INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

CIMAT

Tesis de Maestría

Imputación del ingreso por trabajo de la Población Ocupada del Estado de Aguascalientes, México

(XII Censo General de Población y Vivienda, 2000)

Un método alternativo

Rubén Darío Herrera Morfín
e-mail: dario.herrera@inegi.org.mx

Asesor
Doctor James M. Lepkowski

Instituto Nacional de Estadística y Geografía
Centro de Investigación en Matemáticas, A.C.

Enero 2011

Quieres devenir filósofo
Desde ahora prepárate a ser ridiculizado y persuádete de que las gentes ordinarias quieren burlarse de ti y decirte: "¡De un día para otro se volvió filósofo ¿De dónde tanta arrogancia?!"
Dentro de ti, que no haya soberbia; trabaja fuertemente en las cosas que te hayan parecido mejor y que sean las más bellas.
Recuerda que, si perseveras en tus propósitos, aquellos que en principio se burlaron de ti, enseguida te aceptarán; mientras que si cedes a sus insultos, serás burlado por segunda vez.

Epíteto, Enchiridion, 22

Agradecimientos

Esta tesis nunca hubiera sido terminada sin el apoyo de ciertas personas.

Primero, debo expresar mi sincero agradecimiento a mi asesor, el doctor Jim Lepkowski. Él siempre me brindó su ayuda y conocimiento de manera desinteresada, y a pesar de la distancia existente entre ambos, la comunicación siempre fluyó en buenos términos. Le estoy especialmente agradecido por darme la oportunidad de escribir mi trabajo de tesis bajo su tutela.

También quiero agradecer a los profesores John Van Hoewyk, Rogelio Ramos y Elías Rodríguez por su apoyo y buena voluntad para leer y criticar mi trabajo

Las siguientes personas contribuyeron enormemente para poder desarrollar este documento:

- Federico Romo, Elsa Resano, Cecilia Martínez y Guadalupe Macías fueron indispensables para poder ubicar el problema original y conocer acerca del procesamiento del INEGI.
- Andrés Ríos, Raymundo Bailón, Eduardo Ríos y Jesús Torres ayudaron enormemente en el acceso a la información.
- Rosy Dávalos, Graciela Farías y Gilberto Calvillo que con su ayuda desinteresada lograron resolver lo necesario para concertar el proyecto con ISR de Michigan, USA.
- Juan Manuel García y Juan Ramón Mena quienes me facilitaron en todo momento el soporte computacional requerido.
- Abdón Sánchez, Walter Rangel, José Vences, Gabriel Alcolea†, Paul Carrasco y Eric Rodríguez quienes me apoyaron para enfrentar el gran reto que implica estudiar una maestría y, en particular, desarrollar esta tesis.
- Patricia Muñoz, David Celis y Edilberto Aldán, con su gran ayuda se logró corregir la versión del documento en español
- Víctor Gabriel Morfín Aguilar quien tradujo este material al idioma inglés.

Agradezco al CIMAT y al INEGI el soporte técnico y económico sin el cual la investigación no tendría sentido alguno. Para todo el personal de ambas instituciones mis más sinceras gratitudes.

Mis gracias especiales a todos mis amigos y compañeros de maestría por estar ahí y por el apoyo que me brindaron, de igual forma por todos los momentos felices compartidos.

Finalmente, quiero expresar mi gratitud a mi familia. En principio, a mis padres que siempre estuvieron brindando su apoyo, a mi esposa Ana Bertha, quién sabe lo que esto significa para ambos, a mis dos hijos Rubén y Benjamín quienes a la fecha aún preguntan acerca del grado de avance de este trabajo.

Gracias mil.

Aguascalientes, Ags. Enero 2011 Rubén Darío Herrera Morfín

Contenido

Lista de tablas

Lista de Gráficas

Capítulo 1

Introducción

1.1 Planteamiento del Problema

Esta investigación retoma la imputación aplicada por el Instituto Nacional de Estadística y Geografía (INEGI) a los ingresos por trabajo[1] de la Población Ocupada[2] del Estado de Aguascalientes (situado al centro de México, ver Anexo A), captados durante el XII Censo General de Población y Vivienda, 2000 (XII CGPyV, 2000). Según datos del propio Censo, en el año 2000, Aguascalientes contaba con una Población Total de 944,285 personas; de ésta, 331,083 eran consideradas como Población Ocupada, misma que presentó una NO RESPUESTA en ingresos y posterior a la imputación de 14,432 registros.

La reciente capacitación del personal del INEGI en temas relacionados con la imputación de datos y el desarrollo constante de nuevas técnicas para analizar la NO RESPUESTA, permiten proponer una metodología alterna que principalmente adicione elementos que conduzcan hacia una mejora continua.

El objetivo fundamental de esta investigación consiste en aplicar la Imputación Múltiple, vía la Regresión Secuencial[3] (Raghunathan et al., 2001a) a la NO RESPUESTA del ingreso. Aunque el hecho de contar con la imputación generada por el INEGI permite plantear como un objetivo adicional la comparación de ambas alternativas; en este sentido, es importante mencionar lo siguiente:

- El origen de la NO RESPUESTA se debe principalmente a que el informante no reportó sus ingresos y, en menor medida se da por inconsistencia de los ingresos reportados contra otras variables (por ejemplo, periodo en el que recibe el ingreso) lo cual se detecta en la fase de validación; en ambos casos no se conoce el valor real del ingreso.
- El INEGI imputó sólo la NO RESPUESTA del ingreso, aunque para lograrlo se apoyó en un grupo de variables predictoras que aportaron información de manera conjunta; por su parte, el método RSIM imputa la NO RESPUESTA del ingreso y de todas las variables predictoras factibles.
- La imputación realizada por el INEGI cubrió un 70% de la NO RESPUESTA del ingreso, mientras que la opción RSIM se aplica al 100% de la NO RESPUESTA tanto del ingreso como de las variables predictoras a emplear.

En nuestros días, la existencia de la NO RESPUESTA provoca que el usuario de la información genere por sí mismo (consciente o inconscientemente) ciertos problemas entre los cuales se encuentran:

- Selecciona y analiza diferentes subconjuntos de variables con los problemas de dimensión correspondientes.

[1] Se define como la percepción total monetaria (dinero en pesos mexicanos) que obtiene la persona ocupada en la semana de referencia por su(s) trabajo(s) o su desempeño en la actividad económica bajo un periodo de pago específico. Se consideran los ingresos por concepto de sueldos, comisiones, propinas y otras cosas.

[2] Persona de 12 años o más que realizó alguna actividad económica, al menos una hora en la semana de referencia, a cambio de un sueldo, salario, jornal u otro tipo de pago en dinero o en especie

[3] A la técnica resultante de la aplicación conjunta de ambos métodos se le denotará con las siglas RSIM

- Puede emplear múltiples técnicas de análisis de la NO RESPUESTA generando diferentes resultados.
- Presenta distintos niveles de conocimiento y habilidad para tratar con la NO RESPUESTA, desarrollando con esto análisis que van desde lo trivial hasta lo más complejo posible.

El tratamiento de la NO RESPUESTA consiste en ignorarla o bien imputarla; bajo la primera alternativa, se puede optar por realizar un análisis de casos completos sujeto a pérdida de información reportada o bien aplicar un estudio de casos incompletos bajo la presencia de limitaciones para realizar los estudios estadísticos tradicionales; en la segunda opción, se toma como base la información reportada para definir el valor por imputar, este puede ser extraído directamente de algún estadístico básico (media, moda, mediana, entre otros) de dicha información o mediante un modelo matemático soportado en variables que expliquen el origen de la NO RESPUESTA.

En particular, la imputación de datos pretende reducir el sesgo debido a la NO RESPUESTA y generar archivos de información completos; de esta manera, el tamaño de muestra será el mismo sin importar las variables que se elijan para un análisis estadístico específico; además, los usuarios generalmente conocen el proceso a desarrollar cuando la información está completa y cuentan con paquetes estadísticos diseñados exprofeso; finalmente, el imputar información oficial o pública, permite que el productor de la información incorpore conocimiento especializado acerca de la razón de la NO RESPUESTA dentro del propio procedimiento de imputación.

Es importante aclarar que aunque se sabe que siempre habrá una alternativa para tratar la NO RESPUESTA, lo ideal es buscar los medios para eliminar o al menos reducir su presencia.

1.2 El XII Censo General de Población y Vivienda, 2000

El XII CGPyV, 2000 presentó características metodológicas específicas:

- Fue un Censo de derecho (o *jure*), lo que significa censar a la Población en su lugar de residencia habitual.
- Un periodo de dos semanas para la captación de la información (del 7 al 18 de febrero del año 2000).
- Se captó la información a partir de una entrevista directa a un informante adecuado, definido como una persona de 15 o más años cumplidos, que viviera en la vivienda y que conociera los datos de todos los residentes habituales.
- La utilización de dos tipos de cuestionario: uno básico y otro ampliado y de un inventario de viviendas. El cuestionario ampliado se aplicó a una muestra probabilística de viviendas (denominada "Muestra Censal") y el básico a todas las viviendas restantes del país; por otra parte, el inventario sirvió para registrar datos de la propia vivienda que permitirían su ubicación e identificación.

En este sentido, las dos principales unidades de análisis del Censo son los residentes habituales[4] y las viviendas[5], al respecto, la población estimada en México en el año

[4] Un residente habitual es toda persona que vive normalmente en la vivienda, esto es, que en ella duerme, prepara sus alimentos, come y se protege del ambiente, y por ello la reconoce como su lugar de residencia

[5] Una vivienda es todo espacio delimitado normalmente por paredes y techos de cualquier material, con entrada independiente, que se utiliza para vivir, esto es, dormir, preparar los alimentos, comer y protegerse del ambiente.

2000 era de 100 millones de personas y de 20 millones de viviendas aproximadamente.

- **Temática Censal**

  La temática censal se estableció tomando en cuenta los resultados de estudios preliminares y los siguientes aspectos:

  - Prioridades de interés nacional.
  - Desglose geográfico de la información (insumo indispensable para la planeación en los ámbitos estatal y municipal).
  - Ausencia o deficiencia de información estadística.
  - Recomendaciones internacionales.
  - Comparabilidad histórica.

  Los temas generados se agruparon en tres grandes bloques: viviendas; número de residentes y de hogares; y características demográficas, sociales, educativas y económicas.

  El bloque de características de la vivienda incluye:

  - Tipo y clase de vivienda.
  - Materiales de construcción en paredes, techos y recubrimiento del piso.
  - Disponibilidad de espacios: total de cuartos, cuartos dormitorio y cocina.
  - Disponibilidad y **frecuencia del servicio**[6] de agua entubada.
  - Disponibilidad y exclusividad de servicio sanitario y conexión de agua.
  - Disponibilidad de drenaje y electricidad.
  - Combustible utilizado para cocinar.
  - Tenencia de la vivienda.
  - **Antigüedad de la vivienda**[6]**.**
  - **Eliminación de basura**[6]**.**
  - Bienes en la vivienda.

  El bloque del número de residentes habituales y de hogares en la vivienda maneja:

  - Total de residentes habituales de la vivienda.
  - Gasto común y número de hogares.

  El bloque de características demográficas, sociales, educativas y económicas de la Población se sub clasifica en:

  - Características demográficas:
    - ✓ Sexo, edad y relación de parentesco de los integrantes del hogar con el jefe(a) del mismo.
    - ✓ Fecundidad y mortalidad: número de hijos nacidos vivos, hijos fallecidos, hijos sobrevivientes, fecha de nacimiento del último hijo nacido vivo y, de éste, sobrevivencia y edad al morir.
    - ✓ Migración: lugar de nacimiento, lugar de residencia en 1995 (entidad o país y municipio o delegación) y **causa de la emigración**[6]**.**
    - ✓ **Migración internacional**[6]: el Censo captó la migración de las personas que se fueron a vivir a otro país entre enero de 1995 y el momento de la captación, y distingue a los migrantes que aún viven en otro país y a los que ya regresaron.
  - Características sociales:

[6] Tema adicional incluido en el cuestionario ampliado

- ✓ Étnicas: Población hablante de lengua indígena, condición de habla española, tipo de lengua y **pertenencia étnica[7].**
- ✓ Religión.
- ✓ Servicios de salud: derechohabiencia y **uso de servicios de salud[7].**
- ✓ Discapacidad: tipo y **causa de la discapacidad[7].**
- ✓ Estado conyugal.

▪ Características educativas:
- ✓ Alfabetismo, asistencia escolar, **causa de abandono escolar[7]**, nivel académico, antecedente escolar y nombre de la carrera.

▪ Características económicas:
- ✓ Condición de actividad, ocupación principal, situación en el trabajo, sector de actividad, ingresos por trabajo, horas trabajadas, **prestaciones laborales[7], lugar de trabajo[7] (municipio o delegación, entidad o país) y otros ingresos[7]** (ingreso que recibe en forma regular la población de 12 años y más proveniente de fuentes diferentes al desempeño de un trabajo).

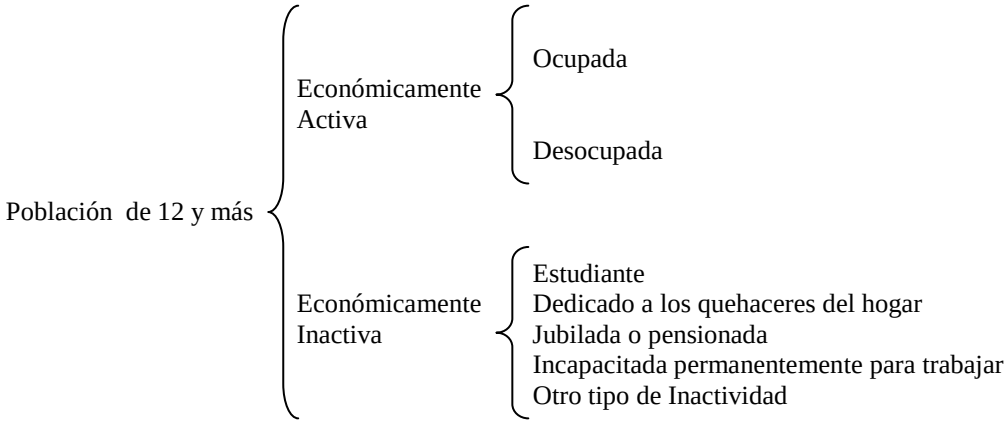
o **Los cuestionarios**

Definida la temática censal, se diseñaron los cuestionarios básico y ampliado; el primero, presenta 47 preguntas mientras que el segundo capta 23 preguntas adicionales para tener finalmente, un total de 70 (para consultar el cuestionario básico, ver Anexo B); nótese que, dado la estrategia de aplicación de ambos cuestionarios, las preguntas del cuestionario básico están incluidas dentro del cuestionario ampliado.

Las preguntas que integran a los cuestionarios, son representadas mediante una serie de variables que permiten conformar una base (o tabla) de datos para cada uno de los bloques ya mencionados; en el Anexo C, se presenta la tabla de datos correspondiente al bloque de características demográficas, sociales, educativas y económicas de la Población para el cuestionario básico.

o **La captación y validación del ingreso**

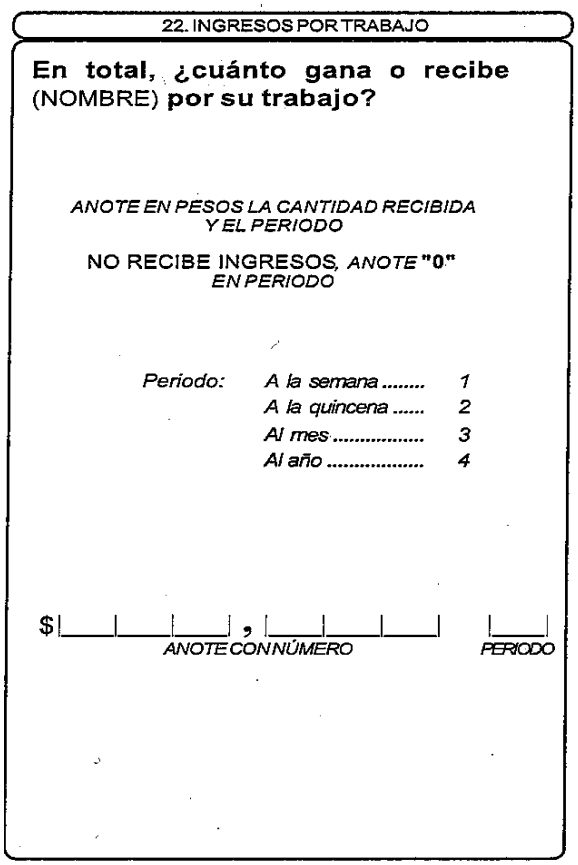
Para identificar las características económicas de la Población, las preguntas correspondientes se aplicaron a las personas de 12 años y más, bajo esta situación, este grupo poblacional se puede clasificar en:



Gráfica 1.1: Clasificación de la Población de 12 años y más según características económicas

[7] Tema adicional incluido en el cuestionario ampliado

En la Gráfica 1.1, se observa que la Población Económicamente Activa Ocupada (o simplemente Población Ocupada) es el universo económico en que se estudia el comportamiento del ingreso y de la frecuencia (el periodo) con que se recibe.
Para la captación del ingreso por trabajo y del periodo en que se recibe se emplea la pregunta 22 del cuestionario (Anexo B), según se aprecia en la Gráfica 2.1.


22. INGRESOS POR TRABAJO

En total, ¿cuánto gana o recibe (NOMBRE) **por su trabajo?**

ANOTE EN PESOS LA CANTIDAD RECIBIDA Y EL PERIODO

NO RECIBE INGRESOS, *ANOTE* **"0"** *EN PERIODO*

Periodo:
A la semana 1
A la quincena 2
Al mes 3
Al año 4

$|___|___|___| , |___|___|___| |___|
ANOTE CON NÚMERO *PERIODO*

Gráfica 1.2: Ingresos por trabajo Pregunta 22 del cuestionario básico.

Durante la entrevista, el ingreso por trabajo puede presentar valores específicos posibles:

- Entre 0 y 999,999; donde la cota superior es NO RESPUESTA dada por el informante.
- Blanco (b) que puede ser por pase de pregunta (menores de 12 años, desocupados o inactivos) o bien porque la persona no recibe ingresos.

Por su parte, el periodo presenta un comportamiento similar:

- Códigos de respuesta: 1, 2, 3 o 4
- Código de NO RESPUESTA: 9 o blanco (b) por pase.

Dentro de la revisión de la consistencia a la base de datos ya capturada, se aplicaron diversos criterios de validación en forma automatizada, en particular, tres de ellos podían asignar el código de NO RESPUESTA al ingreso por trabajo, los casos fueron:

- ✓ Asignar al ingreso el valor de NO ESPECIFICADO (999,999) cuando éste venía en blanco para la Población Ocupada.
- ✓ Asignar al ingreso el valor de cero (0) o de NO ESPECIFICADO (999,999) dependiendo de la declaración de las variables situación en el trabajo y periodo en el que recibe el ingreso.
- ✓ Asignar al ingreso el valor de NO ESPECIFICADO (999,999) cuando el periodo declarado era NO ESPECIFICADO (9) o venía en blanco.

Una vez culminada la etapa de validación, el ingreso por trabajo se recalculó en forma mensual apoyándose en el periodo reportado, los criterios aplicados se muestran en la Tabla 1.1.

Periodo	**Variable**	**Ingreso mensual**
Mensual	PERINGRE=3	INGRESOS
Semanal	PERINGRE=1	(INGRESOS/7)*30
Quincenal	PERINGRE=2	INGRESOS*2
Anual	PERINGRE=4	INGRESOS/12
NO ESPECIFICADO	PERINGRE=9	999,999

Tabla 1.1: Criterios para mensualizar el ingreso por trabajo

El valor calculado fue redondeado y en caso de que rebasara la cantidad 999,999 se asignó el código 999,998.

Al revisar el resultado de la validación automática y de la mensualización se encontró que el nivel de NO ESPECIFICADO del ingreso resultaba muy elevado, por lo que se procedió a examinar los posibles motivos de este comportamiento.

El ingreso se modificó por el código de NO ESPECIFICADO, principalmente, en las situaciones siguientes:

- ✓ Ingreso especificado pero no se tuvo la frecuencia con que se recibía, es decir, no se indicó el periodo.
- ✓ Ingreso en blanco para la Población Ocupada
- ✓ Ingreso cero o el periodo en que se recibe fue cero y no se trataba de un trabajador sin pago en el negocio o predio familiar.

La frecuencia con la cual se aplicaron estos tratamientos reportó que el último caso presentaba la tasa mayor de cambio, este criterio daba por hecho que la declaración sobre la situación del trabajo estaba asociada con el ingreso cero sólo cuando se trataba de trabajadores sin pago, y que todos aquéllos como empleado u obrero, jornalero o peón, patrón y trabajador por su cuenta, necesariamente tenían que recibir un ingreso por su trabajo.

Esta situación se revisó y finalmente se acordó en dar de baja el criterio y reasignar el ingreso cero declarado para todos aquellos que su situación en el trabajo no era trabajadores sin pago.

Después de este ajuste, la NO RESPUESTA en el ingreso continuó siendo considerada como alta, por lo que se decidió realizar una imputación como alternativa para reducir o corregir esta NO RESPUESTA.

1.3 Imputación del INEGI

La imputación del ingreso por trabajo (mensualizado) de la Población Ocupada, que desarrolló el INEGI, se apoyó en las siguientes variables predictoras:

- ✓ Sexo.
- ✓ Parentesco.
- ✓ Edad.
- ✓ Nivel académico.
- ✓ Situación en el trabajo.
- ✓ Ocupación principal.

El procedimiento realizado consistió en los siguientes pasos:

1. Para la Población Ocupada incluida en la base de datos censal, se filtraron aquellos registros con respuesta especificada en las variables predictoras, generando una Población Ocupada "especificada".
2. Las variables predictoras se asociaron de acuerdo con sus códigos de respuesta, a cada combinación generada se le llamó "imagen". Cabe mencionar que, previo a la asociación, algunas de las variables fueron agrupadas bajo criterios predefinidos (por ejemplo, parentesco de tres dígitos se agrupó en uno), reduciendo así el número total de combinaciones posibles a una cantidad razonable. Cada imagen obtenida se identificó con un número entero único.
3. La Población Ocupada "especificada" se agrupó en función de las imágenes generadas.
4. Para cada grupo conformado (uno para cada imagen), se obtuvo el ingreso modal particular excluyendo y contabilizando aquellos registros con el ingreso NO ESPECIFICADO.
5. Todas las imágenes fueron ordenadas en forma descendente acorde con la frecuencia del ingreso NO ESPECIFICADO; se definió como imágenes a imputar al 70 % de los casos con mayor frecuencia.
6. A los registros con el ingreso NO ESPECIFICADO de las imágenes a imputar, se les asignó el ingreso modal de su imagen correspondiente.

Debe notarse que aún después de aplicar este procedimiento de imputación, la variable ingresos por trabajo siguió presentando NO RESPUESTA (en un nivel máximo del 30% de las imágenes construidas) en la base de datos censal (para tener un mayor detalle del procedimiento, ver el Anexo D).

Detección de los registros imputados

Por la propia naturaleza de este trabajo, el ubicar los registros imputados por el INEGI, se vuelve una cuestión por demás trascendente, al grado incluso de provocar la cancelación de la investigación en caso de no lograr detectarlos plenamente.

A diferencia de lo que se pudiera pensar, esta actividad resultó ser una tarea complicada, puesto que el INEGI no identificó de alguna forma a los registros que imputó y los respaldos de información generados durante el proceso de imputación resultaron extraviados, además, se presentaron otras adversidades como son:

- El personal que desempeñó las actividades principales de la imputación ya no trabajaba para el INEGI durante el periodo de la investigación.

- El hecho de que el INEGI fuera reestructurado después del año 2000, implicó que los elementos empleados (bases de datos, programas de cómputo, entre otros) en la imputación fueran difíciles de ubicar.

La detección de los registros imputados requirió recopilar los insumos empleados por el INEGI y fue necesario desarrollar algunos procesos informáticos que se basaron en:

- Comparar una base de datos empleada como insumo por el INEGI que incluía a las seis variables predictoras, los registros imputados y un grupo de registros con ingreso cero contra la base de datos censal.
- Para los registros no coincidentes, aplicar tablas de actualización cartográfica, para su correcta comparación.

Con lo anterior, se logró ubicar bajo condiciones aceptables 21,854 registros de los 21,865 imputados (99.95 %); finalmente, como prueba de evaluación del proceso de búsqueda, se comparó contra la base de datos generada durante la etapa de codificación, corroborándose que los registros definidos como imputados por el proceso de detección tenían en dicha base el código de NO RESPUESTA.

1.4 Métodos de imputación

Cuando se va a realizar una imputación, es recomendable investigar previamente el origen y el tipo de la NO RESPUESTA a imputar, ya que estos dos aspectos están fuertemente relacionados con la selección del método a emplear.

- El origen de la NO RESPUESTA se puede clasificar en tres casos según el mecanismo que generó su presencia:
  - NO RESPUESTA completamente aleatoria (MCAR). Se da cuando la probabilidad de que el valor de una variable X_j, para que sea observado para un individuo i, no depende ni del valor de esa variable, x_{ij} ni del valor de las demás variables consideradas $x_{ik}, k \neq j$; es decir, la NO RESPUESTA no es originada por las variables presentes en la matriz de datos. Por ejemplo, en el caso de tener en un estudio las variables ingreso y edad, estaremos bajo un modelo MCAR cuando al analizar conjuntamente edad e ingresos, suponemos que la falta de respuesta es independiente del verdadero valor de los ingresos y la edad, en símbolos:

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

    Donde *R* es la variable indicadora de respuesta de la variable Ingresos y valdrá 1 en el caso de haber respuesta y 0 en otro caso
  - NO RESPUESTA aleatoria (MAR). Se da cuando la probabilidad de que el valor de una variable X_j, para que sea observado para un individuo i, no depende del valor de esa variable, x_{ij} pero quizás sí del valor que toma alguna otra variable observada $x_{ik}, k \neq j$; es decir, la NO RESPUESTA está asociada a variables presentes en la matriz de datos. En el ejemplo anterior si suponemos que los ingresos son independientes de los ingresos del miembro

del hogar pero puede depender de la edad estaremos bajo un modelo MAR; en términos de una ecuación:

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

- NO RESPUESTA no aleatoria (NMAR). Se da cuando la probabilidad de que el valor de una variable X_j, para que sea observado para un individuo i, depende del valor de esa variable, x_{ij} siendo este valor desconocido. En el ejemplo, se obtiene que la función respuesta de la variable ingresos depende del propio valor de la variable ingresos, además de poder depender de otros factores; es decir:

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

- El tipo de NO RESPUESTA, consiste en considerar que la NO RESPUESTA se puede agrupar según el número de variables en las que esté presente, de forma general existen dos tipos:
  - NO RESPUESTA por unidad. La integran aquellos registros que presentan NO RESPUESTA en todas las variables; se da cuando no hay contacto con el informante o bien se rechaza la entrevista.
  - NO RESPUESTA por variable. La definen aquellos registros que presentan NO RESPUESTA en cierta(s) variable(s), pero no en todas.

En general, la NO RESPUESTA por unidad podrá ser excluida exceptuando el caso de datos provenientes de una encuesta lo cual implica que se ajusten los factores de expansión mediante alguna técnica de elevación (weighthing).

Existen dos formas de actuar ante la presencia de la NO RESPUESTA por variable; la primera opción, consiste en ignorarla (aunque debe reportarse su presencia) y usar únicamente la información reportada; la segunda alternativa, implica corregirla mediante la aplicación de algún método de imputación.

- **Métodos que emplean la información reportada (ignorando la NO RESPUESTA)**
  Estos métodos consisten en la eliminación de registros donde se presente la NO RESPUESTA y aunque pueden ser útiles cuando la NO RESPUESTA es pequeña, generalmente si el mecanismo de generación es MAR o NMAR conducen a estimaciones sesgadas. Existen dos métodos por aplicar.
  - Eliminación por lista
    Es una solución conservadora. Consiste en emplear solamente los registros que tengan respuesta en todas las variables. Las ventajas de este método son su simplicidad, su constante presencia en los paquetes estadísticos y la posibilidad de comparar los estadísticos univariados, dado que se obtienen

con la misma información reportada. Los inconvenientes son: los análisis pierden potencia al reducirse el número de elementos y se desperdicia una importante cantidad de información reportada por el informante.

- Eliminación por pares
  Sólo se emplean los registros que tienen respuesta en todas las variables analizadas para un cálculo específico, los registros con NO RESPUESTA se eliminan, además, este método tiene la desventaja de no poder asegurar que la matriz de correlaciones sea positiva definida, condición indispensable para invertir dicha matriz; por otra parte, los posibles cálculos estadísticos a realizar se generan con diferentes tamaños de muestra lo que limita la comparación de resultados.

- **Métodos típicos de Imputación**

Laaksonen (2000) clasifica a los métodos de imputación en tres tipos:

A. Imputación deductiva o lógica.
B. Imputación empleando un modelo donante.
C. Imputación empleando un registro donante.

A continuación, se detalla la clasificación anterior exponiendo algunos métodos de imputación incluidos en cada caso.

A. Imputación deductiva o lógica
   Es un método determinístico, el cual consiste en asignar valores tras definir con un cierto grado de certidumbre a la información reportada y con ello, construir funciones entre dicha información y la NO RESPUESTA. Actualmente, este método se aplica en registros en los que la NO RESPUESTA se puede deducir a partir de los valores del resto de variables del mismo registro. Una imputación deductiva toma generalmente el siguiente formato condicional:

   Sí (condición) entonces (acción)

B. Imputación empleando un modelo donante
   Estos métodos asignan valores que son extraídos a partir de modelos matemáticos construidos con la información reportada; debido a su propia definición, es posible que el valor por asignar no esté incluido en el rango de respuesta de la información reportada. Sin pérdida de generalidad, se considera que los modelos se generan según un procedimiento de regresión estadística.
   <u>Procedimientos de Regresión</u>
   El modelo se obtiene a partir de aplicar una Regresión Simple o Múltiple; en este caso, el origen de la NO RESPUESTA supone un mecanismo de generación MAR.
   - ✓ Imputación de la Media
     Es el modelo más sencillo de los pertenecientes a los procedimientos de regresión. Consiste en asignar el valor medio de la información reportada a todos los valores con NO RESPUESTA. Tiene como desventajas que modifica la distribución original de la variable a imputar y subestima su

varianza, es decir, no conserva la relación entre las variables ni la distribución de frecuencias original.

Presenta como consecuencia de su aplicación que la media de la información completa (reportada e imputada) con o sin reemplazo de la NO RESPUESTA siga siendo la media de la información reportada.

- ✓ Regresión Múltiple

  La variable dependiente es imputada en función de variables regresoras, las cuales pueden ser cualitativas o cuantitativas, generalmente estas variables están altamente correlacionadas con la variable dependiente. En particular, a todos los registros con los mismos valores en las variables regresoras se les asigna el mismo valor presentándose en parte las desventajas del método anterior.
- ✓ Regresión Logística

  Método aplicable a variables dependientes binarias.

Los siguientes métodos de regresión modifican en menor medida la distribución de la variable a imputar y subestiman en menor grado la varianza.

- ✓ Regresión Aleatoria

  Este método resuelve el problema de la distorsión de la distribución tras la imputación y se basa en añadir una perturbación aleatoria a las estimaciones generadas por el modelo de Regresión Simple o Múltiple.
- ✓ Regresión Logística Aleatoria

  Se conforma incluyendo una perturbación aleatoria al modelo generado por la Regresión Logística.

C. Imputación empleando un registro donante

El valor por asignar al registro a imputar (candidato o receptor) es extraído de un registro completo (donante) incluido en la información reportada; usualmente y previo a la imputación, la información es conformada en grupos (o estratos).

Entre las ventajas de estos métodos se pueden destacar:

- ✓ El valor imputado está contenido dentro del rango de respuesta de la información reportada.
- ✓ Son sencillos de implementar.

La principal desventaja radica en que conserva el rango de respuesta de la información reportada lo cual no es adecuado cuando dicho rango difiere con el rango de respuesta de diseño de la variable.

Existe un gran número de métodos entre los que se destacan los siguientes:

- ✓ <u>Cold-Deck</u>

  Se define un donante por grupo (o estrato) a partir de fuentes de información externas: datos históricos, distribuciones de frecuencias, entre otras. El método asigna a la NO RESPUESTA del candidato los valores del donante correspondiente al mismo grupo (o estrato). Una desventaja adicional de este método es que la calidad de los resultados dependerá de la calidad de la información externa disponible.
- ✓ <u>Hot-Deck</u>

  Toda la información, se clasifica en grupos disjuntos, de tal forma que sean lo más homogéneos posibles; a la NO RESPUESTA de cada

candidato, se le asigna el valor de un donante particular del mismo grupo. Este método supone que dentro de cada grupo, la NO RESPUESTA sigue la misma distribución que la de la información reportada, si esta hipótesis no es cierta sólo se reducirá parte el sesgo debido a la NO RESPUESTA; además, tiene algunas características interesantes a destacar:

- No se presentan problemas al momento de ajustar varias variables.
- No se requiere de fuertes supuestos para estimar los valores individuales de la NO RESPUESTA.
- Conserva la distribución inicial de la variable a imputar.

Sin embargo, tiene algunas desventajas:

- Distorsiona la relación con el resto de las variables.
- Carece de un mecanismo probabilístico para seleccionar al donante.
- Requiere tomar decisiones subjetivas que afectan a la calidad de los datos, lo que imposibilita calcular su confianza.
- Los grupos deben ser definidos en base a un número reducido de variables, con la finalidad de asegurar que habrá suficientes donantes en todos los grupos.
- Existe la posibilidad de usar varias veces al mismo donante.

✓ Hot-Deck secuencial

El donante, además de pertenecer al mismo grupo (o estrato), debe ser el inmediato anterior al candidato según el orden prefijado al momento de conformar el grupo correspondiente, en esta imputación la estratificación previa debe producir una autocorrelación positiva entre las variables sujetas a la imputación, de esta forma se asegura una mayor similitud entre el donante y el candidato. Las desventajas de este método son:

- Hay que facilitar valores iníciales cuando el primer registro es candidato.
- Ante una racha continua de candidatos, se emplea el mismo donante.
- Es difícil de estudiar la precisión de las estimaciones.

✓ Hot-Deck con donante aleatorio

Consiste en elegir aleatoriamente a uno o varios donantes para cada candidato; hay diferentes versiones de este método, el caso más simple es elegir aleatoriamente un donante e imputar el candidato con dicha información, también se puede elegir una muestra de donantes mediante distintos tipos de muestreo e imputar al valor medio obtenido con todos ellos.

✓ Hot-Deck modificado

Consiste en clasificar y ajustar los donantes potenciales y candidatos utilizando un considerable número de variables. El ajuste se hace sobre bases jerárquicas.

✓ Donor

Se emplea una función distancia definida entre las variables para que se mida el grado de proximidad entre cada posible donante y el candidato. En este caso se asignan en bloque los valores del donante en la NO

RESPUESTA del candidato. Es necesaria una transformación previa de los datos para anular los efectos de escala en la función distancia.

- **Nuevos métodos de imputación**

  Recientemente, se han desarrollado nuevos métodos en el área del análisis de la NO RESPUESTA, entre los cuales destacan:

  - Máxima Verosimilitud

    Asumiendo que la información reportada sigue un modelo multivariado particular, se aplica el principio de máxima verosimilitud el cual es simple pero presenta en términos computacionales una solución compleja; al respecto, el algoritmo EM (Dempster et al., 1977) permite resolver numéricamente y en forma iterativa el cálculo del estimador máximo verosímil mediante dos pasos:

    - ✓ Paso E (Valor esperado). Calcula el valor esperado de la información reportada basándose en la función de verosimilitud.
    - ✓ Paso M (Maximización). Se asigna a los datos con NO RESPUESTA el valor esperado obtenido en el paso anterior (E) y entonces se calcula de nueva cuenta la función de máxima verosimilitud pero ahora considerando todos los registros (reportados e imputados).
  - Regresión Secuencial

    Este método se considera como una generalización de los procedimientos de regresión, ya que permite imputar varias variables a la vez; estas, pueden tener distribuciones diferentes (continuas, categóricas, enteras o dicotómicas, entre otras). La técnica se basa en un proceso iterativo en que se combina una secuencia de modelos de regresión y se usan los valores imputados en cierta variable particular para la predicción de la NO RESPUESTA de las otras variables (Raghunathan et al., 2001b).
  - Imputación Múltiple

    Partiendo del hecho de que la Imputación Simple (generada por los métodos que sólo pueden asignar un valor) tiende a sobreestimar la precisión, ya que no se toma en cuenta la variabilidad de las componentes entre las distintas imputaciones realizadas. La Imputación Múltiple, sustituye a la NO RESPUESTA por m (m>1) valores obtenidos a través de la estimación de un modelo aleatorio apropiado. Posteriormente, se lleva a cabo el análisis estadístico ordinario con los m archivos completos, acto seguido, se combinan los resultados empleando fórmulas específicas (Little y Rubin, 1987). Asume un mecanismo de generación tipo MAR.
  - Redes Neuronales. Árboles de Clasificación y de Regresión. Lógica Difusa

    Técnicas propuestas en proyectos europeos (AUTIMP[8] y EUREDIT[9]) y que se están desarrollando en la actualidad (Puerta, 2002).

---

[8] Acrónimo de las palabras "automatic" e "imputation". AUTIMP, es un sistema de cómputo para la imputación automática en encuestas económicas y censos de población. Fue financiado por la Sociedad de Información Tecnológica dentro del 4º Framework Programme de la Unión Europea.

[9] Acrónimo de las palabras "europe" y "editing"; EUREDIT, es un proyecto que desarrolla y evalúa los nuevos métodos de validación e imputación. Es financiado por la Sociedad de Información Tecnológica dentro del 5th Framework Programme de la Unión Europea http://www.cs.york.ac.uk/euredit

El siguiente capítulo, se dedica exclusivamente a describir con mayor detalle los métodos de Regresión Secuencial y de Imputación Múltiple ya que son las técnicas que se van a emplear en esta investigación. En el capítulo 3, se habla sobre el sistema de cómputo IVEware que implementa ambas técnicas, se explican, entre otras cosas, la estructura, las capacidades, el funcionamiento y cómo se usa el sistema para imputar. En el capítulo 4, se detalla cómo IVEware fue aplicado para imputar el ingreso por trabajo de la Población Ocupada del Estado de Aguascalientes. En el capítulo 5, se presentan los principales resultados de la aplicación y finalmente, en el capítulo 6, se mencionan las conclusiones relevantes generadas por la investigación.

