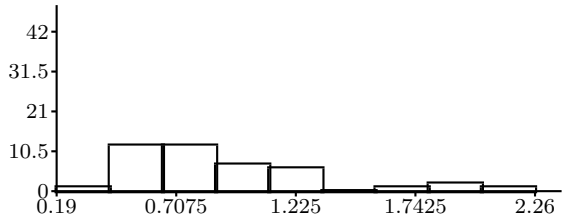
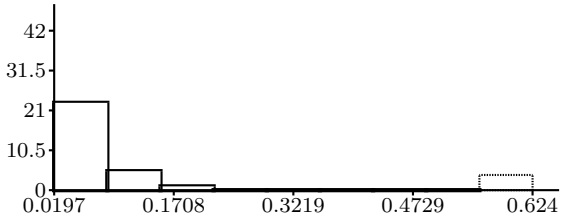
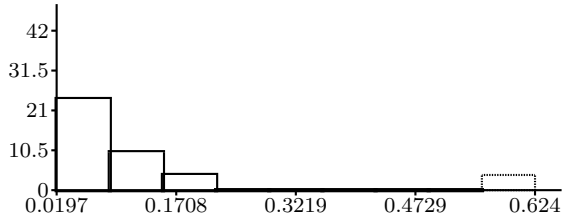
idClase	Region	Incgroup	totprofmh	treatpre
C19	2	2	0	2
C33	4	4	5	3
C17	0	1	0	2
C36	1	1	1	0
C22	0	0	1	0
C34	1	1	1	1
C30	0	0	0	0

idClase	lundpararectrail	comcarewor	usmhexperca	d1f5i2exmhos
C19	1	3	0	0
C33	4	3	7	7
C17	1	1	1	1
C36	1	0	0	0
C22	1	2	0	0
C34	1	0	0	0
C30	0	0	0	1

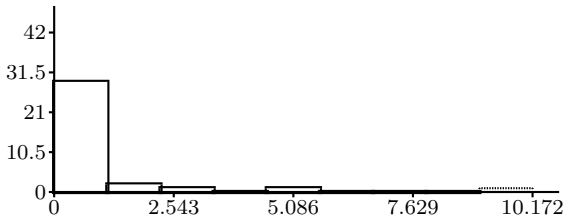
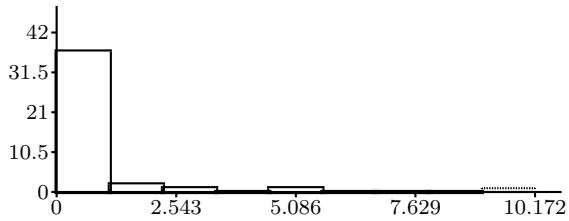
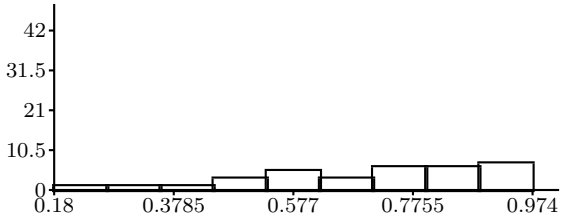
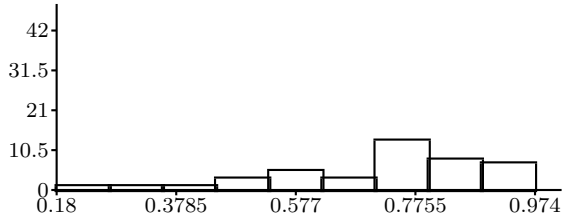
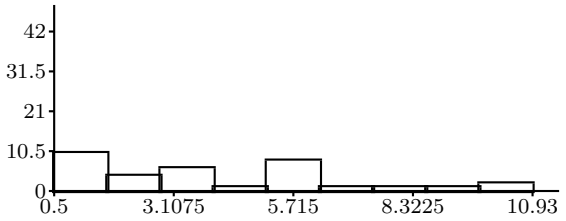
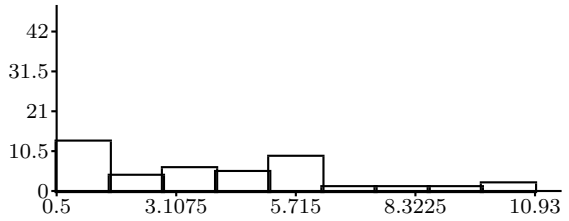
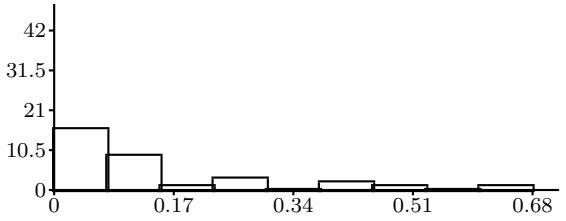
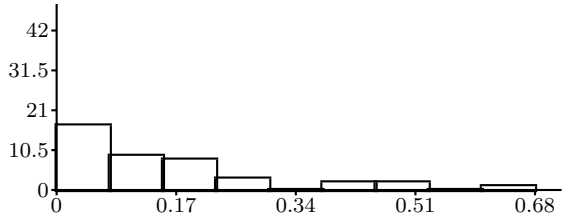
idClase	d2f11i1closepsybeds	d2f6i71mhrec10y	capratiosch
C19	0	0	0
C33	4	7	6
C17	1	0	0
C36	1	1	0
C22	0	0	1
C34	1	1	0
C30	1	0	2

