ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

IMPUTACIÓN ESTADÍSTICA: UNA APLICACIÓN AL SISTEMA NACIONAL INTERCONECTADO DEL ECUADOR

TESIS PREVIO A LA OBTENCIÓN DEL GRADO DE MAGÍSTER EN ESTADÍSTICA APLICADA

ADRIANA JANET PACHECO TOSCANO
apacheco@cenace.org.ec

DIRECTOR: DR. HOLGER CAPA SANTOS
hcapa@epn.edu.ec

Quito, abril 2008

DECLARACIÓN

Yo Adriana Janet Pacheco Toscano, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentada para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

ADRIANA JANET PACHECO TOSCANO

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por Adriana Janet Pacheco Toscano, bajo mi supervisión.

_____

Dr. Holger Capa Santos
DIRECTOR DE TESIS

DEDICATORIA

A mis hijos Dianita y Martín,
mis grandes tesoros.

.

AGRADECIMIENTOS

A mi esposo por su paciencia y cariño

A mis padres y hermano por su apoyo incondicional

A la Corporación CENACE por el apoyo brindado

Al Doctor Holger Capa por su paciencia y estimulo para culminar el presente trabajo

CONTENIDO

RESUMEN

En el desarrollo teórico de la mayoría de técnicas y modelos estadísticos no se tienen en cuenta algunas cuestiones que surgen en su aplicación práctica, en concreto, un problema al que con seguridad se ha enfrentado cualquier analista de datos es el de los datos faltantes, también denominados perdidos o incompletos. Disponer de un archivo de datos completos es ideal, pero aplicar métodos de imputación inapropiados para lograrlo, puede generar más problemas de los que se resuelve. Durante las últimas décadas se han desarrollado procedimientos que tienen mejores propiedades estadísticas que las opciones tradicionales como la eliminación de los datos, el método de las medias y el hot-deck, tal es el caso de los algoritmos de imputación múltiple, los que se pueden aplicar utilizando paquetes comerciales. Existen implicaciones en el análisis secundario de datos que deben ser evaluados con mucho cuidado y en este trabajo se revisa y se concluye que los métodos imputación encontrados para la potencia activa instantánea de las barras de carga del Sistema Nacional Interconectado pueden ser generalizados para más variables eléctricas pero se debe considerar que cada situación es diferente y la tasa de no respuesta y su distribución espacial cambia para cualquier variable por lo que no es conveniente adoptar a priori el mismo procedimiento de imputación para todas la variables eléctricas. En la primera fase se analiza la teoría en la que se sustentan los procedimientos de imputación utilizados y en la segunda fase se aplican los siete métodos alternativos para imputar distintos conceptos de potencia activa instantánea de las barras de carga del Sistema Nacional Interconectados del Ecuador y se analiza con datos reales que método estima el valor perdido con un error inferior al 1% por la precisión requerida por los procesos técnicos y

comerciales que se realizan en CENACE. Se demuestra que es factible emplear técnicas de imputación a la variable potencia activa instantánea de las barras de carga del Sistema Nacional y que el 66% de los datos perdidos pueden ser reemplazados a través de métodos de imputación múltiple o simple y los datos reemplazados por los mejores métodos no subestiman la varianza.

CAPÍTULO 1

INTRODUCCIÓN

En el análisis de datos que realiza el Área de Análisis de la Operación de la Dirección de Operaciones del Centro Nacional de Control de Energía “CENACE” es habitual encontrarse con matrices de datos incompletas, especialmente en relación a las potencias activas instantáneas de las barras de carga del Sistema Nacional Interconectado del Ecuador de los datos extraídos de sistema de Manejo de Energía (EMS). Esta situación dificulta la preparación de la información para los procesos técnicos y comerciales que ejecuta la Corporación, además de que el tratamiento y análisis de los datos impide la utilización de los procedimientos estadísticos básicos.

El objetivo de este trabajo es abordar el problema de datos faltantes dentro del marco de la extracción de la información de las bases de datos del Sistema de Manejo de Energía (EMS), que registra información de las variables eléctricas del Sistema Nacional Interconectado del Ecuador. Debido a la inmensa cantidad de información recopilada no es factible la liberación de la información sin un proceso previo de validación de los datos y consecuente reemplazo del dato por uno de mejores características, de ser necesario. Por este motivo, es necesario realizar imputación de la información faltante y para ello es muy importante tener en cuenta la estructura espacio temporal que presentan los datos observados.

Para comprender el problema, a continuación se presentan las funciones de CENACE, descripción de la adquisición de datos en el sistema de Manejo de Energía y el almacenamiento y la extracción de los datos de las variables eléctricas.

1.1 CENACE

1.1.1 PRINCIPALES RESPONSABILIDADES

El Centro Nacional de Control de Energía - CENACE, es una Corporación Civil de derecho privado, de carácter eminentemente técnico, sin fines de lucro, cuyos miembros son todas las empresas de generación, transmisión, distribución y los grandes consumidores. Se encarga del manejo técnico y económico de la energía en bloque, garantizando en todo momento una operación adecuada que redunde en beneficio del usuario final.

El CENACE tiene a su cargo la administración de las transacciones técnicas y financieras del Mercado Eléctrico Mayorista, debiendo resguardar las condiciones de seguridad de operación del Sistema Nacional Interconectado, responsabilizándose por el abastecimiento de energía al mercado, al mínimo costo posible, preservando la eficiencia global del sector.

Sus principales funciones son:

a) La coordinación de la operación en tiempo real del Sistema Nacional Interconectado en condiciones de operación normal y de contingencia, ateniéndose a los criterios y normas de seguridad y calidad que determina el CONELEC;

b) Ordenar el despacho de los equipos de generación para atender la demanda al mínimo costo marginal horario de corto plazo de todo el parque de

generación y controlar que la operación de las instalaciones de generación la efectúe cada titular de la explotación, sujetándose estrictamente a su programación;

c) Mantener informado al CONELEC sobre el cumplimiento de las disposiciones normativas;

d) Asegurar la transparencia y equidad de las decisiones que adopte;

e) Coordinar los mantenimientos de las instalaciones de generación y transmisión, así como las situaciones de racionamiento en el abastecimiento que se puedan producir;

f) Preparar los programas de operación para los siguientes doce meses, con un detalle de la estrategia de operación de los embalses y la generación esperada mensualmente de cada central.

1.1.2 ADQUISICIÓN DE DATOS EN EL SISTEMA DE MANEJO DE ENERGÍA

El Centro Nacional de Control de Energía (CENACE), en su calidad de Administrador Técnico y Comercial del Mercado Eléctrico Mayorista (MEM) del Ecuador, cuenta con diversos sistemas tecnológicos que le permiten cumplir con calidad, seguridad y economía sus funciones.

El sistema de tiempo real EMS, constituye la herramienta tecnológica con la cual el CENACE cumple la función de Coordinador del Sistema Nacional Interconectado e Interconexiones Internacionales; este sistema dispone de una infraestructura de red para la adquisición y transporte de datos, soportada sobre los servidores eLAN (Front Ends Remotos) y la red de telecomunicaciones de TRANSELECTRIC.

La implementación del nuevo Sistema de Control de Energía, EMS Network Manager del CENACE contempló la instalación de 4 pares de servidores eLAN en puntos geográficos estratégicos del país (Santa Rosa, Quevedo, Pascuales y Molino); utiliza como medio de comunicaciones los sistemas de fibra óptica y PLC (Power Line Carrier) de la Empresa TRANSELETRIC, facilitando de esta manera la conexión de los Agentes del MEM (Empresas Generadoras, Empresas de Distribución y Grandes Consumidores) al sistema EMS y permitiendo la recolección distribuida de la información proveniente de las Unidades Terminales Remotas (RTU), de las subestaciones de transmisión y generadores del país y su posterior envío hasta los Centros de Control del CENACE y TRANSELETRIC.

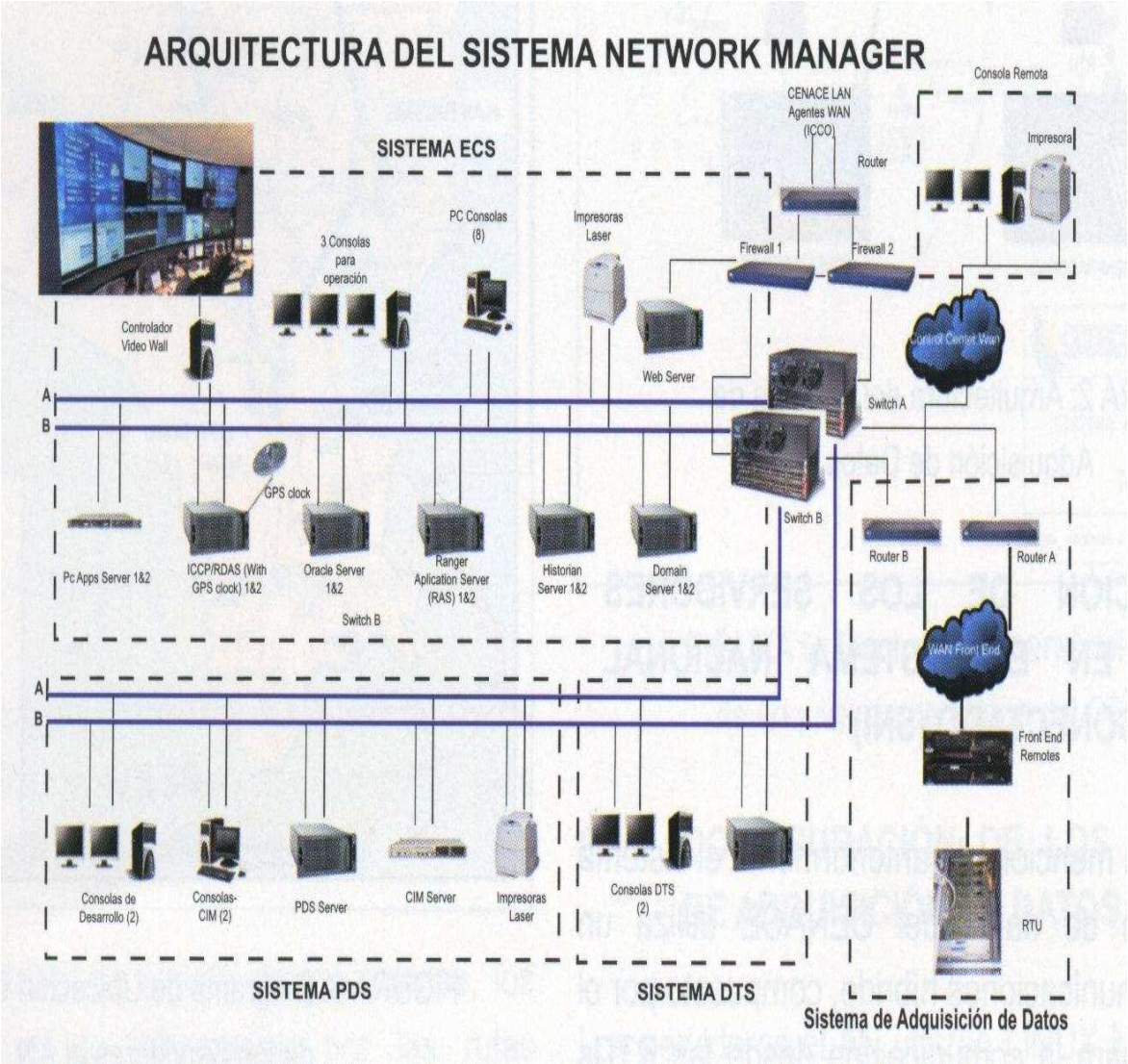
1.1.2.1 Sistema de Manejo de Energía (EMS)[1]

El sistema de manejo de Energía cuyo nombre comercial es Network Manager (NM), fue suministrado por la empresa ABB Inc. de Estados Unidos.

Este sistema está compuesto por varios subsistemas; entre los más importantes: el ECS que procesa y sustenta las aplicaciones del sistema en tiempo real, el PDS que permite el desarrollo y pruebas de nuevas aplicaciones antes de ser instaladas en el sistema ECS, el subsistema DTS que es un simulador para el entrenamiento de operadores y el subsistema de adquisición de datos. La Figura 1.1 muestra la arquitectura del sistema NM.

[1] Revista Energía- Adquisición de Datos en el EMS

FIGURA 1.1
ARQUITECTURA DEL SISTEMA NETWORK MANAGER



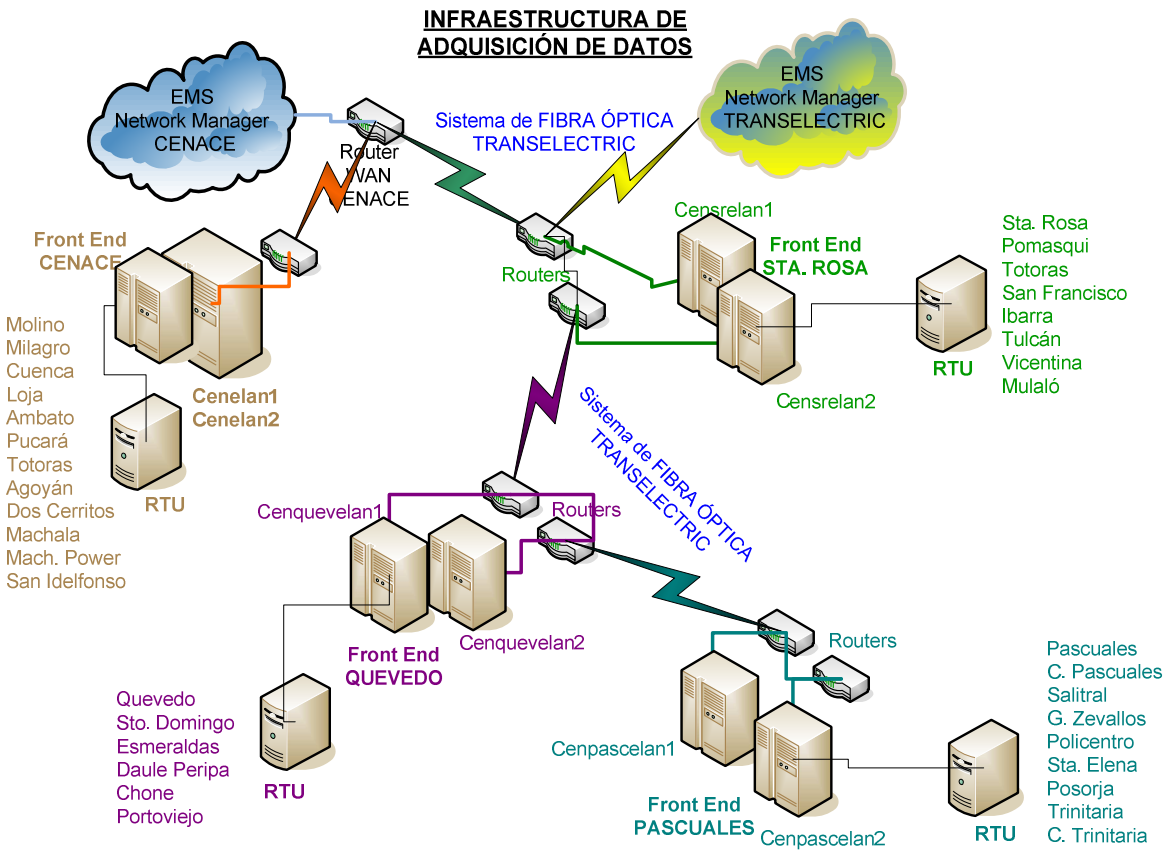
1.1.2.2 ARQUITECTURA DEL SISTEMA DE ADQUISICIÓN DE DATOS

La adquisición de datos del sistema Network Manager inicia con la recolección de la información desde los distintos generadores y subestaciones de transmisión eléctrica, y su envío hacia los servidores eLAN; estos últimos encargados de concentrar la información, para luego enviarla simultáneamente mediante software a los dos centros de control (CENACE y TRANSELECTRIC), proceso que se realiza creando dos unidades terminales remotas virtuales (Virtual

Terminal Unit VRTU) por cada RTU física real, cada una de las cuales se direcciona hacia cada centro de control.

Este proceso está soportado en los sistemas de comunicaciones Power Line Carrier (PLC) y el sistema de fibra óptica como se puede apreciar en la Figura 1.2.

FIGURA 1.2
ARQUITECTURA DEL SISTEMA DE ADQUISICIÓN DE DATOS



La ubicación de los cuatro pares de servidores se puede verificar en la figura 1.3.

1.1.3 BASE DE DATOS DE VARIABLES ELÉCTRICAS

Los datos recopilados desde campo de los Agentes del MEM son almacenados en el Histórico de RANGER y guardados en una serie temporal diseñada para la base de datos, especialmente para archivar gran cantidad de datos, desde miles a

cientos de miles de valores. La facilidad de operación permite que cualquier señal de algún período sea recuperada en segundos.

El sistema histórico puede almacenar los siguientes tipos de datos:

- Estados y correspondientes banderas (alertas) de calidad de las monitorizaciones.

- Valores en unidades de ingeniería y correspondientes banderas (alertas) de calidad para puntos analógicos.

- Valores del acumulador de puntos y banderas de calidad correspondiente, para valores del acumulador de puntos.

- Cualquier valor de dato (real o entero) puede ser especificado para ser almacenado en el registro histórico.

FIGURA 1.3
UBICACIÓN GEOGRÁFICA DE LOS SERVIDORES eLAN[1]

Existe un enlace entre el servidor de datos históricos del RANGER y el Cliente histórico del RANGER a través de:

- *PI-ProcessBook*: Provee una interfase de usuario con capacidad de creación de gráficos, tendencias y resumen de datos.

- *Datalink*: Provee una interfase con hojas de cálculo para la extracción y análisis de datos.

[1] Revista Energía- Adquisición de Datos en el EMS

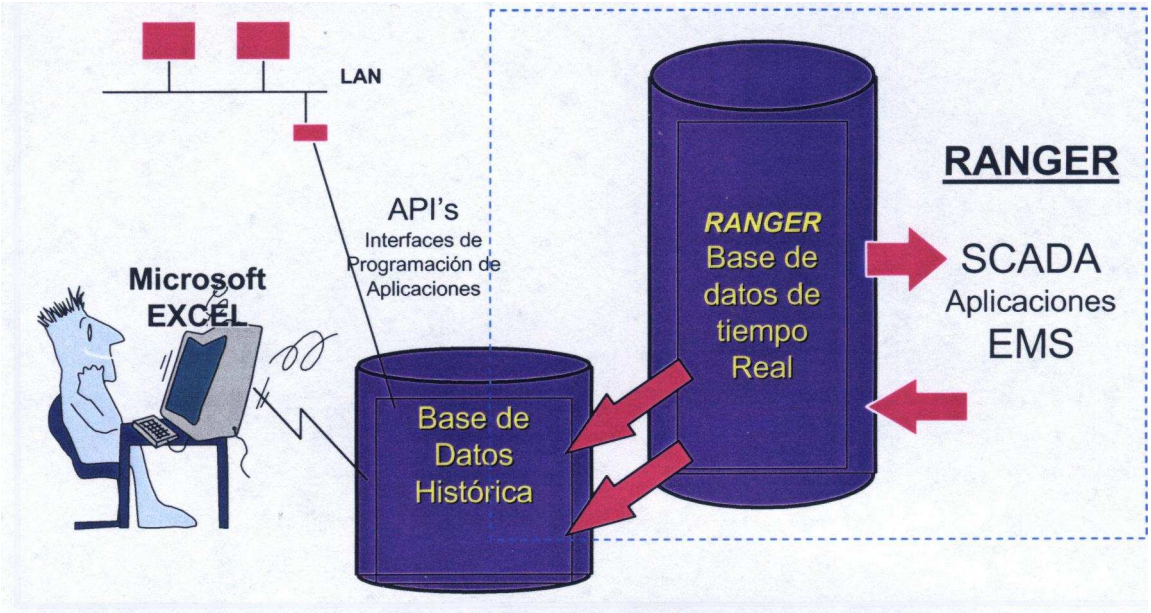
- *ODBC Driver*: Provee acceso mediante el uso del estándar de conectividad de datos de Microsoft.

- *API* : La Interfase de Programación de Aplicaciones permite el acceso a los datos mediante programas externos y aplicaciones específicas.

- PI-ODBC Driver: Se emplea para llevar los datos históricos y de tiempo real a cualquier aplicación cliente y que cumpla con la norma ODBC.

A continuación un esquema explicativo:

FIGURA 1.4
RELACIONES ENTRE BASES DE DATOS



1.2 ANTECEDENTES DEL PROBLEMA

La Corporación Centro Nacional de Control de Energía - CENACE, se encuentra equipada actualmente con un Sistema de Manejo de Energía (EMS) de última

generación, el que permite supervisar el sistema eléctrico de potencia del Sistema Nacional Interconectado y sus Interconexiones Internacionales.

Este sistema, dispone de una infraestructura de red para la adquisición y transporte de datos que utiliza un sistema de comunicaciones híbrido, compuesto por un sistema de PLC para la comunicación desde las RTUs de los Agentes del MEM hasta los servidores eLAN y el sistema de fibra óptica para la comunicación de los servidores eLAN hasta el Centro de Control de CENACE.

La transmisión de los datos de los Agentes del MEM al Centro de Control es vulnerable a daños tanto de las RTUs, sistema de comunicaciones y de los servidores, que en varias ocasiones no pueden ser superados inmediatamente. Este problema provoca que los datos que son almacenados en la Base de Datos Histórica no cuenten con la calidad requerida, especialmente de consistencia y homogeneidad del dato.

Ante esta falta de datos, el Área de Análisis de la Operación de la Dirección de Operaciones debe realizar el reemplazo del dato, ya sea solicitándolo directamente al Agente del MEM o, en muchas ocasiones, estimando el dato que no fue factible extraer desde el campo; esto con el objetivo de disponer de información confiable, de calidad y en el menor tiempo posible, lo que es extremadamente importante debido a que estos datos son utilizados en muchos procesos que realiza el CENACE, como son Planificación, Liquidaciones Comerciales y Estadística Operativa del Sistema Eléctrico Ecuatoriano.

En este contexto el presente trabajo de tesis pretende establecer los fundamentos teóricos de la imputación estadística de datos aplicados a la información de potencia activa de las barras de carga disponible en el EMS, evaluando la aplicabilidad de los métodos conocidos de imputación estadística y determinando la mejor opción para el EMS de Ecuador.

1.3 OBJETIVOS DE LA INVESTIGACIÓN

1.3.1 OBJETIVO GENERAL

Realizar un estudio sobre los diferentes métodos de imputación estadística de datos y proponer alternativas de imputación de datos de potencia activa instantánea provenientes del sistema de manejo de energía EMS de las barras de carga del Sistema Nacional Interconectado del Ecuador.