Capítulo 2

La Regresión Secuencial y la Imputación Múltiple

Este capítulo permite conocer el proceso iterativo en el que se basa la Regresión Secuencial, además, explica el funcionamiento de la Imputación Múltiple y dado que ambas técnicas suelen aplicarse de manera conjunta para imputar la NO RESPUESTA (método RSIM), se incorpora una explicación que ilustra dicha combinación.

2.1 Regresión Secuencial (RS)

Esta técnica es considerada una extensión de los métodos de regresión, ya que para imputar una variable en particular usa a todas las variables incluidas en la investigación, es decir, emplea los valores ya imputados en cierta(s) variable(s) para predecir la NO RESPUESTA de otra variable, logrando con esto preservar la estructura de la matriz de covarianzas. Por otra parte, el método permite manejar escalas diversas en la(s) variable(s) a imputar, incluyendo la lineal, dicotómica, politómica y entera; también acepta el uso de restricciones que definen el número de casos a ser imputados y avala el empleo de cotas que delimitan el rango de la imputación; estos dos últimos aspectos son aplicados a nivel variable (Raghunathan et al., 2001b).

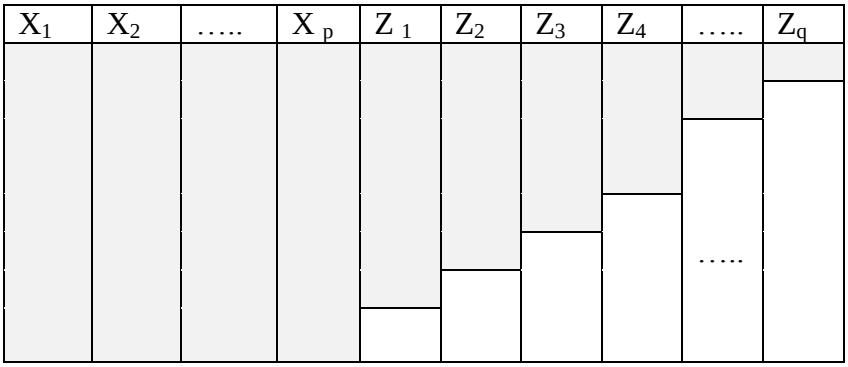
La RS, emplea un procedimiento secuencial basado en reproducir **C** ciclos para imputar varias variables a la vez; además, usa métodos bayesianos para perturbar los coeficientes de regresión empleados en la estimación agregando residuales aleatorios (Raghunathan et al., 2001b).

Enseguida, se explica la Regresión Secuencial apoyándose en ecuaciones matemáticas y herramientas gráficas.

Inicialmente, las variables de la investigación se clasifican en:

- *p* variables sin presencia de NO RESPUESTA (denotadas con la letra X).
- *q* variables con presencia de NO RESPUESTA (denotadas con la letra Z).

Sin pérdida de generalidad se asume que Z_1 y Z_q presentan la menor y la mayor NO RESPUESTA respectivamente, de manera gráfica se tiene:



Gráfica 2.1: Regresión Secuencial (clasificación inicial)

En la Gráfica 2.1, el tono sombreado representa a la información reportada por el informante o bien generada a partir de la aplicación de algún criterio de validación (imputación lógica o deductiva) y el tono blanco representa la información a

imputar. Cada una de las variables Z_i (i=1,…, q) define horizontalmente dos zonas: una de respuesta (r) y otra de NO RESPUESTA (m); nótese que la partición en principio, tiene sentido reflejarla sobre las variables X_j (j=1,…, p) pero después lo tendrá sobre las variables Z_i conforme vayan siendo imputadas.
Combinando las dos consideraciones anteriores, cada una de las variables Z_i requiere del uso de una notación bidimensional; así por ejemplo, para la variable Z_1, se tiene que:

$z_{r,1}$, representa la zona con respuesta en la variable Z_1.

$z_{m,1}$, representa la zona de NO RESPUESTA de la variable Z_1.

$x_{r,1}$, representa la zona con respuesta de X_1, definida por la zona de respuesta de Z_1.

$x_{m,1}$, representa la zona con respuesta de X_1, definida por la zona de NO RESPUESTA de Z_1.

$x_{r,2}$, representa la zona con respuesta de X_2, definida por la zona de respuesta de Z_1.

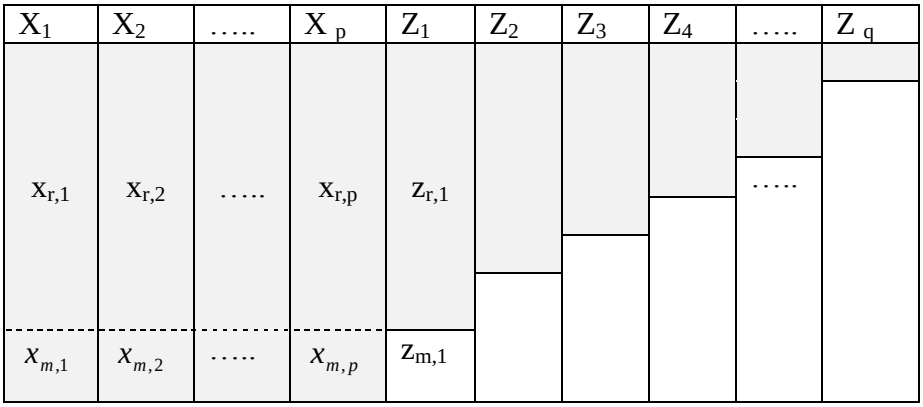
$x_{m,2}$, representa la zona con respuesta de X_2, definida por la zona de NO RESPUESTA de Z_1.

.
.
.

$x_{r,p}$, representa la zona con respuesta de X_p, definida por la zona de respuesta de Z_1.

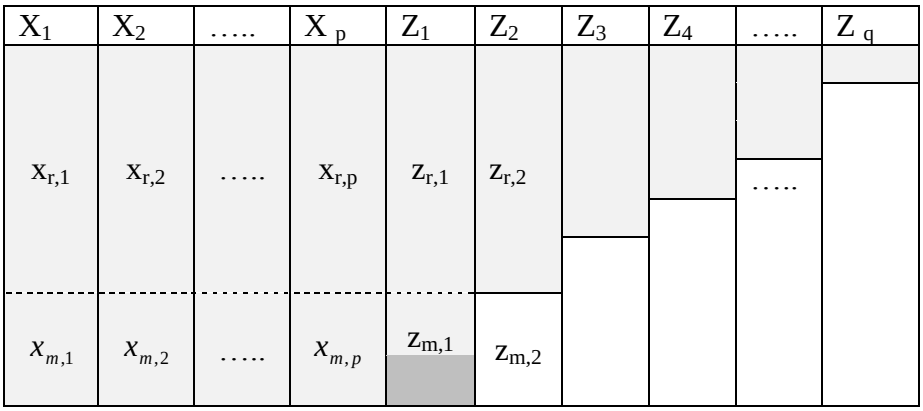
$x_{m,p}$, representa la zona con respuesta de X_p, definida por la zona de NO RESPUESTA de Z_1.

Esta notación es idéntica para el resto de las variables Z_i; aunque una vez que una variable Z_i en particular haya sido imputada podrá ser considerada como una variable sin presencia de NO RESPUESTA (X_p) con lo que el listado anterior se incrementa. El resultado de la partición para la variable Z_1, se muestra en la Gráfica 2.2.



Gráfica 2.2: Regresión Secuencial (partición empleando a la variable Z_1)

Y en la gráfica 2.3, lo referente a la variable Z_2.



Gráfica 2.3: Regresión Secuencial (partición empleando a la variable Z_2)

El ciclo uno consta de imputar cada una de las variables Z_i (i=1,…, q) iniciando con Z_1 (la de menor NO RESPUESTA) y terminando con Z_q (la de mayor NO RESPUESTA); primero, se genera un modelo de regresión (función *f*) con la información reportada de cada variable Z_i considerando inicialmente a las variables X_j como predictoras pero incorporando a las variables Z_i conforme vayan siendo imputadas; posteriormente se estima la zona de NO RESPUESTA correspondiente con dicho modelo, en notación matemática se tiene:

- para Z_1:

  $z_{r,1} = f_{1(1)}\left(x_{r,1}, x_{r,2}, ..., x_{r.p} \right)$ (modela para obtener la función *f*).

  $\hat{z}_{m,1} = \hat{f}_{1(1)}\left(x_{m,1}, x_{m,2}, ..., x_{m,p} \right)$ (estima bajo *f*, estocásticamente).

- para Z_2:

  $z_{r,2} = f_{2(1)}\left(x_{r,1}, x_{r,2}, ..., x_{r,p}, z_{r,1} \right)$ (obtiene el modelo e incorpora a Z_1).

  $\hat{z}_{m,2} = \hat{f}_{2(1)}\left(x_{m,1}, x_{m,2}, ..., x_{m,p}, \hat{z}_{m,1} \right)$ (estima incorporando a Z_1).

.
.
.
.
.

- para Z_q:

  $z_{r,q} = f_{q(1)}\left(x_{r,1}, x_{r,2}, ..., x_{r,p}, z_{r,1}, ..., z_{r,q-1} \right)$ (modela e incorpora a $Z_{1,\ldots,}Z_{q-1}$).

  $\hat{z}_{m,q} = \hat{f}_{q(1)}\left(x_{m,1}, x_{m,2}, ..., x_{m,p}, \hat{z}_{m,1}, ..., \hat{z}_{m,q-1} \right)$ (estima, incorpora a $Z_{1,\ldots,}Z_{q-1}$).

En la Gráfica 2.4, se ilustra la situación de la tabla de datos para la variable Z_1.

X_1	X_2	…..	X_p	Z_1	Z_2	Z_3	Z_4	…..	Z_q
$x_{r,1}$	$x_{r,2}$	…..	$x_{r,p}$	$z_{r,1}$				…..	
$x_{m,1}$	$x_{m,2}$	…..	$x_{m,p}$	$\hat{z}_{m,1}$					

Gráfica 2.4: Regresión Secuencial (después del ciclo uno, variable Z_1)

La gráfica 2.5, presenta el comportamiento de la tabla de datos después de imputar a la variable Z_2.

X_1	X_2	…..	X_p	Z_1	Z_2	Z_3	Z_4	…..	Z_q
$x_{r,1}$	$x_{r,2}$	…..	$x_{r,p}$	$z_{r,1}$	$z_{r,2}$			…..	
$x_{m,1}$	$x_{m,2}$	…..	$x_{m,p}$	$\hat{z}_{m,1}$	$\hat{z}_{m,2}$				

Gráfica 2.5: Regresión Secuencial (después del ciclo uno, variable Z_2)

Hasta este momento se han imputado todas las variables posibles; con esto, se tiene un primer archivo con casos completos.

Los ciclos dos a **C** equivalen a repetir el ciclo uno **C-1** veces, pero considerando como variables predictoras a todas las variables de la investigación; es decir, además de considerar a las variables X_j se incluyen todas las variables Z_i adoptando la estimación del ciclo inmediato anterior, matemáticamente se tiene:

- para Z_1:

  $z_{r,1} = f_{1(C)}\left(x_{r,1}, x_{r,2},..., x_{r,p}, \hat{z}_{r,2}, \hat{z}_{r,3},..., \hat{z}_{r,q} \right)$ (modela la función *f*).

  $\hat{z}_{m,1} = \hat{f}_{1(C)}\left(x_{m,1}, x_{m,2},..., x_{m,p}, \hat{z}_{m,2}, \hat{z}_{m,3},..., \hat{z}_{m,q} \right)$ (estima estocásticamente).

- para Z_2:

$$z_{r,2} = f_{2(C)}\left(x_{r,1}, x_{r,2}, \ldots, x_{r,p}, z_{r,1}, \hat{z}_{r,3} \ldots, \hat{z}_{r,q} \right)$$

$$\hat{z}_{m,2} = \hat{f}_{2(C)}\left(x_{m,1}, x_{m,2}, \ldots, x_{m,p}, \hat{z}_{m,1}, \hat{z}_{m,3}, \ldots, \hat{z}_{m,q} \right)$$

.
.
.

- para Z_q:

$$z_{r,q} = f_{q(C)}\left(x_{r,1}, x_{r,2}, \ldots, x_{r,p}, z_{r,1}, \ldots, z_{r,q-1} \right)$$

$$\hat{z}_{m,q} = \hat{f}_{q(C)}\left(x_{m,1}, x_{m,2}, \ldots, x_{m,p}, \hat{z}_{m,1}, \ldots, \hat{z}_{m,q-1} \right)$$

La Gráfica 2.6 presenta el proceso para la variable Z_1.

X_1	X_2	…..	X_p	Z_1	Z_2	Z_3	Z_4	…..	Z_q
$x_{r,1}$	$x_{r,2}$	…..	$x_{r,p}$	$z_{r,1}$	$\hat{z}_{r,2}$	$\hat{z}_{r,3}$	$\hat{z}_{r,4}$	…..	$\hat{z}_{r,q}$
$x_{m,1}$	$x_{m,2}$	…..	$x_{m,p}$	$\hat{z}_{m,1}$	$\hat{z}_{m,2}$	$\hat{z}_{m,3}$	$\hat{z}_{m,4}$		$\hat{z}_{m,q}$

Gráfica 2.6: Regresión Secuencial (variable Z_1, ciclos dos a **C**)

La gráfica 2.7 ilustra lo correspondiente a la variable Z_2.

X_1	X_2	…..	X_p	Z_1	Z_2	Z_3	Z_4	…..	Z_q
$x_{r,1}$	$x_{r,2}$	…..	$x_{r,p}$	$z_{r,1}$	$z_{r,2}$	$\hat{z}_{r,3}$	$\hat{z}_{r,4}$	…..	$\hat{z}_{r,q}$
$x_{r,1}$	$x_{r,2}$	…..	$x_{r,p}$	$\hat{z}_{m,1}$	$\hat{z}_{m,2}$	$\hat{z}_{m,3}$	$\hat{z}_{m,4}$		$\hat{z}_{m,q}$

Gráfica 2.7: Regresión Secuencial (variable Z_2, ciclos dos a **C**)

Del proceso anterior, se tienen **C** archivos completos; sin embargo, sólo el valor generado en el último ciclo es adoptado como la imputación de la Regresión Secuencial.

El modelo de regresión (función *f*) se obtiene de acuerdo con el tipo de variable al que corresponde cada una de las variables a imputar, luego se tiene que si:

- Z_i es continua, se aplica una regresión lineal normal.
- Z_i es binaria, se aplica una regresión logística.
- Z_i es categórica, se aplica una regresión politómica o una regresión logística generalizada.
- Z_i es discreta, se aplica un modelo Poisson log lineal.
- Z_i es mixta o semicontinua (toma un valor 0 o entre un mínimo y un máximo), se aplica un modelo de dos etapas; para definir el estado 0 o no 0, se usa un modelo de regresión logística y para valores imputados con un valor distinto de cero se emplea una regresión lineal normal.

2.2 Imputación Múltiple (IM)

La Imputación Múltiple consiste en repetir *m* veces un algoritmo de imputación que por su propio diseño sólo puede asignar un único valor ("Imputación Simple"). Así, se obtienen *m* conjuntos de datos completos que son analizados en forma individual; posteriormente, los resultados se combinan para obtener estimaciones definitivas a nivel global (Little y Rubin, 1987).

Suponiendo que se desea estimar el parámetro θ (media, proporción o coeficiente de regresión, entre otros) y dada la aplicación de una imputación simple *m* veces, se tienen $\hat{\theta}_k$ estimaciones de θ y $\hat{W}_k$ estimaciones respectivas de la varianza de $\hat{\theta}_k$ (con k=1,.., m) obtenidas según el diseño particular de la investigación.

Se tiene que la estimación de la Imputación Múltiple para θ, está dada por:

$$\hat{\theta}_{IM} = \frac{1}{m}\sum_{k=1}^{m}\hat{\theta}_k$$

Y que la estimación de la varianza correspondiente se obtiene con:

$$\hat{\text{var}}(\theta)_{IM} = \frac{1}{m}\sum_{k=1}^{m}\hat{W}_k + \left(1+\frac{1}{m}\right)\left(\frac{1}{m-1}\right)\sum_{k=1}^{m}\left(\hat{\theta}_k - \hat{\theta}_{IM}\right)^2$$

Nótese que para considerar el efecto de la imputación en la estimación de la varianza total, la variación dentro de cada conjunto (primer sumando) es combinada con la variación entre los *m* conjuntos (parte del segundo sumando). De acuerdo con esta observación, la expresión anterior, suele escribirse de manera económica como:

$$\hat{\text{var}}(\theta)_{IM} = \bar{W}_k + \left(1+\frac{1}{m}\right)B_k$$

Donde al término $\bar{W}_k$ se le conoce como la varianza dentro (within) de la imputación y a B_k, como la varianza entre (between) la imputación.

Este método permite además, obtener intervalos de confianza para el parámetro θ, bajo la siguiente expresión:

$$\hat{\theta}_{IM} \pm t_{\nu,1-\alpha/2} \sqrt{\hat{\text{var}}(\theta)_{IM}}$$

Los grados de libertad ν, se obtienen de:

$$\nu = \left(m-1\right)\left(\frac{\left(1+1/m\right)B_K}{\bar{W}_K + \left(1+1/m\right)B_K}\right)^2$$

2.3 Regresión Secuencial e Imputación Múltiple (RSIM)

El método RSIM resulta de aplicar las dos técnicas anteriores de manera conjunta, la combinación se basa en considerar que la Imputación Simple mencionada en la sección 2.2, es generada mediante la ejecución de la Regresión Secuencial; en otras palabras, RSIM consiste en repetir *m* veces (*m*>1) los **C** ciclos (**C**>1) de la Regresión Secuencial para así obtener m imputaciones de los registros con NO RESPUESTA.

RSIM, fue desarrollado por T. E. Raghunathan, J. M. Lepkowski, Peter Solenberger y John Van Hoewyk de la Universidad de Michigan en el año 2001 y presenta un par de características interesantes a destacar:

- Permite cierta flexibilidad en el sentido que todos los tipos de modelos pueden ser ajustados
- No requiere suponer un comportamiento específico para la distribución conjunta de las variables consideradas en la investigación.

Además, este método supone entre algunas cuestiones que el mecanismo de generación de la NO RESPUESTA es aleatorio (MAR), lo cual significa que la probabilidad de presencia de NO RESPUESTA en una variable particular *X* depende de otras variables pero no de *X* (Little y Rubin, 1987); cabe mencionar que en la práctica, resulta común que por diversas razones dicho supuesto no se cumpla y que desafortunadamente no hay otras alternativas viables de aplicación que generen resultados adecuados (Allison, 2000).

Con el fin de proveer elementos que fortalezcan la explicación del método RSIM, en el siguiente capítulo se detalla una descripción del sistema de cómputo IVEware, el cual presenta dentro de sus aplicaciones la automatización de dicha técnica.

Capítulo 3

IVEware

En este capítulo, se ilustran los elementos básicos que permiten conocer el funcionamiento del sistema IVEware (acrónimo de las palabras "imputación", "varianza", "estimación" y "software"); en particular, se detalla un procedimiento para poder realizar una imputación según las técnicas RS o RSIM mediante el uso del sistema, aunado a esto, se incluye el desarrollo de un ejemplo didáctico.

IVEware fue desarrollado dentro del programa de Metodología de Encuestas por el Instituto de Investigación Social de la Universidad de Michigan y está disponible sin costo en la dirección http://www.isr.umich.edu/src/smp/ive/. El sistema se desarrolla mediante un conjunto de rutinas escritas en los lenguajes C y FORTRAN que se pueden ejecutar dentro del ambiente del paquete estadístico SAS (versión 6.12 o superiores) o bien como un archivo independiente (versión denominada Srcware); para ambos casos existen las aplicaciones en Linux (o Unix) y Windows. A la par del desarrollo informático, se han generado documentos auxiliares como la guía del usuario, el manual de instalación y diversos ejemplos numéricos.

Partiendo del hecho de que el INEGI no cuenta con el software SAS y que, típicamente, la mayoría del personal del Instituto trabaja en ambiente Windows, la explicación a detallar en el presente capítulo se particulariza para la aplicación Srcware (versión ejecutable beta 0.1 publicada en el año 2005); la Grafica 3.1, muestra la pantalla principal del sistema bajo dicha versión.

Gráfica 3.1: Pantalla principal de Srcware (versión independiente de IVEware)

3.1 Capacidades del sistema

IVEware presenta en su diseño ciertas capacidades a destacar:

- Desarrolla una Imputación Simple o Múltiple en combinación con el método de Regresión Secuencial.

- Realiza análisis descriptivos y análisis basados en un modelo considerando características de un diseño de muestra complejo, como conglomerados, estratificación y factores de expansión.
- Desarrolla el análisis de la Imputación Múltiple empleando estadísticas descriptivas y estadísticas propias para encuestas basadas en un modelo.
- Pretende preservar la estructura de la matriz de covarianza de la información reportada y con esto obtener mejores predicciones de la NO RESPUESTA.

Adicionalmente, el sistema puede imputar ciertas subpoblaciones colocando para esto restricciones sobre las variables a imputar; además, permite asignar cotas para definir un rango de imputación e incorporar las posibles interacciones del modelo de regresión a generar.

3.2 Estructura del sistema

El sistema incluye cuatro módulos denotados con una palabra que representa la función que se realiza dentro de cada uno de ellos.

- Módulo IMPUTE. Imputa la NO RESPUESTA mediante una aproximación basada en la Regresión Secuencial, además puede crear múltiples conjuntos de datos imputados.
- Módulo DESCRIBE. Estima: la media poblacional, proporciones, diferencias entre subgrupos, contrastes y combinaciones lineales de medias y proporciones. Dado un diseño de muestra complejo, la estimación de la varianza la realiza mediante la aproximación de series de Taylor. Si existe presencia de la NO RESPUESTA puede desarrollar un análisis de la Imputación Múltiple.
- Modulo REGRESS. Ajusta modelos lineales, logísticos, politómicos, Poisson, Tobit y proporcionales para datos que provengan de un diseño de muestra complejo. Para estimar la varianza muestral, se usa una aproximación de replicas repetidas. Cuando la NO RESPUESTA existe, puede efectuar un análisis de la Imputación Múltiple.
- Modulo SASMODE. Si los datos son analizados con ciertos procedimientos del paquete estadístico SAS, el usuario puede combinarlos con las características del diseño de muestra. Bajo presencia de NO RESPUESTA puede desarrollar un análisis de la Imputación Múltiple. A diferencia de los módulos anteriores, SASMODE requiere tener instalado SAS para su aplicación.

La ejecución de los tres primeros módulos se realiza mediante el uso de comandos representados por las palabras IMPUTE, DESCRIBE y REGRESS en combinación con instrucciones de programación las cuales consisten en una serie de términos reservados necesarios para completar la ejecución del comando (una explicación del comando IMPUTE se presenta en la sección 3.4).

Por otra parte, el sistema puede emplear comandos adicionales, en particular hay dos que se deben destacar ya que complementan el funcionamiento de los ya mencionados.

- Comando GETDATA. Importa datos de otros sistemas informáticos al ambiente de IVEware; es el primer comando a ejecutar.
- Comando PUTDATA. Permite asignar en archivos individuales a los *m* conjuntos completos obtenidos después de una Imputación Múltiple, el formato de los archivos de salidas es tipo texto (extensión .TXT).

Todos los comandos requieren como insumo dos archivos, uno de ellos, contiene la información a imputar o analizar (generalmente esta en formato texto) y el otro incluye la sintaxis de ejecución (comandos e instrucciones de programación) que detalla la imputación o análisis por realizar (extensión .SET).
En general, la sintaxis de ejecución tiene el siguiente formato:

% COMANDO (name= , dir = , setup=)
Instrucciones de programación

Name identifica al nombre de los archivos de salida asociados a la ejecución del comando y *dir* la ubicación de los mismos; por su parte, *setup* puede adoptar dos opciones excluyentes: "new" que significa que las instrucciones de programación están incluidas dentro de la propia sintaxis y "old" que implica que están definidas en un archivo externo.
Los archivos de salida presentan un nombre único pero con diferentes extensiones y formatos; normalmente, para un comando en particular se generan cinco archivos de salida:

- .TXT: incluye una bitácora de la ejecución del comando (tiempos de procesamiento).
- .LST: presenta un reporte que incluye estadísticas descriptivas para el archivo imputado o analizado.
- .MET: detalla los metadatos del archivo imputado o analizado.
- .SET: consiste en una copia de la sintaxis ya ejecutada.
- .TXT: incluye el archivo imputado o una copia del archivo analizado.

Todos los archivos son importantes, pero aquellos con extensiones "LST" y "TXT" (último de la lista) se les debe dar especial atención debido a que presentan la información de mayor interés; el primero, reportando diversos estadísticos (media, desviación estándar, porcentaje del código de respuesta entre otros) de cada variable imputada o analizada y el segundo presentando la información ya imputada o bien analizada.
Una aclaración importante con relación a la sintaxis de ejecución, consiste en que puede estar conformada por uno o varios comandos; en esta situación, se hablará de una sintaxis "general" la cual producirá tantos "grupos" de archivos de salida como comandos la integren. La sintaxis "general" tiene el siguiente formato:

% COMANDO 1 (name= , dir = , setup=)
Instrucciones de programación

% COMANDO 2 (name= , dir = , setup=)
Instrucciones de programación
.
.
.

% COMANDO n (name= , dir = , setup=)
Instrucciones de programación

3.3 Funcionamiento del sistema

El funcionamiento de IVEware como tal es muy sencillo, sólo se requiere crear (o abrir, en caso de ya existir) el archivo de sintaxis de ejecución y presionar el botón que contiene al símbolo **">"**; como señalamiento de que está en ejecución, el sistema presenta una serie de ventanas que interactúan (abren y cierran muy rápidamente), cuando dejan de hacerlo, el procesamiento ha terminado y la sintaxis solicitada, ya ejecutada, se muestra en pantalla.

Para determinar si la ejecución fue realizada satisfactoriamente, se debe confirmar que la(s) bitácora(s) emitida(s), excluya(n) mensajes con códigos de error.

Las causas principales que comúnmente conducen a un error de ejecución son:

- Problemas con la lectura del archivo a imputar o analizar debido a que el formato no es compatible con IVEware. Para modificar el formato del archivo a imputar o analizar, se recomienda emplear los paquetes comerciales Excel o preferentemente StatTransfer.
- Una declaración incorrecta de la longitud del archivo a imputar o analizar, es decir, la descripción del archivo dentro del comando GETDATA presenta inconsistencias (faltan o sobran variables).
- Una inconsistencia en la escritura de los nombres de las variables dentro o entre el(los) comando(s) que integra(n) la sintaxis.
- Existencia de espacios en blanco que no son válidos (por ejemplo, entre el nombre del comando y su paréntesis).
- Una ubicación incorrecta (o falta) del carácter **";"**, el cual debe escribirse al terminar la redacción de cada instrucción de programación.

Con la idea de facilitar la detección de errores y en el caso de una imputación a la vez estimar el periodo de procesamiento de cómputo, es recomendable realizar pruebas preliminares, éstas, inician con la selección de una muestra de datos y de algunas variables a imputar o analizar, posteriormente, se va incrementando la cantidad de variables y el volumen de registros hasta cubrir el total de variables y a todos los datos. Bajo una imputación, las pruebas preliminares permiten identificar aquella(s) variable(s) que ocupa(n) el mayor tiempo de procesamiento la(s) cual(es) suele(n) ser de tipo categórica(s); al respecto, la forma de disminuir el efecto consiste en reducir el número de categorías aplicando una agrupación al rango de respuesta de la(s) variable(s) antes de imputar.

3.4 Imputando con el sistema

En general, el proceso de imputación con IVEware consiste en desarrollar secuencialmente cinco etapas:

1. Definir la estructura del archivo a imputar.
2. Importar el archivo a imputar.
3. Diseñar y procesar el archivo de sintaxis de ejecución.
4. Extraer los *m* archivos imputados.
5. Calcular la estimación y la varianza de la Imputación Múltiple.

Las tres primeras etapas son requeridas al aplicar RS o RSIM, la cuarta y quinta sólo aplican con RSIM; por otra parte, aunque la primera y la quinta etapa se realizan fuera del sistema, ambas son necesarias para el resto; la primera funciona como insumo para el resto y la última como cierre del proceso de imputación vía RSIM.

Existe una etapa adicional no incluida en el listado anterior definida como: Variables "auxiliares" (recodificadas); su realización, depende de que la imputación

se aplique en variables “auxiliares” y no directamente en las variables “originales” a imputar; ésta, se efectúa a la par que la etapa 1, es externa a IVEware y consiste básicamente en definir el número de variables “auxiliares” a emplear y ejecutar un proceso de recodificación.

A continuación se detalla en qué consiste cada una de las etapas.

- **Etapa 1. Definición de la estructura del archivo a imputar**

  Usualmente el archivo de datos por analizar está “bien” definido desde el inicio de la investigación, en el sentido de que se conocen las variables que lo integran y de ellas, cuales serán variables indicadoras y cuales variables en estudio (incluye a la(s) variable(s) de interés y a la(s) variable(s) predictora(s)); tradicionalmente, ambos tipos de variables conforman la estructura básica del archivo a imputar.

  Por otra parte, la presencia de falta de normalidad de las variables en estudio, implica la transformación de los datos (usualmente a escalas de tipo logarítmicas) como corrección a esta situación; luego, se deben incorporar variables “adicionales” para almacenar el resultado de la transformación y registrar la imputación una vez aplicada, pero convertida en la escala original de la variable; así, por cada variable que no presente una distribución normal, se deben anexar dos variables más al archivo a imputar.

Una vez definida la estructura del archivo a imputar, se tiene que cada variable que lo integra puede determinar tres bloques de información:

- La información reportada u observada. Es la fuente de datos que permite construir el modelo estadístico a emplear en la imputación.
- La información a imputar. Típicamente representa a la NO RESPUESTA.
- La información a ignorar. Son los datos a excluir: blancos por pase, NO RESPUESTA por unidad, entre otros.

Con el fin de dimensionar el problema, justificar la aplicación de la imputación y tener elementos para evaluar el funcionamiento del sistema, resulta relevante contabilizar el número de casos incluidos en cada bloque de información y en particular detectar la(s) variable(s) a imputar.

- **Etapa opcional. Variables “auxiliares” (recodificadas)**

  Cuando se opta por escribir el resultado de la imputación directamente sobre la(s) variable(s) a imputar, la sintaxis a ejecutar suele ser sencilla y corta, además la diferencia entre el tamaño del archivo a imputar y el archivo ya imputado es mínima; por contraparte, además de perder el valor original del registro imputado, éste, no puede identificarse de manera directa en la base de datos imputada; con esto, no resulta fácil evaluar el funcionamiento del sistema.

  Para preservar los valores originales de las variables, una alternativa, radica en incorporar una estructura adicional al archivo a imputar que consiste en agregar dos variables “auxiliares” por cada una de las variables a imputar; acto seguido, se debe desarrollar y ejecutar un proceso de recodificación, que se encarga de insertar códigos especiales a las variables “auxiliares”. Después de la recodificación, una de las dos “nuevas” variables contiene a la información reportada y códigos particulares (que representan la información a imputar y a ignorar) de la variable a imputar, esta variable “auxiliar” es la que realmente imputa el sistema; la otra, incluye códigos especiales que representan a los tres bloques de información de la variable a imputar y es empleada para simplificar

las restricciones que tenga dicha variable (esta segunda variable no sufre cambio alguno después de la imputación).
Para ilustrar didácticamente el desarrollo de esta etapa, considérese una variable X que presenta el siguiente comportamiento:

- Sólo se capta para personas con 18 años cumplidos (está en blanco para menores de 18 años).
- Su respuesta válida es: 1, 2, 3, 4, 5, 6 o 9.
- El código 9 representa la NO RESPUESTA.

Así, el proceso de recodificación, debe añadir las dos variables "auxiliares" al archivo a imputar e incluir una programación que permita construir los casos que se muestran en la Tabla 3.1, ambos aspectos suelen desarrollarse empleando un manejador de base de datos.

Variable X (valor)	**Variable "auxiliar" 1 (valor)**	**Variable "auxiliar" 2 (valor)**
1,2,3,4,5,6 (información reportada)	1,2,3,4,5,6	1
9 (información a imputar)	**.**	0
blanco (información a ignorar)	**.I**	9

Tabla 3.1: Bloques de información para la variable X

Es importante notar que después de la imputación, la variable "auxiliar" 1 presentará dos cambios:

- En los registros donde está el carácter "**.**", habrá un valor imputado.
- En donde hay un código ".I" habrá un valor numérico constante, este será un "0" (cero) si la variable es continua o el valor de la máxima categoría incrementado en uno si es de tipo categórica.

Claramente, esta opción incrementa de manera sustancial el tamaño tanto del archivo a imputar como el del archivo ya imputado, pero permite agregar elementos que mejoran y evalúan el propio proceso de imputación; así, puede obtenerse un reporte estadístico alterno al generado por IVEware (previo desarrollo y ejecución de un programa de cómputo en ambiente externo al sistema) que corrige los cálculos generados erróneamente por el sistema al considerar la información ignorada como imputada; también pueden producirse con cierta facilidad gráficos para el análisis y evaluación de la imputación, además es posible reportar estadísticas adicionales a las consideradas originalmente por el sistema. Estos aspectos, son tratados con mayor detalle en el capítulo 4 en el que se describe la aplicación de esta investigación y que incluye la realización formal de esta etapa adicional.

- **Etapa 2. Importación del archivo a imputar**

  Convertir el archivo a imputar a un formato adecuado para IVEware, equivale a realizar una importación mediante la ejecución del comando GETDATA; para la elaboración de la sintaxis correspondiente, se deben definir aspectos

relacionados con la ubicación y estructura del archivo a imputar entre los cuales se encuentran:

- Nombre y ubicación (dirección o ruta) de los archivos de salida del comando GETDATA.
- Nombre, ubicación y formato del archivo a imputar (si es tipo texto aclarar qué tipo de carácter lo delimita).
- Nombre y tipo de todas las variables que integran el archivo a imputar.

La ubicación de los aspectos anteriores dentro de la sintaxis se ilustra en el ejemplo que se presenta más adelante, además, se explica la relación con las instrucciones de programación correspondientes (para una consulta de la sintaxis completa de este comando ver el Anexo E).

Elaborado el archivo de sintaxis de la importación, resta ejecutarlo (ver sección 3.3) para obtener así el archivo a imputar bajo formato de IVEware y poder continuar con la siguiente etapa.

- **Etapa 3. Diseño y ejecución del archivo de sintaxis (imputación)**

Redactar el archivo de sintaxis para ejecutar la imputación no es otra cosa que construir la sintaxis del comando IMPUTE; para desarrollarla, se deben conocer algunos aspectos:

- Nombre y ubicación (dirección o ruta) de los archivos de salida del comando IMPUTE.
- Nombre y ubicación del archivo a imputar, importado previamente por el comando GETDATA.
- Nombre y tipo de la(s) variable(s) que será(n) imputada(s).
- Restricciones. Indicar para cada variable a imputar las condiciones que define los tres bloques de información.
- Cotas numéricas. Definir para cada variable a imputar el soporte o dominio de los valores imputados.
- Número de ciclos. Determinar el número de ciclos que la Regresión Secuencial debe aplicar.
- Número de múltiplos. Declarar el número *m* de imputaciones a generar, se define una Imputación Simple cuando *m* es uno, en otro caso se denomina Imputación Múltiple
- Semilla de arranque aleatorio. Si la semilla se escribe dentro de paréntesis y se aplica una Imputación Múltiple, se genera la misma imputación las *m* veces.

En el ejemplo que se presenta más adelante se revisan detalles relacionados con la sintaxis y las instrucciones de programación respectivas (para una consulta de la sintaxis completa de este comando ver el Anexo E).

Una vez elaborada la sintaxis de la imputación, se procede a su ejecución (ver sección 3.3).

Liberada la imputación a nivel de sintaxis, ahora debe cuestionarse cómo se aprobarían los resultados numéricamente; al respecto, existen algunos criterios a nivel de recomendación que auxilian en esta labor. Para cada variable en estudio e Imputación Simple aplicada, debe revisarse que:

- El total de los bloques de información obtenidos al definir la estructura del archivo a imputar, debe coincidir con lo reportado por el archivo de salida (extensión .LST) del comando IMPUTE; cabe recordar, que el uso de

restricciones implica que IVEware considere como información imputada a la ignorada, esta situación debe tomarse en cuenta para la correcta comparación.

- Exista diferencia poco significativa entre la información reportada y la imputada.
- La presencia de valores atípicos en la información imputada no se debe dar a menos de que ya existan en la información reportada.

En caso de una Imputación Múltiple, además de revisar lo anterior se debe checar que a nivel registro, los valores imputados no cambien sustancialmente entre las distintas Imputaciones Simples (propiedad de estabilidad de la IM).

- **Etapa 4. Extracción de los "*m*" archivos imputados**

  Bajo la realización de una Imputación Múltiple, se requiere llevar a cabo la extracción de los *m* archivos imputados correspondientes; esta acción, consiste en aplicar el comando PUTDATA. La sintaxis de este comando, debe incluir ciertos elementos:

  - Nombre y ubicación (dirección o ruta) de los archivos de salida del comando PUTDATA.
  - Nombre y ubicación del archivo de salida del comando IMPUTE.
  - Incluir (o excluir) títulos a los campos del archivo de salida, es decir, definir si se desea (o no) agregar los nombres de las variables como cabeceras.

  Es importante aclarar que la sintaxis de esta etapa, debe incluir *m* aplicaciones del comando PUTDATA, diferenciándose en la redacción del nombre del archivo de salida y en el número asignado a la instrucción MULT (para la ejecución de la sintaxis, ver sección 3.3).

  El ejemplo que se ilustra enseguida, detalla el uso de las instrucciones de programación (la sintaxis completa de este comando se puede consultar en el Anexo E).

