Señor… si me das fortuna,
no me quites la razón

Capítulo 3
- Propuesta -

3.1 Introducción

La Fusión tiene como objetivo combinar datos de diferentes fuentes para poder disponer de toda la información en un solo archivo. Esto se consigue imputando a unos individuos la información proveniente de otros individuos con los cuales comparten aspectos en común que se relacionan con la información que se quiere estimar.

Como ya se ha mencionado, el interés por la fusión se debe a ciertos casos en los cuales es imposible disponer de la información específica para un solo archivo o a la reducción de costos entre otras causas.

La fusión integra muestras (encuestas) emparejando respuestas individuales. Dos muestras de una misma población (por ejemplo, uso de medios y uso de productos) seleccionarán diferentes individuos para representar esa población. La fusión intenta conectar un individuo dado en una muestra a un(os) individuo(s) en la otra muestra cuyas características sean lo más parecidas posible sobre un conjunto preseleccionado de variables. Se define una medida de distancia para establecer la semejanza. Cuanto menor sea la distancia, mejor es el parecido entre individuos. Una vez seleccionados los individuos que aportaran la información requerida, esta se puede imputar de diferentes formas.

La forma de imputación se puede hacer de manera determinística o estocástica. Los distintos métodos de imputación se pueden aplicar para imputar un valor para cada valor ausente (imputación simple) o en algunos casos, imputar más de un valor para permitir una evaluación apropiada de la incertidumbre de la imputación (imputación múltiple)[1] en la cual se utilizan m ($3 \leq m \leq 10$) valores imputados (cf. 1.2.2.2).

[1] Little, R.J.; Rubin, D.B. (2002)

La imputación conlleva variación adicional que debe ser estimada. Con frecuencia se emplea la imputación múltiple para medir esta variación adicional, aunque existen otros métodos[2].

Tiene que existir evidencia de que el método empleado es lo suficientemente preciso para el objetivo deseado. En un nivel agregado, interesará saber si:

- Se han preservado las distribuciones marginales
- Se ha conservado la estructura correlacional de los distintos pares de variables
- La distribución conjunta de las variables se ha conservado incluyendo algunos valores de estadísticos puntuales

En un nivel individual interesa saber si los valores imputados se han aproximado a los valores reales no observados.

Así pues, los objetivos principales que se esperan de la operación de fusión y de los cuales se deberán establecer sus indicadores son: coherencia, lo que significa que los valores imputados deberán ser realistas; precisión, lo que quiere decir que los valores imputados se aproximen lo más posible a los valores reales desconocidos y que reproduzca f (Y|X) en el archivo receptor.

Adicionalmente se pueden evaluar estadísticos marginales, así como ciertas medidas estadísticas entre las variables específicas en el archivo receptor suponiendo que tanto el archivo donante como el archivo receptor son muestras de la misma población.

Existen diversas técnicas de fusión, de entre las cuales se ha decidido escoger para este trabajo de investigación la técnica de fusión por búsqueda de los vecinos más cercanos (método de imputación Hot deck) en un espacio de referencia factorial.

Esta decisión permite obtener imputaciones condicionadas a variables comunes, reduciendo de esta manera el sesgo y preservando la asociación entre las variables específicas y comunes. Se puede adaptar para el desarrollo de imputaciones multivariadas, preservando así la asociación entre variables específicas. Se prevén valores sin sentido

[2] Rao y Shao, 1992

(incoherentes) y se pueden hacer imputaciones por extracciones aleatorias de una distribución predictiva para simular valores reales y efectuar imputación múltiple.
El algoritmo de búsqueda de vecinos seleccionado para este trabajo ha sido el algoritmo de Fukunaga/Narendra. Es un algoritmo fácil de llevar a cabo y muy controlable. Las estructuras requeridas para la operación del algoritmo de búsqueda de vecinos se pueden obtener con ayuda de paquetes estadísticos. Tomando ventajas de las prestaciones existentes en paquetes como SPAD, se puede permanecer dentro del marco de estos.
En la imputación por donante resulta de vital importancia la determinación del número de vecinos (donantes) que se habrán de tomar, así como de la forma empleada para seleccionarlos. La calidad de la fusión depende en parte de la calidad del método empleado y de la calidad predictiva de las variables comunes. Se debe tener en cuenta también el propósito para el cual se hace la fusión. El establecer una comparación entre las variables específicas en los donantes y los receptores resultaría un poco artificial dado que las variables específicas en los receptores no han sido observadas. No hay que olvidar que el principal objetivo de la fusión es reconstruir una base de datos lo mas parecida posible a la que se hubiera obtenido con los datos completos. Las etapas para el proceso de fusión quedan definidas de la siguiente manera:

- Preproceso
- Selección de las variables comunes.
- Posicionamiento de donantes y receptores en el mismo subespacio factorial definido por las variables comunes.
- Determinación de la tabla de vecinos en la que se relaciona cada receptor con sus k vecinos más cercanos.
- Imputación de los valores.
- Validación de la imputación.

En la etapa de preproceso, se analizan las características de las muestras. No es necesario que las dos muestras representen a la misma población, pero si que las distribuciones condicionales de las variables a transferir con respecto a las variables comunes sean iguales. Otra característica a revisar es la naturaleza de variables a transferir (normalidad en el caso de las variables continuas, y multinomial en el caso de las discretas). El hecho de no ser normales, implicaría un uso impropio del método estocástico basado en normalidad.

Uso de la ponderación de los individuos para caracterizar la muestra. Deben ser ponderados de tal manera que la muestra de donantes sea representativa de la población. Pesos inversamente proporcionales a la probabilidad de selección (Probabilidad de que un individuo esté presente en la muestra).

En la etapa de selección de variables comunes, se debe asegurar el poder predictivo sobre las variables que se desean imputar. La no existencia de este poder de predicción significará que las imputaciones que se hagan serán equivalentes a una imputación aleatoria.

En la selección del subespacio factorial, se necesita enmarcar las dos muestras en un mismo espacio de referencia a través de las técnicas de análisis factoriales (Análisis de Componentes Principales, Análisis de Correspondencias Múltiples). Las variables activas son las variables comunes en las dos muestras. Así mismo, los individuos en la muestra donante serán considerados como individuos activos. De esta manera entonces, se hace una proyección en suplementario de los individuos en la muestra receptora.

$$\Psi_\alpha = \mathbf{Z}\mathbf{u}_\alpha = \begin{cases} \mathbf{X}\mathbf{S}^{-1}\mathbf{u}_\alpha & \text{(ACP Normalizado)} \\ \mathbf{X}\mathbf{u}_\alpha & \text{(ACP no Normalizado)} \end{cases} \quad [3]$$

$\mathbf{u}_\alpha$ = vectores propios calculados a partir de la muestra de donantes

Una vez que se ha establecido el espacio común, lo siguiente a realizar es la búsqueda de los vecinos más cercanos. El tiempo de búsqueda estará en función del número de variables conservadas y el número de donantes.

Finalmente, con el conjunto de vecinos más cercanos para cada receptor, se puede efectuar el proceso de imputación, para el cual, existen diferentes procedimientos. Esta etapa va a requerir de alguna herramienta para medir su efectividad.

[3] cf. Aluja Banet, T.; Morineau, A. (1999)

3.2 Métodos de imputación

En función de los objetivos de la fusión, existen ciertos asuntos que se deben considerar para la imputación:

- Complejidad, es decir, cuantos vecinos deberán ser tomados en cuenta.
- Condiciones sobre la imputación (restringida o irrestricta).
- Determinística o estocástica.
- Univariada o multivariada.

Cada método seleccionado tiene propiedades que se deben de considerar en función de los objetivos de la fusión.
La imputación determinística (por la media local o la moda local para variables categóricas) reduce la variabilidad de los valores imputados, pero mejora la precisión. Tiende a inflar las correlaciones. La imputación estocástica (extracciones aleatorias de una distribución predictiva f (Y|X) mantiene la variabilidad aunque con una mala precisión. La imputación multivariada preserva las correlaciones entre las variables específicas.
Los distintos métodos (enfoques) que se investigan y desarrollan en la implantación del sistema que se va a presentar toman en cuenta las siguientes características:
Las posibilidades de imputación se muestran en la siguiente tabla:

Tabla 3.1 Esquemas de imputación

		Un vecino	K vecinos
Determinístico		T1DM	TkDM
Estocástico	Univariante	T1SU	TKSU
	Multivariante	T1SM	TKSM

3.2.1 T1DM
(**T**ake **O**ne **D**eterministic **M**ultivariate)

En este método de imputación, se busca para el i-ésimo receptor, el j-ésimo donante (con un perfil semejante), de tal manera que la distancia d (i, j) en el espacio p-dimensional sea mínima. Del conjunto de vecinos más cercanos para el individuo en cuestión, se tomará el más cercano. La imputación se hará copiando todas las variables del donante al receptor.

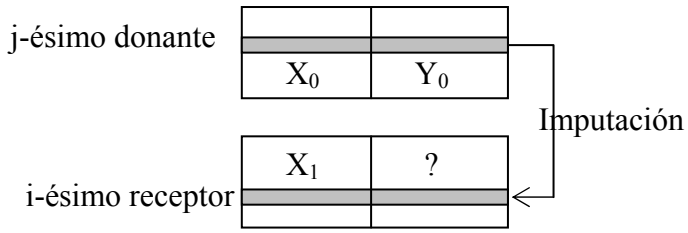


Figura 3.1 Imputación T1DM

Este enfoque evita estimaciones incoherentes ya que los valores imputados pertenecen a valores reales observados. Una desventaja con este método es la repetición de los donantes. Es posible que un mismo donante sea el vecino más cercano de muchos receptores y por lo tanto los valores de las variables se repetirán para cada receptor. El efecto inmediato en la imputación por este método será la reducción de la variabilidad. Para evitar la pérdida de variabilidad, se puede aplicar la idea de una función de penalización, de tal manera que un mismo donante no sea usado demasiadas veces. Esta idea queda incluida en el sistema con el uso de un parámetro P que indica la probabilidad de que un donante pueda ser repetido. Cuando se presenta el evento de que el vecino ya ha sido utilizado, se repetirá con una probabilidad P y por lo tanto será desechado con una probabilidad de 1-P. De esta manera, si el valor de P = 0, el método se convierte en la asignación por el vecino recíproco (cf. 1.3.3.1). Del mismo modo, si P = 1, el método se convertirá en la asignación por el vecino más cercano.

Tenemos entonces:

- Imputación sin restricciones (P = 1)

- Imputación con restricciones (P < 1)

Una definición para el concepto de vecinos recíprocos es la siguiente:

Dos individuos, un donante $\mathbf{d_1}$ y un receptor $\mathbf{r_1}$, son vecinos recíprocos, si $\mathbf{d_1}$ es el donante más cercano a $\mathbf{r_1}$ y $\mathbf{r_1}$ es el receptor más cercano a $\mathbf{d_1}$.

