IMPUTACIÓN BASADA EN ÁRBOLES DE CLASIFICACIÓN

AITOR PUERTA GOICOECHEA

JUNIO 2002

INDICE

Capítulo
1

Objetivos

El objetivo principal de este cuaderno técnico es presentar la aplicación de los árboles de clasificación y regresión como parte de un proceso de imputación de datos. Los árboles de clasificación y regresión generan subgrupos de población que contienen elementos homogéneos dentro de ellos y heterogéneos entre distintos subgrupos con respecto a la variable a discriminar, en nuestra situación dicha variable será la variable que se desea imputar. Los resultados obtenidos mediante dicha técnica proporcionan mejoras con respecto a otras estrategias de imputación que emplean unas determinadas variables para crear a priori los subgrupos de población; mediante este método se seleccionan los grupos que mejor discriminación proporcionan. Dentro de los subgrupos de población generados se pueden aplicar infinidad de métodos de imputación existentes, que variaran dependiendo del tipo de la variable a imputar, y distintas estrategias (imputación univariante, multivariante, múltiple,...).

Capítulo
2

Introducción y antecedentes

Introducción a la Imputación de Datos

En una investigación estadística, tanto parcial como exhaustiva, es frecuente que individuos encuestados no respondan a una o más preguntas del cuestionario. Cuando esto ocurre se dice que se tienen datos ausentes o missing y estamos bajo un problema de no-respuesta. La no-respuesta puede introducir sesgo en la estimación e incrementar la varianza muestral debido a la reducción del tamaño muestral.

La imputación de datos es la etapa final del proceso de depuración de datos, tras el proceso de edición, en el cual los valores missing o que han fallado alguna regla de edición del conjunto de datos son reemplazados por valores aceptables conocidos. La razón principal por la cual se realiza la imputación es obtener un conjunto de datos completo y consistente al cual se le pueda aplicar las técnicas de estadística clásicas.

Para la aplicación de la imputación de datos se recibe de la etapa anterior un fichero de datos con ciertos campos marcados por "falta de respuesta" ó "borrados" en la fase de edición por no cumplir alguna regla de edición propuesta. Tras la imputación de todas las variables del estudio se obtiene un fichero completo.

Encontrar un buen método de imputación es una tarea importante ya que errores cometidos en las imputaciones de datos individuales pueden aparecer aumentados al realizar estadísticas agregadas. Por todo esto parece razonable estudiar métodos de imputación que conserven características de la variable como pueden ser: conservación de la distribución real de la variable, relación con el resto de variables en estudio,... Los métodos de imputación para datos faltantes varían según el tipo del conjunto de datos, extensión, tipo de no-respuesta,…

De forma general existen dos grandes grupos de no-respuesta:

- Registros que tienen todos los campos missing.
- Registros que tienen ciertos campos con valor missing.

Registro	Sexo	Edad	Estado Civil	Región	Ingresos	Gastos
1	1	34	1	Araba/Álava	34.567	6.859
2	2	26	1	Bizkaia	78.686	7.635
3	2	45	2	Gipuzkoa	68.763	67.875
4	*	*	*	*	*	*
5	1	18	3	Araba/Álava	38.947	6.859
6	2	36	1	Bizkaia	6.886	7.635
7	2	25	2	Gipuzkoa	6.763	*

Elevación o Weigthing

Imputación

Para el primer caso de no-respuesta se aplica la técnica conocida como elevación o weighting, mientras que para el caso en el cual aparece no-respuesta en ciertos campos se aplica la técnica de imputación.

En décadas anteriores era habitual, a la hora de analizar los datos, ignorar aquellos registros que poseían algún valor missing en alguna variable. Se empleaban los métodos de eliminación por lista (listwise deletion) o por pares (pairwise deletion). Esto suponía que aquellos individuos que no habían contestado a alguna de las preguntas del análisis eran ignorados y esto podía provocar ciertos problemas en los resultados. Por una parte, las estimaciones pueden estar sesgadas, ya que la eliminación de los que no responden, supone asumir que la no-respuesta se distribuyen de forma aleatoria entre los distintos tipos de entrevistados. En el mejor de los casos, aquel en el que la no-respuesta se distribuye de forma aleatoria, estamos perdiendo una cantidad importante de información al eliminar también la información que estos individuos dieron a otras preguntas del cuestionario.

En las últimas décadas, se han desarrollado gran variedad de métodos de imputación para evitar los problemas derivados de la no-respuesta parcial y obtener un fichero de datos completos. Las razones para utilizar estos procedimientos en el tratamiento estadístico de los datos son básicamente:

1. Reducir el sesgo de las estimaciones. (sesgo debido a la no-respuesta).

2. Facilitar procesos posteriores de análisis de los datos.

3. Facilitar la consistencia de los resultados entre distintos tipos de análisis.

4. Enriquecer el proceso de estimación con fuentes auxiliares de información.

La imputación de datos ha sido cuestionada durante mucho tiempo debido en mayor medida a que se desconoce en realidad el impacto que provoca en la calidad de los resultados. Actualmente existen técnicas mediante las cuales se pueden obtener estimaciones sobre el error que incluye la imputación en la estimación (mediante la imputación múltiple por ejemplo).

Las argumentos principales contrarios a la imputación son:

1. Falsa sensación de confianza en el usuario. Debido a que realmente no se aumenta la información disponible sino que se genera a partir de la información que se posee.

2. Descuido de fases anteriores. Puede generar, en las fases previas, descuidos debido a la confianza de que la imputación solucionará los problemas que surjan en fases previas.

3. En el caso de estar bajo falta de respuesta no aleatoria: si se procede a realizar imputaciones de registros enteros para solucionarlo, estamos introduciendo sesgos.

4. En el caso de realizar imputación a registros con falta de respuesta total se están 'fabricando' datos.

5. Un procedimiento de imputación basado en supuestos poco realistas o con una metodología pobre, puede provocar un empeoramiento en la calidad de los datos.

Muchos de estos argumentos contrarios a la realización de la imputación se deben a la mala utilización de esta técnica y por tanto si se emplea correctamente no aparecen dichas desventajas.

- Con respecto al apartado 2. Una idea básica que se debe de tener es que la imputación no debe sustituir ni descuidar a ninguna fase previa. Hay que intentar obtener el valor real de las distintas variables por todos los medios disponible y en el caso de no obtenerlo se recurrirá a la imputación de datos.

- En el caso del argumento 3., decir que actualmente hay métodos desarrollados para tratar variables en las cuales el mecanismo de no-respuesta es no aleatorio.

- Con respecto al argumento 5 contrario a la imputación, hay que indicar la importancia que posee la realización de estudios previos a la imputación para conocer ante qué mecanismo de no-respuesta estamos y que método de imputación nos proporcionará mejores resultados dependiendo del tipo de variable, asociación con el resto de las variables del estudio,…

Recientemente la imputación está teniendo mucho auge y se están investigando distintos métodos que tienen en cuenta el comportamiento de los no respondientes y utiliza una amplia gama de técnicas para estimar la información faltante con precisión. Los defensores de la imputación argumentan entre otras ventajas las siguientes:

1. Ganancia en credibilidad ante el usuario. Al cual se le ofrece una información fiable y completa tras realizar la validación e imputación de datos.

2. Uniformidad y comparabilidad de los datos que utilizan usuarios. Parece más razonable que se aproveche la información disponible y realizar la imputación a cada registro, frente a la idea de dar la información que se dispone sin realizar imputaciones. En esta situación, los usuarios no expertos suelen hacen caso omiso a la información sobre no-respuesta presentada, y en realidad se está suponiendo que los no respondientes se comportan de igual forma que los respondientes.

3. Posibilidad de aprovechar otras informaciones. El organismo que produce los resultados es el más adecuado para realizar la depuración e imputación de los

datos ya que son lo que poseen información auxiliar que puede ayudar a mejorar las estimaciones.

Asunciones de no-respuesta

Cuando se va a realizar una imputación de datos se debe tener en cuenta con que tipo de datos faltantes se está trabajando y para esto se debe conocer si el mecanismo que genera la ausencia de datos faltantes es aleatorio o no.

Hay tres tipos de mecanismos:

- MCAR, Missing completely at Random (Completamente aleatorio). Se da este tipo cuando la probabilidad de que el valor de una variable X_j, sea observado para un individuo i no depende ni del valor de esa variable, x_{ij}, ni del valor de las demás variables consideradas, x_{ik}, $k \neq j$. Es decir, la ausencia de información no está originada por ninguna variable presente en la matriz de datos. Por ejemplo en el caso de tener en un estudio las variables ingreso y edad. Estaremos bajo un modelo MCAR cuando al analizar conjuntamente edad e ingresos, suponemos que la falta de respuesta en el campo ingresos es independiente del verdadero valor de los ingresos y la edad. Es decir:

  $$\Pr(R(Ingresos) \mid Edad, Ingresos) = \Pr(R(Ingresos))$$

  Donde R es la variable indicadora de respuesta de la variable Ingresos, valdrá 1 en el caso de haber respuesta y 0 en el caso de poseer valor missing.

- MAR, Missing at Random. (Aleatorio): Se da este tipo si la probabilidad de que el valor de una variable X_j sea observado para un individuo i no depende del valor de esa variable, x_{ij}, pero quizá sí del que toma alguna otra variable observada x_{ik}, $k \neq j$. Es decir, la ausencia de datos está asociada a variables presentes en la matriz de datos. En el ejemplo anterior si suponemos que los ingresos son independientes de los ingresos del miembro del hogar pero puede depender de la edad estaremos bajo un modelo MAR. Es decir:

  $$\Pr(R(Ingresos) \mid Edad, Ingresos) = \Pr(R(Ingresos) \mid Edad)$$

- NMAR, No missing at Random. Se produce este tipo de mecanismo en el caso en el cual la probabilidad de que un valor x_{ij} sea observado depende del propio valor x_{ij}, siendo este valor desconocido. En el ejemplo anterior, se obtiene que la función respuesta de la variable ingresos depende del propio valor de la variable ingresos, además de poder depender de otros factores.

  $$\Pr(R(Ingresos) \mid Edad, Ingresos) = \Pr(R(Ingresos) \mid Edad, Ingresos)$$

Generalmente, los supuestos anteriores de MAR y MCAR para el conjunto de la encuesta son difícilmente sostenibles, en cambio para el caso de realizar la imputación basada en estratos o grupos, dentro de éstos sí es mas acertado suponer los modelos

MAR y MCAR. Esta es una de las causas para que las imputaciones tiendan a hacerse dividiendo la población en subgrupos disjuntos.

Para el estudio de la imputación mediante árboles de clasificación, se necesita la suposición de que se está trabajando bajo, al menos, no-respuesta aleatoria (MAR), o completamente aleatoria (MCAR).

Si tenemos acceso a las variables que explican por qué es missing, tendremos acceso a los mecanismos de no-respuesta. Por ejemplo si personas con estudios superiores tienden a no responder a preguntas referentes a los ingresos del hogar, entonces la variable estudios realizados será una explicación de por qué el ingreso es missing. Si incluimos dicha variable en alguna ecuación como una variable 'mecanismo', aliviaremos el sesgo causado por la no-respuesta en los ingresos.

Tratamiento de la no-respuesta

Como se ha comentado anteriormente hay dos tipos de no-respuesta que van a ser tratadas de distinta forma. Por un lado, tenemos registros con todos los campos missing a los cuales se le va a aplicar alguna técnica de elevación o weigthing, mientras que por el contrario, en el caso de estar ante registros con solamente algunos campos missing les aplicaremos técnicas de imputación.

Las técnicas de elevación o weigthing principales son: Ponderación, duplicación, sustitución y Tasa RAD (Raking Ratio). Se puede obtener más información de dichas técnicas, entre otras, en la publicación **"Procedimientos de depuración de datos estadísticos".** Seminario Internacional de Estadística en Euskadi .1990. I. Villan Criado, M. S. Bravo Cabria.

Existen distintas formas de actuación ante la falta de respuesta parcial:

- No realizar imputación y usar únicamente la información disponible tras la depuración. En esta situación cuando hay valores misssing no serán imputados, por tanto, para el análisis posterior solo se consideran los valores con respuesta.

- Aplicar imputación a los registros con campos missing. Hay numerosas técnicas de imputación que Laaksonen (2000) clasifica de la siguiente manera:

    1. Imputación deductiva o lógica: En el caso de tener funciones conocidas entre ciertos valores observados y valores missing.

    2. Imputación modelo donante. Los valores imputados son generados a partir de un modelo. Es decir, los valores imputados pueden no haber sido observados.

    3. Imputación donante real. Los valores imputados son generados a partir de valores observados, de un registro donante respondiente real .

Destacar que mediante el método 2, también se pueden proporcionar valores reales, sin embargo no recibe el valor directamente de un registro donante real. Emplear registros donantes no siempre es una ventaja, por ejemplo en el caso en el que los valores observados no cubren todos los valores potenciales exhaustivamente. La imputación mediante donante real es imposible de aplicar correctamente en el caso en el que no

haya respondientes dentro de ciertas áreas, siendo también problemático en el caso en el que se tenga una baja respuesta dentro de algún grupo.

Debido a la complejidad de la imputación no se puede decir qué método es mejor que otro, en general, ya que depende en gran medida del tipo de variable que estemos tratando, comportamiento de la no-respuesta,... Los softwares de edición e imputación automática no son capaces de resolver el problema de la imputación por sí mismos, sin embargo pueden ayudar en la imputación práctica. Otro problema que suele surgir es que mientras un método puede ser muy ventajoso para algunas estimaciones estadísticas puede no serlo para otras.

Los métodos de imputación deben presuponer tres condicionantes básicos:

- Debe superar todos los controles de validación definidos, o lo que es lo mismo, no producir errores que previamente se habían eliminado.
- Deben cambiar el menor número posible de campos.
- Deben mantener en la medida de lo posible, siempre que no sean manifiestamente sesgadas o erróneas, las distribuciones de frecuencias de las variables, extraídas de las unidades que han superado los controles de validación.

Métodos que emplean toda la información de los Respondientes

Esta forma de trabajar consiste en considerar para los sucesivos análisis únicamente la información disponible tras la recogida y validación de la información. Estos métodos, que consisten en la eliminación de registros, aunque pueden ser aceptables en los casos en los que la proporción de casos incompletos es pequeña, conducen generalmente a estimaciones sesgadas, puesto que indirectamente se asume que el proceso de falta de respuesta se comporta mediante un proceso completamente aleatorio (MCAR). Existen dos métodos que se comentan a continuación:

Listwise Deletion (Eliminación por lista)

Es una solución muy conservadora que consiste en emplear solamente los registros que tengan respuesta en todas las variables del estudio. Las ventajas de este método son su simplicidad y la posibilidad de comparar los estadísticos univariantes, dado que se realizan con las mismas observaciones. Por el contrario los inconvenientes de este método son elevados, los análisis pierden potencia al reducirse el número de elementos y existe el riesgo de que los estimadores estén sesgados si el proceso de no-respuesta no es completamente aleatorio (MCAR). Además este método desperdicia una importante cantidad de información que se conoce.

Pairwise Deletion (eliminación por pares) o método de casos disponibles

En este caso se emplean todas las observaciones que tienen valores válidos para las variables de interés en cada momento. Por ejemplo, para el estudio de la correlación o covarianza entre las distintas variables el número de elementos variará según el número de registros que tengan valor no missing en dichas variables. Este método tiene la desventaja de no poder asegurar que la matriz de correlaciones sea definida positiva,

condición indispensable para invertir la matriz de correlaciones. Esta situación es debida a que se emplean distintas submuestras para el cálculo de las distintas correlaciones. De la misma forma que el método Listwise Deletion se obtienen buenos resultados únicamente en el caso de estar bajo un proceso de no-respuesta completamente aleatorio.

En el caso del estudio de los resultados de un censo estas soluciones propuestas anteriormente no son válidas debido a que la misión del censo es ser una investigación estadística exhaustiva.

Métodos de Imputación

La solución al problema del sesgo de las estimaciones consiste en imputar los datos faltantes, sustituyéndolos por valores estimados mediante algún método de imputación.

Durante las décadas anteriores se empleaban procedimientos de imputación basados en la experiencia, la intuición y la oportunidad. Se suponía uniforme la probabilidad de que las unidades respondiesen y se ignoraba frecuentemente el sesgo causado por la no-respuesta.

Actualmente se emplean infinidad de métodos de imputación y se generan nuevos métodos empleando distintas técnicas estadísticas. Gran parte de los métodos de imputación se pueden expresar mediante la siguiente formula:

$$y_{vi} = f(y_{nm}) + e$$

Donde y_{vi} representa el valor imputado, y_{nm} representa las observaciones con valores válidos (no missing), mientras que el e se refiere al residuo aleatorio.

En el caso de métodos determinísticos se asigna e =0 y es variable en el caso de métodos estocásticos. Los primeros proporcionan mejores resultados si se tiene en cuenta los estimadores puntuales como la media, mediana,… sin embargo provocan distorsiones en la distribución de la variable.

A continuación se comentan las características de los principales métodos de imputación junto con las ventajas y desventajas de cada uno de ellos, siguiendo la clasificación propuesta por Laaksonen (2000).

Imputación deductiva o lógica

Es un método de imputación determinístico que consiste en la asignación de valores a las celdas faltantes tras deducir con un cierto grado de certidumbre los valores más plausibles. Actualmente este método se aplica en situaciones en las que las respuestas que faltan se pueden deducir a partir de los valores del resto de variables de dicho registro.

Una imputación determinística toma generalmente el siguiente formato:

If (condición) then (acción)

Por ejemplo, en el caso de tener no-respuesta en la variable situación profesional y tener una edad menor de 16 años, se imputa a la categoría inactivo, debido a la normativa vigente que prohíbe trabajar a menores de dicha edad.

Imputación mediante registro donante

Son procedimientos que asignan a los campos a imputar de un registro el valor que en tales campos tiene otro registro de la encuesta. A los registros completos se les denomina registros donantes y los registros con campos a imputar se denominan registros receptores o candidatos. A los campos que se utilizan para establecer la relación entre registro donante y candidatos se les denominan campos de control. Dichos campos pueden ser tanto cualitativos como cuantitativos o de ambos tipos. En el caso de tratar variables exclusivamente cualitativas, el cruce de las distintas variables para dividir la población en subgrupos disjuntos se denominan estratos y la relación entre los registros candidatos y los donantes se establecen por igualdad de los códigos del estrato. Entre las ventajas de estos métodos se pueden destacar: 1. se imputa un valor posible y realizado y 2. es sencillo de implementar. Mientras que el principal problema se debe a que puede no haber respondientes con todo el rango de valores necesario en la variable a imputar. Existen gran número de métodos entre los que se destacan los siguientes:

Procedimiento Cold-Deck

Se define un registro donante por estrato como "registro tipo" en base a fuentes de información externas: datos históricos, distribuciones de frecuencias, etc… El método asigna a los campos a imputar de todos los registros candidatos los valores del registro donante correspondiente al mismo estrato. A partir de este método se originó el procedimiento hot-deck. La desventaja principal de este método es que la calidad de los resultados dependerá de la calidad de la información externa disponible.

Procedimientos Hot-deck

Este método es un procedimiento de duplicación . Cuando falta información en un registro se duplica un valor ya existente en la muestra para reemplazarlo. Todas las unidades muestrales se clasifican en grupos disjuntos de forma que sean lo más homogéneas posible dentro de los grupos. A cada valor que falte, se le asigna un valor del mismo grupo. Se está suponiendo que dentro de cada grupo la no-respuesta sigue la misma distribución que los que responden. Este supuesto incorpora una fuerte restricción al modelo, si esta hipótesis no es cierta se reducirá sólo en parte el sesgo debido a la no-respuesta. El método Hot-deck tienen ciertas características interesantes a destacar: 1. los procedimientos conducen a una post-estratificación sencilla, 2. no presentan problemas a la hora de encajar conjuntos de datos y 3. no se necesitan supuestos fuertes para estimar los valores individuales de las respuestas que falten. Otra ventaja de este método es la conservación de la distribución de la variable. Sin embargo estos métodos tienen algunas desventajas, ya que distorsiona la relación con el resto de las variables, carece de un mecanismo de probabilidad y requieren tomar decisiones subjetivas que afectan a la calidad de los datos, lo que imposibilita calcular su confianza. Otros de los inconvenientes son: 1. que las clases han de ser definidas en base a un número reducido de variables, con la finalidad de asegurar que habrá suficientes observaciones completas en todas las clases y 2. la posibilidad de usar varias veces a una misma unidad que ha respondido. Existen diversas variantes de dicho método:

Procedimiento Hot-deck o Fichero caliente secuencial

El registro donante es el registro sin valor missing, perteneciente al mismo estrato e inmediatamente anterior al registro candidato. Para aplicar esta imputación previamente se debe clasificar el fichero de tal forma que produzca una autocorrelación positiva entre los campos sujetos a imputación, de esta forma se asegura una mayor similitud entre registro donante y candidato. Las desventajas de este método son considerables: 1. hay que facilitar valores iniciales para el caso de tener valores missing en el primer registro, 2. ante una racha de registros a imputar, se emplea el mismo registro donante y 3. es difícil de estudiar la precisión de las estimaciones.