- **Etapa 5. Cálculo de la estimación y de la Varianza (en caso de aplicar RSIM)**

  Definido previamente un parámetro de interés y nuevamente bajo el supuesto de la aplicación de una Imputación Múltiple, la quinta etapa del proceso se refiere a calcular la estimación y la varianza de la Imputación Múltiple efectuada. La actividad es sencilla, sólo consiste en aplicar las expresiones matemáticas dadas en el capítulo 2 empleando los *m* archivos imputados, usualmente se emplea Excel o una calculadora científica para computar.

Con la idea de fortalecer lo anterior, a continuación se desarrolla un ejercicio para un conjunto pequeño de datos, el énfasis del ejemplo radica en el manejo de las sintaxis para las etapas de importación, imputación y extracción; además, se aprovecha para presentar los diversos reportes de salida y lo referente al cálculo de las estimaciones correspondientes a una Imputación Múltiple.

- **Ejemplo**
  El archivo a imputar *VEINTE.TXT* presenta la siguiente estructura:

STUDYID	*X1*	*X2*	*Y*	*STUDYID*	*X1*	*X2*	*Y*
1	*0*	*28.00781*		**11**	*0*	*22.53533*	
2	*1*	*19.16682*	*217.74124*	**12**	*0*	*29.23859*	
3	*1*	*36.87843*	*235.47414*	**13**	*1*	*44.47300*	*220.86140*
4	*0*	*24.05821*	*149.53657*	**14**	*0*		
5	*1*	*17.58699*		**15**	*0*	*21.33619*	
6	*1*		*175.79420*	**16**	*1*	*40.90243*	*203.87273*
7	*1*			**17**	*0*	*40.23866*	
8	*0*			**18**	*1*	*36.54201*	*207.53087*
9	*0*	*28.41465*		**19**	*0*		
10	*1*	*46.14381*	*195.27791*	**20**	*1*	*42.77463*	

Tabla 3.2: Estructura del archivo VEINTE.txt

De la Tabla 3.2, se observa que el ejemplo maneja una población de 20 registros y 4 variables, donde:

- STUDYID es el campo de identificación por registro y no requiere de análisis numérico alguno.
- X1 es categórica con códigos de respuesta 0 o 1 y NO presenta NO RESPUESTA.
- X2 es numérica y continua, con 15 registros reportados y 5 a imputar.
- Y es numérica y continua, con 8 registros reportados y 12 a imputar.

Así, hay 3 variables en estudio (X1, X2 e Y) y tanto X2 como Y presentan registros con NO RESPUESTA.

Por otra parte, supóngase que:

- El archivo a imputar consiste en una copia de VEINTE pero excluyendo los nombres de las variables (primer renglón).
- No existe la información a ignorar (blancos por pase o NO RESPUESTA por unidad) o equivalentemente restricción alguna.
- Ninguna variable presenta cotas en su rango de imputación.
- Se asignan los valores imputados directamente en las variables X2 e Y.
- Se requiere aplicar una Imputación Múltiple, 5 ciclos en la Regresión Secuencial y 2 múltiplos para la Imputación Múltiple.
- Para la variable Y, se debe obtener la estimación de la media y de su varianza mediante la técnica RSIM.

El proceso de Imputación Múltiple solicitado consta de las siguientes tres fases:

- Diseño y Ejecución de la sintaxis
  Las instrucciones de programación necesarias para la ejecución de las tres etapas (importación, Imputación Múltiple y extracción) se integran en una sintaxis "general" la cual se presenta en la Tabla 3.3, la cual, incluye dos columnas: en la primera, se indica la instrucción de programación que integra cada uno de los comandos a ejecutar, en la segunda se incluye una explicación o comentario acerca del funcionamiento de la propia instrucción.

/* importation */	Comentario Explicativo
%getdata(name=veinte, setup=new);	Inicia sintaxis de importación
datain veinte.txt;	Archivo a imputar (su ubicación debe coincidir con la del archivo de sintaxis)
metadata;	Define estructura del archivo a imputar
delim "\t";	Tipo texto y delimitado por tabulador
Variables	
name=STUDYID type=char;	Definición de las variables y de su tipo respectivo
name=X1 type=num;	
name=X2 type=num;	
name=Y type=num;	
end;	Cierra la ejecución METADATA
run;	Termina sintaxis de importación
/* multiple imputation */	Comentario Explicativo
%impute(name=impute, setup=new);	Inicia sintaxis de Imputación Múltiple
title Multiple imputation;	Titulo en el archivo de salida LST
datain veinte;	Archivo a imputar (generado por GETDATA)
dataout impute;	Archivos de salida
default continuous;	Las 4 variables son candidatas a imputarse
transfer STUDYID;	No será imputada, pero si incluida en el archivo de salida
categorical X1;	Variable a imputar pero tipo categórica
minrsqd .01;	
iterations 5;	5 iteraciones (o ciclos) en la Regresión Secuencial
multiples 2;	2 imputaciones
seed 2001;	Semilla de arranque aleatorio
run;	Termina sintaxis de imputación
/* extracting the remaining two multiply imputed datasets */	Comentario Explicativo
%putdata(name=impute_mult1, setup=new);	Inicia sintaxis de extracción
imputation impute;	Archivo insumo (generado por IMPUTE)
dataout impute_mult1;	Archivo de salida (sin cabeceras)
mult 1;	Guarda la primera imputación de IM
run;	Termina sintaxis de extracción
%putdata(name=impute_mult2, setup=new);	Inicia sintaxis de extracción
imputation impute;	Archivo insumo (generado por IMPUTE)
table impute_mult2;	Archivo de salida (con cabeceras)
mult 2;	Guarda la segunda imputación de IM
run;	Termina sintaxis de extracción

Tabla 3.3: Sintaxis “general” de ejecución para el archivo VEINTE.txt

- Liberación de la imputación

  Posterior a la ejecución de la sintaxis, se verifica que las bitácoras de ejecución (archivos: IMPVEINTE.txt, VEINTE.txt, IMPUTE.txt, IMPUTE_MULT1.txt e IMPUTE_MULT2.txt) estén libres de error.

  En particular, el archivo IMPVEINTE.txt, reporta el desarrollo del proceso completo y tiene la siguiente presentación:

```
SRCware SRC SMP Statistical Software
Survey Research Center, Institute for Social Research
University of Michigan
Version 1.0, Copyright (c) 2005
Mon Jan 12 19:16:48 2009
Begin SRCware execution
Mon Jan 12 19:16:48 2009
Begin veinte
Mon Jan 12 19:16:48 2009
Begin getdata execution
Normal termination
Mon Jan 12 19:16:48 2009
Begin impute
Mon Jan 12 19:16:48 2009
Begin iveset execution
Normal termination
Mon Jan 12 19:16:48 2009
Begin impute execution
Normal termination
Mon Jan 12 19:16:48 2009
Begin putdata execution
Normal termination
Mon Jan 12 19:16:48 2009
Begin impute_mult1
Mon Jan 12 19:16:48 2009
Begin putdata execution
Normal termination
Mon Jan 12 19:16:48 2009
Begin impute_mult2
Mon Jan 12 19:16:48 2009
Begin putdata execution
Normal termination
Mon Jan 12 19:16:48 2009
End SRCware execution
```

  El reporte anterior no presenta errores, por lo que la imputación se libera en su sintaxis.

  Para evaluar la consistencia de los resultados de la imputación numéricamente, se deben revisar los archivos de salida: IMPUTE.lst, IMPUTE_MULT1.txt e IMPUTE_MULT2.txt los cuales incluyen elementos que auxilian para lograr esta actividad.

En particular, el archivo IMPUTE.lst presenta el reporte estadístico de las dos imputaciones solicitadas bajo el siguiente formato:

IVEware Setup Checker, Mon Jan 12 19:16:48 2009
1
Setup listing:
Title multiple imputations;
Datain veinte;
Dataout impute;
Default continuous;
Transfer STUDYID;
Categorical x1;
Minrsqd .01;
Iterations 5;
Multiples 2;
Seed 2001;
Run;
IVEware Iterative Imputation Procedure, Mon Jan 12 19:16:48 2009
1
Multiple Imputations
Imputation 1

Variable	Observed	Imputed	double counted
X1	20	0	0
X2	15	5	0
Y	8	12	0

Variable X2

	Observed	Imputed	Combined
Number	15	5	20
Minimum	17.587	25.0733	17.587
Maximum	46.1438	48.3762	48.3762
Mean	31.8865	34.5217	32.5453
Std Dev	9.77713	9.43581	9.51583

Variable Y

	Observed	Imputed	Combined
Number	8	12	20
Minimum	149.537	163.058	149.537
Maximum	235.474	217.377	235.474
Mean	200.761	186.12	191.976
Std Dev	27.3345	18.7899	23.1049

IVEware Iterative Imputation Procedure, Mon Jan 12 19:16:48 2009
2
Multiple imputation
Imputation 2

Variable	Observed	Imputed	Double counted
X1	20	0	0
X2	15	5	0
Y	8	12	0

Variable X2

	Observed	Imputed	Combined
Number	15	5	20
Minimum	17.587	11.1609	11.1609

```
Maximum        46.1438        32.4675        46.1438
Mean           31.8865        23.7609        29.8551
Std Dev         9.77713        8.24513        9.88836
Variable Y
                Observed       Imputed        Combined
Number          8              12             20
Minimum         149.537        148.532        148.532
Maximum         235.474        212.928        235.474
Mean            200.761        193.621        196.477
Std Dev          27.3345        18.4286        22.0175
Srcware putdata procedure, Mon Jan 12 19:16:48 2009
1
Multiple imputation
Dataset: impute
Delimiters: "\t"
Variables: 6
Multiple variable: _MULT_
ID variable: _OBS_
Observations: 20
```

Nótese que los totales de la información reportada e imputada indicados por IMPUTE.lst son los mismos que los obtenidos a partir del archivo a imputar, además, se observa que la imputación genera pequeñas diferencias entre la información reportada (observed) y la combinada (combined); considerando ambos aspectos es posible liberar en términos numéricos a la Imputación Múltiple (un tratamiento completo sobre como validar una imputación se discute en el siguiente capítulo).

Por su parte, IMPUTE_MULT1.txt e IMPUTE_MULT2.txt representan a los dos archivos completos producto de las dos iteraciones de la Imputación Múltiple.

El archivo IMPUTE_MULT1.txt consiste en:

```
1     0     28.00781      167.5116946411     1
2     1     19.16682      217.74124     1     2
3     1     36.87843      235.47414     1     3
4     0     24.05821      149.53657     1     4
5     1     17.58699      217.3772482711     5
6     1     48.376179327  175.7942      1     6
7     1     38.215672254  175.0239971441     7
8     0     34.244277693817 0.2923120761     8
9     0     28.41465      183.6558727831     9
10    1     46.14381      195.27791     1     10
11    0     22.53533      175.9026932241     11
12    0     29.23859      203.5232265171     12
13    1     44.473        220.8614      1     13
14    0     25.0732627609171.7904294061     14
15    0     21.33619      203.2429534441     15
16    1     40.90243      203.87273     1     16
17    0     40.23866      213.2449474281     17
18    1     36.54201      207.53087     1     18
19    0     26.6989248065163.0579090151     19
20    1     42.77463      188.8144127861     20
```

Y el archivo IMPUTE_MULT2.txt es:

STUDYID	X1	X2	Y	_MULT_	_OBS_
1	0	28.00781	199.687210283	2	1
2	1	19.16682	217.74124	2	2
3	1	36.87843	235.47414	2	3
4	0	24.05821	149.53657	2	4
5	1	17.58699	192.764150842	2	5
6	1	27.7816067391	175.7942	2	6
7	1	32.4674834929	212.928130601	2	7
8	0	20.4424066747	148.53182336	2	8
9	0	28.41465	180.86085324	2	9
10	1	46.14381	195.27791	2	10
11	0	22.53533	208.299465811	2	11
12	0	29.23859	187.262215874	2	12
13	1	44.473	220.8614	2	13
14	0	11.1608675543	194.96647411	2	14
15	0	21.33619	205.025224597	2	15
16	1	40.90243	203.87273	2	16
17	0	40.23866	210.342643714	2	17
18	1	36.54201	207.53087	2	18
19	0	26.9522478762	206.427814633	2	19
20	1	42.77463	176.352898526	2	20

Nótese que, IMPUTE_MULT2.txt incluye en el primer renglón los nombres de las variables e IMPUTE_MULT1 no; en ambos, la información imputada está subrayada y se imprime a 8 o más dígitos.

- Estimaciones de la Imputación Múltiple

  La estimación solicitada de la media de la variable Y (parámetro θ) y de su varianza emplea la información del reporte IMPUTE.lst el cual indica que:
  - ✓ Para la primera imputación: la media de Y es 191.976, la desviación estándar es 23.1049 y la varianza es 533.8364.
  - ✓ Para la segunda imputación: la media de Y es 196.477, la desviación estándar es 22.0175 y la varianza es 484.7703.

  Retomando las expresiones dadas en el capítulo 2 tenemos que las estimaciones de RSIM son:

$$\hat{\theta}_{RSIM} = \frac{1}{m}\sum_{k=1}^{m}\hat{\theta}_k = \frac{191.976 + 196.477}{2} =$$

$$194.2265$$

$$\hat{\text{var}}(\theta)_{RSIM} = \frac{1}{m}\sum_{k=1}^{m}\hat{W}_k + \left(1+\frac{1}{m}\right)\left(\frac{1}{m-1}\right)\sum_{k=1}^{m}\left(\hat{\theta}_k - \hat{\theta}_{RSMI}\right)^2 =$$

$$\frac{533.8364+484.7703}{2}+1.5*10.1295=$$

$$509.3034+15.1942=$$

$$524.4976$$

En el siguiente capítulo, se discuten con mayor profundidad los aspectos de la técnica RSIM bajo una nueva aplicación del sistema IVEware, pero ahora con información real del Estado de Aguascalientes correspondiente al XII CGPyV, 2000.

Capítulo 4

Diseño de la investigación: datos del INEGI y su imputación múltiple

El objetivo primordial de esta investigación consiste en aplicar la técnica RSIM al ingreso por trabajo (mensualizado) de la Población Ocupada del Estado de Aguascalientes, apoyándose para tal efecto en el uso del sistema IVEware; en relación a esto, es prudente realizar algunos comentarios previos a la aplicación formal.

1.- El uso de variables predictoras en la conformación del modelo de imputación, permite que éstas también puedan ser imputadas; con esto, se da una mayor utilidad a los resultados de la investigación, ya que la información en estudio proviene de un evento censal.

2.- La información en estudio, presenta una característica ventajosa para su procesamiento la cual consiste en que la información a imputar está identificada con códigos constantes ("9", "99",…, "999999").

3.- Se desconoce el valor reportado de la información a imputar, lo cual resulta en una desventaja a la investigación ya que limita el desarrollo de la evaluación y el análisis a realizar de la imputación.

4.- Aunque la no respuesta por unidad no se imputa, ésta debe ser detectada para ser excluida del análisis mediante el uso de restricciones sobre la información a imputar

4.1 Estructura del archivo a imputar

La base de datos que soporta la investigación representa a la Población del Estado de Aguascalientes captada por el XII CGPyV, 2000 y está conformada por 944,285 registros y 70 variables; 48 de ellas, están representadas por la tabla de Población (Anexo C) y de las 22 restantes, 18 se refieren a la identificación geográfica, 2 identifican al estrato sociodemográfico asignado por INEGI y 2 más indican la imputación aplicada por el INEGI.

El proceso de construcción del archivo a imputar, toma como fuente de información la base de datos anterior y su estructura definitiva, requiere de realizar las fases que se describen a continuación.

- **Fase 1. Definición de la variable de interés**
  Retomando el objetivo fundamental de esta investigación, se define que la variable de interés es:
  - El ingreso por trabajo (mensualizado) de la Población Ocupada (TOT_ING).
- **Fase 2. Definición de las variables indicadoras**
  Estas variables identifican geográficamente a cada uno de los registros o bien indican si estos fueron imputados o no por el estudio del INEGI.
  Para la identificación geográfica se emplean 14 variables:
  - Identificación de la persona (ID_PERSO).
  - Identificación del hogar (ID_HOGAR).
  - Identificación de la vivienda (ID_VIV).
  - Entidad federativa (ENT).
  - Municipio (MUN).

- Localidad (LOC).
- Área geoestadística básica (AGEB).
- Manzana (MZA).
- Segmento (SEG).
- Número de persona del hogar (NUM_PER).
- Tipo de cuestionario (TIPO_CUEST).
- Llave de identificación única (LLAVE_UN).
- Numero consecutivo de vivienda (NUM_VIV).
- Apellido (APELLIDO).

Para la imputación realizada por el INEGI se tienen 2 variables:

- Imputación del INEGI (IMPUTE).
- Criterio por el cual se detecta al registro como imputado por INEGI (SOBRA).

- **Fase 3. Selección de las variables predictoras**

  El grupo de variables predictoras se conforma por dos subgrupos; el primero de ellos, está integrado por las seis variables consideradas en la imputación del INEGI, con esto, se aprovecha lo invertido en esta fase por esa investigación y a la par se aportan elementos para poder comparar posteriormente ambas imputaciones; mientras que, el segundo subgrupo se conforma por seis variables seleccionadas por recomendación del asesor del proyecto; luego, las doce variables predictoras son:

  Grupo I.

  - Sexo (SEXO).
  - Edad (EDAD).
  - Nivel de escolaridad (NIV_ESC).
  - Ocupación (OCUPAC).
  - Posición en el trabajo (POS_TRA).
  - Parentesco (PARENTES).

  Grupo II.

  - Estrato sociodemográfico (ESTRATO+SUBESTRA).
  - Zona (ZONA).
  - Estado civil (EDO_CONY).
  - Número de hijos nacidos vivos (VIVOS).
  - Discapacidad del tipo motriz (MOTRIZ).

Dentro del paréntesis se presenta el mnemónico correspondiente empleado en la base de datos.

Hasta esta fase el archivo a imputar está compuesto de 944,285 registros y 29 variables extraídas de la base de datos que soporta la investigación, de las cuales 13 son variables en estudio (una variable de interés y doce variables predictoras).

- **Fase 4. Generación del logaritmo del ingreso**

  El ingreso por trabajo de la Población Ocupada no cumple con el supuesto distribucional de normalidad, por lo que es necesario aplicar una transformación logarítmica para lograrlo; para esto, se requiere crear un par de variables adicionales; la primera contiene el logaritmo del ingreso y se convierte consecuentemente en una variable a imputar, la segunda contendrá el resultado de aplicar la transformación inversa (función exponencial) a los valores

imputados del ingreso; es claro, que esta segunda variable se obtiene hasta después de realizar la imputación.
La Población Ocupada sin ingresos (TOT_ING=0) genera una dificultad de cálculo en la transformación logarítmica (log de 0 no está definido), este inconveniente se resuelve al replantear la transformación incrementando el ingreso en una unidad antes de calcular el logaritmo.
Con base en lo anterior, se anexan dos variables a la estructura del archivo a imputar las cuales para su generación, requieren de cálculo adicional; nótese que ambas no están incluidas en la base de datos de la investigación, las variables son:

- Logaritmo del ingreso (LOG_ING).
- Ingreso imputado (TOT_ING_IM).

- **Fase 5. Detección de las restricciones y cotas**

  El cuestionario (Anexo B) está integrado por bloques temáticos y la presencia de pases de preguntas es frecuente; en ambos casos, para lograr el manejo correcto de la información implica que se identifiquen con claridad los filtros necesarios para las variables en estudio relacionadas, creándose así una serie de restricciones inherentes a dichas variables.
  Otras situaciones que también generan restricciones son:
  - La NO RESPUESTA por unidad.
  - La exclusión de registros para mantener la consistencia de la información (viviendas que no presentaron información de sus ocupantes, parentesco: sólo un jefe por hogar; PARENTES<>”100”).

  Una vez detectada la presencia de restricciones (o información a ignorar), como consecuencia el emplear variables “auxiliares” (recodificadas) se vuelve obligado; la causa de esta decisión obedece fundamentalmente a que contar con este tipo de variables permite:
  - Producir un reporte estadístico alterno (recuérdese que bajo la presencia de restricciones, IVEware erróneamente considera como información imputada a la información ignorada).
  - En el caso particular del ingreso, se logra conservar el valor imputado por el INEGI lo que permite compararlo contra el valor que genere la técnica RSIM.
  - Simplificar la redacción de las propias restricciones.

  Con el fin de evitar duplicidad en la escritura, las restricciones presentes en las variables predictoras y en el ingreso se muestran en la siguiente fase.

  Por otro lado, es posible que las variables en estudio estén acotadas en su respuesta, ya sea por alguna condición particular o bien por la longitud de diseño de la respuesta de la variable; esto, se detecta al revisar la estructura en el archivo a imputar o al consultar los rangos válidos en la tabla de Población (Anexo C).

Las cotas numéricas que se requieren en esta investigación, se muestran en la Tabla 4.1.

VARIABLE	COTA
EDAD	12,…,130
VIVOS	0,…,25
LOG_ING	0,…,13.815508 (0,…,999997 en escala original)

Tabla 4.1: Cotas de las variables predictoras y del ingreso

Es importante mencionar que normalmente esta fase consume una buena parte del tiempo del proyecto, debido a la intensa interrelación entre las variables en estudio.

- **Fase 6. Generación de las variables "auxiliares" (recodificadas)**

Tomando como base la descripción dada en el capítulo anterior (sección 3.4) en la cual se indican los aspectos generales para la creación y uso de las variables "auxiliares", en esta fase se contempla el desarrollo de las siguientes actividades:

- Adición de variables "auxiliares"

  Para las variables en estudio con presencia de NO RESPUESTA, se agregan al archivo a imputar 2 variables "auxiliares" denotadas de la siguiente manera:
  - ✓ Primera variable "auxiliar". El mnemónico de la variable en estudio a imputar más la terminación "_CM", existe la excepción para las variables EDO_CONY y PARENTES donde la terminación cambia por "_C".
  - ✓ Segunda variable "auxiliar". El mnemónico del ingreso o de la variable predictora más la terminación "_F".
- Definición de los bloques de información

  La conformación de los bloques de información se desarrolla bajo tres niveles de análisis:
  - ✓ Individualmente

    Las variables en estudio de acuerdo con sus restricciones, producen una definición inicial de los tres bloques de información, la cual queda conformada bajo los criterios mostrados en la Tabla 4.2.

Variable	Información a Imputar	Información a Ignorar	Información Reportada
TOT_ING	IMPUTE<>b o TOT_ING=999998,999999	EDAD < 12	TOT_ING=0,…,999997
SEXO	Ninguna	Ninguna	SEXO= 1,2
EDAD	EDAD = 999	EDAD < 12	11<EDAD<131
NIV_ESC	NIV_ESC = 9	EDAD < 5	NIV_ESC = 0,.., 8
OCUPAC	OCUPAC = 9999	OCUPAC = b	OCUPAC = 1100,…, 8390
POS_TRA	POS_TRA = 9	POS_TRA = b	POS_TRA = 1,…, 5
PARENTES	PARENTES = 999	PARENTES = b	PARENTES = 100,…, 624
ESTRATO	Ninguna	Ninguna	ESTRATO = 1, 2, 3, 4

SUBESTRA	Ninguna	Ninguna	SUBESTRA = 0, 1, 2, 3
ZONA	Ninguna	Ninguna	ZONA = 01, 35 ,45 ,55 ,60
EDO_CONY	EDO_CONY = 9	EDAD < 12	EDO_CONY = 1,…, 8
VIVOS	VIVOS = 98, 99	VIVOS = b	VIVOS = 0,…, 25
MOTRIZ	MOTRIZ = 9	MOTRIZ = b	MOTRIZ = 1, 2

Tabla 4.2: Bloques de información para las variables predictoras y el ingreso (criterios individuales)

- ✓ Colectivamente
  Los bloques de información definidos en la Tabla 4.2, se deben ajustar al considerar la NO RESPUESTA por unidad (registros donde todas las variables en estudio captadas mediante entrevista presentan NO RESPUESTA conjunta); un total de 4,442 registros cuyo origen se relaciona directamente con aquellas viviendas que no presentaron información de sus residentes se deben transferir de:
  - La información a imputar (columna 1) hacia la información a ignorar (columna 2) siempre que existan registros a imputar.
  - La información reportada (columna 3) hacia la información a ignorar (columna 2) siempre que no existan registros a imputar (SEXO, ESTRATO, SUBESTRATO, ZONA).
- ✓ A nivel hogar
  Efectuada la revisión en forma individual y colectiva, los bloques de información también deben conservar cierta consistencia a nivel hogar; en particular, la variable parentesco (PARENTES) presenta un ajuste en este sentido al cambiar 208,167 registros reportados (columna 3) con código de respuesta "100" (jefe o jefa del hogar) a un status de información a ignorar (columna 2) y con esto respetar el hecho de que en cada hogar exista un jefe el cual debe ser único.

- Recodificación de las variables "auxiliares"

La recodificación es el término empleado para definir el proceso de llenado de las variables "auxiliares"; la asignación de valores o códigos, se rige de acuerdo con los criterios dados en la Tabla 4.3:

Variables en estudio a imputar	**Primera variable "auxiliar" (terminación _CM o _C)**	**Segunda variable "auxiliar" (terminación _F)**
Información Reportada	Información Reportada	1
Información a Imputar	.	0
Información a Ignorar	.I	9

Tabla 4.3: Bloques de información para las variables "auxiliares" (códigos)

Donde la información reportada, a imputar y a ignorar (primera columna) se obtiene de acuerdo con la Tabla 4.2 (sin olvidar aplicar el ajuste por la presencia de la NO RESPUESTA por unidad y por criterios de consistencia).

La segunda variable "auxiliar" (tercera columna) funciona como indicadora para diferenciar a los tres bloques de información y para simplificar las restricciones (al reescribirlas) mientras que la primera variable "auxiliar" (segunda columna) será la que realmente se imputará y después de ello, incluirá además de la información reportada, a los valores imputados que sustituyen al carácter "." y a una cantidad constante asignada por IVEware que reemplaza al texto ".I" (la constante es 0 para variables continuas y la máxima categoría reportada incrementada en una unidad para categóricas).
La recodificación de las variables "auxiliares" se obtiene aplicando el programa de cómputo RECODEINEGI.PRG desarrollado en Visual Fox (para consulta, ver Anexo F).
Con relación a las cotas, sólo deben actualizarse cambiando el nombre de las 3 variables originales por la variable "auxiliar" recodificada correspondiente (terminación "_F") para quedar según la indicado por la Tabla 4.4

VARIABLE	COTAS
EDAD_F	12,…,130
VIVOS_F	0,…,25
LOG_ING_F	0,…,13.815508

Tabla 4.4: Cotas de las variables "auxiliares"

De lo anterior, el archivo a imputar definitivo está compuesto de 944,285 registros y 49 variables, de las cuales 29 pertenecen a la base de datos que soporta la investigación y 20 fueron agregadas durante la aplicación de las fases 4 a 6.

En este momento, es relevante dimensionar el problema por resolver contabilizando los bloques de información definitivos para cada variable en estudio; en particular, es importante identificar el volumen de registros a imputar.
Los totales definitivos del archivo a imputar para las variables en estudio se presentan en la Tabla 4.5.

Variable	Información a Imputar	Información a Ignorar	Información Reportada
LOG_ING_CM	*36,854*	*610,900*	*296,531*
SEXO	*0*	*4,442*	*939,843*
EDAD_CM	*2,302*	*282,120*	*659,863*
NIV_ESC_CM	*7,859*	*120,579*	*815,847*
OCUPAC_CM	*8,209*	*611,983*	*324,093*
POS_TRA_CM	*8,487*	*611,983*	*323,815*
PARENTES_C	*1,460*	*215,562*	*727,263*
ESTRATO	*0*	*4,442*	*939,843*
SUBESTRA	*0*	*4,442*	*939,843*
ZONA	*0*	*4,442*	*939,843*
EDO_CONY_C	*1,626*	*282,120*	*660,539*
VIVOS_CM	*6,003*	*595,801*	*342,481*
MOTRIZ_CM	*3,691*	*13,959*	*926,635*

Tabla 4.5: Bloques de información para las variables predictoras y el ingreso (totales)

Lógicamente, el conteo se obtiene a partir de la segunda variable "auxiliar" aunque cuando no exista información a imputar o restricción en alguna variable específica (no existen las variables "auxiliares"), se debe recurrir directamente a la variable original sin olvidar omitir la exclusión de la NO RESPUESTA por unidad.

4.2 La Imputación: el procesamiento y la validación

El hecho de no tener antecedentes sobre la aplicación de IVEware en Censos de Población y, consecuentemente, desconocer la duración del procesamiento computacional, implica que este aspecto se diseñe bajo la ejecución de una serie de etapas secuenciales definidas en función del volumen de datos a manejar y de posibles recortes que se apliquen sobre los códigos de respuesta de las variables predictoras; en este sentido, se debe poner atención especial en estudiar los tiempos de proceso empleados por el sistema y en procurar mantener el óptimo funcionamiento del equipo de cómputo (procesamiento local, ejecución exclusiva de IVEware, máxima velocidad de proceso y máximo espacio disponible en disco duro); en particular, en esta investigación se usa una computadora marca LANIX modelo CORP3190 con un procesador INTEL CORE 2 DUO de 2.44 GHz, una memoria RAM de 2 GB y un espacio disponible en disco duro de 130 GB.

Enseguida, se describen las etapas a realizar.

- **Etapa 1. Pruebas preliminares**

  Se desarrollan 3 pruebas preliminares con la idea de construir y validar la sintaxis de ejecución pero sobre todo para obtener una estimación del tiempo de cómputo que se requiere para procesar la Imputación Simple y el primer ciclo de la Regresión Secuencial (MULTIPLES=1 e ITERATIONS=1).

  - La prueba inicial consiste en tomar los primeros 1,000 registros del archivo a imputar sin considerar orden alguno. Esta prueba arroja dos resultados relevantes:
    - ✓ Se reduce el número de categorías de la variable OCUPAC_CM mediante un recorte en su longitud (de 2 a 4 posiciones) conservando los dos dígitos iníciales (de izquierda a derecha) acorde con lo recomendado por el INEGI en su proceso de imputación. Esta medida permitirá acortar substancialmente el tiempo de cómputo que IVEware requiere para imputar dicha variable al pasar de 597 categorías a sólo 18; además, contribuye positivamente al realizar posteriormente la comparación con el procedimiento del INEGI.
    - ✓ Obtener una sintaxis de ejecución "general" libre de errores, que incluye lo referente a la importación, la Imputación Múltiple y la extracción de los archivos completos (para su consulta, ver Anexo G).
  - La segunda prueba maneja 30,000 registros conformados por los primeros 20,000, que presentan información reportada en ingresos más los primeros 10,000 registros con NO RESPUESTA en la misma variable; ejecutada esta prueba, se genera una primera aproximación real de los tiempos de procesamiento; de entrada se establece que la unidad de medida del tiempo de cómputo será la hora.
  - La tercera prueba preliminar, maneja 244,285 registros integrados por los primeros 222,431 registros con información reportada en ingresos combinados con los 21,854 registros imputados por el INEGI en dicha variable; de esta prueba se concluye que:

- ✓ Los tiempos de procesamiento arrojados, pueden emplearse para estimar el tiempo de cómputo necesario para procesar el archivo a imputar completo.
- ✓ Disminuir el número de categorías de la variable PARENTES_C, conservando solamente el primer dígito (de izquierda a derecha); al igual que en OCUPAC_CM, el ajuste aplicado es recomendado por el INEGI. El número de categorías pasa de 46 a 6 esperando con esto una mejora importante en el tiempo de cómputo definitivo.

Las estadísticas generadas por las tres pruebas preliminares se integran en la Tabla 4.6.

	Pruebas preliminares (Primer ciclo de la Regresión Secuencial e Imputación Simple)					
	muestra (1,000)		muestra (30,000)		muestra (244,285)	
Variable	información imputada	tiempo	Información imputada	tiempo	Información imputada	tiempo
parentes_c	*0*	*1 s*	*36*	*35 m*	*357*	*2 h 18 m*
pos_tra_cm	*11*	*1 s*	*139*	*1 m*	*1360*	*3 m*
ocupac_cm (2 dígitos)	*6*	*1 m*	*103*	*12 m*	*1091*	*53 m*
edo_cony_c	*0*	*1 s*	*40*	*1 m*	*400*	*4 m*
edad_cm	*3*	*1 s*	*43*	*1 m*	*449*	*1 m*
motriz_cm	*1*	*1 s*	*112*	*1 m*	*909*	*1 m*
vivos_cm	*5*	*1 s*	*172*	*1 m*	*1504*	*1 m*
niv_esc_cm	*8*	*1 s*	*157*	*12 m*	*1729*	*1 h 5 m*
log_ing_cm	*1*	*1 s*	*10007*	*1 m*	*21885*	*1 m*
tiempo total	*1 minuto*		*1 h 5 m*		*4 h 27 m*	

Tabla 4.6: Tiempos de procesamiento (pruebas preliminares, primer ciclo de la Regresión Secuencial y la Imputación Simple)

A partir de las pruebas preliminares, se concluye que procesar una corrida múltiple combinada con una Regresión Secuencial (MULT=5 e ITER=2) para una gran muestra (244,285 registros), requiere alrededor de 44 horas de tiempo continuo; con esto, se pudiese suponer que aplicar la Imputación Múltiple vía la Regresión Secuencial al archivo a imputar completo ocuparía cerca de 6 días aunque esta estimación se debe reducir significativamente al aplicar la recomendación de recortar la variable PARENTES_C.

- **Etapa 2. Prueba final (Imputación Simple y la Regresión Secuencial)**
  Esta prueba contempla la ejecución y validación numérica de la Imputación Simple vía la Regresión Secuencial (MULTIPLES=1 e ITERATIONS=2) empleando el archivo a imputar completo y sujeto a los ajustes aportados por las pruebas preliminares para las variables OCUPAC_CM y PARENTES_C.
  La información imputada y los tiempos de proceso de la prueba se muestran en la Tabla 4.7.

	Prueba final (Regresión Secuencial e Imputación Simple) 944,285 registros	
Variable	información imputada	Tiempo
parentes_c (1 dígito)	*1460*	*50 m*
pos_tra_cm	*8487*	*32 m*
ocupac_cm (2 dígitos)	*8209*	*10 h*
edo_cony_c	*1626*	*45 m*
edad_cm	*2302*	*1 m*
motriz_cm	*3691*	*1 m*
vivos_cm	*6003*	*1 m*
niv_esc_cm	*7859*	*5 h 27 m*
log_ing_cm	*36854*	*1 m*
tiempo total	*17 h 38 m*	

Tabla 4.7: Tiempos de procesamiento (prueba final, Regresión Secuencial y la Imputación Simple)

Según datos de la Tabla 4.7, procesar la Imputación Múltiple vía la Regresión Secuencial (MULTIPLES=5 e ITERATIONS=2) requerirá de un máximo de 90 horas (alrededor de 4 días) de tiempo de cómputo, cifra bastante sensata en el entorno de una aplicación censal.

La validación (liberación) de los resultados numéricos de la imputación efectuada en esta etapa, se realiza según las recomendaciones dadas en el capítulo anterior (sección 3.4); así para las nueve variables imputadas en estudio, se revisa que los bloques de información (Tabla 4.5) coincidan con el reporte estadístico alterno; una vez aprobada la comparación anterior, se procede a evaluar si la diferencia distribucional entre la información reportada e imputada es poco significativa empleando para esto los elementos estadísticos dados por el propio reporte y algunas herramientas gráficas adicionales.

