Señor… si me das éxito,
No me quites la humildad

Capítulo 1
- Fusión de Datos -

1.1 Introducción

Las actitudes, el conocimiento y las acciones generalmente se basan en muestras. Algunos basan sus conclusiones en muestras pequeñas y pocas veces toman en cuenta la magnitud de lo que se desconoce. Generalmente se carece de recursos para estudiar más de una parte del problema de interés que pudiera aumentar nuestro conocimiento. Algunas razones para el uso de las técnicas de muestreo son: costo reducido, mayor velocidad, mayor enfoque o perspectiva y mayor exactitud.[1]

El muestreo ha venido a jugar un papel importante en los censos. En Estados Unidos, se introdujo una muestra del 5% en el censo de 1940 con preguntas adicionales acerca de la ocupación, fertilidad etc. El uso del muestreo se extendió ampliamente en 1950. El proceso continuó en los censos de 1960 y 1970. El censo total se levantó con base en muestras excepto por cierta información básica requerida de cada persona por razones legales o constitucionales.

Algunas oficinas de gobierno hacen uso del muestreo para obtener información actual, como los estudios de la población actual que proporcionan datos mensuales de la composición y tamaño de la fuerza laboral.

La investigación de mercado depende altamente del enfoque de muestreo. Continuamente se mantiene bajo escrutinio la estimación de las audiencias en radio y televisión para diferentes programas y lectores de periódicos y revistas (incluyendo los anuncios). A los productores y distribuidores les interesa conocer la reacción de la gente a nuevos productos o métodos de empaque o la razón para la preferencia de un producto sobre otro. Son populares las encuestas de opinión, de actitud y en las elecciones.

[1] Cochran, W.G. (1977)

Cochran clasifica la investigación por muestreo de manera amplia en: descriptiva y analítica. En el caso descriptivo, el objetivo es simplemente obtener cierta información acerca de grupos grandes (por ejemplo, número de hombres, mujeres y niños que vieron determinado programa de televisión). En el caso analítico, se hacen comparaciones entre diferentes subgrupos de la población para descubrir si existen diferencias entre ellos y para formar o verificar hipótesis acerca de las razones de esas diferencias.

Toda la información contenida en las encuestas ayuda a entender mejor a una población, sin embargo, las encuestas tienen algunas limitaciones como son la falta de respuestas, el número máximo de preguntas (no deben ser demasiado grandes) y el costo que puede llegar a tener.

La fusión de datos surge como una alternativa a la fuente única de datos frente a la necesidad de conseguir el máximo de información posible al menor costo. 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. Es importante hacer notar que no se esta generando nueva información.

Son muchos los motivos que pueden llevar a generar información. Por mencionar algunos, puede ayudar a controlar, optimizar, administrar, examinar, investigar, planificar, predecir, negociar o tomar decisiones según sea el área de interés.

La información por si misma es un bien patrimonial, de manera que deber ser protegida, pero también explotada. Debido al desarrollo tecnológico tanto en el área computacional como en la de transmisión de datos, ha aumentado espectacularmente la capacidad de manipular y almacenar la información. Es tal la información que nos llega que resulta difícil asimilarla.

La integración de datos a partir de diferentes fuentes es y ha sido tópico de interés en distintas áreas de la investigación. Se requiere información confiable y oportuna.

Existen distintos procedimientos desarrollados para enfrentar el problema de la fusión de datos. El objetivo principal es el de proporcionar información conjunta sobre variables observadas en distintas fuentes.

La fusión de datos es esencialmente multidisciplinaria y está en el cruce de diferentes ciencias. Reúne un gran número de métodos y herramientas matemáticas.

No es específica de un área específica de conocimiento o una aplicación.
En el área de la estadística, los orígenes de la fusión de datos son los estudios de mercado, especialmente las encuestas de consumo y encuestas de medios.

"La fusión de fuentes de datos, es un conjunto de técnicas desarrolladas a partir de los años 80 en el dominio de los estudios de medios".[2]

Alguna de las razones de la fusión de datos:

- Reducción de costos en las encuestas.

- Limitación en la recolección de datos

- "Diversificación transversal de mercados que lleva a buscar un conocimiento del comportamiento global del consumidor y no más a estudiar un mercado específico" [7]

El objetivo de la fusión de datos es el de utilizar lo mejor de la información existente en un archivo para reconstruir la información ausente en otro archivo. La idea es estimar los valores de las variables no informadas (valores faltantes) a partir de un bloque de variables informadas correlacionadas con el bloque de variables a reconstituir.
Sus aplicaciones son variadas:

- Estudios de mercado
  Paneles de medición de audiencia televisiva, estudios de consumo de determinados productos. Resultado: una sola fuente de datos que proporciona información detallada de audiencia televisiva para los usuarios de productos y servicios (Figura 1.1).

- Encuestas
  Las encuestas a nivel nacional, permiten enriquecer las encuestas locales.

[2] Lebart L. ;Lejeune M. (1995)

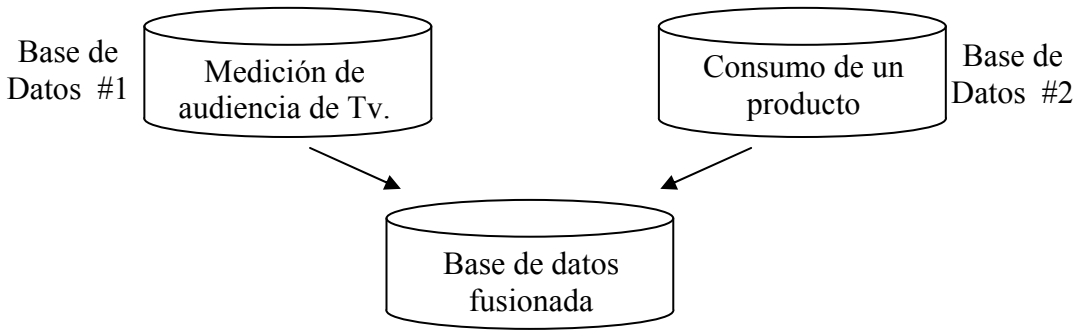


Figura 1.1 Ejemplo de fusión

- Estudios de clientela

  Tiendas departamentales, bancos, etc. enriquecen sus archivos de clientes (receptores), efectuando sondeos (de calidad en el servicio, por ejemplo) en opinión de ciertos clientes (donantes).

Como toda técnica, tiene ventajas y desventajas. Entre las ventajas, se puede mencionar la obtención de resultados a tiempo y la reducción de carga de la respuesta. Se ha demostrado que cuanto más grande es el cuestionario, mayor es la tasa de no respuestas y menor es la exactitud de las respuestas. Se puede por lo tanto considerar a cada archivo individual más confiable y completo que un subconjunto completo de datos.
Entre las desventajas está el hecho de que el procedimiento se basa en la consideración de la relación entre las variables que se van a fusionar. Se requiere por lo tanto de un cierto conocimiento previo.

La fusión de datos abarca situaciones variadas teniendo una base común: dos muestras distintas son tomadas de un mismo universo y por cada muestra se dispone de un cierto número de variables comunes y de variables específicas (Lebart L. Lejeune M 1995). El método consiste en transferir variables de una muestra llamada donante (D) hacia otra llamada receptora(R).

Las operaciones en este método son:

- Inyección.- D y R son dos submuestras de una misma encuesta.

- Fusión unilateral.- D y R son extraídas de dos encuestas distintas. Comparten un cierto número de variables.

- Fusión recíproca.- Cada muestra es donante para la otra en lo que se refiere a las variables específicas.

En esta tesis se restringirá el trabajo de investigación a la fusión unilateral, dado que la fusión recíproca se puede reducir a dos fusiones unilaterales en un proceso secuencial..

Se tiene por lo tanto:

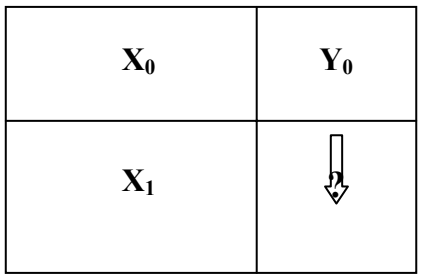


Figura 1.2 Formalización del problema

X representa las variables comunes. El primer archivo contiene observaciones para un conjunto de (p + q) variables tomadas sobre N_0 unidades. El segundo archivo contiene observaciones solo de un subconjunto de p variables sobre N_1 unidades. En este caso, el objetivo es llenar la parte vacía en el diagrama, lo cual representa un tipo especial de estimación de datos ausentes, debido a que no fueron recolectados. La parte desconocida **Y** del segundo archivo, se va a predecir usando el par $(\mathbf{X_0}, \mathbf{Y_0})$. De esta manera, se llamará al primer archivo, donante y al segundo archivo, receptor.

Algunas definiciones en este contexto:

- Variables comunes: variables presentes en las dos muestras, donantes y receptoras.

Dos tipos de variables comunes:

- críticas: subconjunto de las variables comunes que dan lugar a valores idénticos entre los donantes y los receptores.
- activas: utilizadas para la transferencia sobre las restricciones impuestas por las variables críticas.

- Variables específicas: variables de la muestra donante que van a reconstruir la muestra receptora.

Básicamente se consideran dos esquemas para la transferencia de las variables específicas al archivo receptor de acuerdo al modelo de distribución de las variables del archivo donante; - explícito o implícito- [3].

1.2 Tratamiento de los datos incompletos

Los métodos estadísticos estándares se han desarrollado para analizar conjuntos de datos rectangulares. Las filas representan unidades, casos o sujetos dependiendo del contexto y las columnas representan variables medidas para cada unidad. El problema de los datos faltantes presenta entonces el análisis de estas matrices de datos en las cuales algunas celdas no tienen observaciones. La falta de respuesta en una encuesta se puede deber a que el individuo no ha podido o no ha querido responder. Se dice que hay datos no completos. La fusión de datos es un caso particular de análisis de datos no completos (figura 1.3b y 1.3e).

Se pueden mencionar algunas diferencias básicas entre el problema de los datos faltantes como problema general y el problema de la fusión de datos en particular.

- En el caso del problema de los datos faltantes, se maneja una sola fuente de datos. En el caso de la fusión, se combinan datos de diferentes fuentes para poder disponer de toda la información en un solo archivo.