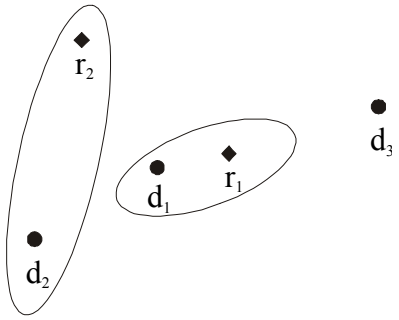


Figura 3.2 Vecinos recíprocos

Se puede ver en la figura que el donante $\mathbf{d_1}$ se ha asignado al receptor $\mathbf{r_1}$ dado que cumplen la definición de vecinos recíprocos. De la misma forma, es necesario asignar un vecino recíproco al receptor $\mathbf{r_2}$. Se puede ver que aunque el donante más cercano a $\mathbf{r_2}$ es $\mathbf{d_1}$, no se puede considerar como su vecino recíproco dado que ya lo es para $\mathbf{r_1}$. Por lo tanto, $\mathbf{d_1}$ quedará excluido de la búsqueda de vecinos recíprocos porque ya ha sido utilizado y en su lugar, se asignará a $\mathbf{d_2}$ como vecino recíproco para $\mathbf{r_2}$.

Asignando los vecinos de esta manera, se asegura la variabilidad en la imputación. Esta alternativa, sin embargo, no esta libre de defectos, ya que se puede presentar el caso no deseado de asignar un vecino muy lejano a un receptor debido a que sus vecinos más cercanos son vecinos recíprocos de otros receptores, lo cual provocaría una estimación poco acertada.

¿Que pasa si existen más receptores que donantes?

Si se tienen n donantes y m receptores ($n < m$), se asignarán las n mejores relaciones de vecinos recíprocos. Una vez hecho esto, se tendrán 0 donantes y n-m receptores. Se reinsertarán los n donantes y se repetirá el proceso hasta que todos los receptores tengan su vecino recíproco. Debido a la diferencia entre donantes y receptores, algunos donantes se repetirán pero lo harán el mismo número de veces, de manera que la pérdida de variabilidad se atenúa.

Para la implantación de este algoritmo se requiere de una estructura que datos que permita hacer inserciones y eliminaciones continuas y que proporcione el par con distancia mínima de forma eficiente. El costo de inserción en una lista ordenada es lineal, buscar el mínimo

en una lista desordenada también tiene costo lineal y los algoritmos rápidos de ordenación tienen un costo de O(nLog n). El costo de estas tres opciones hace poco viable el algoritmo. La estructura de datos que proporciona el comportamiento deseado es el heap. En esta estructura se pueden hacer inserciones y eliminaciones con un costo logarítmico.
Con este algoritmo, los donantes se repiten cuando es absolutamente necesario, ya sea porque hay mas receptores que donantes o porque existen receptores para los que todos sus vecinos ya son donantes de otro receptor.

Pseudo código del algoritmo Fast RN

```
fastRN (receptor[n], m', neighbour[n][k])
maxDonations = max (1, n div m');
for      i=1 To n Do
        reciprocal[i] = -1;
        actNeighbour[i] = 1;
key = neighbour[i][1].distance;
        pair[1] = receptor[i];
        pair[2] = neighbour[i][1].donor;
        workingHeap.insert(key, pair);
endFor;
while   not workingHeap.isEmpty() Do
              for          i=1 To m Do
                     nDonations[i] = 0;
              endFor;
              while      not workingHeap.isEmpty() Do
                           pair = workingHeap.getFirst(&key);
                           workingHeap.deleteFirst();
                           r = pair[1];
                           if      reciprocal[r] = -1 Then
                                  d = pair[2];
                                  if       nDonations[d] < maxDonations Then
                                         reciprocal[r] = d;
                                         nDonations[d] = nDonations[d] + 1;
                                  else
                                         v = actNeighbour[r] + 1;
                                         actNeighbour[r] = v;
                                         if          v > k Then
                                                actNeighbour[r] = 0;
                                                key = neighbour[r][0].distance;
                                                pair[1] = receptor[r];
                                                pair[2] = neighbour[r][0].donor;
                                                recyclingHeap.insert(key, pair);
                                         else
```

```
                                    key = neighbour[r][v].distance;
                                        pair[1] = receptor[r];
                                        pair[2] = neighbour[r][v].donor;
                                        workingHeap.insert(key, pair);
                                    endIf;
                                endIf;
                            endIf;
                    endWhile;
                    workingHeap = recyclingHeap;
                    recyclingHeap.remove();
endWhile;
return reciprocal;
end;
```

3.2.2 T1SU

(Take **O**ne **S**tochastic **U**nivariate)

En este método, se selecciona de manera aleatoria un vecino para cada variable seleccionada para imputación. Se copia el valor de la variable en cuestión del vecino seleccionado a la variable correspondiente en el receptor. Si se establece que los k donantes se disponen en un arreglo llamado **V**, el procedimiento de imputación es el siguiente:

- Generar $\min(i) \mid \sum_{r=1}^{i} p_r \geq u \sim U(0,1)$ $\quad p_r = f(w)$

  w = peso de cada donante

  i = vecino entre los k existentes

- y = **V**[i] (valor de la variable imputada)

El peso de los donantes varía inversamente con la distancia, de tal manera que los puntos más cercanos tienen mayor peso.

$$w_i = K\,[d\,(x_r, x_d)] \qquad (3.1)$$

K es la función Kernel que determina el peso de cada punto basado en la distancia al punto de referencia (cf. 2.2).

Ejemplo 3.1

En este caso, se tiene para el receptor **r**, el conjunto de sus k = 5 vecinos más cercanos **V**. La función Kernel que determina el peso de cada punto basado en la distancia al punto de referencia es: $w_i = \frac{1}{d_0 + d}$ con $d_0 = 0$.

Tabla 3.2 Distancias y pesos de los vecinos

	Vecinos	$d = d(r, V_i)$	$1/d_i$	$p(V_i)$	$\Sigma\, p(V_i)$
	V_1	1.0	1.000	0.385	0.385
	V_2	1.5	0.667	0.256	0.641
i = 3 →	V_3	2.5	0.400	0.154	0.795
	V_4	3.0	0.333	0.128	0.923
	V_5	5.0	0.200	0.077	1.000

$$p(V_i) = \frac{1/d_i}{\sum_i 1/d_i}$$

- Valor generado de $U \sim U(0, 1) = 0.645$

- Valor mínimo de i para el cual la suma acumulada de probabilidades es mayor que U = 0.645 es 3.

El vecino seleccionado para imputar el valor de la variable es el V_3. Por lo tanto el valor de la variable en cuestión del vecino V_3 se imputará a la variable correspondiente en el receptor. Este procedimiento se repite para cada variable específica.

Este método reduce el efecto de la reducción de la variabilidad por variable por su naturaleza aleatoria, pero no toma en cuenta las relaciones entre las variables.

3.2.3 T1SM

(**T**ake **O**ne **S**tochastic **M**ultivariate)

En este caso, se selecciona de manera aleatoria un vecino para cada receptor. Son extracciones probabilísticas locales. Ahora se copiará el valor de todas las variables seleccionadas para imputación del vecino seleccionado al individuo receptor. El procedimiento es semejante al empleado en el caso T1SU. Con los k donantes en el arreglo **V**, el procedimiento de imputación es:

- Generar $\min(i) \mid \sum_{r=1}^{i} p_r \geq u \sim U(0,1)$ $\qquad p_r = f(w)$

  w = peso de cada donante

  i = vecino en los k existentes

- **Y** = **V**[i] (vector de valores de las variables para imputación)

Al igual que en el caso anterior, el peso de los donantes varía inversamente con la distancia, de tal manera que los puntos más cercanos tienen mayor peso.

De esta manera se conservan las correlaciones existentes entre las variables específicas, dado que los valores provienen del mismo vecino. Al mismo tiempo se obtienen estimaciones coherentes puesto que los valores imputados son valores reales observados y la característica aleatoria del método reduce el efecto de la reducción de variabilidad.

3.2.4 TKDM

(**T**ake **K** **D**eterministic **M**ultivariate)

En este método, el valor imputado de cada variable, será el resultado de un valor centrado entre los K vecinos seleccionados. Hay que hacer notar, que esta centralización se hace de manera ponderada.

Este método de imputación preserva la media pero distorsiona la dispersión y las relaciones entre las variables.

En el caso de las variables continuas:

el valor imputado será la media ponderada de los K vecinos más cercanos, y el peso de cada vecino vendrá especificado de acuerdo con la función Kernel que determina el peso de cada punto basado en la distancia al punto de referencia (cf. 2.2).
Así, el valor para la j-ésima variable del receptor, será:

$$y_j = \sum_{m=1}^{K} w_m \mathbf{V}_{mj} \qquad (3.2)$$

en donde:

K = número de vecinos
w_m = peso del m-ésimo vecino
$\mathbf{V}_{mj}$ = valor de la j-ésima variable en el m-ésimo vecino

De esta manera se toma en consideración la información de los K vecinos y por lo tanto, el resultado será un buen estimador puntual. Sin embargo, por el hecho de asignar un valor por la media de los vecinos, se verá reducida la variabilidad.

En el caso de las variables nominales:

el valor imputado será la moda local ponderada (calculada con los K vecinos). Se construye un histograma en el que cada vecino aportará su propio peso. De esta manera, la selección ya no estará solo en función de la ocurrencia de cada modalidad, sino además del peso de cada vecino.

Ejemplo 3.2
En este caso, se tiene para el receptor **r**, el conjunto de sus K = 6 vecinos más cercanos **V**. La función Kernel que determina el peso de cada punto basado en la distancia al punto de referencia es:

$$w_i = \frac{1}{d_0 + d} \text{ con } d_0 = 0.$$

Se considera una variable nominal con 3 modalidades.

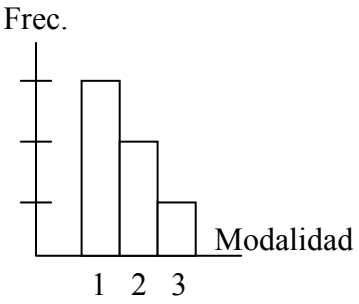


Figura 3.3 Histograma de modalidades para una variable

Tabla 3.3 Vecinos más cercanos

Vecinos	$d_i = d(r, V_i)$	$1/d_i$	Modalidad
V_1	1.0	1.000	2
V_2	1.5	0.667	1
V_3	2.5	0.400	2
V_4	3.0	0.333	3
V_5	5.0	0.200	1
V_6	5.8	0.172	1

Se puede ver en la figura 3.3 que la modalidad con mayor frecuencia de ocurrencia es la 1. Sin embargo, al considerar los pesos correspondientes de los vecinos, el resultado será diferente como se muestra en la siguiente tabla.