- Validación de los bloques de información

  El reporte alterno se genera mediante la ejecución del programa EST_LLENAINEGI.PRG que emplea el archivo ya imputado como insumo y es desarrollado en Visual Fox (para su consulta, ver Anexo H), la estructura de salida del reporte alterno es:

```
MULTIPLE 1
edad_cm      nnumber     minimun       maximun       mean      stddev
-------------------------------------------------------------------
imputed      2302          12.00        121.00        49.89     20.39
observed   659863          12.00        130.00        33.13     16.88
combined   662165          12.00        130.00        33.19     16.92
motriz_cm           Observed             Imputed              Combined
                  Freq     Per         Freq     Per          Freq      Per
-------------------------------------------------------------------------
code 1           7,504      0.81          47      1.27       7,551      0.81
code 2         919,131     99.19       3,644     98.73     922,775     99.19
total          926,635    100.00       3,691    100.00     930,326    100.00
niv_esc_cm          Observed             Imputed              Combined
                  Freq     Per         Freq     Per          Freq      Per
-------------------------------------------------------------------------
code 0          43,974      5.39         440      5.60      44,414      5.39
```

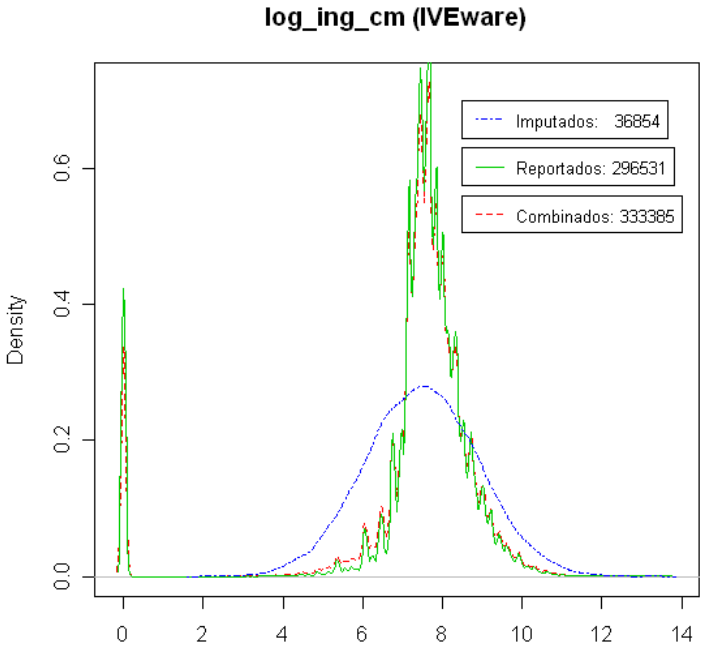
```
code 1       40,430      4.96        156      1.98     40,586      4.93
code 2      373,598     45.79      3,781     48.11    377,379     45.81
code 3      175,049     21.46        569      7.24    175,618     21.32
code 4       76,335      9.36        162      2.06     76,497      9.29
code 5        3,433      0.42          5      0.06      3,438      0.42
code 6       35,792      4.39        103      1.31     35,895      4.36
code 7       62,848      7.70         77      0.98     62,925      7.64
code 8        4,388      0.54      2,566     32.65      6,954      0.84
total       815,847    100.00      7,859    100.00    823,706    100.00
ocupac_cm          Observed            Imputed             Combined
                  Freq    Per         Freq    Per          Freq     Per
------------------------------------------------------------------------
code 11      11,871      3.66        312      3.80     12,183      3.67
code 12      11,239      3.47        342      4.17     11,581      3.49
code 13      14,434      4.45        325      3.96     14,759      4.44
code 14       2,653      0.82        352      4.29      3,005      0.90
code 21       7,827      2.42        351      4.28      8,178      2.46
code 41      22,637      6.98        351      4.28     22,988      6.92
code 51       8,401      2.59        358      4.36      8,759      2.64
code 52      54,827     16.92      1,135     13.83     55,962     16.84
code 53      33,678     10.39        605      7.37     34,283     10.32
code 54      15,251      4.71        445      5.42     15,696      4.72
code 55      16,094      4.97        475      5.79     16,569      4.99
code 61       9,620      2.97        357      4.35      9,977      3.00
code 62      23,885      7.37        400      4.87     24,285      7.31
code 71      45,932     14.17        657      8.00     46,589     14.02
code 72       5,683      1.75        335      4.08      6,018      1.81
code 81      21,522      6.64        544      6.63     22,066      6.64
code 82      11,246      3.47        431      5.25     11,677      3.51
code 83       7,293      2.25        434      5.29      7,727      2.33
total       324,093    100.00      8,209    100.00    332,302    100.00
pos_tra_cm         Observed            Imputed             Combined
                  Freq    Per         Freq    Per          Freq     Per
------------------------------------------------------------------------
code 1      234,740     72.49      2,104     24.79    236,844     71.27
code 2       19,299      5.96      1,072     12.63     20,371      6.13
code 3       10,551      3.26        135      1.59     10,686      3.22
code 4       51,120     15.79      5,090     59.97     56,210     16.92
code 5        8,105      2.50         86      1.01      8,191      2.46
total       323,815    100.00      8,487    100.00    332,302    100.00
vivos_cm    nnumber    minimun    maximun    mean     stddev
------------------------------------------------------------
imputed       6003       0.00      16.00     2.48      2.20
observed    342481       0.00      25.00     2.65      3.37
combined    348484       0.00      25.00     2.65      3.35
parentes_c         Observed            Imputed             Combined
                  Freq    Per         Freq    Per          Freq     Per
------------------------------------------------------------------------
code 2      161,488     22.20          0      0.00    161,488     22.16
code 3      481,942     66.27        491     33.63    482,433     66.20
code 4        1,009      0.14        496     33.97      1,505      0.21
code 5        2,472      0.34          1      0.07      2,473      0.34
code 6       80,352     11.05        472     32.33     80,824     11.09
total       727,263    100.00      1,460    100.00    728,723    100.00
edo_cony_c         Observed            Imputed             Combined
                  Freq    Per         Freq    Per          Freq     Per
------------------------------------------------------------------------
code 1       28,290      4.28         12      0.74     28,302      4.27
code 2       10,848      1.64         92      5.66     10,940      1.65
code 3        6,406      0.97         17      1.05      6,423      0.97
code 4       24,193      3.66         63      3.87     24,256      3.66
code 5       34,389      5.21         32      1.97     34,421      5.20
code 6        3,873      0.59          0      0.00      3,873      0.58
code 7      295,249     44.70        397     24.42    295,646     44.65
code 8      257,291     38.95      1,013     62.30    258,304     39.01
total       660,539    100.00      1,626    100.00    662,165    100.00
log_ing_cm   nnumber    minimun    maximun    mean     stddev
------------------------------------------------------------
imputed       36854       2.05      13.38     7.48       1.42
observed     296531       0.00      13.71     7.36       1.92
combined     333385       0.00      13.71     7.37       1.87
```

Para cada variable imputada, los totales indicados por el reporte alterno para la información imputada (columna “nnumber” y fila “imputed” - variable continua o columna “Imputed/Freq” y fila “total” - variable categórica) y la información reportada (columna “nnumber” y fila “observed” - variable continua o columna “Observed/Freq” y fila “total” - variable categórica) coinciden con los correspondientes de la Tabla 4.5. La información a ignorar se válida automáticamente al ser el complemento natural de los dos bloques ya verificados anteriormente.

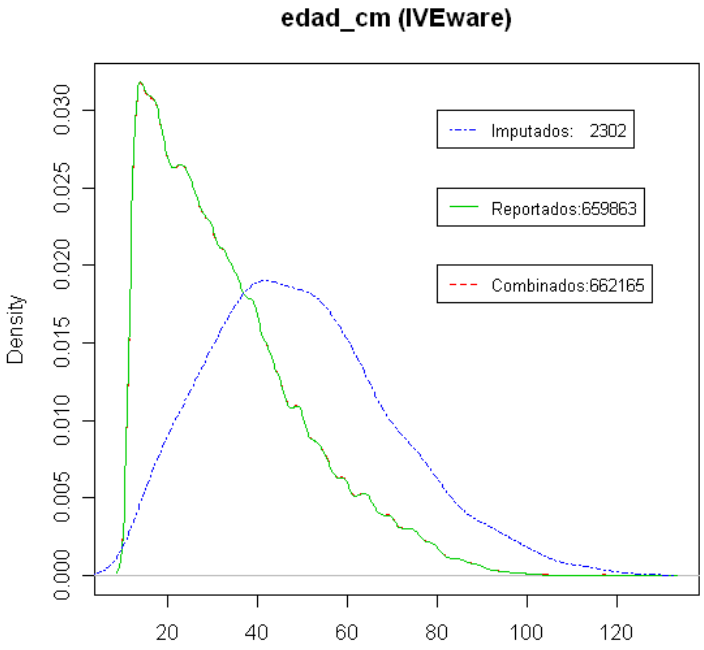
- Comparación de la información reportada e imputada

  Los estadísticos a emplear para comparar la información imputada y la reportada son: la media, la desviación estándar (ambas para una variable continua) y del porcentaje del código de respuesta (variable categórica); estos elementos, son proporcionados por el reporte estadístico alterno.

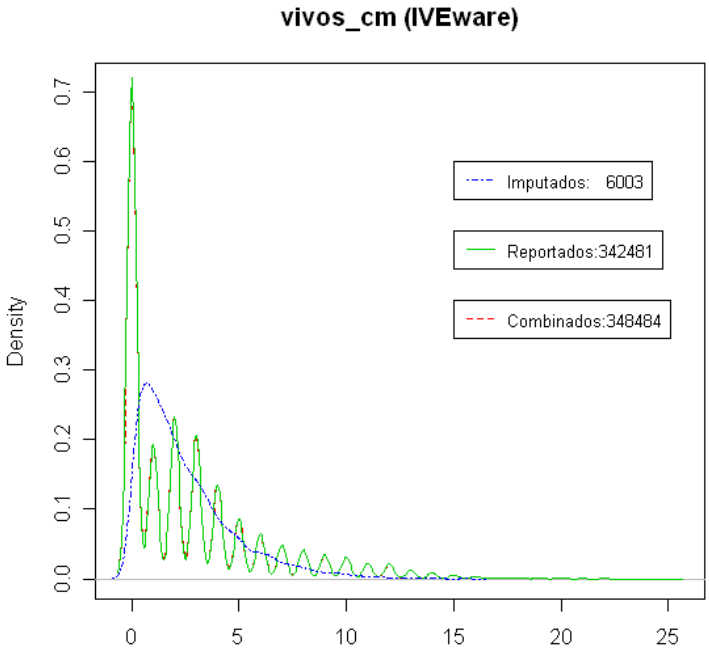
  Revisando la salida impresa del reporte alterno para cada una de las variables imputadas en estudio, se aprecia que la media, la desviación estándar (columnas “mean”, “stddev”) y del porcentaje del código de respuesta (columna “Per”) entre la información reportada (“Observed”) y la combinada (“Combined”) presentan cambios mínimos; es decir, el efecto de la información imputada (“Imputed”) sobre la información reportada (“Observed”) es prácticamente nulo; en las Gráficas 4.1 a 4.9 se puede confirmar la aseveración.



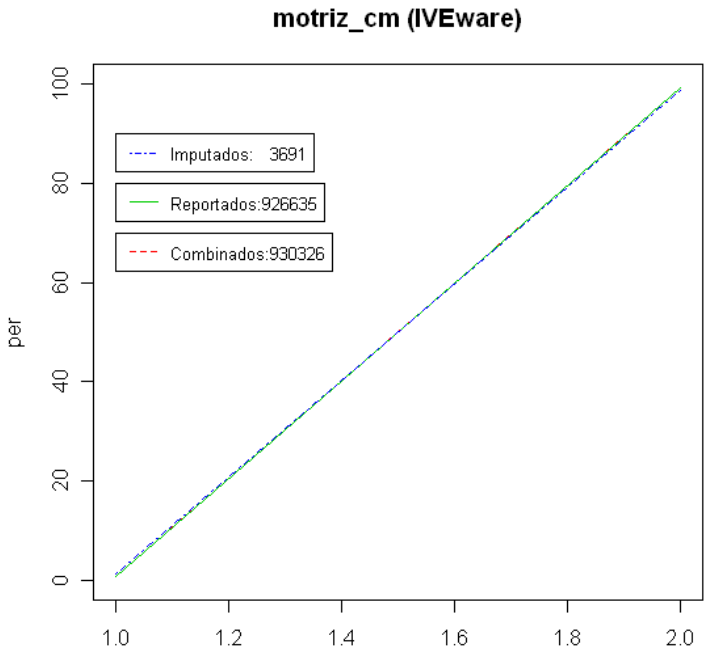
Gráfica 4.1: Distribución de la información reportada, imputada y combinada de LOG_ING_CM (Regresión Secuencial y la Imputación Simple)



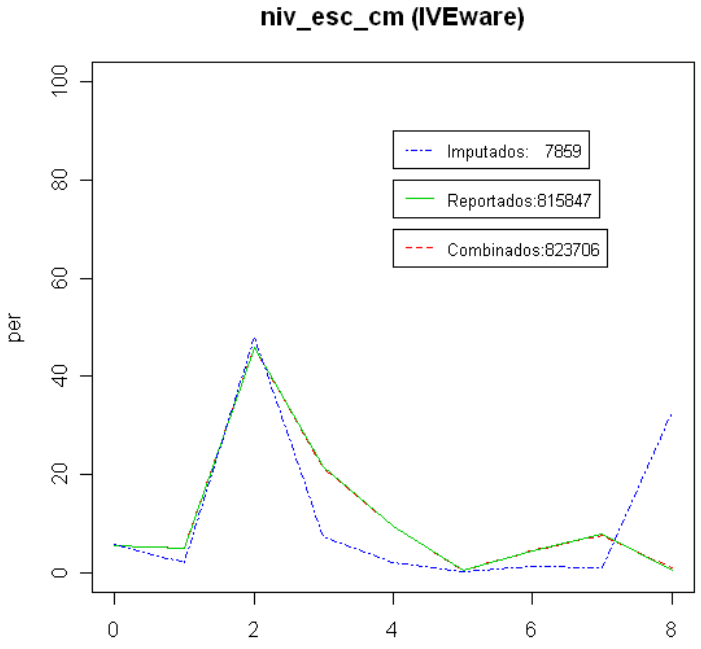
Gráfica 4.2: Distribución de la información reportada, imputada y combinada de EDAD_CM (Regresión Secuencial y la Imputación Simple)



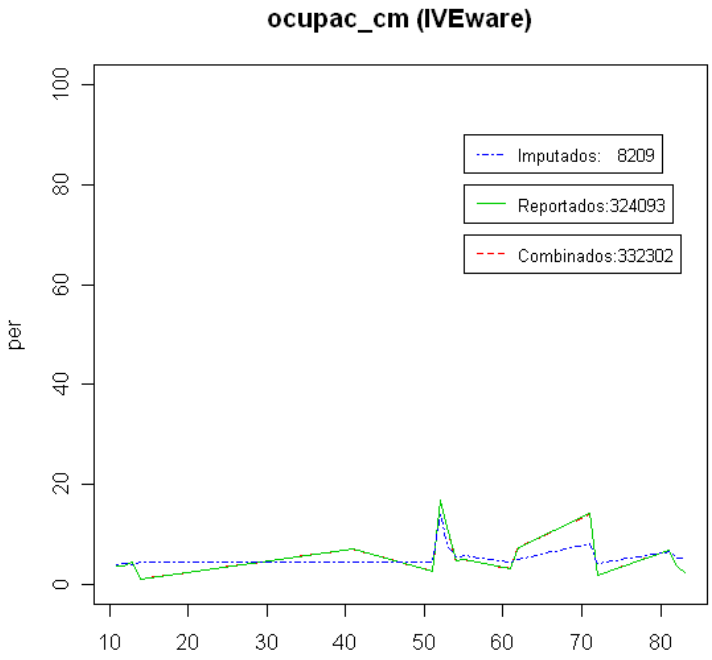
Gráfica 4.3: Distribución de la información reportada, imputada y combinada de VIVOS_CM (Regresión Secuencial y la Imputación Simple)



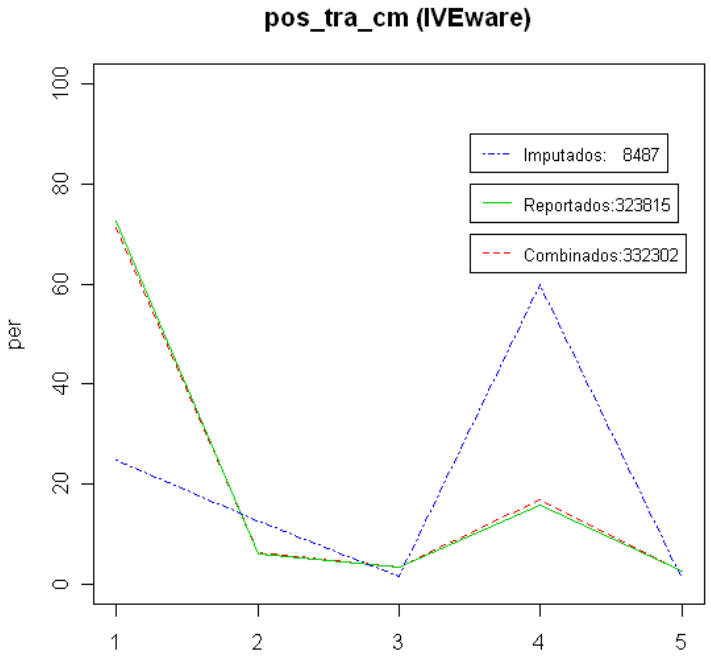
Gráfica 4.4: Distribución de la información reportada, imputada y combinada de MOTRIZ_CM (Regresión Secuencial y la Imputación Simple)



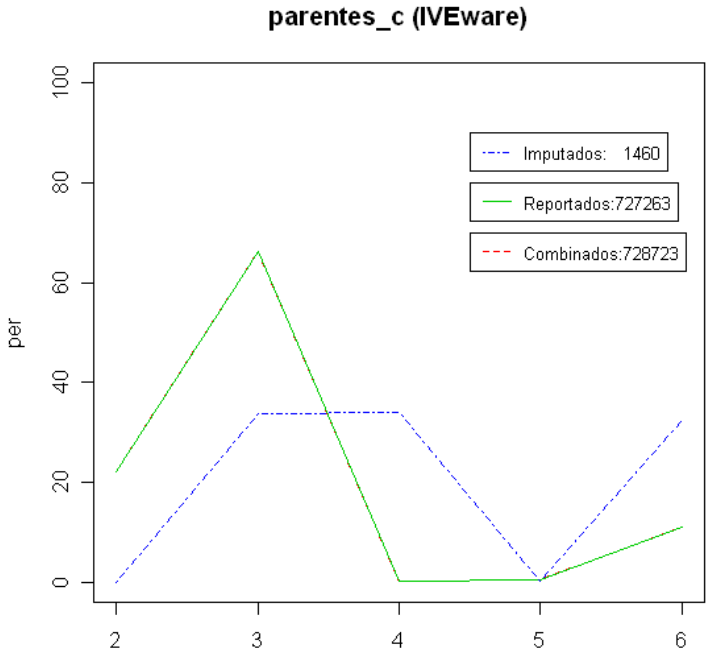
Gráfica 4.5: Distribución de la información reportada, imputada y combinada de NIV_ESC_CM (Regresión Secuencial y la Imputación Simple)



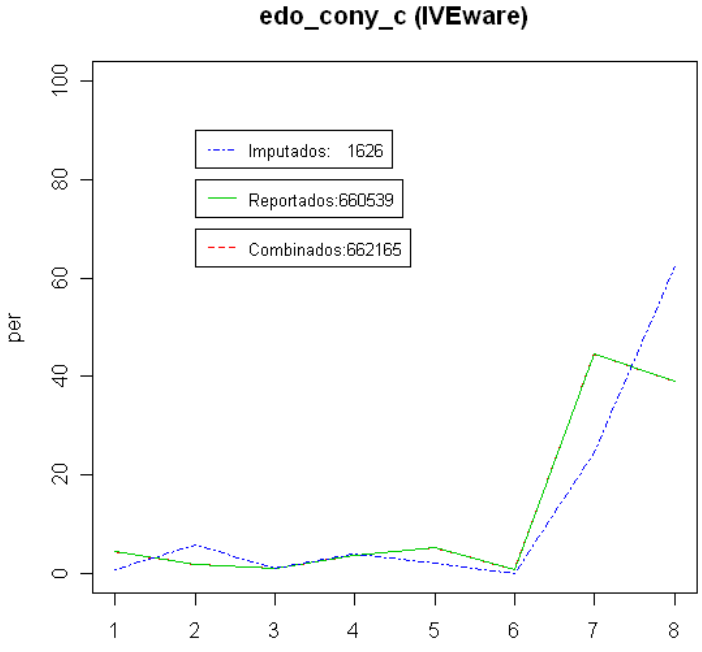
Gráfica 4.6: Distribución de la información reportada, imputada y combinada de OCUPAC_CM (Regresión Secuencial y la Imputación Simple)



Gráfica 4.7: Distribución de la información reportada, imputada y combinada de POS_TRA_CM (Regresión Secuencial y la Imputación Simple)



Gráfica 4.8: Distribución de la información reportada, imputada y combinada de PARENTES_C (Regresión Secuencial y la Imputación Simple)



Gráfica 4.9: Distribución de la información reportada, imputada y combinada EDO_CONY_C (Regresión Secuencial y la Imputación Simple)

Además de confirmar el idéntico comportamiento entre la información reportada y la combinada, en general, las Gráficas 4.1 a 4.9 muestran que la información imputada también presenta una distribución similar, aunque es notable que las gráficas 4.7 y 4.8 destacan cierta diferencia se espera que esta disminuya al procesar la imputación múltiple; luego, se tiene que la diferencia distribucional entre la información reportada y la imputada es poco significativa.

Una vez aprobada la igualdad de los bloques de información y aceptadas como poco significativas las diferencias distribucionales entre la información reportada y la imputada, esta etapa, culmina su proceso liberando satisfactoriamente los resultados numéricos de la Imputación Simple combinada con la Regresión Secuencial.

- **Etapa 3. Prueba final (Imputación Múltiple y la Regresión Secuencial)**

Esta prueba equivale a repetir la etapa anterior en varias ocasiones (expertos en la materia recomiendan 5 veces), es decir, consiste simplemente en adecuar la sintaxis de ejecución modificando el parámetro MULTIPLES de 1 a 5, además, implica importar los cinco archivos imputados. Los tiempos de esta prueba son:

	Prueba final (Regresión Secuencial e Imputación Múltiple) 944,285 registros	
Variable	información imputada	Tiempo
parentes_c (1 dígito)	*1460*	*4 h 2 m*
pos_tra_cm	*8487*	*2 h 42 m*
ocupac_cm (2 dígitos)	*8209*	*50 h 30 m*
edo_cony_c	*1626*	*3 h 35 m*
edad_cm	*2302*	*10 m*
motriz_cm	*3691*	*10 m*
vivos_cm	*6003*	*10 m*
niv_esc_cm	*7859*	*29 h 45 m*
log_ing_cm	*36854*	*10 m*
tiempo total	*91 h 14 m*	

Tabla 4.8: Tiempos de procesamiento (prueba final, Regresión Secuencial y la Imputación Múltiple)

De la Tabla 4.8, se observa que el tiempo de procesamiento requerido se incrementa de manera notable con respecto al empleado por la segunda etapa en alrededor de 5 veces más; con esto, la estimación del tiempo total indicada por dicha etapa (90 horas), resultó ser bastante acertada.

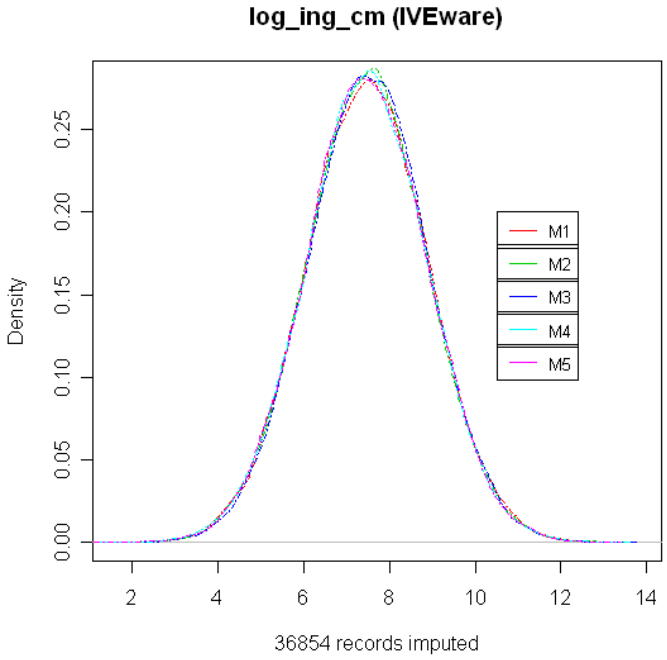
Partiendo del hecho de que la Imputación Simple ya fue aprobada y que representa a la primera corrida del proceso múltiple, la liberación numérica en esta fase sólo valida la repetibilidad del sistema, lo cual consiste en revisar que los valores imputados no cambien substancialmente entre las distintas imputaciones simples (propiedad de estabilidad de la Imputación Múltiple). Luego, para liberar los resultados numéricos en esta etapa, es suficiente con asegurar que:

- Los tres bloques de información sean iguales en tamaño para las cinco imputaciones, esto se puede verificar en el reporte estadístico alterno.
- Para cada registro imputado, las cinco imputaciones simples converjan a (o repitan) un mismo valor, lo que puede comprobarse al verificar que la media y la desviación estándar de la información imputada son también convergentes.

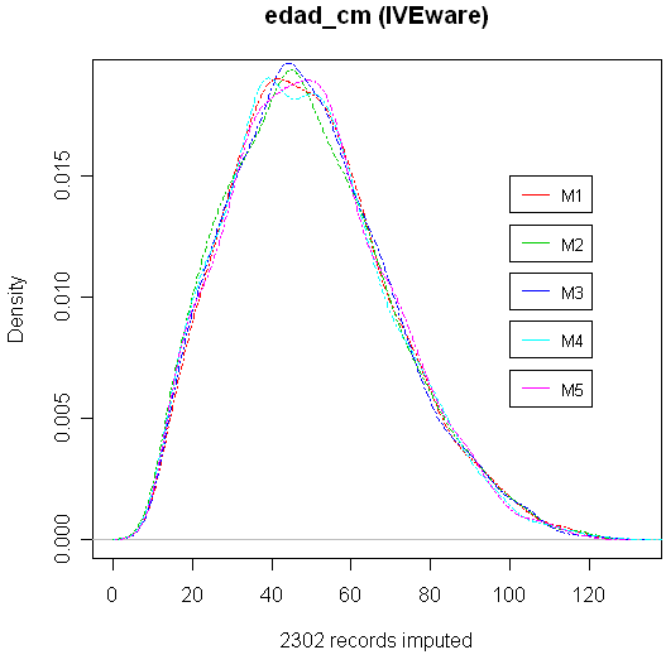
El reporte estadístico alterno para las variables EDAD e INGRESO (ver reporte completo en Anexo I) presenta la siguiente estructura:

```
MULTIPLE 1 (M1)
edad_cm    nnumber  minimun       maximun     mean    stddev
-------------------------------------------------------------
imputed     2302   12.00         121.00      49.89    20.39
observed  659863   12.00         130.00      33.13    16.88
combined  662165   12.00         130.00      33.19    16.92
log_ing_cm  nnumber   minimun     maximun     mean    stddev
-------------------------------------------------------------
imputed     36854      2.05        13.38      7.48     1.42
observed   296531      0.00        13.71      7.36     1.92
combined   333385      0.00        13.71      7.37     1.87
MULTIPLE 2 (M2)
edad_cm    nnumber    minimun    maximun     mean     stddev
-------------------------------------------------------------
imputed      2302      12.00      119.00     49.33     20.77
observed   659863      12.00      130.00     33.13     16.88
combined   662165      12.00      130.00     33.19     16.92
log_ing_cm   nnumber   minimun   maximun   mean    stddev
-------------------------------------------------------------
imputed       36854     1.45      13.30    7.48     1.40
observed     296531     0.00      13.71    7.36     1.92
combined     333385     0.00      13.71    7.37     1.87
MULTIPLE 3 (M3)
edad_cm    nnumber   minimun    maximun   mean    stddev
-------------------------------------------------------------
imputed      2302    12.00     125.00    49.71   20.32
observed   659863    12.00     130.00    33.13   16.88
combined   662165    12.00     130.00    33.19   16.92
log_ing_cm  nnumber    minimun   maximun   mean   stddev
-------------------------------------------------------------
imputed      36854      1.95      13.29    7.51    1.39
observed    296531      0.00      13.71    7.36    1.92
combined    333385      0.00      13.71    7.38    1.87
MULTIPLE 4 (M4)
edad_cm    nnumber   minimun    maximun    mean    stddev
-------------------------------------------------------------
imputed     2302      12.00      127.00    49.27    20.37
observed  659863      12.00      130.00    33.13    16.88
combined  662165      12.00      130.00    33.19    16.92
log_ing_cm  nnumber   minimun    maximun     mean   stddev
-------------------------------------------------------------
imputed      36854     1.71       13.24      7.47    1.41
observed    296531     0.00       13.71      7.36    1.92
combined    333385     0.00       13.71      7.37    1.87
MULTIPLE 5 (M5)
edad_cm    nnumber   minimun   maximun    mean     stddev
-------------------------------------------------------------
imputed     2302     12.00     119.00    49.69     20.23
observed  659863     12.00     130.00    33.13     16.88
combined  662165     12.00     130.00    33.19     16.92
log_ing_cm  nnumber    minimun   maximun   mean    stddev
-------------------------------------------------------------
imputed      36854      1.45      12.91    7.47     1.41
observed    296531      0.00      13.71    7.36     1.92
combined    333385      0.00      13.71    7.37     1.87
```

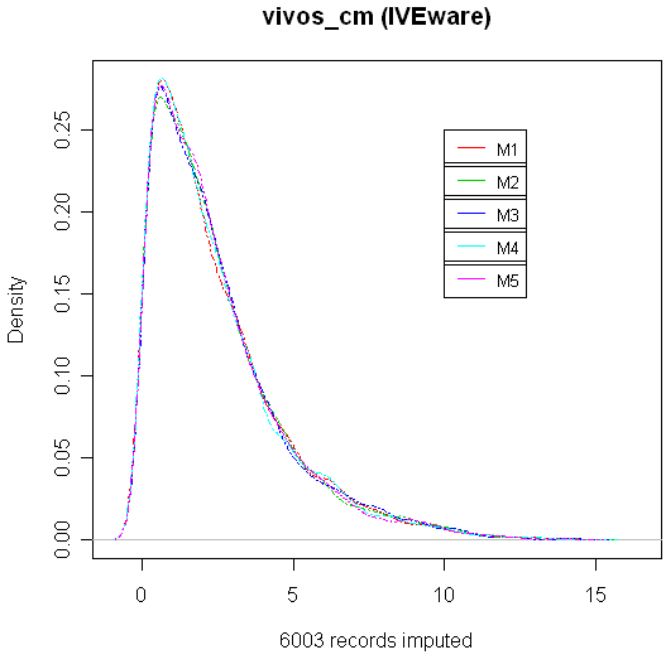
Como es de esperarse, los totales de los bloques de información (imputada, reportada y por ignorar) son iguales para las cinco corridas.
La convergencia de la media y de la desviación estándar de las variables imputadas en estudio para la información imputada, se evidencia en las Graficas 4.10 a 4.17.



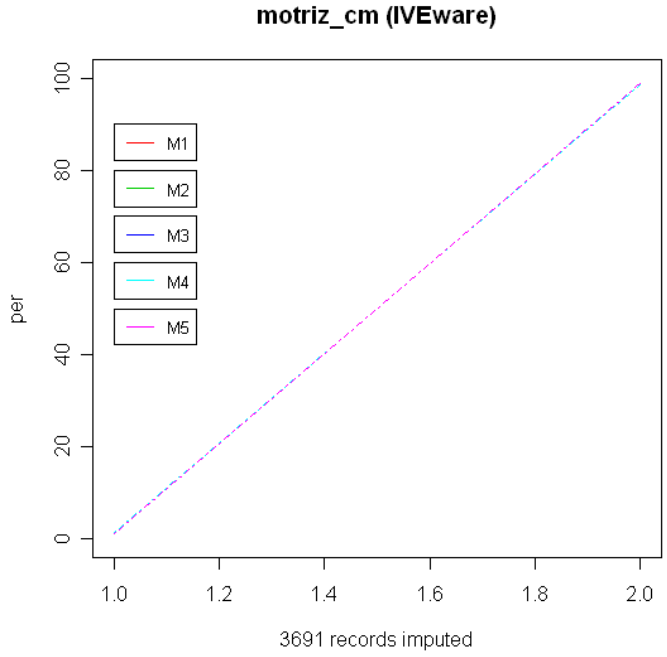
Gráfica 4.10: Distribución de la información imputada de LOG_ING_CM (Regresión Secuencial y la Imputación Múltiple)



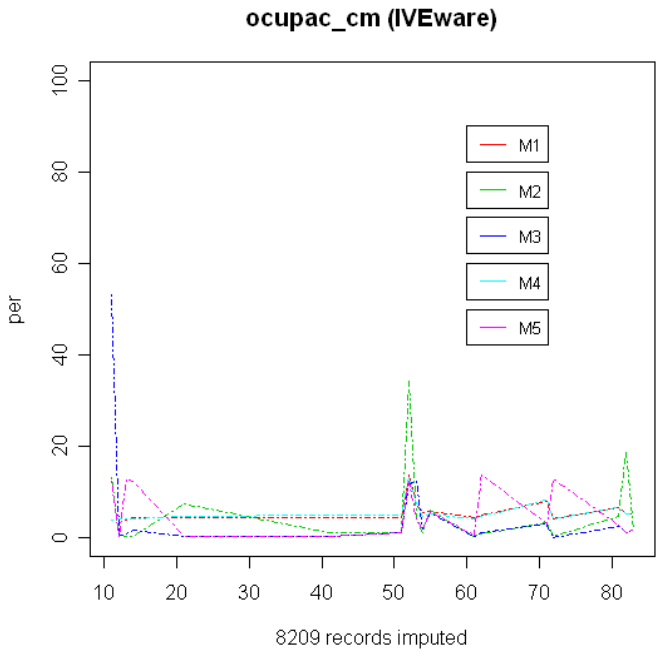
Gráfica 4.11: Distribución de la información imputada de EDAD_CM (Regresión Secuencial y la Imputación Múltiple)



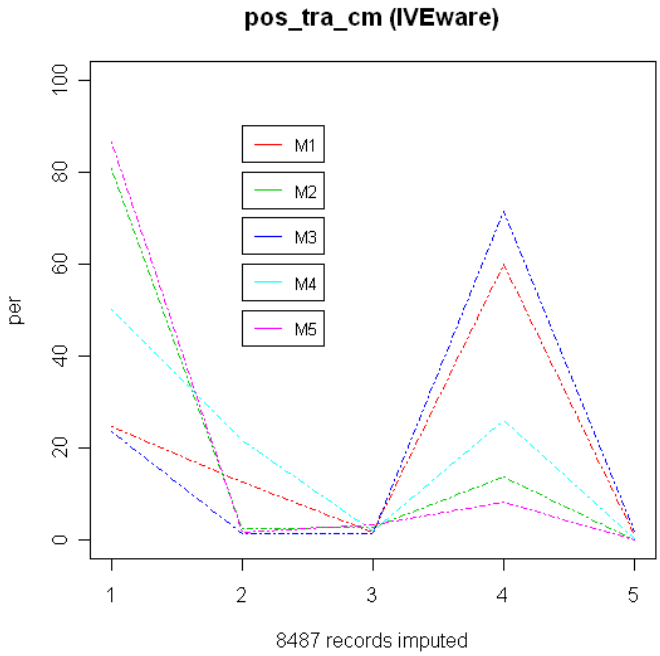
Gráfica 4.12: Distribución de la información imputada de VIVOS_CM (Regresión Secuencial y la Imputación Múltiple)



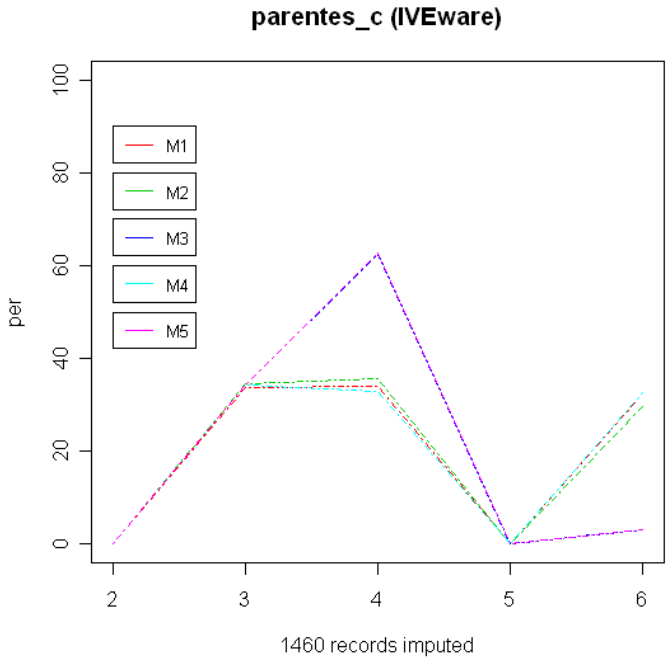
Gráfica 4.13: Distribución de la información imputada de MOTRIZ_CM (Regresión Secuencial y la Imputación Múltiple)



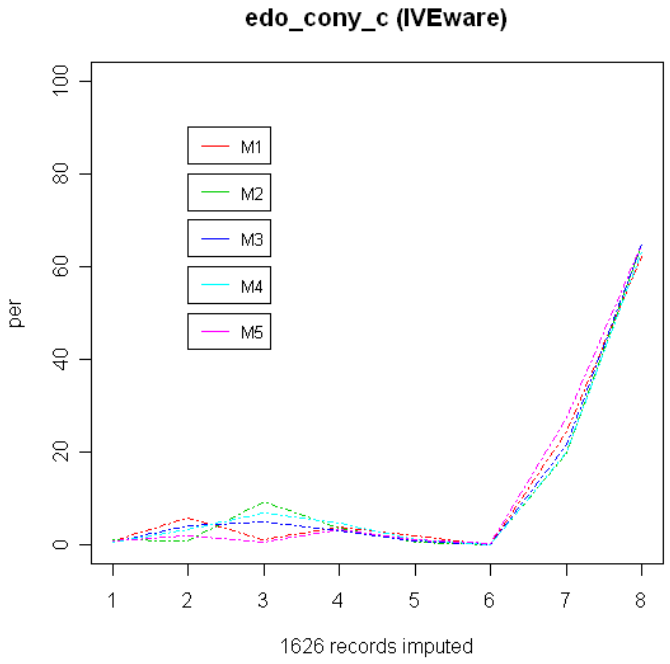
Gráfica 4.14: Distribución de la información imputada de OCUPAC_CM (Regresión Secuencial y la Imputación Múltiple)



Gráfica 4.15: Distribución de la información imputada de POS_TRA_CM (Regresión Secuencial y la Imputación Múltiple)



Gráfica 4.16: Distribución de la información imputada de PARENTES_C (Regresión Secuencial y la Imputación Múltiple)



Gráfica 4.17: Distribución de la información imputada de EDO_CONY_C (Regresión Secuencial y la Imputación Múltiple)

En todas las variables continuas y en algunas categóricas (motriz_cm y edo_cony_c) la repetibilidad es evidente; en el resto de las variables aunque la información imputada presenta diferencias (a lo más en 2 códigos), los cinco valores persiguen una misma tendencia.

De las conclusiones emitidas a partir del análisis del reporte y de la revisión de las gráficas, se considera a la Imputación Múltiple vía la Regresión Secuencial como liberada.

- **Etapa 4. Transformación inversa del ingreso**

  La última etapa del procesamiento, radica en aplicar una transformación numérica a la variable LOG_ING_CM, dicha operación consiste en convertir la información reportada e imputada a la escala natural del ingreso mediante la aplicación de la función exponencial; el resultado de dicha operación, se registra en la variable TOT_ING_IM la cual se diseñó ex profeso en la fase 4 de la sección 4.1.

Una vez realizadas las etapas anteriores, el procesamiento finaliza su ejecución reportando cinco archivos (M1, M2, M3, M4 y M5) ya imputados y completos, listos para su explotación y análisis.

En el siguiente capítulo, se discute una serie de resultados arrojados por el análisis de la imputación ya realizada, los cuales serán fundamentales para emitir las conclusiones de la investigación.

Capítulo 5

Principales resultados de la investigación

Los resultados de la imputación efectuada mediante RSIM (método alternativo) se analizan en dos sentidos; inicialmente bajo un enfoque interno, revisando el comportamiento de algunas variables en estudio imputadas según las estimaciones generadas por la propia técnica y, posteriormente, se efectúa un contraste con las estimaciones del ingreso aportadas por la imputación del INEGI (fuente externa) sujeto a una serie de consideraciones debido a las características de aplicación de cada método.