1.3.2 OBJETIVOS ESPECÍFICOS

1. Describir los principales métodos de imputación de datos cualitativos, categóricos y mixtos.

2. Identificar qué método es aplicable a cada una de las barras de carga del sistema nacional del Ecuador.

3. Realizar un ejemplo de aplicación de imputación de datos de las barras de carga del sistema nacional del Ecuador.

4. Evaluar la efectividad de la imputación de los datos de las barras de carga del Sistema Nacional del Ecuador.

1.4 JUSTIFICACIÓN DEL PROYECTO

1.4.1 JUSTIFICACIÓN TEÓRICA

Los usuarios del sistema de manejo del energía EMS del Centro de Control de Energía CENACE que analizan la base de datos histórica, no cuentan con elementos que permitan realizar un reemplazo del dato erróneo detectado y que

afecta la calidad de los datos en su conjunto. Para enfrentar el problema, y dado que no es factible ignorar las deficiencias pues se alteran los procesos que realiza la Corporación, se plantea el tratamiento de los datos incompletos empleando técnicas de imputación que rescatan la idea intuitiva de reemplazar los valores perdidos por otros, seleccionados con alguna metodología científica.

Por esta razón es necesario realizar un análisis técnico de los diferentes métodos de imputación de datos, así como determinar qué método es el mejor para el tratamiento de los datos incompletos de los registros de potencia activa instantánea, proveniente del Sistema de Manejo de Energía (EMS), de las barras de carga del Sistema Nacional Interconectado del Ecuador.

1.4.2 JUSTIFICACIÓN METODOLÓGICA

La madurez que han alcanzado los métodos de imputación estadística aplicados a diversas ramas de la industria y la investigación científica hacen plausible la aplicación de estos modelos al Sector Eléctrico

1.4.3 JUSTIFICACIÓN PRÁCTICA

La aplicación de los métodos para tratamiento de datos incompletos o inconsistentes permitirá al personal de ingenieros del CENACE complementar la base de datos con valores ajustados en lugar de valores erróneos. La bondad del método de imputación deberá garantizar en la medida de lo posible que el valor reemplazado sea muy similar al real. La incorporación del valor faltante a la base de datos permitirá a los usuarios del sistema de manejo de energía (EMS) del CENACE posteriormente, emplear los métodos de análisis estadísticos estándares para datos completos en los procesos que se realizan en la Corporación.

1.5 METODOLOGÍA

En este trabajo de tesis se presentan cinco capítulos:

El capítulo 1, contiene la Introducción que presenta el contexto en el que nace el problema de la falta de datos en las bases de datos del Sistema de Manejo de Energía CENACE y la necesidad de enfrentar la solución a través de procedimientos de imputación de datos.

En el capítulo 2, Imputación Simple, se exponen los fundamentos teóricos de un conjunto de métodos de imputación univariante y la forma en la que se aplican haciendo énfasis en su bondades y limitaciones.

En el capítulo 3, Imputación Múltiple, se describen los fundamentos teóricos de los métodos de imputación multivariante y estadística Bayesiana y una visión general de imputación de series temporales.

El capitulo 4, Aplicación de la imputación estadística de datos al Sistema Nacional Interconectado del Ecuador, se aplican 6 algoritmos de imputación con el objetivo de sustituir los valores faltantes en la variable potencia activa instantánea de las barra de carga del Sistema Nacional Interconectado del Ecuador. Posteriormente, se evalúa cuál de los métodos de imputación estadística de datos es aplicable a la información disponible en el EMS.

Finalmente, en el capítulo 5 se presentan las conclusiones y recomendaciones.

CAPÍTULO 2

IMPUTACIÓN SIMPLE

El objetivo principal de la imputación de datos es realizar estimaciones no sesgadas de valores ausentes en una determinada observación, que permitan mantener el tamaño de la muestra, de tal manera que no condicione la potencia estadística del estudio, a la vez que permita controlar posibles sesgos en las series con pérdida de datos.

El método apropiado de imputación de datos a aplicar depende del patrón y mecanismo de pérdida, que ha de ser conocido previamente. Por tanto, es esencial identificar claramente cual es el patrón y los mecanismos por los que se produce la ausencia de datos en la variable de análisis, de manera que se pueda seleccionar el método más adecuado para proceder a la imputación de datos ausentes.

2.1 PATRONES DE DATOS PERDIDOS

Los métodos estadísticos estándares se han desarrollado para analizar conjuntos de datos rectangulares. Las filas representan unidades, casos o sujetos dependiendo del contexto y las columnas representan variables medidas para cada unidad. En la resolución del problema de datos faltantes se aborda el análisis de estas matrices de datos, en las cuales algunas celdas no tienen observaciones.

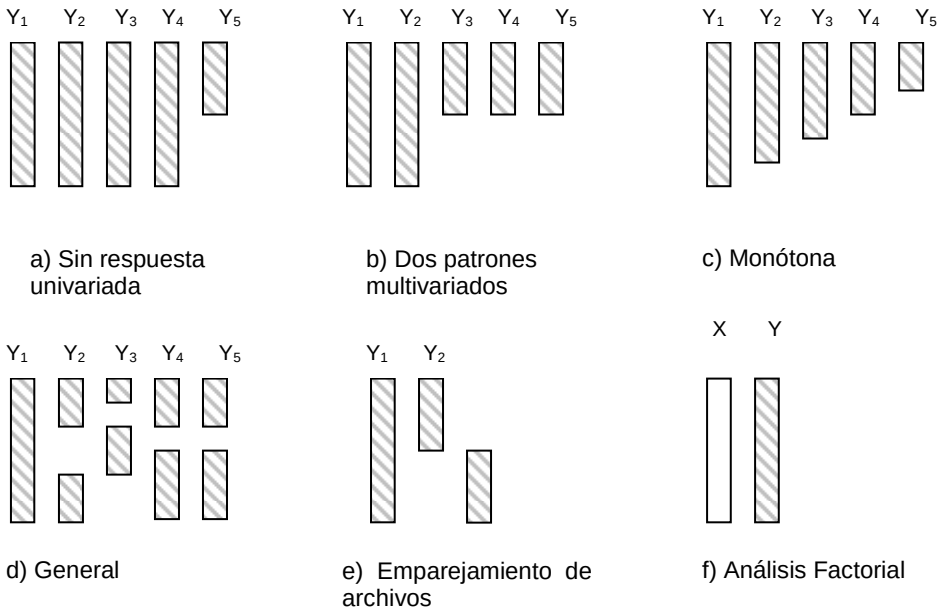
En la resolución del problema de datos faltantes se consideran los siguientes aspectos:

- Se maneja una sola fuente de datos
- Se analizan los datos en una forma global
- Se estudian las causas por las cuales ocurren los datos faltantes y se proponen métodos para su inferencia.

Es de mucha utilidad distinguir el patrón de ausencia de datos y el mecanismo de ausencia de datos, debido a que algunos métodos de análisis dependen de estos patrones.(Little R., Rubin D.B 2000)

La figura 2.1 muestra algunos ejemplos de patrones de datos perdidos, donde la filas representan observaciones y las columnas representan a las variables.

FIGURA 2.1
PATRONES DE AUSENCIA DE DATOS



La figura 2.1 a) representa a los datos perdidos sin respuesta univariada, donde los datos perdidos aparecen en una de las variables del vector aleatorio.

La figura 2.1 b) representa a los datos perdidos de dos patrones multivariados, donde los datos perdidos se encuentran en mas de una variables del vector aleatorio.

La figura 2.1 c) representa a los datos perdidos monótonos, donde las variables del vector aleatorio con datos perdidos puede ser ordenada de acuerdo a los datos perdidos por variable. La creación de patrones monótonos involucra la pérdida de una cantidad substancial de información.

La figura 2.1 d) representa a los datos perdidos de manera general, donde los datos perdidos pueden presentarse en cualquier variables del vector aleatorio sin ninguna lógica en especial.

La figura 2.1 e) representa a los datos perdidos de emparejamiento de archivos, donde Y_1 representa al conjunto de variables que es común a ambas fuentes de datos, los mismos que son totalmente observados; Y_2 representa al conjunto de variables observadas para la primera fuente de datos pero no para la segunda y; Y_3 es el conjunto de datos de variables observadas para la segunda fuente de datos pero no para la primera. Es factible observar que no existe información en este patrón de datos con relación a las asociaciones parciales de Y_2 y Y_3 dado Y_1.

La figura 2.1 f) representa a los datos perdidos con análisis factorial, donde X representa al conjunto de variables aleatorias que son completamente perdidas y Y al conjunto de variables observadas totalmente.

El papel importante de los mecanismos de ausencia era ignorado hasta que el concepto fue formalizado en la teoría de Rubbin.

2.2 LA NATURALEZA DE LA DATOS PERDIDOS[2]

Cuando en una muestra aparecen valores perdidos por razones fuera del control del investigador, es necesario establecer unos supuestos sobre el proceso que los ha generado. Estos supuestos serán en general no verificables, y por ello deberán hacerse explícitos y analizar la sensibilidad del procedimiento general de estimación frente a desviaciones de los mismos.

Si se entiende la presencia de valores perdidos como un fenómeno probabilístico, que necesita un mecanismo matemático que describa las leyes que rigen su aparición, y que capte aproximadamente las posibles relaciones entre la aparición de valores perdidos y los datos no observados en sí mismos.

De manera general, se considera un vector aleatorio Y k-dimensional que genera los datos y un vector R, también k-dimensional, formado por variables aleatorias binarias tomando valores 0 ó 1 para indicar valor observado o no observado. Se llamará mecanismo de no respuesta a la distribución de probabilidad de R. Adicionalmente, si se extrae una muestra del vector Y se tendrá una muestra de R, cuya forma dependerá de la complejidad del patrón de no respuesta.

Ejemplo 1: Si el patrón de datos es univariante, se tendrá una variable aleatoria binaria unidimensional que indica si el valor concreto es observado o perdido. Si el patrón de datos es general se tendrá entonces una matriz de dimensiones n x k con elementos r_{ij} tomando valor 0, si x_{ij} es observado, ó 1 si x_{ij} es no observado.

Por otro lado, se define Y como una muestra multidimensional, tal que se realice una partición de la matriz Y de la forma (Y_{obs}, Y_{aus}), donde Y_{obs}, y Y_{aus} denotan la parte observada y la parte no observada, perdida o ausente, respectivamente. Estas premisas permitirán definir los mecanismos de no repuesta que a continuación se presentan.

[2] LONGFORD NICHOLAS (2005) , LITTLE\RUBIN (2002)

2.2.1 MCAR (MISSING COMPLETELY AT RANDOM) - PERDIDOS COMPLETAMENTE AL AZAR

Se dice que los datos están *perdidos completamente al azar* cuando la probabilidad de que el valor de una variable Y_j, sea observado para un individuo i no depende ni del valor de esa variable, y_{ij}, ni del valor de las demás variables consideradas $y_{ij,}$ $i \neq j$. Es decir, la ausencia de la información no está originada por ninguna variable presente en la matriz de datos. Por ejemplo en el caso de tener en un estudio las variables ingreso y edad, se estará en un modelo MCAR cuando al analizar conjuntamente edad e ingresos, la falta de respuesta en el campo ingresos es independiente del verdadero valor de los ingresos y edad, es decir:

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

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

En los resultados de Rubin (1976), no es necesario que se satisfaga para todas las posibles realizaciones de R, basta que se verifique en la muestra dada.

2.2.2 MAR (MISSING AT RANDOM) – PERDIDOS AL AZAR

Los datos perdidos al azar, se presentan cuando la probabilidad de que el valor de una variable Y_j sea observado para un individuo i no depende del valor de esa variable, y_{ij}, pero quizá sí del que toma alguna otra variable observada $y_{ij,}$ $k \neq j$. Es decir, la ausencia de datos está asociada a variables presentes en la matriz de datos. En el ejemplo anterior, si se supone que los ingresos totales de un hogar son independientes del ingreso individual de sus miembros pero si puede depender de la edad, en este caso se trata de un modelo MAR, es decir:

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

2.2.3 NMAR (NOT MISSING AT RANDOM) –PERDIDOS NO AL AZAR (SI EL MECANISMO DE AUSENCIA DEPENDE DE Y_I)

La hipótesis de datos *perdidos no al azar* (NMAR) es general y se produce cuando la probabilidad de que un valor y_{ij} sea observado depende del propio valor y_{ij}, siendo este valor desconocido. En el ejemplo mencionado, se obtiene que la función respuesta de la variable ingresos depende del propio valor de la variable ingresos, además dependen de otros factores.

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

Las imputaciones permiten obtener distribuciones predictivas de los valores perdidos, requiriendo para ello métodos de creación de este tipo de distribuciones basados en datos observados [3].

2.3 MODELOS EXPLÍCITOS

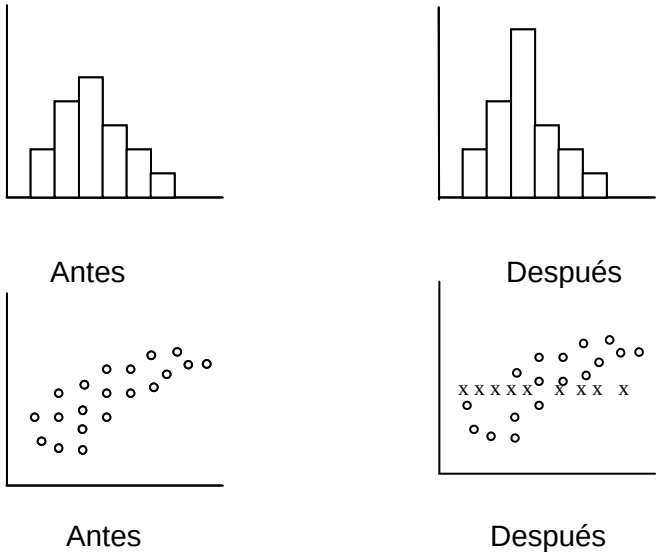
La distribución predictiva se basa en un modelo estadístico formal y por tanto la suposiciones son explícitas.

2.3.1 MÉTODOS DE IMPUTACIÓN POR LA MEDIA

Este método trabaja bajo un procedimiento MCAR y consiste en sustituir los valores perdidos de la variable *y* por la media de los valores observados; es decir, la variable incompleta *y* es imputada para cada valor perdido de *y.*

[3] Little/Rubin (2002)

FIGURA 2.2
CARACTERÍSTICA DE LA IMPUTACIÓN POR MEDIAS[4]



La imputación por la media provoca que la distribución de los nuevos valores sea una representación incorrecta de los valores de la población, debido a que la forma de la distribución es distorsionada por la adición de los valores iguales a la media, tal como se observa en la figura 2.2.

Adicionalmente, este procedimiento distorsiona la distribución fundamental de los datos, provocando que la distribución tenga un pico pronunciado alrededor de la media y reduciendo la varianza.

2.3.1.1 MEDIA NO CONDICIONAL

Una forma simple de imputación de los valores perdidos es estimar la media de los valores observados correspondientes a la variable analizada, entonces:

[4] JUÁREZ CARLOS (2004)

Sean y_{ij} los valores de la variable Y_j para un individuo i. El valor perdido y_{ij} es estimado por $\bar{y}_j^{(j)}$ que es la media de los valores registrados de la variable $Y_{j.}$; por tanto, el promedio de los valores observados e imputados se encuentran en $\bar{y}_j^{(j)}$

La varianza de los valores observados e imputados es:

$$\frac{s_{jj}^{(j)}(n^{(j)}-1)}{(n-1)} \qquad \textbf{(2.1)}$$

Donde $s_{jj}^{(j)}$ es la varianza estimada de los $n^{(j)}$ casos observados[3].

Si el tipo de datos es MCAR, $s_{jj}^{(j)}$ es estimador consistente de la varianza, pero la varianza de la muestra del conjunto de datos completos se ve subestimada por el factor $\frac{(n^{(j)}-1)}{(n-1)}$. Adicionalmente, debido a esta subestimación los valores perdidos e imputados se encuentran en el centro de la distribución, por tanto la imputación de la media distorsiona la distribución empírica de los valores de la variable Y. Por otro lado, las estimaciones de cantidades que no son lineales en los datos como la varianza, percentiles o medidas de forma, no son estimadas consistentemente usando los métodos estándar para datos completos.

El mismo caso ocurre si los valores de la variable Y_j son agrupados en subclases por tablas de contingencia, por tanto, los valores perdidos en una celda son todos reemplazados por un valor medio común y entonces son clasificados en la subclase de $Y_{j.}$

La covarianza de las muestras para Y_j y Y_k con datos completados se calcula por:

[3] Little/Rubin(2002)

$$\frac{\tilde{s}_{jk}^{(jk)}(n^{(jk)}-1)}{(n-1)} \qquad \textbf{(2.2)}$$

Donde $n^{(jk)}$ corresponde al número de casos con Y_j y Y_k observados y $\tilde{s}_{jk}^{(jk)}$ se calcula con la siguiente expresión:

$$\sum_{i \in I_{jk}} \frac{(y_{ij}-\bar{y}_j^{(j)})(y_{ik}-\bar{y}_k^{(k)})}{(n^{(jk)}-1)} \qquad \textbf{(2.3)}$$

Donde I_{jk} es el conjunto de los $n^{(jk)}$ casos con Y_j y Y_k observados y $\bar{Y}_j^{(jk)}, \bar{Y}_k^{(jk)}$ son calculados sobre el conjunto de los casos.

Por tanto, para modelos MCAR, $\tilde{s}_{jk}^{(jk)}$ es un estimador consistente de la covarianza. La estimación para datos completos subestima la magnitud de la covarianza por el factor $\frac{(n^{(jk)}-1)}{(n-1)}$. Entonces, aunque la matriz de la covarianza de datos completos sea semi-definida positiva, la varianza y la covarianza son sistemáticamente atenuadas.

Se debe observar que los resultados de la matriz de covarianza pueden no ser definida positiva, particularmente cuando las variables son altamente correlacionadas.

2.3.1.2 MEDIA CONDICIONAL

Una mejora en la imputación media no condicional es la imputación media condicional dada por los valores observados; es decir, imputa medias condicionadas a valores observados. El método más común es el de agrupar valores observados y no observados de acuerdo al peso por clases basadas en las variables observadas e imputa los valores faltantes con valores observados en la misma clase.

Por ejemplo, en el caso de regresión logística, la variable de respuesta Y induce dos clases C_o y C_1 compuestas por n_o y n_1 elementos respectivamente. Entonces, para cada variable Y_j en la clase C_o de tamaño n_o se tienen r_o valores observados y $n_o - r_o$ valores faltantes. Similarmente en la clase C_1 de tamaño n_1 se tienen r_1 valores faltantes para la misma variable. Este tipo de imputación es similar a la imputación por la media no condicional pero aplicada a ambas clases. La imputación condicional por la media obedecería a la siguiente ecuación:

$$y_{(j),C_k} = \frac{\sum_{i=1} r_k \, y_{i,obs(j),C_k}}{r_k} \quad \text{para k=0 , 1} \qquad \textbf{(2.4)}$$

2.3.2 MÉTODOS DE IMPUTACIÓN POR LA MEDIANA

Además de la imputación por la media, otra medida de tendencia central utilizada es la mediana. Este método de imputación es empleado para variables continuas y categóricas ordinales. El método de la mediana es de fácil aplicación, además de que es factible encontrar en la mayoría de programas estadísticos. El método consiste en que cada dato faltante es sustituido por la mediana de una misma variable. La imputación por la mediana se define:

$$Y_j = Mediana(Y_{obs(j)}) \qquad \textbf{(2.5)}$$

Ejemplo 2: Considerando los siguientes datos:

CUADRO 2.1

DATOS DE MUESTRA DEL EJEMPLO 2:

Datos de potencias activas horarias posición Ambato.

Unidad	Y
1	6.5
2	5.4
3	8.4
4	6.2
5	6.5
6	
7	6.2
8	
9	7.4
10	

Donde:

Corresponden a valores perdidos.

Extrayendo la estadística descriptiva de los datos:

CUADRO 2.2

ESTADÍSTICA DESCRIPTIVA DE LOS DATOS DE MUESTRA DEL EJEMPLO 2

	Y
Media	6.66
Error típico	0.37
Mediana	6.50
Moda	6.50
Desviación estándar	0.97
Varianza de la muestra	0.94
Curtosis	1.01
Coeficiente de asimetría	0.90
Rango	3.00
Mínimo	5.40
Máximo	8.40
Suma	46.60
Cuenta	7.00

A continuación, la mediana se reemplaza en los datos perdidos, tal como se indica a continuación:

CUADRO 2.3

DATOS DE MUESTRA DEL EJEMPLO 2 IMPUTADOS

Unidad	Y
1	6.5
2	5.4
3	8.4
4	6.2
5	6.5
6	6.5
7	6.2
8	6.5
9	7.4
10	6.5

Como puede observarse en este método, la desventaja principal es que la calidad del resultado dependerá de la calidad de información, carece de un mecanismo

de probabilidad y el valor estimado es reemplazado varias veces subestimando la varianza.

2.3.3 MÉTODOS DE IMPUTACIÓN POR LA MODA

La imputación por la moda se utiliza para variables categóricas. En caso de haber más de una moda se escoge aleatoriamente entre las modas y se imputa el valor faltante por ese valor. Se utiliza la imputación por la moda para este tipo de variable en vez de la media muestral o la mediana, para evitar que el número de clases en la variable no se afecte.

Al igual que la imputación por la mediana la desventaja principal es que la calidad del resultado dependerá de la calidad de información, carece de un mecanismo de probabilidad y el valor estimado es reemplazado varias veces, subestimando la varianza.

2.3.4 MÉTODOS DE IMPUTACIÓN POR REGRESIÓN

Los valores faltantes se sustituyen por valores estimados desde una regresión realizada sobre los valores observados.

Considerando una variable Y_i que presenta n_{aus} valores perdidos o ausentes y $n_i = n - n_{aus}$ valores observados. Se supone además que las k -1 variables restante Y_j con $j \neq i$, no presentan valores perdidos. Con este método se estima la regresión de la variable Y_i sobre las variables Y_j para toda $j \neq i$, a partir de los n_i casos completos y se imputa cada valor perdido con la predicción dada por la ecuación de regresión estimada. Así, si para el caso I el valor y_{li} no se observa, entonces se imputa mediante:

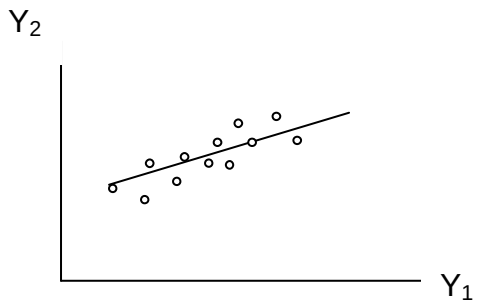
$$\hat{y}_{li} = \hat{\beta}_{o_{obs}} + \sum_{j \neq i} \hat{\beta}_{j_{obs}} y_{lj} \qquad \textbf{(2.6)}$$

Donde $\hat{\beta}_{0_{obs}}$ y $\beta_{j_{obs}}$ con $j \neq i$, representan los coeficientes de la regresión de Y_i sobre Y_j, $\forall\, j \neq i$, basada en las n_i observaciones completas. Frente a la imputación

mediante la media, este método incorpora la información que sobre Y_i contienen el resto de las variables.