Tabla 3.4 Asignación de pesos

Modalidad	Frec.	Pesos	
1	3	0.667 + 0.200 + 0.172	1.039
2	2	1.000 + 0.400	1.400
3	1	0.333	0.333

La mayor proximidad de la modalidad 2 y en consecuencia el peso asignado a estos individuos, ocasiona que el valor imputado a la variable en cuestión del receptor sea la modalidad 2.

3.2.5 TKSU

(**T**ake **K** **S**tochastic **U**nivariate)

De manera independiente, para cada variable existirán K vecinos. Se busca reproducir de manera empírica la distribución que sigue cada una de las variables seleccionadas para imputación. La reconstrucción de la distribución se hará a partir de los valores de los vecinos y de sus distancias con respecto al receptor. Se emplea la idea de las distribuciones empíricas. La construcción de una distribución empírica esta básicamente en función de los datos disponibles[4]. Estos pueden ser:

- Observaciones individuales
- Datos agrupados

En el caso de observaciones individuales, se define una función de distribución acumulada a partir de segmentos de recta. Se ordenan de manera creciente las observaciones.

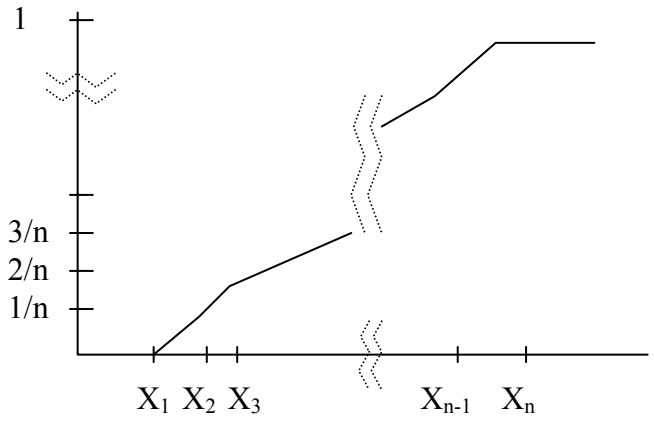


Figura 3.4 F.D.A. empírica

$$F(x)=\begin{cases}0 & x < X_1 \\ \dfrac{i-1}{n-1}+\dfrac{x-X_i}{(n-1)(X_{i+1}-X_i)} & X_i \le x < X_{i+1} \\ 1 & X_n \le x\end{cases} \qquad (3.3)$$

[4] Law, A.M.; Kelton, W.D. (1991)

En este caso se está suponiendo que las observaciones son equiprobables ($F(X_i) = \frac{i-1}{n-1}$), sin embargo, esto se puede modificar, asignándole a cada observación una probabilidad proporcional a su distancia.

En el caso de que el número de observaciones de acuerdo con el usuario sea grande, estas se pueden agrupar en k intervalos adyacentes generalmente equidistantes:

$$[a_0, a_1), [a_1, a_2), \ldots, [a_{k-1}, a_k].$$

La idea de que se considere grande el número de observaciones, corresponde al número de segmentos de línea que se tendrán que formar y no a un concepto probabilístico ya que las distribuciones que se están formando son empíricas.

Cada intervalo contiene n_j observaciones de tal manera que: $n_1 + n_2 + \ldots + n_k = n$

Se define entonces una función de distribución acumulada de la siguiente manera:

$$G(a_0) = 0$$

$$G(a_j) = \frac{(n_1 + n_2 + \ldots + n_j)}{n} \qquad j = 1,2,\ldots k$$

$$G(x) = \begin{cases} 0 & x < a_0 \\ G(a_{j-1}) + \dfrac{x - a_{j-1}}{a_j - a_{j-1}} & a_{j-1} \le x < a_j \\ 1 & a_k \le x \end{cases} \qquad (3.4)$$

A partir de estas definiciones, la imputación se hace empleando el enfoque de la transformada inversa para generar valores de variables aleatorias distribuidas de manera empírica.

$$x = \tilde{F}^{-1}(p) \qquad (3.5)$$

en donde p representa un número aleatorio ~ U (0, 1).

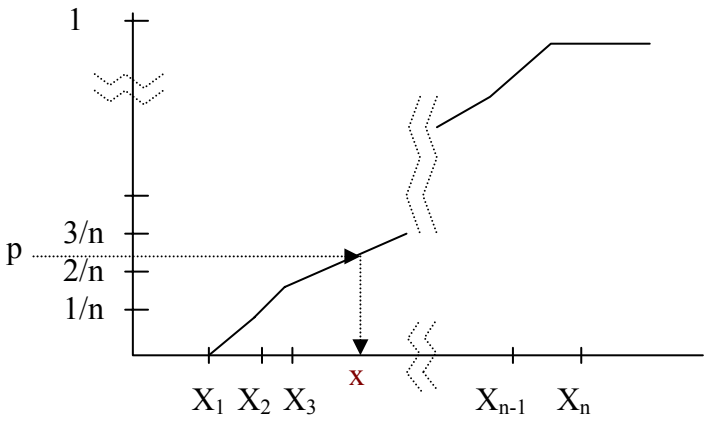


Figura 3.5 Imputación con una FDA empírica

Ejemplo 3.3

En este caso, se tiene para el receptor **r**, el conjunto de sus K = 6 vecinos más cercanos **V**. La función Kernel que determina el peso de cada punto basado en la distancia al punto de referencia es:

$$w_i = \frac{1}{d_0 + d} \text{ con } d_0 = 0.$$

Tabla 3.5 Vecinos más cercanos

Vecinos	$\mathbf{d_i = d(r, V_i)}$	$\mathbf{1/d_i}$	**X**	$\mathbf{p(x_i)}$
$\mathbf{V_1}$	1.0	1.000	38	0.361
$\mathbf{V_2}$	1.5	0.667	41	0.240
$\mathbf{V_3}$	2.5	0.400	47	0.144
$\mathbf{V_4}$	3.0	0.333	50	0.120
$\mathbf{V_5}$	5.0	0.200	25	0.072
$\mathbf{V_6}$	5.8	0.172	75	0.062

De acuerdo a lo antes mencionado, se deben ordenar los vecinos en orden creciente con respecto al valor de la variable. Se muestran en la tabla 3.5 para cada vecino, sus distancias y pesos. El valor imputado para una variable empleando este método se hará siguiendo el enfoque de transformada inversa, que de acuerdo a la ecuación (3.3) será:

$$x = \frac{[U - p(X_i)](X_{i+1} - X_i)}{\Delta p} + X_i \qquad (3.6)$$

U ~ U (0, 1)

i = intervalo al que pertenece u

Δp = diferencia de probabilidades acumulados en los puntos (X_{i+1}, X_i) => p (X_i).

Tabla 3.6 Vecinos ordenados por su valor

	Vecinos	X	$p(X_i)$	$\Sigma\ p(x_i)$
	V_5	25	0.072	0.072
i = 2 →	V_1	38	0.361	0.433
	V_2	41	0.240	0.673
	V_3	47	0.144	0.818
	V_4	50	0.120	0.938
	V_6	75	0.062	1.000

Existe un punto X_0 que será el valor mínimo para esa variable y un intervalo 0 cuya probabilidad acumulada será 0.

- Valor generado de U ~ U (0, 1) = 0.45
- Valor máximo de i para el cual la suma acumulada de probabilidades es menor que U = 0.45 es 2.
- $X_i = 38$ $X_{i+1} = 41$

$$X = \frac{(0.45 - 0.433)(41 - 38)}{0.24} + 38 = 38.31$$

Para el caso de las variables nominales, se busca reproducir su función de distribución empírica a través de su función masa de probabilidad p(x), enfoque similar al empleado en el caso de las variables continuas. Esto se hace construyendo un histograma de modalidades, asignando a cada modalidad una probabilidad en función del Kernel que determina el peso de cada punto basado en la distancia al punto de referencia (cf. 2.2).

La modalidad seleccionada se hace de acuerdo al enfoque de la transformada inversa para variables categóricas.

El procedimiento de asignación es el siguiente:

- Generar $\min(i) \mid \sum_{r=1}^{i} p_r \geq U \sim U(0,1)$ $\qquad p_r = f(w)$

  w = peso de cada modalidad

  i = modalidades de la variable

- y = i (valor de la modalidad imputada)

Ejemplo 3.4

En este caso, se tiene para el receptor **r**, el conjunto de sus K = 6 vecinos más cercanos **V**. La función Kernel que determina el peso de cada punto basado en la distancia al punto de referencia es:

$$w_i = \frac{1}{d_0 + d} \text{ con } d_0 = 0.$$

Tabla 3.7 Vecinos más cercanos

Vecinos	$\mathbf{d_i = d(r, V_i)}$	$\mathbf{1/d_i}$	**Modalidad**
$\mathbf{V_1}$	1.0	1.000	2
$\mathbf{V_2}$	1.5	0.667	1
$\mathbf{V_3}$	2.5	0.400	2
$\mathbf{V_4}$	3.0	0.333	3
$\mathbf{V_5}$	5.0	0.200	1
$\mathbf{V_6}$	5.8	0.172	1

Tabla 3.8 Histograma ponderado

	Modalidad	Frec.	**Pesos**		$\mathbf{p(x_i)}$
i = 1 →	**1**	3	0.667 + 0.2 + 0.172	1.039	0.375
	2	2	1 + 0.4	1.400	0.505
	3	1	0.333	0.333	0.120

- Valor generado de $U \sim U(0, 1) = 0.35$

- Valor mínimo de i para el cual la suma acumulada de probabilidades es mayor que U = 0.35 es 1.

La modalidad seleccionada para imputar la variable en cuestión será entonces 1.

3.2.6 TKSM

(**T**ake **K** **S**tochastic **M**ultivariate)
Se efectúan imputaciones por extracciones aleatorias de una distribución multivariada de variables específicas. Se supone normalidad multivariada para variables continuas o distribución multinomial para variables categóricas. Se busca con este enfoque reproducir las relaciones existentes entre las variables específicas en los donantes, conservando las características locales de los K vecinos más cercanos (representadas en el caso de variables continuas por la media y la desviación estándar).

El procedimiento de imputación para variables continuas es el siguiente:
(Generación aleatoria de observaciones $\sim N_p(\overline{\mathbf{Y}}, \mathbf{S})$, en donde $\overline{\mathbf{Y}}$ representa el vector de medias muestral $\left(\bar{y}_1 \quad \bar{y}_2 \quad \ldots \quad \bar{y}_p\right)$ y $\mathbf{S} = \mathbf{V}^{1/2}\rho\mathbf{V}^{1/2}$ es la matriz de varianzas-covarianzas muestral. V es la matriz diagonal de desviaciones estándar).