El análisis se hace para grupos de información muy específicos, ya que el hecho de desconocer el valor observado de la NO RESPUESTA del ingreso, no permite comparar ni mucho menos evaluar la imputación a un nivel individual (por registro); así, se estudian el comportamiento de los tres bloques de información (reportada, a imputar y a ignorar) revisando ciertos estadísticos particulares tanto en forma tabular como de manera gráfica.

5.1 Análisis interno

En un sentido estricto, este análisis tiene inicio en el capítulo 4 (etapas 2 y 3 de la sección 4.2) al momento de contrastar la distribución de la información imputada con la reportada; cabe recordar que los resultados de tal comparación, permitieron declarar como liberada a la imputación; así, este análisis desarrolla aspectos adicionales que en parte confirman dicha decisión pero que a la vez permiten lograr un mayor nivel de explotación de la información.

- **Estimaciones individuales**

  A partir de la selección de un parámetro de interés (media, tasa, razón, proporción, total entre otros), se comparan las estimaciones correspondientes de las cinco imputaciones simples (M1, M2, M3, M4 y M5) entre si y contra la estimación individual de la información reportada, también, se revisa el efecto de la imputación sobre la información reportada al contrastar las cinco estimaciones de la información combinada (reportada e imputada) nuevamente contra la estimación individual de la información reportada.

  Sin pérdida de generalidad, se elige como parámetros a:

  - La media (promedio) para las variables continuas
  - El porcentaje del código de respuesta para las variables categóricas

  La Tabla 5.1, ilustra las estimaciones obtenidas a partir del reporte estadístico alterno (Anexo I) para el logaritmo del ingreso (LOG_ING_CM) y las variables número de hijos nacidos vivos (VIVOS_CM), edad (EDAD_CM) y parentesco (PARENTES_C), para las cinco imputaciones simples se agregan algunas estadísticas integradas como son la media, la desviación estándar y los intervalos de confianza.

Variable	Información imputada							LI	LS	Información reportada
	M1	M2	M3	M4	M5	media	desviación std			
LOG_ING_CM (media)	7.48	7.48	7.51	7.47	7.47	7.482	0.016431677	7.43270497	7.53129503	7.36
VIVOS_CM (media)	2.48	2.49	2.5	2.46	2.46	2.48	0.017888544	2.424334369	2.531665631	2.65
EDAD_CM (media)	49.89	49.33	49.71	49.27	49.69	49.58	0.266308092	48.77907572	50.37692428	33.13
PARENTES_C (porcentaje del código de respuesta)										
2	0	0	0.07	0	0.07	0.03	0.038340579	-0.08702174	0.143021737	22.2
3	33.63	34.52	34.38	34.32	34.25	34.22	0.344456093	33.18663172	35.25336828	66.27
4	33.97	35.68	62.47	33.01	62.6	45.55	15.53828916	-1.06886748	92.16086748	0.14
5	0.07	0.07	0	0.07	0.07	0.06	0.031304952	-0.03791486	0.149914855	0.34
6	32.33	29.73	3.08	32.6	3.01	20.15	15.65483791	-26.8145137	67.11451373	11.05
Variable	Información combinada (imputada y reportada)							LI	LS	Información reportada
	M1	M2	M3	M4	M5	media	desviación std			
LOG_ING_CM (media)	7.37	7.37	7.38	7.37	7.37	7.37	0.004472136	7.358583592	7.385416408	7.36
VIVOS_CM (media)	2.65	2.65	2.65	2.65	2.65	2.65	0	2.65	2.65	2.65
EDAD_CM (media)	33.19	33.19	33.19	33.19	33.19	33.19	0	33.19	33.19	33.13
PARENTES_C (porcentaje del código de respuesta)										
2	22.16	22.16	22.16	22.16	22.16	22.16	0	22.16	22.16	22.2
3	66.2	66.2	66.2	66.2	66.2	66.20	0	66.2	66.2	66.27
4	0.21	0.21	0.26	0.2	0.26	0.23	0.029495762	0.139512713	0.316487287	0.14
5	0.34	0.34	0.34	0.34	0.34	0.34	0	0.34	0.34	0.34
6	11.09	11.09	11.03	11.09	11.03	11.07	0.032863353	10.96740994	11.16459006	11.05

Tabla 5.1: Estimación de la media y del porcentaje del código de respuesta para el ingreso y algunas variables predictoras (Imputación Simple)

La Tabla 5.1 ilustra varias situaciones de interés:

- Para las variables LOG_ING_CM, VIVOS_CM y EDAD_CM de la información imputada (primera parte de la tabla), las cinco imputaciones simples (columnas 2 a 6) presentan una variación muy pequeña (columna 8), es decir, la imputación resultó ser estable.
- Por otro lado, al comparar la estimación de la media (columna 7) de la información imputada (primera parte de la tabla) contra la estimación de la información reportada se aprecian diferencias; en el caso del ingreso, el modelo empleado por RSIM supone que la información imputada debería tener mayores ingresos que los de la información reportada, este comportamiento probablemente se debe a que la NO RESPUESTA en esta variable normalmente está asociada a la Población que tiene un ingreso alto.

- Ahora bien, en la información combinada (segunda parte de la tabla) las diferencias mencionadas en el punto anterior, se vuelven prácticamente nulas (para todas las variables) lo cual se debe a que la proporción entre la información imputada y la información reportada es pequeña y a que distribucionalmente dichos bloques de información son muy similares.

- **Estimaciones con RSIM**

  Empleando nuevamente como insumo el reporte estadístico alterno (Anexo I) y de acuerdo con las expresiones indicadas en el capítulo 2, se obtienen las estimaciones de la media del ingreso y de su varianza según RSIM para, con esto, lograr dimensionar la magnitud de las estimaciones al compararlas contra las correspondientes de la información reportada.

  En particular, las estimaciones según RSIM de la información imputada y de la combinada (imputada y reportada) se muestran en la Tabla 5.2.

Variable (parámetro)	Información imputada							Información reportada
	M1	M2	M3	M4	M5	media	Varianza	
LOG_ING_CM (media)	*7.48*	*7.48*	*7.51*	*7.47*	*7.47*	*7.4817*	*.00022*	*7.36*
LOG_ING_CM (varianza)	*2.0164*	*1.96*	*1.9321*	*1.9881*	*1.9881*	*1.97694*	*1.977204*	*3.6864*
Variable (parámetro)	Información combinada (imputada y reportada)							Información reportada
	M1	M2	M3	M4	M5	media	Varianza	
LOG_ING_CM (media)	*7.37*	*7.37*	*7.38*	*7.37*	*7.37*	***7.372***	*.00002*	*7.36*
LOG_ING_CM (varianza)	*3.4969*	*3.4969*	*3.4969*	*3.4969*	*3.4969*	*3.4969*	***3.496924***	*3.6864*

Tabla 5.2: Estimación del ingreso (Imputación Múltiple)

Según la Tabla 5.2, se tiene que las estimaciones con RSIM de la media del ingreso y de su varianza correspondiente (ambas subrayadas) son muy similares con las de la información reportada; bajo estos términos, la Imputación Múltiple efectuada presenta un efecto numérico prácticamente nulo sobre la información reportada, además, el hecho de que la estimación de la varianza esté acotada superiormente por la varianza de la información reportada, implica que dicha estimación sea considerada como aceptable en magnitud.

Así pues, el enfoque interno confirma que la técnica RSIM genera resultados numéricos satisfactorios, avalando con esto el uso de cualquiera de las cinco imputaciones simples como representativo de la Imputación Múltiple.

5.2 Análisis externo

Una vez analizada la aplicación en forma individual, el trabajo se enfoca en desarrollar actividades encaminadas a satisfacer el objetivo secundario de la investigación, el cual consiste en comparar la Imputación Múltiple (RSIM) contra el método de imputación que aplicó el INEGI en el año 2000 (fuente externa).

Previo a la comparación, resulta fundamental conocer la estructura que presenta la Población Ocupada dentro de ambas investigaciones en función de los códigos de

respuesta del ingreso y de los bloques de información; con esto, se logra determinar la existencia de diferencias entre ambas opciones además de distinguir sus causas.
Para el INEGI considerada como fuente externa, la estructura de la Población Ocupada es:

<table>
<tr><td colspan="5">POBLACIÓN OCUPADA (332,302)</td></tr>
<tr><td colspan="5">¿Filtro del INEGI?</td></tr>
<tr><td colspan="3">OK (314,524)</td><td colspan="2">~~OK~~ (17,778)</td></tr>
<tr><td colspan="3">¿$=999,999?</td><td>¿$=999,999?</td><td>¿$=999,998?</td><td>¿$<>999,999 y <>999,998?</td></tr>
<tr><td rowspan="3">OK (9,313) no repuesta actual</td><td colspan="2">~~OK~~ (305,211)</td><td rowspan="3">OK (5,572) no repuesta actual</td><td rowspan="3">OK (8)</td><td rowspan="3">OK (12,198)</td></tr>
<tr><td colspan="2">¿Impute= b?</td></tr>
<tr><td>OK (283,357) <u>información reportada</u> (incluye 107 registros con $=999998)</td><td>~~OK~~ (21,854) <u>información a imputar</u></td></tr>
</table>

Tabla 5.3: Estructura de la Población Ocupada (INEGI)

Para la técnica RSIM, la estructura de la Población es:

<table>
<tr><td colspan="4">POBLACIÓN OCUPADA (332,302)</td></tr>
<tr><td colspan="4">¿Impute = b?</td></tr>
<tr><td colspan="3">OK (310,448)</td><td rowspan="3">~~OK~~ (21,854) <u>información a imputar</u></td></tr>
<tr><td>¿$=999,999?</td><td>¿$=999,998?</td><td>¿$<>999,999 y <>999,998?</td></tr>
<tr><td>OK (14,885) <u>información a imputar</u></td><td>OK (115) <u>información a imputar</u></td><td>OK (295,448) <u>información reportada</u> (para obtener los 296,531 del reporte alterno, resta agregar 1,083 registros que no forman parte de la Población Ocupada pero reportan ingreso 0)</td></tr>
</table>

Tabla 5.4: Estructura de la Población Ocupada (RSIM)

De acuerdo con las Tablas 5.3 y 5.4, las dos metodologías presentan ciertas particularidades que las hacen distintas, aunque también es cierto que el método alternativo se ha desarrollado procurando lograr la máxima similitud con la opción del INEGI; como consecuencia, la propia comparación y la interpretación de los resultados deben tomar muy en cuenta esta situación.

La comparación entre ambas opciones se realiza sujeta a las acotaciones siguientes:

- La variable por analizar es exclusivamente el ingreso, puesto que el INEGI sólo imputó dicha variable.
- La información a imputar se ajusta a la definida por el INEGI, ya que es la común entre ambas opciones, es decir, el total de registros imputados se modifica de 36,854 a 21,854, la diferencia (15,000 registros) pasa a formar parte de la información a ignorar (Tablas 5.3 y 5.4).

- La información reportada es distinta, el INEGI maneja 283,357 y el método alternativo 296,531 (Tablas 5.3 y 5.4).
- La técnica de imputación aplicada por el método alternativo es la Regresión Secuencial combinada con la Imputación Múltiple mientras que el INEGI imputa bajo una técnica de tipo determinista.
- El INEGI emplea 6 variables predictoras y el método alternativo adiciona 6 más, usando un total de 12 variables predictoras (en 8 de ellas se aplica al menos una imputación).

El proceso comparativo a realizar es similar al de la sección anterior pero incorporando las estimaciones generadas a partir de la imputación del INEGI y considerando el ajuste en la información a imputar; así, primero se define un parámetro de interés, a continuación se revisan las estimaciones de la media y varianza respectivas dentro de los bloques de información y finalmente se analiza el comportamiento distribucional mediante el apoyo de gráficos.

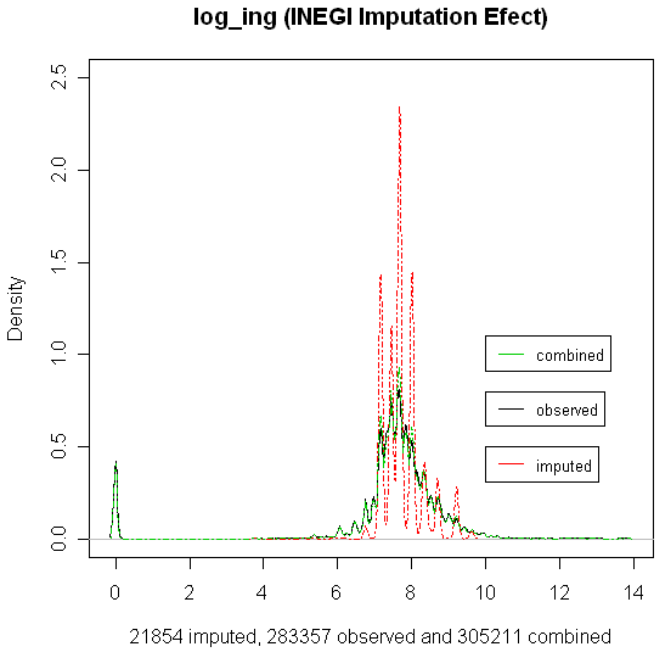
Con la finalidad de aportar elementos que enriquezcan la comparación, resulta natural previamente cuestionarse el patrón de comportamiento del ingreso en ambas técnicas. Cabe aclarar que, aunque la Imputación Múltiple ya fue revisada en la sección anterior, se requiere checar de nueva cuenta, debido al ajuste aplicado en la información a imputar.

- **Imputación del INEGI (fuente externa)**

  La estructura del ingreso para la información del INEGI es:

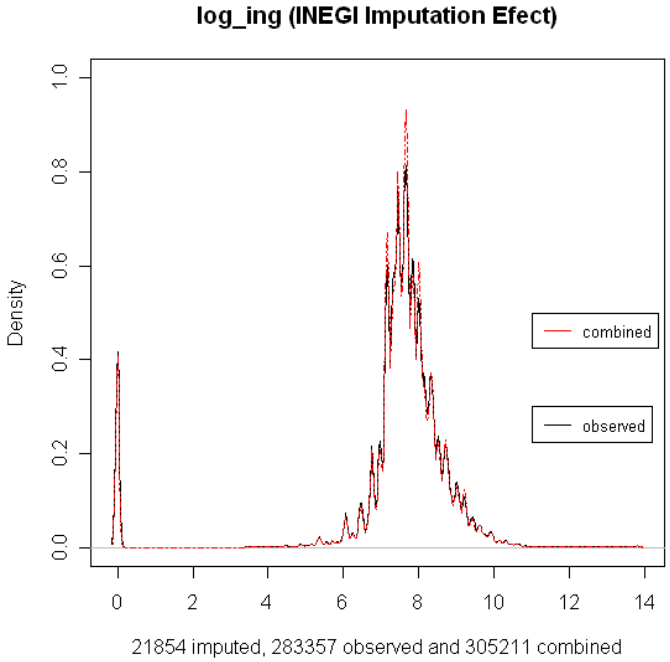
log_ing	nnumber	minimun	maximun	mean	stddev
imputed	21854	3.76	9.62	7.75	0.52
observed	283357	0.00	13.71	7.39	1.87
combined	305211	0.00	13.82	7.42	1.81

  De la estructura anterior, se tiene que las estimaciones de la media del ingreso y de su desviación estándar correspondientes para la información reportada (observed) y la combinada (combined) son muy similares; luego, la imputación del INEGI presenta un efecto numérico poco significativo sobre la información reportada; llama la atención que la información imputada presente una desviación estándar (0.52) demasiado pequeña con respecto a las restantes, gráficamente se tiene el siguiente comportamiento:



Gráfica 5.1: Distribución de la información reportada, imputada y combinada del ingreso (imputación del INEGI)

La Gráfica 5.1, confirma la diferencia mencionada entre las estimaciones de la varianza; además, muestra con claridad el comportamiento determinístico de la imputación aplicada (picos pronunciados y con mayor presencia alrededor de la media), con el fin de evaluar con mayor cuidado el efecto de este fenómeno en la información combinada, se presenta el siguiente gráfico:



Gráfica 5.2: Distribución de la información reportada y la combinada del ingreso (imputación del INEGI)

En la Gráfica 5.2, se observa que la imputación en general, tiende a respetar la distribución de la información reportada pero también que para un número discreto de puntos la diferencia entre distribuciones es notable (esto no es posible detectarlo al revisar numéricamente la estructura del ingreso).

Retomando el proceso del análisis externo, resta ahora comparar la imputación del INEGI versus la imputación generada por RSIM. En la Tabla 5.5, se presentan las estimaciones generadas.

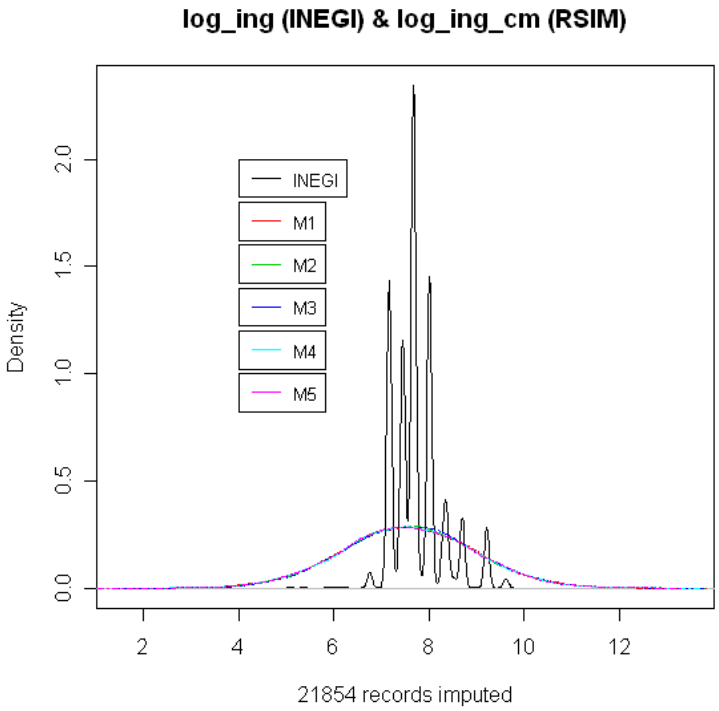
Variable (parámetro)	MÉTODO ALTERNATIVO (RSMI)								INEGI	
	Información imputada							Información reportada	Información imputada	Información reportada
	M1	M2	M3	M4	M5	Media	varianza			
LOG_ING_CM (media)	*7.55*	*7.56*	*7.57*	*7.55*	*7.55*	*7.556*	*0.00008*	*7.36*	*7.75*	*7.39*
LOG_ING_CM (varianza)	*1.96*	*1.9321*	*1.9044*	*1.9321*	*1.9321*	*1.93214*	*1.932236*	*3.6864*	*0.2704*	*3.4969*
Variable (parámetro)	MÉTODO ALTERNATIVO (RSMI)								INEGI	
	Información combinada (imputada y reportada)							Información reportada	Información combinada (imputada y reportada)	Información reportada
	M1	M2	M3	M4	M5	media	varianza			
LOG_ING_CM (media)	*7.37*	*7.37*	*7.37*	*7.37*	*7.37*	***7.37***	*0*	*7.36*	*7.42*	*7.39*
LOG_ING_CM (varianza)	*3.5721*	*3.5721*	*3.5721*	*3.5721*	*3.5721*	*3.5721*	***3.5721***	*3.6864*	*3.2761*	*3.4969*

Tabla 5.5: Estimación del ingreso (RSIM versus el INEGI)

Según datos de la Tabla 5.5, se presentan varios aspectos a destacar:

- Dada la buena repetibilidad de las cinco imputaciones simples y la poca diferencia entre la información reportada y la combinada para cada imputación, la técnica RSIM se considera como satisfactoria para el nuevo universo de la información imputada (primera parte de la tabla).
- La estimación para la media del ingreso y de su varianza respectiva de RSIM es confiable (ambas subrayadas); en particular resalta que la estimación de la varianza entre (between) imputaciones es 0.
- La estimación del INEGI para la media del ingreso, también se puede considerar numéricamente como aceptable.
- La estimación de la varianza del INEGI presenta una subestimación relativa de aproximadamente el doble con respecto a RSIM (6.3% del INEGI contra 3.1% de RSIM).

Con la idea de apoyar las conclusiones anteriores, en la Gráfica 5.3 se analiza la distribución de la información imputada para ambas opciones.



Gráfica 5.3: Distribución de la información imputada del ingreso (el INEGI vs RSIM)

En el capítulo siguiente, se presentan las conclusiones relevantes de la investigación.

Capítulo 6

Conclusiones

Es inevitable que la NO RESPUESTA no sea un tema de discusión en estos tiempos, por otra parte también es cierto que en la medida en que ésta se controle y preferentemente no se presente, los proyectos estadísticos se verán más fortalecidos.

Bajo la presencia de la NO RESPUESTA, la imputación se vuelve una alternativa interesante digna de analizar; al respecto, esta investigación pretende difundir el uso de la técnica de Imputación Múltiple en combinación con la Regresión Secuencial (RSIM).

La aplicación de RSIM al ingreso de la Población Ocupada captado durante el XII CGPyV, 2000 para el Estado de Aguascalientes, genera una serie de productos que se ponen a consideración del INEGI para su análisis, entre los cuales destacan:

- Cinco bases de datos completas; es decir, sin presencia de NO RESPUESTA excepto en los casos donde los registros formen parte de la NO RESPUESTA por unidad (para variables en estudio captadas mediante entrevista) y en aquellas variables que no fueron consideradas dentro de la investigación.
- Un procedimiento que explica el funcionamiento de la técnica de RSIM (capítulo 2).
- Una aplicación didáctica del uso del sistema IVEware (capítulo 3).
- Un procedimiento a manera de recomendación para desarrollar un proyecto de Imputación Múltiple de información a nivel censal el cual incluye la programación y desarrollos necesarios para imputar las 13 variables en estudio y que permiten el uso de herramientas informáticas: sistema IVEware (Srcware), Visual FoxPro y R (capítulos 4 y 5), sin olvidar omitir que se incluyen cifras del tiempo de cómputo necesario.

Numéricamente hablando, la imputación efectuada presenta un nivel de modificación a nivel registro que va desde un 0.2% hasta un 12.42% de la información reportada; además permite recuperar 65,816 registros (7%) los cuales fueron imputados por al menos una de las 9 variables imputadas; el 60% de estos registros se deben al ingreso y el resto a las variables predictoras. Dentro de este grupo de datos, se ubican dos situaciones especiales:

- Se rescatan 1,219 registros (0.37%) captados como Población Ocupada pero que son excluidos de dicha Población dado que su edad presenta NO RESPUESTA; así, este importante universo económico se modifica de 331,083 registros a 332,302.
- Se detectan 115 registros ocupados cuyo ingreso declarado es “999,998” el cual representa una respuesta insuficientemente no especificada (asignada al momento de mensualizar el ingreso cuando el resultado de este cálculo excedía el valor de 999,999); cabe mencionar, que estos registros son considerados por el INEGI como ingresos válidos lo cual conlleva a la creación de valores extremos erróneos y al cálculo de índices económicos incorrectos, entre otras situaciones.

En términos del número de imputaciones efectuadas, se aplican 76,482 modificaciones destacando: el ingreso con un 48.18%, la posición en el trabajo con 11.10%, la ocupación principal con 10.73%, el nivel de escolaridad con 10.27% y las cinco variables predictoras restantes en conjunto, con un 20%.

Por otra parte, es prudente resaltar que además de aportar buenas estimaciones, la técnica RSIM preserva la distribución (o estructura) de la Población Ocupada que declaró su ingreso; de hecho, esta conclusión es extensiva a todas las variables predictoras en estudio.

Con relación a la comparación con la investigación desarrollada por el INEGI, más allá de los ajustes aplicados para poder realizar el contraste y de los resultados ya emitidos, resalta el hecho de que se provee de un valor del ingreso como una alternativa al ya aportado por el Instituto.

Evidentemente, existen algunos aspectos que limitan la investigación ya sea porque no se tiene referencia de solución o bien porque resultan poco relevantes para el desarrollo de la misma; en cualquier sentido, las limitaciones como tal quedan fuera del propósito de este estudio y son postergadas para su tratamiento posterior; con fines meramente aclaratorios, los principales aspectos son:

- Definición del mecanismo que genera la NO RESPUESTA. A pesar de tener una excelente repetibilidad en la imputación del ingreso, esto no basta para concluir que la NO RESPUESTA es aleatoria (MAR) por lo que es necesario estudiar con mayor cuidado el origen de la NO RESPUESTA.
- La selección de variables predictoras. La estrategia para elegir las variables predictoras puede ser un buen motivo de discusión; consecuentemente, debe evaluarse con mayor cuidado si las variables seleccionadas fueron las más adecuadas. Al respecto, una propuesta consiste en replicar la aplicación para diversas muestras de datos cuya característica principal es que el ingreso esté especificado; con esto, se podrá tanto de manera individual como agrupada evaluar la imputación.
- Desconocer el valor observado de la NO RESPUESTA. Este aspecto limita el análisis y evaluación de la propia imputación ya que sólo permite considerar resultados agrupados; en esta situación y debido a las características de la propia información en estudio, no es posible proponer alternativas de mejora.
- La validación de la imputación. El proceso de liberación aplicado es relativamente sencillo y digno de mejorar; se deben incorporar diversas pruebas en la comparación distribucional y en la detección de registros atípicos; en este último caso, el trato va desde su respectiva exclusión, su imputación particular y su posterior incorporación a la información en estudio.

Esta investigación, necesariamente se convierte en punta de lanza para trabajos futuros que permitan dar un seguimiento a los resultados emitidos; en este sentido, el INEGI debe considerar evaluar la conveniencia de llevar a cabo las siguientes actividades:

- **A corto plazo:**

  1.- Reproducir el proceso de imputación del ingreso para diversas muestras no probabilísticas cuya característica primordial es que el ingreso debe estar especificado y con esto, fortalecer los procesos de validación y análisis.

  2.- Desarrollar la imputación del ingreso bajo otras técnicas; en particular, con las que el INEGI actualmente tiene en desarrollo, como son: Árboles Jerárquicos y Redes Neuronales.

  3.- Actualizar el valor de las variables que el INEGI no capta por entrevista sino a través de cálculos y que estén relacionadas con las variables en estudio que fueron imputadas; por ejemplo: el ingreso por hogar, el tipo de hogar, los años de estudio acumulados, entre otros.

  4.- Procesar los índices que involucren el uso del ingreso y de las variables predictoras imputadas, para así estar en posibilidad de analizar las variaciones contra las cifras publicadas.

  5.- Replicar el proceso de imputación pero incrementando el número de variables predictoras hasta lograr manejar el total de variables contenidas en la base de datos censal; esto, obviamente estaría sujeto a un análisis de costo beneficio.

- **A mediano plazo**:

  1.- Reproducir el proceso de imputación del ingreso para la Muestra Censal (encuesta probabilística aplicada al 10% de la Población censada); en este caso, se estudiaría la relación entre el diseño de muestra empleado y la propia imputación aplicada; además, se aprovecharía para comparar nuevamente (a nivel muestral) contra la opción del INEGI y dar un uso mayor al sistema IVEware al emplear otros comandos que incluye para esto fines.

  2.- Replicar el proceso de imputación para muestras probabilísticas, las cuales preferentemente presenten el esquema de selección empleado por la Muestra Censal; adicional a lo anterior, los registros seleccionados deben cumplir que su ingreso esté especificado. El desarrollo de estas acciones pretende enriquecer el proceso de imputación que arrojase la actividad anterior.

- **A largo plazo:**

  1.- Aplicar la técnica RSIM dentro de la fase de validación del proyecto censal; de esta manera, se contará con un valor alternativo al típicamente aportado por los criterios de validación (imputación determinística) que brindaría elementos para evidenciar la calidad del propio criterio.

  2.- Conducir la imputación de datos mediante el uso de la técnica RSIM en información generada por los otros niveles de captación del Censo como son hogares, viviendas, entre otros.

  3.- Reproducir el proceso de imputación del ingreso en las entidades federativas restantes (30 Estados y un Distrito Federal); así, la NO RESPUESTA podría evaluarse en un contexto nacional.

Finalmente, al margen de los resultados de la investigación y del cumplimiento de las actividades ya mencionadas, este trabajo considerará cubiertos sus objetivos en la medida en que los productos generados, sean aprovechados por el INEGI. En este tenor, debe aclararse que no se pretende cambiar el accionar actual del Instituto en la materia sino más bien adicionar elementos que colaboren hacia la mejora continua; por otra parte, queda como una propuesta final la aplicación de los productos generados durante el próximo XIII Censo General de Población y Vivienda a desarrollarse en el año 2010.

Anexo A

Mapa de México (ubicación geográfica del Estado de Aguascalientes)


United States of America
GULF
OF
MEXICO
MEXICO CITY
AGUASCALIENTES
PACIFIC OCEAN
Central
America

Anexo B

El cuestionario censal (básico)

XII CENSO DE POBLACIÓN Y VIVIENDA 2000

Cuestionario básico

INSTITUTO NACIONAL DE ESTADÍSTICA GEOGRAFÍA E INFORMÁTICA

1. IDENTIFICACIÓN GEOGRÁFICA

ENTIDAD FEDERATIVA _____ |__|__|

MUNICIPIO O DELEGACIÓN _____ |__|__|__|

CLAVE DE AGEB ..... |__|__|__|__| — |__|

LOCALIDAD _____ |__|__|__|__|

MANZANA ..... |__|__|__|

SEGMENTO ..... |__|

2. CONTROL DE VIVIENDA Y CUESTIONARIOS

CONSECUTIVO DE LA VIVIENDA ..... |__|__|__|

NÚMERO DE HOGAR ..... |__|

TOTAL DE HOGARES EN LA VIVIENDA ..... |__|

TOTAL DE CUESTIONARIOS EN LA VIVIENDA ..... |__|

3. DIRECCIÓN DE LA VIVIENDA

_____

CALLE, AVENIDA, CALLEJÓN, CARRETERA, CAMINO

_____ _____ _____

NÚMEROEXTERIOR NÚMEROINTERIOR COLONIA, FRACCIONAMIENTO, BARRIO, UNIDAD HABITACIONAL

4. CONTROL DE PAQUETE

FOLIO DE PAQUETE ... |__|__|__|__|

CONSECUTIVO DEL CUESTIONARIO EN EL PAQUETE ..... |__|__|

5. CLASE DE VIVIENDA

CIRCULE UN SOLO CÓDIGO

CASA INDEPENDIENTE	1
DEPARTAMENTO EN EDIFICIO	2
VIVIENDA O CUARTO EN VECINDAD	3
VIVIENDA O CUARTO EN LA AZOTEA	4
LOCAL NO CONSTRUIDO PARA HABITACIÓN	5
VIVIENDA MÓVIL	6
REFUGIO	7

6. NOMBRE DE LOS RESPONSABLES

ENTREVISTADOR(A)

_____ |__|

JEFE (A) DE ENTREVISTADORES

_____ |__|

RESPONSABLE DE AGEB

_____ |__|

VALIDADOR(A)

_____ |__|

7. RESULTADO DE LA VALIDACIÓN

VALIDADO	1		
A VERIFICACIÓN POR ERROR EN:			
IDENTIFICACIÓN GEOGRÁFICA	2	GASTO COMÚN, NÚMERO DE HOGARES / CONTROL DE VIVIENDA.	5
CONTROL DE VIVIENDA Y CUESTIONARIOS	3	LISTA DE PERSONAS / CARACTERÍSTICAS DE LAS PERSONAS ...	6
NÚMERO DE PERSONAS / LISTA DE PERSONAS ..	4	SEXO, EDAD / NÚMERO DE HIJOS	7

INEGI. Información a la población: llame sin costo al (01) 800 491 0100. En Aguascalientes, 9 10 53 63.

I. Características de la vivienda

1. PAREDES

¿De qué material es la mayor parte de las paredes o muros de esta vivienda?

CIRCULE UN SOLO CÓDIGO

Material de desecho	1
Lámina de cartón	2
Lámina de asbesto o metálica	3
Carrizo, bambú o palma	4
Embarro o bajareque	5
Madera	6
Adobe	7
Tabique, ladrillo, block, piedra, cantera, cemento o concreto	8

2. TECHOS

¿De qué material es la mayor parte del techo de esta vivienda?

CIRCULE UN SOLO CÓDIGO

Material de desecho	1
Lámina de cartón	2
Lámina de asbesto o metálica	3
Palma, tejamanil o madera	4
Teja	5
Losa de concreto, tabique, ladrillo o terrado con viguería	6

3. PISOS

¿De qué material es la mayor parte del piso de esta vivienda?

CIRCULE UN SOLO CÓDIGO

Tierra	1
Cemento o firme	2
Madera, mosaico u otros recubrimientos	3

4. COCINA

¿Esta vivienda tiene un cuarto para cocinar?

CIRCULE UN SOLO CÓDIGO

Sí 1

No ..... 2 → PASE A 5

En el cuarto donde cocinan, ¿también duermen?

CIRCULE UN SOLO CÓDIGO

Sí 3

No 4

5. NÚMERO DE CUARTOS

¿Cuántos cuartos se usan para dormir sin contar pasillos?

|_|_| *ANOTE CON NÚMERO*

Sin contar pasillos ni baños, ¿cuántos cuartos tiene en total esta vivienda? Cuente la cocina.

|_|_| *ANOTE CON NÚMERO*

6. DISPONIBILIDAD DE AGUA

¿En esta vivienda tienen:

LEA LAS OPCIONES HASTA OBTENER UNA RESPUESTA AFIRMATIVA Y CIRCULE UN SOLO CÓDIGO

agua entubada dentro de la vivienda?	1
agua entubada fuera de la vivienda, pero dentro del terreno?	2
agua entubada de llave pública (o hidrante)?	3
agua entubada que acarrean de otra vivienda?	4
agua de pipa?	5
agua de un pozo, río, lago, arroyo u otra?	6

7. SERVICIO SANITARIO

¿Esta vivienda tiene:

excusado o sanitario?
retrete o fosa?
letrina?
hoyo negro o pozo ciego?

CIRCULE UN SOLO CÓDIGO

Sí 1

No ..... 2 → PASE A 10

8. USO EXCLUSIVO

¿Este servicio lo usan solamente las personas de esta vivienda?

CIRCULE UN SOLO CÓDIGO

Sí 1

No 2

9. CONEXIÓN DE AGUA

¿Este servicio sanitario:

LEA LAS OPCIONES HASTA OBTENER UNA RESPUESTA AFIRMATIVA Y CIRCULE UN SOLO CÓDIGO

tiene conexión de agua?	1
le echan agua con cubeta?	2
¿No se le puede echar agua?	3

Continúe con la pregunta 10 ➡

10. DRENAJE

¿Esta vivienda tiene drenaje o desagüe de aguas sucias:

LEA LAS OPCIONES HASTA OBTENER UNA RESPUESTA AFIRMATIVA Y CIRCULE UN SOLO CÓDIGO

a la red pública?	1
a una fosa séptica?	2
a una tubería que va a dar a una barranca o grieta?	3
a una tubería que va a dar a un río, lago o mar?	4
¿No tiene drenaje?	5

11. ELECTRICIDAD

¿Hay luz eléctrica en esta vivienda?

CIRCULE UN SOLO CÓDIGO

Sí	1
No	2

12. COMBUSTIBLE

¿El combustible que más usan para cocinar es:

LEA LAS OPCIONES HASTA OBTENER UNA RESPUESTA AFIRMATIVA Y CIRCULE UN SOLO CÓDIGO

gas?	1
leña?	2
carbón?	3
petróleo?	4
electricidad?	5

13. TENENCIA

¿Esta vivienda es propiedad de alguna persona que vive aquí?

CIRCULE UN SOLO CÓDIGO

Sí ..... 1 → *PREGUNTE*

LEA LAS OPCIONES HASTA OBTENER UNA RESPUESTA AFIRMATIVA Y CIRCULE UN SOLO CÓDIGO

¿Está pagándose?	3
¿Está totalmente pagada?	4
¿Está en otra situación?	5

No ..... 2 → *PREGUNTE*

LEA LAS OPCIONES HASTA OBTENER UNA RESPUESTA AFIRMATIVA Y CIRCULE UN SOLO CÓDIGO

¿Está rentada?	6
¿Está prestada, la cuidan o en otra situación?	7

14. BIENES EN LA VIVIENDA

¿En esta vivienda tienen:

LEA TODAS LAS OPCIONES Y CIRCULE EL CÓDIGO SEGÚN LA RESPUESTA

	Sí	No
radio o radiograbadora?	1	2
televisión?	3	4
videocasetera?	5	6
licuadora?	7	8
refrigerador?	1	2
lavadora?	3	4
teléfono?	5	6
calentador de agua (boiler)?	7	8
automóvil o camioneta propios?	1	2
computadora?	3	4

Continúe con la siguiente sección ➡

II. Residentes, hogares y lista de personas

1. NÚMERO DE PERSONAS	2. GASTO COMÚN	3. NÚMERO DE HOGARES
¿Cuántas personas viven normalmente en esta vivienda contando a los niños chiquitos y a los ancianos (cuente también a los sirvientes que duermen aquí)**?** _____ *ANOTE CON NÚMERO*	**¿Todas las personas que viven en esta vivienda comparten un mismo gasto para la comida?** *CIRCULE UN SOLO CÓDIGO* Sí ..... **1** *PASE 4 LISTA DE PERSONAS* No **2**	**Entonces ¿Cuántos hogares o grupos de personas tienen gasto separado para la comida, contando el de usted?** _____ *ANOTE CON NÚMERO*

CUANDO EN LA VIVIENDA EXISTA MÁS DE UN HOGAR O GRUPO DE PERSONAS, APLIQUE UN CUESTIONARIO PARA CADA HOGAR A PARTIR DE LA LISTA DE PERSONAS

4. LISTA DE PERSONAS EN EL HOGAR
Por favor, dígame el nombre de las personas que viven en su hogar, empezando por el jefe o la jefa; déme también el nombre de los niños chiquitos y los ancianos (incluya a los sirvientes que duermen aquí):
PERSONA **1** _____ *ANOTE EL NOMBRE DEL JEFE(A)*
PERSONA **2**
PERSONA **3**
PERSONA **4**
PERSONA **5**
PERSONA **6**

SI EN EL HOGAR Y MÁS DE 6 PERSONAS, UTILICE OTRO CUESTIONARIO Y CONTINÚE CON LA LISTA

Copie el nombre de todas las personas en los espacios destinados para ello en la Sección III y haga las preguntas usando el nombre de cada una de las personas.