La imputación por regresión es ilustrada gráficamente para k=2 variables en la figura 2.3:

FIGURA 2.3
IMPUTACIÓN POR REGRESIÓN PARA k = 2



Como puede observarse en la figura 2.3 los puntos representan los casos en los que los valores son observados tanto para Y_1 como para Y_2, y estos puntos son empleados para calcular la regresión lineal de Y_2 sobre Y_1. Con esta información es posible obtener la estimación mínima cuadrática $\hat{y}_{21} = \hat{\beta}_{20,1} + \hat{\beta}_{21,1} y_{i1}$

Ejemplo 3: Se consideran los siguientes datos:

CUADRO 2.4
DATOS DE MUESTRA DEL EJEMPLO 3

Unidad	Y	X
1	6,5	6,9
2	5,4	4,5
3	8,4	12,2
4	6,2	5,3
5	6,5	6,6
6		2,6
7	6,2	3,4
8		11
9	7,4	10,2
10		9,7

donde:

_ Corresponden a valores perdidos.

Eliminando los valores perdidos se obtiene lo siguiente:

CUADRO 2.5
DATOS DE MUESTRA PROCESADOS DEL EJEMPLO 3

UNIDAD	**Y**	**X**
1	6,5	6,9
2	5,4	4,5
3	8,4	12,2
4	6,2	5,3
5	6,5	6,6
7	6,2	3,4
9	7,4	10,2

Con estos datos, se procede a realizar la regresión lineal, obteniéndose los siguientes parámetros:

CUADRO 2.6
ESTADÍSTICA DESCRIPTIVA DE LOS DATOS DEL CUADRO 2.5

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*
Término Independiente	4.64	0.38	12.30	6.3E-05
X	0.29	0.05	5.77	0.0022

Por lo tanto, la ecuación de la regresión lineal resulta:

$$\hat{y}_i = 4.64 + 0.29\, x_i \qquad \textbf{(2.7)}$$

A continuación se sustituyen los valores perdidos de la variable *Y*, a través de la regresión lineal:

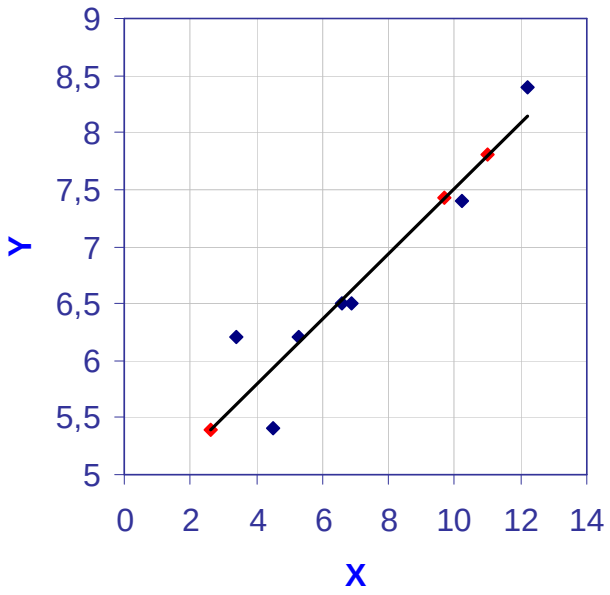
CUADRO 2.7

APLICACIÓN DE LA IMPUTACIÓN EN BASE A REGRESIÓN LINEAL

Unidad	Y	X
1	6.5	6.9
2	5.4	4.5
3	8.4	12.2
4	6.2	5.3
5	6.5	6.6
6	**5.4**	2.6
7	6.2	3.4
8	**7.8**	11
9	7.4	10.2
10	**7.4**	9.7

GRÁFICO 2.1

REGRESIÓN LINEAL DE LOS DATOS DEL EJEMPLO 3



A continuación se realiza un cuadro comparativo de medias, desviaciones estándar y correlaciones:

CUADRO 2.8

RESUMEN DE MEDIA, DESVIACIÓN ESTÁNDAR Y CORRELACIONES

CASO ELIMINADO VALORES PERDIDOS			
VARIABLES	N	MEDIA	Desv. Estándar
Y	7	6,66	0,97
X	7	7,01	3,15
Correlación (X,Y)		0,93	
CASO COMPLETO CON VALORES PERDIDOS			
VARIABLES	N	MEDIA	Desv. Estándar
Y	7	6,66	0,97
X	10	7,24	3,36
CASO CON REEMPLAZO DE VALORES PERDIDOS CON REGRESIÓN LINEAL			
VARIABLES	N	MEDIA	Desv. Estándar
Y	10	6,72	1,00
X	10	7,24	3,36
Correlación (X,Y)		0,96	

2.3.4.2 IMPUTACIÓN POR REGRESIÓN ESTOCÁSTICA

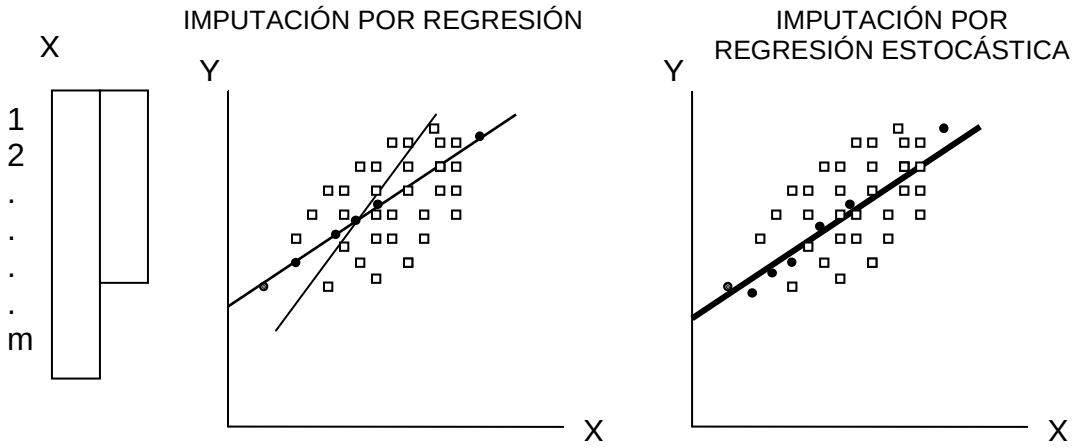
Reemplaza valores perdidos por valores estimados mediante la imputación por regresión añadiéndole un residual, reflejando así la incertidumbre en la predicción del valor. En los modelos de regresión lineal normal, los residuales serán naturalmente normales con media cero y varianza igual a la varianza residual en la regresión, esto es:

$$\hat{y}_{li} = \hat{\beta}_{o_{obs}} + \sum_{j \neq i} \hat{\beta}_{j_{obs}} y_{lj} + \varepsilon_{li} \qquad (2.8)$$

Donde $\varepsilon_{li} \approx N(0, \sigma^2_{resid})$, siendo σ^2_{resid} la varianza residual de la regresión de Y_i sobre Y_j $\forall j \neq i$

FIGURA 2.4

COMPARACIÓN DE IMPUTACIÓN POR REGRESIÓN CON REGRESIÓN ESTOCÁSTICA



Como puede observarse en la Figura 2.4, la imputación por regresión estocástica reemplaza valores perdidos por puntos cercanos a la línea de la regresión, pero mejora la variabilidad.

Ejemplo 4: Continuando el Ejemplo 3, se observa que con la regresión lineal obtenida y con la inclusión del error cuadrático medio de la regresión se obtiene la siguiente ecuación:

$$\hat{y}_i = 4.64 + 0.29x_i + e_i \qquad \textbf{(2.9)}$$

Donde e_i es generado aleatoriamente de una distribución normal con media 0 y varianza igual a la varianza residual

Se presenta un cuadro de medias, desviaciones estándar y correlaciones:

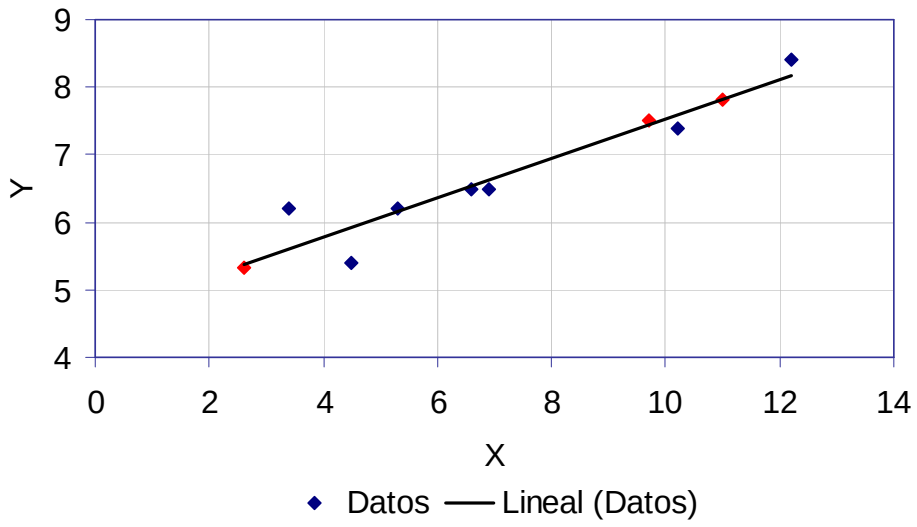
CUADRO 2.8

RESUMEN DE MEDIA, DESVIACIÓN ESTÁNDAR Y CORRELACIONES

CASO CON REEMPLAZO DE VALORES PERDIDOS CON REGRESIÓN LINEAL ESTOCÁSTICA			
VARIABLES	N	MEDIA	Desv. Estándar
Y	10	6,73	1,02
X	10	7,24	3,36
Correlación (X,Y)		0,95	

GRÁFICO 2.1

EJEMPLO DE REGRESIÓN LINEAL ESTOCÁSTICA



2.3.5 MÉTODOS DE IMPUTACIÓN POR MÁXIMA VEROSIMILITUD

Los métodos de imputación por máxima verosimilitud tienen como objetivo realizar estimaciones máximo verosímiles de los parámetros de una distribución cuando existen datos faltantes. Para ello se utiliza un valor inicial del vector de parámetros y se llenan los datos faltantes mediante valores estimados; así con los datos completos se obtiene una nueva estimación del vector de parámetros y con la nueva estimación del vector de parámetros se vuelven a llenar los datos faltantes. Se siguen realizando las iteraciones hasta lograr convergencia.

Considerando un fenómeno multivariante real cuyo comportamiento viene descrito por un vector aleatorio k-dimensional $Y = (Y_1, ..., Y_k) \in R^k$ con distribución de probabilidad $P[Y;\theta]$, siendo θ el vector de parámetros desconocidos.

Cuando se dispone de una muestra completa de Y, una amplia clase de métodos de inferencia se justifican en la interpretación de $P[Y;\theta]$ como una función de verosimilitud que resume la evidencia que sobre θ, hay en los datos. Pero en presencia de valores perdidos, solo disponemos de Y_{obs} cuya distribución se obtiene como:

$$P[Y_{obs};\theta] = \int P[Y;\theta]\, dY_{aus} \qquad \textbf{(2.10)}$$

Si se realiza inferencia sobre θ a partir de los datos observados, es necesario comprobar que la ecuación anterior (2.10) es una verosimilitud adecuada. Rubin (1976) identifica las condiciones para que así sea, estableciendo que basta que se verifique la hipótesis MAR (perdidos al azar) como se comprueba a continuación:

Según se ha formalizado el problema de las muestras incompletas, es necesario especificar un modelo para Y, $P[Y;\theta]$ y un modelo para la no respuesta $P[R \mid Y_{obs}, Y_{aus}, \xi]$. Por otro lado, mediante el producto $P[R \mid Y_{obs}, Y_{aus}, \xi].P[Y;\theta]$ se obtiene la distribución conjunta $P[Y, R;\theta,\xi]$. Ahora la verosimilitud basada en la parte observada puede expresarse como:

$$P[Y_{obs}, R;\theta,\xi] = \int P[Y, R;\theta,\xi]\, dY_{aus} = \int P[R \mid Y_{obs}, Y_{aus}, \xi].P[Y;\theta]\, \mathrm{dY_{aus}} \qquad (2.11)$$

Con el supuesto MAR, la ecuación (2.10) se convierte en:

$$P[Y_{obs}, R;\theta,\xi] = P[R \mid Y_{obs}, \xi] \int P[Y;\theta]\, dY_{aus} = P[R \mid Y_{obs}, \xi].P[Y_{obs};\theta] \qquad (2.12)$$

De modo que la verosimilitud (2.11) bajo el supuesto MAR queda factorizada en dos partes, una relativa la vector θ y otra relativa al vector ξ. Si además θ y ξ

proporcionan poca información la una sobre la otra, entonces las inferencias sobre θ basadas en verosimilitudes no se verán afectadas por $P[R \mid Y_{obs}, \xi]$, esto es, el mecanismo de no respuesta puede ser ignorado y la función de verosimilitud L de θ será $L(\theta|Y_{obs})$ proporcional a $P[Y_{obs}; \theta]$.

Este resultado, destaca que bajo ignorabilidad se pueden realizar inferencias sobre el vector de parámetros θ de la distribución de Y a partir de la verosimilitud $L(\theta|Y_{obs})$.

Por otro lado, desde el punto de vista bayesiano, todas las inferencias se basan en la distribución de probabilidad a posteriori de los parámetros conocidos, que pueden escribirse utilizando el teorema de Bayes como:

$$P[\theta, \xi \mid R, Y_{obs}] = \frac{P[R, Y_{obs} \mid \theta, \xi]\varphi(\theta, \xi)}{\iint P[R, Y_{obs} \mid \theta, \xi]\varphi(\theta, \xi)\, d\theta\, d\xi} \qquad \textbf{(2.13)}$$

Donde φ representa a la distribución a priori de (θ, ξ). Bajo el supuesto MAR, se puede sustituir (2.12) en (2.13), obteniendo que $P[\theta, \xi \mid R, Y_{obs}]$ es proporcional a $P[R \mid Y_{obs}, \xi]\ P[Y_{obs} \mid \theta]\ \varphi(\theta, \xi)$. Si además θ y ξ proporcionan poca información la una sobre la otra, entonces la distribución marginal a posteriori θ queda como:

$$P[\theta \mid Y_{obs}, R] = \int P[\theta, \xi \mid R, Y_{obs}]\, d\xi \propto P[Y_{obs} \mid \theta]\varphi_\theta(\theta) \int P[R \mid Y_{obs}, \xi]\varphi_\xi(\xi)\, d\xi \ \alpha\ L(\theta \mid Y_{obs})\, \varphi_\theta(\theta) \qquad \textbf{(2.14)}$$

$\propto$ representa proporcionalidad

Por tanto, bajo la hipótesis de ignorabilidad, toda la información sobre θ se recoge en la distribución a posteriori que ignora el mecanismo de no respuesta observado, $P[\theta \mid Y_{obs}] \alpha\ L(\theta \mid Y_{obs})\, \varphi_\theta(\theta)$.

2.3.5.1 ALGORITMO EM

El algoritmo EM es un algoritmo iterativo general por estimación de máxima verosimilitud en problema de datos faltantes. El enfoque general para calcular estimadores de máxima verosimilitud fue proporcionado por Dempster, Laird y Rubin. La técnica empleada es conocida como el algoritmo EM que se basa en cálculos iterativos en dos pasos: el paso E (predicción) y el paso M (maximización).

Sea Y = (Y_{obs}, Y_{aus}) una muestra incompleta, a partir de la cual se desea obtener el estimador de máxima verosimilitud (EMV) de θ. Para tal efecto se factoriza P[Y; θ] como:

$$P[Y;\theta] = P[Y_{obs};\theta]\ P[Y_{aus} \mid Y_{obs},\theta] \qquad \textbf{(2.15)}$$

A continuación se puede deducir:

$$\log L(\theta \mid Y_{obs}) = \log L(\theta \mid Y) - \log P(Y_{aus} \mid Y_{obs},\theta) \qquad \textbf{(2.16)}$$

Siendo log L(θ|Y_{obs}) el logaritmo de verosimilitud para los datos observados y log L(θ|Y) el logaritmo de verosimilitud de los datos completos de Y. En presencia de datos faltantes, se estima θ mediante la maximización de log L(θ|Y_{obs}) con respecto a θ dada Y_{obs}. Se puede observar que el algoritmo EM relaciona el EMV de θ a partir de log L(θ|Y_{obs}) con el EMV de θ a partir de log L(θ|Y). A continuación se toman las esperanza respecto a P[Y_{aus}| Y_{obs},θ] a los dos lados de la ecuación (2.16) y dado un estimador $\theta^{(t)}$ de θ, se obtiene:

$$\log L(\theta \mid Y_{obs}) = Q(\theta;\theta^{(t)}) - H(\theta;\theta^{(t)}) \qquad \textbf{(2.17)}$$

donde:

$$Q(\theta;\theta^{(t)}) = \int \log L(\theta \mid Y)\ P[Y_{aus} \mid Y_{obs},\theta^{(t)}]\ \mathrm{dY}_{aus} \qquad \textbf{(2.18)}$$

y

$$H(\theta;\theta^{(t)}) = \int \log P[Y_{aus} \mid Y_{obs}, \theta]\, P[Y_{aus} \mid Y_{obs}, \theta^{(t)}]\, \mathrm{dY_{aus}} \qquad \textbf{(2.18)}$$

El paso E (predicción) del algoritmo EM calcula $Q(\theta, \theta^{t})$, a través del reemplazo de los valores perdidos, o una función de ellos, por su esperanza condicionada dados Y_{obs} y $\theta^{(t)}$.

El paso M (maximización) determina el EMV de $\theta^{(t+1)}$ que maximiza $Q(\theta, \theta^{t})$ como si no hubiera datos perdidos.

Los pasos E y M se repiten alternativamente generando una sucesión de estimadores $\{\theta^{(t)}\}$. La diferencia en el valor del log-verosimilitud log $L(\theta|Y_{obs})$ en las dos iteraciones sucesivas viene dada por:

$$\log L(\theta^{t+1} \mid Y_{obs}) - \log L(\theta^{(t)} \mid Y_{obs}) = Q(\theta^{t+1};\theta^{(t)}) - Q(\theta^{(t)};\theta^{(t)}) + H(\theta^{(t)};\theta^{(t)}) - H\theta^{(t+1)};\theta^{(t)}) \qquad \textbf{(2.19)}$$

El estimador $\theta^{(t+1)}$ es elegido de manera que $Q(\theta^{(t+1)};\theta^{(t)}) \geq Q(\theta^{(t)};\theta^{(t)})$ y $H(\theta^{(t)};\theta^{(t)}) \geq H(\theta^{(t+1)};\theta^{(t)})$, debido a la desigualdad de Jensen y la concavidad de la función logarítmica. Adicionalmente el log $L(\theta|Y_{obs})$ se va incrementando en cada iteración logrando la convergencia hacia el EMV θ.
En Little y Rubin (2002) se pueden encontrar resultados teóricos y condiciones de la convergencia del algoritmo. Un criterio de convergencia habitual en la practica consiste en detener el proceso cuando la diferencia entre dos estimaciones sucesivas de θ sea suficientemente pequeña.

Una ventaja del algoritmo EM es su convergencia puntual. En este sentido bajo condiciones generales, cada iteración incrementa el logaritmo de máxima verosimilitud $l(\theta|X_{obs})$, y si $l(\theta|X_{obs})$ se encuentra en el límite, la secuencia $l(\theta^{(t)} | X_{obs})$ converge a un valor estacionario de $l(\theta|X_{obs})$. Generalmente, si la secuencia $\theta^{(t)}$ converge, ésta converge a un máximo local o al punto máximo de la curva $l(\theta^{(t)} | X_{obs})$. Una desventaja del método EM es que su tasa de convergencia puede

ser de un crecimiento muy lento cuando existe una gran cantidad de información perdida.

A continuación se presentarán unos ejemplos:

Ejemplo 5*:* Datos normales univariados[3]

Suponiendo X_i independiente e idénticamente distribuida con $N(\mu,\sigma^2)$, donde X_i corresponde a los valores observados cuando i=1.....,r, y X_i está relacionada con los valores perdidos cuando i = r+1,.....,n; además se asume que el mecanismo de los datos perdidos es ignorable, entonces:

La esperanza de cada valor perdido X_i está dada por X_{obs} y $\theta = (\mu , \sigma^2)$.
Por otro lado, el logaritmo de máxima verosimilitud $l(\theta|X_{obs})$ basado en todos los X_i cuando i=1.....,n, es lineal en los estadístico suficientes $\sum_1^n X_i$ y $\sum_1^n X_i^2$.

Entonces:

El *paso E* del algoritmo, calcula:

$$E(\sum_{i=1}^{n} X_i \mid \theta^{(t)}, X_{obs}) = \sum_{i=1}^{r} X_i + (n-r)\mu^{(t)} \qquad \textbf{(2.20)}$$

$$E(\sum X_i^2 \mid \theta^{(t)}, X_{obs}) = \sum_{i=1}^{r} X_i^2 + (n-r)[(\mu^{(t)})^2 + (\sigma^{(t)})^2] \qquad \textbf{(2.21)}$$

En la primera estimación los parámetros serian: $\theta^{(t)} = (\mu^{(t)}, \sigma^{(t)})$

Si los datos no están perdidos, el estimador MV de μ es $\sum_{i=1}^{n} \frac{X_i}{n}$ y el estimador MV de σ^2 es $\sum_{i=1}^{n} \frac{X_i^2}{n} - \left(\sum_{i=1}^{n} \frac{X_i}{n} \right)^2$.

[3] Little/Rubin (2002)

El *paso M* usa las mismas expresiones con las esperanzas de los cálculos de los estadísticos suficientes calculados en el paso E, pero sustituidas por los estadísticos suficientes de los datos incompletos. Es así que en el paso M se calcula:

$$\mu^{(t+1)} = \frac{E(\sum_{i=1}^{n} X_i \mid \theta^{(t)}, X_{obs})}{n} \qquad \textbf{(2.22)}$$

$$(\sigma^{t+1)})^2 = \frac{E(\sum_{i=1}^{n} X_i^2 \mid \theta^{(t)}, X_{obs})}{n - (\mu^{(t+1)})^2} \qquad \textbf{(2.23)}$$

Encerando $\mu^{(t)} = \mu^{(t+1)}$ y $\sigma^{(t)} = \sigma^{(t+1)} = \hat{\sigma}$ en las ecuaciones (2.20) – (2.23), se observa que los valores fijos de estas iteraciones siempre son:

$$\hat{\mu} = \sum_{i=1}^{r} \frac{x_i}{r} \qquad \textbf{(2.24)}$$

$$\hat{\sigma}^2 = \sum_{i=1}^{r} \frac{x_i^2}{r - \hat{\mu}^2} \qquad \textbf{(2.25)}$$

Los que son estimadores de MV de μ y σ^2 desde los X_{obs}, asumiendo que son MAR.

Ejemplo 6: Ejemplo Multinomial[3]

Suponiendo que el vector de datos observados contiene los siguientes valores X_{obs} = (38, 34, 125) los que provienen de una función multinomial con probabilidades (1/2 - θ/2 , θ/2 , θ/2). El objetivo es encontrar el estimador de ML de θ.

Entonces el vector aleatorio $X=(x_1,x_2,x_3)$ se distribuye multinomialmente con la función de probabilidad:

[3] Little/Rubin (2002)

$$f(x/\theta) = \frac{(x_1 + x_2 + x_3)!}{x_1 x_2 x_3}(\frac{1}{2} - \frac{\theta}{2})X_1(\frac{\theta}{2})X_2(\frac{\theta}{2})X_3 \qquad \textbf{(2.26)}$$

Luego el logaritmo de la función de verosimilitud será:

$$L(\theta/X) = \log\ f(x/\theta) = x_1 \log(1-\theta) + x_2 \log\theta + x_3 \log\theta + z \qquad \textbf{(2.27)}$$

donde z incluye los términos constantes.