[3] G. Santini, 1986

- En el problema de los datos faltantes, se analizan los datos en una forma global y en el caso de la fusión, el interés radica en las respuestas de los individuos.

- En el caso del problema de los datos faltantes, se estudian las causas por las cuales ocurren los datos faltantes y se proponen métodos para su inferencia. En el caso de la fusión, se habla de ausencia de datos en bloques completos. Se puede hablar de datos ausentes por diseño.

Resulta útil distinguir el patrón de ausencia de datos y el mecanismo de ausencia de datos, debido a que algunos métodos de análisis dependen de estos patrones (Little R, Rubin D.B. 2000).
M representa la matriz indicadora de ausencia de datos con $m_{ij} = 1$ si el valor es ausente y 0 si está presente, y define el patrón de ausencia de datos.

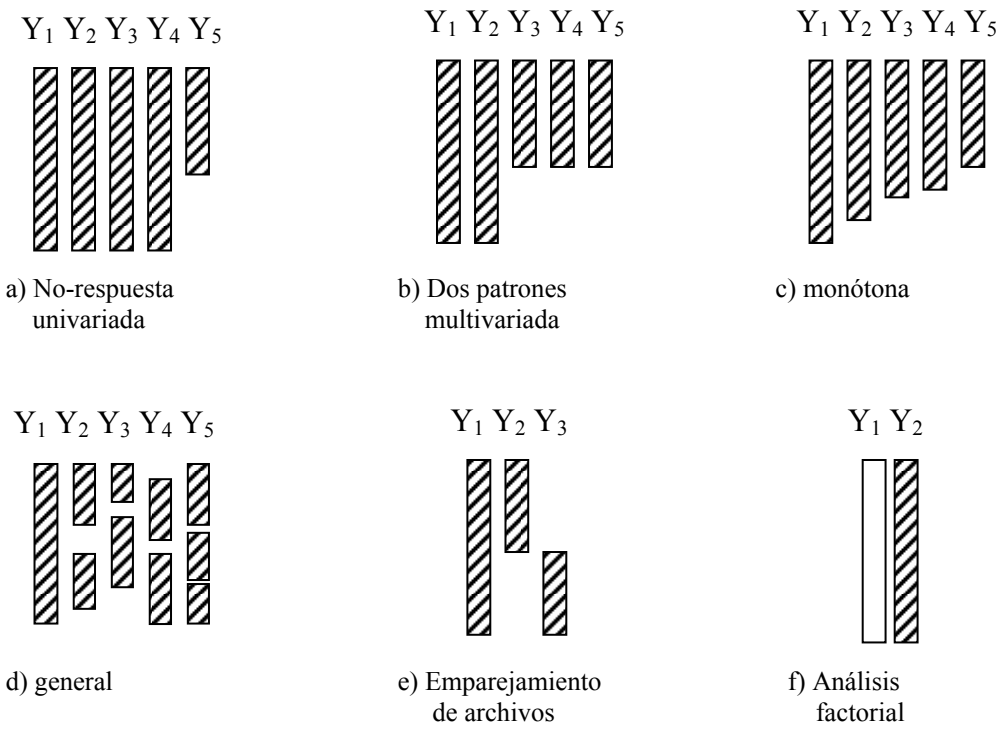


Figura 1.3 Patrones de ausencia de datos

El papel importante del mecanismo de ausencia dejó de ser ignorado hasta que el concepto fue formalizado en la teoría de Rubin.

Si: $f(M|Y,\phi) = f(M|\phi)$ para todo Y, ϕ MCAR (missing completely at random)

Si la probabilidad de que el valor de una variable Y_j sea observado para un individuo i no depende ni del valor de esa variable, y_{ij}, ni del valor de las demás variables consideradas, y_{ik} $k \neq j$.

Si: $f(M|Y,\phi) = f(M|Y_{obs},\phi)$ para todo Y_{aus}, ϕ MAR (missing at random)

Si la probabilidad de que el valor de una variable Y_j sea observado para un individuo i no depende del valor de esa variable, y_{ij}, pero tal vez del que toma alguna otra variable observada y_{ik} $k \neq j$.

Si: el mecanismo de ausencia depende de y_i, NMAR (not missing at random)

Si la probabilidad de que un valor y_{ij} sea observado depende del propio valor y_{ij}.

Ejemplo 1.1[4]:

K = 2, Y_1 = Edad, Y_2 = Ingreso

- Si la probabilidad de que el Ingreso sea faltante es la misma para todos los individuos, sin considerar su Edad o Ingreso, entonces los datos son MCAR.

- Si la probabilidad de que el Ingreso sea faltante varía de acuerdo a la edad del respondiente, entonces los datos son MAR.

- Si la probabilidad de que el Ingreso se registre varía de acuerdo al ingreso para aquellos con la misma edad, entonces los datos son NMAR.

[4]Fuente: Little/Rubin, (2002),pag. 16

Los métodos para tratar los datos incompletos pertenecen a una de las siguientes categorías: (taxonomía de los métodos de datos faltantes -Little R, Rubin D.B. 2000).

- Procedimientos basados en unidades registradas completamente.
  Cuando algunas variables no se registran para algunos individuos, se descartan y se analizan los registros completos. Puede dar buenos resultados con cantidades pequeñas de datos faltantes.

- Procedimientos de ponderación.
  Estos procedimientos modifican los pesos para ajustar las no respuestas como si fueran parte del diseño de muestreo. En un proceso sin datos faltantes, los pesos de diseño son inversamente proporcionales a su probabilidad de selección. Se trata de disminuir el sesgo.

- Procedimientos basados en imputaciones.
  Los datos faltantes se llenan y los datos completos son analizados por métodos estándares.

- Procedimientos basados en modelos
  Se define un modelo para los datos observados y se basa la inferencia en la verosimilitud o distribución posterior bajo ese modelo, con parámetros estimados por procedimientos como máxima verosimilitud.

Las imputaciones son extracciones de una distribución predictiva de los datos faltantes y requiere de un método para crear una distribución predictiva para la imputación basada en datos observados. [5]

La imputación es un método general y flexible para manejar problemas de datos faltantes. Sin embargo, tiene fallas. En palabras de Dempster y Rubin (1983):

"La idea de la imputación es seductora y peligrosa. Seductora porque puede adormecer al usuario en un estado placentero de creer que los datos están completos después de todo y esto es peligroso porque junta situaciones en donde el problema es suficientemente menor

[5] Little/Rubin, (2002)

que puede ser legítimamente manejado en esta forma y situaciones en donde estimadores estándares aplicados a los datos reales e imputados tienen sesgos substanciales ".

Hay dos enfoques genéricos para los métodos de imputación:

- **Modelos explícitos**
  - Métodos de imputación por la media
    - Media incondicional
    - Media Condicional
  - Métodos de imputación por regresión
    - Regresión Lineal por mínimos cuadrados
    - Regresión estocástica
    - Regresión PLS (y PLS estocástica)
  - Métodos de imputación por maximización de verosimilitud
    - Algoritmo EM (y EM estocástico)
    - Algoritmo de aumento de datos
- **Modelos implícitos**
  - Métodos de imputación Hot-deck
    - Muestreo aleatorio simple
    - Por clase
    - Secuencial
    - Vecino más cercano

1.2.1 Modelos explícitos

La distribución predictiva se basa en un modelo estadístico formal y por lo tanto las suposiciones son explícitas.

1.2.1.1 Métodos de imputación por la media

Los valores faltantes se sustituyen por las medias de las unidades observadas.

a) Media incondicional[6]

Una forma particularmente simple de imputación estimando la media de los valores observados.

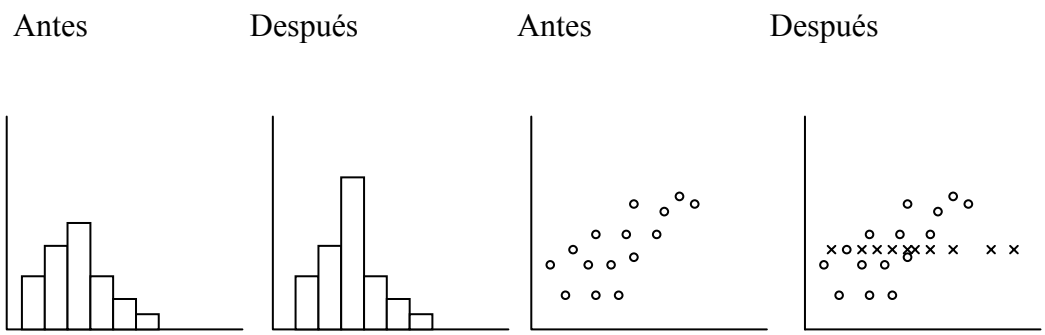


Figura 1.4 Características de la imputación por medias

Algunas características:

- Se conserva la media
- Distorsiona algunas características de la distribución (varianza, cuantiles)
- Distorsiona las relaciones entre las variables

b) Media condicional

Imputa medias condicionadas a valores observados. Un método común consiste en agrupar los valores observados y no observados en clases ajustadas e imputa los valores faltantes de los valores observados en la misma clase.

[6] Schafer, J.L. (2000)

1.2.1.2 Métodos de imputación por regresión

Los valores faltantes se sustituyen por valores predichos a partir de una regresión sobre los valores observados.

A partir de X = conjunto de variables predictoras (independientes) y Y = conjunto de variables explicadas (dependientes), se tiene:

a) Regresión Lineal por mínimos cuadrados

$$\hat{y}_i = \tilde{\beta}_0 + \sum_{j=1}^{r} \tilde{\beta}_j x_j \qquad \text{en donde:} \qquad \tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) \quad (1.1)$$

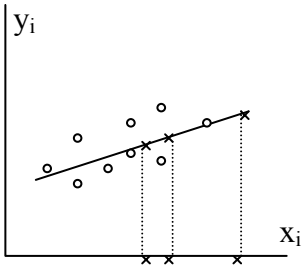


Figura 1.5 Imputación por regresión

- Sobreestima las correlaciones

b) Regresión estocástica

$$\hat{y}_i = \tilde{\beta}_0 + \sum_{j=1}^{r} \tilde{\beta}_j x_j + \mathbf{Z} \quad (1.2)$$

en donde **Z** representa el residual distribuido normalmente N (0, S^2); $S^2 = MS_E$ [7] (La variabilidad en una regresión: $SS_T = SS_R + SS_E$)

La imputación por regresión emplea las observaciones con valores completos de (**X, Y**) para ajustar modelos de regresión de Y sobre X y usar estas regresiones para imputar los valores faltantes.