Procedimiento Hot-deck con donante aleatorio

Consiste en elegir aleatoriamente a uno o varios registros donantes para cada registro candidato. Hay diferentes modificaciones de este método. El caso más simple es elegir aleatoriamente un registro donante e imputar el registro candidato con dicha información. Se puede elegir una muestra de registros donantes mediante distintos tipos de muestreo e imputar al valor medio obtenido con todos ellos. Este último tipo tiene un elemento de variabilidad añadida debido al diseño de elección de la muestra que incorporan.

Procedimiento Hot-deck modificado

Consiste en clasificar y encajar los donantes potenciales y receptores utilizando un considerable número de variables. El encaje se hace sobre bases jerárquicas del siguiente modo: si no se encuentra un donante para encajar con un receptor en todas las variables de control, se eliminan algunas variables consideradas como menos importantes y de esta forma conseguir el encaje a un nivel superior.

Procedimiento DONOR

En este método se emplea una función distancia definida entre las variables para que se mida el grado de proximidad entre cada posible registro donante y el registro candidato. En este caso se imputa en bloque los valores del registro donante en los campos sin respuesta del candidato. Es necesaria una modificación previa de los datos para anular los efectos de escala en la función distancia.

Imputación mediante modelos donantes

Son procedimientos que asignan a los campos a imputar de un registro valores generados a partir de modelos ajustados a los valores observados de los registros respondientes. Existen diversos métodos de imputación, los principales se comentan a continuación:

Procedimientos de regresión

Se incluyen en dicho grupo aquellos procedimientos de imputación que asignan a los campos a imputar valores en función del modelo:

$$y_{vi} = a + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k + e$$

Donde y_{vi} es la variable dependiente a imputar y las variables $\{x_j | j = 1,..,n\}$ son las regresoras que pueden ser tanto cualitativas como cuantitativas, generalmente variables altamente correladas con la dependiente. Las variables cualitativas se incluyen en el modelo mediante variables ficticias o dummy. En este tipo de modelos se supone aleatoriedad MAR, donde e es el término aleatorio. A partir de este modelo se pueden generar distintos métodos de imputación dependiendo de: 1. Subconjunto de registros a los que se aplique el modelo. 2. Tipo de regresores 3. Los supuestos sobre la distribución y los parámetros del término aleatorio e .

Imputación de la media:

El modelo basado en la imputación de la media es el modelo más sencillo de los pertenecientes a los procedimientos de regresión. Sigue el siguiente modelo: $y_{vi} = a + e$

Este método de imputación es muy sencillo y consiste en la asignación del valor medio de la variable a todos los valores missing de la población o el estrato según se haga la imputación global o a partir de subgrupos construidos a partir de las categorías de otras variables que intervienen en el estudio. En la versión estocástica se incluye un residuo aleatorio. Este método tiene como desventajas que modifica la distribución de la variable reduciendo la varianza de la variable, como consecuencia en el caso de realizar análisis bivariantes reduce la covarianza entre las variables. Es decir, este método no conserva la relación entre las variables ni la distribución de frecuencias original. Además en este modelo se supone estar bajo un procedimiento MCAR.

Los modelos mas generales de regresión tienen ciertas mejoras con respecto a la imputación de la media que se comentan a continuación:

1. Asume el supuesto menos estricto de aleatoriedad, modelo MAR.

2. Infraestima el valor de la varianza y covarianza en menor medida que en el caso de imputación a la media..

3. Modifica en menor medida la distribución de las variables.

Modelos de regresión aleatoria

Este método se originó con la finalidad de resolver el problema de la distorsión de la distribución tras la imputación. Se propone añadir una perturbación aleatoria a las estimaciones del modelo de regresión:

$$\hat{x}_{im} = a + b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik} + \hat{e}_i$$

Donde las perturbaciones e_i se calculan mediante alguno de los siguientes métodos:

Se obtiene una muestra aleatoria de tamaño s de los r residuos observados $\hat{e}_i = \hat{x}_{im} - x_{im}$ y se suman a los valores x_{im} estimados.

Se obtienen aleatoriamente s valores de una distribución con media cero y varianza $\hat{s}_2$, donde $\hat{s}_2$ es la varianza residual correspondiente a los valores observados de x_m.

Método de imputación mediante regresión logística

Método de imputación similar al de regresión aplicable a variables binarias. Se realiza con los registros respondientes una regresión logística y en base a esta regresión se imputan los registros con no-respuesta. De la misma forma que en otros métodos está la versión determinística y la aleatoria que incluye una perturbación aleatoria. Recientemente se está aplicando el método de regresión logística basada en técnicas de registros donantes, con la idea de imputar los registros sin respuesta mediante los registros respondientes, implementado en el programa de imputación SOLAS (1999).

Método Regression-based nearest neighbour hot decking (RBNNHD)

Método propuesto por Laksonen (2000) que combina la imputación mediante métodos de regresión con los métodos de ficheros donantes. Consiste en construir una regresión lineal multivariante con los registros con respuesta. Clasificar los registros con no-respuesta añadiéndoles un termino error y posteriormente ordenarlos según el valor imputado. Tras esto, se aplica la regla del vecino más próximo a los registros con valor imputado y se modifica por el asignado mediante este método donante.

Método de estimación de Buck y método iterativo de Buck

El objetivo de este método, propuesto por Buck en 1960, es estimar la matriz de covarianzas de cualquier población K -variante cuando no hay datos completos. Consiste en estimar los datos faltantes en la muestra mediante técnicas de regresión y calcular la matriz de covarianzas considerando los datos imputados como reales. El algoritmo posee varias fases: 1. estimar el vector de medias con las r observaciones completas, 2. estimar el valor de la variable a imputar en la observación i-ésima regresando esta variable sobre las variables con dato en dicha observación y 3. añadir un termino de corrección a los términos de la matriz de varianzas y covarianzas, con el objeto de obtener estimaciones insesgadas de tales términos.

El método iterativo de Buck consiste en iterar los pasos 1 a 3 del método de Buck hasta obtener convergencia de las estimaciones.

Estimación máximo verosímil con datos no completos

En este tipo de métodos se supone que los datos completos siguen un determinado modelo multivariante. Es importante elegir un modelo que sea suficientemente flexible para reflejar las características de los datos estudiados. Estos métodos están desarrollados ampliamente en la tesis doctoral de María Jesús Barcena de UPV / EHU. Se comentan brevemente los principales métodos:

Método basado en factorizar la verosimilitud

Método basado en factorizar la función de verosimilitud aplicable a modelos y estructuras de datos no completos en el caso en el que el logaritmo de la función de verosimilitud puede descomponerse de la siguiente manera:

$$l(f \mid X_{obs}) = l_1(f_1 \mid X_{obs}) + l_2(f_2 \mid X_{obs}) + \ldots + l_j(f_j \mid X_{obs})$$

Donde f es la función de verosimilitud , X_{obs} el conjunto de datos observados y $l(x)$ el logaritmo de x. En esta situación se puede resolver aplicando resultados conocidos cuando se trabaja con datos completos.

Algoritmo EM (Expectation Maximization). Little & Rubin (1987)

Método basado en factorizar la función de verosimilitud que permite obtener estimaciones máximo verosímiles (MV) de los parámetros cuando hay datos no completos con unas estructuras determinadas. A diferencia del anterior, es válido para cualquier estructura de datos no completos. El algoritmo EM permite resolver de forma iterativa el calculo del estimador máximo verosímil mediante dos pasos en cada iteración:

1. Paso E (Valor esperado): Calcula el valor esperado de los datos completos basándose en la función de verosimilitud.

2. Paso M (Maximización): Se asigna a los datos missing el valor esperado obtenido en el paso anterior (E) y entonces se calcula la función de máxima verosimilitud como si no existiesen valores missing. Ambos pasos se realizan de forma iterativa hasta obtener convergencia.

Algoritmo de aumento de datos

Procedimiento iterativo que permite obtener valores simulados de los datos ausentes y de los parámetros desconocidos q , para algunas clases de modelos multivariantes. De la misma forma que el algoritmo EM, trata de solucionar un problema difícil con datos incompletos resolviendo repetidas veces problemas accesibles con datos completos. Consiste en un proceso iterativo que tiene dos fases:

Paso I de imputación de los datos ausentes. Simula valores para los datos ausentes mediante la distribución obtenida en la fase anterior y los valores observados.

Paso P o posteriori, que consiste en simular nuevos valores de los parámetros a partir de la distribución a posteriori condicionada a los datos completados en la fase anterior.

Muestreo de Gibbs

El muestreo de Gibbs es otro procedimiento para estimar los parámetros del modelo e imputar los datos ausentes. Se emplea el muestreo de Gibbs cuando se modela el problema de falta de datos mediante una metodología bayesiana.

Otros métodos

Recientemente se han propuesto, y se siguen estudiando, diversas técnicas estadísticas aplicadas en la fase de imputación de datos, como pueden ser:

Imputación de datos basadas en distintas técnicas de redes neuronales. Que está siendo estudiado actualmente en el proyecto EUREDIT.

Imputación basado en árboles de clasificación y regresión. Propuesta en el proyecto europeo AUTIMP como una técnica adecuada de imputación. Este método se comentará más detalladamente con posterioridad.

Imputación basada en la lógica difusa. Técnica propuesta en el proyecto europeo EUREDIT y se está desarrollando en la actualidad.

Estrategias de Imputación

Antes de realizar la imputación surge el problema de qué criterios se deben tener en cuenta para seleccionar el modelo de imputación a aplicar. Esta respuesta no es sencilla y hay que tener en cuenta los siguientes cinco aspectos que se detallan a continuación:

1. **La importancia de la variable a imputar.** Si la variable es de elevada importancia, es natural que se elija mas cuidadosamente la técnica de imputación a aplicar.

2. **Tipo de la variable a imputar.** Hay que considerar en este contexto el tipo de la variable, es decir, si es continua ó categórica tanto nominal como ordinal. Teniendo en cuenta para el primer grupo el intervalo para el cual está definido y para los segundos las distintas categorías de la variable.

3. **Estadísticos que se desean estimar.** En el caso que solamente nos interese conocer el valor medio y el total, se pueden aplicar los métodos más sencillos como son: imputación al valor medio o mediano y en base a las proporciones pueden ser razonables. Sin embargo al aplicar estos métodos habrá problemas en la estimación de la varianza, debido a que se infraestima su valor real.

   En el caso en el que se requiera la distribución de frecuencias de la variable, la varianza y asociaciones entre las distintas variables, se deben emplear métodos mas elaborados y analizar el fichero de datos. El problema en este caso se incrementa cuando hay una elevada tasa de no-respuesta.

4. **Tasas de no-respuesta y exactitud necesaria.** No se debe abusar de los métodos de imputación y menos cuando tenemos una elevada tasa de no-respuesta de la cual se desconoce el mecanismo. El problema no es tan grave en el caso en que se proporciona la correcta información sobre la precisión de las medidas estadísticas. En el artículo de Seppo Laaksonen (2000) se considera tasa de no-respuesta elevada cuando dicha tasa supera un tercio del total.

5. **Información auxiliar disponible.** La imputación puede mejorar al emplear información auxiliar disponible. En el caso de no disponer información auxiliar una técnica muy recomendada a aplicar es la imputación mediante el método hot deck aleatorio.

   La tarea de la imputación varía en gran medida dependiendo del tamaño del conjunto de datos. Cuando se dispone de un fichero de datos pequeño es problemático en el caso de tener valores missing en unidades cruciales, al aplicar hot deck aleatorio se pueden producir errores graves. Este caso se suele dar en muchas muestras económicas. En cambio cuando se posee un conjunto de datos de grandes dimensiones surgen menos problemas y se pueden aplicar distintos métodos de imputación.

La imputación se puede considerar como un proceso de varias etapas:

Paso 1: El proceso de imputación empieza cuando se dispone de un fichero de datos con valores faltantes, que ha debido pasar anteriormente la fase de edición.

Paso 2: Recopilar y validar para el proceso de imputación toda la información auxiliar que pueda ayudar en la imputación.

Paso 3: Estudiar los distintos modelos de imputación para las variables que van a ser imputadas. Seleccionar la técnica de imputación a aplicar pudiendo ser: imputación univariante, en el caso de imputar una sola variable en cada momento ó imputación multivariante en el caso de imputar simultáneamente un conjunto de variables de la investigación estadística. En esta fase es interesante observar los patrones de no-respuesta que aparecen en dicho estudio, y comprobar si hay gran número de registros que simultáneamente tienen no-respuesta en un conjunto de variables, en este caso puede ser interesante aplicar una imputación multivariante.

Paso 4: Seleccionar varios métodos de imputación posibles. En esta fase según el tipo de la variable a imputar, información auxiliar disponible, tipo de no-respuesta,… se seleccionan los métodos apropiados para dicha variable. Es conveniente seleccionar más de uno para poder contrastar los resultados que se obtienen mediante los distintos métodos.

Paso 5: Estimación puntual y varianza muestral para los distintos métodos de imputación empleados y su evaluación. El objetivo es obtener estimaciones con el mínimo sesgo y la mejor precisión.

Paso 6: Tras estos se pasa a calcular la varianza de la imputación, la cual se puede calcular mediante diferentes técnicas. Durante los últimos años, se han presentado varios métodos para el cálculo de la estimación de la varianza de los datos imputados.

- ✓ Imputación múltiple. Propuesto por Rubin (1987,1996)
- ✓ Imputación de pesos fraccionada (Fractionally weighted imputation) basada en la imputación múltiple pero para la estimación de la varianza toma los beneficios de aplicar el método Jack-Knife propuesto por Rao y Shao (1992).
- ✓ Analítica. Shao (1997) presenta algunos nuevos desarrollos referentes a algunos métodos de imputación para el calculo de la varianza de los valores imputados.

Paso 7: Resultados de la imputación.

1. Estimaciones puntuales y estimación final de la varianza.
2. Micro ficheros con valores reales e imputados.

Criterios de cumplimiento por la Imputación

El proceso de imputación debe ser capaz de reproducir eficientemente un fichero de datos completo al cual se le pueda aplicar un análisis estadístico para datos completos. Con la finalidad de obtener unos resultados adecuados tras la imputación se deben

calcular una serie de estadísticos que nos corroboren que estamos ante una imputación adecuada para el estudio en cuestión.

A continuación se proponen una serie de medidas que son deseables para obtener una buena imputación de datos, propuestas en el proyecto europeo de Edición e Imputación de datos (EUREDIT). Para el caso en el cual se desean producir estimaciones agregadas los criterios 1. y 2. son irrelevantes.

1. Precisión en la predicción: El proceso de imputación debe preservar el valor real lo máximo posible, es decir, debe resultar un valor imputado que sea lo más cercano posible al valor real.

2. Precisión en el ranking: El proceso de imputación debe maximizar la preservación del orden en los valores imputados. Es decir, debe resultar una ordenación que relacione el valor imputado con el valor real o sea muy similar. Esta medida se refiere a variables numéricas o categóricas ordinales.

3. Precisión en la distribución: El proceso de imputación debe preservar la distribución de los valores reales. Es decir, las distribuciones marginales y de orden superior de los datos imputados debe ser esencialmente la misma que la correspondiente de los valores reales.

4. Precisión en la estimación: El proceso de imputación debe reproducir los momentos de órdenes menores de la distribución de los valores reales. En particular, debe producir parámetros insesgados e inferencias eficientes de la distribución de los valores reales.

5. Imputación plausible: El proceso de imputación debe conducir a valores imputados que sean plausibles. En particular, deben ser valores aceptables al aplicarles el proceso de edición.

Las medidas propuestas anteriormente dependen del tipo de variable que estemos considerando, según el tipo de las variables a imputar hay criterios que no hay que tener en cuenta.

Existen distintas medidas propuestas para los distintos tipos de variables (nominales, ordinales, continuas,..) que se pueden consultar en el artículo de EUREDIT **"Interm Report on Evaluation Criteria for Statistical Editing and Imputation"**. Principalmente las características que se desean obtener de la imputación realizada son: la conservación de los momentos de la distribución original y la semejanza entre los valores reales y los imputados asignados a cada uno de ellos.

.

Capítulo
3

Imputación múltiple

En las últimas décadas, se ha desarrollado un nuevo método en el área del análisis de datos incompletos: la imputación múltiple. Tras la publicación de los trabajos de Little y Rubin (1986-87) han aparecido otros muchos artículos estudiando esta técnica de imputación.

La imputación múltiple es una técnica en la que los valores perdidos son sustituidos por m >1 valores simulados. Consiste en la imputación de los casos perdidos a través de la estimación de un modelo aleatorio apropiado realizada m veces y, como resultado, se obtienen m archivos completos con los valores imputados. Posteriormente, se lleva a cabo el análisis estadístico ordinario con las m matrices de datos completas y se combinan los resultados con una serie de fórmulas específicas proporcionadas por Little y Rubin (1987).

El objetivo de la imputación múltiple es hacer un uso eficiente de los datos que se han recogido, obtener estimadores no sesgados y reflejar adecuadamente la incertidumbre que la no-respuesta parcial introduce en la estimación de parámetros. En el caso de imputación simple tiende a sobreestimar la precisión ya que no se tiene en cuenta la variabilidad de las componentes entre las distintas imputaciones realizadas.

Para llevar a cabo la imputación múltiple de los valores perdidos, procederíamos del siguiente modo:

- ✓ En primer lugar se seleccionan las variables que se emplearán en el modelo de imputación. Es imprescindible que todas las variables que se van a utilizar conjuntamente en posteriores análisis se incluyan en dicho modelo, también se deben incluir todas aquellas variables que puedan ayudar a estimar los valores missing.

- ✓ En segundo lugar, se decide el número de imputaciones que se desea realizar. En general según se indica en la publicación de Rubin, entre 3 y 5 imputaciones son suficientes.

- ✓ Decidir el método de imputación a aplicar a los distintos ficheros de datos. Hay que tener en cuenta que esta fase es muy importante y se debe hacer un estudio del método a aplicar en función de las características de las variables a imputar, información auxiliar disponible, variables explicativas,… Para poder aplicar la imputación múltiple, el método seleccionado debe contener algún componente de imputación aleatoria. Con esta propiedad se asegura la posibilidad de obtener, para cada registro a imputar, modificaciones entre los valores imputados al completar los distintos ficheros de datos. Por ejemplo, no se va a poder aplicar la imputación múltiple en el caso de realizar métodos determinísticos, como pueden ser la imputación deductiva, al valor medio,…

- ✓ El siguiente paso será el de llevar a cabo los análisis estadísticos (univariantes, bivariantes o multivariantes) necesarios para la investigación. El análisis se realizará

con las matrices generadas tras la imputación y los resultados se combinarán con las distintas fórmulas proporcionadas por Little y Rubin.

Observando las distintas matrices generadas tras la imputación múltiple se puede hacer una idea respecto a la precisión del método de imputación, si no se observan grandes variaciones entre los valores imputados de las distintas matrices se tiene una gran precisión de las estimaciones. Sin embargo hay técnicas estadísticas mas adecuadas para el estudio de la precisión de los estimadores.

Combinación de los m ficheros de datos generados

Se obtiene un único coeficiente Q que combina los m estimadores $\hat{Q}_j$ obtenidos de los j $(j=1,..m)$ ficheros de datos completos generados y U_j es la varianza estimada del parámetro $\hat{Q}_j$:

$$\overline{Q} = \frac{1}{m}\sum_{j=1}^{m}\hat{Q}_j$$

Para calcular el error estándar, primero debemos calcular la varianza dentro de cada conjunto de datos:

$$\overline{U} = \frac{1}{m}\sum_{j=1}^{m}U_j$$

Y la varianza entre las imputaciones es:

$$B = \frac{1}{m-1}\left[\sum_{j=1}^{m}(\hat{Q}_j - \overline{Q})^2\right]$$

Siendo la varianza total:

$$T = \overline{U} + \left(1+\frac{1}{m}\right)B = \frac{1}{m}\sum_{j=1}^{m}U_j + \left[\sum_{j=1}^{m}(\hat{Q}_j - \overline{Q})^2 \Big/ (m-1)\right]\left(1+\frac{1}{m}\right)$$

A partir de esta información, se pueden construir los intervalos de confianza mediante la distribución t de Student con df grados de libertad. Donde:

$$df = (m-1)\left(1+\frac{m\overline{U}}{(m+1)B}\right)^2$$

Eficiencia de la estimación

La eficiencia de una estimación depende del número de ficheros de datos imputados realizados, parece razonable que a mayor número de imputaciones realizadas mejor se conocerá la variabilidad de la imputación. Rubin en 1987 indicaba que aproximadamente 3-10 imputaciones eran suficientes para obtener buenos resultados.

Propuso la siguiente medida aproximada para el cálculo de la eficiencia de la estimación en el caso de realizar m imputaciones:

$$\left(1+\frac{g}{m}\right)^{-1}$$

Donde g es la tasa de información faltante por la cantidad que ha sido estimada. Se calcula de la siguiente forma: $g = \frac{r+2/(df+3)}{r+1}$

Donde $r = \frac{\left(1+m^{-1}\right)B}{\bar{U}}$ es el relativo incremento en la varianza debido a la no-respuesta.