Derivando con respecto a θ e igualando a cero se obtiene:

$$\frac{x_2}{\hat{\theta}} + \frac{x_3}{\hat{\theta}} - \frac{x_1}{1-\hat{\theta}} = 0 \qquad \textbf{(2.28)}$$

Entonces el estimador de máxima verosimilitud de θ sería:

$$\hat{\theta} = \frac{x_2 + x_3}{x_1 + x_2 + x_3} \qquad \textbf{(2.29)}$$

Nótese que el logaritmo de máxima verosimilitud l(θ|X) es lineal en X, de la misma forma, el encontrar la esperanza de l(θ|X) dado θ y X_{obs} involucra los mismos cálculos que encontrar la esperanza de X dado θ y X_{obs}, con lo cual la estimación de los valores perdidos será:

$E(x_1 \mid \theta, X_{obs}) = 38$

$E(x_2 \mid \theta, X_{obs}) = 34$

$E(x_3 \mid \theta, X_{obs}) = 125\ (\theta/4) / (1/2+\theta/4)$

$E(x_4 \mid \theta, X_{obs}) = 125\ (1/2) / (1/2+\theta/4)$

Por tanto, el *paso E* considerando la t-ésima iteración con la estimación de $\theta^{(t)}$, sería:

$$x_3^t = 125\frac{\left(\frac{\theta^{(t)}}{4}\right)}{\left(\frac{1}{2}+\frac{\theta^{(t)}}{4}\right)} \qquad \textbf{(2.30)}$$

Y el *paso M*, a través de la ecuación (2.29) sería:

$$\theta^{(t+1)} = \frac{34 + x_3^{(t)}}{72 + x_3^{(t)}} \qquad \textbf{(2.31)}$$

Inicializando $\theta^{(t+1)} = \theta^{(t)} = \hat{\theta}$ y combinando las ecuaciones (2.30) y (2.31) se encuentra una ecuación cuadrática que permite obtener $\hat{\theta}$, esto es:

$$x_3^{(t)} = \frac{\frac{125}{4}\hat{\theta}}{\frac{1}{2}+\frac{\hat{\theta}}{4}} \qquad \textbf{(2.32)}$$

$$\hat{\theta} = \frac{34 + x_3^{(t)}}{72 + x_3^{(t)}} \qquad \textbf{(2.33)}$$

Combinando las dos ecuaciones anteriores, se obtiene la siguiente ecuación cuadrática:

$$197\hat{\theta}^2 - 15\hat{\theta} - 68 = 0 \qquad \textbf{(2.34)}$$

De donde se puede obtener el valor de $\hat{\theta}$ que es de 0.6268, iniciando con $\theta^{(0)} = \frac{1}{2}$ e iterando a partir de las ecuaciones (2.30) y (2.31) se obtiene la convergencia lineal del algoritmo EM, que se indica en la tabla siguiente:

CUADRO 2.9
RESULTADOS DE LA APLICACIÓN DEL ALGORITMO EM

t	$\hat{\theta}$	$\theta^{(t)}$	$x_3^{(t)}$	$\theta^{(t+1)}$	$\theta^{(t)} - \hat{\theta}$	$\frac{\theta^{(t+1)} - \hat{\theta}}{\theta^{(t)} - \hat{\theta}}$
0	0,6268215	0,5	25	0,60824742	-0,1268215	0,14645841
1	0,6268215	0,60824742	29,1501976	0,62432105	-0,01857408	0,1346203
2	0,6268215	0,62432105	29,7372653	0,62648888	-0,00250045	0,13302371
3	0,6268215	0,62648888	29,8158924	0,62677732	-0,00033262	0,13281127
4	0,6268215	0,62677732	29,8263445	0,62681563	-4,4176E-05	0,13278305
5	0,6268215	0,62681563	29,8277325	0,62682072	-5,8658E-06	0,13277931
6	0,6268215	0,62682072	29,8279168	0,62682139	-7,7885E-07	0,13277881
7	0,6268215	0,62682139	29,8279413	0,62682148	-1,0341E-07	0,13277874
8	0,6268215	0,62682148	29,8279445	0,6268215	-1,3731E-08	0,13277874
9	0,6268215	0,6268215	29,827945	0,6268215	-1,8232E-09	0,13277873
10	0,6268215	0,6268215	29,827945	0,6268215	-2,4209E-10	0,13277879

2.4 MODELOS IMPLÍCITOS

El estudio se basa en algoritmos cuyas suposiciones son implícitas, y no a modelos estadístico formales, necesitándose valoración para asegurarse de su razonabilidad.

En cada algoritmo, para que un valor faltante pueda ser imputado usualmente se asume dependencia de otras variables auxiliares dentro del conjunto de datos.
Los métodos se pueden utilizar en cualquier tipo de variable. Se genera mayor variabilidad que en los métodos del modelo explícito pues provee una selección aleatoria de los valores a utilizarse en la imputación, características que hacen en cierta forma superior los modelos implícitos sobre los explícitos[3].
Adicionalmente los valores imputados por los métodos bajo este modelo preserva la distribución de los datos observados. Por lo tanto, para generar los aproximaciones no se requieren fuertes supuestos distribucionales; por consiguiente, el conjunto que incluye los datos imputados puede analizarse por métodos estadísticos tradicionales con resultados aceptables.

[3] Little/Rubin (2002)

Para este tipo de métodos se requiere alguna programación para poder implantarlos, además de información completa en las variables auxiliares. Por tanto, los estimados dependen de sobremanera de los valores de las variables auxiliares; es así que, no existe algoritmo fijo para imputar mediante estos modelos.

2.4.1 IMPUTACIÓN "HOT DECK"

Este método es un procedimiento de duplicación. Cuando falta información en un registro se duplica un valor ya existente en la muestra para reemplazarlo. Todas las unidades muestrales se clasifican en grupos disjuntos de forma que sean lo más homogéneas posibles dentro de los grupos. A cada valor que falta, se le asigna un valor del mismo grupo. Se supone que dentro de cada grupo el dato faltante sigue la misma distribución que los datos observados.

Esta suposición principal pesa en la variable de clasificación. Estas variables deberán estar bien relacionadas con los valores registrados y también deben estar bien relacionadas con los valores faltantes.

El método Hot-Deck tiene ciertas características interesantes a destacar:

1. Los procedimientos conducen a una post-estratificación sencilla. Por ejemplo: en el caso de encuestas; aunque algunas veces las personas pueden rehusarse a cooperar en una encuentra de entrevista personal, los entrevistadores pueden observar rasgos cualitativos acerca de la persona, tales como raza, sexo o rango de edad, así como variables geográficas. Estas variables pueden ser usadas como variables de clasificación.
2. No presentan problemas a la hora de encajar conjuntos de datos
3. No necesitan supuestos fuertes para estimar los valores individuales de las respuestas que faltan
4. Generalmente se conserva la distribución de las variables.

En la imputación “Hot-Deck”, los valores perdidos son reemplazados por valores correspondientes a unidades similares en la muestra. Se supone que se cuenta con n valores de un conjunto de N unidades, posteriormente al realizarse el muestreo se obtienen r valores de una muestra de n valores de una variable Y, donde n, N y r son tratados como valores fijos.

Entonces los primeros r < n datos, corresponden a los datos observados y si se asume igual probabilidad de muestreo, la media Y puede ser estimada como la media de las respuestas de las unidades imputadas. Esto es[1]:

$$\bar{y}_{HD} = \frac{r\bar{y}_R + (n-r)\bar{y}^*_{NR}}{n} \qquad \textbf{(2.35)}$$

Donde $\bar{y}_R$ es la media de las unidades de respuesta y

$$\bar{y}^*_{NR} = \sum_{i=1}^{r} \frac{H_i y_i}{n-r} \qquad \textbf{(2.36)}$$

Donde H_i es el número de veces que y_i es usado como sustituto de un valor perdido de Y, con $\sum_{i=1}^{r} H_i = n-r$ como el número de unidades perdidas.

Las propiedades de $\bar{y}_{HD}$ dependen del procedimiento usado para generar los números $\{H_1,.....,H_r\}$.

La media y la varianza de $\bar{y}_{HD}$ serían:

$$E(\bar{y}_{HD}) = E[E(\bar{y}_{HD} \mid Y_{obs})] \qquad \textbf{(2.37)}$$

$$Var(\bar{y}_{HD}) = Var[E(\bar{y}_{HD} \mid Y_{obs})] + E[Var(\bar{y}_{HD} \mid Y_{obs})] \qquad \textbf{(2.38)}$$

[3] Little/Rubin(2002)

Donde la esperanza y la varianza interna son sobre las distribuciones de $\{H_1,.....,H_r\}$ de los datos observados Y_{obs}, y la esperanza y varianza externa son sobre los modelos de las distribuciones de Y. El segundo término de la ecuación (2.32) representa la adición de la varianza desde el procedimiento de imputación estocástica.

Ejemplo 7: En el siguiente ejemplo se imputa valores perdidos para la variable Y.

CUADRO 2.10
CONJUNTO DE DATOS DEL EJEMPLO 7

Personas	Sexo	Años del Grupo	Estado Civil	Ingresos	Carro propio
1	M	2	S	50	N
2	F	1	C	80	S
3	F	2	C	90	S
4	M	2	S	60	-
5	M	2	C	40	S
6	F	1	C	20	-
7	M	1	C	30	S
8	F	2	C	-	-
9	F	2	S	100	S
10	F	1	C	40	-

Donde:
F representa femenino y M masculino.
S significa soltero y C casado.
N es No y S Si.

Aplicando la metodología:

1. Se encuentra un conjunto de variables categóricas X que se asocian con la variable Y.
2. Se forma una tabla de contingencia basada en la variable X.
3. Donde existan casos con valores perdidos de la variable Y, en una celda particular de la tabla, se toma uno o mas casos de valores observados en

la misma celda y se imputa el valor perdido por el valor observado de la variable Y;

Por tanto, la persona 4 puede ser imputada desde la persona 1 y la variable “carro propio” sería N. La persona 6 puede ser imputada desde la persona 2 y la variable “carro propio” sería S. La persona 8 puede ser imputada desde la persona 3 y la variable “ingresos” sería 90 y variable “carro propio” sería S. Finalmente, la persona 10 puede ser imputada desde la persona 2 y la variable “carro propio” sería S.
Existen diversas variantes de este método, que se describen a continuación:

2.4.1.1 IMPUTACIÓN “HOT DECK CON MUESTREO ALEATORIO SIMPLE”

Este tipo de imputación funciona para cualquier tipo de variable. Consiste en tomar un muestreo aleatorio simple con reemplazo de los valores observados en una variable y usarlos como reemplazos o sustituciones para los valores faltantes dentro de la misma variable.

Sea $\bar{y}_{HDI}$ el estimador Hot Deck descrito en la ecuación (2.35) cuando los $\{H_i\}$ son obtenidos por muestreo aleatorio simple con reemplazo desde los valores observados de la variables Y. La distribución de los $\{H_1,.....,H_r\}$ en el procedimiento “Hot Deck” es multinomial con tamaño de la muestra *n-r* y probabilidades *(1/r,.....,1/r)*

Es así que los momentos de la distribución de los $\{H_1,.....,H_r\}$ dados los datos observados de Y son[3]:

$$E(H_i \mid Y_{obs}) = \frac{(n-r)}{r} \qquad \textbf{(2.39)}$$

[3] Little/Rubin (2002)

$$Var(H_i \mid Y_{obs}) = \frac{(n-r)(1-\frac{1}{r})}{r} \quad \textbf{(2.40)}$$

$$Cov(H_i, H_j \mid Y_{obs}) = -\frac{(n-r)}{r^2} \quad \textbf{(2.41)}$$

Por tanto, tomando la esperanza de la distribución de $\bar{y}_{HD}$ sobre la distribución de $\{H_1,.....,H_r\}$:

$$E(\bar{y}_{HDI} \mid Y_{obs}) = \bar{y}_R \quad \textbf{(2.42)}$$

$$Var(\bar{y}_{HDI} \mid Y_{obs}) = \frac{(1-\frac{1}{r})(1-\frac{r}{n})s_{yR}^2}{n} \quad \textbf{(2.43)}$$

En general, si se asume muestreo aleatorio simple desde una población finita de tamaño N y datos perdidos del tipo MCAR, tenemos:

$$E(\bar{y}_{HDI}) = \overline{Y} \quad \textbf{(2.44)}$$

$$Var(\bar{y}_{HDI}) = (\frac{1}{r} - \frac{1}{N})S_y^2 + \frac{(1-\frac{1}{r})(1-\frac{r}{n})S_y^2}{n} \quad \textbf{(2.45)}$$

Donde el primer término de la varianza corresponde a la varianza del muestreo aleatorio simple de $\bar{y}_R$ y el segundo término está asociado al incremento en la varianza desde el procedimiento Hot Deck.

Una de las ventajas de este método, tal como se menciona en la introducción de mismo, es que los valores imputados no distorsionan la distribución de los valores de la muestra de la variables Y.

2.4.1.2 IMPUTACIÓN "HOT DECK" CON AJUSTES DE CELDAS

Es factible estimar el ajuste de celdas. Los valores perdidos en cada celda pueden llenarse con valores observados para la misma celda. La estimación puede hacerse a través de los pesos de las clases.

La media y la varianza de los resultados estimados por Hot – Deck de $\bar{y}$ se pueden encontrar aplicando previamente fórmulas independientes en cada celda y luego obtener combinaciones sobre las celdas. El ajuste de las celdas se consigue por niveles desde una de las variables categóricas.

2.4.1.3 IMPUTACIÓN "HOT DECK" POR VECINO MÁS CERCANO

En un intento por buscar un método general que fuera más certero en sus estimaciones se han creado métodos que hacen uso de métricas para medir distancias entre unidades, basadas en los valores de las variables asociadas dentro del mismo conjunto de datos. Posteriormente se procede a calcular las distancias e imputar los valores faltantes utilizando las unidades del conjunto de unidades completas más cercanas según la métrica.

El algoritmo del vecino más cercano imputa los valores utilizando la métrica euclidiana entre las unidades.

A continuación se explican brevemente los pasos del algoritmo:

- Se particiona el conjunto de datos D en dos partes: Las unidades completamente observadas D_c y las unidades con valores faltantes $D_{m.}$

- Para cada unidad y_i en D_m se calculan las distancias entre y_i y las unidades completas y_c, y se escogen las k unidades más cercanas según la distancia euclidiana. El conjunto escogido para y_i se llama el conjunto D_k de los vecinos más cercanos a y_i, el cual contiene valores faltantes para una o varias variables o atributos.

- Con los valores en D_k se imputan los valores faltantes en y_i, en cada variable j dependiendo del tipo de ésta. Si la variable j es del tipo continuo entonces se hace un promedio de los k vecinos en esa variable j y ese promedio pasa a ser el valor de imputación para $x_{i,j.}$ Por otro lado, si la variable es de tipo binaria u ordinal se buscará el valor que más se repita dentro de los k vecinos y ese pasará a ser el valor de imputación.

- El proceso termina cuando las unidades en D_m se han imputado completamente.

El método del vecino más cercano toma en consideración la estructura de la correlación del conjunto de datos. Sin embargo para este método no se provee un criterio específico de selección de la métrica. En lugar de la distancia euclidiana pueden emplearse las distancias de Maniatan, la de Mahalanobis y la de Pearson, entre otras[3].

Entre las desventajas del método del vecino más cercano se cita que imputa valores escogiendo k vecinos dentro de una misma clase, por lo que este método es condicional. Adicionalmente, tampoco provee un criterio específico para la selección de un número k de vecinos.

Una aproximación más general es la definición de una métrica para la medición de la distorsión entre unidades, basadas en los valores de la covarianza y entonces se escogen los valores imputados que vienen desde las unidades de respuesta a las unidades con valores perdidos.

Por ejemplo $y_i = (y_{i1},.....,y_{ik})^T$ serán los valores de k correspondientes a la covarianza de la unidad i para lo cual yi es perdido. Si estos variables fueran usadas para el ajuste de celdas, la métrica sería:

[3] Little/Rubin (2002)

$$d(i,j)=\begin{cases}0 & \text{i, j en la misma celda}\\ 1 & \text{i, j en celdas diferentes}\end{cases}$$

Otras métricas posibles son:

Desviación máxima:

$$d(i,j)=\max_k\left|y_{ik}-y_{jk}\right| \qquad \textbf{(2.46)}$$
$$d(i,j)=(y_i-y_j)^T S^{-1}_{yy}(y_i-y_j)$$

Donde Syy es un estimador de la matriz de covarianza de y_i

Finalmente, se debe considerar que si el número de vecinos es pequeño, la estimación se hará sobre una muestra pequeña y por lo tanto el efecto será una mayor varianza en la estimación. Por otro lado, si la imputación se hace a partir de un número grande de vecinos, el efecto puede ser la introducción de sesgo en la estimación por información de individuos alejados.

Por ejemplo, suponemos que en una celda en particular de la tabla de contingencia, existen n_1 casos con datos completos en la variable Y y n_o casos con valores perdidos en misma variable Y; por tanto, los pasos a seguir se resumen en:

1. Desde los n_1 casos con datos completos, se toma una muestra aleatoria simple con reemplazo de los n_1 casos.
2. Desde esta muestra, se toma una muestra aleatoria (con reemplazo) de los n_o casos.
3. Se asigna los n_o valores observados de la variable Y en los n_o casos con valores perdidos sobre la variable Y.
4. Se repiten los pasos 1 hasta al 3 para cada celda en la tabla de contingencia.

Estos cuatro pasos producen un conjunto de datos completos cuando se aplican a todas las celdas de la tabla de contingencia.

2.4.1.4 IMPUTACIÓN "HOT DECK" SECUENCIAL

El procedimiento secuencial Hot Deck se usa cuando la muestra tiene algún tipo de orden dentro de cada grupo de clasificación, en este caso, para cada uno de los valores faltantes se duplica el valor previo. Por ejemplo, el ordenamiento puede ser basado en una distribución geográfica. El resultado de un ordenamiento geográfico es que el valor registrado duplicado para un valor faltante es hecho de una unidad, la cual es geográficamente cercana a la unidad del valor faltante.

Por otro lado, se deberá tener en cuenta que si el primer registro tiene un dato faltante, este es remplazado por un valor inicial para imputar, pudiendo ser obtenido de la información externa. Si el valor no está perdido, éste será al valor inicial y es usado para imputar el subsiguiente dato faltante.

Es así que, que un valor perdido de Y es reemplazado por un valor observado cercano anterior a la secuencia. Por ejemplo, si n=6 y r=3, esto es, y_1, y_4 y y_5 son observados y y_2, y_3 y y_6 son perdidos; entonces y_2, y_3 son reemplazados por y_1 y y_6 es reemplazo por y_5. de la siguiente forma:

CUADRO 2.11
REEMPLAZO DE VALORES APLICANDO HOT DECK SECUENCIAL

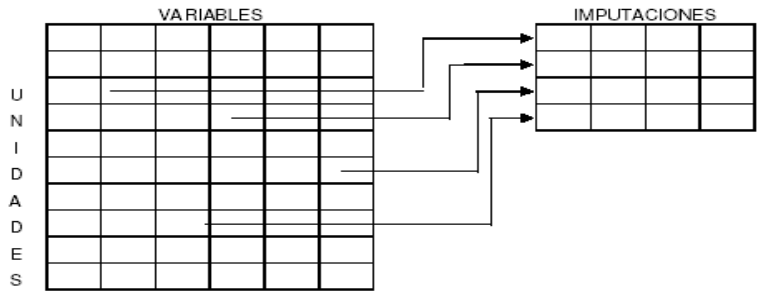
TOTAL VARIABLES	DATOS OBSERVADOS	DATOS PERDIDOS	HOT DECK SECUENCIAL
y_1	y_1		y_1
y_2		y_2	y_1
y_3		y_3	y_1
y_4	y_4		y_4
y_5	y_5		y_5
y_6		y_6	y_5

CAPÍTULO 3

IMPUTACIÓN MÚLTIPLE

La imputación múltiple es la técnica para reemplazar un valor perdido con dos o más valores aceptables, representados por una distribución de probabilidad. La figura 3.1 describe un conjunto de datos con información faltante en el cual se ha utilizado imputación múltiple, reemplazando cada valor perdido por un punto del vector de m valores posibles, que son almacenados en una matriz auxiliar con una fila para cada valor faltante y m columnas que representan la cantidad de imputaciones realizadas. Dichos valores están ordenados de manera que la primera columna de la matriz auxiliar contiene los primeros valores que se utilizan para sustituir los valores faltantes, generando así el primer un conjunto de datos "completos". La segunda columna genera el segundo conjunto de datos "completos" y así sucesivamente. En la práctica este método es útil cuando la fracción de valores perdidos no es excesiva, es decir cuando m esta entre 2 y 10.

FIGURA 3.1
CONJUNTO DE DATOS CON m IMPUTACIONES PARA CADA DATO FALTANTE



La estrategia general de la imputación múltiple se puede resumir en cuatro grandes pasos: 1) selección del método de imputación (explícito o implícito), 2) generación de conjuntos de valores imputados, 3) análisis de los diferentes conjuntos de valores imputados y 4) combinación de los resultados de estos análisis para obtener una estimación promedio.

En este capítulo se aborda la generación de imputaciones múltiples y la realización de inferencias desde el conjunto de datos imputados múltiples.

3.1 VENTAJAS Y DESVENTAJAS DE LA IMPUTACIÓN MÚLTIPLE

La imputación múltiple comparte dos ventajas básicas con la imputación simple[4]:

1. Factibilidad para generar un conjunto completo de datos para llevar a cabo cualquier tipo de análisis

2. Posibilidad para incorporar el conocimiento del analista de datos

Existen tres ventajas adicionales importantes de la imputación múltiple sobre la imputación simple:

1. Cuando las imputaciones son aleatorias y se intenta representar la distribución de los datos, la imputación múltiple incrementa la eficiencia de la estimación.

2. Cuando la imputación múltiple representa imputaciones repetidas bajo un modelo de no respuesta que implica validar inferencias, es decir, que reflejan la variabilidad adicional debido a los valores perdidos bajo algún

[4] Donald B. Rubin (2004)

modelo, estos son obtenidos simplemente por la combinación de inferencia de datos completos en una forma total.

3. En la imputación múltiple es factible la generación repetida de imputaciones aleatorias de un modelo, esto permite estudiar directamente las sensitividades de los varios modelos de no respuesta inferidos, simplemente usando métodos de datos completos.