---

[7] $SS_E = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$ $\quad E(SS_E) = \sigma^2 (n - p)$ $\quad \hat{\sigma}^2 = \frac{SS_E}{n - p}$ $\quad$ n-p = Grados de Libertad

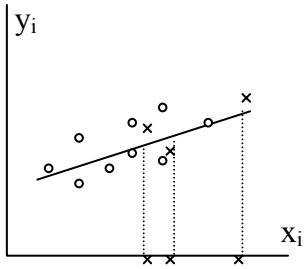


Figura 1.6 Regresión estocástica

Supone que permanece constante la relación de las variables predictoras X y las de respuesta Y en los dos archivos. Se puede considerar además:

- Regresión simple tomando la variable más correlacionada.
- Regresión múltiple tomando el mejor subconjunto de variables predictoras.
- Análisis de varianza (caso particular de la regresión en donde la variable predictora X es categórica y la variable de respuesta es cuantitativa.
- Modelo lineal generalizado para distintos tipos de variables de respuesta.

c) Regresión PLS[8]

El objetivo es descomponer la tabla X orientada a la explicación de la tabla Y.
Si **Y** = variables de respuesta para un conjunto de datos y **X** = variables predictoras, la regresión PLS trata de encontrar dos conjuntos de pesos, **w** y **c** para crear una combinación lineal de las columnas de **X** y **Y** respectivamente, de tal manera que su covarianza sea máxima.
El algoritmo PLS está diseñado específicamente alrededor de las sustituciones interdependientes $\mathbf{u}_1 \Rightarrow \mathbf{t}_1$ y $\mathbf{t}_1 \Rightarrow \mathbf{u}_1$ en una forma iterada hasta que ocurra la convergencia. En la convergencia, se han calculado un conjunto final de vectores (**t** , **w**)

[8] cf. Tenenhaus M. (1998)

y los correspondientes (**u** , **c**) para el h-ésimo componente PLS para el espacio **X** e **Y** respectivamente.
Se desea expresar la regresión de **Y** sobre los componentes PLS $t_1,\ldots, t_h$ en función de las variables X.

La regresión PLS descompone las matrices **X** y **Y** como un producto de un conjunto común de factores ortogonales y un conjunto de cargas específicas (loadings). De esta manera, las variables independientes se descomponen como **X** = **TP**' con **T'T** = **I**.

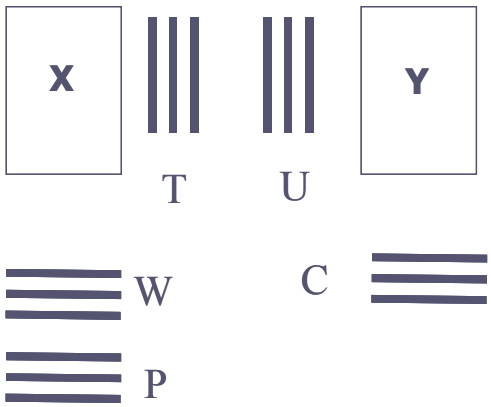


Figura 1.7 Esquema PLS

Se tiene:

$$\mathbf{X} = \mathbf{TP}' \qquad \mathbf{Y} = \mathbf{UC}' \tag{1.3}$$

X separada en dos matrices; **T** (Factor scores) **P** (loadings) coeficientes de regresión de **X** sobre **T**.

Y separada en dos matrices; **U** (Factor scores) **C** (loadings) coeficientes de regresión de **Y** sobre **U**.

Las columnas de **T** son los vectores latentes. Cuando el número de vectores latentes es igual al rango de **X**, desarrollan una descomposición exacta de **X**.

Específicamente, se trata de obtener un primer par de vectores **t** = **Xw** y **u** = **Yc** con las restricciones **w'w** = 1, **t't** = 1 y **t'u** máximo.

Cuando se obtiene el primer vector latente, se elimina mediante deflacción de **X** y **Y** y el proceso continua hasta que la matriz **X** sea nula.

Una regresión usando extracción de factores da como resultado:

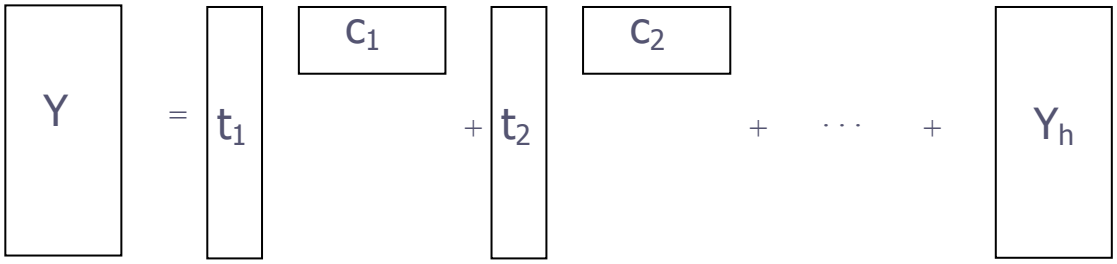
$$\mathbf{T} = \mathbf{XW} \qquad \text{para una matriz adecuada de pesos W} \tag{1.4}$$

Un modelo de regresión lineal

$$\mathbf{Y} = \mathbf{TC'} + \mathbf{R} \tag{1.5}$$

C = matriz de coeficientes de regresión de **Y** sobre **T**

R = error o ruido



$$Y = t_1 c_1^{'} + t_2 c_2^{'} + \ldots + t_h c_h^{'} + Y_h \tag{1.6}$$

Se tiene de manera simplificada:[9]

$$\mathbf{T} = \mathbf{XP'^{-1}} = \mathbf{XI_pP'^{-1}} = \mathbf{XWW^{-1}P'^{-1}} = \mathbf{XW(P'W)^{-1}} \tag{1.7}$$

Entonces:

$$\mathbf{Y} = \mathbf{XW(P'W)^{-1}C'} + \mathbf{R} \tag{1.8}$$

Una vez calculada **C'**, el modelo es equivalente a:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{R} \tag{1.9}$$

De donde

$$\mathbf{B} = \mathbf{W(P'W)^{-1}C'} \tag{1.10}$$

[9] La demostración detallada se puede consultar en: (Tenenhaus M. 1998)

De esta forma el modelo original planteado:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{Y}_h \qquad (1.11)$$

permite predecir los valores faltantes (imputación)

d) Regresión PLS estocástica

En este caso, al final de la estimación se agrega un componente aleatorio distribuido normalmente como se hizo en (1.2).

1.2.1.3 Métodos de imputación por maximización de verosimilitud

Son métodos que tienen como objetivo, estimaciones Máximo Verosímiles de los parámetros de una distribución cuando hay datos faltantes. Utilizando un valor inicial del vector de parámetros, rellena los datos faltantes mediante valores estimados. Con los datos completos, se obtiene una nueva estimación del vector de parámetros y con la nueva estimación del vector de parámetros se vuelven a llenar los datos faltantes. Esto se repite hasta lograr convergencia.

Para un conjunto de datos $X = (X_{obs}, X_{aus})$ en donde X_{obs} indica los valores observados y X_{aus}, los valores ausentes, estableciendo que los datos han sido generados según un modelo descrito mediante la función de probabilidad o densidad conjunta de X_{obs} y X_{aus} con un vector de parámetros θ, se tiene:

$$f(X|\theta) = f(X_{obs}, X_{aus}|\theta) \qquad (1.12)$$

La densidad de probabilidad marginal de X_{obs} se obtiene integrando con respecto a los valores ausentes X_{aus}:

$$f(X_{obs}|\theta) = \int f(X_{obs}, X_{aus}|\theta) dX_{aus} \qquad (1.13)$$

Se define la verosimilitud de θ basada en datos X_{obs} ignorando el mecanismo de ausencia como una función de θ proporcional a $f(X_{obs}|\theta)$:

$$L(\theta|X_{obs}) \propto f(X_{obs}|\theta) \qquad \theta \in \boldsymbol{\Omega}_\theta \qquad (1.14)$$

Se obtienen inferencias Bayesianas para θ ignorando el mecanismo de ausencia basadas en X_{obs}, incorporando una función de probabilidad a priori p(θ) y basando la inferencia en la distribución a posteriori: $p(\theta|X_{obs}) \propto p(\theta)\, L(\theta|X_{obs})$.

Suponiendo que el modelo propuesto sea el adecuado y que el mecanismo que genera la ausencia de datos sea ignorable, toda la información estadística relevante sobre θ está contenida en la función de verosimilitud de los datos observados, $L(\theta|X_{obs})$, o en la función de distribución a posteriori $p(\theta|X_{obs})$.

Estas funciones suelen ser bastante complicadas y es necesario utilizar alguna técnica específica para obtener las estimaciones máximo verosímiles.

a) Algoritmo EM

Un enfoque general para calcular estimadores de máxima verosimilitud a partir de datos incompletos está dado por Dempster, Laird y Rubin. Su técnica, conocida como el algoritmo EM[10], consiste de cálculos iterativos en dos pasos – predicción y estimación.

El algoritmo aplicado al caso de datos normales multivariantes está muy relacionado con una versión iterativa del método que imputa estimaciones por regresión.

Predicción.- Dado $\tilde{\theta}$, predice la contribución de cualquier observación faltante al estadístico suficiente (datos completos).

Estimación.- Uso del estadístico suficiente para calcular una estimación revisada del parámetro.

Cuando las observaciones $X_1, X_2, \ldots X_n$ son una muestra aleatoria de una población normal multivariada, el algoritmo de estimación predicción se basa en el estadístico suficiente de datos completos:

$$T_1 = \sum_{j=1}^{n} X_j = n\bar{X} \qquad \text{y} \qquad T_2 = \sum_{j=1}^{n} X_j X_j' = (n-1)S + n\bar{X}\bar{X}' \qquad (1.15)$$

[10] Johnson, R.A. ; Wichern, D.W. (1998)

Predicción

Para cada vector x_j con valores faltantes, $x_j^{(1)}$ son los valores faltantes y $x_j^{(2)}$ los valores disponibles.