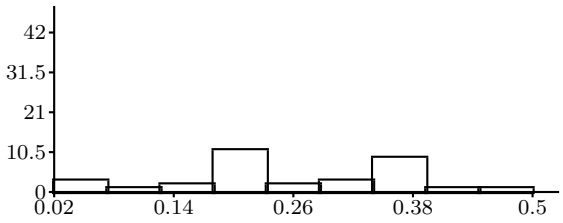
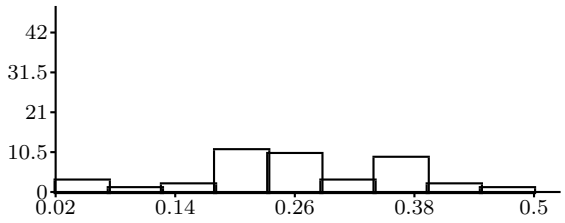
Descriptiva de les Propietats Modificades

Histograma

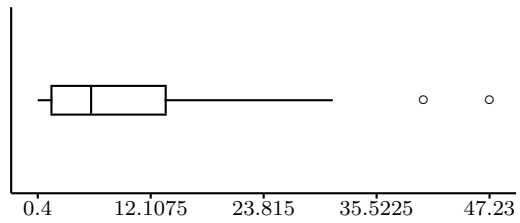
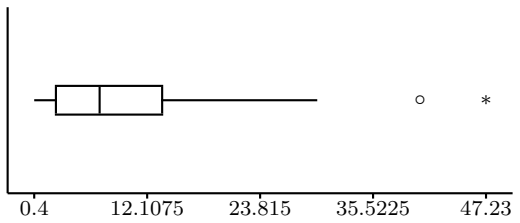
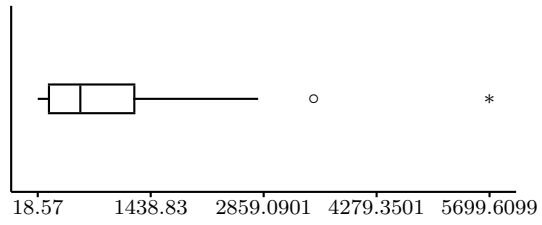
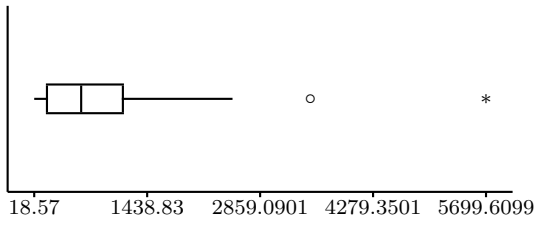
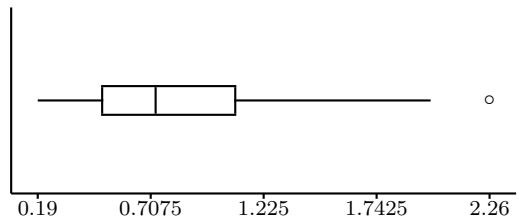
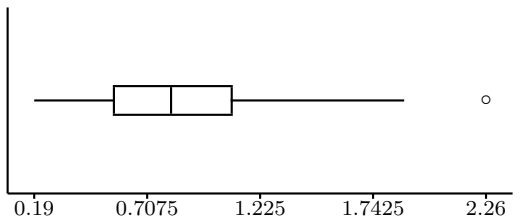
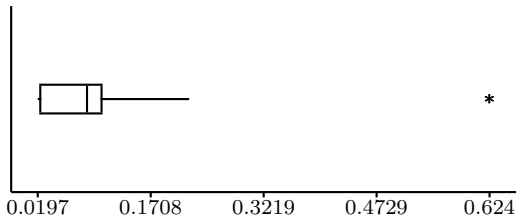