- Obtención de los valores y vectores propios (λ_i, **U**) de las variables específicas en el archivo donante.
  λ_i = valores propios**,** i = 1, . . . p (número de variables específicas continuas)
  U = matriz de vectores propios

- Generación de un vector Normal Multivariante $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{\Sigma})$ con $\mathbf{\Sigma} = \mathbf{I}_p$

- Producto matricial $\mathbf{Y} = \mathbf{Z}\,\mathbf{D}_\lambda$
  $\mathbf{D}_\lambda$ = matriz diagonal con los valores propios.
- Imputación (normalizada)

$$\tilde{\mathbf{X}} = \mathbf{YU'}$$

El procedimiento anterior representa la extracción aleatoria de un individuo en el espacio factorial (distribución global). A esta extracción (individuo) se le asignan las características contenidas en los K vecinos más cercanos (distribución local).
Las variables se transforman de acuerdo a las características que define el espacio original (media local y desviación estándar local). De esta manera la imputación estará afectada por los valores de los vecinos seleccionados.

$$\hat{x}_{ij} = \tilde{x}_{ij}\,\sigma_j^l + \bar{x}_j^l \qquad (3.7)$$

σ_j^l = desviación estándar local para la j-ésima variable.

$\bar{x}_j^l$ = medial local para la j-ésima variable.

$\hat{\mathbf{X}}$ representa entonces los valores imputados de las variables continuas seleccionadas para imputación.

Si alguna de las variables que se desean imputar empleando este enfoque no sigue una distribución normal, se puede aplicar alguna de las transformaciones de normalidad incluidas en el sistema (**ln**(x), √x, **logit**(x) [5]).

En el caso de las variables nominales el procedimiento es el siguiente:

(Extracciones aleatorias de un hiperespacio de variables multinomiales con:

$$P(y_1=n1, \ldots, y_G=n_G)=\frac{n!}{n_1!n_2!\ldots n_G!}p_1^{n_1}\ldots p_G^{n_G}$$

G es el número de modalidades para la variable)

[5] $\log it(x) = \ln\left(\frac{x}{1-x}\right)$

Se imputarán variables categóricas conservando las relaciones que existen entre ellas. Se construye ahora un hipercubo de p dimensiones. Se tendrán tantas dimensiones como variables categóricas específicas existan. Cada dimensión tendrá tantas divisiones como modalidades tenga la variable que representa. Cada celda en el hipercubo tendrá diferente probabilidad de ocurrencia. Se harán extracciones aleatorias de este hipercubo. A cada individuo le corresponderá una celda del hipercubo que depende de los valores de sus variables categóricas. Entre más individuos haya en una celda, mayor será su probabilidad de ocurrencia. Se posicionan ahora los K vecinos más cercanos de un receptor en el hipercubo (local). Cada individuo aporta a la celda correspondiente un peso proporcional a la distancia al receptor en función del kernel empleado (cf. 2.2). El hipercubo así generado, contendrá información local sobre las relaciones que existen entre las diferentes categorías de las variables. Se hacen ahora extracciones aleatorias del hipercubo.

El procedimiento de imputación para variables categóricas es el siguiente:

- Construcción del hipercubo a partir de los K vecinos más cercanos

- Asignación de probabilidades a las celdas del hipercubo.

- Generar $\min(i) \mid \sum_{r=1}^{i} p_r \geq u \sim U(0,1)$ $\qquad p_r = f(w)$

  w = peso de cada celda

  i = celda en el hipercubo

- **Y** = Modalidades correspondientes a la i-ésima celda (celda seleccionada)

Debido a la naturaleza de los datos, existirán algunas celdas vacías considerando el número de vecinos seleccionados. Esto significa la no ocurrencia de esa combinación de modalidades para las variables específicas en la muestra de K vecinos, generando por lo tanto un hipercubo disperso.

Ejemplo 3.5

En este caso, se tiene para el receptor **r**, el conjunto de sus K = 10 vecinos más cercanos. La función Kernel que determina el peso de cada punto basado en la distancia al punto de referencia es:

$$w_i = \frac{1}{d_0 + d} \text{ con } d_0 = 0.$$

Se manejan dos variables y tres modalidades cada una.

Tabla 3.9 Valores de los vecinos

Vecino	Y_1	Y_2	distancia	Peso
1	1	1	1,00	1,00
2	1	3	1,50	0,67
3	2	1	2,00	0,50
4	2	2	2,50	0,40
5	2	3	3,00	0,33
6	2	1	3,50	0,29
7	3	1	4,00	0,25
8	3	1	4,50	0,22
9	3	2	5,00	0,20
10	1	3	5,50	0,18

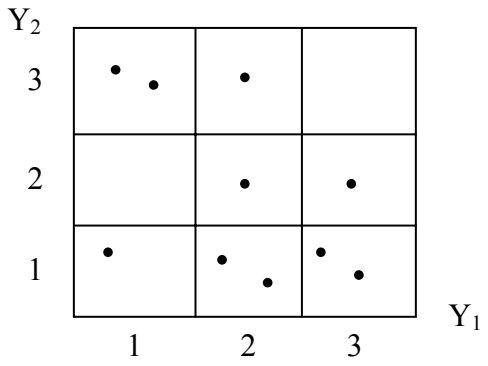


Figura 3.6 Hipercubo (2 dimensiones)

Tabla 3.10 Probabilidad de las modalidades

	Y_1	Y_2	p_r	Σp_r
	1	1	0,248	0,248
	2	1	0,196	0,444
i = 3 →	3	1	0,116	0,560
	1	2	0,000	0,560
	2	2	0,099	0,659
	3	2	0,050	0,709
	1	3	0,210	0,919
	2	3	0,082	1,000
	3	3	0,000	1,000

- Valor generado de $U \sim U(0, 1) = 0.560$

- Valor mínimo de i para el cual la suma acumulada de probabilidades es mayor que U = 0.56 es 3.

Las modalidades seleccionadas para imputar la variable en cuestión serán entonces (3, 1).

3.3 Otras opciones

3.3.1 Variables mixtas

La definición de variables mixtas responde a la realidad de conceptos cotidianos. Por ejemplo, la idea de patrimonio; no tendría sentido medir el patrimonio de un individuo que no lo tiene. Se debe investigar en primer lugar si lo tiene o no. En caso de tenerlo se intentará medirlo.

Otro ejemplo de variable mixta es: compras mensuales por Internet. Se puede presentar el caso de un comprador habitual por Internet que durante un mes determinado no haya realizado compras.

También es posible que se trate de un individuo que nunca compra por Internet o que no tenga acceso a la red. En este último caso entonces, no tiene sentido estimar cuantas compras por Internet realiza un individuo que no tiene conexión a la red.

Este tipo de variables, que llamamos mixtas tienen un comportamiento categórico y continuo al mismo tiempo. Requieren de un tratamiento especial.

- Decidir si se les dará un valor numérico o no (Tratamiento categórico).

- En caso afirmativo: tratamiento continuo correspondiente a tipo de imputación seleccionado.

Los valores imputados para variables continuas pueden ser redondeados al entero más próximo, lo que permitiría efectuar imputaciones para variables categóricas por métodos diseñados para las variables continuas. Sin embargo, esto no es muy recomendable, ya que se puede incurrir en errores de juicio al momento de evaluar los resultados de la imputación. No obstante, la opción como tal, otorga versatilidad al diseño del sistema.

3.3.2 Redondeo

Otra característica implantada en este sistema, es la capacidad para filtrar los valores imputados. Existe la opción de asignar valores mínimos que pueden ser imputados, así como valores máximos. Esta capacidad de establecer umbrales, evita la imputación de valores extremos a las variables.
Es posible también que el interés esté en la imputación de valores continuos dentro de un rango preestablecido.

3.4 El parámetro π

El método de imputación mediante el uso de los K vecinos más cercanos, es un método de imputación local. Los valores que se imputan a cada receptor se generan con la información contenida en un subespacio determinado por el número de los vecinos más cercanos seleccionados. La información local esta representada por estos K vecinos. Mediante el uso del parámetro π se harán las imputaciones considerando tanto los estadísticos locales como globales. Tomando como referencia los conceptos de estimación para áreas pequeñas, se pueden modificar algunos de los métodos empleados para imputación.

3.4.1 Estimadores para áreas pequeñas

Los estimadores compuestos combinan estimadores de área local y nacional con pesos determinados para optimizar la precisión de la combinación. El método propuesto es computacionalmente elemental y no demandante. Se puede aplicar una versión multivariante del estimador compuesto. Cuando las medias están correlacionadas a nivel de área local, el estimador multivariado es más eficiente que el univariado aplicado separadamente a cada variable [6] .

A partir de un diseño de muestreo aleatorio simple la información nacional contiene los valores y_{il} de la variable y para los sujetos i = 1 , 2 , . . . n_l para áreas locales l = 1 , 2 , . . . , L. $N = n_1 + \ldots + n_L$ $q_l = n_l / N$.

Las medias muestrales de área local son: $$\hat{p}_l = \frac{\sum_{i=1}^{n_l} y_{il}}{n_l} \qquad (3.8)$$

y la media muestral nacional es: $$\hat{p} = \frac{1}{N}\sum_l \sum_i y_{il} = \sum_l q_l \hat{p}_l \,. \qquad (3.9)$$

El estimador compuesto es entonces: $$\tilde{p}_l = (1 - b_l)\hat{p}_l + b_l p \qquad (3.10)$$

Los coeficientes b_l se escogen para minimizar EMSE[7] = $$\text{EMSE} = E_s\left[E_l\left\{ \left(\tilde{p}_l - p_l \right)^2 | p_l \right\} \right]$$

en donde E_l denota la esperanza dada el área local y E_s denota la esperanza (incondicional) sobre las áreas locales. EMSE es una función cuadrática de b_l; si $q_l < ½$, el mínimo se obtiene para:

$$b_l^* = \frac{v_l(1 - q_l)}{v_l(1 - 2q_l) + \text{var}(\hat{p}) + \sigma^2} \qquad v_l = \text{var}(\hat{p}_l | p_l) \,. \qquad (3.11)$$

[6] cf. Nicholas T. Longford (1999)
[7] La mejora de la estimación es en valor esperado, en términos de la reducción en el error cuadrático medio esperado (EMSE)

La teoría estadística de estimación en áreas pequeñas propone una forma de combinar estimadores directos (obtenidos de los datos de muestra pertenecientes al área pequeña) e indirectos (obtenido a través de información auxiliar de otras áreas, periodos o fuentes estadísticas). Los estimadores directos son insesgados pero poco precisos y los indirectos son generalmente sesgados con menor variación.

El estimador compuesto es la combinación lineal de los estimadores directa e indirecta que minimiza el error cuadrático medio. Representa un compromiso entre la ausencia de sesgo y mínima varianza.

$$\tilde{\theta}_j = \pi_j \hat{\theta}_* + (1 - \pi_j)\hat{\theta}_j \tag{3.12}$$

Se pueden considerar tres estimadores compuestos (cf. Costa, A. et. al. 2003)

- Teórico (usa factores de peso óptimos)
- Clásico (usa factores de peso estimados, supone homogeneidad de varianza y sesgo para todos las áreas).
- Alternativo (usa valores específicos de área de sesgo y varianza)

Para el uso del estimador compuesto clásico, se tiene lo siguiente:

$$\hat{\theta}_j \sim N(\theta_j, \sigma^2), \quad j = 1, 2, \ldots, J \qquad \text{con} \quad \theta_j \sim N(\theta_*, b^2)$$

siendo b^2 una estimación común del sesgo cuadrado.