En resumen, el método de imputación múltiple corrige la desventaja de la imputación simple con respecto al tratamiento de la variabilidad de los datos faltantes pues, como las *m* imputaciones son repeticiones bajo un modelo de predicción de los valores faltantes, el análisis en los datos completos puede combinarse fácilmente para crear inferencias que tomen en consideración la variabilidad del muestreo, y por ende la variabilidad de los valores faltantes. Además, como las *m* imputaciones provienen de más de un modelo, la incertidumbre de hallar el modelo correcto queda superada por la variabilidad en las inferencias a través de los modelos utilizados para las *m* imputaciones[3]

La imputación múltiple tiene tres desventajas con relación a la imputación simple:

1. Se requiere mayor esfuerzo computacional para procedimientos de imputación múltiple que de imputación simple.

2. Se requiere mayor espacio para almacenar un conjunto de datos de imputación múltiple. Esto es, se debe almacenar la matriz completa de datos y la matriz auxiliar. La matriz auxiliar es de tamaño igual al producto (número de imputaciones por valor perdido) X (número de valores perdidos); por ejemplo, si se supone *m* imputaciones por valor perdido y *g* el porcentaje de valores perdidos, entonces la matriz de datos auxiliares es *mg* en porcentaje más grande que la matriz de datos.

[3] Little/Rubin (2002)

3. Se requiere mayor trabajo para analizar el conjunto de datos con imputación múltiple que el conjunto de datos con imputación simple.

En resumen, la imputación múltiple presenta varios inconvenientes. La necesidad de hallar un modelo probabilístico en el conjunto de datos completos puede representar una ardua tarea, pues en la mayoría de los casos resulta muy complicado ajustar un modelo que tome en consideración la información de los valores faltantes y que al mismo tiempo consiga convergencia en los estimados. Además, contrario a la imputación simple, la imputación múltiple requiere de un gran esfuerzo computacional y de una cantidad considerable de tiempo y la disponibilidad de *software* es limitada.

3.2 FUNDAMENTOS DE LA TEORIA BAYESIANA

La justificación directa para el uso de la imputación múltiple proviene de la teoría Bayesiana, la que no solamente provee las relaciones teóricas generales sino que también analiza el cómo crear imputaciones múltiples con análisis de los datos resultantes.

3.2.1 VARIABLES EN LA POBLACIÓN FINITA

Se definen dos clases de variables en la población finita de N unidades.
La primera clase incluye variables que describen las características de las unidades que son de interés intrínsico para el investigador: covariables X y su variable de salida Y.
La segunda clase incluye indicadores de las variables que son necesarias para identificar los valores observados, los no observados y las inferencias de las cantidades de la población: indicadores de muestreo e indicadores de respuesta R.

3.2.2 COVARIABLE “X”

X se refiere a la covariables observadas totalmente, tales como indicadores de estrato o tamaños de unidades de medida, registradas para todas las N unidades en la población.

Generalmente X es un vector fila, es decir: $X = \begin{pmatrix} X_1 \\ . \\ . \\ . \\ X_N \end{pmatrix}$

Pero dependiendo del problema X puede ser una matriz en el que se encuentren involucrados muchos componentes.

3.2.3 VARIABLE DE SALIDA “Y”

Y es la variable cuyos valores no son conocidos por las unidades de la población.

La variable Y será un vector si existe una variable de salida $Y = \begin{pmatrix} Y_1 \\ . \\ . \\ . \\ Y_N \end{pmatrix}$, caso contrario puede ser una matriz.

La variable Y debe incluir solamente variables de interés primario en el estudio.

3.2.4 INDICADOR DE INCLUSIÓN “I”

I es el indicador para inclusión/exclusión de un estudio. Si existe una sola variable de salida, el indicador I es binario, esto es:

I_i=1 indicador en el que la última variable i_{th} está incluida en el estudio
I_i=0 indicador en el que la última variable i_{th} está excluida del estudio.

Cuando Y_i esté compuesta de más de una variable, es decir p variables, entonces I_i es también un vector con p componentes y por tanto se tendrán I_{ij}=1 ó un I_{ij}=0

Lo importante es considerar que I es totalmente observado.

3.2.5 INDICADOR PARA RESPUESTAS

Corresponde a la variable que indica la respuesta y la no respuesta. Si existe una sola variable de salida, el indicador R es binario, esto es:

R_i=1 indicador en el que la $i_{ésima}$ unidad respondió
R_i=0 indicador en elque la $i_{ésima}$ unidad no respondió

Cuando Y_i esté compuesta de más de una variable, es decir p variables, entonces R_i es también un vector con p componentes y por tanto se tendrán R_{ij}=1 ó R_{ij}=0.

Es importante considerar que cuando R_{ij} es conocida cada vez que I_{ij}= 1 y desconocida cuando I_{ij}=0. Esto es, el estado de la respuesta es conocido para todos los Y_{ij} incluidos en el estudio y desconocido para todos los Y_{ij} excluidos del estudio.

3.2.6 NOTACIÓN

Sea Q la cantidad de interés en el estudio de investigación; por ejemplo, la media de la población Y. Generalmente Q es un vector de k filas. Además se asume que con datos completos, la inferencia para Q deberá basarse en:

$$(Q - \hat{Q}) \ \sim \ N(0,U)$$

donde:

$\hat{Q}$ Corresponde al valor estimado del estadístico Q

U corresponde a la varianza que generalmente es una matriz de covarianza k x k de $(Q-\hat{Q})$

N(0,U) es la distribución normal de k variables con media 0 y varianza U.

Se supone que bajo el modelo bayesiano específico, el conjunto de las *m* imputaciones repetidas han sido realizadas y usada la construcción del conjunto de m datos completos, donde $\hat{Q}_{*1},.....\hat{Q}_{*m}$ **y** $U_{*1},.....,U_{*m}$ corresponden a los valores de los estadísticos $\hat{Q}$ y U encontrados para cada uno de este conjunto de datos.

La teoría que se describe a continuación se aplica directamente cuando $\hat{Q}$ y U son obtenidos de datos completos que proveen la media y la varianza de Q datos (X,Y_{com},R_{com},I) bajo el mismo modelo bayesiano para crear imputaciones repetidas.

3.2.7 COMBINACIÓN DE LA ESTIMACIÓN Y VARIANZA DE DATOS COMPLETOS REPETIDOS

A continuación se definen las expresiones para los estadísticos de datos completos $\hat{Q}$ y U.

Los *m* estimadores de datos completos repetidos y la varianza asociada de los datos completos para Q bajo un modelo de datos perdidos pueden ser combinados de la siguiente manera:

Sea la media de las *m* estimaciones de datos completos igual a:

$$\overline{Q}_m = \sum_{l=1}^{m} \frac{\hat{Q}_{*l}}{m} \qquad \textbf{(3.1)}$$

Sea el promedio de las *m* varianzas de los datos completos igual a:

$$\overline{U}_m = \sum_{l=1}^{m} \frac{U_{*l}}{m} \quad \textbf{(3.2)}$$

Sea la varianza entre las m estimaciones de datos completos igual a:

$$B_m = \sum_{l=1}^{m} \frac{(\hat{Q}_{*l} - \overline{Q}_m)^t (\hat{Q}_{*l} - \overline{Q}_m)}{m-1} \quad \textbf{(3.3)}$$

donde el superíndice t indica transposición cuando Q es un vector.

Además, se define al total de la varianza de $(Q - \hat{Q})$ como:

$$T_m = \overline{U}_m + (1 + \frac{1}{m}) B_m \quad \textbf{(3.4)}$$

3.2.8 ESCALAR Q

La estimación del intervalo y los niveles de significancia para el escalar Q son formados usando una distribución de referencia t, con:

$$\nu = (m-1)(1 + \frac{1}{r_m})^2 \quad \textbf{(3.5)}$$

Los grados de libertad, donde r_m representa el incremento relativo en la varianza debido a los datos perdidos corresponde a:

$$r_m = (1 + \frac{1}{m}) \frac{B_m}{\overline{U}_m} \quad \textbf{(3.6)}$$

Entonces, de una manera estándar, al 100(1-α)% el intervalo estimado de Q sería:

$$\overline{Q}_m \pm t_\nu (\frac{\alpha}{2}) T_m^{1/2} \quad \textbf{(3.7)}$$

Donde $t_\nu(\frac{\alpha}{2})$ está sobre 100α/2 puntos de porcentaje de la distribución t con ν grados de libertad. Por otro lado, el nivel de significancia asociado con los valores nulos de Q_o está dado por:

$$\Pr ob\left\{F_{1,\nu} > \frac{(Q_o - \overline{Q}_m)^2}{T_m}\right\} \qquad \textbf{(3.8)}$$

La fracción de información acerca de los Q valores perdidos debido a no respuesta sería:

$$\gamma_m = \frac{r_m + \frac{2}{\nu + 3}}{r_m + 1} \qquad \textbf{(3.9)}$$

3.2.9 NIVELES DE SIGNIFICANCIA BASADOS EN LA ESTIMACIÓN COMBINADA Y VARIANZA

Cuando m es mayor que k (por ejemplo m ≥ 5 k), el nivel de significancia de los valores nulos de Q_o de Q pueden ser encontrados por:

$$D_m = \frac{(Q_o - \overline{Q}_m)T_m^{-1}(Q_o - \overline{Q}_m)^t}{k} \qquad \textbf{(3.10)}$$

y permitiendo que el nivel de significancia sea la probabilidad que una variable aleatoria F sobre k y ν grados de libertad sea mayor que D_m; ν está dado por (3.5) con r_m generalizado a ser el incremento relativo promedio en la varianza debido a la no respuesta.

$$r_m = \frac{(1 + \frac{1}{m})Tr(B_m \overline{U}_m^{-1})}{k} \qquad \textbf{(3.11)}$$

Donde Tr(A) es la suma de los elementos de la diagonal en la matriz A de orden k x k. Cuando m es pequeño en comparación a k, el mejor estadístico de prueba corresponde a:

$$\tilde{D}_m = \frac{\frac{1}{(1+r_m)}(Q_o - \overline{Q}_m)\frac{1}{\overline{U}_m}(Q_o - \overline{Q}_m)^t}{k} \qquad \textbf{(3.12)}$$

referido a la distribución de probabilidad $F_{k,\frac{k+1}{2}\nu}$

Si Q es escalar, D_m , $\tilde{D}_m$ son idénticas y sus distribuciones de referencia están dadas por (3.8) para el nivel de significancia de Q_o.

3.2.10 NIVELES DE SIGNIFICANCIA BASADOS EN LOS NIVELES DE SIGNIFICANCIA DE DATOS COMPLETAMENTE REPETIDOS

Se puede calcular un equivalente asintótico estadístico de los m niveles de significancia de los datos completos asociados con Q_o y el valor r_m. Específicamente $d_{*1},.....,d_{*m}$ serán los *m* valores repetidos de los datos completos, χ^2 es el estadístico asociado con los valores nulos Q_o; esto es, el nivel de significancia para el *iésimo* conjunto de datos completos es la probabilidad que una variable aleatoria χ^2 con k grados de libertad sea mayor que d_{*1} . Entonces:

$$\hat{D}_m = \frac{\frac{\overline{d}_m}{k} - \frac{m-1}{m+1}r_m}{1+r_m} \qquad \textbf{(3.13)}$$

donde

$$\overline{d}_m = \sum_{l=1}^{m}\frac{d_{*1}}{m} = \textit{ al promedio del estadístico } \chi^2 \textit{ repetido}$$

referido a la distribución $F_{k,\frac{k+1}{2}\nu}$. El estadístico $\hat{D}_m$ es más útil que D_m o $\tilde{D}_m$ porque éste depende del escalar χ^2 asociado con $d_{*1},.....,d_{*m}$ y del escalar $r_{m.}$

En algunos casos, solamente la fracción estimada de información perdida para un componente particular de Q puede ser conocida, quizás porque un intervalo estimado es creado solamente para ese componente; usando ésta en lugar de r_m en $\hat{D}_m$, lo que implique usar ν en lugar de los grados de libertad del numerador $\frac{k+1}{2}\nu$ en la distribución de referencia para la estimación de $\hat{D}_m$, asumiendo que la fracción de información perdida sobre estos componentes es típica de otros componentes. Existe alguna información en $d_{*1},.....,d_{*m}$ acerca de r_m, pero no está claro como usarlo para obtener los *p* valores; para esto, un método de estimador de momentos de r_m, puede ser empleado para estimar tanto $\hat{D}_m$ como ν, de la siguiente forma:

$$\hat{r}_m = \frac{(1+\frac{1}{m})s_d^2}{2\bar{d}_m + [4\bar{d}_m^2 - 2ks_d^2]_+^{1/2}} \qquad \textbf{(3.14)}$$

donde:

$$s_d^2 = \frac{\sum_{l=1}^{m}(d_{*1} - \bar{d}_m)^2}{m-1} \qquad \textbf{(3.15)}$$

Los resultados de pruebas estadísticas son:

$$\hat{\tilde{D}}_m = \frac{\frac{\bar{d}_m}{k} - \frac{m-1}{m+1}\hat{r}_m}{1+\hat{r}_m} \qquad \textbf{(3.16)}$$

La cual es referida a la distribución F sobre los grados de libertad k y $\frac{1+\frac{1}{k}}{2}\hat{v}$, donde:

$$\hat{v} = (m-1)\left(1+\frac{1}{\hat{r}_m}\right)^2 \qquad \textbf{(3.17)}$$

con $\hat{r}_m$ equivalente a la ecuación (3.14).

3.3 CONDICIONES GENERALES PARA LA VALIDACIÓN DE ALEATORIEDAD DE INFERENCIAS DE *m* IMPUTACIONES REPETIDA INFINITAS

Se supone que las inferencias son realizadas desde un número finito de imputaciones múltiples bajo el modelo de uso de estadístico de datos completos estándares, esto es $\hat{Q} = \hat{Q}(X, Y_{com}, I)$ y $U = U(X, Y_{com}, I)$ y la distribución de referencia normal $\left(Q - \overline{Q}_\infty \mid X, Y_{obs}, R_{com}, I\right) \approx N\left(0, T_\infty\right)$[5]. Los resultados de las inferencias de imputación repetida se validan desde respuestas aleatorias bajo perspectivas basadas en aleatoriedad de los posibles mecanismos y mecanismos de muestreo específicos si:

$$(\overline{Q}_\infty \mid X, Y) \approx N(Q, T_o) \qquad \textbf{(3.18)}$$

y

$$(T\infty \mid X, Y) \approx (T_o, << T_0) \qquad \textbf{(3.19)}$$

Donde Q=Q(X,Y) y T_0=T_0(X,Y) son fijados por los valores verdaderos de (X,Y); $T_\infty = \overline{U}_\infty + \beta_\infty$; y $\overline{Q}_\infty, \overline{U}_\infty$ y B_∞ son funciones de (X, Y_{obs}, R_{com}, I). Las

[5] $\overline{Q}_\infty = E(Q \mid X, Y_{obs}, R_{com}, I)$, $T_\infty = V(Q \mid X, Y_{obs}, R_{com}, I)$

variables aleatorias en (3.18) y (3.19) son (R , I) y la notación << significa que si A → (B, << C) la distribución de A tiende a ser centrada en B y cada componente tiene una variabilidad substancial menor que cada componente positivo de C. Cuando (3.18) y (3.19) se presentan, se dice que la inferencia de imputaciones repetidas está validada aleatoriamente. Por tanto, la estimación del intervalo de confianza debe ser al 95% y si Q=Qo, el p-valor dado por $p-value(Q_o \mid X, Y_{obs}, R_{com}, I) = \Pr ob\{\chi_k^2 > kD_\infty\}$[5], será uniformemente distribuido en (0,1)

Es importante enfatizar que todos estos requerimientos para las inferencias de validación aleatoria son realizada por (3.18) y (3.19) bajo el mecanismo de respuesta positiva, es decir, mecanismo de muestreo específico y datos de población (X,Y) verdaderos.

3.3.1 CONDICIONES GENERALES PARA LA VALIDACIÓN DE ALEATORIEDAD

Existen dos condiciones suficientes para la inferencia de *m* imputaciones repetidas infinitas para la validación de aleatoriedad:

1. La inferencia de datos completos tiene que ser validada aleatoriamente:

$$\left(\hat{Q} \mid X, Y\right) \approx N\left(Q, U_0\right) \qquad \textbf{(3.20)}$$

$$\left(U \mid X, Y\right) \approx \left(U_{0,} << U_0\right) \qquad \textbf{(3.21)}$$

Donde Q =Q(X, Y) y U_0=U_0(X, Y) son fijados por los valores reales de X y Y, y la variable aleatoria fundamental en (3.20) y (3.21) es I con distribución de probabilidad de Pr(I | X, Y) = Pr (I | X).

---

[5] $D_\infty = \dfrac{\left(Q_o - \overline{Q}_\infty\right) T_\infty^{-1} \left(Q_o - \overline{Q}_\infty\right)^t}{k}$ *k* grados de libertad (k dimensiòn de Q)

2. El procedimiento de imputación múltiple es propio, básicamente en el sentido de que los estadísticos m-infinitos $\left(\overline{Q}_\infty, \overline{U}_\infty, B_\infty\right)$ produzcan inferencias para la validación de aleatoriedad para los estadísticos de datos completos $\hat{Q}$ y U, bajo el mecanismo de respuesta positiva.

3.4 MÉTODOS DE IMPUTACIÓN MÚLTIPLE PROPIOS

Un procedimiento de imputación múltiple es propio para el conjunto de datos estadísticos completos $\left\{\hat{Q}, U\right\}$, si las tres condiciones son satisfechas:

1. El conjunto (X,Y,I) es fijo, bajo los posibles mecanismos de respuesta, los m=∞ procedimientos de imputación múltiple proveen inferencias de validación aleatorias para los estadísticos de datos completos $\hat{Q} = \hat{Q}(X, Y_{inc}, I)$ basados en los estadísticos $\overline{Q}_\infty$ y B_∞ :

$$(\overline{Q}_\infty \mid X, Y, I) \approx N(\hat{Q}, B) \qquad \textbf{(3.22)}$$

$$(B_\infty \mid X, Y, I) \approx (B, << B) \qquad \textbf{(3.23)}$$

Donde B=B (X, Y_{inc}, I) definida por:

$$B = V(\overline{Q}_\infty \mid X, Y, I) \qquad \textbf{(3.24)}$$

2. El conjunto (X,Y,I) es fijo, bajo los posibles mecanismos de respuesta, las m=∞ imputaciones estimadas de los estadísticos de datos completos U (X, Y_{inc}, I), esto es, $\overline{U}_\infty$ es centrada en U con variabilidad de menor orden que $\overline{Q}_\infty$:

$$\overline{U}_\infty(X, Y, I) \approx (U, << B) \qquad \textbf{(3.25)}$$

3. El conjunto (X,Y) es fijo, sobre muestras repetidas la variabilidad de B es de menor orden que $\hat{Q}$:

$$(B \mid X,Y) \approx (B_O, << U_O) \qquad \textbf{(3.26)}$$

Donde $B_o=B_o(X, Y)$ está definida por:

$$B_O = E(B \mid X,Y) \qquad \textbf{(3.27)}$$

y $U_o=U_o(X,Y)$ es definida por (3.21)

La variable aleatoria fundamental en (3.22) – (3.25) es R con distribución especificada por el mecanismo de respuesta Pr(R | X, Y), mientras que la variable aleatoria fundamental en (3.26) y (3.27) es I con distribución de probabilidad Pr(I|X).

Se observa que (3.20) y (3.22) implican:

$$\left(\overline{Q}_\infty \mid X,Y\right) \approx N\left(Q, U_0 + E\left(E(B \mid X,Y)\right)\right) \qquad \textbf{(3.28)}$$

Empleando la definición de la ecuación (3.27) se obtiene:

$$\left(\overline{Q}_\infty \mid X,Y\right) \approx N\left(Q, U_0 + B_0\right) \qquad \textbf{(3.29)}$$

De la misma forma con las ecuaciones (3.23) y (3.25) se obtiene:

$$\left(\overline{U}_\infty + B_\infty \mid X,Y,I\right) \approx (U + B, << 2B) \qquad \textbf{(3.30)}$$

Lo mismo ocurre con las ecuaciones (3.21) y (3.26):

$$\left(\overline{U}_{\infty} + B_{\infty} \mid X, Y\right) \approx \left(U_0 + B_0, << \left(2B_0 + 2U_0\right)\right) \quad \textbf{(3.31)}$$

3.5 MÉTODOS DE IMPUTACIÓN PROPIA E IMPROPIA CON DATOS PERDIDOS IGNORABLES

Se considera una muestra aleatoria simple de tamaño n para una población de N valores con media finita $Q = \overline{Y}$ y varianza S^2. Con datos perdidos, la inferencia para $\overline{Y}$ se basa en los estadísticos de datos completos estándares $\hat{Q} = \bar{y}$, la media de la muestra, y $U = s^2(\frac{1}{n} - \frac{1}{N})$, donde s^2 es la varianza de la muestra. Se asume que N y n son los bastante grandes (tal que se cumpla la ley de los grandes números) para llevar a cabo la inferencia de datos completos estándares, basados en la validación de la aleatoriedad de $\hat{Q}$ y U, de modo que se cumplen las ecuaciones (3.20) y (3.21) con $U_0 = S^2(\frac{1}{n} - \frac{1}{N})$.

Debido al gran volumen de datos perdidos en la que solamente n_1 (n_1 es también grande) de los n valores en Y_{com} son observados, se trata por simplicidad como valor fijo. Además la media muestral observada es $\bar{y}_1$ y la varianza muestral es s_1^2. Por tanto, n_1 y n son grandes, de allí que la muestra se trata como fija y se promedia sobre los mecanismos de respuesta, los que son semejantes al nivel de muestreo aleatorio simple desde los n valores incluidos en Y_{com} dados por (3.20) y (3.21)

$$(\bar{y}_1 \mid X, Y, I) \approx N\left(\bar{y}, s^2\left(\frac{1}{n_1} - \frac{1}{n}\right)\right) \quad \textbf{(3.32)}$$

$$(s_1^2 \mid X, Y, I) \approx (s^2 << S^2) \quad \textbf{(3.33)}$$

3.5.1 IMPUTACIÓN MULTIPLE ALEATORIA SIMPLE

Se toma como referencia la imputación simple “Hot Deck”, para revisar el proceso de imputación múltiple, con el ejemplo siguiente:

Se supone que se cuenta con una muestra aleatoria simple de n=10 unidades de una población de N=1000 unidades. Se conoce el valor de la covariable X (cuyo tamaño es de 1970) para cada una de las N unidades y se trata de realizar una observación de la variable Y (cuyo tamaño es de 1980) para cada uno de las n unidades incluidas en la muestra, pero dos unidades son datos perdidos. El objetivo del estudio es estimar $\bar{Y}$, es decir, la media de Y en la población.

Se asume que con datos completos, el estimador $\frac{\bar{X}\ \bar{y}}{\bar{x}}$ debe ser usado con un intervalo de confianza del 95%, esto es: $\frac{\bar{X}\ \bar{y}}{\bar{x}} \pm 1.96\frac{\sigma^2}{\sqrt{n}}$, donde $\bar{X}$ es conocida como la media de la población, que para el ejemplo es 12. Adicionalmente $\bar{x}$ y $\bar{y}$ son las medias de las variables X y Y producto del muestreo aleatorio de las n unidades; y

$$\sigma^2 = \sum_{i=1}^{n} \frac{(Y_i - \frac{X_i \bar{y}}{\bar{x}})^2}{n-1} \qquad \textbf{(3.34)}$$

La tabla 3.1 presenta los valores de Y_i , X_i para las 10 unidades de la muestra, donde ‘-‘ representan los datos perdidos.