Así, $$x_j' = [\, x_j^{(1)'},\ x_j^{(2)'}]. \tag{1.16}$$

Dadas las estimaciones $\tilde{\mu}, \tilde{\Sigma}$ del paso de estimación, se usa la media de la distribución normal condicional de $x_j^{(1)}$ dado $x_j^{(2)}$ para estimar el valor faltante.

Esto es: $$\tilde{x}_j^{(1)} = E(X_j^{(1)} \,|\, x_j^{(2)}; \tilde{\mu}, \tilde{\Sigma}) = \tilde{\mu}^{(1)} + \tilde{\Sigma}_{12}\tilde{\Sigma}_{22}^{-1}(x_j^{(2)} - \tilde{\mu}^{(2)}) \tag{1.17}$$

estima la contribución de $x_j^{(1)}$ a T_1.

La contribución de $x_j^{(1)}$ a T_2 es:

$$\overbrace{x_j^{(1)} x_j^{(1)'}} = E(X_j^{(1)} X_j^{(1)'} \,|\, x_j^{(2)}; \tilde{\mu}, \tilde{\Sigma}) = \tilde{\Sigma}_{11} - \tilde{\Sigma}_{12}\tilde{\Sigma}_{22}^{-1}\tilde{\Sigma}_{21} + \tilde{x}_j^{(1)}\tilde{x}_j^{(1)'} \tag{1.18}$$

y

$$\overbrace{x_j^{(1)} x_j^{(2)'}} = E(X_j^{(1)} X_j^{(2)'} \,|\, x_j^{(2)}; \tilde{\mu}, \tilde{\Sigma}) = \tilde{x}_j^{(1)}\tilde{x}_j^{(2)'} \tag{1.19}$$

Estimación

Cálculo del estimador de máxima verosimilitud revisado:

$$\tilde{\mu} = \frac{\tilde{T}_1}{n} \qquad \tilde{\Sigma} = \frac{1}{n}\tilde{T}_2 - \tilde{\mu}\tilde{\mu}' \tag{1.20}$$

Si todos los componente x_j son ausentes, entonces,

$$\tilde{x}_j = \tilde{\mu} \quad y \quad \tilde{S} = \tilde{x}_j \tilde{x}_j' = \tilde{\Sigma} + \tilde{\mu}\tilde{\mu}' \tag{1.21}$$

b) Regresión EM estocástica

De la misma forma como se trató el caso en la imputación con el algoritmo PLS, se puede aplicar la misma idea en este caso. Al final de la estimación se puede incluir un

componente de ruido por medio de una variable aleatoria distribuida normalmente. En este caso se hace uso del cálculo de la variación residual que forma parte del algoritmo (1.15). Existe también una versión estocástica del algoritmo EM, llamada SEM. [11]

c) Algoritmo de aumento de datos

Cuando el tamaño de muestra es pequeño, una aproximación alterna útil para los estimadores de máxima verosimilitud es agregar una distribución previa para los parámetros y calcular la distribución posterior de los parámetros de interés.

La distribución posterior para un modelo con un mecanismo de ausencia de datos que puede ser ignorado es:

$$p(\theta|Y_{obs}, M) = p(\theta|Y_{obs}) \propto p(\theta) f(Y_{obs}|\theta) \qquad (1.22)$$

en donde $p(\theta)$ es la distribución previa y $f(Y_{obs}|\theta)$ es la densidad de los datos observados.

El aumento de datos[12] es un método iterativo de simular la distribución posterior de θ que combina características del algoritmo EM y la imputación múltiple. Se puede ver como un refinamiento para muestras pequeñas del algoritmo EM usando simulación, con el paso de imputación (I) correspondiendo al paso E y el paso posterior (P) Correspondiendo al paso M.[13]

- Paso I Dados los valores de los parámetros obtenidos en la iteración t, simula valores para los datos ausentes.

$$X_{aus}^{(t+1)} \sim f(X_{aus}|X_{obs}, \theta^{(t)}) \qquad (1.23)$$

- Paso P o posteriori, simula nuevos valores de los parámetros

$$\theta^{(t+1)} \sim f(\theta|X_{obs}, X_{aus}^{(t+1)}) \qquad (1.24)$$

Es un algoritmo de la clase MCMC (Markov Chain Monte Carlo), útil en problemas de datos faltantes.

[11] Celeux y Diebolt, 1987; Celeux, 1988, cit. en Comyn, M. 1999, 24
[12] Tanner y Wong, 1987
[13] cf. Schafer, J.L. (1997)

A partir de valores iniciales de θ^0 se repiten sucesivamente los pasos I y P. Se obtiene una secuencia aleatoria $\left\{\left(\theta^t, X^t_{aua}\right): t = 1,2...\right\}$, la cual converge en distribución a P (Y_{aus}, $\theta|Y_{obs}$).

- Es muy similar al algoritmo EM

- La tasa de convergencia está relacionada con las fracciones de datos ausentes.

1.2.2 Modelos implícitos

El enfoque está sobre un algoritmo, el cual implica un modelo subyacente; las suposiciones son implícitas, pero aún necesitan ser valoradas cuidadosamente para asegurarse que son razonables.

Quedan comprendidos dentro de este grupo, los métodos de imputación Hot deck (imputación por donante). En estos métodos se sustituyen valores individuales extraídos de unidades observadas similares. Es una práctica común que involucra esquemas elaborados para la selección de las unidades similares para la imputación[14]. La Imputación Hot Deck se puede definir como un método en el cual, el valor imputado se selecciona de una distribución estimada para cada valor faltante, a diferencia de la imputación por medias, en donde la media de la distribución es sustituida.

Generalmente la distribución empírica consiste de valores de las unidades donantes, de tal manera que la imputación hot-deck involucra la sustitución de valores individuales extraídos de unidades donantes similares.

Algunos métodos de imputación Hot deck:

- Muestreo aleatorio simple

- Por clase

- Secuencial

- Vecino más cercano

[14] D.B. Rubin (1987)

1.2.2.1 Imputación Hot deck: Muestreo aleatorio simple

Los donantes se extraen de manera aleatoria. Dado un esquema de muestreo equiprobable, la media se puede estimar como la media de los receptores y los donantes.

Esto es:
$$\bar{y}_{HD} = \frac{\{n_D \bar{y}_D + n_R \bar{y}_R\}}{n_D + n_R} \tag{1.25}$$

con
$$\bar{y}_R = \sum_{i=1}^{n_D} \frac{H_i y_i}{n_R} \tag{1.26}$$

en donde H_i es el número de veces que y_i se usa como valor imputado, n_D representa el número de donante, n_R el número de receptores con $\sum_{i=1}^{n_D} H_i = n_R$

Las propiedades de $\bar{y}_{HD}$ dependen del procedimiento empleado para generar los números $\{H_1, H_2, \ldots, H_{nD}\}$.

1.2.2.2 Imputación Hot deck: Por clase

El donante se escoge al azar de la clase a la que pertenece el receptor. Los valores faltantes dentro de cada celda se reemplazan por los valores registrados de la misma celda. La oficina de censos emplea este método para imputar ingresos en el suplemento de ingresos de la encuesta de la población actual (CPS) (Hanson, 1978), basada en variables observadas (edad, raza, sexo, relación familiar, hijos, edo. civil, ocupación, escolaridad, tipo de residencia) de individuos semejantes, de tal manera que su clasificación crea una gran matriz.

1.2.2.3 Imputación Hot deck: Secuencial

- Se hace uso de variables concomitantes (covariate) para clasificar los registros faltantes en clases.
- Se arreglan los registros en cada clase en un cierto orden y se obtiene un valor por cold deck [15] para cada clase.

[15] Reemplaza el valor faltante por un valor constante de una fuente externa. Hace uso de un conjunto fijo de valores que se puede construir con datos históricos, expertos en el área, etc.

Por ejemplo, si para la clase h, las unidades estas arregladas en la secuencia: i1; i2; i3; i4; i5; i6; i7; i8; i9; i10 y los valores actuales disponibles son: yi3; yi6; yi9 solamente y el resto son faltantes.
Con xh como el valor por cold deck en la clase h, entonces los valores imputados serán: xh; xh; yi3; yi3; yi3; yi6; yi6; yi6; yi9; yi9:

1.2.2.4 Imputación Hot deck: Vecino más cercano

Es un procedimiento no paramétrico basado en la suposición de que los individuos cercanos en un mismo espacio tienen características similares.
Requiere de la definición de medida de distancia (generalmente euclídea). El uso de la distancia euclídea tiene el inconveniente de tratar todas las variables de la misma forma. Esto implica estandarizar las variables antes de calcular la distancia. Reemplaza los valores ausentes en una observación por aquellos de otra(s) observación(es) de alguna forma cercana a ella, de acuerdo a una idea predefinida de cercanía en el espacio de las variables comunes **X**.
Un elemento importante a considerar es el número de vecinos. Si el número de vecinos es pequeño, la estimación se hará sobre una muestra pequeña y por lo tanto el efecto será una mayor varianza en la estimación. Por otro lado, si la imputación se hace a partir de un número grande de vecinos, el efecto puede ser la introducción de sesgo en la estimación por información de individuos alejados.

1.2.3 Imputación múltiple

Cuando los valores faltantes se reemplazan por un conjunto de valores imputados, los análisis posteriores no reflejan la incertidumbre de los datos faltantes.
La imputación múltiple es una técnica Monte Carlo en la cual los valores faltantes se reemplazan por $m > 1$ versiones simuladas ($3 \leq m \leq 10$). La pregunta de cómo obtener inferencias válidas a partir de datos imputados se trata en el libro de Rubin sobre imputación múltiple (1987). En su método, cada conjunto de datos completo simulado se analiza por métodos estándares y los resultados se combinan para producir estimaciones e intervalos de confianza que incorporan la incertidumbre de los datos faltantes.

No es el único método para tratar valores faltantes ni es necesariamente el mejor para un problema dado.

Rubin muestra que la eficiencia de una estimación basada en m imputaciones es aproximadamente

$$\left(1+\frac{\gamma}{m}\right)^{-1} \tag{1.27}$$

en donde γ es la tasa de información faltante para la cantidad que esta siendo estimada.

A partir de cada análisis, se debe calcular y almacenar las estimaciones y los errores estándares.