III. Características de las personas

Ahora le voy a preguntar por (NOMBRE): **PERSONA 1** _____ *Anote el nombre de la persona*

1. PARENTESCO

¿Qué es (NOMBRE) **del jefe(a) del hogar?**

SI ES EL JEFE(A) SÓLO CONFIRME Y CIRCULE UN SOLO CÓDIGO

Jefe(a) 1

Esposo(a) o compañero(a) 2

Hijo(a) 3

Otro _____ *ANOTE EL PARENTESCO*

2. SEXO

(NOMBRE) **es mujer**

(NOMBRE) **es hombre**

CIRCULE UN SOLO CÓDIGO

Hombre 1

Mujer 2

3. EDAD

¿Cuántos años cumplidos tiene (NOMBRE)?

MENOR DE UN AÑO, ANOTE "000"

|_|_|_| *ANOTE CON NÚMERO*

4. LUGAR DE NACIMIENTO

¿En qué estado de la República o en qué país nació (NOMBRE)?

Aquí, en este estado 1

En otro estado _____ *ANOTE EL ESTADO*

En otro país _____ *ANOTE EL PAÍS*

5. DERECHOHABIENCIA

¿Tiene (NOMBRE) **derecho a servicio médico en:**

LEA TODAS LAS OPCIONES Y CIRCULE LAS RESPUESTAS AFIRMATIVAS

el Seguro Social (IMSS)? 1

el ISSSTE? 2

Pemex, Defensa o Marina? 3

otra institución?

_____ *ANOTE LA INSTITUCIÓN*

Entonces, no tiene derecho a servicio médico 5

6. TIPO DE DISCAPACIDAD

¿(NOMBRE) **tiene limitación para:**

LEA TODAS LAS OPCIONES Y CIRCULE LAS RESPUESTAS AFIRMATIVAS

moverse, caminar o lo hace con ayuda? 1

usar sus brazos y manos? 2

¿Es sordo(a) o usa un aparato para oír? 3

¿Es mudo(a)? 4

¿Es ciego(a) o sólo ve sombras? 5

¿Tiene algún retraso o deficiencia mental? 6

¿Tiene otra limitación física o mental?

_____ *ANOTE LA LIMITACIÓN*

Entonces, no tiene limitación física o mental 8

PARA PERSONAS DE 5 AÑOS CUMPLIDOS O MÁS

7. ENTIDAD O PAÍS DE RESIDENCIA EN 1995

Hace 5 años, en enero de 1995, ¿en qué estado de la República o en qué país vivía (NOMBRE)?

Aquí, en este estado 1

En otro estado _____ *ANOTE EL ESTADO*

En otro país _____ *ANOTE EL PAÍS*

PASE A 9

8. MUNICIPIO DE RESIDENCIA EN 1995

¿En qué municipio (delegación) vivía (NOMBRE) **en enero de 1995?**

Aquí, en este municipio o delegación 2

En otro municipio o delegación _____ *ANOTE EL MUNICIPIO O DELEGACIÓN*

Continúe con la pregunta 9

PARA PERSONAS DE 5 AÑOS CUMPLIDOS O MÁS **PERSONA 1**

9. LENGUA INDÍGENA

¿(NOMBRE) **habla algún dialecto o lengua indígena?**

CIRCULE UN SOLO CÓDIGO

Sí 1

No ..... 2 → *PASE A 10*

¿**Qué dialecto o lengua indígena habla** (NOMBRE)?

_____

ANOTE LA LENGUA INDÍGENA

¿(NOMBRE) **habla también español?**

CIRCULE UN SOLO CÓDIGO

Sí 3

No 4

10. ALFABETISMO

¿(NOMBRE) **sabe leer y escribir un recado?**

CIRCULE UN SOLO CÓDIGO

Sí 1

No 2

11. ASISTENCIA

¿(NOMBRE) **actualmente va a la escuela?**

CIRCULE UN SOLO CÓDIGO

Sí 1

No 2

12. ESCOLARIDAD

¿**Hasta qué año o grado aprobó (pasó)** (NOMBRE) **en la escuela?**

ANOTE CON NÚMERO EL ÚLTIMO GRADO Y CIRCULE EL CÓDIGO DE NIVEL

	Grado	*Nivel*	
Ninguno *(anote "0")*	___	0	
Preescolar o kinder	___	1	
Primaria	___	2	→ *PASE A 15*
Secundaria	___	3	
Preparatoria o bachillerato	___	4	
Normal	___	5	
Carrera técnica o comercial	___	6	→ *PASE A 13*
Profesional	___	7	
Maestría o doctorado	___	8	→ *PASE A 14*

13. ANTECEDENTE ESCOLAR

¿**Para entrar a la carrera** (normal, técnica, comercial o profesional) **qué estudios le pidieron como requisito?**

CIRCULE UN SOLO CÓDIGO

Primaria terminada 1

Secundaria terminada 2

Preparatoria terminada 3

14. NOMBRE DE LA CARRERA

¿**Cuál es el nombre de la carrera** (normal, técnica, comercial, profesional, maestría o doctorado)?

_____

_____

_____

_____

ANOTE LA CARRERA

15. RELIGIÓN

¿**Cuál es la religión de** (NOMBRE)?

CIRCULE UN SOLO CÓDIGO

Ninguna 1

Católica 2

Otra religión

_____

_____

_____

ANOTE LA RELIGIÓN

Continúe con la pregunta 16 ➡

PERSONA 1 — **PARA PERSONAS DE 12 AÑOS CUMPLIDOS O MÁS**

16. ESTADO CONYUGAL

¿Actualmente (NOMBRE):

LEA LAS OPCIONES HASTA OBTENER UNA RESPUESTA AFIRMATIVA Y CIRCULE UN SOLO CÓDIGO

- **vive con su pareja en unión libre?** .. 1
- **está separado(a)?** 2
- **está divorciado(a)?** 3
- **es viudo(a)?** 4
- **está casado(a)?**
  - **¿Sólo por el civil?** 5
  - **¿Sólo religiosamente?** 6
  - **¿Civil y religiosamente?** 7
- **está soltero(a)?** 8

17. CONDICIÓN DE ACTIVIDAD

¿La semana pasada (NOMBRE):

LEA LAS OPCIONES HASTA OBTENER UNA RESPUESTA AFIRMATIVA Y CIRCULE UN SOLO CÓDIGO

- **trabajó?** ..... 1 → PASE A 19
- **tenía trabajo, pero no trabajó?** ..... 2 → PASE A 19
- **buscó trabajo?** 3
- **¿Es estudiante?** 4
- **¿Se dedica a los quehaceres de su hogar?** ... 5
- **¿Es jubilado(a) o pensionado(a)?** 6
- **¿Está incapacitado(a) permanentemente para trabajar?** ..... 7 → PASE A 24
- **¿No trabaja?** 8

18. VERIFICACIÓN DE ACTIVIDAD

Además de (RESPUESTA DE 17), **¿la semana pasada** (NOMBRE):

LEA LAS OPCIONES HASTA OBTENER UNA RESPUESTA AFIRMATIVA Y CIRCULE UN SOLO CÓDIGO

- **ayudó en un negocio familiar?** 1
- **vendió algún producto?** 2
- **hizo algún producto para vender?** 3
- **ayudó trabajando en el campo o en la cría de animales?** 4
- **a cambio de un pago realizó otro tipo de actividad?** Por ejemplo: lavó o planchó ajeno, cuidó coches 5
- **¿No trabaja?** ..... 6 → PASE A 24

19. OCUPACIÓN U OFICIO

¿Qué hizo (NOMBRE) **en su trabajo de la semana pasada?**

_____

_____

ANOTE LAS ACTIVIDADES O TAREAS

¿Cuál es el nombre de su ocupación, oficio o puesto?
Por ejemplo: campesino(a), maestro(a) de primaria, vendedor(a) ambulante.

_____

_____

ANOTE LA OCUPACIÓN, OFICIO O PUESTO

20. SITUACIÓN EN EL TRABAJO

¿(NOMBRE) **en su trabajo de la semana pasada fue:**

LEA LAS OPCIONES HASTA OBTENER UNA RESPUESTA AFIRMATIVA Y CIRCULE UN SOLO CÓDIGO

- **empleado(a) u obrero(a)?** 1
- **jornalero(a) o peón?** 2
- **patrón(a)?** (contrata trabajadores) 3
- **trabajador(a) por su cuenta?** 4
- **trabajador(a) sin pago en el negocio o predio familiar?** 5

Continúe con la pregunta 21 ➡

PERSONA 1

21. HORAS TRABAJADAS

En total, ¿cuántas horas trabajó (NOMBRE) **la semana pasada?**

|_|_|_| *ANOTE CON NÚMERO*

22. INGRESOS POR TRABAJO

En total, ¿cuánto gana o recibe (NOMBRE) **por su trabajo?**

ANOTE EN PESOS LA CANTIDAD RECIBIDA Y EL PERIODO

NO RECIBE INGRESOS, *ANOTE* **"0"** *EN PERIODO*

Periodo: *A la semana* 1
A la quincena 2
Al mes 3
Al año 4

$|_|_|_|,|_|_|_| *ANOTE CON NÚMERO* |_| *PERIODO*

23. ACTIVIDAD ECONÓMICA

¿En dónde trabajó (NOMBRE) **la semana pasada?**
Por ejemplo: en el campo, en una fábrica, en un taller mecánico.

ANOTE EN DÓNDE TRABAJÓ

El negocio, empresa o lugar donde trabajó ¿a qué se dedica?
Por ejemplo: a cultivar maíz, a hacer muebles, a vender ropa.

ANOTE A QUÉ SE DEDICA

PARA MUJERES DE 12 AÑOS CUMPLIDOS O MÁS

24. NÚMERO DE HIJOS

En total, ¿cuántas hijas e hijos que nacieron vivos ha tenido (NOMBRE)?

NINGUNO, *ANOTE* **"00"** *Y PASE A LA SIGUIENTE PERSONA*

|_|_| *ANOTE CON NÚMERO*

25. HIJOS FALLECIDOS

De las hijas e hijos que nacieron vivos ¿cuántos han muerto?

NINGUNO, *ANOTE* **"00"**

|_|_| *ANOTE CON NÚMERO*

26. HIJOS SOBREVIVIENTES

¿Cuántas de las hijas e hijos de (NOMBRE) **viven actualmente?**

NINGUNO, *ANOTE* **"00"**

|_|_| *ANOTE CON NÚMERO*

27. FECHA DE NACIMIENTO

¿En qué mes y año nació la última hija o hijo nacido vivo de (NOMBRE)?

ANOTE EL MES Y EL AÑO

Mes ..... |_|_|

y

Año |_|_|_|_|

28. SOBREVIVENCIA

Esta última hija o hijo de (NOMBRE) **¿vive actualmente?**

CIRCULE UN SOLO CÓDIGO

Sí ... **1** *PASE A LA SIGUIENTE PERSONA*

No .. **2**

29. EDAD AL MORIR

¿Qué edad tenía cuando murió?

ANOTE SÓLO UNA RESPUESTA EN: DÍAS O MESES O AÑOS

SI VIVIÓ MENOS DE UN DÍA *ANOTE* **"00"** *EN DÍAS*

Días .... |_|_|

o

Meses . |_|_|

o

Años ..... |_|_|

Pase a la persona 2 ➡

OBSERVACIONES

CONFIDENCIALIDAD

Conforme a las disposiciones del **Artículo 38 de la Ley de Información Estadística y Geográfica en vigor,** "Los datos e informes que los particulares proporcionen para fines estadísticos o provengan de registros administrativos o civiles, serán manejados, para efectos de esta Ley, bajo la observancia de los principios de confidencialidad y reserva y no podrán comunicarse, en ningún caso, en forma nominativa o individualizada, ni harán prueba ante autoridad administrativa o fiscal, ni en juicio o fuera de él."

OBLIGATORIEDAD

De acuerdo con el **Artículo 42, párrafo primero, de la Ley de Información Estadística y Geográfica en vigor,** "Los informantes estarán obligados a proporcionar con veracidad y oportunidad los datos e informes que les soliciten las autoridades competentes para fines estadísticos, censales y geográficos, y a prestar el auxilio y cooperación que requieran las mismas."

Anexo C

Descripción de la tabla de Población (base de datos censal de explotación)

Descripción		Núm. Preg.	Mnemónico	Rangos válidos	Long.
TABLA DE POBLACIÓN					
Identifica al jefe del hogar o a la relación de parentesco que tiene cada integrante del hogar en relación al jefe.		1B	PARENTES	{100,200,300,401-412,420,430, 440,501-503, 601-624,b}	3
Identifica si la persona es hombre o mujer.		2	SEXO	{1,2}	1
Identifica los años cumplidos que la persona tiene al momento del Censo.		3	EDAD	{000..130,999}	3
Identifica la entidad federativa o país de nacimiento de la persona.		4B	LUGNACR	{001-032,100-535,600,999}	3
Identifica si la personas es derechohabiente al seguro social ó si de la persona no se sabe si es derechohabiente o no.		5A	IMSS	{1,9,b}	1
Identifica si la personas es derechohabiente al ISSSTE.		5B	ISSSTE	{2,b}	1
Identifica si la persona es derechohabiente a PEMEX, Defensa o Marina.		5C	PEMEX	{3,b}	1
Identifica si la persona tiene derechohabiencia a otra institución diferente de IMSS, ISSSTE, PEMEX, Ejército o Marina.			OTRA_INS	{4,b}	1
Identifica si la persona es no derechohabiente a servicio médico.		5E	NOTIENE	{5,b}	1
Identifica si la persona tiene una discapacidad que le impide moverse o caminar, también si de la persona no se sabe si tiene alguna discapacidad.		6A	MOTRIZ	{1,9,b}	1
Identifica si la persona tiene una discapacidad que le limita usar sus brazos y manos.		6B	BRAZOS	{2,b}	1
Identifica si la persona tiene una discapacidad auditiva.		6C	AUDITIVA	{3,b}	1
Identifica si la persona tiene una discapacidad de lenguaje.		6D	LENGUAJE	{4,b}	1
Identifica si la persona tiene una discapacidad de tipo visual.		6E	VISUAL	{5,b}	1
Identifica si la persona tiene algún retraso o deficiencia mental.		6F	MENTAL	{6,b}	1
Identifica si la persona tiene alguna discapacidad que por su descripción no fue factible reasignarlas a algún tipo de las pre codificadas.		6G	DISCAP	{131, 199, 299, 399, 421, 422, 430, 499, 960,b}	3
Identifica si la persona tiene dos limitaciones y una no se pudo clasificar.			OT_DISC	{7,b}	1
Identifica si la persona no tiene ninguna discapacidad.		6H	NO_DISC	{8,b}	1
Identifica la entidad federativa o país de residencia de la persona en enero de 1995.		7B	LUGRES95	{001-032,100-535,600,999,b}	3
Identifica el municipio de residencia de la persona en enero de 1995. Existe relación con lo que contenga la entidad de residencia en enero de 1995.		8B	MUN_1995	{001-570,999,b}	3
Identifica si la persona habla algún dialecto o lengua indígena.		9A	LEN_INDI	{1,2,9,b}	1
Identifica el dialecto o lengua indígena que la persona habla.		9B	TIPO_LEN	{0111-1311,5010,5020,9999,b}	4

Descripción		Núm. Preg.	Mnemónico	Rangos válidos	Long.
Identifica si la persona habla lengua indígena y si también sabe hablar español.		9C	HAB_ESP	{3,4,9,b}	1
Identifica si la persona sabe leer y escribir un recado.		10	ALFAB	{1,2,9,b}	1
Identifica si la persona asiste actualmente a la escuela.		11	ASISTEN	{1,2,9,b}	1
Identifica el último grado aprobado por la persona. Está relacionado con el nivel académico.		12A	GDO_APR	{0..9,b}	1
Identifica el último nivel académico alcanzado por la persona.		12B	NIV_ESC	{0..8,9,b}	1
Identifica el tipo de antecedente escolar necesario para cursar la carrera técnica o profesional de la persona.		13	ANT_ESCO	{1..3,9,b}	1
Identifica el nivel académico de la persona, así como el tipo de antecedente escolar necesario para cursar su carrera. (Fusión de nivel académico y antecedente escolar)			NIV_ACA	{00,10,20,30,40,51,52,61,62,63,73,80,90,b}	2
Número de años que ha aprobado la persona desde que entro a la escuela y hasta el momento del Censo.			ESC_ACU	{00..24,99,b}	2
Identifica la clave de la carrera técnica o comercial, profesional, maestría o doctorado que tiene la persona.		14	CARRERA	{0011-2993,3111-4990,5110-6990, 7110,9999,b}	4
Identifica la clave de la religión que profesa la persona.		15B	RELIGION	{0001,1101-9001,9100,9999,b}	4
Identifica el estado conyugal o civil en el que se encuentra la persona.		16	EDO_CONY	{1..8,9,b}	1
Identifica la condición de actividad económica que tiene la persona, la semana anterior a la fecha del Censo.		17	ACTI_INA	{10,13…16,18…20,30,40,50,60,70,80,99,b}	2
Identifica la clave de ocupación u oficio que tiene la persona, en la semana anterior al Censo.		19	OCUPAC	{1100-8390,9999,b}	4
Identifica la situación laboral que la persona tiene en relación a su lugar de trabajo, en la semana anterior al Censo.		20	POS_TRA	{1..5,9,b}	1
Identifica el número de horas trabajadas la semana anterior al Censo.		21	HRAS_TRA	{000..999,b}	3
Identifica la cantidad en pesos que la persona percibe mensualmente por su trabajo.		22A	TOT_ING	{000000..999999,b}	6
Identifica la actividad económica del lugar donde trabajó la persona en la semana anterior al Censo.		23	RAMA	{110-939,999,b}	3
Identifica el número de hijos nacidos de la mujer de 12 años o más.		24	VIVOS	{00..25,98,99,b}	2
Identifica el número de hijos nacidos vivos de la mujer de 12 años o más que al momento del Censo habían fallecido.		25	FALLECID	{00..25,99,b}	2
Identifica el número de hijos nacidos vivos de la mujer de 12 años o más que al momento del Censo estaban con vida.		26	SOBREVI	{00..25,99,b}	2
Mes de nacimiento del último hijo nacido vivo de la mujer de 12 años o más de edad.		27A	MES_NAC	{01..12,99,b}	2
Año de nacimiento del último hijo nacido vivo de la mujer de 12 años o más de edad.		27B	ANIO_N	{1930..2000,9999,b}	4

Descripción		Núm. Preg.	Mnemónico	Rangos válidos	Long.
Identifica la condición de sobrevivencia del último hijo nacido vivo de la mujer de 12 años o más de edad, al momento del Censo.		28	ULTIMO_H	{1,2,9,b}	1
Identifica la edad en días que tenía el último hijo nacido vivo al momento de morir.		29A	DIA_EDAD	{00..29,98,99,b}	2
Identifica la edad en meses que tenía el último hijo nacido vivo al momento de morir.		29B	MES_EDAD	{01..11,98,b}	2
Identifica la edad en años cumplidos que tenía el último hijo nacido vivo al momento de morir.		29C	ANI_EDAD	{01..98,b}	2

Anexo D

El proceso de imputación del INEGI

Census 2000 Aguascalientes´s Employment Population

STEP 1.

Set filter to:
Edad=(12,130) and
Sexo=1,2 and
Pos_tra=1,2,3,4,5 and
Parentes<>'999' and
Niv_esc<> '9' and
Ocupac<>'9999'

STEP 2.

Forming groups: (separated by commas and color)

Parentes: 100, 200, 300, 401-440, 501-503, 601-624, 701-703 (7 groups)

Edad: 12-17, 18-24, 25-54, 55-130 (4 groups)

Sexo: 1, 2 (2 groups)

Pos_tra: 1, 2, 3, 4, 5 (5 groups)

Niv_esc: 0-1, 2, 3 or (Niv_esc=6 and Ant_esco=1) , 4 or (Niv_esc=6 and Ant_esco=2) , 5, 7-8 or (Niv_esc=6 and Ant_esco=3) (6 groups)

Ocupac: 4100-4190, 1100-1190, 2100-2190, 5100-5190, 5200-5290 or 5300-5390 or 5400-5491, 8200-8209, 8300-8390, 1200-1290 or 1300-1390 or 1400-1490' or 5500-5590 or 6100-6190 or 6200-6290 or 7100-7190 or 7200-7290 or 8100-8190 (8 groups)

Then we get 13440 (as 7*4*2*5*6*8) possible different groups (denoted by Image)
Note: "- is the same as to"
For each group or image, we have total of resident which income was not specified (RNE)
Sum RNE for all images in the variable TOTAL
Order all groups by RNE (minimum to maximum)

STEP 3.

To ignore Images when: Accumulated(**RNE**) < = 0.30***TOTAL**
Inegi only considered imputing the rest 70%

Example: (suppose 5 groups or images)

IMAGE	RNE	Accumulated of RNE
12892	2	2
22893	3	5
23598	3	8
55532	6	14
00054	6	20

TOTAL=20
Imputing images are: 23598, 55532 and 00054 since red number (Accumulated of RNE) are minus to 20*0.30=6

STEP 4.

Imputing Income (images obtained in step 3)
For each group calculate **income mode.**
For all residents into groups, which have not specify income assign income mode

Anexo E

Los comandos de IVEware (proceso de imputación)

Comando GETDATA

El comando GETDATA importa el archivo por imputar o analizar al ambiente de IVEware, su sintaxis debe contener:

- Instrucciones para manejar datos, metadatos o bien instrucciones para trabajar una tabla y
- La instrucción RUN.

En general, la sintaxis de ejecución tiene el siguiente formato:

% GETDATA (name=, dir =, setup=)
Instrucciones de programación

Instrucciones de programación para el comando **GETDATA**

- DATAIN *archivo*: identifica la ubicación y el nombre del archivo de datos a imputar o analizar.
- DELIMITER *carácter(es)*: se usa la opción "CSV" para indicar que las variables están separadas por una coma u otro(s) carácter(es) para delimitar la longitud de las variables. Si los datos incluyen un delimitador, este debe encerrarse dentro de comillas. "/t" indica como delimitador al tabulador.
- METADATA *archivo*: indica el formato del archivo de metadatos a ser leído. Si el archivo es omitido, las opciones para emplear la instrucción METADATA debe escribirse inmediatamente (las opciones se explican abajo).
- END: indica que terminan las instrucciones de METADATA.
- NOBS *número*: representa el número de observaciones a ser incluidas en el archivo de salida. Si esta instrucción es omitida, se incluirán todas las observaciones del archivo a imputar.
- PRINT *none/standard/details/all*: se define las características de la impresión:
  - *None*: la opciones estándar.
  - *standard*: genera un resumen acerca de la base de datos.
  - *details* y *all*: produce un resumen de la base de datos e información detallada acerca de las variables y de la codificación respectiva.
    La opción por default es *standard.*
- SUBSET *lista de variables*: define las variables a incluir en el archivo de salida. Si esta instrucción se omite, se incluirán todas las variables definidas dentro de METADATA.
- TABLE *archivo(s)*: indica el formato de la tabla de datos a ser leída. Los datos debe estar en una hoja de cálculo en forma de tabla, el primer renglón debe contener los nombres de las variables; los nombres de tipo carácter deben estar precedidos por el símbolo "$". Las celdas deben estar delimitadas por blancos, comas o caracteres tipo tabulador, aquellas cuyos valores incluyan blancos, comas o caracteres tipo tabulador, deben encerrarse dentro de comillas sencillas o dobles.
- TITLE *texto*: define el texto a ser empleado como título.
- RUN: instrucción requerida, debe ser la última instrucción de este comando

Instrucciones de programación para la opción **METADATA** (puede estar dentro la sintaxis de GETDATA o bien en un archivo externo).

- DELIMITER *carácter(es)*: se usa la opción "CSV" para indicar que las variables están separadas por una coma u otro(s) carácter(es) para delimitar la

longitud de las variables. Si los datos incluyen un delimitador, este debe encerrarse dentro de comillas. "/t" indica como delimitador al tabulador.

- ID *variable*: define el nombre del campo del identificador. _OBS_ es el default para el número de observación.
- MULTIPLE *variable*: para una Imputación Múltiple esta variable contiene el nombre del campo donde se indica el número de imputación al que pertenece el registro, el default es _MULT_.
- RECORDLENGTH *número*: indica la longitud del registro para el caso binario
- STANDAR: especifica que el archivo de datos/metadatos están en el formato estándar de IVEware (producidos por GETDATA).
- VARIABLE(S) *descripción*: incluye la descripción de variables (cuyos parámetros se definen abajo. La palabra "VARIABLE" puede omitirse después de escribirse una vez.
- CODEFRAME(S) *descripción*: presenta la descripción del CODEFRAME (cuyos parámetros se definen abajo). La palabra "CODEFRAME" puede omitirse después de escribirse una vez.

Instrucciones de programación para la opción **VARIABLE**

- NAME *nombre de variable*: indica el nombre de la variable. Requerido.
- LABEL *texto*: define el nivel de la variable, el valor por default es blanco.
- TYPE *character/numeric/integer/floating*: representa el tipo de la variable: tipo ASCII, numérico, entero binario, o binario punto flotante. Los tipos: entero binario y punto flotante no pueden usarse con datos delimitados.
- CODEFRAME *nombre del catálogo*: específica el uso de un catalogo de códigos para la variable.
- LOCATION *número*: indica la localización inicial para la variable. Se puede omitir para datos delimitados. Para datos no delimitados, la localización por default es 1 para la primera variable; para el resto, se incrementa la localización de acuerdo con el ancho de la variable anterior.
- WIDTH *número*: denota el ancho de la variable. Se puede omitir para datos delimitados. Para datos no delimitados, los anchos por default son 1 para la primera variable; para el resto, se incrementa de acuerdo con el ancho de la variable anterior.
- DECIMALS *número*: define el número de posiciones decimales para la variable. El default es 0 para variables carácter o no consecutivas y para las variables distintas a carácter que estén en serie el número indicado.
- MISSING *valor*: especifica el valor del código de la NO RESPUESTA. puede ser "." o cualquier otro carácter (o grupo de caracteres).

Instrucciones de programación para la opción **CODEFRAME**

- NAME *nombre del catálogo*: define el nombre del catálogo de códigos. Requerido.
- LABEL *texto*: especifica el nivel del catálogo, el default es blanco.
- VALUE *texto*: indica el valor del código y de su nivel, al menos se requiere tener un par de valores/niveles.

Módulo IMPUTE

El modulo IMPUTE es un procedimiento general de Imputación Multivariada que puede manejar estructuras complejas (las variables a imputar presentan diversas formas

distribucionales, típicamente algunas son continuas, otras discretas, muchas son dicotómicas, politómicas o semicontinuas) siempre que estas presenten NO RESPUESTA aleatoria. Este módulo produce valores imputados para cada lectura individual del conjunto de datos condicionado sobre todos los valores reportados de dicha lectura, la estrategia básica es crear imputaciones a través de una secuencia de regresiones múltiples; el tipo de modelo de regresión varía de acuerdo al tipo de variable a imputar. La secuencia de valores imputados puede continuar de manera cíclica, cada vez sobrescribiendo el último valor asignado, construyendo independencia entre los valores imputados y explotando la estructura de correlación entre variables. IMPUTE supone que las variables son de los siguientes tipos: continuas, binarias, categóricas (politómicas, es decir, con más de dos categorías), discretas, o mixtas; los tipos de modelos de regresión empleados son lineal, logístico, Poisson, logit generalizado o mixto logístico-lineal, dependiendo del tipo de variable a imputar. Además, IMPUTE puede aceptar dos tipos de características comunes en los datos que le agregan complejidad al modelo: la restricción de imputación a subpoblaciones y la acotación de los valores imputados.
En general, la sintaxis de ejecución tiene el siguiente formato:

%IMPUTE (name=, dir =, setup=)
Instrucciones de programación

Instrucciones de programación para el modulo IMPUTE.
Requeridas:

- DATAIN *archivo*: identifica la ubicación y el nombre del archivo de datos a imputar.
- DATAOUT *archivo/ALL*: identifica la ubicación y el nombre del archivo de salida que contiene a los datos imputados. ALL especifica que se incluirán las *m* imputaciones en el mismo archivo.
- *Declaración de los tipos de variables: si no se específica tipo alguno se asumirá que todas las variables son continuas. El tipo de variable debe registrarse antes de las instrucciones: BOUNDS, INTERACT o RESTRICT.*
  - CONTINUOS *lista de variables*: se emplea un modelo lineal normal, quizás se requiera transformar los datos para lograr normalidad e imputar sobre la escala transformada, en este caso después de la imputación se debe aplicar la transformación inversa para tener la variable en su escala original.
  - CATEGORICAL *lista de variables*: se usa un modelo logístico o logístico generalizado.
  - MIXED *lista de variables*: se emplea un modelo de dos etapas, primero se usa un modelo logístico para imputar el estado 0 o no 0, y para valores no cero, se usa un modelo de regresión lineal normal.
  - COUNT *lista de variables*: se usa un modelo de regresión Poisson.
  - DROP *lista de variables*: excluye variables del procedimiento de imputación y tampoco las incluye en el archivo de salida.
  - TRANSFER *lista de variables*: excluye variables del procedimiento de imputación pero si las incluye en el archivo de salida, pueden usarse para definir restricciones o cotas.

  - DEFAULT *tipo de variable*: asume que todas las variables tienen el tipo de variable indicado, se recomienda para fin de eliminar la necesidad de escribir una larga lista de variables de un mismo tipo.
- RUN: cierra la sintaxis del modulo IMPUTE.

Opcionales:

- RESTRICT *variable (expresión lógica)*: se usa para restringir la imputación en aquellas observaciones que satisfacen la expresión lógica.
- BOUNDS *variable (expresión lógica)*: es útil para restringir el rango de valores a ser imputados.
- INTERACT *variable*variable*: le permite al usuario especificar los términos de interacción a ser incluidos en el modelo de regresión.
- MAXPRED *número*; OR MAXPRED *lista de variables (número)*: define el número máximo de variables predictoras a ser incluidas en el modelo. Se usa un procedimiento de regresión por pasos para seleccionar a las mejores predictoras
- MINRSQD *decimal*: especifica el r^2 (mínimo marginal) para una regresión por pasos.
- MAXLOGI *número*: indica el número máximo de algoritmos iterativos a ser desarrollado bajo un modelo de regresión logístico o multilogit, el valor por default es 50.
- MINCODI *decimal*: especifica el cambio mínimo proporcional en cualquier coeficiente de regresión para continuar un proceso iterativo de regresión logística.
- ITERATIONS *número*: define el número de ciclos a realizar (Regresión Secuencial).
- MULTIPLES *número*: indica el número de imputaciones a realizar (IM).
- PERTURB *instrucción*: seguida de una instrucción (COEF/SIR) permite al usuario controlar las perturbaciones de los valores imputados.
- SEED *número*: especifica una semilla para la selección aleatoria de la distribución predictiva posterior (debe ser mayor a cero).
- NOBS *número*: indica el número de observaciones a ser usadas en el análisis, se puede elegir por esta opción mientras el proceso esté en pruebas.
- OFFSETS *variables discretas (variable OFFSET*): se usa para especificar una variable compensatoria cuando se ajusta un modelo de regresión Poisson.
- PRINT *instrucciones*: indica características de reporte de salida.
- TITLE *texto\n texto*: indica el título a ser impreso en la cabecera de cada página, \n indica que el texto debe ser impreso después de la primera línea.

Comando PUTDATA

Dada una Imputación Múltiple, el comando PUTDATA permite asignar en distintos archivos a todos los *m* conjuntos de datos imputados, su sintaxis debe contener:

- una instrucción DATAIN o una instrucción de imputación y
- la instrucción RUN.

En general, la sintaxis de ejecución tiene el siguiente formato:

%PUTDATA (name=, dir =, setup=)
Instrucciones de programación

Instrucciones de programación para el comando **PUTDATA.**

- DATAIN *archivo*: identifica la ubicación y el nombre del archivo de datos a imputar o analizar.
- DATOUT *archivo*: identifica la ubicación y el nombre del archivo de datos a ser generado. Si la instrucción se omite, los archivos de salida estarán en el directorio y con el nombre que se especifiquen dentro del paréntesis del comando, se emplean las extensiones .MET y .DAT para archivos de metadatos y datos, respectivamente.
- DELIMITER *carácter(es)*: se usa la opción "CSV" para indicar que las variables están separadas por una coma u otro(s) carácter(es) para delimitar la longitud de las variables. Si los datos incluyen un delimitador, este debe encerrarse dentro de comillas. "/t" indica como delimitador al tabulador.
- IMPUTATION *archivo*: especifica el nombre y ubicación del archivo imputado a leer.
- MULT *número/all*: indica el número de Imputación Múltiple a incluir en el archivo de salida. ALL especifica que se incluirán las *m* imputaciones en el mismo archivo. Si la instrucción es omitida, sólo se incluye la primera Imputación Múltiple.
- NOBS *número*: representa el número de registros a ser incluidas en el archivo de salida. Si esta instrucción es omitida, se incluirán todos los registros del archivo a imputar.
- PRINT *none/standard/details/all*: se define las características de la impresión
  - *None*: la opciones estándar.
  - *standard*: genera un resumen acerca de la base de datos.
  - *details* y *all*: produce un resumen de la base de datos e información detallada acerca de las variables y de la codificación respectiva.
    La opción por default es *standard.*
- TABLE *archivo(s)*: indica el formato de la tabla de datos a ser leída. Los datos debe estar en una hoja de cálculo en forma de tabla, el primer renglón debe contener los nombres de las variables, los nombres de las variables de tipo carácter deben estar precedidos por el símbolo "$". Las celdas deben estar delimitadas por blancos, comas o caracteres tipo tabulador. Las celdas cuyos valores incluyan blancos, comas o caracteres tipo tabulador, deben encerrarse dentro de comillas sencillas o dobles.
- TITLE *texto*: define el texto a ser empleado como título.
- RUN: instrucción requerida, debe ser la última instrucción de este comando.