CUADRO 3.1
DATOS OBSERVADOS

No.	Y	X
1	10	8
2	-	9
3	14	11
4	-	13
5	16	16
6	15	18
7	20	6
8	4	4
9	18	20
10	22	25

En este ejemplo se creará la imputación múltiple a través de muestreo aleatorio simple con reemplazo desde los n_1 valores observados.

Debido a que $Q_\infty = \bar{y}_1$, y en conjunción con la ecuación (3.32), se obtiene:

$$(\overline{Q}_\infty \mid X,Y,I) \approx N(\hat{Q},B) \qquad \textbf{(3.35)}$$

Donde B, está definido por $V(\overline{Q}_\infty \mid X,Y,I)$, y corresponde a:

$$B = s^2\left(\frac{1}{n_1} - \frac{1}{n}\right) \qquad \textbf{(3.36)}$$

Con lo cual se satisface la ecuación (3.22)
Adicionalmente se puede observar que el promedio de la varianza de la muestra de los datos completos es:

$$s_1^2\left(1-\frac{1}{n_1}\right)\left[1+\frac{n_1}{n}\left(\frac{1}{n-1}\right)\right] \qquad \textbf{(3.37)}$$

Pero para muestras grandes en el que n_1 y n son grandes:

$$\overline{U}_\infty = s_1^2\left(\frac{1}{n}-\frac{1}{N}\right) \qquad \textbf{(3.38)}$$

y de la ecuaciones (3.33) y (3.36) se observa que:

$$(\overline{U}_\infty \mid X,Y,I) \approx (U, << \mathrm{B}) \qquad \textbf{(3.39)}$$

Donde,

$$U = s^2\left(\frac{1}{n}-\frac{1}{N}\right) \qquad \textbf{(3.40)}$$

Con ello se satisface lo establecido en la ecuación (3.25), al igual que la ecuación (3.26) para n grande. Por tanto, empleando las ecuaciones (3.36) y (3.40) se deduce que:

$$V(B \mid X,Y) << E(U \mid X,Y) \qquad \textbf{(3.41)}$$

Estos resultados consideran la ecuación (3.23), esto es, que a través de la imputación múltiple se considera B_∞ y la varianza de datos completos estimados $\bar{y}_{*l}$, l=1,2,….. igual a B. Por otro lado, la varianza actual de los datos completos estimados $\overline{Q}_\infty = \lim_{m\to\infty}\sum_1^m \frac{\bar{y}_{*l}}{m}$ sobre los mecanismos de respuesta está dado por (3.36), por lo tanto se tiene que:

$$B_\infty = \left(1-\frac{n_1}{n}\right)\left(1-\frac{1}{n_1}\right)\frac{s_1^2}{n} \qquad \textbf{(3.42)}$$

Así desde (3.36), para n grande se obtiene que:

$$E(B_{\infty} \mid X,Y,I) = \frac{Bn_1}{n} \qquad \textbf{(3.43)}$$

Entonces, B_{∞} subestima a B por la tasa de respuesta y por ende la ecuación (3.23) no se cumple.

Por tanto, se puede concluir que la imputación múltiple hot deck con muestreo aleatorio simple de datos perdidos desde datos observados, es impropia para $\left\{\bar{y}, s^2\left(\frac{1}{n}-\frac{1}{N}\right)\right\}$ y para cualquier población de valores de Y, debido a que no cuenta con suficiente variabilidad entre imputaciones.

Para un m grande, se tiene:

$$E(T_{\infty} \mid X,Y) = E(\overline{U}_{\infty} \mid X,Y) + E(B_{\infty} \mid X,Y) \qquad \textbf{(3.44)}$$

De las ecuaciones (3.36),(3.38) y (3.40) es factible obtener:

$$E(T_{\infty} \mid X,Y) = S^2\left(\frac{1}{n}-\frac{1}{N}\right) + S^2\frac{\left(1-\frac{n_1}{n}\right)}{n} \qquad \textbf{(3.45)}$$

$$V(\overline{Q}_{\infty} \mid X,Y) = S^2\left(\frac{1}{n_1}-\frac{1}{N}\right) \qquad \textbf{(3.46)}$$

$$E(T_{\infty} \mid X,Y) = V(\overline{Q}_{\infty} \mid X,Y) - \left(1-\frac{n_1}{n}\right)S^2\left(\frac{1}{n_1}-\frac{1}{n}\right) \qquad \textbf{(3.47)}$$

Con lo cual se demuestra que T_{∞} subestima a $T_0 = V(\overline{Q}_{\infty} \mid X,Y)$.

3.5.2 IMPUTACIÓN REPETIDA BAYESIANA NORMAL

Considerando el método de imputaciones repetidas normales, se desprende que $\overline{Q}_\infty = \bar{y}_1$, de modo que las ecuaciones (3.35) y (3.36) se cumplen y entonces la ecuación (3.22) se verifica. Adicionalmente, se observa que para muestras grandes, el promedio de la varianza de los datos completos es igual a s_1^2, de modo que (3.38) y (3.40) se cumplen y entonces las ecuaciones (3.25) y (3.26) se satisfacen. Trabajando en la parte algebraica se obtiene para muestras grandes:

$$B_\infty = s_1^2\left(\frac{1}{n_1} - \frac{1}{n}\right) \qquad \textbf{(3.48)}$$

el cual es un estimador insesgado para B con variabilidad de orden baja, por tanto, satisface la condición (3.23).

Finalmente la imputación repetida bayesiana normal es propia para $\left\{\bar{y}, s^2\left(\frac{1}{n} - \frac{1}{N}\right)\right\}$ en muestras bastante grandes con datos perdidos para cualquier población de valores de Y.

Sustituyendo (3.38) y (3.48) en (3.44) se obtiene:

$$E(T_\infty \mid X, Y) = V(\overline{Q}_\infty \mid X, Y) \qquad \textbf{(3.49)}$$

Este procedimiento sugiere que cuando la inferencia estándar es usada con datos completos y los datos perdidos son ignorables, entonces la imputación múltiple deberá ser realizada con repeticiones bajo un modelo bayesiano normal.

3.5.3 IMPUTACIÓN BOOTSTRAP BAYESIANA (BB)

Un paso simple para generar imputaciones repetidas de Y_{per} consiste en repetir los siguientes dos pasos m veces independientemente. Por simplicidad en la notación $Y_{obs} = (Y_1,.....Y_{n_1})$.

***Paso 1*:** Extraer $n_1 - 1$ números aleatorios entre 0 y 1, y luego ordenar estos valores $a_1,.....,a_{n_1-1}$; también se toma $a_0 = 0$ y $a_{n_1} = 1$.

***Paso 2*:** Extraer cada uno de los valores perdidos n_o en $Y_{per,}$ por sorteo desde $Y_1,.....,Y_{n_1}$ con probabilidades $(a_1 - a_0),(a_2 - a_1),.....(1 - a_{n_1-1})$, esto es, independientemente n_o veces, adicionalmente extraer un número aleatorio uniforme u e imputar Y_i si $a_{i-1} < u \leq a_i$

Realizando un tratamiento algebraico para muestras grandes, se obtiene: $\overline{Q}_\infty = \bar{y}_1$, $\overline{U}_\infty = s_1^2\left(\frac{1}{n} - \frac{1}{N}\right)$, $B_\infty = s_1^2\left(\frac{1}{n_1} - \frac{1}{n}\right)$**.** Adicionalmente las imputaciones de Bootstrap son asintóticamente propias $\left\{\bar{y}, s^2\left(\frac{1}{n} - \frac{1}{N}\right)\right\}$ con datos perdidos.

3.5.4 APROXIMACIÓN BOOTSTRAP BAYESIANA (ABB)

Los métodos de imputación no Bayesiana pueden también ser propios si estos incorporan apropiadamente la variabilidad entre imputaciones. Tales métodos se construyen fácilmente como modificación de los métodos Bayesianos que generan repeticiones aproximadas. La imputación por aproximación bootstrap Bayesiana, primeramente extrae los n_1 valores aleatorios con reemplazo desde las Y_{obs} para crear los Y_{obs}^*, y entonces calcula los $n_0 = n - n_1$ componentes de los Y_{per} perdidos aleatoriamente con reemplazo desde Y_{obs}^*.

La diferencia entre la ABB y los BB se fundamenta en los parámetros de los datos, los cuales dan las probabilidades de cada componente en Y_{obs}. Estas distribuciones tienen las mismas medias y correlaciones, pero se diferencian en las varianzas solamente por el factor $\left(1+\frac{1}{n_1}\right)$. Consecuentemente, la imputación ABB repetida es también un método propio para $\left\{\bar{y}, s^2\left(\frac{1}{n}-\frac{1}{N}\right)\right\}$.

3.5.5 MEDIA Y VARIANZA AJUSTADA AL MÉTODO HOT-DECK

Modificaciones menores del método de imputación no Bayesiana al método de imputación normal dado en el punto 3.5.2 dan como resultado la media y la varianza (MV) ajustada al método hot deck.

En primer lugar se debe extraer la media μ y la desviación estándar σ^2. Posteriormente se modifica μ de modo que ésta sea cero; de la misma forma, se modifica σ^2 de modo que ésta sea 1 (por ejemplo para cada componente en Y_{obs} restar $\bar{y}_1$ y dividir por $s_1\sqrt{1-\frac{1}{n_1}}$). La ventaja potencial de esta modificación sobre la imputación simple es que los valores generados conservan la misma forma distribucional de los valores observados en Y_{obs}.

Por ejemplo, si el componente de Y_{obs} es sesgado, entonces el método de media y varianza tiende a imputar valores en Y_{per} con los mismos sesgos, mientras que el método de imputación simple tiende a imputar valores en Y_{per} que son simétricos y no importa la forma de distribución en Y_{obs}.

El método de media y varianza es asintóticamente propio $\left\{\bar{y}, s^2\left(\frac{1}{n}-\frac{1}{N}\right)\right\}$.

3.6 EVALUACIÓN DE LOS NIVELES DE SIGNIFICANCIA DESDE LOS ESTADÍSTICOS BASADOS EN LOS MOMENTOS D_m Y $\tilde{D}_m$ CON ESTIMACIÓN DE MULTICOMPONENTES

3.6.1 NIVELES DE UN PROCEDIMIENTO DE PRUEBA SIGNIFICANTE

Desde la perspectiva basada en la aleatoriedad de repuesta, bajo la hipótesis nula de que $Q = Q_0$, los *p* valores deberían tener una distribución uniforme dada por: (X,Y), el promedio sobre (R,I) y el conjunto de imputaciones múltiples. Específicamente se deben considerar los estadísticos D_m y $\tilde{D}_m$, los valores nulos $Q = Q_0$ y los porcentajes de los puntos 100α de la distribución de referencia, $F_{k,k'\nu}(\alpha)$ donde k' es 1 para D_m y $\frac{k+1}{2}$ para $\tilde{D}_m$**.** Entonces el procedimiento tiene el nivel correcto α si:

$$\Pr ob\{D > F_{k,k'\nu}(\alpha) \mid X, Y; Q = Q_0\} = \alpha \qquad \textbf{(3.50)}$$

Valores referenciales de α que son de particular interés común incluyen 10%, 5% y 1%.

3.6.2 NIVELES DE D_m - ANÁLISIS PARA MÉTODOS DE IMPUTACIÓN PROPIA Y MUESTRAS GRANDES

El procedimiento para α, que usa el estadístico D_m y la distribución de referencia $F_{k,\nu}$ está dada en la ecuación (3.50) con $D = D_m$ y k'=1. De manera general, y asumiendo: método de imputación propia, inferencia de datos completos validos y

la validación de distribuciones de muestreo asintóticas de (Q_{*l}, U_{*l}), se encuentra una probabilidad igual a:

$$\Pr ob\left\{ Z_0 \frac{1}{\left[I + \left(1 + \frac{1}{m} \right) W \right]} Z_0^t > kF_{k,\nu}(\alpha) \right\} \quad \textbf{(3.51)}$$

donde:

$$\nu = (m-1)\left[1 + \frac{1}{\left(1 + \frac{1}{m} \right)} \frac{k}{Tr(W)} \right]^2 \quad \textbf{(3.52)}$$

$$Z_0 \approx N\left((0, I + \left(1 + \frac{1}{m} \right) \rho_0 \right) \quad \textbf{(3.53)}$$

$$W = \sum_{i=1}^{m-1} \frac{Z_i^t Z_i}{m-1} \quad \textbf{(3.54)}$$

$$Z_i \overset{i.i.d}{\approx} N(0, \rho_0) \text{ independiente de } Z_0 \quad \textbf{(3.55)}$$

y,

ρ_0 es una matriz diagonal k x k de valores propios de B_0 con relación a U_0, $\rho_0 = diag(\rho_{01}, \ldots\ldots\ldots\ldots, \rho_{0k})$

La expresión (3.51) es evaluada por métodos de Monte Carlo como una función de α, m , k y ρ_0 por estimación de la variable aleatoria normal k variada en cada replicación.

3.6.3 Nivel de D_m - Resultados Numéricos

Al observar la ecuación (3.51) se encuentra que el nivel de D_m puede ser relativamente insensible a la variabilidad en los valores propios $\rho_{01},.....,\rho_{0k}$, los que son equivalentes a la fracción de la población de información perdida debido a la no respuesta, $\gamma_{01},.....,\gamma_{0k},\gamma_0 = \rho_0 \frac{1}{I+\rho_0}$.

Si m es bastante grande, es claro que desde (3.51) el nivel de D_m será exacto; únicamente resta la pregunta de saber qué valor de m constituye los bastante grande, y se encuentra en el Anexo No. 1, a través de la que es posible encontrar una regla aproximada para que el nivel de D_m sea exacto, y esto se da cuando $m > 10\gamma_0 k$. Por tanto, el 50% de información perdida es una situación extrema. Se indica además, que la regla de D_m es exacta si $m \geq 5k$.

3.6.4 Nivel de $\tilde{D}_m$ - Análisis

El procedimiento para α, que usan el estadístico $\tilde{D}_m$ y la distribución de referencia $F_{k,k'\nu}$ está dada por la ecuación (3.50) con $D = D_m$. La probabilidad puede ser escrita de acuerdo a la notación de la ecuación (3.51) como:

$$\Pr ob\left\{\frac{1}{k+Tr(W)} Z_0 Z_0^t > F_{k,k'\nu}(\alpha)\right\} \qquad \textbf{(3.56)}$$

donde:

$$Z_0 Z_0^t = \sum_{i=1}^{k}\left[1+\left(1+\frac{1}{m}\right)\rho_{oi}\right]\chi_{l,i}^2 \qquad \chi_{\text{l,i}}^2 \overset{i.i.d}{\sim} \chi_i^2 \qquad \textbf{(3.57)}$$

$$Tr(W) = \sum_{i=1}^{k}\frac{\rho_{oi}\chi_{m-1}^2}{m-1} \qquad \chi_{\text{m-1,i}}^2 \overset{i.i.d}{\sim} \chi_{m-1}^2 \qquad \textbf{(3.58)}$$

Si algunos de los valores propios ρ_{0i} son iguales, el número de variables aleatorias χ^2 pueden reducirse. Por ejemplo, en el caso extremo cuando todos los $\rho_{0i} = \rho_0$, se encuentra:

$$Z_0 Z_0^t = \left[1 + \left(1 + \frac{1}{m}\right)\rho_{0.}\right]\chi_K^2 \qquad \textbf{(3.59)}$$

$$Tr(W) = \frac{\rho_{0.}}{m-1}\chi^2_{k(m-1)} \qquad \textbf{(3.60)}$$

y entonces (3.56) puede ser evaluado por:

$$\Pr ob\left\{\frac{\left[1 + \left(1 + \frac{1}{m}\right)\rho_{0.}\right]\frac{\chi_k^2}{k}}{1 + \frac{\rho_o}{k(m-1)}\chi^2_{k(m-1)}}\right\} > F_{k,k'\nu}(\alpha) \qquad \textbf{(3.61)}$$

Se debe considerar que si el 50% de información es perdida, $\rho_{0i} = 1$ y si no existe información perdida $\rho_{oi} = 0$.

3.7 PROCEDIMIENTOS CON RESPUESTAS QUE NO PUEDEN SER IGNORABLES

En la teoría revisada hasta el momento se enfocan casos sencillos con referencia al muestreo aleatorio simple de una variable de salida, sin covariables y con datos perdidos con respuestas no ignorables y donde el objetivo es estimar la media de la población usando inferencias estándares. A continuación se revisarán casos generales que consideren muestras grandes y mecanismos de muestreo.

3.7.1 MODELO DE REGRESIÓN LINEAL NORMAL CON UNA VARIABLE DE SALIDA Y_i

El método más común de predicción de una variable Y_i a través de un grupo de predictores X_i, corresponde a la regresión lineal:

$$Y_i \approx N(X_i\beta, \sigma^2) \quad \textbf{(3.62)}$$

La expresión anterior constituye la especificación para $f(Y_i \mid X_i, \theta)$, $\theta = (\beta, \log\sigma)$ donde β es un vector de q componentes y σ un escalar.

Modelación.- continuando con la tarea de modelación, se asume que

- θ es impropio
- Pr(θ) constante
- $n_1 > q$, donde n_1 es el número de repuestas

Estimación.- Se debe considerar que la distribución posterior de θ involucra solamente las unidades con Y_i observadas.
Además, a través de los cálculos bayesianos estándares para modelos lineales normales se obtiene que σ^2 es $\hat{\sigma}_1^2(n_1 - q)$ dividida por $\chi^2_{n_1-q}$ y β dado σ^2 es normal con media $\hat{\beta}_1$ y matriz de varianza – covarianza $\sigma^2 V$, donde en términos de los estadísticos de los mínimos cuadrados de los n_1 vectores (Y_i, X_i), i $\in$ obs, corresponden a:

$$\hat{\sigma}_1^2 = \sum_{obs} \frac{(Y_i - X_i\hat{\beta}_1)^2}{n_1 - q} \quad \textbf{(3.63)}$$

$$\hat{\beta}_1 = V\left[\sum_{obs} X_i^t Y_i\right] \quad \textbf{(3.64)}$$

donde:

$$V = \frac{1}{\left[\sum X_i^t X_i\right]} \qquad \textbf{(3.65)}$$

Imputación.- mediante los siguientes pasos a seguir en la imputación del modelo:

1. Sea g una variable aleatoria $\chi^2_{n_1-q}$,tal que:

$$\sigma_*^2 = \hat{\sigma}_1^2 \frac{(n_1 - q)}{g} \qquad \textbf{(3.66)}$$

2. Sea q la variable independiente N(0,1), para obtener el vector de q componentes de Z, se procede a calcular:

$$\beta_* = \hat{\beta}_1 + \sigma_* \sqrt{V} Z \qquad \textbf{(3.67)}$$

3. Calcular los valores n_0 de los Y_{per} como:

$$Y_{i*} = X_i \beta_* + z_i \sigma_* \qquad \textbf{(3.68)}$$

donde las desviaciones normales n_0 de z_i son independientes

Un nuevo valor imputado para Y_{per} es iniciado por la estimación de un nuevo valor del parámetro σ_*^2. Entonces, si *m* imputaciones repetidas son deseadas, estos tres pasos deberán ser ejecutados m veces independientemente.

3.7.2 MODELO DE REGRESIÓN LOGÍSTICA PARA VARIABLES DE SALIDA DICOTOMICAS Yi

Se supone que la variable Y_i es dicotómica (0 - 1) y que:

$$f(Y_i \mid X_i, \theta) = \text{logit}^{-1}(X_i\theta)^{\gamma_i}\left[1-\text{logit}^{-1}(X_i\theta)\right]^{1-\gamma_i} \quad \textbf{(3.69)}$$

donde la función inversa logit es:

$$\text{logit}^{-1}(a) = \frac{\exp(a)}{1+\exp(a)} \quad \textbf{(3.70)}$$

por tanto, la función logit es:

$$\text{logit}(a) = \log\left(\frac{a}{1-a}\right) \quad \textbf{(3.71)}$$

Donde θ es el vector columna con el mismo número de componentes de Xi.

En la práctica común se usa la aproximación normal para muestras grandes asumiendo que la probabilidad Pr(θ) es constante y así aproximar la media de θ , $E(\theta \mid X, Y_{obs})$, al estimador de máxima verosimilitud $\hat{\theta}$ definido por:

$$\prod_{i\in obs} f(Y_i \mid X_i, \hat{\theta}) \geq \prod_{i\in obs} f(Y_i \mid X_i, \theta) \text{ para todo } \theta \quad \textbf{(3.72)}$$

Y la varianza posterior de θ, $V(\theta \mid X, Y_{obs})$ por la inversa negativa de la segunda derivada de la matriz de la distribución log-posterior en $\theta = \hat{\theta}$:

$$\hat{V}(\hat{\theta}) = \left[\frac{\partial^2}{\partial\theta\,\partial\theta}\log\prod_{i\in \text{obs}} f(Y_i \mid X_i, \theta)\Big|_{\theta=\hat{\theta}}\right]^{-1} \quad \textbf{(3.73)}$$

Usando estas aproximaciones y calculando $\hat{\theta}$ y luego $V(\hat{\theta})$, se define el proceso de estimación de la siguiente manera:

1. Calcular θ desde $N(\hat{\theta},\hat{V}(\hat{\theta}))$, a este calculo se le identifica como θ_*
2. Para $i \in per$ se calcula la función $\log it^{-1}(X_i\theta_*)$
3. Calcular n_0 a través de los números aleatorios de N (0,1), u_i, $i \in per$

Estos pasos se repiten para cada una de las m imputaciones con los nuevos cálculos de los números aleatorios.

3.7.3 PATRONES DE MONOTONÍA DE DATOS PERDIDOS EN VARIABLES DE SALIDA MULTIVARIABLE Yi

Existen muchos casos prácticos en los cuales la creación de imputaciones múltiples con variables de salida tipo multivariante Y_i presentan algo de dificultad con respecto a los casos con una variable de salida Y_i. Usualmente estos casos involucran patrones especiales de datos perdidos en Y llamados monótonos.

Con el fin de contar con una clara definición de datos perdidos monótonos, se considerarán como valores perdidos todos los valores no observados para una unidad incluida en un estudio. Además se asumirá esencialmente que el mecanismo de muestreo es tal que si la unidad $i_{ésima}$ es seleccionada para la inclusión en el estudio, es decir, todos los componentes de Y_i están destinados a ser observados, estos es I_i=(1,……..,1) o I_i=(0,……,0) y así cualquier valor no observado para una unidad incluida en la muestra es debido a su no respuesta. Después de asumir respuestas no ignorables tanto a los valores no observados en Yi debido a mecanismos de muestreo ignorables, $Y_{i,exc}$ y a valores no observados en Y_i debido a no respuestas ignorables $Y_{i,per}$, estos pueden ser imputados desde la modelación de la distribución de (X,Y) necesaria para imputar $Y_{i,per}$ o las imputaciones para $Y_{i,exc}$ siempre pueden ser ignoradas.

3.7.3.1 DEFINICIÓN – DATOS PERDIDOS MONÓTONOS EN Y_i

La figura 3.2 presenta un ejemplo de patrones de de datos perdidos en Y monótonos. 1 significa “Observados” y 0 “Perdidos”.