Sea $\hat{Q}_j$ la estimación de una cantidad escalar de interés obtenida de un conjunto de datos j (j = 1, 2, . . . , m) y U_j el error estándar asociado con $\hat{Q}_j$. La estimación global es el promedio de las estimaciones individuales:

$$\overline{Q} = \frac{1}{m}\sum_{j=1}^{m}\hat{Q}_j \tag{1.28}$$

Para el cálculo del error estándar global, se calcula primero la varianza dentro de la imputación (within-imputation):

$$\overline{U} = \frac{1}{m}\sum_{j=1}^{m}U_j \tag{1.29}$$

y la varianza entre las imputaciones:

$$B = \frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{Q}_j-\overline{Q}\right)^2 \tag{1.30}$$

de tal manera que la varianza total es:

$$T = \overline{U}+\left(1+\frac{1}{m}\right)B \tag{1.31}$$

Los intervalos de confianza se pueden obtener tomando la estimación global ± un número de errores estándares, en donde el número es un cuantil de la distribución t con df grados de libertad:

$$df = (m-1)\left(1+\frac{m\overline{U}}{(m+1)B}\right)^2 \tag{1.32}$$

La tasa de información faltante se define como:

$$\gamma = \frac{r+\frac{2}{df+3}}{r+1} \tag{1.33}$$

con $$r = \frac{(1+m^{-1})B}{\overline{U}} \tag{1.34}$$

Ejemplo 1.2[16]

Grado	Sexo	Ed. Especial	Punt. Local	Punt. Nacional	1ª Imputac.	2ª Imputac	3ª Imputac
8	F	No	345	?	42.91	42.27	36.23
8	M	No	325	30			
8	M	Si	300	?	10.02	8.25	14.38
7	F	No	314	45			
7	M	Si	291	?	13.26	27.43	11.09
7	F	No	303	10			
7	F	No	334	?	18.15	38.97	29.74
6	M	Si	383	32			
6	F	No	376	60			
6	F	No	310	?	27.70	29.18	23.34
6	F	No	383	?	53.57	67.13	55.78

(M = 1, F = 0; Ed. Especial: No = 0, Si = 1)

[16] Fuente :Wayman, J. C. (2003)

1ª Imputación = -135.78+0.31Grado+1.14Sexo-10.68Ed.Especial+0.50Punt.Local+error
2ª Imputación = -131.51+0.11Grado+1.43Sexo-10.19Ed.Especial+0.49Punt.Local+error
3ª Imputación = -133.40+0.34Grado+0.81Sexo- 9.96Ed.Especial+0.50Punt.Local+error

Análisis:

Conjunto 1: Media = 37.8105 Varianza = 0.0187
Conjunto 1: Media = 37.8488 Varianza = 0.0185
Conjunto 1: Media = 37.8166 Varianza = 0.0185

$\overline{Q} = 37.8253$ $\overline{U} = 0.0186$ $B = 0.00042$ $T = 0.0192$

error estándar = $\sqrt{T} = 0.138$

1.3 Técnicas de Fusión[17]

La mayoría de las fusiones se basan en el mismo principio. Cada individuo del archivo de receptores se empareja con otro del archivo de donantes tan similar como sea posible en términos de un índice de similitud calculado a partir de las variables activas seleccionadas entre las variables comunes.[18] Esta técnica fue iniciada por Lucien Boucharenc (1981) y Friedrich Wendt (1976, 1984). Se pueden considerar dos familias de técnicas de fusión:

- Emparejamiento de individuos (el papel de las variables activas)

  Es un procedimiento simple y puede, en teoría integrar las relaciones entre las variables transferidas ya que son las características transferidas como un todo. Tiene el inconveniente de reducir la variabilidad de los resultados, en especial cuando el tamaño de la muestra donante es menor que la muestra receptora. La selección de las variables activas es importante y determina la calidad de los resultados obtenidos.

[17] El objetivo de la fusión es usar una especie de imputación hot-deck para transferir variables de una de las muestras llamada donante a la otra muestra llamada receptora. (Lebart L.;Lejeune M. 1995)
[18] Lebart L.;Lejeune M (1995)

- Estimación de las variables

  Se trata de estimar los valores de cada una de las variables. Se establece la relación de estimación de una variable específica en función de las variables comunes en el archivo donante y utilizar esa relación para cada individuo receptor. Se selecciona el mejor subconjunto de variables predicitvas. Sin embargo, para la fusión no existe la certeza de que se recupere la relación existente entre las variables transferidas puesto que el modelo se hace para cada variable.

1.3.1 Emparejamiento aleatorio intra celular

Es el método más utilizado. El método consiste en particionar las muestras de cada encuesta en celdas, reagrupando los individuos semejantes, y destinar a un individuo del archivo receptor un individuo del archivo donante tomado al azar sin reemplazo dentro de la misma celda.

- Selección de los criterios de agrupación. (variables más correlacionadas con las variables a imputar)

- Partición de las dos encuestas según los criterios para crear las celdas.

- Asignación de un individuo donante, tomado al azar sin reemplazo de una celda a un individuo receptor de la misma celda.

Es necesario para esta técnica que las dos muestras tengan una distribución parecida según las variables comunes y que los tamaños de las muestras no sean muy diferentes.

1.3.2 Emparejamiento exacto (emparejamiento estadístico)

El emparejamiento estadístico (statistical matching) se hace sobre la base de características similares, más que sobre información identificativa única. Otros términos han sido usados para esto, como "synthetic, stochastic, attribute, y data matching".

Como técnica ha sido más o menos ampliamente practicada desde la llegada de los archivos de uso público en los 60's. (Kadane 1978)

Por ejemplo, sean A y B dos archivos tomados de dos diferentes encuestas. El archivo A contiene las variables (X, Y) y el archivo B las variables (X, Z). El objetivo será entonces combinar los dos archivos para obtener uno que contenga las variables (X, Y, Z). A diferencia de la coincidencia exacta (record linkage), los dos archivos que se van a combinar no suponen la existencia de los mismos individuos. Los registros con individuos similares se combinan.

Si el archivo A consiste en parte de los registros: (X_1, Y_1), (X_2, Y_2), (X_3, Y_3),
y el archivo B los registros: (X_1, Z_1), (X_3, Z_3), (X_4, Z_4), (X_5, Z_5)
se podría crear: (X_1, Y_1, Z_1), (X_3, Y_3, Z_3)

Emparejar sobre X_1 y X_3 no implica que sean los mismos individuos. Esto sería un emparejamiento parcial de los dos archivos. Sin embargo existen otras estrategias que pueden ser agrupadas en:

- emparejamiento con restricciones
  Se requiere del uso de todos los registros en los dos archivos y básicamente preserva las distribuciones marginales de Y y Z.

- emparejamiento sin restricciones
  No se requiere del uso de todos los registros.

En el primer caso, se deberá terminar con un archivo combinado que tendrá registros adicionales que usarán la información remanente no usada del archivo A: registro (X_2, Y_2) y los dos registros no usados del archivo B: (X_4, Z_4), (X_5, Z_5)

En el segundo caso, aunque todos los registros de uno de los archivos pueden ser usados (emparejados) con registros similares del segundo archivo, algunos registros del segundo archivo pueden ser usados mas de una vez o nunca. En este ejemplo sería: $(X_2, Y_2, ?)$ y (X_4, Z_4), (X_5, Z_5) podrían no ser incluidos.

1.3.3 Fusión mediante descripción factorial

Técnica definida por G. Santini (1984)

- Espacio factorial

  Se desarrolla un análisis factorial (Análisis de Correspondencias Múltiples, Análisis de componentes principales) sobre las variables comunes o una selección de las variables comunes más discriminantes de las variables específicas, para el conjunto de individuos donantes y receptores. Se construye el espacio factorial a partir de los primeros n ejes del análisis. La similitud donante receptor se mide por una distancia definida en R^n

- Vecindad

  Para cada receptor se calcula una vecindad seleccionada entre los donantes. Se puede definir a partir de una esfera de radio r centrada sobre cada receptor o a partir de la búsqueda de los k donantes más próximos.

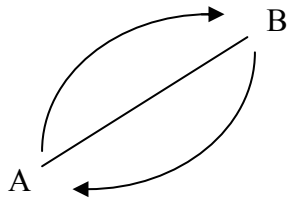
1.3.3.1 Fusión por matrimonio

Se basa en el cálculo de una distancia en el espacio factorial entre los individuos donantes y receptores. G. Santini define "maridaje entre un receptor y un donante:
Los individuos son maridos en función de su proximidad d calculada sobre las coordenadas factoriales. Se debe evitar que un mismo individuo donante sea utilizado muchas veces. Se introduce el concepto de penalización. Es decir, si un donante ya es marido de otros δ individuos receptores, esa distancia d se penaliza como: $d^* = 1-(1-d)^{\delta}$

La propuesta de maridaje de G. Santini:

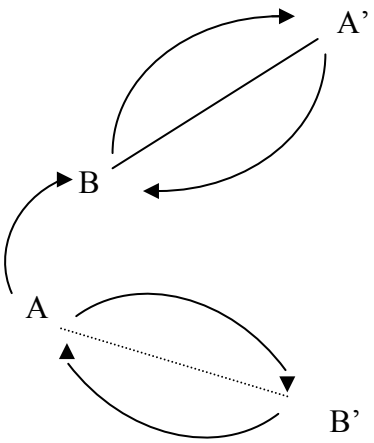
- Matrimonio por vecinos recíprocos

  Si A es el vecino más cercano de B y recíprocamente B es el vecino más cercano de A y no han estado casados jamás, entonces A y B son inmediatamente casados.



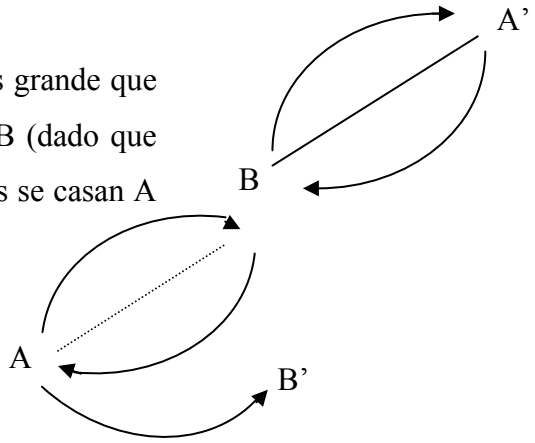
Matrimonio con el amigo de infancia

Si B es el vecino más cercano de A, pero B ya está casado con A', entonces A será casado con B' que es el siguiente vecino más cercano de A.