Selección del método de imputación

El aspecto importante de la imputación múltiple, de la misma forma que en el resto de imputaciones, reside en la definición del modelo de imputación y en el método de imputación. Es fundamental que el modelo empleado en las estimaciones de los valores faltantes contenga las variables que se van a emplear posteriormente en los análisis estadísticos ordinarios, con el fin de preservar las relaciones entre las variables. Cuanto mejor sea el modelo respecto a la predicción, menor será la variación de los valores imputados y más precisos serán los estimadores posteriores. El método de estimación de los valores imputados varía de unas aplicaciones a otras, de modo que las propiedades también varían.

En general, la imputación múltiple es una de las soluciones más adecuadas al problema de no-respuesta parcial debido a su fácil aplicación y a la posibilidad de aplicar dicho método en distintas situaciones y ante diferentes tipos de variables.

Software de Imputación múltiple

En la actualidad existen varias aplicaciones que permiten realizar la imputación múltiple con distintos tipos de matrices de datos.

Entre las aplicaciones exclusivamente dedicadas a la imputación están los programas AMELIA, MICE, NORM-CAT-MIX-PAN y SOLAS. Se encuentra información de dichos softwares en las páginas webs:

http://www.multiple-imputation.com
http://www.utexas.edu/cc/faqs/stat/general/gen25.html Destacan los módulos de imputación múltiple incluidos recientemente en SAS versiones 8.1 y 8.2. También existen macros de SAS que realizan imputación múltiple: sirnorm, em_covar, mvn y macros de Paul Allinson. Existe además una aplicación SAS de imputación denominada IVEvare.

Capítulo
4

Arboles de clasificación y regresión

Se define un árbol de decisión como una estructura en forma de árbol en la que las ramas representan conjuntos de decisiones. Estas decisiones generan sucesivas reglas para la clasificación de un conjunto de datos en subgrupos disjuntos y exhaustivos. Las ramificaciones se realizan de forma recursiva hasta que se cumplen ciertos criterios de parada.

El objetivo de estos métodos es obtener individuos más homogéneos con respecto a la variable que se desea discriminar dentro de cada subgrupo y heterogéneos entre los subgrupos. Para la construcción del árbol se requiere información de variables explicativas a partir de las cuales se va a realizar la discriminación de la población en subgrupos.

El programa AID (Automatic Interaction Detection) de Sonquist, Baker y Morgan (1.971), representa uno de los primeros métodos de ajuste de los datos basados en modelos de árboles de clasificación. AID esta basado en un algoritmo recursivo con sucesivas particiones de los datos originales en otros subgrupos menores y más homogéneos mediante secuencias binarias de particiones. Posteriormente surgió un sistema recursivo binario similar denominado CART (Classification And Regression Tree, Árboles de Clasificación y Regresión) desarrollado por Breiman en 1.984. Un algoritmo recursivo de clasificación no binario, denominado CHAID (Chi Square Automatic Interaction Detection, Detección de Interacción Automática de Chi Cuadrado) fue desarrollado por Kass en 1.980. Recientemente se han propuesto distintos métodos: FIRM propuesto por Hawkins, una simbiosis de construcción de árboles n-arios y análisis discriminante propuesto por Loh y Vanichsetakul y otra alternativa conocida como MARS (Multivarite Adaptive Regression Splines, propuesto por Friedman en 1991.

Dentro de los métodos basados en árboles se pueden distinguir dos tipos dependiendo de tipo de variable a discriminar:

- Árboles de clasificación. Este tipo de árboles se emplea para variables categóricas, tanto nominales como ordinales.

- Árboles de regresión. Este tipo de discriminación se aplica a variables continuas.

Teniendo en cuenta el tipo de variable con que estamos trabajando se calculan distintas medidas para el estudio de la homogeneidad. En todos los casos las variables explicativas son tratadas como variables categóricas. En particular en el caso de tener una variable explicativa continua, salvo que haya sido categorizada previamente, será tratada como una variable categórica con el número de clases igual al número de valores distintos de la variable en el fichero de datos. Por esta razón el conjunto de datos requiere ser tratado previamente.

Dependiendo de la estructura del árbol, del número de ramas que se permiten generar a partir de un nodo, se distinguen dos tipos:

- Árboles basados en la metodología CART: Técnica de árbol de decisión que permite generar únicamente dos ramas a partir de un nodo.

- Árboles basados en la metodología CHAID: genera distinto número de ramas a partir de un nodo.

Entre las ventajas de esta técnica no paramétrica de clasificación de la población están las siguientes:

- Las reglas de asignación son legibles y por tanto la interpretación de resultados es directa e intuitiva.

- Es una técnica no paramétrica que tiene en cuenta las interacciones que pueden existir entre los datos.

- Es robusta frente a datos atípicos o individuos mal etiquetados.

- Es válida sea cual sea la naturaleza de las variables explicativas: continuas, nominales u ordinales.

- Los criterios de construcción del árbol, el método y el algoritmo son los mismos tanto para árboles de clasificación como para los de regresión.

Por el contrario este método de clasificación de los datos tiene una serie de desventajas:

- Las reglas de asignación son fuertes y bastante sensibles a ligeras perturbaciones de los datos.

- Dificultad para elegir el árbol "óptimo".

- Ausencia de una función global de las variables (como pueden ser una ecuación de regresión, función lineal discriminante, …) y como consecuencia pérdida de la representación geométrica.

- Las variables explicativas continuas deben categorizarse previamente.

- Los árboles requieren grandes masas de datos para asegurarse que la cantidad de observaciones de los nodos hoja es significativa.

La estadística básica de Objetos Simbólicos va a consistir en un conjunto de gráficos y medidas resumen que van a depender de que variables formen esos objetos.

Formulación del problema

Se parte de un fichero de datos con una variable Y a discriminar, denominada variable respuesta, y un conjunto finito de variables $X_1, X_2, ..., X_p$ conocidas como variables explicativas.

Se trata de seleccionar entre las variables explicativas aquellas que discriminen mejor a la variable Y. Se obtendrá una partición de la población de forma que se obtengan dos o más subgrupos lo más heterogéneos posibles entre sí con respecto a la variable respuesta Y, y lo más homogéneos posibles dentro. Si se sigue haciendo

sucesivamente esta discriminación para los nuevos nodos generados y aplicando un criterio de parada, obtendremos el árbol de clasificación o regresión.

Un árbol de decisión consta de los siguientes elementos:

- Nodos intermedios: engendran dos o más (dependiendo del método empleado) segmentos descendientes inmediatos. También llamados segmentos intermedios.
- Nodos terminales: Es un nodo que no se puede dividir más. También denominado segmento terminal.
- Rama de un nodo t: Consta de todos los segmentos descendientes de t, excluyendo t.
- Árbol de decisión completo (A_{max}): Árbol en el cual cada nodo terminal no se puede ramificar.
- Sub-árbol: Se obtiene de la poda de una o más ramas del árbol A_{max} .

A pesar de los distintos tipos de árboles de clasificación y regresión existentes la forma de actuar en todos ellos es similar, salvo ligeras modificaciones. En primer lugar se debe tener un conjunto de datos con una variable respuesta (categórica o continua) y un conjunto de variables explicativas, todas ellas categóricas o continuas que han sido previamente categorizadas. Todos los registros del fichero de datos son examinados para encontrar la mejor regla de clasificación de la variable respuesta. Estas reglas se realizan basándose en los valores de las variables explicativas. La secuencia de particiones define el árbol. Cada partición se realiza para optimizar la clasificación del subconjunto de datos. El proceso de división es recursivo y finaliza la ramificación cuando se verifica un criterio de parada que ha debido ser definido previamente.

Espacio de búsqueda

Hay un gran número de posibles formas de efectuar divisiones en función de los valores que tomen las variables explicativas $X_1,...,X_p$, y generalmente no se pueden considerar todas ellas. Dependerá en gran medida del tipo de variable que estemos tratando:

- **Variable X_i cualitativa nominal*:*** En este caso la variable toma K valores distintos entre los que no cabe establecer un orden natural. Si tenemos que discriminar con ayuda de una variable nominal los elementos que van a los distintos nodos hijos en el nodo t, podemos formar todos los subgrupos de los K valores que puede tomar X_i y enviar a un nodo los casos que generan la mejor discriminación con respecto a la variable respuesta y los restantes al otro nodo.
- **Variable X_i cualitativa ordinal:** En este caso si la variable toma n valores, una vez ordenadas las categorías, se consideran como posibles cortes los $n-1$ valores intermedios. Entre estos posibles cortes se considerará el que proporcione grupos más homogéneos con respecto a la variable respuesta.

- **Variable X_i continua:** Se trabaja con estas variables de la misma forma que con las variables ordinales, con la particularidad de que en este caso el número de valores de corte a comprobar será mucho más elevado debido a que pueden aparecer, en el caso de no haber repeticiones, $N-1$ cortes en el caso de ser N el tamaño de la muestra. De este conjunto se seleccionarán los grupos que mejor discriminen los individuos con respecto a la variable respuesta.

Estimación de la tasa de error

La elección de un árbol respecto de otro dependerá en general de una estimación de su tasa de error $R(T)$. El problema es cómo realizar la estimación de dicha tasa. Existen diversas formas de calcular la estimación con una serie de ventajas e inconvenientes que se detallan a continuación:

- **Estimador por resustitución ó estimación intramuestral:** Es el estimador más simple, pero también el más sesgado inferiormente. Consiste en dejar caer por el árbol la misma muestra que ha servido para construirlo. Debido a que los árboles tienen gran flexibilidad para adaptarse a la muestra dada se puede obtener una estimación sesgada inferiormente de la tasa de error, y por tanto desconocer realmente el error real del árbol.

- **Estimador por muestra de validación o muestra de contraste:** Consiste en dejar caer por el árbol una muestra distinta a la empleada para la realización del árbol. Por ello éste no se ha podido adaptar a dichos registros como ocurría en el estimador anterior. Tenemos de esta forma un estimador de $R(T)$ insesgado pero tiene el inconveniente de forzar a reservar, para su uso en la validación, una parte de la muestra que se podía haber empleado en la construcción del árbol. Hay cierta pérdida de información. Se suele emplear dicho estimador en el caso de estar ante un tamaño de muestra elevado, como ocurre en el caso de los censos, debido a que no se pierde mucha información al eliminar del estudio una muestra para la estimación del error.

- **Estimación por validación cruzada:** La idea de la validación cruzada consiste en estimar $R(T)$ procediendo de forma reiterada de forma análoga al estimador por muestra de validación. Se deja cada vez fuera de la muestra para la construcción del árbol a una fracción k^{-1} del tamaño muestral total. Obtendremos de esta forma k estimaciones $R^{(1)}(T),...,R^{(k)}(T)$ y promediándolas de la siguiente forma:

$$R^{cv}(T)=\frac{R^{(1)}(T)+...+R^{(k)}(T)}{k}$$

  Observar que el árbol realizado para cada una de las submuestras podría ser distinto a los demás, en este caso la expresión anterior no sería válida.

- **Estimador bootstrap:** Recientemente se ha propuesto esta técnica de remuestreo para la estimación de la tasa de error. Ripley (1996).

Reglas de parada

Existen distintos criterios de parada que pueden provocar la finalización de los algoritmos que realizan árboles de clasificación o regresión. Las causas que pueden provocar la finalización son:

- Se ha alcanzado la máxima profundidad del árbol permitida.
- No se pueden realizan más particiones, porque se ha verificado alguna de las siguientes condiciones:
    1. No hay variables explicativas significativas para realizar la partición del nodo.
    2. El número de elementos en el nodo terminal es inferior al número mínimo de casos permitidos para poder realizar la partición.
    3. El nodo no se podrá dividir en el caso en el cual el número de casos en uno o más nodos hijos sea menor que el mínimo número de casos permitidos por nodo.

Existen dos técnicas básicas en la construcción de los árboles:

- **"Mirada hacia delante"**. Esta estrategia se basa en subdividir los nodos escogiendo en cada momento la división que produjese la máxima disminución de impureza $i(t)$ mientras un estimador adecuado de la tasa de error $R(T)$ disminuyera. Dado que en cada paso se examinan árboles con un número de nodos muy similar, basta estimar $R(T)$ por $\hat{R}(T)$. En el momento en el cual no se obtiene un descenso de la tasa de error aceptable se para la fase de la ramificación y se considera a este como el árbol óptimo.
- **"Mirada hacia atrás"**. Esta estrategia sugiere construir árboles frondosos, llegando al árbol máximo posible A_{max} sin tener en cuenta las tasas de error y tras su construcción se procede a realizar un trabajo de poda y quedarnos con aquel árbol que proporcione menor tasa de error. Esta teoría se basa en que no se conoce lo que hay tras una ramificación si no se realiza y en el caso de no encontrar resultados satisfactorios siempre estaremos a tiempo de eliminar dicho rama. Tras construir el árbol completo A_{max} se aplica un algoritmo de poda con el cual se obtiene una secuencia de sub-árboles mediante la supresión sucesiva de las ramas que proporcionan menos información en términos de discriminación entre las clases de la variable Y. Finalmente se elige el sub-árbol A^* que proporcione la menor tasa de error.

Una posibilidad de poda para los árboles de clasificación consiste en el uso de la tasa de mala clasificación. Esta es una medida del porcentaje de casos mal clasificados en un nodo terminal. Se crea la función indicadora $c(\bullet)$ que valdrá 1 en el caso en que la condición incluida entre los paréntesis sea cierta y 0 en caso contrario. Por tanto la tasa de mala clasificación $R(d)$ será calculada de la siguiente forma:

$$R(d) = \frac{1}{N}\sum_{i=1}^{N} c\left(d(x_i) \neq j_i\right)$$

Donde: N denota el número total de casos que han sido clasificados.

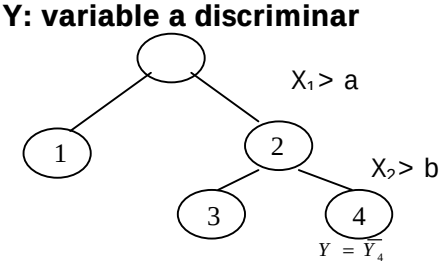
$d(x_i)$ denota la categoría asociada al nodo para el caso i.

j_i denota la verdadera categoría del caso i.

ÁRBOLES BASADOS EN MODELOS DE SEGMENTACIÓN RECURSIVOS BINARIOS

Se define un árbol binario como un grafo formado por nodos y arcos verificando lo siguiente:

1. Hay un solo nodo que no tiene padre y se denomina raíz.
2. Cada nodo distinto de la raíz tiene un único padre.
3. Cada nodo tiene exactamente dos o ningún hijo. En el caso de nodos sin hijos o nodos terminales hablamos también de hojas.



Podemos ver un árbol binario como una representación esquemática de un proceso de partición recursiva, en el cual en cada nodo no terminal tomamos la decisión de dividir la muestra de una cierta manera.

La idea básica de la segmentación recursiva binaria consiste en ir dividiendo el fichero de datos de interés en sucesivas particiones binarias. Tras un nodo padre se generan dos nodos hijos dividiendo los individuos pertenecientes al nodo padre en base a los valores de una variable explicativa. Se emplea para la partición la variable explicativa que mejor discrimina a la variable respuesta. El algoritmo actúa de forma recursiva y los nodos hijos generados pasan a ser potenciales nodos padres que a su vez pueden generar otro par de nodos hijos.

El primer nodo padre que va a ser subdividido es el fichero de datos original. El objetivo de las sucesivas ramificaciones y la construcción del árbol es obtener grupos de elementos homogéneos dentro de los nodos y heterogéneos entre los distintos nodos. El algoritmo procede a evaluar todos los posibles nodos padres candidatos a ramificar y selecciona aquel que más reduce la heterogeneidad dentro del nodo si se procede a

generar dos hijos a partir de él. Esto se realiza sucesivamente, cuanta mayor profundidad del árbol menor será el número de individuos pertenecientes a cada nodo hasta llegar a un punto en el cual no se pueda realizar más ramificaciones y se obtengan los llamados nodos terminales.

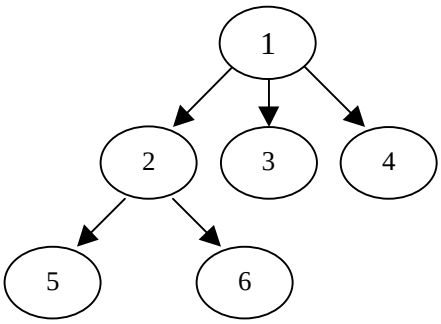
ÁRBOLES BASADOS EN MODELOS DE SEGMENTACIÓN DE K-HIJOS (CHAID)

Se define un árbol con k-hijos como un grafo formado por nodos y arcos verificando lo siguiente:

1. Hay un solo nodo que no tiene padre y se denomina raíz.

2. Cada nodo distinto de la raíz tiene un único padre.

3. Cada nodo tiene ninguno, dos o más hijos. En el caso de nodos sin hijos o nodos terminales hablamos también de hojas.



Este **método se puede aplicar tanto a variables** respuestas categóricas como continuas y permiten a cada nodo padre generar dos o más nodos hijos. Dentro del grupo de algoritmos que realizan árboles de clasificación y regresión no binaria destaca el algoritmo CHAID (Chi-square Automatic Interaction Detection), en este apartado nos vamos a referir principalmente a este algoritmo.

CHAID es una técnica no paramétrica de árboles de clasificación y regresión alternativa a la binaria. La técnica binaria es más restrictiva, ya que solo permite que se realicen dos ramificaciones por cada nodo. En cambio la metodología CHAID estudia distintos números de ramificaciones y selecciona el número de ramificaciones óptimo para obtener menor variabilidad dentro de los nodos con respecto a la variable respuesta. El número de ramificaciones posibles varía entre el rango comprendido entre dos ramificaciones y el número de categorías de la variable explicativa seleccionada para discriminar. CHAID originalmente se desarrolló como un método de detección de combinaciones o interacciones entre las variables. En la actualidad se emplea en marketing directo como una técnica de segmentación de mercados.

De la misma forma que en la metodología CART, las variables explicativas van a ser tratadas como variables categóricas por lo que requiere categorizar determinadas variables continuas empleando algún criterio adecuado.

Para cada nodo padre potencial, el algoritmo CHAID primero evalúa todas las combinaciones de los valores de las posibles variables explicativas empleadas para la discriminación (tratándolas como categóricas), agrupando las categorías que se comportan homogéneamente con respecto a la variable respuesta en un grupo y manteniendo separadas aquellas categorías que se comportan de forma heterogénea. Se selecciona la mejor mezcla de categorías de la variable explicativa formando un conjunto de nodos hijos que pasan a formar potenciales nodos padres. La forma de actuar depende del tipo de variable respuesta:

- En el caso en el que la variable respuesta sea categórica, se realizan las tablas de contingencia, con los registros pertenecientes al nodo padre, de cada variable explicativa con la variable respuesta. Se selecciona la variable explicativa que proporciona mejores resultados (aquella que proporciona menor p -valor al realizar el test chi-cuadrado).

- Para el caso de tratar una variable respuesta continua se calcula un equivalente p -valor de F de Student.

- Para el caso de tratar variables categóricas ordinales, se calcula un estadístico similar al de las variables continuas para calcular el p -valor mediante el test de cociente de probabilidades.

Una vez que la variable explicativa ha sido seleccionada junto con la tabla de contingencia, los nodos hijos son definidos por las clases de la variable explicativa que aparecen en la tabla de contingencia.

Dado que en este cuaderno técnico se presenta la técnica de imputación de datos basada en árboles de clasificación mediante el algoritmo CHAID se ofrece en el anexo I información referente al algoritmo de dicho proceso.

Software de árboles de clasificación y regresión

Existen diversos software relacionados con los árboles de clasificación y regresión que van a ser brevemente comentados. Básicamente se pueden clasificar en dos grupos:

- ✓ Software dedicado exclusivamente a la realización de árboles de clasificación y regresión: dentro de este grupo se incluye el software CART.

- ✓ Módulos o macros de paquetes estadísticos. Dentro de este grupo se incluye el módulo de SPAD SPAD•S, la macro TREEDISC de SAS, el módulo AnswerTree de SPSS y el módulo de árboles de clasificación y regresión de S-PLUS.

Capítulo
5

Imputación mediante árboles de clasificación

La idea básica de un modelo de imputación basado en árboles es muy simple. Dada una variable respuesta categórica o continua cuyo valor es missing y una serie de variables categóricas explicativas, el método emplea en primer lugar los registros con valor conocido en la variable respuesta. Con dichos registros se construye el árbol de clasificación que explica la distribución de la variable respuesta en función de las variables explicativas. Los nodos terminales de este árbol son tratados como clases de imputación. De esta forma, cada registro con valor missing en la variable respuesta llega a un determinado nodo terminal en función de los valores que posea en las variables explicativas empleadas en la construcción del árbol. A la hora de imputar se realizará basándose en los registros con valor conocido en la variable respuesta y que han sido asignados a dicho nodo, pudiéndose aplicar distintos métodos de imputación.

Los métodos a aplicar pueden ser muy diversos: imputación a la categoría más probable, imputación aleatoria en función de la distribución de frecuencias de dicho nodo, imputación hot-deck, imputación al vecino más próximo...