Anexo F

Programa de cómputo RECODEINEGI.PRG

```
SET DECIMALS TO 6
SET SAFETY off
USE c:\dario\dirnormatividad\mceo\tesis\insumoinegi\complede(limpia).dbf
COPY TO d:\dario\completa(inegi)\complede.dbf
CLOSE ALL
USE d:\dario\completa(inegi)\complede.dbf
REPLACE a.edad_cm WITH '.I' FOR a.edad='999' AND a. Edo_cony<>' '
REPLACE a.edad_cm WITH '.I' FOR a.edad='999' AND (a.niv_esc=' ' OR (a.niv_esc<>' ' AND a. Edo_cony=' '))
REPLACE a.edad_cm WITH '.I' FOR VAL (a.edad)<12
Replace a.edad_cm WITH '.' FOR (edad='999' AND a. Edo_cony<>' ' and (acti_ina<>'99' OR parentes<>'999' or
edad<>'999' or niv_esc<>'9' or motriz<>'9' or edo_cony<>'9' and (vivos<>'98' OR vivos<>'  )))
REPLACE a.edad_cm WITH a.edad FOR (a.edad_cm='   ')
REPLACE a.edad _f   WITH '0' FOR   left (a.edad_cm,2)='. '
REPLACE a.edad _f   WITH '9' FOR   left (a.edad_cm,2)='.I'
REPLACE a.edad _f   WITH '1' FOR (a.edad _f=' ')
REPLACE a.motriz_cm WITH '.I' FOR a. Motriz='9'
REPLACE a.motriz_cm WITH '2' FOR a.no_disc='8'
REPLACE a.motriz_cm WITH '.I' FOR a. Motriz=' ' AND a.no_disc=' '
REPLACE a.motriz_cm WITH '.' FOR (motriz='9' AND (acti_ina<>'99' OR parentes<>'999' or edad<>'999' or niv_esc<>'9' or
motriz<>'9' or edo_cony<>'9' or edo_cony<>'9'   and (vivos<>'98' OR vivos<>'  )))
REPLACE a.motriz_cm WITH a. Motriz FOR a.motriz_cm='  '
REPLACE a.motriz_f WITH '0' FOR a.motriz_cm='. '
REPLACE a.motriz_f WITH '9' FOR a.motriz_cm='.I'
REPLACE a.motriz_f WITH '1' FOR (a.motriz_f=' ')
REPLACE a.niv_esc_cm WITH '.I' FOR a.niv_esc='9'
REPLACE a.niv_esc_cm WITH '.I' FOR VAL (a.edad)<5
REPLACE a.niv_esc_cm WITH '.' FOR (niv_esc='9' AND (acti_ina<>'99' OR parentes<>'999' or edad<>'999' or niv_esc<>'9'
or motriz<>'9' or edo_cony<>'9' or edo_cony<>'9'   and (vivos<>'98' OR vivos<>'  )))
REPLACE a.niv_esc_cm WITH a.niv_esc FOR a.niv_esc_cm='  '
REPLACE a.niv_esc_f WITH '0' FOR left (a.niv_esc_cm,2)='. '
REPLACE a.niv_esc_f WITH '9' FOR left (a.niv_esc_cm,2)='.I'
REPLACE a.niv_esc_f WITH '1' FOR (a.niv_esc_f=' ')
REPLACE a.ocupac_cm WITH '.' FOR a. Ocupac='9999'
blanco en ocupac
REPLACE a.ocupac_cm WITH '.I' FOR a. Ocupac='    '
REPLACE a.ocupac_cm WITH LEFT (a.ocupac,2) FOR a.ocupac_cm='  '
REPLACE a.ocupac_f WITH '0' FOR left (a.ocupac_cm,2)='. '
REPLACE a.ocupac_f WITH '9' FOR left (a.ocupac_cm,2)='.I'
REPLACE a.ocupac_f WITH '1' FOR (a.ocupac_f=' ')
REPLACE a.pos_tra_cm WITH '.' FOR a. Pos_tra='9'
blanco en ocupac
REPLACE a.pos_tra_cm WITH '.I' FOR a. Pos_tra=' '
REPLACE a.pos_tra_cm WITH a. Pos_tra FOR a.pos_tra_cm='  '
REPLACE a.pos_tra_f WITH '0' FOR left (a.pos_tra_cm,2)='. '
REPLACE a.pos_tra_f WITH '9' FOR left (a.pos_tra_cm,2)='.I'
REPLACE a.pos_tra_f WITH '1' FOR (a.pos_tra_f=' ')
REPLACE a.vivos_cm  WITH '.I' FOR (a. vivos='99' OR vivos='98')
REPLACE a.vivos_cm WITH '.I' FOR a. Vivos='  '
REPLACE a.vivos_cm WITH '.' FOR ((a. Vivos='99' OR vivos='98') AND (acti_ina<>'99' OR parentes<>'999' or edad<>'999'
or niv_esc<>'9' or motriz<>'9' or edo_cony<>'9' or edo_cony<>'9'   and (vivos<>'98' OR vivos<>'  )))
REPLACE a.vivos_cm  WITH a. vivos FOR a.vivos_cm='  '
REPLACE a.vivos_f WITH '0' FOR left (a.vivos_cm,2)='. '
REPLACE a.vivos_f WITH '9' FOR left (a.vivos_cm,2)='.I'
REPLACE a.vivos_f WITH '1' FOR (a.vivos_f=' ')
REPLACE a.parentes_c WITH '.I' FOR A.PARENTES='999'
REPLACE a.parentes_c WITH '.I' FOR a.seg $'IVJ'
REPLACE a.parentes_c WITH '.I' FOR a. Parentes='100'
REPLACE a.parentes_c WITH '.' FOR (A.PARENTES='999' AND (acti_ina<>'99' OR parentes<>'999' or edad<>'999' or
niv_esc<>'9' or motriz<>'9' or edo_cony<>'9' or edo_cony<>'9'   and (vivos<>'98' OR vivos<>'  )))
REPLACE a.parentes_c WITH a. Parentes FOR a.parentes_c='  '
REPLACE a.parentes_f WITH '0' FOR left (a.parentes_c,2)='. '
REPLACE a.parentes_f WITH '9' FOR left (a.parentes_c,2)='.I'
REPLACE a.parentes_f WITH '1' FOR (a.parentes_f=' ')
REPLACE a.edo_cony_c WITH '.I' FOR a. Edo_cony='9'
REPLACE a.edo_cony_c WITH '.I' FOR VAL(a.edad)<12
REPLACE a.edo_cony_c WITH '.' FOR (a. Edo_cony='9' AND (acti_ina<>'99' OR parentes<>'999' or edad<>'999' or
niv_esc<>'9' or motriz<>'9' or edo_cony<>'9' or edo_cony<>'9'   and (vivos<>'98' OR vivos<>'  )))
REPLACE a.edo_cony_c WITH a. Edo_cony FOR a.edo_cony_c='  '
REPLACE a.edo_cony_f WITH '0' FOR left (a.edo_cony_c,2)='. '
```

```
REPLACE a.edo_cony_f WITH '9' FOR left (a.edo_cony_c,2)='.I'
REPLACE a.edo_cony_f WITH '1' FOR (a.edo_cony_f=' ')
Replace a.log_ing WITH STR (LOG (VAL (a.tot_ing)+1),10,6) FOR (a.tot_ing>='000000' AND VAL(a.tot_ing)<999999)
replace a.log_ing WITH a.tot_ing FOR a.log_ing='          '
REPLACE a.log_ing_cm WITH '.' FOR (a. Impute<>' ')
REPLACE a.log_ing_cm WITH '.' FOR (a.tot_ing='999999' OR a.tot_ing='999998')
REPLACE a.log_ing_cm WITH '.I' FOR (acti_ina='99' and parentes='999' and edad='999' and niv_esc='9' and motriz='9' and
edo_cony='9' and tot_ing='000000' and (vivos='98' or vivos='  '))
REPLACE a.log_ing_cm WITH '.I' FOR (a.log_ing='          ')
REPLACE a.log_ing_cm WITH a.log_ing FOR (a.log_ing_cm='          ')
REPLACE a.log_ing_f WITH '0' FOR left (a.log_ing_cm,2)='. '
REPLACE a.log_ing_f WITH '9' FOR left (a.log_ing_cm,2)='.I'
REPLACE a.log_ing_f WITH '1' FOR (a.log_ing_f=' ')
CLOSE ALL
USE d:\dario\completa(inegi)\complede.dbf
COPY to d:\dario\completa (inegi)\complede.dbf TYPE fox2x
CLOSE all
```

Anexo G

Sintaxis General (importación, imputación y extracción)

```
%getdata (name=insumo, dir=d: \dario\completa (inegi), setup=new);
 Datain d:\dario\completa (inegi)\complede.txt;
 Metadata;
 Delim "\t";
 Variables
   Name=id_perso    type=char;
   Name=id_hogar    type=char;
   Name=id_viv      type=char;
   Name=ENT         type=char;
   Name=mun         type=char;
   Name=loc         type=char;
   Name=Ageb        type=char;
   Name=mza         type=char;
   Name=seg         type=char;
   Name=num_per     type=char;
   Name=parentes    type=num;
   Name=sexo        type=num;
   Name=edad        type=num;
   Name=motriz      type=num;
   Name=niv_esc     type=num;
   Name=edo_cony    type=num;
   Name=ocupac      type=num;
   Name=pos_tra     type=num;
   Name=tot_ing     type=num;
   Name=vivos       type=num;
   Name=zona        type=num;
   Name=estrato     type=num;
   Name=subestra    type=num;
   Name=tipo_cuest  type=char;
   Name=llave_un    type=char;
   Name=num_viv     type=char;
   Name=apellido    type=char;
   Name=impute      type=char;
   Name=sobra       type=char;
   Name=edad_cm     type=num;
   Name=edad_f      type=num;
   Name=motriz_cm   type=num;
   Name=motriz_f    type=num;
   Name=niv_esc_cm  type=num;
   Name=niv_esc_f   type=num;
   Name=ocupac_cm   type=num;
   Name=ocupac_f    type=num;
   Name=pos_tra_cm  type=num;
   Name=pos_tra_f   type=num;
   Name=vivos_cm    type=num;
   Name=vivos_f     type=num;
   Name=parentes_c  type=num;
   Name=parentes_f  type=num;
   Name=edo_cony_c  type=num;
   Name=edo_cony_f  type=num;
   Name=log_ing     type=num;
   Name=log_ing_cm  type=num;
   Name=log_ing_f   type=num;
   Name=tot_ing_im  type=num;
 End;
 Run;
%impute (name=salida, dir=d: \dario\completa (inegi), setup=new);
   Datain                insumo;
   Dataout               darioout;
   Default               transfer;
   Continuous            log_ing_cm edad_cm vivos_cm;
```

```
   Categorical          parentes_c sexo motriz_cm niv_esc_cm edo_cony_c
                        ocupac_cm pos_tra_cm zona estrato subestra;
   Restrict             log_ing_cm (log_ing_f=0, 1)
                         edad_cm (edad_f=0,1)
                        parentes_c (parentes_f=0,1)
                        motriz_cm (motriz_f=0,1)
                        niv_esc_cm (niv_esc_f=0,1)
                        edo_cony_c (edo_cony_f=0,1)
                        ocupac_cm (ocupac_f=0,1)
                        pos_tra_cm (pos_tra_f=0,1)
                        vivos_cm (vivos_f=0,1);
   Bounds               log_ing_cm (>=0, <=13.815508)
                        edad_cm (>=12, <=130)
                        vivos_cm (>=0, <=25);
   Seed                 100;
   Iterations           2;
   Multiples            5;
Run;
%putdata (name=darioout, dir=d: \dario\completa (inegi), setup=new);
   Imputation            d:\dario\completa (inegi)\salida;
   Table                 d:\dario\completa (inegi)\m1;
   Mult                 1;
Run;
%putdata (name=darioout, dir=d: \dario\completa (inegi), setup=new);
   Imputation            d:\dario\completa (inegi)\salida;
   Table                 d:\dario\completa (inegi)\m2;
   Mult                 2;
Run;
%putdata (name=darioout, dir=d: \dario\completa (inegi), setup=new);
   Imputation            d:\dario\completa (inegi)\salida;
   Table                 d:\dario\completa (inegi)\m3;
   Mult                 3;
Run;
%putdata (name=darioout, dir=d: \dario\completa (inegi), setup=new);
   Imputation            d:\dario\completa (inegi)\salida;
   Table                 d:\dario\completa (inegi)\m4;
   Mult                 4;
Run;
%putdata (name=darioout, dir=d: \dario\completa (inegi), setup=new);
   Imputation            d:\dario\completa (inegi)\salida;
   Table                 d:\dario\completa (inegi)\m5;
   Mult                 5;
Run;
```

Anexo H

Programa de cómputo EST_LLENAINEGI.PRG

```
SET DELETE ON
SET EXCLUSIVE OFF
SET TALK OFF
SET STATUS ON
SET CLOCK ON
SET DATE TO FRENCH
SET SAFETY OFF
CLEAR
SET PRINTER TO "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\salidaFOX.lst"
SET PRINTER on
?'IVEware Iterative Imputation Procedure (VISUAL FOX)'
?'FECHA INICIAL',DATE()
? 'tiempo inicial:' + time()
?
?
SELECT c
USE "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\multii1m4.dbf" SHARED
CALCULATE std(c.edad_cm) FOR c.edad_f=0 TO sdvedadi
CALCULATE std(c.edad_cm) FOR c.edad_f=1 TO sdvedado
CALCULATE std(c.edad_cm) FOR c.edad_f=0 OR c.edad_f=1 TO sdvedadc
CLOSE all
SELECT edad_f,cnt(*) as total ,MIN(edad_cm) as mi,MAX(edad_cm) as ma,avg(edad_cm) as me FROM
"d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\multii1m4.dbf" inTO TABLE
"d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\aa.dbf" GROUP BY edad_f WHERE
edad_f<>9
CLOSE all
SELECT a
DECLARE nom(3)
nom[01]='imputed    '
nom[02]='observed   '
nom[03]='combined   '
USE "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\aa.dbf" SHARED
?'edad_cm          nnumber              minimun               maximun                   mean
stddev'
?'-------------------------------------------------------------------------------------------'
R1=1
DO WHILE NOT EOF()
    ?NOM[R1]
    R1=R1+1
    ??a.total
    ??ROUND(a.mi,2)
    ??ROUND(a.ma,2)
    ??ROUND(A.me,2)
    IF r1=2 then
       ??ROUND(sdvedadi,2)
    ELSE
       ??ROUND(sdvedado,2)
    endif
    SKIP IN a
ENDDO
SELECT cnt(*) as total,MIN(edad_cm) as mi,MAX(edad_cm) as ma,avg(edad_cm) as me FROM
"d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\multii1m4.dbf" inTO TABLE
"d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\bb.dbf" where edad_f<>9
CLOSE all
SELECT a
USE "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\bb.dbf" SHARED
?NOM[3]
DO WHILE NOT EOF()
    ??a.total
    ??ROUND(a.mi,2)
    ??ROUND(a.ma,2)
    ??ROUND(a.me,2)
    ??ROUND(sdvedadc,2)
    SKIP IN a
ENDDO
CLOSE database
?
************************************ DECLARACION E INICIALIZACION DE VARIABLES GLOBALES
STORE 3 TO NF                 && NF: NUMERO DE RENGLONES
STORE 6 TO NC                  && NC: NUMERO DE COLUMNAS
DECLARE MAT(NF,NC)
```

```
STORE 0.0 TO MAT
DECLARE nom(3)
nom[01]='code 1  '
nom[02]='code 2  '
nom[03]='total   '
create table Estruct.dbf;
(field_name C(10),field_type C(1),field_len N(3,0),field_dec N(3,0))
FOR B=1 TO NC
   CAMPO='C'+LTRIM(STR(B))
   append blank
   replace field_name with CAMPO
   replace field_type with 'N'
   replace field_len  with 10
   replace field_dec  with 2
ENDFOR
create "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\aa" from Estruct.dbf
USE
ERASE Estruct.dbf
SELECT a
USE "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\multii1m4.dbf" SHARED
************************************************************motriz
?'motriz_cm          Observed            Imputed              Combined     '
?'                 Freq    Per         Freq    Per          Freq     Per   '
?'-------------------------------------------------------------------------------'
R1=1
DO WHILE NOT EOF()
    IF motriz_f=1 AND motriz_cm=1 then
       mat[1,1]=mat[1,1]+1
    ENDIF
    IF motriz_f=0 AND motriz_cm=1 then
       mat[1,3]=mat[1,3]+1
    ENDIF
    IF motriz_cm=1 then
       mat[1,5]=mat[1,5]+1
    ENDIF
    IF motriz_f=1 AND motriz_cm=2 then
       mat[2,1]=mat[2,1]+1
    ENDIF
    IF motriz_f=0 AND motriz_cm=2 then
       mat[2,3]=mat[2,3]+1
    ENDIF
    IF motriz_cm=2 then
       mat[2,5]=mat[2,5]+1
    ENDIF
    SKIP IN a
ENDDO
FOR j= 1 TO nc
   FOR i=1 TO nf-1
      mat[nf,j]=mat[nf,j]+mat[i,j]
  ENDFOR
ENDFOR
FOR i= 1 TO nf
   FOR j=2 TO nc STEP 2
      mat[i,j]=mat[i,j-1]/mat[nf,j-1]*100
  ENDFOR
ENDFOR
   SELE C
   USE "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\aa"
      FOR B=1 TO NF
         APPEND BLANK
      ENDFOR
SELECT c
   FOR REN=1 TO NF
      GO REN IN c
      FOR COL=1 TO NC
         CAMPO=FIELD(COL,'c')
         REPLACE c.&CAMPO. WITH MAT[REN,COL]
      ENDFOR
   ENDFOR
SELECT c
   FOR REN=1 TO RECCOUNT('c')
```

```
      GO REN IN c
      FOR COL=1 TO FCOUNT('c')
         CC=LTRIM(STR(COL))
         IF COL=1
            ? nom[Ren]
            ?? TRANSFORM( c.C&CC., '    999,999')
         ELSE
            if(MOD(col,2)<>0) then
               ?? TRANSFORM( c.C&CC., '    999,999')
            ELSE
               ?? TRANSFORM( c.C&CC., '    999.99')
            endif
         ENDIF
      ENDFOR
   ENDFOR
CLOSE database
?
************************************** DECLARACION E INICIALIZACION DE VARIABLES GLOBALES
STORE 10 TO NF                && NF: NUMERO DE RENGLONES
STORE 6 TO NC                 && NC: NUMERO DE COLUMNAS
DECLARE MAT(NF,NC)
STORE 0 TO MAT
DECLARE nom(10)
nom[01]='code 0  '
nom[02]='code 1  '
nom[03]='code 2  '
nom[04]='code 3  '
nom[05]='code 4  '
nom[06]='code 5  '
nom[07]='code 6  '
nom[08]='code 7  '
nom[09]='code 8  '
nom[10]='total   '
*crea cascaron
create table Estruct.dbf;
(field_name C(10),field_type C(1),field_len N(3,0),field_dec N(3,0))
FOR B=1 TO NC
   CAMPO='C'+LTRIM(STR(B))
   append blank
   replace field_name with CAMPO
   replace field_type with 'N'
   replace field_len  with 10
   replace field_dec  with 2
ENDFOR
create "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\aa" from Estruct.dbf
USE
ERASE Estruct.dbf
SELECT a
USE "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\multii1m4.dbf" SHARED
************************************************************niv_esc
?'niv_esc_cm          Observed            Imputed              Combined     '
?'                 Freq    Per         Freq    Per          Freq     Per   '
?'-------------------------------------------------------------------------------'
R1=1
DO WHILE NOT EOF()
    IF niv_esc_f=1 AND niv_esc_cm=0 then
       mat[1,1]=mat[1,1]+1
    ENDIF
    IF niv_esc_f=0 AND niv_esc_cm=0 then
       mat[1,3]=mat[1,3]+1
    ENDIF
    IF niv_esc_cm=0 then
       mat[1,5]=mat[1,5]+1
    ENDIF
    IF niv_esc_f=1 AND niv_esc_cm=1 then
       mat[2,1]=mat[2,1]+1
    ENDIF
    IF niv_esc_f=0 AND niv_esc_cm=1 then
       mat[2,3]=mat[2,3]+1
    ENDIF
    IF niv_esc_cm=1 then
```

```
      mat[2,5]=mat[2,5]+1
    ENDIF
    IF niv_esc_f=1 AND niv_esc_cm=2 then
      mat[3,1]=mat[3,1]+1
    ENDIF
    IF niv_esc_f=0 AND niv_esc_cm=2 then
      mat[3,3]=mat[3,3]+1
    ENDIF
    IF niv_esc_cm=2 then
      mat[3,5]=mat[3,5]+1
    ENDIF
    IF niv_esc_f=1 AND niv_esc_cm=3 then
      mat[4,1]=mat[4,1]+1
    ENDIF
    IF niv_esc_f=0 AND niv_esc_cm=3 then
      mat[4,3]=mat[4,3]+1
    ENDIF
    IF niv_esc_cm=3 then
      mat[4,5]=mat[4,5]+1
    ENDIF
    IF niv_esc_f=1 AND niv_esc_cm=4 then
      mat[5,1]=mat[5,1]+1
    ENDIF
    IF niv_esc_f=0 AND niv_esc_cm=4 then
      mat[5,3]=mat[5,3]+1
    ENDIF
    IF niv_esc_cm=4 then
      mat[5,5]=mat[5,5]+1
    ENDIF
    IF niv_esc_f=1 AND niv_esc_cm=5 then
      mat[6,1]=mat[6,1]+1
    ENDIF
    IF niv_esc_f=0 AND niv_esc_cm=5 then
      mat[6,3]=mat[6,3]+1
    ENDIF
    IF niv_esc_cm=5 then
      mat[6,5]=mat[6,5]+1
    ENDIF
    IF niv_esc_f=1 AND niv_esc_cm=6 then
      mat[7,1]=mat[7,1]+1
    ENDIF
    IF niv_esc_f=0 AND niv_esc_cm=6 then
      mat[7,3]=mat[7,3]+1
    ENDIF
    IF niv_esc_cm=6 then
      mat[7,5]=mat[7,5]+1
    ENDIF
    IF niv_esc_f=1 AND niv_esc_cm=7 then
      mat[8,1]=mat[8,1]+1
    ENDIF
    IF niv_esc_f=0 AND niv_esc_cm=7 then
      mat[8,3]=mat[8,3]+1
    ENDIF
    IF niv_esc_cm=7 then
      mat[8,5]=mat[8,5]+1
    ENDIF
    IF niv_esc_f=1 AND niv_esc_cm=8 then
      mat[9,1]=mat[9,1]+1
    ENDIF
    IF niv_esc_f=0 AND niv_esc_cm=8 then
      mat[9,3]=mat[9,3]+1
    ENDIF
    IF niv_esc_cm=8 then
      mat[9,5]=mat[9,5]+1
    ENDIF
    SKIP IN a
ENDDO
FOR j= 1 TO nc
   FOR i=1 TO nf-1
      mat[nf,j]=mat[nf,j]+mat[i,j]
  ENDFOR
```

```
ENDFOR
FOR i= 1 TO nf
   FOR j=2 TO nc STEP 2
      mat[i,j]=mat[i,j-1]/mat[nf,j-1]*100
  ENDFOR
ENDFOR
*llena matriz a base
*crea abase
   SELE C
   USE "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\aa"
      FOR B=1 TO NF
         APPEND BLANK
      ENDFOR
*llena base
SELECT c
   FOR REN=1 TO NF
      GO REN IN c
      FOR COL=1 TO NC
         CAMPO=FIELD(COL,'c')
         REPLACE c.&CAMPO. WITH MAT[REN,COL]
      ENDFOR
   ENDFOR
*imprime base para reporte
SELECT c
   FOR REN=1 TO RECCOUNT('c')
      GO REN IN c
      FOR COL=1 TO FCOUNT('c')
         CC=LTRIM(STR(COL))
         IF COL=1
            ? nom[Ren]
            ?? TRANSFORM( c.C&CC., '    999,999')
         ELSE
            if(MOD(col,2)<>0) then
               ?? TRANSFORM( c.C&CC., '    999,999')
            ELSE
               ?? TRANSFORM( c.C&CC., '    999.99')
            endif
         ENDIF
      ENDFOR
   ENDFOR
?
CLOSE database
************************************* DECLARACION E INICIALIZACION DE VARIABLES GLOBALES
STORE 19 TO NF                && NF: NUMERO DE RENGLONES
STORE 6 TO NC                 && NC: NUMERO DE COLUMNAS
DECLARE MAT(NF,NC)
STORE 0 TO MAT
DECLARE nom(19)
nom[01]='code 11  '
nom[02]='code 12  '
nom[03]='code 13  '
nom[04]='code 14  '
nom[05]='code 21  '
nom[06]='code 41  '
nom[07]='code 51  '
nom[08]='code 52  '
nom[09]='code 53  '
nom[10]='code 54  '
nom[11]='code 55  '
nom[12]='code 61  '
nom[13]='code 62  '
nom[14]='code 71  '
nom[15]='code 72  '
nom[16]='code 81  '
nom[17]='code 82  '
nom[18]='code 83  '
nom[19]='total    '
create table Estruct.dbf;
(field_name C(10),field_type C(1),field_len N(3,0),field_dec N(3,0))
FOR B=1 TO NC
   CAMPO='C'+LTRIM(STR(B))
```

```
   append blank
   replace field_name with CAMPO
   replace field_type with 'N'
   replace field_len  with 10
   replace field_dec  with 2
ENDFOR
create "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\aa" from Estruct.dbf
USE
ERASE Estruct.dbf
SELECT a
USE "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\multii1m4.dbf" SHARED
*****************************************************************ocupac
?'ocupac_cm          Observed            Imputed              Combined     '
?'                 Freq    Per         Freq    Per          Freq     Per   '
?'-------------------------------------------------------------------------------'
R1=1
DO WHILE NOT EOF()
    IF (ocupac_f)=1 AND ocupac_cm=11 then
       mat[1,1]=mat[1,1]+1
    ENDIF
    IF (ocupac_f)=0 AND ocupac_cm=11 then
       mat[1,3]=mat[1,3]+1
    ENDIF
    IF (ocupac_cm)=11 then
       mat[1,5]=mat[1,5]+1
    ENDIF
    IF (ocupac_f)=1 AND ocupac_cm=12 then
       mat[2,1]=mat[2,1]+1
    ENDIF
    IF (ocupac_f)=0 AND ocupac_cm=12 then
       mat[2,3]=mat[2,3]+1
    ENDIF
    IF (ocupac_cm)=12 then
       mat[2,5]=mat[2,5]+1
    ENDIF
    IF (ocupac_f)=1 AND ocupac_cm=13 then
       mat[3,1]=mat[3,1]+1
    ENDIF
    IF (ocupac_f)=0 AND ocupac_cm=13 then
       mat[3,3]=mat[3,3]+1
    ENDIF
    IF (ocupac_cm)=13 then
       mat[3,5]=mat[3,5]+1
    ENDIF
    IF (ocupac_f)=1 AND ocupac_cm=14 then
       mat[4,1]=mat[4,1]+1
    ENDIF
    IF (ocupac_f)=0 AND ocupac_cm=14 then
       mat[4,3]=mat[4,3]+1
    ENDIF
    IF (ocupac_cm)=14 then
       mat[4,5]=mat[4,5]+1
    ENDIF
    IF (ocupac_f)=1 AND ocupac_cm=21 then
       mat[5,1]=mat[5,1]+1
    ENDIF
    IF (ocupac_f)=0 AND ocupac_cm=21 then
       mat[5,3]=mat[5,3]+1
    ENDIF
    IF (ocupac_cm)=21 then
       mat[5,5]=mat[5,5]+1
    ENDIF
    IF (ocupac_f)=1 AND ocupac_cm=41 then
       mat[6,1]=mat[6,1]+1
    ENDIF
    IF (ocupac_f)=0 AND ocupac_cm=41 then
       mat[6,3]=mat[6,3]+1
    ENDIF
    IF (ocupac_cm)=41 then
       mat[6,5]=mat[6,5]+1
    ENDIF
```

```
IF (ocupac_f)=1 AND ocupac_cm=51 then
   mat[7,1]=mat[7,1]+1
ENDIF
IF (ocupac_f)=0 AND ocupac_cm=51 then
   mat[7,3]=mat[7,3]+1
ENDIF
IF (ocupac_cm)=51 then
   mat[7,5]=mat[7,5]+1
ENDIF
IF (ocupac_f)=1 AND ocupac_cm=52 then
   mat[8,1]=mat[8,1]+1
ENDIF
IF (ocupac_f)=0 AND ocupac_cm=52 then
   mat[8,3]=mat[8,3]+1
ENDIF
IF (ocupac_cm)=52 then
   mat[8,5]=mat[8,5]+1
ENDIF
IF (ocupac_f)=1 AND ocupac_cm=53 then
   mat[9,1]=mat[9,1]+1
ENDIF
IF (ocupac_f)=0 AND ocupac_cm=53 then
   mat[9,3]=mat[9,3]+1
ENDIF
IF (ocupac_cm)=53 then
   mat[9,5]=mat[9,5]+1
ENDIF
IF (ocupac_f)=1 AND ocupac_cm=54 then
   mat[10,1]=mat[10,1]+1
ENDIF
IF (ocupac_f)=0 AND ocupac_cm=54 then
   mat[10,3]=mat[10,3]+1
ENDIF
IF (ocupac_cm)=54 then
   mat[10,5]=mat[10,5]+1
ENDIF
IF (ocupac_f)=1 AND ocupac_cm=55 then
   mat[11,1]=mat[11,1]+1
ENDIF
IF (ocupac_f)=0 AND ocupac_cm=55 then
   mat[11,3]=mat[11,3]+1
ENDIF
IF (ocupac_cm)=55 then
   mat[11,5]=mat[11,5]+1
ENDIF
IF (ocupac_f)=1 AND ocupac_cm=61 then
   mat[12,1]=mat[12,1]+1
ENDIF
IF (ocupac_f)=0 AND ocupac_cm=61 then
   mat[12,3]=mat[12,3]+1
ENDIF
IF (ocupac_cm)=61 then
   mat[12,5]=mat[12,5]+1
ENDIF
IF (ocupac_f)=1 AND ocupac_cm=62 then
   mat[13,1]=mat[13,1]+1
ENDIF
IF (ocupac_f)=0 AND ocupac_cm=62 then
   mat[13,3]=mat[13,3]+1
ENDIF
IF (ocupac_cm)=62 then
   mat[13,5]=mat[13,5]+1
ENDIF
IF (ocupac_f)=1 AND ocupac_cm=71 then
   mat[14,1]=mat[14,1]+1
ENDIF
IF (ocupac_f)=0 AND ocupac_cm=71 then
   mat[14,3]=mat[14,3]+1
ENDIF
IF (ocupac_cm)=71 then
   mat[14,5]=mat[14,5]+1
```

```
    ENDIF
    IF (ocupac_f)=1 AND ocupac_cm=72 then
       mat[15,1]=mat[15,1]+1
    ENDIF
    IF (ocupac_f)=0 AND ocupac_cm=72 then
       mat[15,3]=mat[15,3]+1
    ENDIF
    IF (ocupac_cm)=72 then
       mat[15,5]=mat[15,5]+1
    ENDIF
    IF (ocupac_f)=1 AND ocupac_cm=81 then
       mat[16,1]=mat[16,1]+1
    ENDIF
    IF (ocupac_f)=0 AND ocupac_cm=81 then
       mat[16,3]=mat[16,3]+1
    ENDIF
    IF (ocupac_cm)=81 then
       mat[16,5]=mat[16,5]+1
    ENDIF
    IF (ocupac_f)=1 AND ocupac_cm=82 then
       mat[17,1]=mat[17,1]+1
    ENDIF
    IF (ocupac_f)=0 AND ocupac_cm=82 then
       mat[17,3]=mat[17,3]+1
    ENDIF
    IF (ocupac_cm)=82 then
       mat[17,5]=mat[17,5]+1
    ENDIF
    IF (ocupac_f)=1 AND ocupac_cm=83 then
       mat[18,1]=mat[18,1]+1
    ENDIF
    IF (ocupac_f)=0 AND ocupac_cm=83 then
       mat[18,3]=mat[18,3]+1
    ENDIF
    IF (ocupac_cm)=83 then
       mat[18,5]=mat[18,5]+1
    ENDIF
    SKIP IN a
ENDDO
FOR j= 1 TO nc
   FOR i=1 TO nf-1
      mat[nf,j]=mat[nf,j]+mat[i,j]
  ENDFOR
ENDFOR
FOR i= 1 TO nf
   FOR j=2 TO nc STEP 2
      mat[i,j]=mat[i,j-1]/mat[nf,j-1]*100
  ENDFOR
ENDFOR
   SELE C
   USE "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\aa"
      FOR B=1 TO NF
         APPEND BLANK
      ENDFOR
SELECT c
   FOR REN=1 TO NF
      GO REN IN c
      FOR COL=1 TO NC
         CAMPO=FIELD(COL,'c')
         REPLACE c.&CAMPO. WITH MAT[REN,COL]
      ENDFOR
   ENDFOR
SELECT c
   FOR REN=1 TO RECCOUNT('c')
      GO REN IN c
      FOR COL=1 TO FCOUNT('c')
         CC=LTRIM(STR(COL))
         IF COL=1
            ? nom[Ren]
            ?? TRANSFORM( c.C&CC., '    999,999')
         ELSE
```

```
            if(MOD(col,2)<>0) then
               ?? TRANSFORM( c.C&CC., '    999,999')
            ELSE
               ?? TRANSFORM( c.C&CC., '    999.99')
            endif
         ENDIF
      ENDFOR
   ENDFOR
?
CLOSE database
************************************* DECLARACION E INICIALIZACION DE VARIABLES GLOBALES
STORE 6 TO NF                && NF: NUMERO DE RENGLONES
STORE 6 TO NC                 && NC: NUMERO DE COLUMNAS
DECLARE MAT(NF,NC)
STORE 0 TO MAT
DECLARE nom(6)
nom[01]='code 1  '
nom[02]='code 2  '
nom[03]='code 3  '
nom[04]='code 4  '
nom[05]='code 5  '
nom[06]='total   '
create table Estruct.dbf;
(field_name C(10),field_type C(1),field_len N(3,0),field_dec N(3,0))
FOR B=1 TO NC
   CAMPO='C'+LTRIM(STR(B))
   append blank
   replace field_name with CAMPO
   replace field_type with 'N'
   replace field_len  with 10
   replace field_dec  with 2
ENDFOR
create "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\aa" from Estruct.dbf
USE
ERASE Estruct.dbf
SELECT a
USE "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\multii1m4.dbf" SHARED
*****************************************************************pos_tra
?'pos_tra_cm          Observed            Imputed              Combined     '
?'                 Freq    Per         Freq    Per          Freq     Per   '
?'-------------------------------------------------------------------------------'
R1=1
DO WHILE NOT EOF()
    IF (pos_tra_f)=1 AND pos_tra_cm=1 then
       mat[1,1]=mat[1,1]+1
    ENDIF
    IF (pos_tra_f)=0 AND pos_tra_cm=1 then
       mat[1,3]=mat[1,3]+1
    ENDIF
    IF (pos_tra_cm)=1 then
       mat[1,5]=mat[1,5]+1
    ENDIF
    IF (pos_tra_f)=1 AND pos_tra_cm=2 then
       mat[2,1]=mat[2,1]+1
    ENDIF
    IF (pos_tra_f)=0 AND pos_tra_cm=2 then
       mat[2,3]=mat[2,3]+1
    ENDIF
    IF (pos_tra_cm)=2 then
       mat[2,5]=mat[2,5]+1
    ENDIF
    IF (pos_tra_f)=1 AND pos_tra_cm=3 then
       mat[3,1]=mat[3,1]+1
    ENDIF
    IF (pos_tra_f)=0 AND pos_tra_cm=3 then
       mat[3,3]=mat[3,3]+1
    ENDIF
    IF (pos_tra_cm)=3 then
       mat[3,5]=mat[3,5]+1
    ENDIF
    IF (pos_tra_f)=1 AND pos_tra_cm=4 then
```

```
      mat[4,1]=mat[4,1]+1
    ENDIF
    IF (pos_tra_f)=0 AND pos_tra_cm=4 then
      mat[4,3]=mat[4,3]+1
    ENDIF
    IF (pos_tra_cm)=4 then
      mat[4,5]=mat[4,5]+1
    ENDIF
    IF (pos_tra_f)=1 AND pos_tra_cm=5 then
      mat[5,1]=mat[5,1]+1
    ENDIF
    IF (pos_tra_f)=0 AND pos_tra_cm=5 then
      mat[5,3]=mat[5,3]+1
    ENDIF
    IF (pos_tra_cm)=5 then
      mat[5,5]=mat[5,5]+1
    ENDIF
    SKIP IN a
ENDDO
FOR j= 1 TO nc
   FOR i=1 TO nf-1
     mat[nf,j]=mat[nf,j]+mat[i,j]
  ENDFOR
ENDFOR
FOR i= 1 TO nf
   FOR j=2 TO nc STEP 2
     mat[i,j]=mat[i,j-1]/mat[nf,j-1]*100
  ENDFOR
ENDFOR
   SELE C
   USE "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\aa"
      FOR B=1 TO NF
        APPEND BLANK
      ENDFOR
SELECT c
   FOR REN=1 TO NF
      GO REN IN c
      FOR COL=1 TO NC
        CAMPO=FIELD(COL,'c')
        REPLACE c.&CAMPO. WITH MAT[REN,COL]
      ENDFOR
   ENDFOR
SELECT c
   FOR REN=1 TO RECCOUNT('c')
      GO REN IN c
      FOR COL=1 TO FCOUNT('c')
        CC=LTRIM(STR(COL))
        IF COL=1
          ? nom[Ren]
          ?? TRANSFORM( c.C&CC., '    999,999')
        ELSE
          if(MOD(col,2)<>0) then
            ?? TRANSFORM( c.C&CC., '    999,999')
          ELSE
            ?? TRANSFORM( c.C&CC., '    999.99')
          endif
        ENDIF
      ENDFOR
   ENDFOR
?
CLOSE database
*************************************************************SELECT c
USE "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\multii1m4.dbf" SHARED
CALCULATE std(c.vivos_cm) FOR c.vivos_f=0 TO sdvvivosi
CALCULATE std(c.vivos_cm) FOR c.vivos_f=1 TO sdvvivoso
CALCULATE std(c.vivos_cm) FOR c.vivos_f=0 OR c.vivos_f=1 TO sdvvivosc
SELECT vivos_f,cnt(*) as total ,MIN(vivos_cm) as mi,MAX(vivos_cm) as ma,avg(vivos_cm) as me
FROM "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\multii1m4.dbf" inTO TABLE
"d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\aa.dbf" GROUP BY vivos_f WHERE
vivos_f<>9
CLOSE all
```

```
SELECT a
DECLARE nom(3)
nom[01]='imputed    '
nom[02]='observed   '
nom[03]='combined   '
USE "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\aa.dbf" SHARED
?'vivos_cm       nnumber               minimun               maximun                 mean
stddev'
?'-------------------------------------------------------------------------------------'
R1=1
DO WHILE NOT EOF()
    ?NOM[R1]
    R1=R1+1
    ??a.total
    ??ROUND(a.mi,2)
    ??ROUND(a.ma,2)
    ??ROUND(A.me,2)
    IF r1=2 then
       ??ROUND(sdvvivosi,2)
    ELSE
       ??ROUND(sdvvivoso,2)
    endif
    SKIP IN a
ENDDO
SELECT cnt(*) as total,MIN(vivos_cm) as mi,MAX(vivos_cm) as ma,avg(vivos_cm) as me FROM
"d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\multii1m4.dbf" inTO TABLE
"d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\bb.dbf" where vivos_f<>9
CLOSE all
SELECT a
USE "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\bb.dbf" SHARED
?NOM[3]
DO WHILE NOT EOF()
    ??a.total
    ??ROUND(a.mi,2)
    ??ROUND(a.ma,2)
    ??ROUND(a.me,2)
    ??ROUND(sdvvivosc,2)
    SKIP IN a
ENDDO
CLOSE database
?
************************************* DECLARACION E INICIALIZACION DE VARIABLES GLOBALES
STORE 6 TO NF                && NF: NUMERO DE RENGLONES
STORE 6 TO NC                 && NC: NUMERO DE COLUMNAS
DECLARE MAT(NF,NC)
STORE 0 TO MAT
DECLARE nom(7)
nom[01]='code 2  '
nom[02]='code 3  '
nom[03]='code 4  '
nom[04]='code 5  '
nom[05]='code 6  '
nom[06]='total   '
create table Estruct.dbf;
(field_name C(10),field_type C(1),field_len N(3,0),field_dec N(3,0))
FOR B=1 TO NC
   CAMPO='C'+LTRIM(STR(B))
   append blank
   replace field_name with CAMPO
   replace field_type with 'N'
   replace field_len  with 10
   replace field_dec  with 2
ENDFOR
create "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\aa" from Estruct.dbf
USE
ERASE Estruct.dbf
SELECT a
USE "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\multii1m4.dbf" SHARED
*************************************************************parentes
?'parentes_c          Observed            Imputed              Combined     '
?'                Freq    Per         Freq    Per          Freq     Per   '
```

```
?'------------------------------------------------------------------------------'
R1=1
DO WHILE NOT EOF()
    IF (parentes_f)=1 AND (parentes_c)=2 then
       mat[1,1]=mat[1,1]+1
    ENDIF
    IF (parentes_f)=0 AND (parentes_c)=2 then
       mat[1,3]=mat[1,3]+1
    ENDIF
    IF (parentes_c)=2 then
       mat[1,5]=mat[1,5]+1
    ENDIF
    IF (parentes_f)=1 AND (parentes_c)=3 then
       mat[2,1]=mat[2,1]+1
    ENDIF
    IF (parentes_f)=0 AND (parentes_c)=3 then
       mat[2,3]=mat[2,3]+1
    ENDIF
    IF (parentes_c)=3 then
       mat[2,5]=mat[2,5]+1
    ENDIF
    IF (parentes_f)=1 AND (parentes_c)=4 then
       mat[3,1]=mat[3,1]+1
    ENDIF
    IF (parentes_f)=0 AND (parentes_c)=4 then
       mat[3,3]=mat[3,3]+1
    ENDIF
    IF (parentes_c)=4 then
       mat[3,5]=mat[3,5]+1
    ENDIF
    IF (parentes_f)=1 AND (parentes_c)=5 then
       mat[4,1]=mat[4,1]+1
    ENDIF
    IF (parentes_f)=0 AND (parentes_c)=5 then
       mat[4,3]=mat[4,3]+1
    ENDIF
    IF ((parentes_c)=5) then
       mat[4,5]=mat[4,5]+1
    ENDIF
    IF (parentes_f)=1 AND (parentes_c)=6 then
       mat[5,1]=mat[5,1]+1
    ENDIF
    IF (parentes_f)=0 AND (parentes_c)=6 then
       mat[5,3]=mat[5,3]+1
    ENDIF
    IF ((parentes_c)=6) then
       mat[5,5]=mat[5,5]+1
    ENDIF
    SKIP IN a
ENDDO
FOR j= 1 TO nc
   FOR i=1 TO nf-1
      mat[nf,j]=mat[nf,j]+mat[i,j]
  ENDFOR
ENDFOR
FOR i= 1 TO nf
   FOR j=2 TO nc STEP 2
      mat[i,j]=mat[i,j-1]/mat[nf,j-1]*100
  ENDFOR
ENDFOR
   SELE C
   USE "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\aa"
      FOR B=1 TO NF
         APPEND BLANK
      ENDFOR
SELECT c
   FOR REN=1 TO NF
      GO REN IN c
      FOR COL=1 TO NC
         CAMPO=FIELD(COL,'c')
         REPLACE c.&CAMPO. WITH MAT[REN,COL]
```

```
      ENDFOR
   ENDFOR
SELECT c
   FOR REN=1 TO RECCOUNT('c')
      GO REN IN c
      FOR COL=1 TO FCOUNT('c')
         CC=LTRIM(STR(COL))
         IF COL=1
            ? nom[Ren]
            ?? TRANSFORM( c.C&CC., '    999,999')
         ELSE
            if(MOD(col,2)<>0) then
               ?? TRANSFORM( c.C&CC., '    999,999')
            ELSE
               ?? TRANSFORM( c.C&CC., '    999.99')
            endif
         ENDIF
      ENDFOR
   ENDFOR
?
CLOSE database
************************************* DECLARACION E INICIALIZACION DE VARIABLES GLOBALES
STORE 9 TO NF                 && NF: NUMERO DE RENGLONES
STORE 6 TO NC                  && NC: NUMERO DE COLUMNAS
DECLARE MAT(NF,NC)
STORE 0 TO MAT
DECLARE nom(09)
nom[01]='code 1  '
nom[02]='code 2  '
nom[03]='code 3  '
nom[04]='code 4  '
nom[05]='code 5  '
nom[06]='code 6  '
nom[07]='code 7  '
nom[08]='code 8  '
nom[09]='total   '
create table Estruct.dbf;
(field_name C(10),field_type C(1),field_len N(3,0),field_dec N(3,0))
FOR B=1 TO NC
   CAMPO='C'+LTRIM(STR(B))
   append blank
   replace field_name with CAMPO
   replace field_type with 'N'
   replace field_len  with 10
   replace field_dec  with 2
ENDFOR
create "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\aa" from Estruct.dbf
USE
ERASE Estruct.dbf
SELECT a
USE "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\multii1m4.dbf" SHARED
************************************************************edo_cony
?'edo_cony_c          Observed            Imputed             Combined     '
?'                 Freq    Per         Freq    Per          Freq     Per   '
?'-------------------------------------------------------------------------'
R1=1
DO WHILE NOT EOF()
    IF (edo_cony_f)=1 AND edo_cony_c=1 then
       mat[1,1]=mat[1,1]+1
    ENDIF
    IF (edo_cony_f)=0 AND edo_cony_c=1 then
       mat[1,3]=mat[1,3]+1
    ENDIF
    IF (edo_cony_c)=1 then
       mat[1,5]=mat[1,5]+1
    ENDIF
    IF (edo_cony_f)=1 AND edo_cony_c=2 then
       mat[2,1]=mat[2,1]+1
    ENDIF
    IF (edo_cony_f)=0 AND edo_cony_c=2 then
       mat[2,3]=mat[2,3]+1
```

```
    ENDIF
    IF (edo_cony_c)=2 then
       mat[2,5]=mat[2,5]+1
    ENDIF
    IF (edo_cony_f)=1 AND edo_cony_c=3 then
       mat[3,1]=mat[3,1]+1
    ENDIF
    IF (edo_cony_f)=0 AND edo_cony_c=3 then
       mat[3,3]=mat[3,3]+1
    ENDIF
    IF (edo_cony_c)=3 then
       mat[3,5]=mat[3,5]+1
    ENDIF
    IF (edo_cony_f)=1 AND edo_cony_c=4 then
       mat[4,1]=mat[4,1]+1
    ENDIF
    IF (edo_cony_f)=0 AND edo_cony_c=4 then
       mat[4,3]=mat[4,3]+1
    ENDIF
    IF (edo_cony_c)=4 then
       mat[4,5]=mat[4,5]+1
    ENDIF
    IF (edo_cony_f)=1 AND edo_cony_c=5 then
       mat[5,1]=mat[5,1]+1
    ENDIF
    IF (edo_cony_f)=0 AND edo_cony_c=5 then
       mat[5,3]=mat[5,3]+1
    ENDIF
    IF (edo_cony_c)=5 then
       mat[5,5]=mat[5,5]+1
    ENDIF
    IF (edo_cony_f)=1 AND edo_cony_c=6 then
       mat[6,1]=mat[6,1]+1
    ENDIF
    IF (edo_cony_f)=0 AND edo_cony_c=6 then
       mat[6,3]=mat[6,3]+1
    ENDIF
    IF (edo_cony_c)=6 then
       mat[6,5]=mat[6,5]+1
    ENDIF
    IF (edo_cony_f)=1 AND edo_cony_c=7 then
       mat[7,1]=mat[7,1]+1
    ENDIF
    IF (edo_cony_f)=0 AND edo_cony_c=7 then
       mat[7,3]=mat[7,3]+1
    ENDIF
    IF (edo_cony_c)=7 then
       mat[7,5]=mat[7,5]+1
    ENDIF
    IF (edo_cony_f)=1 AND edo_cony_c=8 then
       mat[8,1]=mat[8,1]+1
    ENDIF
    IF (edo_cony_f)=0 AND edo_cony_c=8 then
       mat[8,3]=mat[8,3]+1
    ENDIF
    IF (edo_cony_c)=8 then
       mat[8,5]=mat[8,5]+1
    ENDIF
    SKIP IN a
ENDDO
FOR j= 1 TO nc
   FOR i=1 TO nf-1
      mat[nf,j]=mat[nf,j]+mat[i,j]
  ENDFOR
ENDFOR
FOR i= 1 TO nf
   FOR j=2 TO nc STEP 2
      mat[i,j]=mat[i,j-1]/mat[nf,j-1]*100
  ENDFOR
ENDFOR
   SELE C
```

```
   USE "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\aa"
      FOR B=1 TO NF
         APPEND BLANK
      ENDFOR
SELECT c
   FOR REN=1 TO NF
      GO REN IN c
      FOR COL=1 TO NC
         CAMPO=FIELD(COL,'c')
         REPLACE c.&CAMPO. WITH MAT[REN,COL]
      ENDFOR
   ENDFOR
SELECT c
   FOR REN=1 TO RECCOUNT('c')
      GO REN IN c
      FOR COL=1 TO FCOUNT('c')
         CC=LTRIM(STR(COL))
         IF COL=1
            ? nom[Ren]
            ?? TRANSFORM( c.C&CC., '    999,999')
         ELSE
            if(MOD(col,2)<>0) then
               ?? TRANSFORM( c.C&CC., '    999,999')
            ELSE
               ?? TRANSFORM( c.C&CC., '    999.99')
            endif
         ENDIF
      ENDFOR
   ENDFOR
?
CLOSE database
*****************************************************************SELECT c
USE "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\multii1m4.dbf" SHARED
CALCULATE std(c.log_ing_cm) FOR c.log_ing_f=0 TO sdvlog_ingi
CALCULATE std(c.log_ing_cm) FOR c.log_ing_f=1 TO sdvlog_ingo
CALCULATE std(c.log_ing_cm) FOR c.log_ing_f=0 OR c.log_ing_f=1 TO sdvlog_ingc
SELECT log_ing_f,cnt(*) as logal ,MIN(log_ing_cm) as mi,MAX(log_ing_cm) as
ma,avg(log_ing_cm) as me FROM "d:\dario\completa(inegi)\10-
14octubre2008(m5iter2)\m4\multii1m4.dbf" inTO TABLE "d:\dario\completa(inegi)\10-
14octubre2008(m5iter2)\m4\aa.dbf" GROUP BY log_ing_f WHERE log_ing_f<>9
CLOSE all
SELECT a
DECLARE nom(3)
nom[01]='imputed    '
nom[02]='observed   '
nom[03]='combined   '
USE "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\aa.dbf" SHARED
?'log_ing_cm     nnumber               minimun               maximun                 mean
stddev'
?'-------------------------------------------------------------------------------------------'
R1=1
DO WHILE NOT EOF()
    ?NOM[R1]
    R1=R1+1
    ??a.logal
    ??ROUND(a.mi,2)
    ??ROUND(a.ma,2)
    ??ROUND(A.me,2)
    IF r1=2 then
       ??ROUND(sdvlog_ingi,2)
    ELSE
       ??ROUND(sdvlog_ingo,2)
    endif
    SKIP IN a
ENDDO
SELECT cnt(*) as logal,MIN(log_ing_cm) as mi,MAX(log_ing_cm) as ma,avg(log_ing_cm) as me
FROM "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\multii1m4.dbf" inTO TABLE
"d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\bb.dbf" where log_ing_f<>9
CLOSE all
SELECT a
USE "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\bb.dbf" SHARED
```

```
?NOM[3]
DO WHILE NOT EOF()
    ??a.logal
    ??ROUND(a.mi,2)
    ??ROUND(a.ma,2)
    ??ROUND(a.me,2)
    ??ROUND(sdvlog_ingc,2)
    SKIP IN a
ENDDO
CLOSE database
?
************************************************************? 'tiempo FINAL:' + time()
?'FECHA FINAL',DATE()
SET PRINTER OFF
SET PRINTER TO PRN
USE
ERASE "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\aa.DBF"
ERASE "d:\dario\completa(inegi)\10-14octubre2008(m5iter2)\m4\bb.DBF"
CLOSE ALL
DELETE FILE est_llenainegi.BAK
```