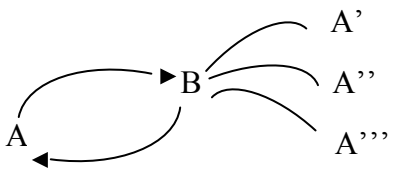
- Matrimonio por adulterio

Si la distancia entre B' y A es más grande que la distancia penalizada entre A y B (dado que B ya esta casado con A'). Entonces se casan A y B.



- Matrimonio por asiduidad

Cuando se quiere unir A y B, pero B es el vecino más cercano de A', A'', A''' con los cuales ya está casado. Al final se casan A y B



- Matrimonio por conveniencia

Estos matrimonios se realizan utilizando los métodos de optimización de distancia a nivel global.

- Matrimonio de los irreducibles

  Individuos restantes que no encontraron pareja. Se buscan otras reglas de optimización que permitan obtener un matrimonio con estos irreducibles.

1.3.3.2 Fusión por búsqueda de socios

Método desarrollado por la sociedad Statiro en 1994. Se basa en la búsqueda del socio de un individuo con la ayuda de una distancia calculada en el espacio factorial.

Etapas:

- Se busca el espacio factorial por un análisis factorial de las variables críticas. Después se mide la similitud de los individuos por una distancia euclidea calculada a partir de sus coordenadas factoriales. Para cada receptor (r) se retienen los m donantes (d) cuya distancia D con el receptor verifica: $D(r, d) < S$ (umbral de distancia).

- Se imponen restricciones para escoger al donante entre los m retenidos.

- Comparación de la semejanza de la señalización entre el donante y el receptor. Se utiliza la agregación multicriterio (cf. Roy B., 1985) .Se obtiene una nota global sobre todas las variables y los individuos más parecidos se retienen. En el caso de igualdad entre dos notas globales, se decide seleccionar al donante que haya sido menos utilizado antes.

Una vez desarrolladas estas etapas, si no se encuentra donante para el receptor, se empieza nuevamente alargando el radio de vecindad S de la primera etapa.

1.3.3.3 Fusión por búsqueda de los vecinos más cercanos[19]

Se tienen dos muestras de datos independientes:

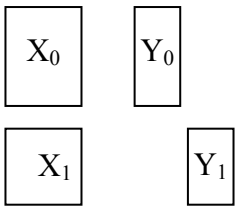


Figura 1. 8 Descripción del problema

X_0 y X_1 son las matrices con las variables comunes. Y_0 y Y_1 son las variables específicas. El objetivo será el de estimar la matriz Y_0 para la segunda muestra. Se define un espacio factorial y se aplica un algoritmo de búsqueda de los k vecinos más cercanos. Es necesario que el subespacio factorial común sea predictivo de las variables suplementarias que se quieren transferir.

Metodología:

- Tanto los donantes como los receptores se posicionan en el mismo subespacio factorial común.

- Para cada receptor de X1 se busca un vecino definido a partir de los k vecinos más cercanos de X0.

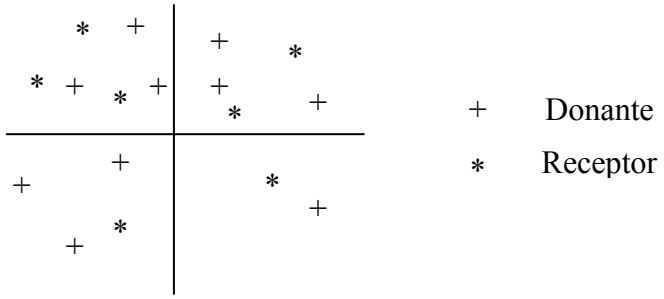


Figura 1.9 Diagrama factorial

[19] cf. Rius, R. et al., (1999) Este trabajo no es de imputación, sino de visualización conjunta de los elementos suplementarios de los dos ficheros.

- Se imputa al receptor de X1 la media de las variables Y0 en la vecindad de los individuos donantes.

Marie Comyn, en su tesis doctoral[20], trata el problema de la fusión y validación de archivos de encuestas partiendo de dos muestras distintas extraídas de un mismo universo. Su modelo se basa en: análisis factorial, búsqueda rápida de k vecinos más cercanos e imputación de datos por frecuencias o probabilidades. El algoritmo que emplea para la búsqueda de vecinos es el de Kittler. La idea de este método es la de seleccionar los puntos candidatos al conjunto de los k vecinos más cercanos siguiendo una métrica euclídea con la ayuda de una métrica para la cual el cálculo de las distancias es más rápida. Propone la utilización de la distancia de Manhattan d_c empleando la relación:

$$d_e(x,y) \le d_c(x,y) \le \sqrt{p}\, d_e(x,y) \qquad (1.35)$$

p es la dimensión del espacio y $d_c(x,y) = \sum |x_i - yi|$

El algoritmo de Kittler opera de la siguiente manera:

1. Calcula la distancia de Manhatan $d_c(x , y_i)$ del receptor a todos los donantes $y_i \in \mathbf{Y}$.

2. Encuentra los k vecinos más cercanos de x ($\tilde{\mathbf{Y}}_x$) siguiendo la métrica d_c.

3. Encuentra $\tilde{y}$ tal que $d_e(x,\tilde{y}) = \max_{\tilde{y}_i \in Y_x} d_e(x,\tilde{y}_i)$

4. Determinar el conjunto de los puntos susceptibles de ser uno de los k vecinos más cercanos de x siguiendo una métrica d_e.

$$Y_x = \left\{ y \,\middle|\, y \in Y, d_c(x,y) \le \sqrt{p} de(x,\tilde{y}) \right\}$$

5. Buscar en el conjunto Y_x, los k vecinos más cercanos siguiendo la métrica de.

[20] Comyn M. (1999)

Richetin modifica este algoritmo introduciendo una tercera distancia, la distancia de Tchebychev: $d_\infty(x,y) = \max_{i=1}^{p}(x_i - y_i)$. Utilizando la siguiente relación;

$$\frac{d_e(x,y)}{\sqrt{p}} \le d_\infty(x,y) \le d_e(x,y)$$

Richetin propone reemplazar el conjunto Y_x de la etapa 4 del algoritmo de Kittler por el conjunto $\mathbf{Y}_x^* = \left\{ y \middle| y \in Y, d_\infty(x,y) \le d_e\left(x, \tilde{y}\right) \right\}$

De esta manera, el algoritmo de Richetin utiliza conjuntamente las distancias de Maniatan y de Tchebychev.

1.3.4 Fusión mediante Árboles de Clasificación[21]

Se presenta a continuación una descripción breve del uso de los árboles de clasificación y regresión (CART Classification and Regression Trees). Es un método de naturaleza no paramétrica. El objetivo es predecir el valor de una variable **Y** en función de los valores de un conjunto de variables **X**. Se necesita para su construcción de una muestra de aprendizaje. Las variables pueden ser de tipo cuantitativo o cualitativo. Si la variable que se va a predecir es de tipo cuantitativo se habla de árboles de regresión. Si Y es una variable cualitativa dividida en J clases y el objetivo es asignar una clase j a cada individuo en función de los valores que forman el vector $\mathbf{x}_i$, entonces se habla de árboles de clasificación. La forma de construir un árbol de regresión o de clasificación es similar. El espacio ***X*** se particiona mediante una secuencia de divisiones binarias en función de las coordenadas de **X**. Un nodo t de un árbol es aquel subconjunto de X determinado por las divisiones realizadas para llegar a él. Los nodos que no se dividen se denominan nodos terminales. A cada nodo terminal se le asigna un valor de la variable de respuesta Y. Un árbol proporciona una regla de asignación que se denota por d(**X**). Si un individuo baja por el árbol, en función de los valores que toma del vector **X**, caerá en un nodo terminal y se le asigna el valor de la variable de respuesta asociado a ese nodo.

[21] cf. Bárcena, M.J. (2000).

En términos generales, el algoritmo queda de la siguiente manera:

- Construir un árbol $\boldsymbol{Y}_X$ con las observaciones i=1, . . . , N_A
  $\boldsymbol{Y}_1$. . $\boldsymbol{Y}_\alpha$ son las hojas del árbol y $\boldsymbol{Y}$ la partición que forman

- Para imputar el valor de **Y** para una observación con i ∈ {N_A+1,... N} se recorre el árbol $\boldsymbol{Y}_X$. Si cae en la hoja $\boldsymbol{Y}_{\delta i}$, se imputa como una función de los valores de **Y** observados en la hoja.

Para imputar simultáneamente, emplean los árboles univariados $\boldsymbol{Y}^{(i)}_{\mathbf{X}}$ construidos para cada una de las variables Y_i i = 1 , 2 , . . . q $\boldsymbol{Y}^{(i)}_{\mathbf{X}}$ es el k-ésimo nodo del j-ésimo árbol de **Y** sobre **X** .