En el proyecto europeo AUTIMP se ha propuesto una forma alternativa de realizar la imputación basada en árboles de clasificación y regresión. Se divide la población respondiente en dos subconjuntos aleatoriamente. El primero de ellos se emplea para la construcción del árbol y el segundo se utiliza para aplicar la imputación. Este segundo subconjunto se clasifica mediante el árbol construido (con la información del primer grupo) y se aplicarán los distintos métodos de imputación según la distribución de frecuencias obtenida dentro de cada nodo terminal mediante este segundo grupo de registros. De esta forma se obtiene una mejor estimación del error de imputación al evitar el error que se puede cometer al imputar mediante información obtenida de los registros que han participado en la construcción del árbol, debido a que los árboles se pueden amoldar a la estructura de la muestra.

Evaluación de la Imputación

Tras realizar la imputación se debe efectuar un estudio de la calidad de la imputación obtenida. Hay dos formas distintas de estudiar dicha calidad:

Comparar las diferencias entre la distribución marginal de los valores reales y la distribución de los valores imputados.

Comparar las diferencias entre valores individuales, es decir, valor real vs. valor imputado por cada registro.

Para el estudio de las comparaciones entre el valor real y el imputado se realiza una tabla de contingencia.

Para comparar entre los distintos estudios esta evaluación se debe hacer para distintos tamaños de árbol, variables explicativas, métodos aplicados, software,….

Estadísticos de conservación de la distribución

Para el estudio de la igualdad entre la distribución marginal de los valores reales y de los valores imputados se propone el calculo de dos estadísticos:

Estadístico de Wald, se calcula de la siguiente forma:

$$W = (R-S)^t \left[diag(R+S) - T - T^t \right]^{-1} (R-S)$$

donde

R es el vector de los totales imputados (por categorías)

S es el vector de los totales reales (por categorías)

T es la matriz correspondiente a la tabla de contingencia formada al cruzar los valores reales e imputados de la variable respuesta.

Bajo la hipótesis de que tanto la distribución marginal de los valores reales e imputados son idénticos W debe comportarse como una distribución chi-cuadrado con p-1 grados de libertad donde p es el orden de la tabla de contingencia entre el valor real vs. imputado.

Estadístico chi-cuadrado de bondad de ajuste. El test chi-cuadrado de bondad de ajuste contrasta si un conjunto de datos se distribuye según una distribución fijada previamente, en nuestra situación la distribución marginal real.

Sea $(x_1,...,x_n)$ m.a.s. proveniente de X v.a. discreta. Siendo el contraste que se realiza el que se detalla a continuación.

$$\begin{cases} H_0 & F_X(x) = F_X^0(x) \quad \forall x \\ H_1 & F_X(x) \neq F_X^0(x) \quad para \quad algún \quad x \end{cases}$$

El estadístico se calcula de la siguiente forma:

$$Q = \sum_{i=1}^{K} \frac{(f_i - e_i)^2}{e_i}$$

siendo:

f_i : Frecuencias observadas. Número de individuos que de la muestra que pertenecen a la categoría i .

e_i : Frecuencias teóricas de la categoría i .

k : Número de categorías de la variable X .

Bajo H_0 Q se distribuye según una chi-cuadrado con $k-1$ grados de libertad.

$$Q \approx c^2_{k-1}$$

Observaciones:

i. En general, si $e_i \geq 5 \quad \forall i$ el contraste funciona correctamente.

ii. Si $\exists e_i \quad t.q. \quad 1{,}5 < e_i < 5$ para algunos valores de i, y estos e_i no superan el 20% del total, el test da en este caso resultados satisfactorios.

Estadísticos de conservación de los valores individuales

Para estudiar como de bien imputa el procedimiento para los valores missing, se proponen los siguientes estadísticos:

Estadístico diagonal: se calcula de la siguiente forma:

$$t_D = \frac{D}{\sqrt{\hat{V}(D)}} \text{ (Diagonal Statistic)}$$

donde D es la proporción de casos imputados incorrectamente y

$$\hat{V}(D) = \frac{1}{n} - \frac{1}{2n^2} 1^t \{diag(R+S) - T - diagT\} 1$$

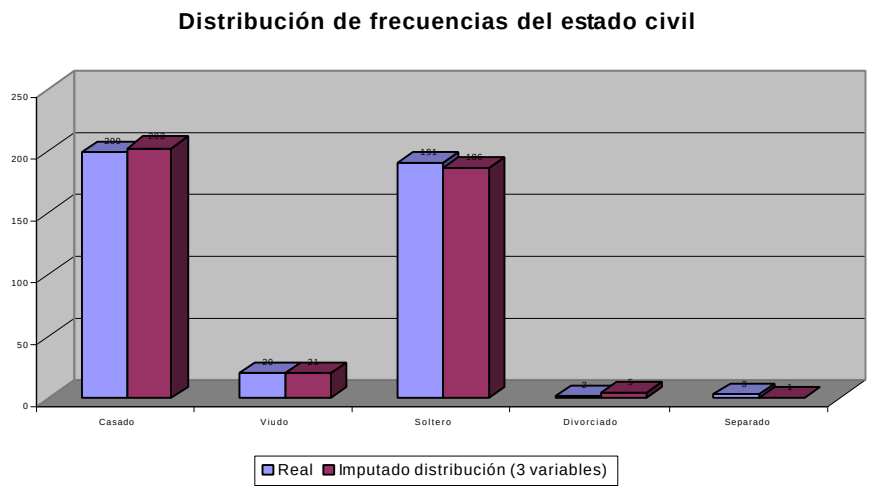
Bajo la hipótesis de que los valores individuales son preservados bajo la imputación, t_D debe aproximarse a la distribución $N(0,1)$.

Estadístico Kappa de Kohen: Este contraste comprueba si existe correspondencia entre las categorías, es decir, si la categoría 1 en la primera variable corresponde a la 1 en la segunda, la categoría 2 de la primera variable con la 2 de la segunda, ... En nuestra situación este contraste es muy útil ya que comprueba si hay relación entre las mismas categorías de la variable seleccionada para imputar contrastando los valores reales con los imputados. En el caso de correspondencia total entre las categorías tendremos una matriz diagonal, que es la situación más favorable posible, debido a que en este caso se realizaría imputación perfecta.

Representación gráfica de la calidad de la imputación

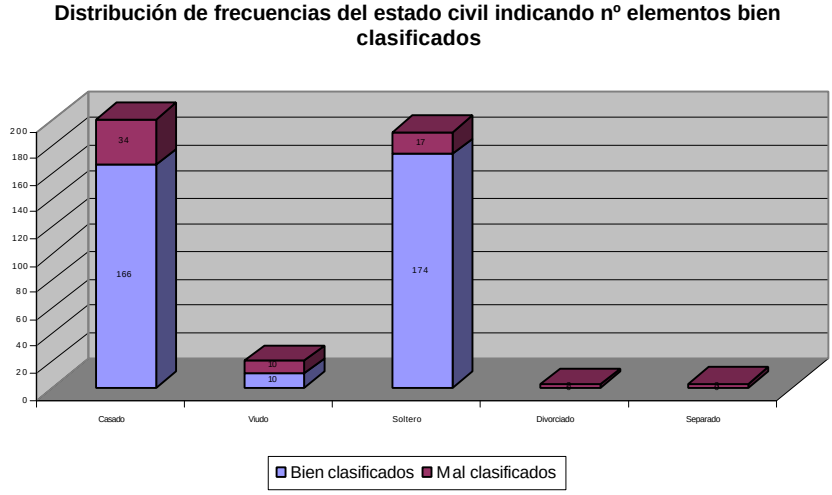
Para mostrar como los diferentes métodos de imputación llevan a cabo esta tarea, a parte de los estadísticos anteriores, hay distintas formas de representación gráfica para cada una de las perspectivas comentadas anteriormente.

Para preservar la distribución marginal se propone el siguiente gráfico, el cual compara la distribución marginal de los valores reales con la correspondiente distribución marginal de los valores imputados.



En este gráfico puede verse un estudio en el que han participado 417 registros de los que se conocía su valor en la variable estado civil y se han imputado mediante un árbol donde han participado las variables sexo, edad y número de hijos. Una vez asociado a cada registro un nodo terminal se realiza una imputación basada en la distribución de frecuencias de dicho nodo.

Para el estudio de la conservación del valor individual se propone el siguiente gráfico que muestra cómo los valores individuales se mantienen tras la imputación en su



categoría correspondiente. Compara, para cada registro, su valor real con el valor asociado a la variable tras la imputación.

La parte inferior de color azul indican los registros cuyo valor de la variable coincide con el valor asociado tras la imputación, mientras que el valor superior, de color rojo, indican

los registros que inicialmente pertenecían a dicha categoría pero que tras realizar la imputación han sido asignadas a una categoría distinta.

WAID 4.0

Entre los objetivos del proyecto Europeo AUTIMP (AUTomatic IMPutation), estaba el desarrollo de un software prototipo de imputación que se denominó WAID. Dicho software está basado en la técnica de árboles Automatic Interaction Detection (AID), presentada por Sonquist, Baker y Morgan en 1971. Debido a que el software proporciona menores pesos a los datos atípicos a la hora de construir el árbol, se ha denominado a la técnica weighted automatic interaction detection (WAID). Dicho programa construye árboles binarios para su posterior aplicación en la imputación de registros de idéntico patrón de no-respuesta. Se puede obtener mayor información de dicho software en la página web del proyecto europeo AUTIMP:

http://www.cbs.nl/en/services/autimp/autimp.htm

Capítulo
6

Aplicación a la estadística de población y vivienda

INTRODUCCIÓN

En este apartado se va explicar el proceso seguido para la aplicación del nuevo método de imputación basado en árboles de clasificación, presentado anteriormente, a la Estadística de Población y Vivienda de 1996. En los siguientes apartados se va a ir describiendo el proceso realizado y los resultados obtenidos.

Los árboles construidos, basados en el algoritmo CHAID, se presentan como herramientas para la imputación de datos. Este desarrollo puede ser un buen método de imputación ya que CHAID crea grupos óptimos homogéneos con respecto a la variable respuesta, que pueden ser considerados como clases de imputación. Este método no provocará inconsistencias en los valores imputados tras aplicar la imputación si se incluyen como variables explicativas aquellas que puedan ocasionarlas.

Como ejemplo se va a desarrollar un análisis para realizar la imputación de la variable relación con la actividad mediante la técnica de árboles de clasificación.

DESCRIPCIÓN DE LOS FICHEROS

Los datos empleados para el estudio son los proporcionados por las variables de la Estadística de Población y Vivienda de Euskadi de 1996, investigación estadística de carácter censal. La inscripción en el padrón y en la Estadística de Población y Vivienda afecta a todos los individuos que residían habitualmente en una vivienda familiar ó colectiva de alguno de los 250 municipios que forman la C.A. de Euskadi.

Se han empleado para el análisis dos ficheros de datos: Uno de ellos contiene los registros validados obtenidos tras el proceso de edición de datos y preparado para aplicar los distintos métodos de imputación planteados y el segundo de ellos, contiene los datos obtenidos tras aplicar los procesos de imputación realizados en la Estadística de Población y Vivienda de 1996. En ambos ficheros toda la información se posee a nivel de individuo y tiene un total de 2.257.924 registros.

Las variables que aparecen en dicho fichero son todas las variables proporcionadas por la Estadística de Población y Vivienda. A continuación se detallan las variables empleadas en el estudio realizado junto con su descripción:

Nombre Simbólico	Descripción
SITR	Situación de residencia
SEXO	Sexo
AGNN	Año de nacimiento
REPP	Relación con la primera persona
CONY	Figura el cónyuge
FIPA	Figura el padre
RELE	Relación con el establecimiento colectivo
ECIV	Estado civil legal
THIJ	Tiene hijos (1-3 y M)

Nombre Simbólico	Descripción
RELA1	Relación con la actividad
EKEN	Entiende euskera
EKHA	Habla euskera
EKLE	Lee euskera
EKES	Escribe euskera
LMAT	Lengua materna
LHAB	Lengua hablada en casa
SEDE	Sedentarios
CEST	Precódigo de estudios

Para la aplicación del nuevo método de imputación propuesto se ha realizado un filtrado del fichero de datos y se han seleccionado únicamente los registros pertenecientes a la Llanada Alavesa, con lo cual se han obtenido 239.494 registros. Con esta reducción de la población se pretende realizar pruebas sobre un subgrupo de la población de tal forma que los cálculos sean menos costosos computacionalmente. Los motivos que nos han llevado a seleccionar dicha comarca son, entre otras razones, que no hay un número elevado de individuos y que posee tasas de no-respuesta superiores a la media de la C. A. de Euskadi.

ANÁLISIS DE LA APLICACIÓN AL CENSO

En este estudio se pretende establecer nuevos métodos con posible aplicación a las variables del censo que en la mayor parte de los casos se refieren a variables categóricas, tanto ordinales como nominales. Por esto los métodos a los que va a ir dirigido este análisis se refieren a variables categóricas.

Se ha desarrollado un método basado en árboles de clasificación. Para la realización del árbol de clasificación se ha empleado la macro de SAS %TREEDISC.

Previamente se ha debido analizar las posibles variables explicativas a incluir en el modelo y selección de aquéllas que proporcionen mayor asociación con la variable a imputar. Para la construcción del árbol se emplean únicamente los registros con respuesta en la variable a imputar.

Una vez construido el árbol se selecciona el conjunto de datos con valor missing en la variable respuesta y se inicia la imputación. Se selecciona cada registro y se deja caer por el árbol y según los valores que posee en las variables explicativas empleadas se va clasificando por distintas ramas del árbol hasta llegar a un nodo terminal. La imputación del registro se va a basar en los registros con valor conocido que han sido clasificados en dicho nodo terminal. En esta situación se han considerado dos alternativas de imputación:

Imputación aleatoria según la distribución de frecuencias obtenido en el nodo terminal

Esta imputación consiste en que una vez clasificado un registro con valor missing a un nodo terminal, se asigna a una categoría aleatoriamente según la distribución de frecuencias dentro de dicho nodo. Con los registros con respuesta, que han participado

en la construcción del árbol, clasificados en dicho nodo terminal, se calculan los porcentajes para cada categoría de la variable respuesta. Se selecciona aleatoriamente en base a estas proporciones una categoría que se asignará al registro missing.

Imputación a la categoría más probable del nodo terminal

Esta imputación consiste en asignar al registro clasificado al nodo terminal la categoría de mayor probabilidad dentro de dicho nodo.

A parte de la estrategia de imputación univariante básica, existen otras estrategias de imputación, como son la imputación múltiple y la imputación multivariante, a las cuales también se les puede aplicar el método de imputación basado en árboles de clasificación.

Imputación múltiple

Consiste, como ya se ha comentado anteriormente, en realizar $m>1$ conjuntos de datos imputados aplicando la misma técnica y estudiar las variaciones entre los valores imputados obtenidos y cuantificar la incertidumbre que la imputación introduce en la estimación de parámetros.

En el caso de aplicar la imputación mediante árboles de clasificación una vez construido el árbol y seleccionado el nodo terminal asociado a cada registro con valor missing se puede aplicar la imputación múltiple únicamente en el caso de existir una fase aleatoria dentro del proceso de imputación.

En nuestra situación solamente se puede aplicar la imputación múltiple cuando se realiza dentro del nodo terminal una imputación aleatoria según la distribución de frecuencias.

Imputación multivariante

Tras haber analizado los patrones de no-respuesta de la Estadística de Población y Vivienda se puede tomar la determinación de aplicar una imputación múltivariante para los patrones más numerosos y con un no elevado conjunto de variables a imputar. En esta situación este tipo de imputación consiste en realizar una variable transformada que contenga tantas categorías como el producto de categorías de las distintas variables que intervienen en el patrón de no-respuesta. Solamente van a poder tratarse patrones de no-respuesta categóricos nominales.

ESTUDIO DESCRIPTIVO DE LAS VARIABLES

TASAS DE NO-RESPUESTA DE EUSKADI Y LLANADA ALAVESA

En la tabla que se incluye a continuación aparece el número de registros missing para cada variable que interviene en la Estadística de Población y Vivienda de Euskadi de 1996. Para cada variable se distinguen dos ámbitos geográficos.

En primer lugar aparecen los resultados obtenidos para la C. A. de Euskadi. Se indica el número de registros con dicha característica missing junto con el porcentaje de no-respuesta. Este ratio se obtiene al realiza el cociente entre el número de registros missing para dicha variable y el número total de registros que intervienen en dicha variable. En el siguiente bloque aparecen los resultados obtenidos para la Llanada Alavesa, comarca sobre la cual se ha desarrollado la mayor parte del estudio. En este grupo, al igual que para la Comunidad completa, se indica el número de registros missing y la tasa de no-respuesta.

Las variables con mayor tasa de no-respuesta en la Llanada Alavesa son las referentes a código de nivel de instrucción con 57,44% de valores faltantes y precódigo de estudios con el 30,69%. Considerando la C. A. de Euskadi son las mismas variables las que proporcionan las mayores tasas de no-respuesta, aunque con porcentajes sensiblemente inferiores: 35,86% para la variable código de instrucción y 13,26% para precódigo de estudios.

TASAS DE NO-RESPUESTA

Variables	Descripción	C. A. de Euskadi		Llanada Alavesa	
		Número de registros missing	Tasa de no-respuesta	Número de registros missing	Tasa de no-respuesta
SITR	Situación de residencia	0	0,00%	0	0,00%
SEXO	Sexo	0	0,00%	0	0,00%
AGNN	Año de nacimiento	1.921	0,09%	19	0,01%
SITUX	Situación actual de la persona	0	0,00%	0	0,00%
REPP	Relación con la primera persona	171.313	7,59%	16.819	7,02%
CONY	Figura el cónyuge	154.993	6,86%	14.221	5,94%
FIPA	Figura el padre	295.687	13,10%	27.825	11,62%
FIMA	Figura el padre y/o la madre	-	-	-	-
ECIV	Estado civil legal	154.839	6,86%	14.253	5,95%
THIJ	Tiene hijos (1-3 y M)	154.865	6,86%	14.214	5,94%
NHIJ3	Número de hijos	140.662	6,23%	1.181	0,49%
RELA1	Relación con la actividad 1	231.231	10,24%	42.711	17,83%
EKEN	Entiende euskera	223.590	9,90%	37.429	15,63%
EKHA	Habla euskera	247.354	10,95%	44.260	18,48%
EKLE	Lee euskera	260.007	11,52%	45.329	18,93%
EKES	Escribe euskera	263.999	11,69%	45.752	19,10%
LMAT	Lengua materna	219.658	9,73%	35.103	14,66%
LHAB	Lengua hablada en casa	220.662	9,77%	34.922	14,58%
SEDE	Sedentarios	227.568	10,08%	35.086	14,65%
CEST	Precódigo de estudios	299.488	13,26%	73.495	30,69%
C_LEST2	Código de nivel de instrucción	809.612	**35,86%**	137.571	**57,44%**

PATRONES DE NO-RESPUESTA

A continuación se incluyen los principales patrones de no-respuesta existentes en los registros de la Estadística de Población y Vivienda para la comarca de la Llanada Alavesa. Los patrones de no-respuesta detectan los grupos de registros con valor missing simultáneamente en un idéntico conjunto de variables. Mediante estos patrones de no-respuesta se puede conocer si existe una tendencia de ciertos grupos de población a no responder a una serie de variables simultáneamente.

Con respecto a los patrones de no-respuesta realizados para la Estadística de Población y Vivienda se han reducido las variables a estudio debido al elevado número de éstas. Se han seleccionado las variables que han sido consideradas interesantes para el estudio de la imputación mediante árboles de clasificación. Estas variables son: año de nacimiento (agnn), relación con la primera persona (repp), figura el cónyuge (cony), figura el padre (fipa), estado civil (eciv), número de hijos (thij), relación con la actividad (rela1), sedentario (sede) y estudios (cest).

Los patrones de no-respuesta aparecen ordenados por el número de registros que tiene cada patrón y el porcentaje de registros sobre el total que tienen algún valor missing. En total se han obtenido 94.178 registros con alguna variable con valor missing y los patrones de respuesta obtenidos son 47. Dichos patrones se detallan a continuación en la siguiente tabla.

Como se puede comprobar aparecen las distintas variables con recuadros en color blanco y negro. El color negro representa la falta de respuesta de dicha variable en el patrón considerado mientras que el color blanco indica que se conoce el valor de dicha variable. Junto a estas categorías aparece el número de registros con dicho patrón y el porcentaje sobre el total de registros con alguna variable missing.

PATRONES DE NO-RESPUESTA										
AGNN	REPP	CONY	FIPA	ECIV	THIJ	RELA1	SEDE	CEST	Frecuencia	%
								■	33.615	35,69%
	■	■	■	■		■	■	■	13.020	13,82%
						■	■	■	12.798	13,59%
			■						9.044	9,60%
						■			6.238	6,62%
						■		■	5.302	5,63%
							■	■	2.449	2,60%
							■		2.380	2,53%
			■					■	2.157	2,29%
	■	■	■	■	■	■	■	■	1.134	1,20%
			■			■	■	■	958	1,02%
						■	■		929	0,99%
	■								880	0,93%
	■					■	■	■	763	0,81%
			■			■			560	0,59%
			■			■		■	464	0,49%
	■							■	368	0,39%
	■					■			258	0,27%
			■				■		172	0,18%
			■				■	■	157	0,17%
	■					■		■	137	0,15%
	■						■	■	125	0,13%
			■			■	■		94	0,10%
	■						■		51	0,05%
				■					27	0,03%
	■	■	■	■	■				21	0,02%
	■					■	■		19	0,02%
■	■	■	■	■		■	■	■	19	0,02%
				■				■	6	0,01%
	■	■	■	■	■	■		■	6	0,01%
	■	■	■	■	■		■		5	0,01%
	■	■	■			■	■	■	4	0,00%
	■	■	■	■	■			■	3	0,00%
				■		■		■	2	0,00%
	■	■	■	■	■	■			2	0,00%
				■			■		1	0,00%
				■				■	1	0,00%
			■	■				■	1	0,00%
		■		■		■	■	■	1	0,00%
		■		■	■	■	■	■	1	0,00%
		■	■	■	■	■	■	■	1	0,00%
	■	■		■	■		■	■	1	0,00%
	■	■	■	■		■	■		1	0,00%
	■	■	■		■	■	■	■	1	0,00%
	■	■	■	■	■		■	■	1	0,00%
■									1	0,00%
									94.178	

Se puede observar que el patrón de no-respuesta más numeroso es el que incluye la variable estudios realizados con un total de 33.615 registros que no han contestado a está variable y sí al resto de las variables en estudio, representando el 35,89% de los registros con algún valor missing.