FIGURA 3.2
PATRONES MONÓTONOS DE DATOS

	Variables						
Unidades	1	1	1	1	1	1	1
	1	1	1	1	1	1	0
	1	1	1	1	1	0	0
	1	1	1	0	0	0	0
	1	1	0	0	0	0	0

Como se puede observar, el primer componente de Y es al menos tan observado como el segundo, el cual es al menos tan observado como el tercer componente y así sucesivamente. Tal patrón de datos perdidos o una aproximación a la misma no es común en la práctica.

3.7.3.2 DESCRIPCIÓN DE TÉCNICAS GENERALES PARA PATRONES DE DATOS MONÓTONOS GENERALES

A fin de describir los procedimientos para patrones monótonos generales, se realiza su notación referida a una columna en particular de Y de la siguiente manera $Y_{[j]} = \left(Y_{1j}, ..., Y_{Nj}\right)^T$, de tal forma que al referirse a un patrón monótono general $Y_{[1]}$, es al menos tan observado como $Y_{[2]}$ y así sucesivamente hasta que el último valor es al menos tan observado como el $Y_{[p]}$, tal como se refleja en la Figura 3.2.

El procedimiento general recomendado es el siguiente:

- El valor perdido de $Y_{[1]}$ es imputado desde los X ignorados de los otros componentes de Y; el valor perdido $Y_{[2]}$ es imputado desde los ($Y_{[1]}$, X) ignorados de los otros componentes de Y, y así sucesivamente.

- Cada una de estos modelos de imputaciones puede ser independientemente aplicado usando métodos desarrollados para una variable de salida Yi. Además el modelo usado puede variar en el tipo tal como se explica en el siguiente ejemplo:

    - El modelo usado para imputar $Y_{[1]}$ desde X puede ser un modelo implícito tal como el método de hot – deck de media y varianza, visto en el punto 3.5.5.

    - El modelo usado para imputar $Y_{[2]}$ desde (X, $Y_{[1]}$) puede ser un modelo de regresión lineal explicito tal como el descrito en el punto 3.7.1.

    - El modelo usado para imputar $Y_{[3]}$ desde (X, $Y_{[1]}$, $Y_{[2]}$) puede ser un modelo de regresión logística explicita, tal como lo revisado en el punto 3.7.2.

3.7.3.3 MODELO DE IMPUTACIÓN IMPLÍCITA CON DOS VARIABLES DE SALIDA Y_i

Para explicar como se puede realizar este tipo de imputaciones, se provee un ejemplo con nueve unidades y un patrón de datos perdidos monótonos, que a continuación se presenta[2]:

[3] Little/Rubin[2002]

CUADRO 3.2
PATRONES MONÓTONOS DE DATOS PERDIDOS

i	X_i	Y_{i1}	Y_{i2}
1	1	1	1
2	1	1	0
3	1	0	0
4	1	1	1
5	2	1	2
6	1	1	2
7	2	2	1
8	2	1	-
9	2	-	-

El procedimiento usado es el siguiente:

- Primero el valor perdido $Y_{[1]}$ se imputa dos veces usando todos los valores observados de X e ignorando los valores no observados de $Y_{[2]}$; puesto que $X_9=2$ y $X_5=X_7=X_8=2$ entonces la quinta, séptima y octava unidad son posible donadores para la novena unidad; dos valores imputados son creados para cada valor perdido por procedimiento de hot-deck con aproximación bootstrap bayesiana desde $Y_{51}=1$ $Y_{71}=2$ y $Y_{81}=1$ encontrándose que pueden ser 1 y 2 para los dos valores imputados de Y_{91}.

- A continuación, los dos valores perdidos en $Y_{[2]}$, esto es Y_{i2} para i=8 e i=9 son cada uno imputados dos veces usando el método para X y $Y_{[1]}$ observados totalmente. Para crear el primer conjunto de valores imputados Y_{i2}, se considera el primer conjunto de valores imputados Y_{i1} como datos reales, esto es se supone primeo que Y_{91} =1. Entonces existe dos unidades que son compatible con la unidad 8 en (X, $Y_{[1]}$), estas son la quinta y novena unidad, pero solamente la quinta unidad tiene $Y_{[2]}$ observado, de modo que este valor $Y_{52}=2$ es imputado para $Y_{82.}$

- De manera similar, existen dos unidades que son compatibles con el valor de novena unidad, considerando (X, $Y_{[1]}$), y estas son la quinta y octava unidad, pero solamente la quinta unidad tiene $Y_{[2]}$ observado, de modo que este valor Y_{52}=2 es imputado para $Y_{92.}$

- Para crear el segundo conjunto de valores imputados Y_{i2}, se considera el segundo conjunto de valores imputados Y_{i1} como datos reales, esto es, se supone primeo que Y_{91} =2. Entonces la octava unidad tiene solamente la quinta unidad como un donador compatible para $Y_{[2]}$ y una vez obtenido este valor, el valor imputado es Y_{82}=2. Ahora la novena unidad tiene X_9=2, Y_{91}=2 y es compatible con la unida 7, cuyo donante es 1 como el valor imputado de $Y_{92.}$

En conclusión, las imputaciones deberán tener el mismo comportamiento en diferentes pasos, es así que la primera imputación encuentra el primer valor para cada valor perdido $Y_{i1,}$ y trata a este dato como un valor real; a continuación imputa un valor para cada valor perdido Y_{i2}, con estos se crea el primer conjunto de imputaciones. La segunda imputación, repite todos los pasos, para encontrar los nuevos valores imputados.

3.7.3.4 MODELO DE REGRESIÓN LINEAL NORMAL EXPLÍCITO CON DOS VARIABLES DE SALIDA Y_i

Con la variable de salida $Y_i = (Y_{i1}, Y_{i2})$ se realizará una regresión lineal con dos variables de salida sobre la variable X_i con Y_{i1} observados sobre las n_1 unidades y Y_{i2} observadas sobre las n_2 unidades, tal que $n_2 \leq n_1$, en un patrón monótono. Correspondiendo a patrones monótonos, se especifica la distribución conjunta de (Y_{i1}, Y_{i2}) dado X_i por especificación de la distribución condicional de Y_{i1} sobre (X_i, θ) como:

$$N(X_i \beta_1, \sigma_1^2) \qquad \textbf{(3.74)}$$

Y entonces la distribución condicional de Y_{i2} sobre (Y_{i1}, X_i, θ) como:

$$N(\gamma Y_{i1} + X_i \beta_2, \sigma_2^2) \quad \textbf{(3.75)}$$

Donde β_1 y β_2 son vectores columna de los q componentes, γ, σ_1^2 y σ_2^2 son escalares, $\theta = (\beta_1, \beta_2, \gamma, \log\sigma_1, \log\sigma_2)$ y Pr(θ) es constante.

Primero se ignora Yi2 y se imputa los valores perdidos de Y_{i1} usando todas las observaciones de X y aplicando los resultados para el modelo lineal dado en el punto 3.7.1, donde por notación Y_i es reemplazado por Y_{i1}, β es reemplazado por β_1, σ^2 por σ_1^2, y obs por obs[1] que es el conjunto de unidades de $Y_{[1]}$ observado. Así los parámetros de la regresión de $Y_{[1]}$ sobre (X, $\theta_1 = (\beta_1, \log\sigma_1)$), son estimados desde las unidades $Y_{[1]}$ observadas, y valores perdidos $Y_{[1]}$ son imputados por la primera estimación desde la distribución posterior de estos parámetros, y entonces se usa X_i para predecir los valores perdidos Y_{i1}. Se supone dos conjuntos de valores perdidos Y_{i1} han sido imputados.

Ahora se trata el primer conjunto de datos imputados Y_{i1} como reales y se imputa un conjunto de valores perdidos Y_{i2} usando los resultados del punto 3.7.1 donde la regresión es de Y_{i2} sobre (Y_{i1}, X_i), y estos parámetros son estimados usando las unidades con Y_{i2} observados. Además, la notación del punto 3.7.1 es cambiada de modo que Y_i es reemplazada por Y_{i2}, β es reemplazada por (γ, β_2), σ^2 por σ_2^2 y obs por obs[2] que es el conjunto de unidades con Y_{i2} observado. A continuación se trata el segundo conjunto de imputaciones Y_{i1} como reales y se imputa el segundo conjunto de valores perdidos Y_{i2} usando el mismo procedimiento pero con los nuevas variables aleatorias.

La distribución posterior de $\theta_2 = (\gamma, \beta_2, \log\sigma_2)$ no involucra cualquier imputación de Y_{i1} puesto que los patrones monótonos implican que todas las unidades con Y_{i2} observado tienen Y_{i1} valores observados. Además, la distribución posterior de θ_2 trata el primer conjunto de imputaciones Y_{i1} como reales, los cuales son idénticas a la distribución posterior de θ_2, tratando así, el segundo conjunto de

imputaciones Y_{i1} como reales. Valores imputados de Y_{i2} pueden, sin embargo, depender de los valores imputados en Y_{i1} desde $Y_{i2,}$ y los Y_{i1} perdidos usará las imputaciones de los valores Y_{i1} en el cálculo de la media condicional de los valores Y_{i2} a ser imputados.

3.7.4 METODO DE IMPUTACION MULTIPLE MARKOV CHAIN MONTE CARLO (MCMC)

Los métodos MCMC, se construye por una cadena de Markov[7] bastante grande para la distribución de los elementos con el fin de estabilizar una distribución común. Esta distribución estacionaria es la distribución de interés.

En la inferencia bayesiana, información acerca del parámetro desconocido es expresada en la forma de una distribución de probabilidad posterior. El método MCMC se aplica para explorar la distribución posterior en inferencia bayesiana. Es decir, a través del método MCMC, se puede simular toda la distribución conjunta de las cantidades desconocidas y obtener simulaciones basadas en la estimación de los parámetros posteriores que son de interés.

Asumiendo que los datos provienen desde una distribución normal multivariable, la agregación de los datos es aplicada desde la inferencia bayesiana a datos perdidos, a través de la repetición de los siguientes pasos:

Primer Paso: Imputación

Con la estimación del vector de la media y matriz de covarianzas, el primer paso consiste en simular los valores perdidos para cada una de las observaciones independientemente. Esto es, si se denotó la variable con valores perdidos para la observación i por $Y_{i(per)}$ y la variables con valores observados como $Y_{i(pbs)}$,

[7]Cadena de Markov: Es una secuencia de variables aleatorias en la cual la distribución de cada uno de los elementos depende de los valores previos.

entonces los valores estimados en el primer paso para $Y_{i(per)}$ desde la distribución condicional de $Y_{i(per)}$ esta dada por los $Y_{i(obs)}$.

P - Pasos: Distribución Posterior

Concluida la simulación de los P-pasos se obtiene, a partir de la estimación de la muestra completa, el vector de la media de la población y de la matriz de covarianza posterior. Entonces estas nuevas estimaciones son usadas en el primer paso. Si no existe información previa acerca de los parámetros, información previa puede ser usada, por ejemplo, información referente a la matriz de la covarianza puede ser útil para estabilizar la inferencia debido al vector de las medias; esta puede ser la matriz singular de la covarianza cercana.

El siguiente paso es realizar varias iteraciones para que los resultados sean confiables, pues se tiene un conjunto de datos imputados.

Por tanto, el objetivo es que estas iteraciones converjan a la distribución estacionaria y entonces se obtiene una estimación aproximada de los valores perdidos.

Estos es, con el estimador de los parámetros $\theta^{(t)}$ en la t$^{\text{ésima}}$ iteración, el primer paso consiste en estimar Y_{per}^{t+1} desde $p(Y_{per} \mid Y_{obs}, \theta^{(t)})$ y en los P-Pasos estimar $\theta^{(t+1)}$ desde $p(\theta \mid Y_{obs}, Y_{per}^{(t+1)})$.

Esto crea una cadena de Markov: $\left(Y_{per}^{(1)}, \theta^{(1)}\right), (Y_{per}^{(2)}, \theta^{(2)}), \ldots\ldots\ldots\right)$

La que converge a la distribución $p\left(Y_{per}, \theta \mid Y_{obs}\right)$

A través del algoritmo EM es factible encontrar el estimador de máxima verosimilitud para modelos paramétricos de datos incompletos. Este algoritmo puede ser empleado para calcular la estimación de los parámetros de la densidad

posterior cuando los datos observados sean grandes. El resultado de la estimación EM provee un buen valor inicial para comenzar el proceso MCMC.

En conclusión, en imputación múltiple el proceso MCMC es usado para crear un número pequeño m de estimaciones independientes o “imputaciones” de datos perdidos Y_{per}

3.7.5 MODELOS DE SERIES DE TIEMPO

Los modelos autoregresivos son relativamente fáciles para ajustar datos de series de tiempo incompletos, con el ajuste del algoritmo EM. Modelos de Box-Jenkins con componentes media móviles son más difíciles de manejarlos, pero la estimación de máxima verosimilitud puede llevarse a cabo una reestructuración de los modelos a través de espacio de estado generales. A continuación se revisarán:

3.7.5.1 MODELOS AUTOREGRESIVOS PARA SERIES DE TIEMPO DE UNA VARIABLE CON VALORES PERDIDOS

Sea $Y = (y_0, y_1, ..., y_T)$ la serie de tiempo de una variable observada completamente con T+1 observaciones. El modelo autoregresivo de orden p AR(p) asume que y_i, en el tiempo *i*, está relacionado a valores en *p* períodos de tiempo previos dado por el modelo[3]:

$$(y_i \mid y_1, y_2, ..., y_{i-1}) \approx N(\alpha + \beta_1 y_{i-1} + ... + \beta_p y_{i-p}, \sigma^2) \qquad (3.100)$$

donde $\theta = (\alpha, \beta_1, \beta_2, ..., \beta_p, \sigma^2)$, α es un término constante, $\beta_1, \beta_2, ..., \beta_p$ son los coeficientes desconocidos de la regresión, y σ^2 corresponde a la varianza del error que también es desconocida. La estimación de mínimos cuadrados de

[3] LITTLE/RUBIN(2002)

$\alpha, \beta_1, \beta_2, ...\beta_p$ y σ^2 puede ser encontrada por la regresión de y_i sobre $x_i = (y_{i-1}, y_{i-2}, ..., y_{i-p})$, empleando las observaciones *i =p, p+1,……,T.*

Si algunas observaciones en la serie están perdidas, se pueden considerar las aplicaciones de los métodos 3.7.1 y del capítulo 2, para regresión con valores perdidos. Esta técnica puede producir aproximaciones toscas, debido a que los procedimientos no son máximo verosímiles, aun si se ignora la distribución marginal de $y_0, y_1, ..., y_{p-1}$, puesto que: 1) los valores perdidos y_i ($i \geq p$) aparecen como variables dependientes e independientes en la regresión y 2) el modelo (3.100) induce una estructura espacial sobre el vector de la media y la matriz de covarianza de Y, la que no es usada en los análisis. Aun así, algoritmos EM especiales se requieren para estimar el modelo AR(p) desde series de tiempo incompletas.

Como ejemplo se analizará el caso en que p=1, es decir, un modelo AR(1) para la serie de tiempo con valores perdidos.

Con p=1 en la ecuación (3.100) se obtiene el siguiente modelo:

$$(y_i \mid y_1, y_2, ..., y_{i-1}, \theta) \approx N(\alpha + \beta\, y_{i-1}, \sigma^2) \qquad \textbf{(3.101)}$$

La serie AR(1) es estacionaria, produce una distribución marginal constante de y_i sobre el tiempo, solo si $|\beta| < 1$. La distribución conjunta de los y_i tiene:

Media marginal constante $\mu = \alpha(1-\beta)^{-1}$

Varianza $\text{var}(y_i) = \sigma^2 (1-\beta^2)^{-1}$

Covarianza $Cov(y_i, y_{i+k}) = \beta^k \sigma^2 (1-\beta^2)^{-1}$ para $k \geq 1$

Ignorando la contribución de la distribución marginal de y_0, el logaritmo de máxima verosimilitud de los datos completos para Y es,

$l(\alpha,\beta,\sigma^2 \mid y) = -\sum_{i=1}^{T} \frac{(y_i - \alpha - \beta\, y_{i-1})^2}{2\theta^2} - \frac{T \log \sigma^2}{2}$, lo cual es equivalente al logaritmo de máxima verosimilitud de la regresión lineal normal de y_i sobre $x_i = y_{i-1}$ con los datos {(x_i , y_i), i=1,……..T}. Los estadísticos suficientes de los datos completos son: S=(s_1, s_2, s_3, s_4, s_5), donde:

$$s_1 = \sum_{i=1}^{T} y_i \qquad s_2 = \sum_{i=1}^{T} y_{i-1} \qquad s_3 = \sum_{i=1}^{T} y_i^2 \qquad s_4 = \sum_{i=1}^{T} y_{i-1}^2 \qquad s_5 = \sum_{i=1}^{T} y_i y_{i-1}$$

Los estimadores de máxima verosimilitud de $\theta = (\alpha, \beta, \sigma)$ son $\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{\sigma})$, donde:

$$\begin{aligned} \hat{\alpha} &= \left(s_1 - \hat{\beta}\, s_2\right) T^{-1} \\ \hat{\beta} &= \left(s_5 - T^{-1} s_1 s_2\right)\left(s_4 - T^{-1} s_2^2\right)^{-1} \\ \hat{\sigma}^2 &= \frac{s_3 - s_1^2 T^{-1} - \hat{\beta}^2 \left(s_4 - s_2^2 T^{-1}\right)}{T} \end{aligned} \qquad \textbf{(3.102)}$$

Ahora se suponen algunas observaciones pérdidas, y que los datos tienen un mecanismo MAR. El estimador de máxima verosimilitud de θ, aun ignora la contribución de la distribución marginal de y_o para máxima verosimilitud, lo que puede ser obtenido por el algoritmo EM.

Sea $\theta^{(t)} = \left(\alpha^{(t)}, \beta^{(t)}, \sigma^{(t)}\right)$ el estimador de θ en la iteración t.

El paso **M** del algoritmo calcula $\theta^{(t+1)}$ desde la ecuación (3.102) con los estadísticos suficientes para datos completos reemplazando S por el estimado $S^{(t)}$ desde el paso **E**.

El paso **E** calcula $S^{(t)} = \left(s_1^{(t)}, s_2^{(t)}, s_3^{(t)}, s_4^{(t)}, s_5^{(t)}\right)$, donde:

$$s_1^t = \sum_{i=1}^{T} \hat{y}_i^{(t)} \qquad s_2^t = \sum_{i=1}^{T} \hat{y}_{i-1}^{(t)} \qquad s_3^t = \sum_{i=1}^{T} \left[\left(\hat{y}_i^{(t)}\right)^2 + c_{ii}^t\right] \qquad s_4^t = \sum_{i=1}^{T} \left[\left(\hat{y}_{i-1}^{(t)}\right)^2 + c_{i-1,i-1}^t\right]$$

$$s_5^t = \sum_{i=1}^{T} \left[\hat{y}_{i-1}^{(t)} \hat{y}_i^{(t)} + c_{i-1,i}^t \right]$$

y

$$\hat{y}_i^{(t)} = \begin{cases} y_i, & \text{si } y_i \text{ es observado} \\ \mathrm{E}\{y_i \mid Y_{obs}, \theta^{(t)}\}, & \text{si } y_i \text{ es perdido} \end{cases}$$

$$c_{ij}^{(t)} = \begin{cases} 0, & \text{si } y_i \text{ o } y_j \text{ son observados} \\ \mathrm{Cov}\{y_i, y_j \mid Y_{obs}, \theta^{(t)}\} & \text{si } y_i \text{ y } y_j \text{ son perdidos} \end{cases}$$

El paso E involucra una operación de extensión estándar sobre la matriz de covarianzas de las observaciones. Sin embargo, esta matriz (TxT) es usualmente grande, de manera que es deseable aplicar propiedades exploratorias del modelo AR(1) para simplificar el cálculo del paso E.

Se supone que $Y_{per}^{*} = (y_{j+1}, y_{j+2}, ..., y_{k-1})$ es una secuencia de valores perdidos, limitada por las observaciones presentes, y_j y y_k. Entonces, 1) los y^*_{per} son independientes de los otros valores perdidos dados por Y_{obs} y θ 2) la distribución de y^*_{per} dados Y_{obs} y θ depende de Y_{obs} solamente por las limitaciones de las observaciones de y_j y y_k. La última distribución es normal multivariada, con matriz de covarianza constante y matriz de media que son los promedios de los pesos de $\mu = \alpha(1-\beta)^{-1}$, y_j y y_k. La matriz de las covarianzas y los pesos dependen solamente del número de valores perdidos en la secuencia y pueden ser encontrados desde la estimación de la matriz de covarianza de $(y_j, y_{j+1}, ..., y_k)$ por extensión de los elementos correspondientes a las variables observadas y_j y y_k.

En general, se supone que y_j y y_{j+2} son observados y y_{j+1} es perdido. La matriz de covarianza de y_j , y_{j+1} y y_{j+2} es:

$$A = \frac{\sigma^2}{1-\beta^2}\begin{bmatrix} 1 & \beta & \beta^2 \\ \beta & 1 & \beta \\ \beta^2 & \beta & 1 \end{bmatrix}$$

Extendiendo sobre y_j y y_{j+2} se obtiene:

$$SWP[j, j+2]A = \frac{1}{1+\beta^2}\begin{bmatrix} -\sigma^{-2} & \beta & -\beta^2\sigma^2 \\ \beta & \sigma^{-2} & \beta \\ -\beta^2\sigma^2 & \beta & -\sigma^{-2} \end{bmatrix} \text{[7]} \qquad \textbf{(3.103)}$$

Por tanto, se puede obtener la esperanza y varianza de y_j empleando la expresión de la ecuación (3.103):

$$E\{y_{j+1} \mid y_j, y_{j+2}, \theta\} = \mu + \beta(1+\beta^2)^{-1}(y_{j+2} - u) + \beta(1+\beta^2)^{-1}(y_1 - \mu)$$

$$E\{y_{j+1} \mid y_j, y_{j+2}, \theta\} = \mu\left\{1 - \frac{2\beta}{1+\beta^2}\right\} + \frac{\beta}{1+\beta^2}(y_j + y_{j+2})$$

$$Var(y_{j+a}, y_{j+2}, \theta) = \sigma^2(1+\beta^2)^{-1}$$

Sustituyendo $\theta = \theta^{(t)}$ en estas expresiones se encuentra $\hat{y}_{j+1}^{(t)}$ y $\hat{c}_{j+1,\, j+1}^{(t)}$ para el paso E. Se observa que $\hat{y}_{j+1}^{(t)}$ es la predicción para los valores perdidos en el final de la iteración del algoritmo.

3.7.5.2 INTERPOLACIÓN ÓPTIMA Y FUNCIÓN DE AUTOCORRELACIÓN INVERSA

Se supone que se tiene una serie estacionaria con observaciones perdidas en el tiempo T. La estimación de los valores perdidos es un problema de interpolación que puede ser resuelto por el cálculo de la esperanza de la variable aleatoria no observada conocidos los demás datos de la misma variable. Grenander y Rosenblatt (1957) encontraron que esta esperanza es:

[7] Revisar Anexo 3

$$E(z_t / Z_{(T)}) = -\sum_{i=1}^{\infty} \delta_i (z_{T+i} + Z_{T-i}) \tag{3.104}$$

donde δ_i corresponde a los coeficientes de autocorrelación inversa y $Z_{(T)}$ incluye a todos los datos excepto los valores perdidos.