La media ponderada de las varianzas muestrales para cada área es:

$$\bar{s}^2 = \frac{\sum_{j=1}^{J} (n_j - 1)s_j^2}{(n - J)} \tag{3.13}$$

n es el tamaño de la muestra completa, n_j el tamaño muestral del área pequeña. Suponiendo que $\sigma_j^2 = \sigma^2$ para todo j, el estimador σ_j^2 es $\frac{\bar{s}^2}{n_j}$.

Para el sesgo cuadrado $(\theta_* - \theta_j)^2$ se define el estimador común:

$$b^2 = \frac{\sum_{j=1}^{J}\left(\hat{\theta}_j - \hat{\theta}_k\right)}{J} \tag{3.14}$$

Se tiene entonces como estimador de π_j: $\hat{\pi}_j = \dfrac{s^{-2}/n_j}{s^{-2}/n_j + b^2}$ (3.15)

3.4.2 Imputación TKDM

3.4.2.1 Variables continuas

Como se mencionó en la sección 3.2.4, el valor imputado será la media ponderada de los K vecinos más cercanos, y el peso de cada vecino estará especificado de acuerdo a la función Kernel que determina el peso de cada punto con base en la distancia al punto de referencia (cf. 2.2).

Así, el valor para la j-ésima variable del receptor r, será:

$$r_j = \sum_{m=1}^{k} w_m v_{mj} \tag{3.16}$$

en donde:

k = número de vecinos

w_m = peso del m-ésimo vecino

v_{mj} = valor de la j-ésima variable en el m-ésimo vecino

El valor obtenido de esta manera será ahora ponderado con la media global, obteniendo así la media *conjunta* definida como:

$$\hat{\mathbf{y}} = \bar{\mathbf{y}}_c = \pi\bar{\mathbf{y}}_l + (1-\pi)\bar{\mathbf{y}}_g \tag{3.17}$$

en donde:

$\bar{\mathbf{y}}_l$ = media local (imputada con el método TKDM)

$\bar{\mathbf{y}}_g$ = media global

$\bar{\mathbf{y}}_c$ = media *conjunta*

El efecto de la distribución global se pone de manifiesto en la ecuación 3.17 a través del valor de π. Si este valor es igual a 1, la imputación estará en función de la muestra local solamente.

3.4.2.2 Variables categóricas

En el caso categórico, la imputación se hará con la moda de la función masa de probabilidad ponderada $p_c(x)$.

$$p_c(x) = \pi p_l(x) + (1-\pi)p_g(x) \qquad (3.18)$$

En la figura 3.7 se muestran las funciones masa de probabilidad local y global. Considerando un valor de π = 0.5 para ponderar las categorías de las variables, se obtiene la función masa de probabilidad *ponderada* mostrada en la figura 3.8.

Al igual que en el caso de las variables continuas, el efecto del valor de π se puede ver en la ecuación 3.18.

Se pretende con esta idea, mezclar de manera suavizada la distribución global y local para desarrollar la imputación.

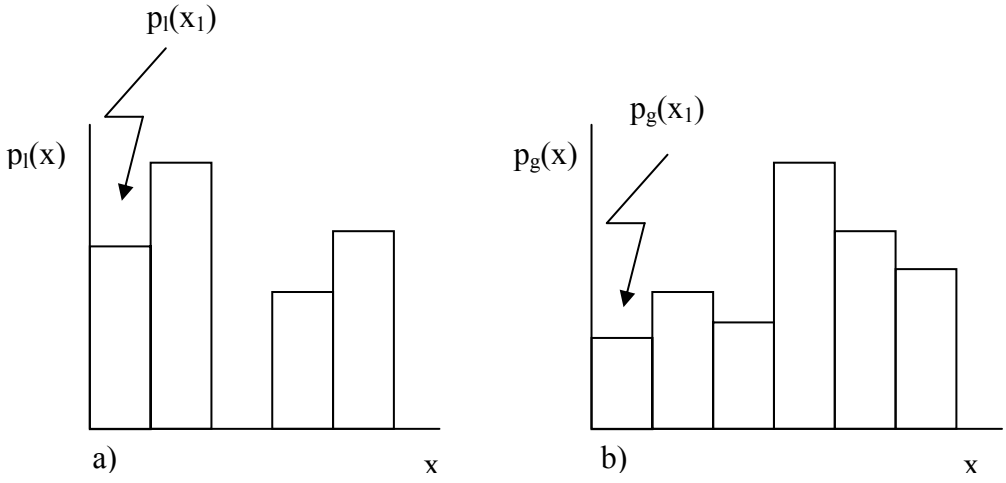


Figura 3.7 Función masa de probabilidad a) local b) global

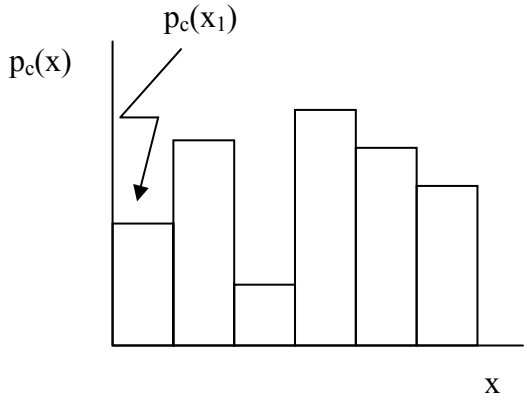


Figura 3.8 Función masa de probabilidad *ponderada*

3.4.3 Imputación TKSU

3.4.3.1 Variables continuas

El valor imputado mediante la distribución empririca (3.7) será desplazado una distancia igual a la diferencia entre la media local de la variable en cuestión y la media *conjunta* (3.18).

$$\hat{y} = \left(\bar{y}_c - \bar{y}_l\right) + y_e \qquad (3.19)$$

y_e = valor imputado con la distribución empírica.

Si el valor de $\pi = 1$, la media conjunta será igual a la media local y por lo tanto, el desplazamiento será nulo. Por lo tanto la imputación se hará unicamente con la información local.

3.4.3.2 Variables categóricas

A partir de la función masa de probabiliad *ponderada* $p_c(x)$ (3.19) se hacen extracciones aleatorias para generar los valores que serán imputados.

3.4.4 Imputación TKSM

3.4.4.1 Variables continuas

Siguiendo el procedimiento visto en la sección 3.2.6 se hará una modificación al momento de establecer el valor que será imputado. En este caso se usará la media *conjunta* y la desviación *conjunta* que se propone de la siguiente forma:

$$\bar{y}_c = \pi\bar{y}_l + (1-\pi)\bar{y}_g \qquad \sigma_c = \sqrt{\pi s_l^2 + (1-\pi)s_g^2} \tag{3.20}$$

Se puede ver en este procedimiento que si el valor de $\pi = 1$, la imputación se basará exclusivamente en la información local. Del mismo modo, con el valor de $\pi = 0$, la imputación se basará exclusivamente en la información global.

3.4.4.2 Variables categóricas

Para las imputaciones con este método, la propuesta es la misma que la vista en la sección 3.2.6 para variables categóricas, con la diferencia de hacer extracciones de un hibercubo *conjunto* $\mathbf{H}_c$ construido como:

$$\mathbf{H}_c = \pi\mathbf{H}_l + (1\text{-}\pi)\mathbf{H}_g \tag{3.21}$$

3.5 Validación

Como se ha mencionado, existen muchas formas de ejecutar la imputación de valores para las variables específicas, pero al final, tiene que existir evidencia de que el método escogido es lo suficientemente exacto para el interés del proceso de imputación. Esto es lo que se conoce como proceso de validación.

Se debe estimar la bondad del procedimiento empleado, analizando los resultados obtenidos a través de diferentes métodos de validación. Si se emplea una prueba incorrecta para la validación, los resultados se pueden convertir en irrelevantes y sin sentido.

Una taxonomía de los niveles de validación es la siguiente:[8]

- Nivel D: en un nivel agregado, se desea saber que tan bien se preservan las distribuciones marginales en los archivos originales.

- Nivel C: en un nivel agregado, se desea saber que tan bien se preserva la estructura de correlaciones de los pares de variables.

- Nivel B: En un nivel agregado, se desea saber que tan bien se reproduce la distribución conjunta real de todas las variables, incluyendo ciertos estadísticos sumarios.

- Nivel A: en un nivel individual, se desea saber que tan bien se preservan los valores reales.

Adicionalmente, para las variables binarias hay cuatro posibles resultados: falso positivo, falso negativo, verdadero positivo y verdadero negativo. Con estos resultados se pueden establecer medidas como: exactitud, precisión, sensitividad. Para las variables continuas existen medidas de dispersión.
Se proponen en este trabajo diferentes medidas de calidad para evaluar el proceso de fusión, resumidas como:

- Comparación de estadísticos marginales sobre variables comunes: E[X0] y E[X1].

- Comparación de estadísticos marginales sobre variables específicas: E[Y0] y E[Y1].

- Comparación de coherencia interna: Cor (Y_0) y $\mathrm{Cor}\left(\hat{Y}_1\right)$.

[8] Rässler, S. (2002)

- Reproducción de las correlaciones(X0, Y0) sobre los valores imputados $(X_1, \hat{Y}_1)$.

- Cálculo de errores (medida de exactitud).

La comparación de los estadísticos marginales solo tendrá sentido si los archivos provienen de la misma población. Se podría establecer el mismo comentario para la coherencia interna. Como la imputación se hace sobre la base de que f (Y|X) es la misma en los donantes como en los receptores, está medida se debe cumplir. Por lo tanto la estructura de correlaciones para las variables comunes como específicas deberá ser semejante en los donantes como en los receptores.

La implantación de estas medidas se hará con el uso de indicadores puntuales como por intervalo. Así mismo, se hará uso de descriptores gráficos para medir la calidad del proceso de fusión.

Una forma de validar la exactitud es haciendo uso de la información disponible (archivo donante). Efectuando la imputación sobre él mismo, es posible medir la desviación $y_{i0} - \hat{y}_{i0}$ y establecer una medida de error.

Así pues, por todo lo antes mencionado, se puede concluir que la calidad de la fusión depende en gran parte del método mismo y en la capacidad predictiva de las variables comunes. Por lo tanto, la evaluación del proceso de imputación estará regida por la siguiente suposición:

El conjunto de los individuos donantes, es una muestra representativa de la población. Se toman los estadísticos donantes como poblacionales. Se probará la suposición de que las dos poblaciones de las cuales fueron extraídos los donantes y los receptores son idénticas con respecto a las leyes de probabilidad para las variables específicas.