Para cada q-tupla (α_1 , α_2 , . . . α_q) tal que α_j (j ∈ { 1, 2 , . . . q}) es la etiqueta del nodo en el árbol $Y_X^{(j)}$

$$C_{\alpha 1,\alpha 2,...,\alpha q} = Y_{\alpha 1}{}^{(1)} \cap Y_{\alpha 2}{}^{(2)} \cap . . . Y_{\alpha q}{}^{(q)} \qquad (1.36)$$

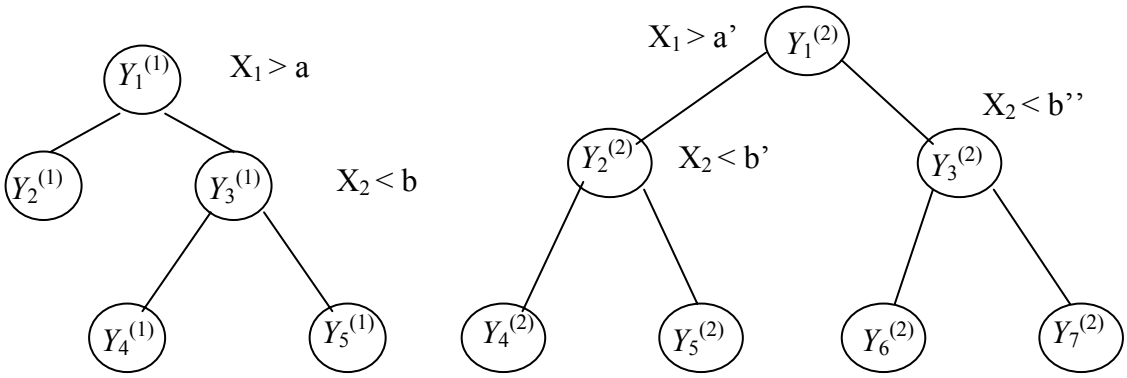


Figura 1.10 Árboles $Y_X^{(1)}$ y $Y_X^{(2)}$

La idea es la siguiente: para imputar $\mathbf{Y}_i$ i ∈ {N_A +1,. . .N} se recorren los árboles construidos para cada variable. Se termina en las hojas $Y_{i1}{}^{(1)}$, $Y_{i2}{}^{(2)}$,. . . $Y_{iq}{}^{(q)}$ y por lo tanto pertenecen a $C_{1, 2,..., q}$ Se imputa $\mathbf{Y}_i$ como una función de los valores de **Y** en la muestra de entrenamiento (archivo A) que también pertenecen a $C_{1,2,...,q}$
Las opciones son: imputar mediante un vector tomado al azar de $C_{1,2,...,q}$, mediante la media de todos, etc.

En este ejemplo, considérese el caso de imputar Y_i tal que a' < X_1 < a y X_2 < b''.
Esto quiere decir que se terminaría en las hojas $Y_2^{(1)}$ y $Y_7^{(2)}$.
Se propone entonces imputar $\mathbf{Y}_i$ usando los valores de Y observados en la muestra de entrenamiento que caigan en $C_{2,7}$.

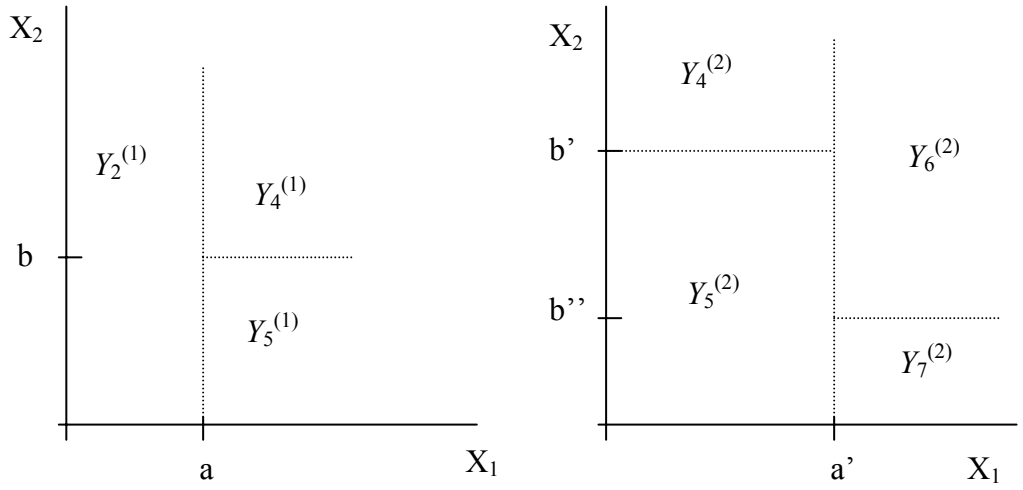


Figura 1.11 Particiones del espacio **X** inducida por los árboles $Y_X^{(1)}$ y $Y_X^{(2)}$

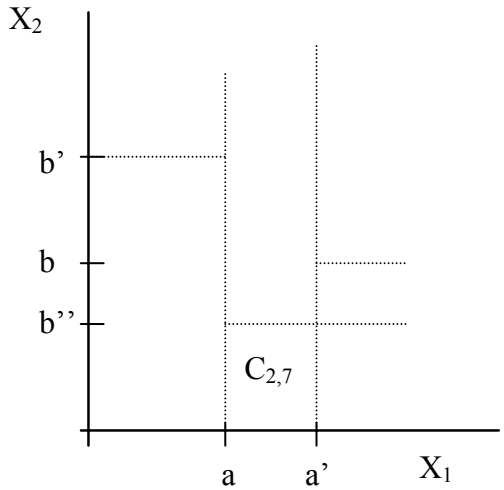


Figura 1.12 Traslape de las particiones de X

M. J. Bárcena y F. Tusell tratan el problema de completar dos archivos con registros que contienen un subconjunto común de variables completas. Con datos simulados, el método trabaja bien si las variables comunes son buenas predictoras de las variables específicas y si la relación funcional entre las predictoras y las específicas pueden estar razonablemente bien aproximadas por un árbol. Probaron su método sobre conjuntos de datos simulados y

reales de tamaño relativamente grande. Establecen que se puede extender para cubrir patrones irregulares de ausencia de datos. El método hace pocas suposiciones, es computacionalmente factible y parece dar buenos resultados.

1.3.5 Fusión mediante Redes Neuronales[22]

Los métodos neuronales son interesantes dentro del cuadro de las fusiones que operan variable por variable. Son procedimientos algorítmicos de cómputo intensivo para transformar entradas en salidas deseadas usando redes altamente conectadas de procesamiento relativamente simple. Sus componentes esenciales son:

- Unidades básicas de computo (nodos o neuronas)
- Arquitectura de la red (describe las conexiones entre los nodos)
- Algoritmo de entrenamiento (determina los parámetros de la red para trabajos particulares)

Los nodos se conectan a otros en el sentido de que la salida de uno puede ser utilizada como parte de la entrada de otro. Cada nodo transforma una entrada en una salida empleando alguna función preestablecida típicamente monótona pero arbitraria. Esta función depende de constantes determinadas con un conjunto de entrenamiento de entradas y salidas.
La arquitectura es la organización de los nodos y los tipos de conexiones permitidas. En aplicaciones estadísticas, los nodos se arreglan en una serie de capas con conexiones entre nodos en diferentes capas pero no entre nodos de la misma capa. La capa que recibe la entrada inicial se llama capa de entrada. La capa final se llama capa de salida. Cualquier capa entre la entrada y la salida se llama capa oculta.
Existen diferentes arquitecturas y cada una utiliza diferentes estrategias de aprendizaje para desarrollar sus tareas.

[22] cf. Johnson, R.A.; Wichern, D.W. (1998)
cf. Tussel, F. (2002)

Las redes neuronales se pueden utilizar para discriminación y clasificación. Las variables de entrada son las características medidas de los grupos y las variables de salida son variables categóricas que indican membresía a un grupo.

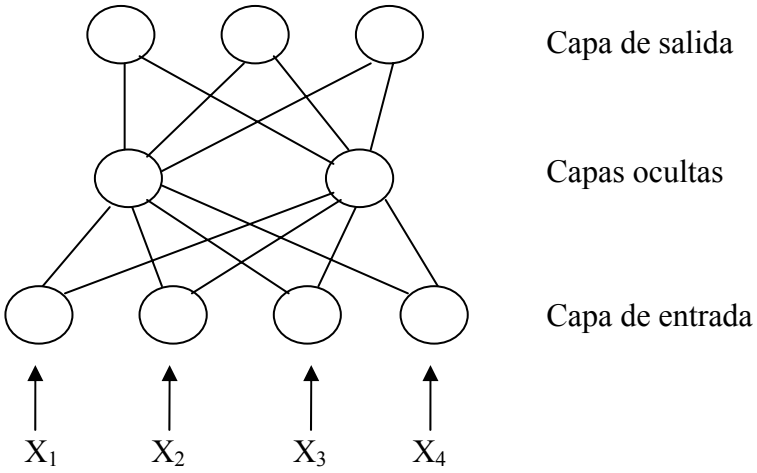


Figura 1.13 Red neuronal con una capa oculta

Se distinguen dos etapas: codificación, en la cual la red se entrena para desarrollar cierta tarea y decodificación, en la cual la red se usa para clasificar ejemplos, predecir o ejecutar cualquier trabajo de aprendizaje.

La etapa de codificación se desarrolla sobre los individuos donantes con las variables comunes como señales de entrada y las variables específicas como señales de salida.

Son un instrumento potente para representar enlaces no lineales entre un conjunto de variables de entrada y otro de salida.

El mecanismo de aprendizaje en el caso de una fusión se construye sobre la información contenida en el archivo de individuos donantes. Como método de fusión variable por variable, no garantiza la conservación de las relaciones entre las variables.

1.3.6 Fusión mediante Análisis homogéneo[23]

El análisis homogéneo permite estimar e imputar datos faltantes para obtener un conjunto de datos lo más homogéneos posible. Se basa en un criterio de optimización que tiende a maximizar la homogeneidad de los datos. Si las variables miden la misma propiedad, es

[23] cf. Co Vila (1997)

posible entonces reemplazarla por una variable sintética sin perder demasiada información. Similar al análisis factorial. Para evaluar la calidad de esta sustitución, se define un criterio de homogeneidad y una función de pérdida. La maximización de la homogeneidad conduce a un análisis homogéneo parecido al análisis de correspondencias múltiples - Co (1997) y Saporta y Co (1999) han estudiado una técnica propuesta por Buuren y Van Rijckevorsel (1992) - . Una condición para la aplicación del análisis es que las variables comunes sean lo más predictivas posible de las variables específicas.
Co Vila trata el problema del tratamiento de datos faltantes con técnicas estadísticas e informáticas. Dedica parte de su trabajo a la fusión de archivos. Considera el problema de la fusión de archivos como un caso particular de datos faltantes. Basa su propuesta en un criterio de optimización (maximiza la homogeneidad de los datos). Su método no copia de manera sistemática el bloque entero de datos de un individuo. Se necesitan altas correlaciones entre las variables. Para las variables con un número grande de categorías, propone una técnica de segmentación dicotómica que reagrupa las modalidades de las variables en dos grupos. Se transforma la fusión de variables que tienen un número grande de modalidades en una serie de fusiones de variables a dos modalidades.

1.3.7 Fusión mediante regresión PLS

Un ejemplo de la aplicación de la metodología PLS se encuentra en el artículo de Nicolás Fischer, Christian Derquene y Gilbert Saporta[24]. Presentan un método para el emparejamiento de datos basado en la generación de individuos virtuales. Se presenta un enfoque multivariado usando regresión PLS.