MEDIDAS DE ASOCIACIÓN

Al tratar variables provenientes de censo se tiene la particularidad de que la mayor parte son variables categóricas nominales y ordinales. Por esta causa no se pueden aplicar las medidas de asociación clásicas para variables continuas (correlación de Pearson). Existen diversos coeficientes obtenidos generalmente a partir de tablas de contingencia de variables categóricas o continuas agrupadas en intervalos, se distinguen según el tipo de variable que estemos tratando: nominal (lambda simétrica, coeficientes de incertidumbre,..) u ordinal (rangos de Spearman, Tau-b de Kendall, Gamma,...). Hay otras medidas calculadas a partir del estadístico chi-cuadrado de Pearson que sirven tanto para variables nominales como ordinales: Phi coeficiente, coeficiente de contingencia y V de Cramer.

Las variables sobre las que se ha aplicado la imputación han sido de tipo nominal (estado civil, relación con la actividad, situación profesional y lugar de trabajo) y por tanto nos vamos a fijar en las medidas de asociación relativas a este tipo de variables.

En el anexo II se incluye una salida de SAS tras realizar una tabla de contingencia entre las variables relación con la actividad (RELA1) y estado civil (ECIV). Junto con la tabla de contingencia aparecen las distintas medidas de asociación que proporciona el software. Como ambas variables son categóricas nominales nos debemos de fijar únicamente en las medidas adecuadas para este tipo de variables. Dichas medidas aparecen resaltadas.

Dicha información se va emplear para decidir qué variables van a incluirse en el modelo como variables explicativas para la construcción del árbol de clasificación. Se han seleccionado tres medidas para dicho estudio debido a que el rango de valores varía en el intervalo (0,1) y esto nos va facilitar la comparación entre las distintas variables. El valor cuanto más próximo a 1 sea mayor será la asociación entre ambas variables y por el contrario, cuanto más próximo a 0 menor será la asociación. Estas son:

- V de Cramer
- Lambda asimétrica $l\left(C|R\right)$.
- Coeficiente de incertidumbre asimétrico $U\left(C|R\right)$.

Se incluyen la lambda y coeficiente de incertidumbre asimétricos, en lugar de las respectivas medidas simétricas, debido a que en esta situación nos interesa estudiar la capacidad de predecir el valor de la variable a imputar a partir de la potencial variable explicativa.

En el ejemplo se intenta estudiar las posibles variables explicativas a incluir para construir el árbol que discrimine la variable relación con la actividad (RELA1, variable columna). En la tabla de contingencia en la cual se enfrentan las variables relación con la actividad (RELA1) y el estado civil (ECIV) se obtienen 0.44616 para la V de Cramer, 0,1973 para la lambda asimétrica $l(C|R)$ y 0,2157 para el coeficiente de incertidumbre asimétrico $U(C|R)$.

A continuación se incluyen, a modo de resumen, las tres medidas de asociación consideradas anteriormente junto con la no-respuesta simultánea de las variables relación con la actividad y la considerada en cada momento. Si se incluye una variable con alta tasa de no-respuesta simultánea puede ocasionar graves problemas a la hora de clasificar los registros con valor missing. Por tanto existen diversas alternativas posibles:

- No incluir dicha variable como explicativa en la construcción del árbol de clasificación, aun en el caso de existir una importante asociación entre ambas variables. Esta opción puede provocar una pérdida de potencial en la clasificación de registros que si posean valor en la variable explicativa.

- Realizar una imputación simultanea de ambas variables mediante una imputación multivariante.

- Realizar dos árboles, uno incluyendo dicha variable explicativa y emplearlo para los registros con valor no missing en esta variable. El segundo árbol no empleará dicha variable y se clasificarán a los registros con no-respuesta simultánea.

- Imputar previamente la variable que provoca dicho problema.

Variable / RELA1	*Descripción*	*V de Cramer*	*Lambda C|R*	*coeficiente de incertidumbre C|R*	*No repuesta conjunta*	*Tasa de no respuesta*
MUNR	Municipio	0,0225	0	0,0013	0	0,00%
SEXO	Sexo	0,4751	0,0404	0,0815	0	0,00%
SITUX	Situación actual de la persona	0,0506	0	0,0019	0	0,00%
REPP	Relación con la primera persona	0,42734	0,3057	0,3116	15.363	35,97%
CONY	Figura el cónyuge	0,64005	0,1559	0,1047	14.189	33,22%
FIPA	Figura el padre	0,78144	0,2289	0,2067	16.263	38,08%
RELE	Relación con el establecimiento colectivo	0,128	0,0026	0,0082	0	0,00%
REFA1	Relaciones familiares en colectivos	0,0691	0,0017	0,0039	0	0,00%
ECIV	Estado civil	0,44616	0,1973	0,2157	14.186	33,21%
THIJ	Número de hijos	0,35511	0,1417	0,1705	14.183	33,21%
EKEN	Entiende euskera	0,3757	0,1126	0,0861	30.879	72,30%
EKHA	Habla euskera	0,4669	0,1261	0,0835	32.005	74,93%
EKLE	Lee euskera	0,3666	0,1121	0,0808	32.206	75,40%
EKES	Escribe euskera	0,3711	0,1207	0,0819	32.216	75,43%
LMAT	Lengua materna	0,0924	0,015	0,0065	30.269	70,87%
LHAB	Lengua hablada en casa	0,0527	0,0035	0,0032	30.261	70,85%
SEDE	Sedentario (sí/no)	0,4081	0,0447	0,0518	29.742	69,64%
CEST	Código de estudios	0,27218	0,2031	0,2089	34.611	81,04%
EDADRELA	Edad categorizada	0,48992	0,4116	0,4	20	0,05%
TIPOLOGIA	Tipología de la sección censal	0,0781	0	0,0143	0	0,00%

A partir de las cifras de las medidas de asociación consideradas se observa como existe una importante relación con la variable a imputar, RELA1, por parte del estado civil, edad, figura el cónyuge, figura el padre,... variando dichas cifras al considerar las distintas medidas de asociación planteadas. Existen otras variables con cierta asociación con la relación con la actividad como pueden ser el código de estudios pero presenta el problema añadido de poseer una elevada no respuesta simultánea entre ambas variables. En este caso el 81,04% de los individuos con no-respuesta en la variable a imputar (RELA1) tampoco han respondido a la variable de nivel de estudio. Esta característica nos indica que no se puede considerar la inclusión de dicha variable en la construcción del árbol directamente y se debería de considerar alguna de las opción consideradas anteriormente cuando se detecta dicha situación.

Una vez analizada la asociación existente entre la relación con la actividad y las variables del estudio se ha decidido construir un árbol de clasificación para la imputación de la variable relación con la actividad donde van a intervenir como variables explicativas: sexo, edad agrupada en siete categorías, la tipología de la sección censal, relación con la primera persona, figura cónyuge, figura el padre, estado civil y número de hijos.

La tasa de error global del árbol de clasificación obtenida a partir de los registros que han participado en el árbol es del 27,69%, se reduce 3,25 puntos si lo comparamos con el caso de emplear el árbol donde participaban tres variables en la clasificación (sexo, edad agrupada en siete categorías y la tipología de la sección censal con 30,942%) . Hay que tener en cuenta que este valor suele ser sesgado inferiormente.

Esta tasa de error se calcula de la siguiente forma:

Nº registros bien clasificados / Nº registros totales

Por este motivo se estudiará posteriormente la calidad del árbol con registros que no participen en su construcción y de los cuales sí tengamos información sobre el valor de la variable relación con la actividad.

CONSERVACIÓN DE LA DISTRIBUCIÓN DE FRECUENCIAS REAL

Una de las propiedades que deben cumplir los métodos de imputación es la conservación de la distribución real tras realizar la imputación. Por esta razón es interesante realizar gráficos o aplicar contrastes que corroboren dicha propiedad. Se han propuesto múltiples estadísticos para estudiar la conservación de la distribución previa a la imputación como pueden ser, entre otros, el estadístico de Wald o el contraste de bondad de ajuste. Para esta investigación se ha decidido aplicar el estadístico de bondad de ajuste debido a que está disponible en el paquete estadístico SAS.

Considerando el ejemplo planteado para la imputación de la variable relación con la actividad a partir del árbol generado incluyendo las variables indicadas anteriormente se ofrecen los siguientes resultados.

Distribución de frecuencias de relación con la actividad a priori

En primer lugar, se presenta la distribución de frecuencias previa a la imputación. En el caso de no encontrarse en situaciones en las cuales la falta de respuesta se deba al propio valor de la variable a imputar se puede considerar la distribución previa a la imputación como la real y analizar las posibles modificaciones que se produzcan tras el proceso de la imputación.

A continuación se detalla la distribución de frecuencias de la variable relación con la actividad antes de imputar, considerando solo válidos aquellos registros que no provocan inconsistencias en las reglas de validación planteadas en el análisis.

The FREQ Procedure

RELA1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Servicio Militar	723	0.37	723	0.37
Ocupado	72129	37.02	72852	37.40
Parado. 1º empleo	5322	2.73	78174	40.13
Parado ya ha trabajado	11480	5.89	89654	46.02
Jubilado	19971	10.25	109625	56.27
Otros pensionistas	5369	2.76	114994	59.03
Incapacitado permanente	840	0.43	115834	59.46
Estudiante	45250	23.23	161084	82.69
Labores del hogar	29723	15.26	190807	97.94
Otra situación	4007	2.06	194814	100.00

Distribución de frecuencias de relación con la actividad tras la imputación realizada en 1.996

Además de los resultados obtenidos mediante el nuevo método planteado se presentan los resultados obtenidos en la Estadística de Población y Vivienda de 1996. La imputación se realizó mediante una asignación aleatoria en base a la distribución obtenida para subgrupos de población: combinando las dos categorías de sexo, siete grupos de año de nacimiento y dos de tipo de municipio (agrícola o no agrícola).

En la tabla que se presenta se observa cómo aparecen pequeñas diferencias entre las categorías de la variable relación con la actividad. Esto junto al gran tamaño de muestra provoca que al aplicar el test de bondad de ajuste de la chi-cuadrado nos indica que se rechaza H_0 ya que el estadístico tiene un valor de 312,9326 y un p-valor inferior a 0,0001, es decir, no se puede aceptar que siga la misma distribución que antes de imputar.

The FREQ Procedure

rela1des	Frequency	Percent	Test Percent	Cumulative Frequency	Cumulative Percent
Servicio Militar	885	0.37	0.37	885	0.37
Ocupado	88751	37.06	37.02	89636	37.43
Parado. 1º empleo	6396	2.67	2.73	96032	40.10
Parado ya ha trabajado	14078	5.88	5.89	110110	45.98
Jubilado	24728	10.33	10.25	134838	56.31
Otros pensionistas	6654	2.78	2.76	141492	59.09
Incapacitado permanente	973	0.41	0.43	142465	59.49
Estudiante	56123	23.44	23.23	198588	82.93
Labores del hogar	34901	14.57	15.26	233489	97.50
Otra situación	5982	2.50	2.06	239471	100.00

Frequency Missing = 23

Chi-Square Test for Specified Proportions

Chi-Square	312.9326
DF	9
Pr > ChiSq	<.0001

Effective Sample Size = 239471
Frequency Missing = 23

Distribución de frecuencias de relación con la actividad tras la imputación mediante árboles de clasificación

A partir del árbol CHAID obtenido se procede a realizar la imputación de datos clasificando los registros con no-respuesta en la variable relación con la actividad. Tras esto se procede al cálculo de las distribuciones de frecuencias obtenidas tanto en el caso de aplicar dentro del nodo terminal asignado a cada registro una imputación aleatoria en base a la distribución de frecuencias de dicho nodo o asignándole a la categoría de mayor probabilidad.

Imputación mediante la selección aleatoria según la distribución de frecuencias

La siguiente tabla representa la distribución de frecuencias de los registros imputados mediante árboles de clasificación al seleccionar aleatoriamente según la distribución del nodo terminal. Al realizar el contraste chi-cuadrado de bondad de ajuste se obtiene un valor del estadístico de 246.349 y le corresponde un p-valor asociado menor a 0,0001. Esto nos obliga a rechazar la hipótesis nula de conservación de la distribución. El valor obtenido es inferior al que se obtuvo con la imputación de 1.996, (312,9326) y nos lleva a concluir que se modifica en menor medida la distribución.

rela1arb	Frequency	Percent	Test Percent	Cumulative Frequency	Cumulative Percent
Servicio Militar	809	0.34	0.37	809	0.34
Ocupado	89474	37.68	37.02	90283	38.02
Parado. 1º empleo	6449	2.72	2.73	96732	40.74
Parado ya ha trabajado	13223	5.57	5.89	109955	46.31
Jubilado	23155	9.75	10.25	133110	56.06
Otros pensionistas	6324	2.66	2.76	139434	58.72
Incapacitado permanente	962	0.41	0.43	140396	59.12
Estudiante	56959	23.99	23.23	197355	83.11
Labores del hogar	35065	14.77	15.26	232420	97.88
Otra situación	5037	2.12	2.06	237457	100.00

Frequency Missing = 2037

The FREQ Procedure

Chi-Square Test for Specified Proportions

Chi-Square	246.3490
DF	9
Pr > ChiSq	<.0001

Effective Sample Size = 237457
Frequency Missing = 2037

Imputación a la categoría de máxima probabilidad

A continuación aparece la distribución de frecuencias obtenidas tras la imputación mediante el árbol de clasificación formado a partir de las variables comentadas anteriormente, con la diferencia de imputar a la categoría más probable dentro del nodo terminal.

De esta forma se obtienen mayores diferencias en la distribución de frecuencias si se compara con la obtenida antes de realizar la imputación. Esto se puede comprobar al observar el valor del estadístico de bondad de ajuste de la chi-cuadrado que tiene un valor de 1374.1390, muy superior al obtenido en los dos casos anteriores. Esto es debido a que mediante este método ha habido categorías a las cuales no han sido imputadas ninguno de los 44.680 registros con valor missing.

maxpr	Frequency	Percent	Test Percent	Cumulative Frequency	Cumulative Percent
Servicio Militar	723	0.30	0.37	723	0.30
Ocupado	92351	38.92	37.02	93074	39.22
Parado. 1º empleo	5332	2.25	2.73	98406	41.47
Parado ya ha trabajado	11548	4.87	5.89	109954	46.33
Jubilado	23090	9.73	10.25	133044	56.06
Otros pensionistas	6533	2.75	2.76	139577	58.82
Incapacitado permanente	841	0.35	0.43	140418	59.17
Estudiante	58370	24.60	23.23	198788	83.77
Labores del hogar	34355	14.48	15.26	233143	98.25
Otra situación	4161	1.75	2.06	237304	100.00

Frequency Missing = 2190

Chi-Square Test for Specified Proportions

Chi-Square	1374.1390
DF	9
Pr > ChiSq	<.0001

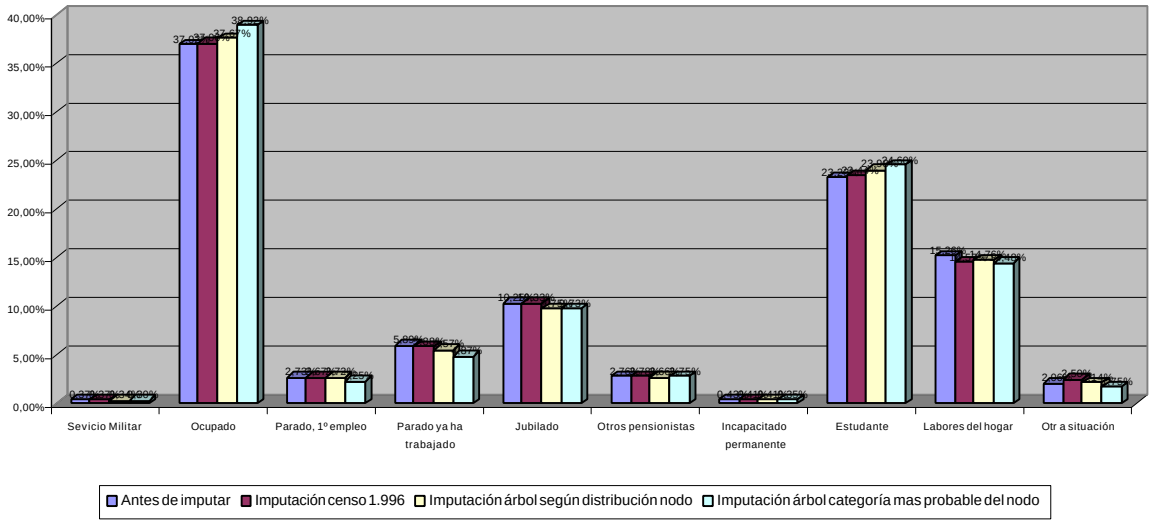
Effective Sample Size = 237304
Frequency Missing = 2190

A modo de resumen de las distintas distribuciones de frecuencias comentadas anteriormente, se incluye la siguiente tabla donde aparecen la distribución de frecuencias antes de aplicar la imputación, tras la imputación realizada en 1.996 y los dos nuevos métodos de imputación basados en árboles de clasificación propuestos: Seleccionando aleatoriamente según la distribución de frecuencias del nodo terminal o asignando a la categoría de mayor probabilidad de dicho nodo.

RELA1	Antes de imputar		Imputación censo 1.996		Imputación árbol según distribución nodo		Imputación árbol categoría más probable del nodo	
	Frecuencia	Porcentaje	Frecuencia	Porcentaje	Frecuencia	Porcentaje	Frecuencia	Porcentaje
Servicio Militar	723	0,37%	885	0,37%	809	0,34%	723	0,30%
Ocupado	72.129	37,02%	88.751	37,06%	89.474	37,67%	92.351	38,92%
Parado, 1º empleo	5.322	2,73%	6.396	2,67%	6.449	2,72%	5.332	2,25%
Parado ya ha trabajado	11.480	5,89%	14.078	5,88%	13.223	5,57%	11.548	4,87%
Jubilado	19.971	10,25%	24.728	10,33%	23.155	9,75%	23.090	9,73%
Otros pensionistas	5.369	2,76%	6.654	2,78%	6.324	2,66%	6.533	2,75%
Incapacitado permanente	840	0,43%	973	0,41%	962	0,41%	841	0,35%
estudiante	45.250	23,23%	56.123	23,44%	56.959	23,98%	58.370	24,60%
Labores del hogar	29.723	15,26%	34.901	14,57%	35.065	14,76%	34.355	14,48%
Otra situación	4.007	2,06%	5.982	2,50%	5.087	2,14%	4.161	1,75%
	Missing=44.680		Missing=23		Missing=2.037		Missing=2.190	
Total	194.814		239.471		237.507		237.304	

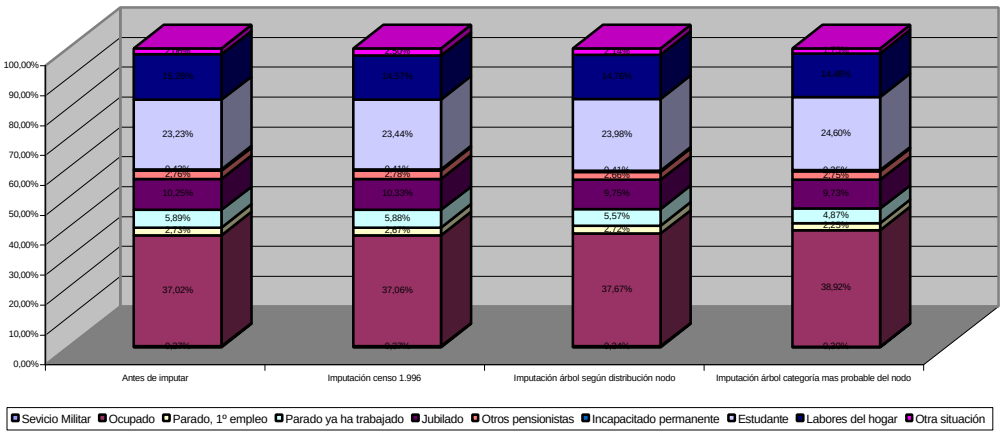
En los gráficos que aparecen a continuación se comparan las distribuciones de frecuencias obtenidas aplicando los métodos comentados anteriormente. Se observa, como ya se había indicado al aplicar el test de bondad de ajuste, que el método que imputa según la categoría más probable dentro del nodo terminal es el que más

Distribución de frecuencias de los valores imputados



modifica la distribución original.