Las pruebas de hipótesis se emplean como una herramienta para medir la distancia que existe entre los valores reales y los imputados y no como una prueba formal y rigurosa de comparación de poblaciones.

3.5.1 Comparación de estadísticos marginales

Variables continuas

En este conjunto de pruebas, se establecen comparaciones entre los estimadores (media y la desviación estándar) para los individuos donantes y receptores. Entre las pruebas propuestas están:

3.5.1.1 Prueba de igualdad de medias

Hipótesis nula: $H_0 : \overline{X}_{di} - \overline{X}_{ri} = \delta_0$ i = 1,. . . , p (número de variables comunes)

$\overline{X}_d$ = Vector de medias en los donantes

$\overline{X}_r$ = Vector de medias en los receptores

Estadístico de prueba: $$t = \frac{(\overline{X}_{di} - \overline{X}_{ri}) - \delta_0}{s} \qquad (3.22)$$

La desviación estándar muestral s, de $\overline{X}_{di} - \overline{X}_{ri}$ depende de la suposición que se haga de las varianzas (varianzas iguales o diferentes). δ_0 representa la diferencia hipotética entre las dos medias poblacionales.

3.5.1.2 Prueba de igualdad de varianzas

La consideración de la igualdad de varianzas, depende de la distribución de las poblaciones de donde provienen las muestras. En el caso de poblaciones normales se tiene:

Hipótesis: $$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_a : \sigma_1^2 \neq \sigma_2^2$$

Estadístico de prueba: $F_0 = \frac{s_1^2}{s_2^2}$ con n_1-1 G.L. en el numerador (3.23)

y n_2-1 G.L. en el denominador

La alternativa para el caso de no Normalidad es la prueba de Levene.

3.5.1.3 Prueba conjunta de igualdad de medias

Sean $\mathbf{X_1, X_2, \ldots, X_n}$ observaciones de una población con media $\boldsymbol{\mu}$ y matriz de varianza – covarianza $\boldsymbol{\Sigma}$ (no singular).

Hipótesis: $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$
$Ha : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$

Estadístico de prueba: $T^2 = N_r\left(\overline{\mathbf{X}}_r - \boldsymbol{\mu}_d\right)'\mathbf{S}^{-1}\left(\overline{\mathbf{X}}_r - \boldsymbol{\mu}_d\right) \sim \chi_p^2$ (3.24)

$$P\left(T^2 \leq \chi_p^2(\alpha)\right) = 1 - \alpha \quad (3.25)$$

3.5.1.4 Intervalos de confianza para la media de las variables.

Es posible entonces, conociendo los valores de las medias y sus desviaciones estándar, establecer intervalos de confianza para cada una de las variables. Se dispones de las siguientes construcciones:

- Intervalos de confianza individuales. No considera la matriz de varianzas-covarianzas.

$$\overline{X}_i \pm t_{n-1}\left(\frac{\alpha}{2}\right)\sqrt{\frac{s_{ii}}{n}} \quad (3.26)$$

- Intervalos de confianza de Bonferroni. Considera la matriz de varianzas-covarianzas.

$$\overline{X}_i \pm t_{n-1}\left(\frac{\alpha}{2p}\right)\sqrt{\frac{s_{ii}}{n}} \quad (3.27)$$

- Intervalos de confianza T^2. Considera la matriz de varianzas-covarianzas.

$$\overline{X}_i \pm \sqrt{\alpha\,\chi_p^2}\sqrt{\frac{s_{ii}}{n}} \quad (3.28)$$

En esta propuesta se ha incluido la construcción de intervalos de confianza de Hotelling.

Las pruebas hasta ahora propuestas pertenecen a las llamadas pruebas de nivel fijo. Se propone la hipótesis nula junto con un nivel, generalmente 0.05. Se listan los posibles resultados del experimento para identificar los resultados extremos que podrían ocurrir menos del 5% de las veces si la hipótesis nula fuera cierta. A este conjunto de valores se les llama región crítica, porque si alguno de ellos ocurre, algo extremo ha pasado. Si esto llega a pasar, se dice que los resultados son significativos al nivel de 0.05. Se rechaza la hipótesis nula con un nivel de significancia de 0.05.

El nivel de significancia observado o valor P, es el menor nivel fijo para el cual la hipótesis nula puede ser rechazada. Si el nivel fijo es mayor o igual que el valor P, se rechaza la hipótesis nula.

Con frecuencia se describe el valor P como la probabilidad de ver resultados tan o mas grandes que aquellos realmente observados si la hipótesis nula fuera cierta.

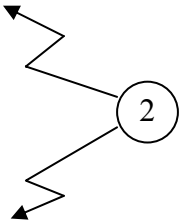
Se desea entonces generar un índice de significancia promedio para todas las variables, lo cual se puede conseguir con la idea de media geométrica.

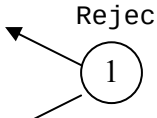
Por cada prueba de comparación de medias, se obtiene el valor P correspondiente y por lo tanto la probabilidad máxima de no rechazo (1-P). Promediando de manera geométrica los distintos valores de (1-P) obtendríamos en promedio la región de rechazo a través del complemento con 1, lo que representaría el nivel promedio de significancia ASL.

Este nivel promedio de significancia puede entonces ser comparado con el nivel fijo de la prueba en un nivel agregado para establecer el rechazo o no de la hipótesis nula. Adicionalmente, para un valor fijo del nivel de la prueba, se puede contabilizar cuantas pruebas fueron rechazadas para generar un índice de porcentaje de rechazo.

Se muestra a continuación un ejemplo de estos índices para las variables específicas (3 en este caso) con un valor límite para el nivel de significancia de la prueba de 1.00 (100%). Este valor se emplea para determinar cuántos valores P son menores que el límite establecido. De acuerdo a estos resultados, se puede establecer que en promedio la hipótesis nula de igualdad de medias y de desviación estándar se acepta (1), para niveles fijos de significancia menores de 0.05, aunque de manera individual existe una prueba significativa en las medias y en las desviaciones estándar (2). En este caso el porcentaje de rechazo es del 100% debido al nivel fijo de la prueba establecido inicialmente. Estos mismos resultados se presentan para las variables comunes.

```
   ******   SPECIFIC VARIABLES   ******

      ***   Continuous variables   ***

                          Donors
Var.          Mean    Std.Dev.      minimum      MAXIMUM

V1        1.450666    0.261024     0.421000     2.766000
V2       -6734.235    13563.52    -57311.19     86265.29
V3        0.756348    0.090292     0.395000     1.041000

                         Receivers
Var.          Mean    Std.Dev.      minimum      MAXIMUM

V1        1.452960    0.260651     0.624000     2.604000
V2       -6714.596    12930.12    -56006.69     86688.00
V3        0.758688    0.088769     0.397000     1.056000

                     Mean Test
Var.     t-Student     P-Value   t interval

V1         1.406764    0.080041      0.012350
V2         0.061835    0.475360     612.64086
V3         12.17437    0.000000      0.004206

              Std. Dev. Test
Var.       F-Fisher     P-Value

V1         1.002861    0.466421
V2         1.100372    0.002395
V3         1.034613    0.157755

Mean Test:
Hotelling's T2: 4.788712  P-Value: 0.187939
Average Significance Level: 0.215589     Rejection Level: 100.00%

Str. Dev. Test:
Average Significance Level: 0.234641     Rejection Level: 100.00%
```





Variables categóricas

En este caso, se puede determinar de manera gráfica la comparación de las variables empleando histogramas de frecuencia, y a partir de ahí, el desarrollo de una prueba de homogeneidad χ^2. El concepto de ASL se aplica de la misma forma en este caso.

A continuación se presenta un ejemplo de este tipo de indicadores para el caso de variables categóricas.

Se presenta el caso de 3 variables categóricas (V4, V5, V6) con 2, 6 y 6 modalidades respectivamente. Se observa al final que en promedio la distribución de las variables en el archivo de los donantes y los receptores son iguales.

```
      ***   Categorical variables   ***

     (=) Donors   (+) Receivers

V4 -
 AA_1 2005 |============================================================
      1982 |++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
 AA_2 1476 |=========================================
      1498 |+++++++++++++++++++++++++++++++++++++++++

Chi2: 0.591754 P-Value: 0.441742

V5 -
 AB_1  857 |========================
       831 |++++++++++++++++++++++++
 AB_2 1205 |==================================
      1258 |++++++++++++++++++++++++++++++++++
 AB_3  527 |===============
       499 |+++++++++++++++
 AB_4  355 |==========
       381 |++++++++++++
 AB_5  338 |==========
       334 |++++++++++
 AB_6  199 |======
       178 |+++++

Chi2: 8.775228 P-Value: 0.118372
```

```
V6 -
 AC_1    1 |
         1 |
 AC_2  180 |=====
       182 |+++++
 AC_3  716 |====================
       666 |++++++++++++++++++
 AC_4 1396 |======================================
      1413 |++++++++++++++++++++++++++++++++++++++
 AC_5 1065 |==============================
      1093 |++++++++++++++++++++++++++++++
 AC_6  123 |====
       126 |++++

Chi2: 4.530061 P-Value: 0.338998

Average Significance Level: 0.312234     Rejection Level: 100.00%
```

3.5.2 Homogeneidad

Al hablar de homogeneidad, se tratará de comparar las relaciones existentes entre las variables en los conjuntos de datos de los donantes y los receptores. Estas relaciones se presentan entre variables específicas (Homogeneidad interna) o entre variables comunes y específicas (Homogeneidad externa).

En los dos casos, el establecimiento de homogeneidad se presentará entre variables continuas; entre variables categóricas; entre variables continuas y categóricas.

El esquema de medición de la homogeneidad queda de la siguiente manera:

- Homogeneidad Interna
  - variable continua-variable continua
  - variable categórica-variable categórica
  - variable continua-variable categórica

- Homogeneidad Externa
  - variable continua-variable continua
  - variable categórica-variable categórica
  - variable continua-variable categórica

La medición de la homogeneidad se puede desarrollar a partir de sus correlaciones. Se tiene entonces:

Hipótesis nula: $H_0: \rho = \rho_0$

ρ_0 = valor del coeficiente de correlación poblacional. Se utiliza la transformada de Fisher para obtener un estadístico $z \sim N(0, 1)$. Con este estadístico se podrá determinar si se acepta o no la hipótesis nula planteada. Por lo tanto, a partir del nivel fijo de significancia, se puede calcular el porcentaje de rechazo para las pruebas que se desarrollen. Con el valor P de cada una de las comparaciones, se puede calcular el nivel promedio de significancia (ASL) descrito anteriormente. Adicionalmente se puede establecer la diferencia entre las correlaciones observadas en el archivo de donantes con las observadas en el archivo de receptores y calcular la diferencia promedio en valor absoluto ACD. Este conjunto de diferencias tendrá variación, la cual se puede utilizar para establecer un intervalo de confianza.