Se dispone de dos muestras de encuestas: una muestra primaria χ y muestras secundarias $\mathbf{Y^1, \ldots, Y^K}$. La muestra primaria incluye las variables (comunes) llamadas X_{MP} y las variables medidas X_M. Las muestras secundarias tienen algunas variables compartidas con las variables $X_{MP,}$ llamadas $Y_{MP}^{(k)}$, y otras variables medidas las cuales son en parte comunes y en otra, no comunes a X_M, llamadas $Y_M^{(k)}$.

[24] cf Fisher, N.; Derquenne, C.; Saporta, S. (2001)

Tabla 1.1 Variables comunes y variables específicas

	CREDOC 1990 (2000 personas)	SOFRES 1990 (8000 clientes)
Variables muestreadas $\mathbf{X}_{MP}$ $\mathbf{Y}_{MP}$	-sexo x edad -ocupación -tamaño de población	-edad -ocupación -tamaño de población
Variables medidas $\mathbf{X}_{M}$ $\mathbf{Y}_{M}$	-características de vivienda -características del hogar -principal fuente de calefacción -opinión con respecto a las estaciones nucleares -opinión con respecto al ambiente	-características de vivienda -características del hogar -principal fuente de calefacción -nivel de satisfacción con respecto a la fuente de calefacción (costo, seguridad, confort térmico, etc.)

El método se basa en dos pasos:

- Generación de la *primera muestra artificial* basada en la muestra primaria χ.

  Reduce la dimensión del espacio empleando Análisis de Correspondencias Múltiples con X_{MP} como variables activas y X_M como variables suplementarias. Selecciona aquellos componentes cuyo valor propio sea mayor de 1/MP(Q). Sustituye el espacio continuo de los componentes principales por un espacio discreto para determinar una distribución empírica. De esta distribución se extraen N individuos artificiales. Estos individuos artificiales se llaman "individuos dummy".
  ($X_{MP,}$ X_M), involucran también una distribución empírica. Se extraen N individuos artificiales conociendo los individuos dummy.

- Inserción estadística de una muestra secundaria $\mathbf{Y}^{1}$ sobre la primera muestra artificial para obtener la segunda muestra artificial.

  G1 representa el número de variables no comunes con la muestra primaria que serán injertadas. P1 son las variables compartidas por la muestra primaria y la muestra secundaria $y_C^{(1)} = \{y_{C(1)}^{(1)}, \ldots, y_{C(P1)}^{(1)}\}$. Utiliza la regresión PLS, ya que permite modelar un bloque de variables de respuesta por un bloque de variables explicativas incluyendo la estructura de correlación.

El segundo paso se repite para tener la *muestra artificial final* o *muestra de individuos virtuales*, injertando de manera progresiva otras muestras secundarias $\mathbf{Y^2, \ldots, Y^K}$.

El método se aplicó a la generación de una muestra de 10000 individuos virtuales. Aplicando Análisis de Correspondencias Múltiples (ACM) reducen la dimensión. Generan un grupo de variables correlacionadas con los componentes principales.

1.4 Validación

Debido a que en general no existe un modelo para los datos, es posible que la única manera para evaluar la calidad de la fusión sea una validación empírica.
Se deben establecer indicadores de calidad. Estos estarán en función de la respuesta que se esté dando al análisis. Es decir, en general, no se está interesado en predicciones individuales y es suficiente con tener predicciones que sean correctas en promedio para un grupo de individuos. Saporta menciona que no es suficiente con recuperar las distribuciones marginales o valores medios, puesto que estos se pueden lograr con un muestreo aleatorio. Es posible que el problema principal sea el conservar la estructura de covarianza, o en el caso de datos categóricos, tener una tabla cruzada correcta entre las variables de interés.
Para esto se va a requerir que exista:

- Un número grande de variables comunes (esto aumenta la dimensionalidad del espacio).

- Altas correlaciones entre las variables comunes y las variables específicas.

- Una estructura común entre el archivo donante y el receptor.

El interés de reproducir los estadísticos marginales puede ser una decisión incorrecta, debido a que solo tiene sentido si las dos fuentes de datos provienen de la misma población. Algo similar se puede decir de la comparación de las correlaciones entre las variables específicas entre los donantes y los receptores.

En el caso de las comparaciones de correlaciones entre las variables comunes y variables imputadas, la situación es diferente, ya que las imputaciones suponen que f (Y|X) es la misma tanto en los donantes como en los receptores.

Por todo esto, será necesario establecer indicadores de la bondad de la fusión, que permitan establecer el resultado de la fusión.

Existen dos peligros relacionados con la fusión de datos que no serán tratados en este trabajo, pero que vale la pena tenerlos en consideración por las consecuencias que esto pueda tener:

- Menos esfuerzo para la recolección de datos.

  La fusión de datos cubre una necesidad con la que se encuentran algunos administradores de datos que quieren entregar a sus usuarios finales un archivo de datos único y completo. Sin embargo, se debe tener cuidado con el uso de estos datos puesto que son estimaciones y no observaciones.

- Confidencialidad y privacidad de los datos.
  Con la fusión de datos se esta en una posición en la cual, información que no fue recolectada, se ha estimado y agregado sin el conocimiento de los individuos.

1.4.1 Propuestas de validación

Marie Comyn[25] define la tasa de error medio de la siguiente manera: calcula un error de reconstitución entre los individuos receptores con los valores reales y los valores obtenidos por la fusión, utilizando la métrica χ^2 entre dos perfiles fila, correspondientes al mismo individuo.

$$d^2(r,f) = \sum_m \frac{IJ}{I_m}\left(\frac{x_{rm}}{J} - \frac{x_{fm}}{J}\right)^2 = \frac{I}{J}\sum_m \frac{1}{I_m}\left(x_{rm} - x_{fm}\right)^2 \qquad (1.37)$$

[25] cf. Comyn, M. (1999)

en donde I es el número de receptores, J es el número de variables específicas, I_m es el número de receptores con la modalidad m. El error total de reconstitución para una fusión de datos se obtiene por la media de las distancias sobre el conjunto de individuos receptores reconstituidos.

En el caso categórico se define una matriz de confusión para establecer otros indicadores

$$L = \frac{\chi^2 - (p-1)^2}{(p-1)\sqrt{2}} \tag{1.38}$$

en donde p es el número de modalidades para una variable específica. La fusión será mejor que la imputación al azar con un error del 5% si L > 1.96

Define también la tasa de concordancia por fusión τ_1 que representa el porcentaje de individuos bien clasificados. Se requiere del uso de la tabla de confusión. El criterio de comparación de la tasa de concordancia:

$$G = \frac{\tau_1 - \tau_0}{\sqrt{\frac{\tau_0(1-t_0)}{n}}} \tag{1.39}$$

en donde τ_0 representa la tasa de concordancia por muestreo aleatorio simple.

Si G > 1.96, la tasa de concordancia por fusión es significativamente mejor que por muestreo aleatorio simple con un margen de error de 5%.

Estudia la eficacia de los métodos de remuestreo como la validación cruzada y bootstrap como técnicas de validación.

- Divide el archivo donante en m partes iguales
- Aplica la fusión, retirando cada vez una de las m partes como receptores.
- Calcula la media de m predictores de error entre las variables específicas reales y reconstituidas.

M. J. Barcena[26], en su tesis doctoral, trabaja con la estimación de la tasa de error de un árbol de regresión y/o clasificación mediante validación cruzada.

Selecciona como árbol T* de tamaño óptimo el que tenga menor tasa de error $R^{cv}(T)$.

Emplea el error cuadrado medio relativo $RE^{cv}(T^*)$:

$$RE^{cv}(T^*) = \frac{R^{cv}(T^*)}{\lambda} \tag{1.40}$$

λ es el valor propio asociado de la variable de respuesta. Mide la proporción que representa el error cuadrado medio respecto a la varianza de la variable de respuesta. En este caso, la varianza de la variable de respuesta corresponde a la varianza del componente principal debido a que efectúa el ensayo con los componentes principales de las variables de respuesta.

Empleando árboles en cascada, utiliza dos indicadores de la calidad de la fusión:

$C_1(k)$ mide la calidad de imputación para cada una de las variables a imputar Y_k

C_2 que indica la calidad de imputación de todo el conjunto de valores observados par un individuo i.

Co Vila[27] utiliza en su trabajo la evaluación a nivel global e individual.

En el caso del nivel global de evaluación, establece que las distribuciones reales y reconstituidas deben estar próximas y presentar poca desviación. Para las variables cualitativas emplea la χ^2 para calcular una distancia entre el marginal real y reconstituido y para las cuantitativas, compara medias y varianzas. Verifica la preservación de las relaciones entre variables. Emplea tablas cruzadas en el caso cualitativo y en el caso cuantitativo la correlación de las variables (la media de la desviación de las covarianzas)

$$\overline{COV} = Cov(X, Y) - COV(\hat{X}, \hat{Y}) \ .$$

A nivel individual, para las variables cuantitativas, el error promedio de imputación:

$$\sum_i \frac{\left| y_i - \hat{y}_i \right|}{n}$$

[26] cf. Barcena, M. J. (2000)
[27] cf. Vila C. (1997)

Para las cualitativas nominales, Lejeune M. y Lebart L. proponen una validación por tasas de buenas reconstituciones que define el porcentaje de bien clasificados.

$\tau = \sum_{i=1}^{k} p_i^2$ En donde p_i representa la probabilidad correspondiente a cada una de las k categorías de una variable, es la tasa correspondiente a una afectación aleatoria.

Para las variables cualitativas ordinales define el coeficiente de proximidad (Saporta G. y Co V.), que penaliza los errores de afectación según la desviación entre la modalidad verdadera y la afectada. Para ello utiliza una matriz que toma en cuenta los costos del error de clasificación $C = [c_{ij}]$. Tiene entonces: $\tilde{d} = \sum_{i=1}^{k}\sum_{j=1}^{k} c_{ij}^{k} p_i p_j$ costo promedio del error global de clasificación sobre la hipótesis de una afectación aleatoria. Si $\bar{d}$ representa la media de los costos de error asociados a una regla de asociación, $\bar{d} - \tilde{d}$ mide la ganancia por comparación con una afectación aleatoria.