Anexo I

Reporte estadístico alterno (prueba final, Regresión Secuencial e Imputación Múltiple)

MULTIPLE 1

edad_cm	nnumber	minimun	maximun	mean	stddev
imputed	2302	12.00	121.00	49.89	20.39
observed	659863	12.00	130.00	33.13	16.88
combined	662165	12.00	130.00	33.19	16.92

motriz_cm	Observed		Imputed		Combined	
	Freq	Per	Freq	Per	Freq	Per
code 1	7,504	0.81	47	1.27	7,551	0.81
code 2	919,131	99.19	3,644	98.73	922,775	99.19
total	926,635	100.00	3,691	100.00	930,326	100.00

niv_esc_cm	Observed		Imputed		Combined	
	Freq	Per	Freq	Per	Freq	Per
code 0	43,974	5.39	440	5.60	44,414	5.39
code 1	40,430	4.96	156	1.98	40,586	4.93
code 2	373,598	45.79	3,781	48.11	377,379	45.81
code 3	175,049	21.46	569	7.24	175,618	21.32
code 4	76,335	9.36	162	2.06	76,497	9.29
code 5	3,433	0.42	5	0.06	3,438	0.42
code 6	35,792	4.39	103	1.31	35,895	4.36
code 7	62,848	7.70	77	0.98	62,925	7.64
code 8	4,388	0.54	2,566	32.65	6,954	0.84
total	815,847	100.00	7,859	100.00	823,706	100.00

ocupac_cm	Observed		Imputed		Combined	
	Freq	Per	Freq	Per	Freq	Per
code 11	11,871	3.66	312	3.80	12,183	3.67
code 12	11,239	3.47	342	4.17	11,581	3.49
code 13	14,434	4.45	325	3.96	14,759	4.44
code 14	2,653	0.82	352	4.29	3,005	0.90
code 21	7,827	2.42	351	4.28	8,178	2.46
code 41	22,637	6.98	351	4.28	22,988	6.92
code 51	8,401	2.59	358	4.36	8,759	2.64
code 52	54,827	16.92	1,135	13.83	55,962	16.84
code 53	33,678	10.39	605	7.37	34,283	10.32
code 54	15,251	4.71	445	5.42	15,696	4.72
code 55	16,094	4.97	475	5.79	16,569	4.99
code 61	9,620	2.97	357	4.35	9,977	3.00
code 62	23,885	7.37	400	4.87	24,285	7.31
code 71	45,932	14.17	657	8.00	46,589	14.02
code 72	5,683	1.75	335	4.08	6,018	1.81
code 81	21,522	6.64	544	6.63	22,066	6.64
code 82	11,246	3.47	431	5.25	11,677	3.51
code 83	7,293	2.25	434	5.29	7,727	2.33
total	324,093	100.00	8,209	100.00	332,302	100.00

pos_tra_cm	Observed		Imputed		Combined	
	Freq	Per	Freq	Per	Freq	Per
code 1	234,740	72.49	2,104	24.79	236,844	71.27
code 2	19,299	5.96	1,072	12.63	20,371	6.13
code 3	10,551	3.26	135	1.59	10,686	3.22
code 4	51,120	15.79	5,090	59.97	56,210	16.92
code 5	8,105	2.50	86	1.01	8,191	2.46
total	323,815	100.00	8,487	100.00	332,302	100.00

vivos_cm	nnumber	minimun	maximun	mean	stddev
imputed	6003	0.00	16.00	2.48	2.20
observed	342481	0.00	25.00	2.65	3.37
combined	348484	0.00	25.00	2.65	3.35

parentes_c	Observed		Imputed		Combined	
	Freq	Per	Freq	Per	Freq	Per
code 2	161,488	22.20	0	0.00	161,488	22.16
code 3	481,942	66.27	491	33.63	482,433	66.20
code 4	1,009	0.14	496	33.97	1,505	0.21
code 5	2,472	0.34	1	0.07	2,473	0.34
code 6	80,352	11.05	472	32.33	80,824	11.09
total	727,263	100.00	1,460	100.00	728,723	100.00

edo_cony_c	Observed		Imputed		Combined	
	Freq	Per	Freq	Per	Freq	Per
code 1	28,290	4.28	12	0.74	28,302	4.27
code 2	10,848	1.64	92	5.66	10,940	1.65
code 3	6,406	0.97	17	1.05	6,423	0.97
code 4	24,193	3.66	63	3.87	24,256	3.66
code 5	34,389	5.21	32	1.97	34,421	5.20
code 6	3,873	0.59	0	0.00	3,873	0.58
code 7	295,249	44.70	397	24.42	295,646	44.65
code 8	257,291	38.95	1,013	62.30	258,304	39.01
total	660,539	100.00	1,626	100.00	662,165	100.00

log_ing_cm	nnumber	minimun	maximun	mean	stddev
imputed	36854	2.05	13.38	7.48	1.42
observed	296531	0.00	13.71	7.36	1.92
combined	333385	0.00	13.71	7.37	1.87

MULTIPLE 2

edad_cm	nnumber	minimun	maximun	mean	stddev
imputed	2302	12.00	119.00	49.33	20.77
observed	659863	12.00	130.00	33.13	16.88
combined	662165	12.00	130.00	33.19	16.92

motriz_cm	Observed		Imputed		Combined	
	Freq	Per	Freq	Per	Freq	Per
code 1	7,504	0.81	48	1.30	7,552	0.81
code 2	919,131	99.19	3,643	98.70	922,774	99.19
total	926,635	100.00	3,691	100.00	930,326	100.00

niv_esc_cm	Observed		Imputed		Combined	
	Freq	Per	Freq	Per	Freq	Per
code 0	43,974	5.39	683	8.69	44,657	5.42
code 1	40,430	4.96	15	0.19	40,445	4.91
code 2	373,598	45.79	3,745	47.65	377,343	45.81
code 3	175,049	21.46	535	6.81	175,584	21.32
code 4	76,335	9.36	163	2.07	76,498	9.29
code 5	3,433	0.42	4	0.05	3,437	0.42
code 6	35,792	4.39	99	1.26	35,891	4.36
code 7	62,848	7.70	50	0.64	62,898	7.64
code 8	4,388	0.54	2,565	32.64	6,953	0.84
total	815,847	100.00	7,859	100.00	823,706	100.00

ocupac_cm	Observed		Imputed		Combined	
	Freq	Per	Freq	Per	Freq	Per
code 11	11,871	3.66	1,072	13.06	12,943	3.89
code 12	11,239	3.47	36	0.44	11,275	3.39
code 13	14,434	4.45	20	0.24	14,454	4.35
code 14	2,653	0.82	23	0.28	2,676	0.81
code 21	7,827	2.42	604	7.36	8,431	2.54
code 41	22,637	6.98	98	1.19	22,735	6.84
code 51	8,401	2.59	93	1.13	8,494	2.56
code 52	54,827	16.92	2,821	34.36	57,648	17.35
code 53	33,678	10.39	350	4.26	34,028	10.24
code 54	15,251	4.71	91	1.11	15,342	4.62
code 55	16,094	4.97	492	5.99	16,586	4.99
code 61	9,620	2.97	31	0.38	9,651	2.90
code 62	23,885	7.37	74	0.90	23,959	7.21
code 71	45,932	14.17	267	3.25	46,199	13.90
code 72	5,683	1.75	35	0.43	5,718	1.72
code 81	21,522	6.64	394	4.80	21,916	6.60
code 82	11,246	3.47	1,546	18.83	12,792	3.85
code 83	7,293	2.25	162	1.97	7,455	2.24
total	324,093	100.00	8,209	100.00	332,302	100.00

pos_tra_cm	Observed		Imputed		Combined	
	Freq	Per	Freq	Per	Freq	Per
code 1	234,740	72.49	6,848	80.69	241,588	72.70
code 2	19,299	5.96	206	2.43	19,505	5.87
code 3	10,551	3.26	246	2.90	10,797	3.25

code 4	51,120	15.79	1,175	13.84	52,295	15.74
code 5	8,105	2.50	12	0.14	8,117	2.44
total	323,815	100.00	8,487	100.00	332,302	100.00

vivos_cm	nnumber	minimun	maximun	mean	stddev
imputed	6003	0.00	15.00	2.49	2.20
observed	342481	0.00	25.00	2.65	3.37
combined	348484	0.00	25.00	2.65	3.35

parentes_c	Observed		Imputed		Combined	
	Freq	Per	Freq	Per	Freq	Per
code 2	161,488	22.20	0	0.00	161,488	22.16
code 3	481,942	66.27	504	34.52	482,446	66.20
code 4	1,009	0.14	521	35.68	1,530	0.21
code 5	2,472	0.34	1	0.07	2,473	0.34
code 6	80,352	11.05	434	29.73	80,786	11.09
total	727,263	100.00	1,460	100.00	728,723	100.00

edo_cony_c	Observed		Imputed		Combined	
	Freq	Per	Freq	Per	Freq	Per
code 1	28,290	4.28	17	1.05	28,307	4.27
code 2	10,848	1.64	15	0.92	10,863	1.64
code 3	6,406	0.97	151	9.29	6,557	0.99
code 4	24,193	3.66	60	3.69	24,253	3.66
code 5	34,389	5.21	10	0.62	34,399	5.19
code 6	3,873	0.59	0	0.00	3,873	0.58
code 7	295,249	44.70	323	19.86	295,572	44.64
code 8	257,291	38.95	1,050	64.58	258,341	39.01
total	660,539	100.00	1,626	100.00	662,165	100.00

log_ing_cm	nnumber	minimun	maximun	mean	stddev
imputed	36854	1.45	13.30	7.48	1.40
observed	296531	0.00	13.71	7.36	1.92
combined	333385	0.00	13.71	7.37	1.87

MULTIPLE 3

edad_cm	nnumber	minimun	maximun	mean	stddev
imputed	2302	12.00	125.00	49.71	20.32
observed	659863	12.00	130.00	33.13	16.88
combined	662165	12.00	130.00	33.19	16.92

motriz_cm	Observed		Imputed		Combined	
	Freq	Per	Freq	Per	Freq	Per
code 1	7,504	0.81	43	1.17	7,547	0.81
code 2	919,131	99.19	3,648	98.84	922,779	99.19
total	926,635	100.00	3,691	100.00	930,326	100.00

niv_esc_cm	Observed		Imputed		Combined	
	Freq	Per	Freq	Per	Freq	Per
code 0	43,974	5.39	820	10.43	44,794	5.44
code 1	40,430	4.96	1,071	13.63	41,501	5.04
code 2	373,598	45.79	3,730	47.46	377,328	45.81
code 3	175,049	21.46	518	6.59	175,567	21.31
code 4	76,335	9.36	178	2.26	76,513	9.29
code 5	3,433	0.42	8	0.10	3,441	0.42
code 6	35,792	4.39	108	1.37	35,900	4.36
code 7	62,848	7.70	59	0.75	62,907	7.64
code 8	4,388	0.54	1,367	17.39	5,755	0.70
total	815,847	100.00	7,859	100.00	823,706	100.00

ocupac_cm	Observed		Imputed		Combined	
	Freq	Per	Freq	Per	Freq	Per
code 11	11,871	3.66	4,367	53.20	16,238	4.89
code 12	11,239	3.47	39	0.48	11,278	3.39
code 13	14,434	4.45	71	0.86	14,505	4.37
code 14	2,653	0.82	129	1.57	2,782	0.84
code 21	7,827	2.42	33	0.40	7,860	2.37
code 41	22,637	6.98	32	0.39	22,669	6.82
code 51	8,401	2.59	98	1.19	8,499	2.56
code 52	54,827	16.92	975	11.88	55,802	16.79

```
code 53        33,678      10.39       1,007      12.27      34,685      10.44
code 54        15,251       4.71         151       1.84      15,402       4.63
code 55        16,094       4.97         457       5.57      16,551       4.98
code 61         9,620       2.97          33       0.40       9,653       2.90
code 62        23,885       7.37          99       1.21      23,984       7.22
code 71        45,932      14.17         255       3.11      46,187      13.90
code 72         5,683       1.75          10       0.12       5,693       1.71
code 81        21,522       6.64         203       2.47      21,725       6.54
code 82        11,246       3.47          90       1.10      11,336       3.41
code 83         7,293       2.25         160       1.95       7,453       2.24
total         324,093     100.00       8,209     100.00     332,302     100.00
pos_tra_cm           Observed            Imputed             Combined
                   Freq     Per          Freq     Per          Freq     Per
-------------------------------------------------------------------------------
code 1       234,740      72.49       2,012      23.71     236,752      71.25
code 2        19,299       5.96         128       1.51      19,427       5.85
code 3        10,551       3.26         114       1.34      10,665       3.21
code 4        51,120      15.79       6,063      71.44      57,183      17.21
code 5         8,105       2.50         170       2.00       8,275       2.49
total        323,815     100.00       8,487     100.00     332,302     100.00
vivos_cm    nnumber  minimun    maximun     mean     stddev
-----------------------------------------------------------
imputed      6003      0.00       15.00      2.50      2.20
observed   342481      0.00       25.00      2.65      3.37
combined   348484      0.00       25.00      2.65      3.35
parentes_c           Observed            Imputed              Combined
                   Freq     Per          Freq     Per          Freq     Per
-------------------------------------------------------------------------------
code 2       161,488      22.20           1       0.07     161,489      22.16
code 3       481,942      66.27         502      34.38     482,444      66.20
code 4         1,009       0.14         912      62.47       1,921       0.26
code 5         2,472       0.34           0       0.00       2,472       0.34
code 6        80,352      11.05          45       3.08      80,397      11.03
total        727,263     100.00       1,460     100.00     728,723     100.00
edo_cony_c           Observed            Imputed              Combined
                   Freq     Per          Freq     Per          Freq     Per
-------------------------------------------------------------------------------
code 1        28,290       4.28          10       0.62      28,300       4.27
code 2        10,848       1.64          67       4.12      10,915       1.65
code 3         6,406       0.97          82       5.04       6,488       0.98
code 4        24,193       3.66          51       3.14      24,244       3.66
code 5        34,389       5.21          13       0.80      34,402       5.20
code 6         3,873       0.59           0       0.00       3,873       0.58
code 7       295,249      44.70         352      21.65     295,601      44.64
code 8       257,291      38.95       1,051      64.64     258,342      39.01
total        660,539     100.00       1,626     100.00     662,165     100.00
log_ing_cm  nnumber    minimun    maximun    mean    stddev
-----------------------------------------------------------
imputed      36854       1.95       13.29     7.51     1.39
observed    296531       0.00       13.71     7.36     1.92
combined    333385       0.00       13.71     7.38     1.87

MULTIPLE 4
edad_cm     nnumber    minimun     maximun     mean     stddev
-----------------------------------------------------------
imputed      2302       12.00       127.00     49.27     20.37
observed   659863       12.00       130.00     33.13     16.88
combined   662165       12.00       130.00     33.19     16.92
motriz_cm            Observed            Imputed              Combined
                   Freq     Per          Freq     Per          Freq     Per
-------------------------------------------------------------------------------
code 1         7,504       0.81          48       1.30       7,552       0.81
code 2       919,131      99.19       3,643      98.70     922,774      99.19
total        926,635     100.00       3,691     100.00     930,326     100.00
niv_esc_cm           Observed            Imputed              Combined
                   Freq     Per          Freq     Per          Freq     Per
-------------------------------------------------------------------------------
code 0        43,974       5.39         659       8.39      44,633       5.42
code 1        40,430       4.96          41       0.52      40,471       4.91
code 2       373,598      45.79       3,722      47.36     377,320      45.81
code 3       175,049      21.46         514       6.54     175,563      21.31
```

code 4	76,335	9.36	170	2.16	76,505	9.29
code 5	3,433	0.42	3	0.04	3,436	0.42
code 6	35,792	4.39	117	1.49	35,909	4.36
code 7	62,848	7.70	79	1.01	62,927	7.64
code 8	4,388	0.54	2,554	32.50	6,942	0.84
total	815,847	100.00	7,859	100.00	823,706	100.00

ocupac_cm	Observed		Imputed		Combined	
	Freq	Per	Freq	Per	Freq	Per
code 11	11,871	3.66	326	3.97	12,197	3.67
code 12	11,239	3.47	345	4.20	11,584	3.49
code 13	14,434	4.45	333	4.06	14,767	4.44
code 14	2,653	0.82	344	4.19	2,997	0.90
code 21	7,827	2.42	388	4.73	8,215	2.47
code 41	22,637	6.98	402	4.90	23,039	6.93
code 51	8,401	2.59	409	4.98	8,810	2.65
code 52	54,827	16.92	1,087	13.24	55,914	16.83
code 53	33,678	10.39	618	7.53	34,296	10.32
code 54	15,251	4.71	391	4.76	15,642	4.71
code 55	16,094	4.97	434	5.29	16,528	4.97
code 61	9,620	2.97	336	4.09	9,956	3.00
code 62	23,885	7.37	382	4.65	24,267	7.30
code 71	45,932	14.17	674	8.21	46,606	14.03
code 72	5,683	1.75	339	4.13	6,022	1.81
code 81	21,522	6.64	542	6.60	22,064	6.64
code 82	11,246	3.47	421	5.13	11,667	3.51
code 83	7,293	2.25	438	5.34	7,731	2.33
total	324,093	100.00	8,209	100.00	332,302	100.00

pos_tra_cm	Observed		Imputed		Combined	
	Freq	Per	Freq	Per	Freq	Per
code 1	234,740	72.49	4,261	50.21	239,001	71.92
code 2	19,299	5.96	1,843	21.72	21,142	6.36
code 3	10,551	3.26	159	1.87	10,710	3.22
code 4	51,120	15.79	2,210	26.04	53,330	16.05
code 5	8,105	2.50	14	0.16	8,119	2.44
total	323,815	100.00	8,487	100.00	332,302	100.00

vivos_cm	nnumber	minimun	maximun	mean	stddev
imputed	6003	0.00	15.00	2.46	2.19
observed	342481	0.00	25.00	2.65	3.37
combined	348484	0.00	25.00	2.65	3.35

parentes_c	Observed		Imputed		Combined	
	Freq	Per	Freq	Per	Freq	Per
code 2	161,488	22.20	0	0.00	161,488	22.16
code 3	481,942	66.27	501	34.32	482,443	66.20
code 4	1,009	0.14	482	33.01	1,491	0.20
code 5	2,472	0.34	1	0.07	2,473	0.34
code 6	80,352	11.05	476	32.60	80,828	11.09
total	727,263	100.00	1,460	100.00	728,723	100.00

edo_cony_c	Observed		Imputed		Combined	
	Freq	Per	Freq	Per	Freq	Per
code 1	28,290	4.28	11	0.68	28,301	4.27
code 2	10,848	1.64	54	3.32	10,902	1.65
code 3	6,406	0.97	112	6.89	6,518	0.98
code 4	24,193	3.66	78	4.80	24,271	3.67
code 5	34,389	5.21	18	1.11	34,407	5.20
code 6	3,873	0.59	0	0.00	3,873	0.58
code 7	295,249	44.70	324	19.93	295,573	44.64
code 8	257,291	38.95	1,029	63.28	258,320	39.01
total	660,539	100.00	1,626	100.00	662,165	100.00

log_ing_cm	nnumber	minimun	maximun	mean	stddev
imputed	36854	1.71	13.24	7.47	1.41
observed	296531	0.00	13.71	7.36	1.92
combined	333385	0.00	13.71	7.37	1.87

MULTIPLE 5

edad_cm	nnumber	minimun	maximun	mean	stddev

```
-------------------------------------------------------------
imputed     2302     12.00     119.00    49.69     20.23
observed  659863     12.00     130.00    33.13     16.88
combined  662165     12.00     130.00    33.19     16.92
motriz_cm          Observed            Imputed              Combined
                 Freq    Per         Freq    Per          Freq     Per
-------------------------------------------------------------------------
code 1          7,504    0.81           39    1.06       7,543    0.81
code 2        919,131   99.19        3,652   98.94     922,783   99.19
total         926,635  100.00        3,691  100.00     930,326  100.00
niv_esc_cm         Observed            Imputed              Combined
                 Freq    Per         Freq    Per          Freq     Per
-------------------------------------------------------------------------
code 0         43,974    5.39        1,859   23.65      45,833    5.56
code 1         40,430    4.96          104    1.32      40,534    4.92
code 2        373,598   45.79        3,858   49.09     377,456   45.82
code 3        175,049   21.46          676    8.60     175,725   21.33
code 4         76,335    9.36          782    9.95      77,117    9.36
code 5          3,433    0.42            3    0.04       3,436    0.42
code 6         35,792    4.39          517    6.58      36,309    4.41
code 7         62,848    7.70           51    0.65      62,899    7.64
code 8          4,388    0.54            9    0.11       4,397    0.53
total         815,847  100.00        7,859  100.00     823,706  100.00
ocupac_cm          Observed            Imputed              Combined
                 Freq    Per         Freq    Per          Freq     Per
-------------------------------------------------------------------------
code 11        11,871    3.66        1,020   12.43      12,891    3.88
code 12        11,239    3.47           28    0.34      11,267    3.39
code 13        14,434    4.45        1,053   12.83      15,487    4.66
code 14         2,653    0.82        1,003   12.22       3,656    1.10
code 21         7,827    2.42           19    0.23       7,846    2.36
code 41        22,637    6.98           32    0.39      22,669    6.82
code 51         8,401    2.59           94    1.15       8,495    2.56
code 52        54,827   16.92        1,019   12.41      55,846   16.81
code 53        33,678   10.39          332    4.04      34,010   10.23
code 54        15,251    4.71          168    2.05      15,419    4.64
code 55        16,094    4.97          474    5.77      16,568    4.99
code 61         9,620    2.97           37    0.45       9,657    2.91
code 62        23,885    7.37        1,136   13.84      25,021    7.53
code 71        45,932   14.17          256    3.12      46,188   13.90
code 72         5,683    1.75        1,066   12.99       6,749    2.03
code 81        21,522    6.64          234    2.85      21,756    6.55
code 82        11,246    3.47           81    0.99      11,327    3.41
code 83         7,293    2.25          157    1.91       7,450    2.24
total         324,093  100.00        8,209  100.00     332,302  100.00
pos_tra_cm         Observed            Imputed              Combined
                 Freq    Per         Freq    Per          Freq     Per
-------------------------------------------------------------------------
code 1        234,740   72.49        7,352   86.63     242,092   72.85
code 2         19,299    5.96          145    1.71      19,444    5.85
code 3         10,551    3.26          283    3.33      10,834    3.26
code 4         51,120   15.79          694    8.18      51,814   15.59
code 5          8,105    2.50           13    0.15       8,118    2.44
total         323,815  100.00        8,487  100.00     332,302  100.00
vivos_cm   nnumber   minimun    maximun   mean    stddev
-------------------------------------------------------------
imputed      6003      0.00       14.00    2.46     2.16
observed   342481      0.00       25.00    2.65     3.37
combined   348484      0.00       25.00    2.65     3.35
parentes_c         Observed            Imputed              Combined
                 Freq    Per         Freq    Per          Freq     Per
-------------------------------------------------------------------------
code 2        161,488   22.20           1    0.07     161,489   22.16
code 3        481,942   66.27         500   34.25     482,442   66.20
code 4          1,009    0.14         914   62.60       1,923    0.26
code 5          2,472    0.34           1    0.07       2,473    0.34
code 6         80,352   11.05          44    3.01      80,396   11.03
total         727,263  100.00        1,460  100.00     728,723  100.00
edo_cony_c         Observed            Imputed              Combined
                 Freq    Per         Freq    Per          Freq     Per
-------------------------------------------------------------------------
```

code 1	28,290	4.28	13	0.80	28,303	4.27
code 2	10,848	1.64	33	2.03	10,881	1.64
code 3	6,406	0.97	9	0.55	6,415	0.97
code 4	24,193	3.66	52	3.20	24,245	3.66
code 5	34,389	5.21	18	1.11	34,407	5.20
code 6	3,873	0.59	4	0.25	3,877	0.59
code 7	295,249	44.70	447	27.49	295,696	44.66
code 8	257,291	38.95	1,050	64.58	258,341	39.01
total	660,539	100.00	1,626	100.00	662,165	100.00

log_ing_cm	nnumber	minimun	maximun	mean	stddev
imputed	36854	1.45	12.91	7.47	1.41
observed	296531	0.00	13.71	7.36	1.92
combined	333385	0.00	13.71	7.37	1.87

Bibliografía

Allison, P. D. (2000), Multiple Imputation for Missing Data: A Cautionary Tale, University of Pennsylvania.

BID (2004). Imputación de bases de datos en las encuestas en hogares. Imputación Múltiple y los missing de ingresos en las encuestas en hogares (presentación).

Brick J. M. y Kalton G. (1996). Handling missing data in Survey research. Statistical Methods in Medical Research 5, 215-238.

CDC.Vital and Health Statistics (2006) series 2, Number 142. National Survey of Family Growth, Cycle 6: Sampling, Design, Weighting, Imputation and Variance Estimation.

Census Bureau Standard (2006). Definitions for Survey and Census Metadata Version 1.4. Supporting Document An Organization of Metadata. http://www.census.gov/quality/S10-2_v1.4_Alphabetized_Metadata.htm.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 39, 1–22.

Garson G. D. (2008). Data Imputation for Missing Values.

Horton N. J. y Kleinman K. P. (2007). Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. The American Statistician 61 (1), 79-90.

Howell D. C. (2007). Treatment of Missing Data.

INEGI (2000). Los tabulados de la Muestra Censal del XII Censo General de Población y Vivienda 2000.

INEGI (2001). Documentación metodológica para la validación automática. Características de las personas (Características económicas).

INEGI (2001). Impacto de los criterios y tratamientos de la validación en la información del XII Censo General de Población y Vivienda 2000. Ingreso por trabajo.

INEGI (2003). Síntesis Metodológica del XII Censo General de Población y Vivienda 2000.

Kalton G. y Kasprzyk D. (1986). The Treatment of Missing Survey Data. Survey Methodology 12 (1), 1-16.

Laaksonen S. (2000). *"How to Find de Best Imputation Technique? Tests with Various Methods."* Statistics Finland.

Lepkowski J.M. (2006). Weighting and Imputation: A US Case Study. Michigan Program in Survey Methodology, CIMAT Seminar.

Little, R. J. A., and Rubin, D. B. (1987), Statistical Analysis With Missing Data, New York; Wiley.

Maindonald J. H. (2008). Using R for Data Analysis and Graphics Introduction, Code and Commentary. Centre for Mathematics and Its Applications, Australian National University.

Medina F. y Galván M. (2007). Imputación de datos: teoría y práctica. Estudios estadísticos y prospectivos. CEPAL (División de Estadística y Proyecciones Económicas).

Platek R. (1986). Metodología y Tratamiento de la NO RESPUESTA. Seminario Internacional de Estadística EUSTAT (cuaderno 10).

Puerta G. A. (2002). Imputación basada en árboles de clasificación EUSTAT.

Raghunathan, T.E., Solenberger P. W. y Van Hoewyk J. (2002). IVEware: Imputation and Variance Estimation Software, User Guide.

Raghunathan, T.E., Solenberger P.W., Van Hoewyk J. y Lepkowski J.M. (2001a). A multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. Survey Methodology, Program Institute for Social Research, university of Michigan.

Raghunathan, T.E., Solenberger P.W., Van Hoewyk J. y Lepkowski J.M. (2001b). Sequential Regression Imputation for Survey Data. Survey Methodology.

Raghunathan, T.E. (2007). Statistical Analysis with Incomplete Data. The 60 Annual Summer Institute. Survey Research Center, Institute for Social Research university of Michigan (presentation).

Scheffer J. (2002). Dealing with Missing Data. Research Letters in the Information and Mathematical Sciences 3, 153-160. http://www.massey.ac.nz/~wwiims/research/letters/volume3number1/scheffer_2.pdf.

Srebotnjak T. (2002). Sequential Regression – A Method for Multiple Imputations of Missing Data. United Nations, Statistics Division WGCI Seminar.

United Nations (2001). Handbook on Population and Housing Census Editing. Department of Economic and Social Affairs series F no. 82 127-130.