En el caso de las relaciones existentes entre variables categóricas, se emplean conceptos relacionados con el análisis de correspondencias simples. Es una técnica de ordenación, inicialmente propuesta por Hirschfeld (1935). El Análisis de Correspondencias se conoce también como "reciprocal averaging", debido a que el algoritmo para encontrar la solución involucra el promediado repetido en renglones y columnas. Las coordenadas de renglones y columnas en el primer eje, se asignan de tal manera que la correlación ponderada entre los dos se maximiza. El valor propio del primer eje es equivalente al coeficiente de correlación entre las coordenadas de renglones y columnas en el primer eje (Gauch 1982, Pielou 1984). No es posible arreglar renglones y/o columnas de tal manera que la correlación sea mayor.

De esta manera se busca posicionar todos los puntos (donantes y receptores) sobre el primer eje a partir de las proyecciones de los puntos sobre el primer eje y con estas coordenadas determinar las correlaciones existentes entre las variables tanto en donantes como en receptores. Se puede por lo tanto, establecer una prueba de hipótesis como se mencionó anteriormente.

Para evaluar las correlaciones existentes entre variables de tipo continuo y variables de tipo categórico, se emplea una idea similar a la ordenación para el caso de las variables categóricas. La idea ahora es la de sustituir cada modalidad por el valor promedio de la variable continua asociada con esa modalidad en el conjunto global (donantes y receptores). De esta manera, el problema de la ordenación quedaría resuelto. Estos valores promedios se usarán entonces para la comparación de correlaciones mediante una prueba de hipótesis.

Ejemplo 3.6

Se presenta a continuación el caso de dos variables, una continua y una categórica con tres modalidades.

Tabla 3.11 Valores de 2 variables, Continua y Categórica

Variable Continua	Variable Categórica
2	1
1	3
3	1
4	2
1	1
3	2
2	2
4	1
5	1
7	3
4	1
2	2
3	1
4	3
2	3

El valor que se le asigne a cada modalidad determinará el valor de la correlación.

Tabla 3.12 Asignación de valor a las modalidades

	Modalidad		
	1	2	3
Valor asignado	1	2	3
	1	3	2
	2	1	3
	2	3	1
	3	1	2
	3	2	1

Calculando la correlación de forma convencional entre estas variables da como resultado:

Tabla 3.13 Resultado de la correlación en función del valor asignado

Valor asignado a las modalidades	123	132	213	231	312	321
Correlación	0.073	-0.083	0.177	-0.177	0.083	-0.073

La correlación máxima se presenta con la asignación de valores 2 1 3 a las modalidades. Se puede ver la correlación en función de la posición asignada a la modalidad.
Sustituyendo las modalidades por su correspondiente valor medio:

Tabla 3.14 Valor medio de cada modalidad

Modalidad	Promedio
1	3.14
2	2.60
3	4.00

se obtiene como resultado una correlación de 0.321. Observando estos valores, las modalidades ordenadas serían 2 1 3 que coincide con lo observado en la tabla 3.13.
Se presenta a continuación un ejemplo de estos índices una vez que se ha definido la mecánica para evaluar las correlaciones.

```
*********   H O M O G E N E I T Y   *********
   ******   INTERNAL HOMOGENEITY   ******
      ***   Continuous variables   ***

Correlation    Donors  Receivers      Z   P-Value

(V1  , V2  ) |  0.734   0.741    -0.683   0.247
(V1  , V3  ) | -0.459  -0.468     0.456   0.324
(V2  , V3  ) | -0.289  -0.296     0.293   0.385

Average Correlation Diference: 0.007494 ± 0.002156
Average Significance Level: 0.321068     Rejection Level: 100.00%

      ***   Categorical variables   ***
Correlation    Donors  Receivers      Z   P-Value

(V4  , V5  ) |  0.574   0.582    -0.535   0.296
(V4  , V6  ) |  0.332   0.333    -0.088   0.465
(V5  , V6  ) |  0.278   0.281    -0.136   0.446

Average Correlation Diference: 0.004471 ± 0.007149
Average Significance Level: 0.406996     Rejection Level: 100.00%

      ***   Continuous-Categorical variables   ***
Correlation    Donors  Receivers      Z   P-Value

(V1  , V4  ) |  0.762   0.769    -0.677   0.249
(V1  , V5  ) |  0.695   0.699    -0.350   0.363
(V1  , V6  ) |  0.444   0.448    -0.233   0.408
(V2  , V4  ) |  0.547   0.559    -0.704   0.241
(V2  , V5  ) |  0.936   0.937    -0.593   0.277
(V2  , V6  ) |  0.289   0.290    -0.040   0.484
(V3  , V4  ) |  0.347   0.350    -0.135   0.446
(V3  , V5  ) |  0.281   0.285    -0.163   0.435
(V3  , V6  ) |  0.950   0.950    -0.143   0.443

Average Correlation Diference: 0.004069 ± 0.006967
Average Significance Level: 0.377890     Rejection Level: 100.00%
```

3.5.3 Exactitud

Variables continuas

Para evaluar la exactitud en la estimación, en el caso de variables continuas, se mide la tasa de error Tau[9]:

$$\tau = \frac{\sum_i (y_{i0} - \hat{y}_{i0})^2}{\sum_i (y_{i0} - \bar{y}_{i0})^2} \qquad (3.29)$$

Se puede interpretar como el reciproco del coeficiente de determinación $1 - R^2$.

Compara el error en la estimación con respecto al error de imputación, si esta hubiera sido hecha sustituyendo los valores por la media.

El cálculo de este indicador se hace con la información completa disponible, información contenida en el archivo donante. Para esto, se repite el proceso de imputación sobre el archivo donante. Para cada individuo donante, se efectúa la imputación sobre la base de sus vecinos más cercano sin incluirlo a él mismo (cf. R1, Pág. 61, Cáp.2).

Variables categóricas

Se puede considerar al proceso de fusión como un problema de clasificación, en el cual los individuos son clasificados en diferentes grupos sobre la base de variables conocidas. Se puede entonces comparar los valores reales con los valores imputados. Para esto se disponen los resultados en una matriz llamada matriz de confusión.

	Imputados		
Reales	Modalidad 1	. . .	Modalidad p
Modalidad 1	n_{11}	. . .	n_{1p}
:	:	:	:
Modalidad p	n_{p1}	. . .	n_{pp}

Figura 3.9 Matriz de confusión

[9] cf. Aluja, Banet (1997)

n_{ii} representa el número de individuos con la modalidad i que fueron clasificados en esa modalidad (aciertos).

- $\frac{\text{Precisión}}{\text{modalidad}} = \frac{\text{aciertos}}{\text{valoresimputados(modalidad)}}$

  De los individuos clasificados con esa modalidad, ¿Que porcentaje de ellos tenían realmente esa modalidad?

- $\frac{\text{Sensitividd}}{\text{modalidad}} = \frac{\text{aciertos}}{\text{valoresreales(modalidad)}}$

  De los individuos con esa modalidad, ¿Que porcentaje de ellos fueron clasificados correctamente?

- Exactitud por variable = $\frac{\text{Elementos en la diagonal}}{\text{Total individuos}}$ = Sensitividad promedio

  Representa el porcentaje de individuos clasificados correctamente.

- Exactitud global = promedio (Exactitud por variable)

Ejemplo 3.7

Variable categórica con dos modalidades

Tabla 3.15 Valores reales e imputados

Y_0	$\hat{Y}_0$		Y_0	$\hat{Y}_0$
1	1		1	1
1	1		2	1
2	1		1	2
1	1		2	2
2	2		2	1
2	2		1	1

Tabla 3.16 Matriz de confusión

	$\hat{C}_1$	$\hat{C}_2$
C_1	5	1
C_2	3	3

```
Variable categórica C
  Precisión  Sensitividad

 C1   5/8       5/6
 C2   3/4       3/6
Sensitividad promedio:  8/12 Exactitud por variable
```

Se define además un indicador Tau' de la siguiente manera:

$$\tau' = \frac{1 - \frac{\text{tr(M.C.)}}{\text{Nl}}}{1 - \frac{\text{frec(moda Real)}}{N_1}} \tag{3.30}$$

El denominador indica el porcentaje de los individuos que no serán imputados por la moda real. Representa un valor constante. El numerador representa el porcentaje de los individuos imputados incorrectamente. Este valor puede variar desde 0 hasta 1. Por lo tanto el cociente tendrá un rango ≥ 0 siendo uno el caso para el cual, la imputación es equivalente a la hecha por la moda.

Un ejemplo del uso de estos indicadores es:

```
*********   A C C U R A C Y   *********
      ***   Continuous variables   ***

Variable    Tau

V1          0.443
V2          0.333
V3          0.560
Average Tau: 0.445399
```

```
      ***   Categorical variables   ***

V4 -
  Precision  Sensitivity

 V4_1  0.803  0.794
 V4_2  0.724  0.735
                   Average Sensitivity:  0.769 Tau':  0.545

V5 -
  Precision  Sensitivity

 V5_1  0.703  0.681
 V5_2  0.552  0.576

V5_3  0.295  0.279
 V5_4  0.252  0.270
 V5_5  0.389  0.385
 V5_6  0.657  0.588
                   Average Sensitivity:  0.508 Tau':  0.753

V6 -
  Precision  Sensitivity

 V6_1  1.000  1.000
 V6_2  0.434  0.439
 V6_3  0.520  0.483
 V6_4  0.514  0.520
 V6_5  0.517  0.531
 V6_6  0.214  0.220
                   Average Sensitivity:  0.501 Tau':  0.833

Global Sensitivity :  0.593
Average Tau': 0.710148
```

3.6 El sistema GRAFT

En el desarrollo del sistema GRAFT se pueden distinguir seis etapas:

- I Preproceso
- II Construcción del archivo base
- III Análisis factorial
- IV Clasificación
- V Búsqueda de vecinos
- VI Imputación
- VII Validación

Estas etapas corresponden a las etapas en un proceso de fusión. Las primeras cuatro tienen como finalidad, posicionar a los individuos donantes y receptores en un mismo espacio factorial en el cual se buscarán los vecinos más cercanos[10]. Posteriormente se efectuarán la búsqueda de vecinos (etapa V) y el proceso de imputación (etapa VI).

Las primeras cuatro etapas no forman parte del sistema presentado en esta tesis, corresponden a un trabajo previo que debe ser desarrollado por el usuario.

Se requiere de un conocimiento de los archivos que se van a fusionar, específicamente, de las variables que se van a manejar, ya que parte del trabajo previo consiste en la selección de las variables que serán seleccionadas como comunes, así como la determinación de aquellas variables que pueden servir como referencias para una búsqueda de vecinos con restricciones.