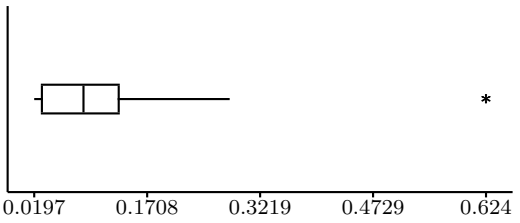
ANTES	DESPUES
totprofmh	
	
treatpre	
	
lundpararectrail	
	
comcarewor	
	

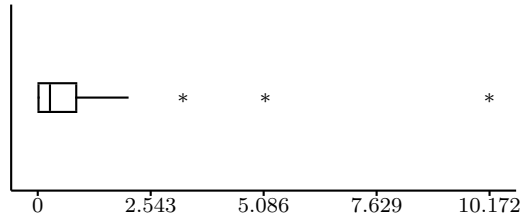
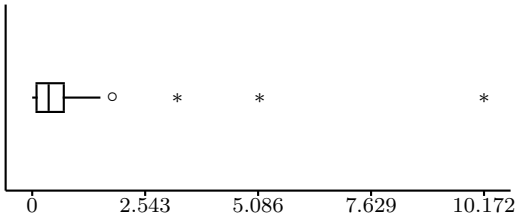
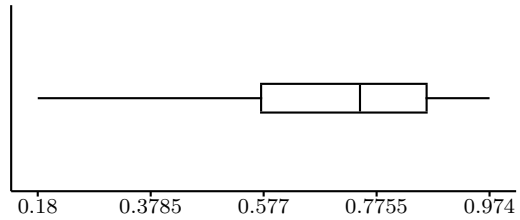
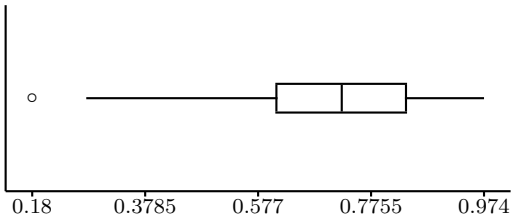
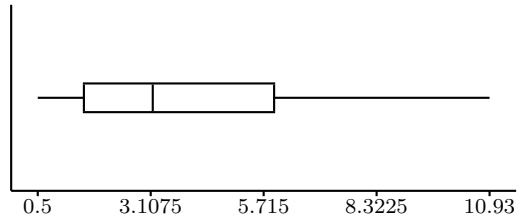
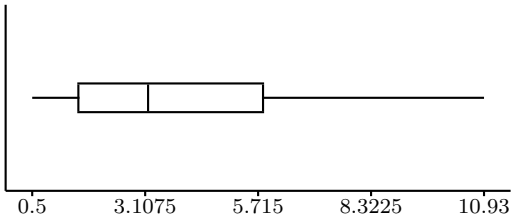
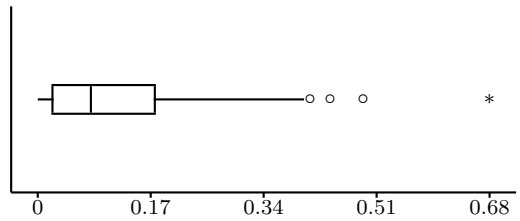
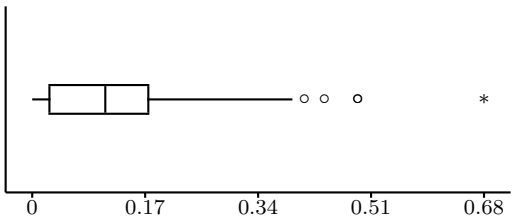
≀