Distribución de frecuencias de los valores imputados



CALIDAD DE LA IMPUTACIÓN

Para conocer la calidad de los valores imputados existen distintas técnicas que comparan el valor real del registro con el valor con que ha sido imputado. En este estudio se ha optado por emplear el estimador por muestra de contraste. Este estimador consiste en dejar caer por el árbol una muestra distinta a la empleada para la realización del árbol y por tanto, no se ha podido adaptar el árbol a dichos registros como ocurría al emplear el estimador intramuestral. Una vez asignados a un nodo terminal, se realiza la imputación mediante los distintos métodos propuestos, en nuestro caso mediante la asignación aleatoria en función de la distribución de frecuencias ó asignando a la categoría de mayor probabilidad de dicho nodo. Hay cierta pérdida de información pero se suele emplear dicho estimador en el caso de estar ante un tamaño de muestra elevado, como ocurre en el caso de los censos.

Siguiendo con la situación anterior se ha construido un árbol para el estudio de calidad del árbol de clasificación para la variable respuesta relación con la actividad donde participan en su construcción las variables explicativas empleadas anteriormente. Para dicho análisis se han dividido los registros con respuesta en la variable relación con la actividad en dos grupos de forma aleatoria:

- El conjunto de datos de mayor tamaño se va a emplear para la construcción del árbol de clasificación donde van a participar las mismas variables explicativas que en la imputación realizada.

- El segundo conjunto de datos va a contener los registros sobre los cuales se va imputar mediante el árbol de clasificación construido con el grupo anterior. De esta forma vamos a tener para cada individuo de este fichero de datos el valor real de la relación con la actividad y el imputado. Dependiendo del número de coincidencias entre ambos valores vamos a obtener la tasa de buena clasificación del árbol.

Se ha seleccionado una muestra aleatoria de 16.234 individuos del total de registros con valor no missing en la variable relación con la actividad (194.814 registros). Con los restantes 178.580 registros se han empleado para la construcción del árbol que se va a emplear para clasificar los registros.

	Arbol según distribución nodo/ censo96	Arbol categoría más probable/censo96
Bien clasificado	9.689	11.559
Mal clasificado	6.545	4.675
Total	16.224	16.174
Missing	10	60
valores con respuesta	16.234	16.234
tasa de buena clasificación	**59,7202%**	**71,4666%**

Los resultados globales obtenidos son: para el caso de imputación según la distribución de frecuencias 9.689 registros han sido bien clasificados y representan el 59,72% de los casos. Mientras que para el caso de imputación según la categoría más probable se clasifican correctamente 11.559 registros que representan el 71,46% de los casos.

Se procede a realizar un estudio de la conservación de la distribución real, en este caso conocida, comparándolo con la distribución de los valores obtenidos tras aplicar el proceso de imputación a esos mismos registros.

Distribución de frecuencias real de relación con la actividad

Se incluye en primer lugar la distribución de frecuencias de los registros con sus valores reales que posteriormente se va a emplear para el estudio de la conservación de la distribución de frecuencias.

The FREQ Procedure

RELA1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Servicio Militar	52	0.32	52	0.32
Ocupado	5984	36.86	6036	37.18
Parado. 1º empleo	456	2.81	6492	39.99
Parado ya ha trabajado	974	6.00	7466	45.99
Jubilado	1636	10.08	9102	56.07
Otros pensionistas	441	2.72	9543	58.78
Incapacitado permanente	82	0.51	9625	59.29
Estudiante	3781	23.29	13406	82.58
Labores del hogar	2494	15.36	15900	97.94
Otra situación	334	2.06	16234	100.00

Distribución de frecuencias de los valores imputados de relación con la actividad

Tras realizar la imputación se procede al cálculo de las distribuciones de ambas imputaciones realizadas (imputación aleatoria en base a la distribución y asignación a la categoría más probable del nodo terminal) realizando el análisis de conservación de la distribución.

Imputación aleatoria según la distribución de frecuencias del nodo terminal

En la siguiente tabla aparece la distribución de frecuencias de la relación con la actividad imputada mediante selección aleatoria dentro del nodo terminal según la distribución de frecuencias. El test de bondad de ajuste de la chi-cuadrado que se ha realizado para contrastar si la distribución de los valores reales se conserva tras aplicarles la imputación nos da un valor del estadístico de 15,831 con un p-valor asociado de 0,0705. Con este p-valor, si tomamos como valor crítico a =0,05, se acepta la hipótesis nula de conservación de la distribución tras la imputación.

The FREQ Procedure

rela1arb	Frequency	Percent	Test Percent	Cumulative Frequency	Cumulative Percent
Servicio Militar	74	0.46	0.32	74	0.46
Ocupado	5980	36.86	36.86	6054	37.32
Parado. 1º empleo	445	2.74	2.81	6499	40.06
Parado ya ha trabajado	1027	6.33	6.00	7526	46.39
Jubilado	1609	9.92	10.08	9135	56.31
Otros pensionistas	434	2.68	2.72	9569	58.98
Incapacitado permanente	70	0.43	0.51	9639	59.41
Estudiante	3781	23.30	23.29	13420	82.72
Labores del hogar	2457	15.14	15.36	15877	97.86
Otra situación	347	2.14	2.06	16224	100.00

Frequency Missing = 10

Chi-Square Test for Specified Proportions

Chi-Square	15.8310
DF	9
Pr > ChiSq	0.0705

Effective Sample Size = 16224
Frequency Missing = 10

Imputación a la categoría de máxima probabilidad del nodo terminal

En la siguiente tabla se observa la distribución de frecuencias tras la imputación mediante el árbol de clasificación y realizando dentro del nodo terminal una asignación al registro de mayor probabilidad. En esta situación no se puede aplicar en SAS directamente el contraste de bondad debido a que no se ha imputado a todas las categorías de la variable relación con la actividad. Este es un problema de este método de imputación, puede existir una categoría con características muy particulares que debido a su reducido tamaño y características heterogéneas de este subconjunto de la

población no se genere ningún nodo terminal que contenta dicha categoría como la que proporcione la mayor probabilidad en ningún nodo terminal. Esta situación ha ocurrido con la categoría de servicio militar. Calculando el valor del estadístico se obtiene un valor de 2.364,23437 y un p -valor inferior a 0,0001, con lo cual nos lleva a rechazar la hipótesis de conservación de la distribución.

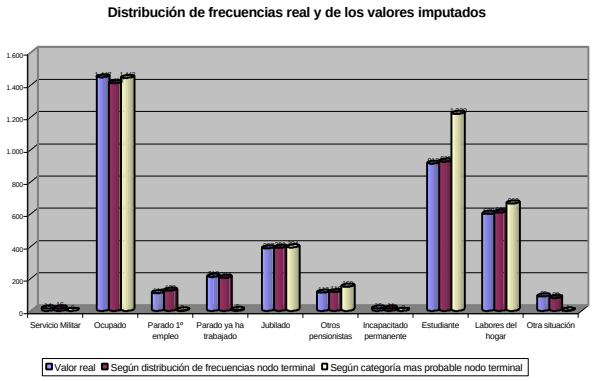
maxpr	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Ocupado	6076	37.57	6076	37.57
Parado. 1º empleo	3	0.02	6079	37.59
Parado ya ha trabajado	13	0.08	6092	37.67
Jubilado	1612	9.97	7704	47.63
Otros pensionistas	600	3.71	8304	51.34
Incapacitado permanente	3	0.02	8307	51.36
Estudiante	5064	31.31	13371	82.67
Labores del hogar	2789	17.24	16160	99.91
Otra situación	14	0.09	16174	100.00

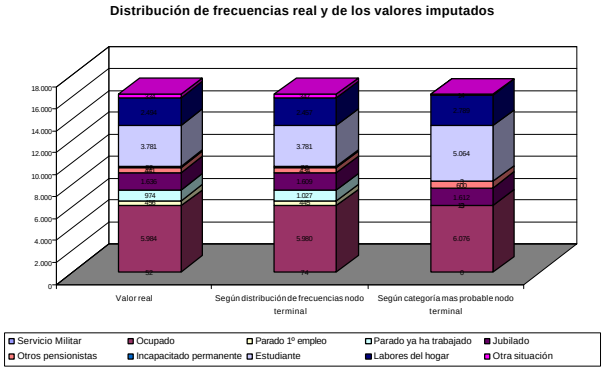
Frequency Missing = 60

A modo de resumen se incluye la siguiente tabla donde aparecen las distribuciones de frecuencias de los valores reales y los dos métodos de imputación comentados anteriormente.

Relación con la actividad	***Valor real***		***Según distribución de frecuencias nodo terminal***		***Según categoría más probable nodo terminal***	
	Frecuencias	Porcentaje	Frecuencias	Porcentaje	Frecuencias	Porcentaje
Servicio Militar	52	0,32%	74	0,46%	0	0,00%
Ocupado	5.984	36,86%	5.980	36,86%	6.076	37,57%
Parado 1º empleo	456	2,81%	445	2,74%	3	0,02%
Parado ya ha trabajado	974	6,00%	1.027	6,33%	13	0,08%
Jubilado	1.636	10,08%	1.609	9,92%	1.612	9,97%
Otros pensionistas	441	2,72%	434	2,68%	600	3,71%
Incapacitado permanente	82	0,51%	70	0,43%	3	0,02%
Estudiante	3.781	23,29%	3.781	23,30%	5.064	31,31%
Labores del hogar	2.494	15,36%	2.457	15,14%	2.789	17,24%
Otra situación	334	2,06%	347	2,14%	14	0,09%
			Missing=10		Missing=60	
Total	16.234		16.224		16.174	

Los gráficos que se incluyen a continuación reflejan la variación que se produce entre la distribución de frecuencias original y la que se obtiene al aplicar los dos métodos de imputación propuestos. Se puede comprobar como la imputación basada en la





categoría más probable del nodo terminal modifica en mayor medida la distribución.

Para el estudio de la conservación del valor real se puede calcular la tabla de contingencia donde se enfrenta la variable que contiene los valores reales con los valores imputados de la relación con la actividad. En el caso de imputación perfecta deben aparecer únicamente elementos en la diagonal. Como esta situación es prácticamente imposible existen distintas herramientas que nos pueden indicar el grado de buena imputación mediante el método aplicado:

Cálculo de la tasa de buena clasificación

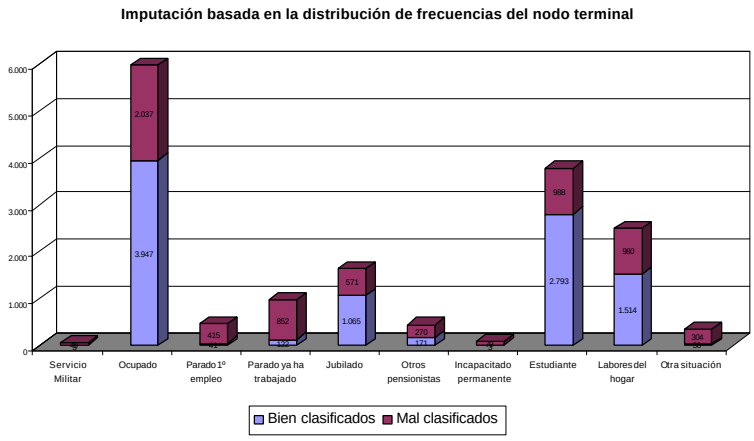
Consiste en comparar el valor imputado con el valor real de cada registro. Se realiza un conteo de todos aquellos registros que han sido bien imputados y se divide por el total de registros imputados. De esta forma se tiene una medida aproximada del porcentaje de elementos bien imputados mediante este método. También se puede realizar dicho cálculo para cada categoría de la variable imputada. De esta forma se puede comprobar la capacidad de imputar correctamente cada categoría de la variable.

Volviendo al estudio de la variable relación con la actividad en la siguiente tabla se incluyen las categorías de dicha variable junto con las tasas de buena y mala clasificación. Las categorías que mejores tasas de buena clasificación proporcionan son estudiante, ocupado y jubilado que superan en ambos métodos el 65%, sin embargo tienen tasa de buena clasificación escasa o incluso nula las categorías servicio militar, parado tanto de primer empleo como los que ya han trabajado anteriormente, otra situación e incapacitado permanente.

Relación con la actividad	Según distribución de frecuencias nodo terminal				Según categoría más probable nodo terminal			
	Bien clasificados		Mal clasificados		Bien clasificados		Mal clasificados	
	Frecuencias	Porcentaje	Frecuencias	Porcentaje	Frecuencias	Porcentaje	Frecuencias	Porcentaje
Servicio Militar	3	5,77%	49	94,23%	0	0,00%	52	100,00%
Ocupado	3.947	65,96%	2.037	34,04%	4.514	75,43%	1.470	24,57%
Parado 1º empleo	41	8,99%	415	91,01%	1	0,22%	455	99,78%
Parado ya ha trabajado	122	12,53%	852	87,47%	2	0,21%	972	99,79%
Jubilado	1.065	65,10%	571	34,90%	1.177	71,94%	459	28,06%
Otros pensionistas	171	38,78%	270	61,22%	293	66,44%	148	33,56%
Incapacitado permanente	3	3,66%	79	96,34%	0	0,00%	82	100,00%
Estudiante	2.793	73,87%	988	26,13%	3.671	97,09%	110	2,91%
Labores del hogar	1.514	60,71%	980	39,29%	1.896	76,02%	598	23,98%
Otra situación	30	8,98%	304	91,02%	5	1,50%	329	98,50%
Total	9.689		6.545		11.559		4.675	

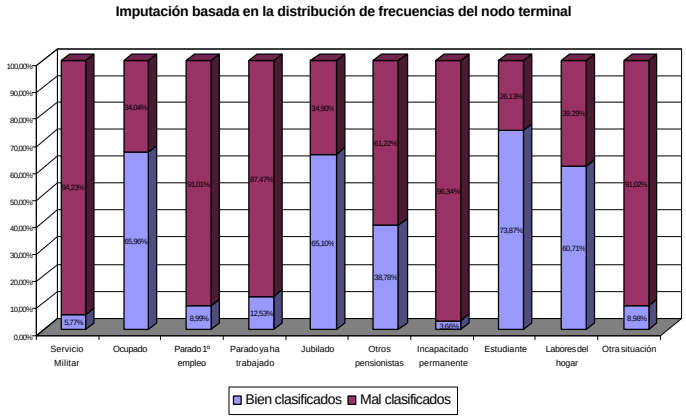
Existe una forma de representar la calidad de la imputación por categoría mediante diversos gráficos que se incluyen a continuación. En estos gráficos aparece la distribución de frecuencias de la variable imputada, en la cual aparece cada categoría dividida en dos franjas. La de color azul indica el número de registros pertenecientes a dicha categoría que han sido imputados correctamente mediante el método basado en el árbol de clasificación. Por el contrario la franja de color rojo indica los registros pertenecientes a la categoría que al aplicarles la imputación han sido asignadas a una categoría distinta.

Imputación basada en la distribución de frecuencias del nodo terminal

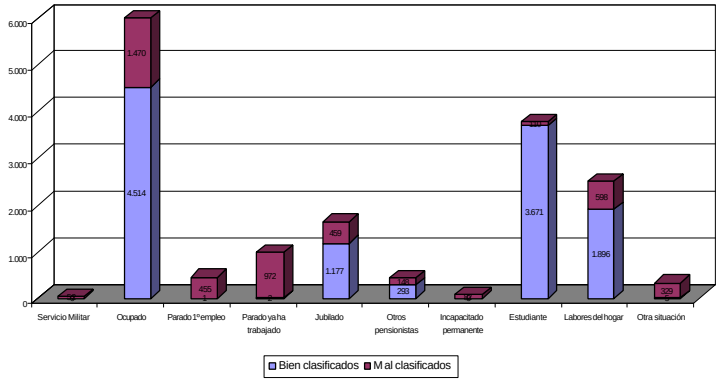


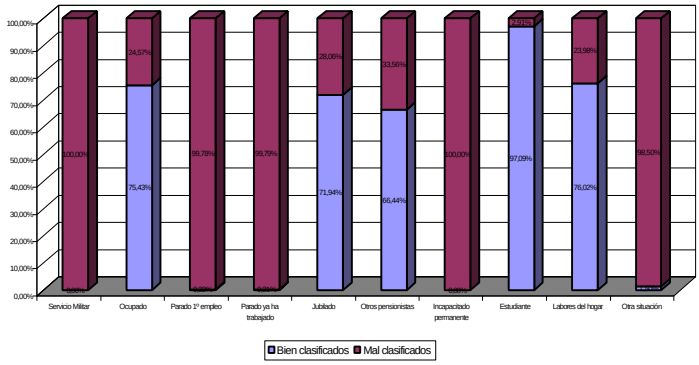
El primero de los gráficos representa la distribución original de la muestra de contraste realizada y se indica el número de registros pertenecientes a cada categoría que han sido correctamente imputados. Mientras que el segundo indica el porcentaje de registros pertenecientes a cada categoría que han sido correctamente imputados.

Imputación mediante la distribución de frecuencias del nodo terminal



Imputación mediante la categoría más probable del nodo terminal





Se puede observar como mediante el método de asignación aleatoria dentro del nodo terminal se obtienen menores porcentajes de buena clasificación que en el caso de asignación a la categoría más probable en el caso de fijarnos en las categorías que mejor se clasifican (ocupado, jubilado y estudiantes). Se obtienen sin embargo mejores porcentajes entre las categorías peor clasificadas mediante el segundo.

Mediante tablas de contingencia

Se construye, como se ha comentado anteriormente, una tabla de contingencia donde se enfrenta para cada registro el valor real y el imputado. Los registros correctamente imputados van a aparecer en la diagonal mientras que por el contrario los mal imputados van a aparecer fuera de ella.

En la tablas de contingencia que aparecen en el anexo III se presenta por filas el valor real de la variable relación con la actividad mientras que en las columnas aparece el valor asignado mediante la imputación basada en la asignación aleatoria según la distribución de frecuencias del nodo terminal del árbol realizado.

Imputación basada en la distribución de frecuencias del nodo terminal

En el anexo III tabla 1 aparece la tabla de contingencia que proporciona el programa SAS tras enfrentar los valores imputados según la distribución de frecuencias del nodo terminal y el valor real de cada registro. SAS proporciona diversas medidas que nos permite obtener unas medidas que indican la calidad de la imputación realizada.

- Una alternativa posible es emplear las medidas de asociación calculadas para los valores reales y los imputados.

  De esta forma, si nos fijamos en la lambda asimétrica $l(R \mid C)$ se puede comprobar la capacidad de predicción de la variable imputada (variable columna) sobre la variable con los valores reales (variable fila). Cuanto más próximo a 1 sea dicho valor mejor imputación se ha realizado. En nuestro ejemplo se observa que el valor de lambda asimétrica $l(R \mid C)$=0.4404.

- Otra alternativa consiste en las medidas de agrupamiento que proporciona SAS. Entre los que destacan la Kappa de Kohen y la simetría de Bowker.

  Test Kappa de Kohen : este contraste comprueba si existe correspondencia entre las categorías, es decir, si la categoría 1 en la primera variable corresponde a la 1 en la segunda, la categoría 2 de la primera variable con la 2 de la segunda, ... En nuestra situación este contraste es muy útil ya que comprueba si hay relación entre las mismas categorías de la variable seleccionada para imputar contrastando los valores reales con los imputados. En el caso de correspondencia total entre las categorías tendremos una matriz diagonal, que es la situación más favorable posible, debido a que en este caso se realizaría una imputación perfecta.

  En el ejemplo que estamos tratando se contrastará si hay relación entre ocupados de la variable con los valores reales con los ocupados imputados, estudiantes de la variable con valores reales con los estudiantes imputados,… En nuestra situación el estadístico Kappa nos proporciona un valor de 0.4774.

  Test de simetría de Bowker : la hipótesis nula de este es que tabla de contingencia satisfaga la simetría, es decir $p_{ij} = p_{ji}$ para todos los pares de celdas de la tabla.

$$\begin{cases} H_0 & p_{ij} = p_{ji} \quad \forall i \neq j \\ H_1 & p_{ij} \neq p_{ji} \quad para \quad a\lg ún \quad i, j \end{cases}$$

La aceptación de dicha hipótesis no nos asegura una buena imputación pero nos indica una medida de estudio de la conservación de la distribución ya que en el caso de producirse malas imputaciones de los registros pertenecientes a la categoría i que han sido imputados a la categoría j, se compensan con las imputaciones de los registros pertenecientes a j asignados a la categoría i.

En el ejemplo que estamos tratando se obtiene un valor del estadístico de simetría de 31.9912 y un p-valor asociado de 0.9277, con lo que se acepta la hipótesis nula de estar bajo una matriz simétrica.

Imputación basada en la categoría más probable del nodo terminal

La tabla de contingencia que aparece en el anexo III tabla 2, contiene el valor real de la relación con la actividad y el asignado tras aplicar la imputación basada en el árbol de clasificación y seleccionando la categoría más probable del nodo terminal. En este caso de la misma forma que en los estudios anteriores hay categorías a las cuales no son imputadas al aplicar dicho método y por tanto no se puede aplicar el contraste de simetría ni calcular el estadístico de Kappa.