[10] a partir del análisis factorial y técnicas de clasificación

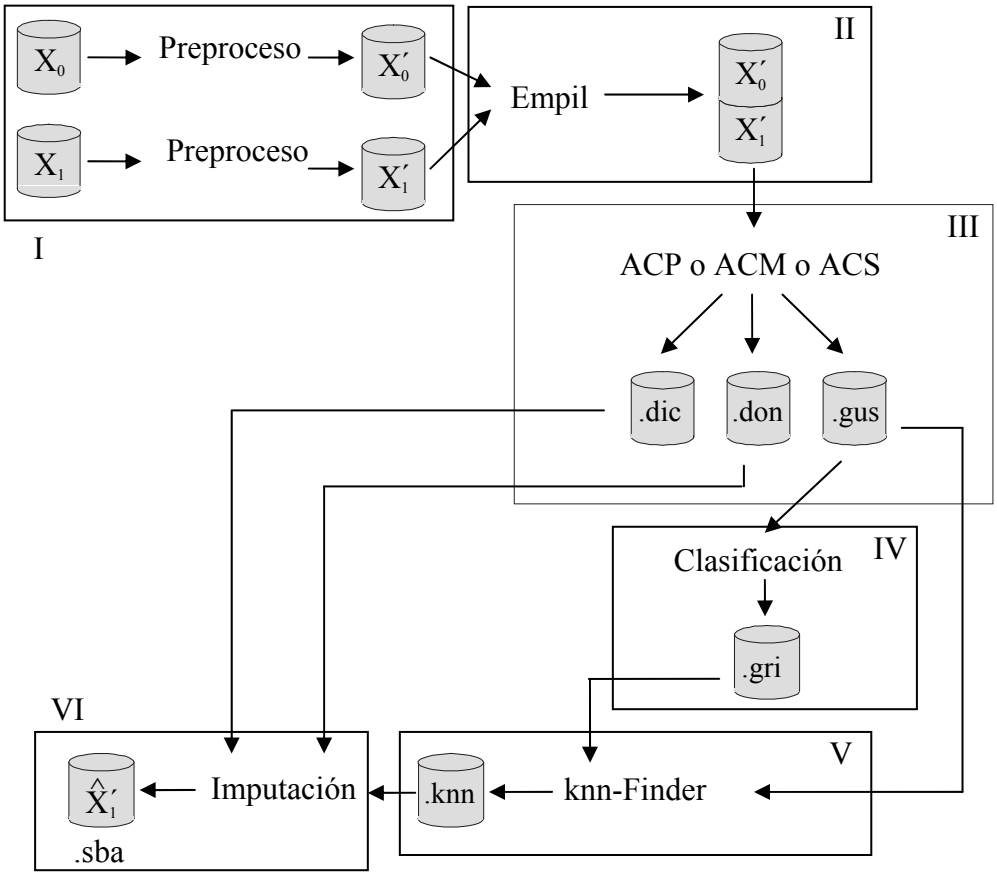


Figura 3.10 Diagrama del sistema GRAFT en el entorno de SPAD

Etapa I. Actividades de preproceso

Una de las primeras actividades en casi todos los procesos de manejo de datos, consiste en adecuar los elementos necesarios para la ejecución del proceso (requisitos y restricciones). Es lo que se puede llamar como actividades de preproceso, las cuales pueden incluir: búsqueda de valores atípicos de acuerdo al contexto de la información que se va a manejar, adecuación de los datos como materia de entrada para el(los) algoritmo(s), validación de condiciones para ejecución, etc. Estas actividades resultan importantes para la buena ejecución del proceso ya que permitirán analizar los resultados finales con relativa confianza.

Una de las actividades que se deben tener en cuenta es la ponderación de los individuos con respecto a la población.

Otra de las actividades de preproceso que resulta fundamental para el proceso de fusión, consiste en la búsqueda del subconjunto mínimo de variables más predictivo. Esto requiere por parte del usuario, de un conocimiento profundo de los archivos que se quieren fusionar y más puntualmente, de las variables que se van a manejar como conjunto de variables comunes. Resulta también importante para la selección de las variables que servirán de referencia para búsquedas con restricciones.

Otra actividad que se debe tener en cuenta en la etapa de preproceso es la llamada transformación de las variables. Es una actividad que se puede llamar de normalización de las variables (**ln**(x), √x, **logit**(x) [11]). Esta actividad está incluida en el sistema desarrollado.

Etapa II. Construcción del archivo base

En esta etapa se construye un solo archivo (archivo base o de referencia). El trabajo consiste en apilar el archivo de donantes ($\mathbf{X}_0^{'}$) y el archivo de receptores ($\mathbf{X}_1^{'}$) para formar un solo archivo. Este nuevo archivo, construido de esta manera, será manejado y procesado posteriormente (análisis factorial). El proceso encargado de esta etapa, llamado **EMPIL,** forma parte del paquete estadístico **SPAD**[12]. En el análisis factorial, se considerarán a los individuos donantes como individuos activos y a los receptores como ilustrativos. De la misma forma se consideran a las variables comunes como variables activas y a las variables específicas como variables ilustrativas. Estas últimas no afectan la construcción de los ejes factoriales, como tampoco influyen los individuos ilustrativos, pero su disposición en una gráfica aporta mucha información acerca de los datos que se están procesando.

Etapa III. Análisis factorial

Uno de los objetivos del análisis factorial es la reducción de dimensiones. Esto afecta directamente al desarrollo del sistema, debido a que la búsqueda de vecinos depende en gran manera del número de variables contenidas en el archivo (búsqueda basada en el

[11] $\log it(x) = \ln\left(\frac{x}{1-x}\right)$

[12] software de **Decisia**, socio del proyecto **ESIS**

cálculo de distancias). Al reducir la dimensionalidad del espacio considerado se reduce también el ruido.
El análisis factorial de **SPAD** genera los siguientes archivos de interés:

- File_Name**.don** archivo de datos que contiene a todos los individuos, refiriéndose a los donantes como activos y a los receptores como suplementarios.

- File_Name**.dic** archivo diccionario que contiene información sobre las variables registradas en el archivo File_Name**.don.**

- File_Name**.gus** archivo que contiene las coordenadas factoriales de todos los individuos registrados en el archivo File_Name**.don.**

Etapa IV. Clasificación

El algoritmo de búsqueda de vecinos utiliza un árbol jerárquico de búsqueda que se deberá construir previamente. Esta construcción se hace con el procedimiento **Classification** de **SPAD**. Tiene dos procedimientos para esto. El proceso **RECIP** (clasificación jerárquica directa), que construye un árbol de agregación jerárquica (dendrograma) de los individuos caracterizados por sus coordenadas factoriales utilizando el criterio de agregación de Ward (basado en la reducción mínima de varianza mediante agregación) que es compatible con el criterio de inercia utilizado para determinar los ejes factoriales. El árbol se construye mediante el algoritmo rápido de búsqueda en cadena de vecinos recíprocos o algoritmo de Benzecri. El otro procedimiento, **SEMIS**, reemplaza a **RECIP** cuando la tabla a analizar es de grandes dimensiones. La clasificación jerárquica se realiza empleando los ejes factoriales y la información del árbol generado se almacena en un archivo File_Name**.gri**. Este archivo contiene información acerca de los nodos generados, los individuos en cada nodo, la distancia más grande de cada punto en el nodo a la media, entre otra. Esta información se usará para reconstruir el árbol requerido por el algoritmo de Fukunaga/Narendra. Resulta importante la elección del número de niveles en el árbol porque esto afecta al procedimiento de búsqueda.

Etapa V. Búsqueda de vecinos

Esta etapa se desarrolla con el módulo llamado Knn-Finder, el cual implementa el algoritmo de búsqueda de vecinos de Fukunaga/Narendra (cf. 2.4.1). El costo de búsqueda para este algoritmo es del orden $O(nLog_2\ n)$. Esto se consigue utilizando un árbol jerárquico de búsqueda construido previamente. El modulo Knn-Finder requiere como es de esperarse, del número de vecinos que se desea localizar para cada receptor. Esta búsqueda puede ser condicionada al cumplimiento de restricciones impuestas sobre algunas variables comunes (cf. 2.4.3). El resultado de este proceso es la generación de archivos con información para el proceso de imputación.

- File.**knn** archivo binario que contiene para cada receptor, información sobre sus vecinos más cercanos: distancia a cada vecino, número de restricciones cumplidas, índice en la tabla de almacenamiento.

- File_knn.**txt** archivo texto con la información mencionada en el punto anterior.

La información almacenada anteriormente, también se almacena para cada individuo donante.

Etapa VI Imputación

Finalmente, con las etapas anteriores ejecutadas y la información obtenida, se puede proceder a la ejecución del proceso de imputación (cf. 3.2). El proceso de imputación se lleva a cabo con el módulo imputation.exe. Requiere para su ejecución del archivo File.knn, puesto que se requiere de la lista de los vecinos más cercanos y de sus distancias. Otra información requerida para este proceso es: File_Name**.don** y File_Name**.dic.** El resultado de este proceso son los archivos:

- File.**sba** base de datos completa en formato dentro del esquema de SPAD.

- File.**txt** base de datos completa en formato texto.

La información de estos archivos representan $(\hat{\mathbf{X}}'_1)$ en el diagrama contenido en la figura 3.9.

Etapa VII Validación

Esta etapa es ejecutada también por el módulo imputation.exe. Desarrolla los distintos criterios seleccionados para medir la calidad de la fusión (cf. 3.5). Genera como resultado los archivos:

- File.**lst** archivo en formato texto que contiene los resultados en forma detallada de las pruebas de validación implantadas.

- File.**lst**.**sum** archivo en formato texto que muestra un resumen de los resultados de la validación de la fusión.

Las etapas V, VI y VII (búsqueda de vecinos, imputación y validación) son los elementos desarrollados para el sistema GRAFT. El algoritmo de búsqueda de vecinos seleccionado para este desarrollo ha sido el de Fukunaga / Narendra. Como ya se mencionó, resulta ser un algoritmo fácil de implantar y bastante controlable. Requiere para su ejecución de la construcción previa de estructuras. Sin embargo, esta aparente desventaja se puede aprovechar, trabajando dentro de un entorno como el que proporciona el paquete estadístico SPAD. Dentro de este entorno se genera el árbol jerárquico (proceso de clasificación) para el procedimiento de Branch and Bound, esencial en el algoritmo de F/N. Puede parecer una desventaja el estar dentro de un entorno como el de SPAD, lo hace dependiente del paquete, sin embargo, existen ventajas, ya que se pueden desarrollar otras actividades previas al proceso de fusión con bastante simplicidad y que ahorran tiempo para desarrollar funciones equivalentes fuera de este entorno. Las distintas técnicas de imputación, están basadas en el concepto de espacio factorial y búsqueda de vecinos más cercanos. Esto remite de nueva cuenta a la idea de trabajar dentro de un entorno estadístico, pues las actividades relacionadas con el análisis factorial vienen ya incluidas. Aprovechando las caracteristicas del entorno, se pueden analizar los archivos antes de su manejo en el sistema GRAFT.