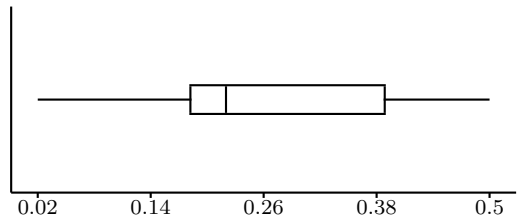
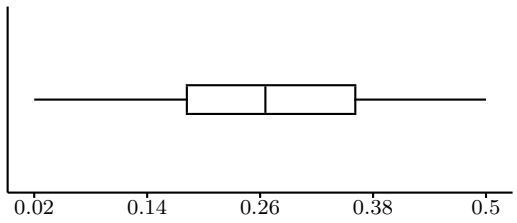
ANTES	DESPUES
usmhexperca	
	
d1f5i2exmhos	
	
d2f11i1closepsybeds	
	
d2f6i71mhrec10y	
	

ANTES	DESPUES
capratiosch	
	

Boxplot

ANTES	DESPUES
totprofmh	
	
treatpre	
	
lundpararectrail	
	
comcarewor	
	

ANTES	DESPUES
usmhexperca	
	
d1f5i2exmhos	
	
d2f11i1closepsybeds	
	
d2f6i71mhrec10y	
	

ANTES	DESPUES
capratiosch	
	

Estad'istics Sumaris

totprofmh

ANTES

Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	8
Nombre d'observacions útils	34
Mitjana	9.7289
Mediana	5.93
Primer quartil (Q1)	1.93
Tercer quartil (Q3)	13.5507
Mínim	0.4
Màxim	47.23
Quasi-desviació típica	10.8552
Coeficient de variació	1.0992

DESPUES

Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	9.9607
Mediana	7.1795
Primer quartil (Q1)	2.76
Tercer quartil (Q3)	13.5507
Mínim	0.4
Màxim	47.23
Quasi-desviació típica	10.0722
Coeficient de variació	0.9991

treatpre

Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	8
Nombre d'observacions útils	34
Mitjana	865.43
Mediana	553.9938
Primer quartil (Q1)	172.77
Tercer quartil (Q3)	1219.4614
Mínim	18.57
Màxim	5699.6099
Quasi-desviació típica	1118.3401
Coeficient de variació	1.2731

Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	848.8199
Mediana	609.0175
Primer quartil (Q1)	192.22
Tercer quartil (Q3)	1120.13
Mínim	18.57
Màxim	5699.6099
Quasi-desviació típica	1010.8624
Coeficient de variació	1.1766

lundpararectrail

Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	9
Nombre d'observacions útils	33
Mitjana	0.8523
Mediana	0.73
Primer quartil (Q1)	0.49
Tercer quartil (Q3)	1.09
Mínim	0.19
Màxim	2.26
Quasi-desviació típica	0.44
Coeficient de variació	0.5083

Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	0.8778
Mediana	0.817
Primer quartil (Q1)	0.56
Tercer quartil (Q3)	1.09
Mínim	0.19
Màxim	2.26
Quasi-desviació típica	0.4248
Coeficient de variació	0.4781

comcarewor

Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	9
Nombre d'observacions útils	33
Mitjana	0.1313
Mediana	0.0856
Primer quartil (Q1)	0.0245
Tercer quartil (Q3)	0.1036
Mínim	0.0197
Màxim	0.624
Quasi-desviació típica	0.1905
Coeficient de variació	1.4285

Estadístics sumaris	
Nombre d'objectes	42
Nombre de dades mancants	0
Nombre d'observacions útils	42
Mitjana	0.1284
Mediana	0.0856
Primer quartil (Q1)	0.0314
Tercer quartil (Q3)	0.1313
Mínim	0.0197
Màxim	0.624
Quasi-desviació típica	0.1695
Coeficient de variació	1.3044

ANTES	DESPUES
usmhexperca	
Estadístics sumaris	**Estadístics sumaris**
Nombre d'objectes 42	**Nombre d'objectes** 42
Nombre de dades mancants 8	Nombre de dades mancants 0
Nombre d'observacions útils 34	Nombre d'observacions útils 42
Mitjana 0.9031	**Mitjana** 0.8485
Mediana 0.2751	**Mediana** 0.3702
Primer quartil (Q1) 0.041	**Primer quartil (Q1)** 0.1213
Tercer quartil (Q3) 0.8426	**Tercer quartil (Q3)** 0.687
Mínim 0	**Mínim** 0
Màxim 10.172	**Màxim** 10.172
Quasi-desviació típica 1.9355	**Quasi-desviació típica** 1.7421
Coeficient de variació 2.1114	**Coeficient de variació** 2.0287
d1f5i2exmhos	
Estadístics sumaris	**Estadístics sumaris**
Nombre d'objectes 42	**Nombre d'objectes** 42
Nombre de dades mancants 9	Nombre de dades mancants 0
Nombre d'observacions útils 33	Nombre d'observacions útils 42
Mitjana 0.705	**Mitjana** 0.7146
Mediana 0.7463	**Mediana** 0.724
Primer quartil (Q1) 0.5744	**Primer quartil (Q1)** 0.611
Tercer quartil (Q3) 0.8615	**Tercer quartil (Q3)** 0.835
Mínim 0.18	**Mínim** 0.18
Màxim 0.974	**Màxim** 0.974
Quasi-desviació típica 0.1913	**Quasi-desviació típica** 0.1717
Coeficient de variació 0.2672	**Coeficient de variació** 0.2374
d2f11i1closepsybeds	
Estadístics sumaris	**Estadístics sumaris**
Nombre d'objectes 42	**Nombre d'objectes** 42
Nombre de dades mancants 8	Nombre de dades mancants 0
Nombre d'observacions útils 34	Nombre d'observacions útils 42
Mitjana 4.0169	**Mitjana** 3.8718
Mediana 3.155	**Mediana** 3.175
Primer quartil (Q1) 1.59	**Primer quartil (Q1)** 1.59
Tercer quartil (Q3) 5.932	**Tercer quartil (Q3)** 5.804
Mínim 0.5	**Mínim** 0.5
Màxim 10.93	**Màxim** 10.93
Quasi-desviació típica 2.8456	**Quasi-desviació típica** 2.6719
Coeficient de variació 0.6979	**Coeficient de variació** 0.6818
d2f6i71mhrec10y	
Estadístics sumaris	**Estadístics sumaris**
Nombre d'objectes 42	**Nombre d'objectes** 42
Nombre de dades mancants 9	Nombre de dades mancants 0
Nombre d'observacions útils 33	Nombre d'observacions útils 42
Mitjana 0.1354	**Mitjana** 0.1474
Mediana 0.08	**Mediana** 0.11
Primer quartil (Q1) 0.0237	**Primer quartil (Q1)** 0.0275
Tercer quartil (Q3) 0.1746	**Tercer quartil (Q3)** 0.1732
Mínim 0	**Mínim** 0
Màxim 0.68	**Màxim** 0.68
Quasi-desviació típica 0.1653	**Quasi-desviació típica** 0.1575
Coeficient de variació 1.2017	**Coeficient de variació** 1.0557