Con respecto a las medidas de asociación se obtiene un valor de la V de Cramer de 0.4911, mientras que para la lambda asimétrica tiene un valor de 0.6661.

IMPUTACIÓN MULTIPLE DE LA RELACIÓN CON LA ACTIVIDAD

Se ha realizado una imputación con el fichero completo de la estadística de población y vivienda de 1996 (toda la Comunidad Autónoma) para la variable relación con la actividad (RELA1) mediante un proceso de imputación múltiple basado en árboles de clasificación.

Para la construcción del árbol de clasificación, con el cual se realizará la imputación de la variable relación con la actividad, han intervenido como variables explicativas: sexo, edad agrupada en siete categorías, la tipología de la sección censal, relación con la primera persona, figura cónyuge, figura el padre, estado civil y número de hijos.

Estudio de la conservación de la distribución

Uno de las características que se desea de un buen método de imputación es que produzcan los menores cambios posibles de la distribución de frecuencias previa, siempre que no exista relación entre el valor de la variable y la falta de respuesta.

Para el estudio de la conservación de la distribución de frecuencias se requiere conocer la distribución de frecuencias de la variable a imputar, en este caso relación con la actividad, tras el proceso de edición. En esta situación se poseen todos los registros con valor que han pasado todas las reglas de validación. Dichos registros son los que se

han seleccionado para que participen en la construcción de árbol. En total en el árbol han participado 2.009.362 registros y los 248.612 registros con valor missing en esta variable serán imputados mediante el árbol construido.

relat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Servicio Militar	4910	0.24	4910	0.24
Ocupado	660822	32.89	665732	33.13
Parado. 1º empleo	81360	4.05	747092	37.18
Parado ya ha trabajado	121806	6.06	868898	43.24
Jubilado	226125	11.25	1095023	54.50
Otros pensionistas	88889	4.42	1183912	58.92
Incapacitado permanente	9854	0.49	1193766	59.41
Estudiante	444179	22.11	1637945	81.52
Labores del hogar	328244	16.34	1966189	97.85
Otra situación	43173	2.15	2009362	100.00

Frequency Missing = 248612

Una vez conocida la distribución de frecuencias previa a la imputación se puede aplicar el contraste de bondad de ajuste de la chi-cuadrado a las distribuciones de frecuencias obtenidas tras aplicar los distintos métodos de imputación propuestos. En la siguiente tabla aparece la imputación que se realizó en el censo de 1.996. Consistió en imputar según una asignación aleatoria en base a la distribución obtenida para subgrupos de población combinando las dos categorías de sexo, siete grupos de año de nacimiento y dos de tipo de municipio (agrícola o no agrícola).

En la tabla se observa cómo aparecen pequeñas diferencias entre las categorías de la variable relación con la actividad. Esto provoca que al aplicar el test de bondad de ajuste de la chi-cuadrado nos indica que se rechaza H_0 ya que el estadístico tiene un valor de 931,2131 y un p-valor inferior a 0,0001, es decir, no se puede aceptar que siga la misma distribución que antes de imputar.

relades	Frequency	Percent	Test Percent	Cumulative Frequency	Cumulative Percent
Servicio Militar	5375	0.24	0.24	5375	0.24
Ocupado	749058	33.20	32.89	754433	33.44
Parado. 1º empleo	90375	4.01	4.05	844808	37.45
Parado ya ha trabajado	137961	6.12	6.06	982769	43.56
Jubilado	254576	11.28	11.25	1237345	54.85
Otros pensionistas	98822	4.38	4.42	1336167	59.23
Incapacitado permanente	10833	0.48	0.49	1347000	59.71
Estudiante	499337	22.13	22.11	1846337	81.84
Labores del hogar	356558	15.81	16.34	2202895	97.65
Otra situación	53049	2.35	2.15	2255944	100.00

Frequency Missing = 2030

Chi-Square Test
for Specified Proportions

Chi-Square	931.2131
DF	9
Pr > ChiSq	<.0001

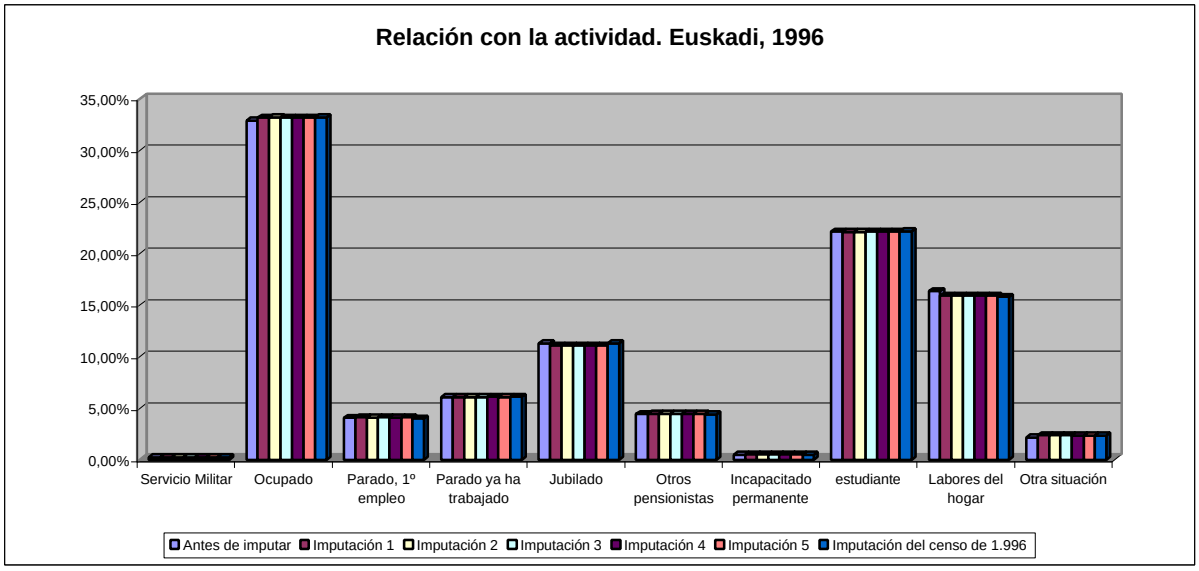
Effective Sample Size = 2255944
Frequency Missing = 2030

De forma similar se realizaron las distribuciones de frecuencias de los distintos conjuntos de datos imputados al seleccionar aleatoriamente según la distribución del

nodo terminal asignado mediante el árbol de clasificación donde participan como variables explicativas las comentadas anteriormente. Se han realizado en total cinco conjuntos de datos que posteriormente serán combinados. Al realizar el contraste chi-cuadrado de bondad de ajuste se obtienen valores del estadístico en torno a 900 y le corresponde un p-valor asociado menor a 0,0001 en todos los conjuntos de datos. Esto nos obliga a rechazar la hipótesis nula de conservación de la distribución. Las distribuciones de frecuencias obtenidas tras la imputación de los distintos conjuntos de datos se pueden consultar en el Anexo IV, tabla 1.

Relación con la actividad	*Estadístico Chi-cuadrado*	*P-Valor*
Imputación 1.996	931,2131	<0,0001
Imputación según selección aleatoria		
Imputación 1	925,2147	<0,0001
Imputación 2	982,8033	<0,0001
Imputación 3	945,9616	<0,0001
Imputación 4	886,0861	<0,0001
Imputación 5	912,8497	<0,0001
Imputación a la categoría más probable	8.715,4240	<0,0001

De forma más directa se observan las variaciones con respecto a la distribución real en el gráfico que representan las distintas distribuciones de frecuencias totales obtenidas para los distintos conjuntos de datos, junto con la distribución de frecuencias antes de imputar y la realizada en 1996.



Tras realizar la combinación de los cinco ficheros de datos completos imputados, según el desarrollo propuesto por Rubin en 1987, se obtienen los siguientes resultados: Una tabla en la que aparece para cada categoría la estimación de la proporción (en tanto por

1), el error estándar asociado, teniendo en cuenta la varianza dentro de cada conjunto de datos y entre los distintos conjuntos de datos. Se realiza un contraste para estudiar si se acepta que el parámetro tiene valor nulo, en este caso no se acepta esta hipótesis en ninguna categoría. La última columna que se proporciona indica la proporción de pérdida de información del parámetro debido a la falta de respuesta.

	ESTIMATE	STD.ERR	T-RATIO	DF	P-VAL.	%MISS.INF.
SEVICIO_MILITAR	0.00239	0.00004	64.0684	68.6433	0	26.2574
OCUPADO	0.33168	0.00035	955.325	116.158	0	19.9239
PARADO_1º_EMPLEO	0.04097	0.00017	236.365	22.5826	0	46.614
PARADO_YA_HA_TRABAJADO	0.06079	0.0002	303.806	29.4648	0	40.7356
JUBILADO	0.11059	0.00021	517.579	1901.51	0	4.68668
OTROS_PENSIONISTAS	0.04457	0.00017	267.601	39.0067	0	35.2594
INCAPACITADO_PERMANENTE	0.00512	0.00005	98.9719	161.77	0	16.7476
ESTUDANTE	0.22102	0.00028	786.368	3305.76	0	3.53686
LABORES_DEL_HOGAR	0.15917	0.00025	630.875	844.072	0	7.10384
OTRA_SITUACIÓN	0.02369	0.00013	188.431	32.278	0	38.8763

Además de la imputación múltiple anterior se ha imputado a la categoría más probable dentro del nodo terminal asignado. De esta forma se obtienen las mayores diferencias en la distribución de frecuencias si se compara con la obtenida antes de realizar la imputación. Esto se puede comprobar al observar el valor del estadístico de bondad de ajuste de la chi-cuadrado que tiene un valor de 8.715,4240, muy superior al obtenido al aplicar los métodos de imputación anteriores.

into_	Frequency	Percent	Test Percent	Cumulative Frequency	Cumulative Percent
Servicio Militar	4910	0.22	0.24	4910	0.22
Ocupado	784147	34.76	32.89	789057	34.98
Parado. 1º empleo	81435	3.61	4.05	870492	38.59
Parado ya ha trabajado	121879	5.40	6.06	992371	43.99
Jubilado	245147	10.87	11.25	1237518	54.86
Otros pensionistas	107991	4.79	4.42	1345509	59.64
Incapacitado permanente	9916	0.44	0.49	1355425	60.08
Estudiante	512370	22.71	22.11	1867795	82.80
Labores del hogar	344934	15.29	16.34	2212729	98.09
Otra situación	43190	1.91	2.15	2255919	100.00

Frequency Missing = 2055

Chi-Square Test
for Specified Proportions

Chi-Square	8715.4240
DF	9
Pr > ChiSq	<.0001

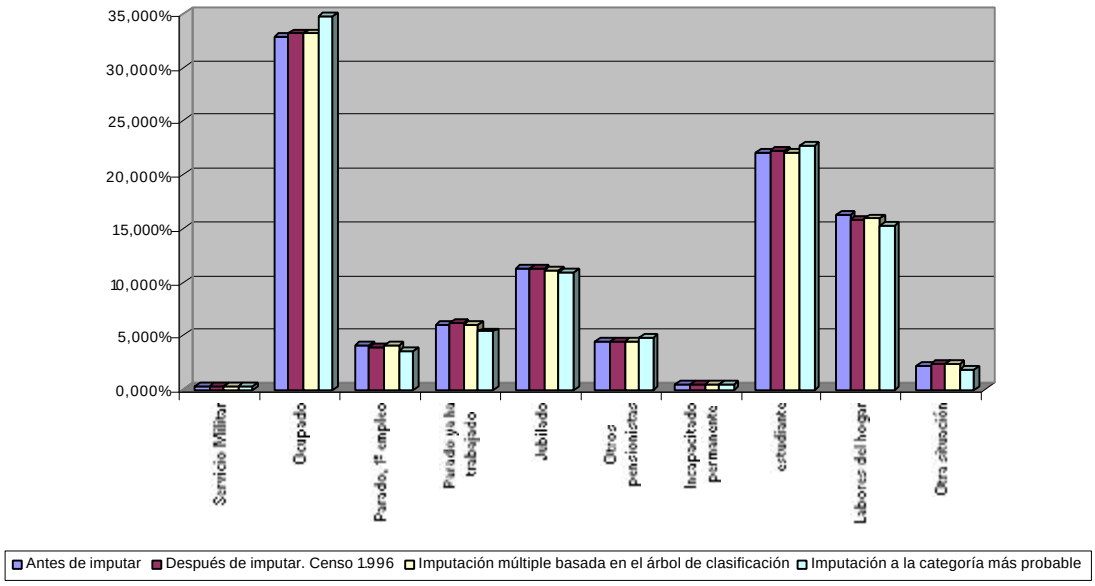
Effective Sample Size = 2255919
Frequency Missing = 2055

En la siguiente tabla se comparan los porcentajes de las distintas categorías que se obtiene antes de realizar la imputación y tras aplicar los distintos métodos de imputación propuestos: imputación realizada en 1.996, imputación múltiple basada en la distribución de frecuencias del nodo terminal asignado e imputación a la categoría más probable del nodo terminal. Junto con el porcentaje se incluye la diferencia existente entre el porcentaje obtenido antes de la imputación y el posterior.

Relación con la actividad	Antes de imputar	Después de imputar. Censo 1.996		Después de imputar. Imputación múltiple		Imputación a la categoría más probable	
		Porcentaje	Diferencia	Porcentaje	Diferencia	Porcentaje	Diferencia
Servicio Militar	0,244%	0,238%	0,006%	0,239%	0,005%	0,218%	0,027%
Ocupado	32,887%	33,204%	-0,317%	33,168%	-0,281%	34,760%	-1,872%
Parado, 1º empleo	4,049%	4,006%	0,043%	4,097%	-0,048%	3,610%	0,439%
Parado ya ha trabajado	6,062%	6,115%	-0,054%	6,079%	-0,017%	5,403%	0,659%
Jubilado	11,254%	11,285%	-0,031%	11,059%	0,195%	10,867%	0,387%
Otros pensionistas	4,424%	4,381%	0,043%	4,457%	-0,033%	4,787%	-0,363%
Incapacitado permanente	0,490%	0,480%	0,010%	0,512%	-0,022%	0,440%	0,051%
Estudiante	22,105%	22,134%	-0,029%	22,102%	0,003%	22,712%	-0,607%
Labores del hogar	16,336%	15,805%	0,530%	15,917%	0,419%	15,290%	1,046%
Otra situación	2,149%	2,352%	-0,203%	2,369%	-0,220%	1,915%	0,234%

A continuación se incluye un gráfico en el cual se comparan los resultados obtenidos al aplicar los distintos métodos de imputación. Se observa cómo se modifica en mayor medida la distribución al emplear el método de imputación a la categoría más probable.

Relación con la actividad. Euskadi 1996



Una vez obtenidos los porcentajes por categorías y el error estándar cometido al realizar la imputación múltiple se procede a la realización de los intervalos de confianza al 95% junto con la distribución de frecuencias antes de imputar y tras la imputación realizada en 1.996 para comprobar las diferencias existentes. Se observa que las proporciones antes de imputar de las categorías servicio militar, parados que ya ha trabajado, otros pensionistas y estudiantes aparecen dentro de los intervalos de confianza realizados. Mientras que si se compara con la imputación realizada en 1996 se incluyen dentro del intervalo las categorías servicio militar, ocupado, parado que ya ha trabajado, estudiante, labores del hogar y otra situación.

Relación con la actividad	Antes de imputar	Imputación múltiple				Después de imputa, censo 1.996
	Porcentaje	Porcentaje	Error	intervalo de confianza al 95%		Porcentaje
				extremo inferior	extremo superior	
Servicio Militar	0,244%	0,239%	0,004%	0,231%	0,247%	0,238%
Ocupado	32,887%	33,168%	0,035%	33,099%	33,237%	33,204%
Parado, 1º empleo	4,049%	4,097%	0,017%	4,064%	4,130%	4,006%
Parado ya ha trabajado	6,062%	6,079%	0,020%	6,040%	6,118%	6,115%
Jubilado	11,254%	11,059%	0,021%	11,018%	11,100%	11,285%
Otros pensionistas	4,424%	4,457%	0,017%	4,424%	4,490%	4,381%
Incapacitado permanente	0,490%	0,512%	0,005%	0,502%	0,522%	0,480%
Estudiante	22,105%	22,102%	0,028%	22,047%	22,157%	22,134%
Labores del hogar	16,336%	15,917%	0,025%	15,868%	15,966%	15,805%
Otra situación	2,149%	2,369%	0,013%	2,344%	2,394%	2,352%

Capítulo

7

Conclusiones

La conclusión principal que se obtiene de dicho método de imputación basado en árboles de clasificación es que estamos ante una buena técnica por diversos motivos: se crean clases de imputación lo más homogéneas posibles con respecto a la variable a imputar y basadas en otras variables que intervienen en el estudio ó información auxiliar, además de obtener valores imputados que no provocarán inconsistencias con respecto a las reglas de edición construidas.

Con respecto a los métodos de imputación aplicados una vez clasificados los elementos mediante el árbol de clasificación indicar que el método aleatorio según la distribución de frecuencias dentro de dicho nodo conserva en mayor medida la distribución inicial que el método que asigna a la categoría con mayor probabilidad dentro del nodo. Si bien este último proporciona mejores resultados con respecto a la correcta imputación de los registros. Por el contrario en aquellas categorías con escaso peso con respecto al conjunto de la población y formado por grupos heterogéneos puede no existir ningún nodo terminal con dicha categoría como la de mayor probabilidad. Por tanto en esta situación no será imputado ningún registro a dicha categoría.

Capítulo

8

Bibliografía

[1] ACOCK, ALAN C.

Working with Missing Data

Oregon State *University*.

[2] AUTIMP: http://www.cbs.nl/en/services/autimp/autimp.htm

[3] BARCENA, J.M.

Técnicas Multivariantes para *el Enlace de Encuestas*

Tesis doctoral, UPV *2001*.

[4] BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A., STONE, C.J.

Classification an Regression Trees

Wadsorth *International* Group, 1984

[5] BRUCE RATNER, PH.

Chaid as a Method for Filling in Missing Values

D. DM STAT-1 CONSULTING.

[6] CALVO, PATRICIA

Segmentación por Árbol Binario

EUSTAT 1997.

[7] *Classification and* Regression *Trees*

Manual S-PLUS.

[8] Congresos de Edición e Imputación:

http://www.unece.org/stats/documents/1999.06.sde.htm

http://www.unece.org/stats/documents/2000.10.sde.htm

[9] *Curso de Depuración* e Imputación *de la Escuela de Estadística*

9-11 de Mayo de 2000.

[10] DE WAAL, TON, PLOMP, R.

Manual WAID (version 4.0).

Statistics Netherlands.

[11] EUREDIT: http://www.cs.york.ac.uk/euredit

[12] *European Community* Household Panel *(*ECHP*) Imputation Wave.*

EUROSTAT 1995.

[13] GUEGUEN, A., NAKACHE, J.P., NICOLAU-MOLINA, J.

SPAD·S. Segmentation par Arbre de décision binaire Discrimination et Régression. Manual de Reference

Inserm. CISIA 1996.

[14] HOOGLAND, J., PANNEKOEK, J.

Evaluation of SPSS *Missing Value Analysis 7.5*

Statistics *Netherlands*.

[15] *Imputa. Versión 2.0.*

Eustat

[16*] Imputación múltiple: http://www.multiple-imputation.com*

[17] *Interm Report on* Evaluation *Criteria for Statistical Editing and Imputation*

Proyecto EUREDIT.

[18] KASS, G.V.

An Exploratory Technique for Investigating Large Quantities of Categorical Data.

Applied Statistics, 1980

[19] LAAKSONEN, SEPPO

"How to Find de Best Imputation Technique? Tests with Various Methods."

Statistics Finland.

[20] *Manual de la Macro* TREEDISC*.*

[21] MESA, D.M., TSAI, P., CHAMBERS, R.L.,

Using Tree-based models for Missing Data Imputation: An Evaluation using UK Census Data

Proyecto europeo AUTIMP, Universidad de Southampton 2000

[22] *Padrón Municipal* de Habitantes y *Censos de Población y Viviendas 1991*

EUSTAT

[23] PLATEK, R.

"Metodología y tratamiento de la no-respuesta".

Seminario Internacional de Estadística en Euskadi. 1986.

[24] SCHAFER CHAPMAN&HALL, J.L.

Analysis of Incomplete Multivariate Data

CRL 1997.

http://stat.psu.edu/~jls

[25] SPSS: AnswerTree Algorithm Summary

[26] STEEN LARSEN, B., MADSEN, B.

"Evaluation of SOLAS 2.0" for Imputing *missing values*

Statistics Denmark.

[27] VILLAN, CRIADO, BRAVO CABRIA, M.S.

"Procedimientos de depuración de datos estadísticos".

Seminario Internacional de Estadística en Euskadi .*1990*. I

ANEXO I

EL ALGORITMO CHAID

Pasos del **Algoritmo CHAID** en el cual se desea clasificar la variable Y y se tiene como variables explicativas $X_1, X_2, ..., X_k$:

1. Calcular la distribución de la variable respuesta Y en el nodo raíz.

2. Para cada variable explicativa X_i $(i = 1,...,k)$, hay que encontrar el par de categorías que tienen menores diferencias significativas respecto a la distribución de Y dentro del nodo. Es decir, aquel que tiene el mayor p -valor. El método emplea para calcular dicho p -valor depende del tipo de variables que estemos tratando en cada momento. Vamos a considerar que estamos tratando en nuestro estudio variables categóricas.

   i. La relación entre la variable explicativa X_1 y la variable respuesta Y dentro del nodo se representa mediante una tabla de contingencia. Se consideran todas las sub-tablas de contingencia posibles que se puedan formar con dos categorías de la variable explicativa.

   ii. El algoritmo identifica el par de categorías de X_1 con mayor p -valor (p_1) asociado y lo compara con el nivel α predeterminado, normalmente $\alpha_{unión}$ =0,05. Si el valor p -valor p_1 es mayor que este valor $\alpha_{unión}$ se agrupan dichas categorías. Se repite el apartado i) considerando el par de categorías agrupadas como una única para calcular las sub-tablas de contingencia. En el caso de no obtener superar el valor de $\alpha_{unión}$ no se realiza ninguna agrupación de las categorías y se pasa al apartado 3.

   iii. De nuevo se selecciona el par de categorías con mayor p -valor y se compara con el valor $\alpha_{unión}$. Si es superior se vuelve a agrupar y se vuelven a calcular las sub-tablas de contingencia. El proceso termina en el caso en el cual el p -valor es inferior a $\alpha_{unión}$ o se llega a dos categorías.

   iv. El algoritmo calcula un ajustado p -valor empleando las categorías agrupadas obtenidas de X_1 y la categoría Y usando el ajuste de Bonferroni.

3. Los pasos i) a iv) se repiten de nuevo con el resto de variables explicativas.

4. El paso final es dividir el nodo basado en la variable explicativa, con las categorías agrupadas, con menor p-valor ajustado si el valor es menor al prefijado $\alpha_{separacion}$. En el caso de obtener un valor superior dicho nodo no se ramifica y será un nodo terminal.

5. Se continua ramificando el árbol hasta que se satisfaga el criterio de parada.

ANEXO II

Tabla de contingencia y medidas de asociación entre las variables relación con la actividad (RELA1) y estado civil (ECIV).

ECIV	RELA1											
Frequency Percent Row Pct Col Pct	.,Militar,	Servicio,1º emple,	Ocupado ,Parado. ,a ha bajado ,	Parado y tra,	Jubilado,nsionist,	Otros pe,tado per,	Incapaci,te manente ,	Estudian,r ,	Labores ,del hoga,	Otra sit,uación ,	Total
	14186	0	28	1	2	15	3	2	9	7	0	.
	.	.	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.	.	.
Casado/a	10529	153	45407	581	5243	13676	879	357	219	26980	309	93804
	.	0.08	23.08	0.30	2.67	6.95	0.45	0.18	0.11	13.72	0.16	47.68
	.	0.16	48.41	0.62	5.59	14.58	0.94	0.38	0.23	28.76	0.33	
	.	17.29	62.96	10.63	45.67	68.29	15.82	39.71	0.48	89.32	6.66	
Divorciado/a	184	0	798	21	213	94	31	8	7	112	10	1294
	.	0.00	0.41	0.01	0.11	0.05	0.02	0.00	0.00	0.06	0.01	0.66
	.	0.00	61.67	1.62	16.46	7.26	2.40	0.62	0.54	8.66	0.77	
	.	0.00	1.11	0.38	1.86	0.47	0.56	0.89	0.02	0.37	0.22	
Separado/a	470	5	1267	52	371	196	56	28	13	268	35	2291
	.	0.00	0.64	0.03	0.19	0.10	0.03	0.01	0.01	0.14	0.02	1.16
	.	0.22	55.30	2.27	16.19	8.56	2.44	1.22	0.57	11.70	1.53	
	.	0.56	1.76	0.95	3.23	0.98	1.01	3.11	0.03	0.89	0.75	
Soltero/a	15820	721	23812	4790	5512	2643	552	456	45190	964	4206	88846
	.	0.37	12.10	2.43	2.80	1.34	0.28	0.23	22.97	0.49	2.14	45.16
	.	0.81	26.80	5.39	6.20	2.97	0.62	0.51	50.86	1.09	4.73	
	.	81.47	33.02	87.65	48.01	13.20	9.94	50.72	99.46	3.19	90.63	
Viudo/a	1522	6	837	21	142	3416	4038	50	8	1882	81	10481
	.	0.00	0.43	0.01	0.07	1.74	2.05	0.03	0.00	0.96	0.04	5.33
	.	0.06	7.99	0.20	1.35	32.59	38.53	0.48	0.08	17.96	0.77	
	.	0.68	1.16	0.38	1.24	17.06	72.68	5.56	0.02	6.23	1.75	
Total	.	885	72121	5465	11481	20025	5556	899	45437	30206	4641	196716
	.	0.45	36.66	2.78	5.84	10.18	2.82	0.46	23.10	15.36	2.36	100.00

Frequency Missing = 42778

Statistics for Table of ECIV by RELA1

Statistic	DF	Value	Prob
Chi-Square	36	156632	<.0001
Likelihood Ratio Chi-Square	36	146820	<.0001
Mantel-Haenszel Chi-Square	1	7407	<.0001
Phi Coefficient		0.89232	
Contingency Coefficient		0.66579	
Cramer's V		0.44616	

Statistic	Value	ASE
Gamma	0.1606	0.0027
Kendall's Tau-b	0.1182	0.0020
Stuart's Tau-c	0.0977	0.0016
Somers' D C|R	0.1381	0.0023
Somers' D R|C	0.1011	0.0017
Pearson Correlation	0.1940	0.0022
Spearman Correlation	0.1431	0.0023
Lambda Asymmetric C|R	0.1973	0.0020
Lambda Asymmetric R|C	0.5555	0.0017
Lambda Symmetric	0.3593	0.0017
Uncertainty Coefficient C|R	0.2157	0.0008
Uncertainty Coefficient R|C	0.3915	0.0015
Uncertainty Coefficient Symmetric	0.2781	0.0010

Effective Sample Size = 196716
Frequency Missing = 42778

WARNING: 18% of the data are missing.

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	7407	<.0001
2	Row Mean Scores Differ	4	8627	<.0001
3	General Association	36	156631	<.0001

Effective Sample Size = 196716
Frequency Missing = 42778

ANEXO III

Tabla 1. Imputación basada en la distribución de frecuencias del nodo terminal

Tabla de contingencia y medidas de asociación entre los valores reales e imputados de la variable relación con la actividad (RELA1).

RELA1	rela1arb										
Frequency Percent Row Pct Col Pct	. Militar	Servicio, 1º emple	Ocupado o	Parado. bajado	Parado y a ha tra	Jubilado	Otros pe nsionist	Incapaci tado per	Estudian te	Labores del hoga	Otra sit uación	Total
Servicio Militar	0 . 0.00 .	3 0.02 5.77 4.05	14 0.09 26.92 0.23	5 0.03 9.62 1.12	4 0.02 7.69 0.39	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	26 0.16 50.00 0.69	0 0.00 0.00 0.00	0 0.00 0.00	52 0.32
Ocupado	2 . 0.15 0.42 33.78	25 24.33 1.05 65.98 66.00	3947 1.05 3.38 2.84 38.20	170 3.38 1.36 9.18 53.46	549 1.36 0.28 3.69 13.74	221 0.28 0.24 0.75 10.37	45 0.24 2.47 0.65 55.71	39 2.47 3.45 6.69 10.58	400 3.45 0.16 9.36 22.79	560 0.16 36.87 0.43 7.49	26 5982	
Parado. 1º emple o	1 . 0.01 0.44 2.70	2 0.01 38.02 2.89	173 1.07 9.01 9.21	41 0.25 7.69 3.41	35 0.22 0.44 0.12	2 0.01 0.44 0.46	2 0.01 0.66 4.29	3 0.02 38.02 4.58	173 1.07 4.84 0.90	22 0.14 0.44 0.58	2 455	
Parado ya ha tra bajado	2 . 0.03 0.51 6.76	5 3.15 52.57 8.55	511 0.29 4.84 10.56	47 0.75 12.55 11.88	122 0.32 5.35 3.23	52 0.05 0.82 1.84	8 0.02 0.41 5.71	4 0.66 11.01 2.83	107 0.70 11.63 4.60	113 0.02 0.31 0.86	3 972	
Jubilado	1 . 0.00 0.00 0.00	0 1.49 14.80 4.05	242 0.00 0.00 0.00	0 0.31 3.06 4.87	50 6.56 65.14 66.19	1065 0.73 7.22 27.19	118 0.04 0.43 10.00	7 0.01 0.12 0.05	2 0.84 8.38 5.58	137 0.09 0.86 4.03	14 1635	
Otros pensionist as	1 . 0.00 0.00 0.00	0 0.29 10.68 0.79	47 0.00 0.00 0.00	0 0.07 2.73 1.17	12 0.72 26.59 7.27	117 1.05 38.86 39.40	171 0.01 0.45 2.86	2 0.04 1.36 0.16	6 0.49 17.95 3.22	79 0.04 1.36 1.73	6 440	
Incapacitado per manente	1 . 0.00 0.00 0.00	0 0.25 50.62 0.69	41 0.01 1.23 0.22	1 0.04 8.64 0.68	7 0.07 14.81 0.75	12 0.02 4.94 0.92	4 0.02 3.70 4.29	3 0.03 6.17 0.13	5 0.04 7.41 0.24	6 0.01 2.47 0.58	2 81	
Estudiante	1 . 0.23 1.01 51.35	38 2.39 10.24 6.47	387 0.97 4.15 35.28	157 0.73 3.15 11.59	119 0.01 0.03 0.06	1 0.01 0.05 0.46	2 0.01 0.11 5.71	4 0.02 73.89 73.87	2793 17.22 0.53 0.81	20 0.12 6.85 74.64	259 1.60 23.30	3780
Labores del hoga r	1 . 0.00 0.00 0.00	0 3.67 23.87 9.95	595 0.13 0.84 4.72	21 0.75 4.89 11.88	122 0.80 5.21 8.08	130 0.49 3.17 18.20	79 0.05 0.32 11.43	8 0.12 0.76 0.50	19 9.33 60.73 61.62	1514 0.03 0.20 1.44	5 15.37	2493
Otra situación	0 . 0.01 0.30 1.35	1 0.14 6.89 0.38	23 0.02 0.90 0.67	3 0.04 2.10 0.68	7 0.06 2.69 0.56	9 0.03 1.50 1.15	5 0.00 0.00 0.00	0 1.54 74.85 6.61	250 0.04 1.80 0.24	6 0.18 8.98 8.65	30 2.06	334
Total	74 0.46	5980 36.86	445 2.74	1027 6.33	1609 9.92	434 2.68	70 0.43	3781 23.30	2457 15.14	347 2.14	16224 100.00	

Frequency Missing = 10

The FREQ Procedure

Statistics for Table of RELA1 by rela1arb

Statistic	DF	Value	Prob
Chi-Square	81	22874.6068	<.0001
Likelihood Ratio Chi-Square	81	17671.8155	<.0001
Mantel-Haenszel Chi-Square	1	4459.9121	<.0001
Phi Coefficient		1.1874	
Contingency Coefficient		0.7649	
Cramer's V		0.3958	

Statistic	Value	ASE	95% Confidence Limits	
Gamma	0.5243	0.0070	0.5106	0.5379
Kendall's Tau-b	0.4430	0.0063	0.4307	0.4554
Stuart's Tau-c	0.3794	0.0054	0.3689	0.3899
Somers' D C|R	0.4432	0.0063	0.4308	0.4556
Somers' D R|C	0.4428	0.0063	0.4304	0.4552
Pearson Correlation	0.5243	0.0068	0.5109	0.5377
Spearman Correlation	0.5020	0.0071	0.4881	0.5159
Lambda Asymmetric C|R	0.4404	0.0062	0.4283	0.4524
Lambda Asymmetric R|C	0.4455	0.0061	0.4335	0.4575
Lambda Symmetric	0.4429	0.0058	0.4315	0.4543
Uncertainty Coefficient C|R	0.3161	0.0041	0.3081	0.3241
Uncertainty Coefficient R|C	0.3170	0.0041	0.3090	0.3251
Uncertainty Coefficient Symmetric	0.3166	0.0040	0.3087	0.3245

Test of Symmetry

Statistic (S)	31.9912
DF	45
Pr > S	0.9277

The FREQ Procedure

Statistics for Table of RELA1 by rela1arb

Simple Kappa Coefficient

Kappa	0.4774
ASE	0.0048
95% Lower Conf Limit	0.4680
95% Upper Conf Limit	0.4868

Test of H0: Kappa = 0

ASE under H0	0.0039
Z	122.4109
One-sided Pr > Z	<.0001
Two-sided Pr > |Z|	<.0001

Tabla 2. Imputación basada en la categoría más probable del nodo terminal

Tabla de contingencia y medidas de asociación entre los valores reales e imputados de la variable relación con la actividad (RELA1).

The FREQ Procedure

Table of RELA1 by maxpr

RELA1 maxpr

Frequency / Percent / Row Pct / Col Pct	.,Ocupado	Parado. 1º emple,o	Parado y,a ha tra,bajado	Jubilado,Otros pe,nsionist,as	Incapaci,tado per,manente	Estudian,te	Labores ,del hoga,r	Otra sit,uación	Total
Servicio Militar	0	9	0	0	0	0	0	42	1	0	52
	.	0.06	0.00	0.00	0.00	0.00	0.00	0.26	0.01	0.00	0.32
	.	17.31	0.00	0.00	0.00	0.00	0.00	80.77	1.92	0.00	
	.	0.15	0.00	0.00	0.00	0.00	0.00	0.83	0.04	0.00	
Ocupado	17	4514	0	4	227	30	1	593	595	3	5967
	.	27.91	0.00	0.02	1.40	0.19	0.01	3.67	3.68	0.02	36.89
	.	75.65	0.00	0.07	3.80	0.50	0.02	9.94	9.97	0.05	
	.	74.29	0.00	30.77	14.08	5.00	33.33	11.71	21.33	21.43	
Parado. 1º emple o	2	149	1	0	1	1	0	285	17	0	454
	.	0.92	0.01	0.00	0.01	0.01	0.00	1.76	0.11	0.00	2.81
	.	32.82	0.22	0.00	0.22	0.22	0.00	62.78	3.74	0.00	
	.	2.45	33.33	0.00	0.06	0.17	0.00	5.63	0.61	0.00	
Parado ya ha tra bajado	6	600	1	2	84	2	0	167	112	0	968
	.	3.71	0.01	0.01	0.52	0.01	0.00	1.03	0.69	0.00	5.98
	.	61.98	0.10	0.21	8.68	0.21	0.00	17.25	11.57	0.00	
	.	9.87	33.33	15.38	5.21	0.33	0.00	3.30	4.02	0.00	
Jubilado	11	173	0	3	1177	149	1	0	117	5	1625
	.	1.07	0.00	0.02	7.28	0.92	0.01	0.00	0.72	0.03	10.05
	.	10.65	0.00	0.18	72.43	9.17	0.06	0.00	7.20	0.31	
	.	2.85	0.00	23.08	73.01	24.83	33.33	0.00	4.20	35.71	
Otros pensionist as	6	34	0	1	67	293	0	10	29	1	435
	.	0.21	0.00	0.01	0.41	1.81	0.00	0.06	0.18	0.01	2.69
	.	7.82	0.00	0.23	15.40	67.36	0.00	2.30	6.67	0.23	
	.	0.56	0.00	7.69	4.16	48.83	0.00	0.20	1.04	7.14	
Incapacitado per manente	3	47	0	0	15	3	0	7	7	0	79
	.	0.29	0.00	0.00	0.09	0.02	0.00	0.04	0.04	0.00	0.49
	.	59.49	0.00	0.00	18.99	3.80	0.00	8.86	8.86	0.00	
	.	0.77	0.00	0.00	0.93	0.50	0.00	0.14	0.25	0.00	
Estudiante	3	96	0	1	1	0	1	3671	8	0	3778
	.	0.59	0.00	0.01	0.01	0.00	0.01	22.70	0.05	0.00	23.36
	.	2.54	0.00	0.03	0.03	0.00	0.03	97.17	0.21	0.00	
	.	1.58	0.00	7.69	0.06	0.00	33.33	72.49	0.29	0.00	
Labores del hoga r	10	427	1	2	29	117	0	12	1896	0	2484
	.	2.64	0.01	0.01	0.18	0.72	0.00	0.07	11.72	0.00	15.36
	.	17.19	0.04	0.08	1.17	4.71	0.00	0.48	76.33	0.00	
	.	7.03	33.33	15.38	1.80	19.50	0.00	0.24	67.98	0.00	
Otra situación	2	27	0	0	11	5	0	277	7	5	332
	.	0.17	0.00	0.00	0.07	0.03	0.00	1.71	0.04	0.03	2.05
	.	8.13	0.00	0.00	3.31	1.51	0.00	83.43	2.11	1.51	
	.	0.44	0.00	0.00	0.68	0.83	0.00	5.47	0.25	35.71	
Total	.	6076	3	13	1612	600	3	5064	2789	14	16174
	.	37.57	0.02	0.08	9.97	3.71	0.02	31.31	17.24	0.09	100.00

Frequency Missing = 60

The FREQ Procedure

Statistics for Table of RELA1 by maxpr

Statistic	DF	Value	Prob
Chi-Square	72	31206.4956	<.0001
Likelihood Ratio Chi-Square	72	23234.2136	<.0001
Mantel-Haenszel Chi-Square	1	6351.8127	<.0001
Phi Coefficient		1.3890	
Contingency Coefficient		0.8116	
Cramer's V		0.4911	

WARNING: 46% of the cells have expected counts less
than 5. Chi-Square may not be a valid test.

Statistic	Value	ASE	95% Confidence Limits	
Gamma	0.6477	0.0066	0.6348	0.6606
Kendall's Tau-b	0.5611	0.0064	0.5485	0.5736
Stuart's Tau-c	0.4699	0.0054	0.4594	0.4804
Somers' D C|R	0.5424	0.0062	0.5302	0.5547
Somers' D R|C	0.5803	0.0066	0.5673	0.5934
Pearson Correlation	0.6267	0.0063	0.6144	0.6390
Spearman Correlation	0.6016	0.0069	0.5881	0.6151
Lambda Asymmetric C|R	0.6661	0.0053	0.6557	0.6764
Lambda Asymmetric R|C	0.5482	0.0061	0.5363	0.5600
Lambda Symmetric	0.6068	0.0054	0.5961	0.6175
Uncertainty Coefficient C|R	0.5125	0.0051	0.5025	0.5225
Uncertainty Coefficient R|C	0.4185	0.0044	0.4099	0.4270
Uncertainty Coefficient Symmetric	0.4607	0.0046	0.4517	0.4698

Effective Sample Size = 16174
Frequency Missing = 60

ANEXO IV

Tabla I

Distribución de frecuencias de los distintos conjuntos de datos al imputar relación con la actividad mediante una imputación aleatoria según la distribución de frecuencias del nodo terminal.

Relación con la actividad	*Antes de imputar*		*Imputación 1*		*Imputación 2*		*Imputación 3*		*Imputación 4*		*Imputación 5*		*Imputación del censo de 1.996*	
	Freq.	**%**	**Freq.**	**%**	**Freq.**	**%**	**Freq.**	**%**	**Freq.**	**%**	**Freq.**	**%**	**Freq.**	**%**
Servicio Militar	4.910	0,24%	5.363	0,24%	5.408	0,24%	5.459	0,24%	5.383	0,24%	5.377	0,24%	5.375	0,24%
Ocupado	660.822	32,89%	749.054	33,17%	749.357	33,19%	748.591	33,15%	748.918	33,17%	748.669	33,16%	749.058	33,20%
Parado, 1º empleo	81.360	4,05%	92.556	4,10%	92.231	4,08%	92.667	4,10%	92.334	4,09%	92.792	4,11%	90.375	4,01%
Parado ya ha trabajado	121.806	6,06%	137.315	6,08%	137.422	6,09%	136.850	6,06%	137.489	6,09%	137.230	6,08%	137.961	6,12%
Jubilado	226.125	11,25%	249.734	11,06%	249.672	11,06%	249.744	11,06%	249.584	11,05%	249.839	11,06%	254.576	11,28%
Otros pensionistas	88.889	4,42%	100.527	4,45%	100.461	4,45%	100.916	4,47%	100.790	4,46%	100.551	4,45%	98.822	4,38%
Incapacitado permanente	9.854	0,49%	11.528	0,51%	11.628	0,51%	11.528	0,51%	11.552	0,51%	11.582	0,51%	10.833	0,48%
estudiante	444.179	22,11%	498.968	22,10%	498.935	22,10%	499.098	22,10%	499.103	22,10%	499.199	22,11%	499.337	22,13%
Labores del hogar	328.244	16,34%	359.437	15,92%	359.190	15,91%	359.496	15,92%	359.523	15,92%	359.331	15,91%	356.558	15,81%
Otra situación	43.173	2,15%	53.492	2,37%	53.670	2,38%	53.625	2,37%	53.298	2,36%	53.404	2,37%	53.049	2,35%
	Missing=248.612												Missing=2.030	
Total	2.009.362		2.257.974		2.257.974		2.257.974		2.257.974		2.257.974		2.255.944	