ANTES	DESPUES
capratiosch	
Estadístics sumaris	**Estadístics sumaris**
Nombre d'objectes 42 Nombre de dades mancants 9 Nombre d'observacions útils 33 **Mitjana** 0.2569 **Mediana** 0.22 **Primer quartil (Q1)** 0.1833 **Tercer quartil (Q3)** 0.3875 **Mínim** 0.02 **Màxim** 0.5 **Quasi-desviació típica** 0.1236 **Coeficient de variació** 0.4737	**Nombre d'objectes** 42 Nombre de dades mancants 0 Nombre d'observacions útils 42 **Mitjana** 0.2625 **Mediana** 0.2656 **Primer quartil (Q1)** 0.1833 **Tercer quartil (Q3)** 0.36 **Mínim** 0.02 **Màxim** 0.5 **Quasi-desviació típica** 0.1128 **Coeficient de variació** 0.4244

Diagrama de frequències

ANTES

Region

Taula de freqüències

Modalitats	Freq. absol.	Freq. acum.	Freq. relat.	Freq. rel. acum.
AFR	6	6	0.1765	0.1765
AMR	7	13	0.2059	0.3824
EMR	4	17	0.1176	0.5
EUR	6	23	0.1765	0.6765
SEAR	7	30	0.2059	0.8824
WPR	4	34	0.1176	1
dades mancants	8	N = 42	0.1905	

Incgroup

Taula de freqüències

Modalitats	Freq. absol.	Freq. acum.	Freq. relat.	Freq. rel. acum.
LOW	9	9	0.2727	0.2727
LOWER	21	30	0.6364	0.9091
UPPER	3	33	0.0909	1
dades mancants	9	N = 42	0.2143	

DESPUES

Region

Taula de freqüències

Modalitats	Freq. absol.	Freq. acum.	Freq. relat.	Freq. rel. acum.
AFR	6	6	0.1429	0.1429
AMR	7	13	0.1667	0.3095
EMR	4	17	0.0952	0.4048
EUR	6	23	0.1429	0.5476
SEAR	7	30	0.1667	0.7143
WPR	4	34	0.0952	0.8095
RegionC19.UNKNOWN	2	36	0.0476	0.8571
RegionC33.UNKNOWN	4	40	0.0952	0.9524
RegionC36.UNKNOWN	1	41	0.0238	0.9762
RegionC34.UNKNOWN	1	42	0.0238	1
dades mancants	0	N = 42	0	

Incgroup

Taula de freqüències

Modalitats	Freq. absol.	Freq. acum.	Freq. relat.	Freq. rel. acum.
LOW	9	9	0.2143	0.2143
LOWER	21	30	0.5	0.7143
UPPER	3	33	0.0714	0.7857
IncgroupC19.UNKNOWN	2	35	0.0476	0.8333
IncgroupC17.UNKNOWN	1	36	0.0238	0.8571
IncgroupC36.UNKNOWN	1	37	0.0238	0.881
IncgroupC33.UNKNOWN	4	41	0.0952	0.9762
IncgroupC34.UNKNOWN	1	42	0.0238	1
dades mancants	0	N = 42	0	

Diagrama de barres