Adicionalmente se define el proceso dual de un modelo ARIMA inversible [3] como un proceso ARMA:

$$\theta(B) z_t = \phi(B) \nabla^d a_t \tag{3.105}$$

el proceso dual es construido por intercambio de roles de los operadores AR y MA. Entonces la función de auto correlación del proceso dual (3.105) es la inversa de la función de auto correlación (IAF) del proceso original[3]. Por ejemplo, a un proceso AR(1) con $(1-\phi B) z_t = a_t$, le correspondería una función de auto correlación (IAF), que en este caso seria la función de auto correlación del proceso MA(1) con $z_t = (1-\phi B) a_t$. Por tanto, la IAF tiene el primer coeficiente de auto correlación igual a $-\phi / (1-\phi^2)$ y todos los otros valores son iguales a cero. Empleando la ecuación (3.104), el interpolador óptimo para una media cero de un proceso AR(1) es:

$$E(z_T / Z_{(T)}) = \frac{\phi}{1+\phi^2} (z_{T+1} + Z_{T-1}) \tag{3.106}$$

Se observa que el proceso dual de un proceso invertible es estacionario y por tanto, la función de auto correlación inversa siempre existe.

Peña y Marawall (1991) demostraron que los resultados de la ecuación (3.104) pueden ser usados tanto por procesos estacionarios como por procesos no estacionarios.

Escribiendo la serie de tiempo en la representación general AR(∞):

$$z_t = \sum_{i=1}^{\infty} \pi_i z_{t-i} + a_t \tag{3.107}$$

---

[3] MODELO ARIMA de la serie y_t , $\phi(B) \nabla^d y_t = \theta(B) a_t$

entonces, si el valor z_T es perdido, se obtiene un estimador insesgado a través de:

$$\hat{z}_T^{(0)} = \sum_{i=1}^{\infty} \pi_i z_{t-i} \tag{3.108}$$

y su estimado, el que es construido desde las observaciones previas de los valores perdidos tendrá una varianza σ_a^2. Sin embargo se debe recordar que se tiene más información en z_T. Esta información está contenida en todas las observaciones posteriores a los valores perdidos. Se puede obtener la siguiente expresión para todos los j, tal que $\pi_j \neq 0$:

$$z_T = \pi_j^{-1}(z_{T+j} - \sum_{\substack{i=j \\ i\neq j}}^{\infty} \pi_i z_{T+t-i}) - \frac{a_{T+j}}{\pi_j} \tag{3.109}$$

y, por tanto, se obtiene un estimador adicional insesgado con retardos de z_T a través de la ecuación:

$$\hat{z}_T^{(j)} = \pi_j^{-1}(z_{T+j} - \sum_{i\neq j} \pi_i z_{T+t-i}) \tag{3.110}$$

con varianza $\sigma_a^2 \big/ \pi_j^{-1}$. Como todas estas estimaciones son condicionalmente insesgadas e independientes dados los valores observados, la mejor estimación lineal insesgada de los valores perdidos $z_{T,}$ será:

$$\hat{z}_T = \sum_{j=0}^{\infty} \left(\pi_j^2 \Big/ \sum \pi_j^2 \right) \hat{z}_T^{(j)} \tag{3.111}$$

donde π_0 =-1.

Se puede combinar los estimadores de adelanto con los n-T estimadores de retraso, a través de la siguiente ecuación para el interpolador simple finito:

$$\hat{z}_{T,F} = \sum_{j=0}^{n-T} \frac{\pi_j^2}{\sum_0^n \pi_i^2} \tag{3.112}$$

Ejemplo:

Se calcula el interpolador óptimo para más de un valor perdido.
Se supone que se tiene un proceso AR(1) en el cual los valores z_T y z_{T+1} son perdidos. Entonces, se calcula el interpolador óptimo de la siguiente manera:

Para z_T se tienen los estimadores en adelanto:

$$\hat{z}_T^{(0)} = \phi z_{T-1}$$

Con error de varianza igual a σ_a^2 y como $z_{T+2} = \phi^2 z_T + \phi a_{T+1} + a_{T+2}$, pueden ser calculados los estimadores con retardo como:

$$\hat{z}_T^{(2)} = \phi^{-2} Z_{T+2}$$

Con varianza $\sigma_a^2(1+\phi^2)/\phi^4$. Por tanto, el mejor estimador linear insesgado será:

$$\hat{z}_T = \frac{\phi(1+\phi^2)}{1+\phi^2+\phi^4} z_{T-1} + \frac{\phi^2}{1+\phi^2+\phi^4} z_{T+2}$$

3.7.5.3 ESTIMACIÓN DE VALORES PERDIDOS: CASO GENERAL

Los análisis previos sugieren los siguientes procedimientos para cálculo de los valores perdidos en series de tiempo:

1. Ejecutar una primera interpolación de los valores perdidos, identificando los modelos ARIMA y estimando sus parámetros por máxima verosimilitud en la serie completa.
2. Obtener los coeficientes de auto correlación inversa, que están directamente dados en el modelo, y calcular el interpolador óptimo de los valores perdidos por la ecuación (3.104).

Estos procedimientos pueden ser iterativos, hasta que las series hayan sido completadas por el interpolador óptimo, es así que se puede calcular otro conjunto de parámetros estimados, los cuales dirigirán la nueva interpolación del valor perdido y así sucesivamente. Las interacciones son importantes cuando el número de valores perdidos son grandes, debido a que el primer parámetro estimado se basa en algunas interpolaciones toscas que pueden dirigir a parámetros estimados sesgados.

Para entender el procedimiento, se ilustrará en un caso simple de una AR(1). La función de máxima verosimilitud calculada por la descomposición del error de predicción asume que z_T son perdidos y se los obtiene por:

$$f(z_1,.....z_{T-1},z_{T+1},.....,z_n)=f(z_1)f(z_2/z_1).....f(z_{T+1}/z_{T-1})f(z_{T+2}/z_{T+1}).....f(z_n/Z_{n-1})$$

Asumiendo normalidad y que los procesos tiene media cero y están condicionados en la primera observación, la función de máxima verosimilitud condicional puede ser fácilmente obtenida para todos los t≥2 pero t≠T. Se tiene que $f(z_t/z_{t-1})$ es $N(\phi_{Z_{T-1}},\sigma^2(1+\phi^2)$.

Entonces la función de máxima verosimilitud condicional a ser maximizada corresponde a:

$$l(\phi,\sigma^2/z_1)=-\frac{(n-2)}{2}\ln\sigma^2-\frac{1}{2}\ln(1+\phi^2)-\sum_{t\in A}\frac{(z^t-\phi z_{t-1})^2}{2\sigma^2}-\frac{(z_{T+1}-\phi^2 z_{T-1})^2}{2\phi^2(1+\phi^2)} \qquad (3.113)$$

donde el conjunto A es {1,2,…….,T-1,T+2,…….,n}. Comparando esta función con la obtenida para series sin valores perdidos pero que posea observaciones discordantes o extremas en el tiempo T, el modelo podría ser:

$$z_t=\phi z_{t-1}+\omega \mathrm{I}_t^{(T)}-\phi\omega \mathrm{I}_{t-1}^{(T)}+a_t \qquad (3.114)$$

y la función de máxima verosimilitud sería:

$$l(\omega,\phi,\sigma^2 / z_1) = -\frac{(n-1)}{2}\ln\sigma^2 - \sum_{t\in A}\frac{(z_t - \phi z_{t-1})^2}{2\sigma^2} - \frac{(z_t - \phi z_{T-1} - \omega)^2}{2\sigma^2} - \frac{(z_{T+1} - \phi(z_T - \omega))^2}{2\sigma^2} \tag{3.115}$$

La estimación de la ecuación (3.115) puede ser realizada en dos pasos: En el primer paso, condicionando ϕ y obteniendo la estimación para ϖ dado ϕ. Obteniendo la derivada de la ecuación (3.115) con respecto a ϖ e inicializando con valores iguales a cero, se obtiene:

$$(z_T - \phi z_{T-1} - \omega) = (z_{T+1} - \phi(z_T - \omega))\phi$$

por lo tanto, se encuentra:

$$\hat{\omega} = z_T - \frac{\phi}{(1+\phi^2)}(z_{T+1} + z_{T-1}) \tag{3.116}$$

Esta estimación se puede interpretar como que la diferencia entre los valores observados y su interpolador óptimo para el AR(1) esta dada por (3.106). Incluyendo este dato en la ecuación (3.115), se obtiene:

$$l(\phi,\sigma^2 / z_1) = -\frac{(n-1)}{2}\ln\sigma^2 - \sum_{t\in A}\frac{(z_t - \phi z_{t-1})^2}{2\sigma^2} - \frac{(z_{T+1} - \phi^2 z_{T-1})^2}{2\sigma^2(1+\phi^2)} \tag{3.117}$$

3.7.5.4 MODELACIÓN POR FILTROS DE KALMAN

Considerando la modelación por los filtros de KALMAN:

$$(y_i \mid A_i, z_i, \theta) \approx N(z_i A_i, B) \tag{3.118}$$

$$(z_0 \mid \theta) \approx N(\mu, \Sigma) \tag{3.119}$$

$$(z_i \mid z_1, \ldots\ldots\ldots\ldots, z_{i-1}, \theta) \approx N(z_{i-1}\phi, Q), \qquad i \geq 1 \tag{3.120}$$

Donde y_i es un vector (1 x q) de variables observadas en el tiempo i, A_i es conocida como la matriz de diseño (p x q) que relaciona la media de y^j con un vector estocástico (1 x p) no observado z_i y $\theta = (B, \mu, \phi, Q)$ representa los parámetros desconocidos donde B, Σ y Q son las matrices de covarianza, μ es la media de z_o, y ϕ es una matriz (p x p) de los coeficientes de autoregresión de z_i en z_{i-1}. La serie aleatoria no observada z_i, la que está modelada como un proceso autoregresivo multivariante de primer orden, es de principal importancia.

Este modelo puede ser empleado cuando una clase de efectos aleatorios se introducen al modelo de series de tiempo, donde z_i tiene una estructura de correlación con el tiempo. El principal objetivo es pronosticar la serie no observada { z_i } para i=1,2,…..,n (suavización) y para I= n+a, n+2,…. (predicción) usando la serie observada y_1, y_2,……,y_n. Si el parámetro θ fuese conocido, el estimador óptimo de z_i debería ser su media condicional, con θ y los datos Y.
Estas cantidades son los llamados estimadores de suavización de Kalman y el conjunto de fórmulas recursivas para su determinación son los llamados filtros de Kalman.

En la práctica el parámetros θ es desconocido, y los procedimientos de predicción y suavización involucran la estimación de máxima verosimilitud de θ , y entonces la aplicación de los filtros de Kalman se realiza a través del reemplazo de θ por su estimación $\hat{\theta}$.

El mismo proceso se aplica cuando los datos Y son incompletos, esto es, se reemplaza Y con su componentes observados Y_{obs}. El estimador de máxima verosimilitud de Q puede ser ejecutado a través de la técnica de Newton – Rapson. Sin embargo, el algoritmo EM provee un método alternativo conveniente con los componentes perdidos Y_{per} de Y y z_1, z_2,……,z_n tratados como datos perdidos. El principal atractivo de esta aproximación es que el paso E del algoritmo EM incluye el calculo del valor esperado de z_i dado Y_{obs} y los estimadores comunes de θ, los cuales son los mismos del proceso de suavización

de Kalman. Los estimadores de ϕ y Q son obtenidos por la aplicación de la autoregresión a los valores esperados de los estadísticos suficientes de los datos completos.

$$\sum_{i=1}^{n} z_i \qquad \sum_{i=1}^{n} z_i^T z_i \qquad \sum_{i=1}^{n} z_{i-1} \qquad \sum_{i=1}^{n} z_{i-1}^T z_{i-1} \qquad \sum_{i=1}^{n} z_{i-1}^T z_i$$

A través del paso E, B es estimado por el valor esperado de la matriz de covarianza residual $\frac{1}{n}\sum (y_i - z_i A)^T (y_i - z_i A)$. Finalmente μ estimada como el valor esperado de z_o, y Σ es el conjunto de consideraciones externas.

CAPÍTULO 4

APLICACIÓN DE LA IMPUTACIÓN ESTADÍSTICA DE DATOS AL SISTEMA NACIONAL INTERCONECTADO DEL ECUADOR

4.1 OBJETIVOS DEL ESTUDIO

El CENACE a través de su Centro de Control y con la ayuda las unidades terminales remotas (UTR) captura datos cada cuatro segundos, de varias señales eléctricas (potencia, voltaje, frecuencia, etc.) de las principales subestaciones eléctricas del País y las envía, a través del sistema de comunicaciones, al Centro de Control ubicado en CENACE. Los datos capturados por UTR son almacenados en una base de datos del EMS llamada HIS.

Los datos que son transmitidos al CENACE, permiten al personal de Sala de Control realizar la supervisión en tiempo real del Sistema Nacional Interconectado, y en muchos casos ejecutan acciones para levantar cualquier restricción del sistema de potencia. Adicionalmente, la información almacenada en la base de datos HIS se utiliza posteriormente en los procesos técnicos y comerciales que ejecuta el CENACE.

Cabe mencionar que la confiabilidad de la supervisión del sistema eléctrico ecuatoriano se fundamenta en la redundancia en las mediciones y a pesar de ello

muchas veces las unidades terminales remotas sufren daños aleatorios que no pueden ser superados inmediatamente, dejando de transmitir datos al CENACE. Ante esta situación, el EMS cuenta con una aplicación adicional, el Estimador de Estado, que estima el dato que no pudo ser medido en campo y cuyo dato en la mayoría de los casos no es almacenado en la base de datos del HIS.

Por esta razón, cuando se validan los datos de potencia activa instantánea se observan inconsistencias (datos incoherentes), que obliga a sustituir un dato por otro de mejores características, que permite al usuario contar con información confiable y segura para ejecutar los procesos de Planificación, Liquidaciones Comerciales y Estadística Operativa del Sistema Eléctrico Ecuatoriano.

En este contexto, la disponibilidad de datos con buenas características, permite a la Dirección de Operaciones del CENACE brindar a sus clientes tanto internos como externos información confiable de calidad y en el menor tiempo posible, cumpliendo así uno de los principales objetivos del Área de Análisis de la Operación.

El término pérdida de datos (data missing), se refiere a la forma global de falta de datos debido a la indisponibilidad de las unidades terminales remotas que traen los datos de potencia activa instantánea desde el campo hacia el Centro de Control, y que por problemas de la misma unidad terminal remota o desde su sistema de comunicaciones o elementos afines a la transmisión y recepción de la información, el dato no cumple con los requisitos básicos de consistencia; por ejemplo, que el dato no contenga la etiqueta de calidad de dato “normal”.

4.2 CONCEPTO DE PERDIDA O AUSENCIA DE DATOS

Hay diversas razones por las que se producen pérdidas de datos relativas a extracción de la medición desde campo a través del sistema de UTR´s. Según donde se encuentre el origen de la pérdida de datos, se identifican dos tipos de

fuentes: el daño de la UTR y todo el sistema asociado a la transmisión y recepción del dato y el congelamiento de la medición.

Las situaciones anteriores son consecuencia de la pérdida de datos, pues las mediciones ya no son recogidas desde el sitio donde se encuentra el dato.

4.3 METODOLOGÍA

Una vez que los datos fueron recopilados desde campo (generadores y subestaciones de transmisión eléctrica) y guardados en un serie temporal en la base de datos del sistema histórico con valores en unidades de ingeniería y con las correspondientes banderas de calidad de la monitorización, se procede a la identificación de la falta de respuesta. La identificación consistió en verificar que los datos de potencia activa y reactiva instantánea provenientes del sistema de manejo de energía EMS de las barras de carga del Sistema Nacional Interconectado del Ecuador contenga la etiqueta de calidad de dato diferente a la etiqueta normal, pues esta caracterización permite identificar al dato erróneo en campo y que fue almacenado bajo esta condición en la base de datos histórica.

Una vez que se determinó la razón para la no presencia de datos, estos fueron reemplazos por el espacio en blanco en la base de datos ya que los programas estadísticos interpretan a este espacio como un carácter ".", que le asignan a los datos omitidos.

El análisis de la falta de respuesta se realiza para el mes de Septiembre del 2007 con una base de datos horaria correspondiente a tres mes anteriores; es decir, la base de datos completa corresponde a los meses de Julio hasta Agosto del 2007.

Como el objetivo de este trabajo es realizar el análisis de la potencia activa instantánea de las barras de carga del Sistema Nacional Interconectado, se ha realizado el análisis de las barra que tienen esta característica, barras que no tienen inmersa generación ni térmica ni hidráulica, constituyéndose en 7

empresas eléctricas con 10 barras, y que serán sujetas al análisis y obtención del dato imputado, a continuación se presenta el cuadro resumen:

CUADRO 4.1
RESUMEN DE LAS EMPRESAS ELÉCTRICAS Y SUS RESPECTIVAS BARRAS, SUJETAS A IMPUTACIÓN

EMPRESA ELÉCTRICA	*BARRA*	*POSICIÓN*
AMBATO	TOTORAS	Ambato
		Montalvo
		Baños
CATEG-SD	PASCUALES	Cervecería
		Vergeles
	TRINITARIA	Guasmo
		Pradera
		Padre Canales
	POLICENTRO	Policentro
QUITO	POMASQUI	Quito 1
		Quito 2
COTOPAXI	AMBATO	Ambato
EMELGUR	DOS CERRITOS	Dos Cerritos
	PASCUALES	Pascuales
EMELNORTE	IBARRA 69	Otavalo
		Cotacachi
		Retorno
EMELSAD	SANTO DOMINGO	S. Domingo 1
		S. Domingo 2

Por lo tanto, como primera fase de este trabajo se cuantifica la falta de respuesta (descripción en punto 4.5.1) y se observa la distribución en la muestra.

En la siguiente fase, para el reemplazo del dato faltante se aplican los siguientes procedimientos: hot – deck, hot – deck con regresión, regresión condicionada, imputación simple y un algoritmo de imputación múltiple.

Debido a que la mayoría de los métodos aplicados se sustentan en técnicas de regresión, se ejecuta como fase previa a la imputación de datos, la especificación de los modelos que relacionan las variables de (potencia activa instantánea) con las variables potencia activa y potencia reactiva instantánea de la misma barra en

que una Empresa Distribuidora retire energía del Sistema Nacional de Transmisión. A través de esta fase se verifica la significancia estadística de los parámetros generados por las especificaciones ajustadas y se procede a imputar los datos faltantes. Posteriormente, con los datos generados para potencia activa instantánea por los procedimientos de imputación se comprobarán cuál de los métodos brinda la mejor estimación.

Cabe señalar que la aplicación de técnicas que utilizan métodos de regresión requiere la especificación funcional que relacione la variable que se desea imputar, con un vector de covariables que expliquen su trayectoria.

4.4 SOFTWARE PARA IMPUTACION DE DATOS

En la actualidad existen varios programas computacionales que permite realizar imputación de datos con distintos tipos de matrices de datos. Entre los principales programas computacionales dedicados a la imputación están los programas SPSS, SAS y STATA. En el cuadro 4.2 se resumen las principales características de los métodos de imputación que utilizan estos programas, adicionalmente se señalan los supuestos en que se sustenta cada método con relación al patrón observado de datos faltantes, la forma en la que se aplica el procedimiento y algunas de sus principales ventajas y desventajas.

En el presente trabajo se emplea el programa STATA 9.0 ya que incorpora gran variedad de algoritmos que permiten efectuar sustituciones de datos mediante distintas opciones metodológicas, además la Corporación CENACE dispone de una licencia de la versión 9.0.

STATA incorpora comandos que permite conocer la manera en que se distribuyen los datos faltantes (*mvpatterns*), opciones de imputación para los procedimientos *hot –deck, hot-deck* con regresión, imputación por regresión (*impute*), imputación simple (*ivis*) e imputaciòn múltiple (*ice*).

CUADRO 4.2

MÉTODOS DE IMPUTACIÓN DISPONIBLES EN EL SOFTWARE ESTADÍSTICO

METODOS DE IMPUTACIÓN					PAQUETES ESTADÍSTICOS		
Método	**Patrón de datos faltantes**	**Aplicación**	**Ventajas**	**Desventajas**	**SPSS**	**SAS**	**STATA**
Medias no condicionadas	Patrón de datos faltantes (MCAR)	Las observaciones faltantes se reemplazan por el valor medio de la variable de análisis.	Fácil de entender y aplicar	Genera estimadores sesgados. Subestima la varianza de los estimadores	x	x	
Medias por subgrupos	Patrón de datos faltantes (MCAR)	Se divide la base de datos en subgrupos utilizando variables correlacionadas. Las observaciones faltantes en el subgrupo de interés se reemplazan por el valor medio de la variable.		Genera estimadores sesgados. Subestima la varianza de los estimadores	x	x	
Hot-Deck (condicionado a covariables)	Patrón de datos faltantes (MCAR)	Se divide la base de datos en subgrupos utilizando variables correlacionadas. Las observaciones faltantes en el subgrupo de interés se sustituyen con los datos de un registro con información similar en las covariables	Los donantes y receptores pertenecen a un mismo subgrupo. El número de donantes se condiciona con el uso de covariables	No siempre es facil definir un criterio de distancia o similitud entre los posibles donantes y receptores.			x
Hot-Deck (con regresión condicionada a covariables)	Patrón de datos faltantes (MCAR). Se requiere especificar un modelo, en donde las covariables estén altamente correlacionadas con la variable a imputar	Se divide la base de datos en subgrupos utilizando variables correlacionadas. Los valores faltantes se sustituyen con el valor medio estimado por la regresiòn efectuada en el subgrupo de interés.	Fácil de aplicar	Todas las observaciones pertenecientes al sugrupo tiene el mismo valor imputado. Se subestima la varianza y se introducen sesgos en la correlación. No siempre es fàcil encontrar un modelo adecuado para la varialbe de interés. Se subestima el error estándar del estimador			x
Regresión	Patrón de datos faltantes (MCAR) Se requiere especificar un modelo, en donde las covariables están altamente correlacionadas			Se subestima el error estándar del estimador. Se subestima la varianza	x		x
Regresión por subgrupos	Patrón de datos faltantes (MCAR) Se requiere especificar un modelo, en donde las covariables están altamente correlacionadas con la variable a imputar.	Se divide la base de datos en subgrupos utilizando variables correlacionadas. Los valores faltantes se sustituyen con el valor medio estimado por la regresiòn efectuada en el subgrupo de interés.	Fácil de aplicar	Genera estimadores sesgados. Subestima la varianza de los estimadores Genera sesgos de correlaciòn. No siempre fácil de encontrar un modelo adecuado para la varialbe de interés. Se subestima el error estándar del estimador.	x		x
Regresión aleatoria	Patrón de datos faltantes (MCAR) Se requiere especificar un modelo, en donde las covariables están altamente correlacionadas			No siempre están disponibles Hay que programar el algoritmo que se desea aplicaar			x
Máxima Verosimilitud (EM)	Patrón de datos faltantes (MCAR)		Genera estimaciones robusta basadas en la muestra observado No efectúa simulaciones	No siempre están disponibles Hay que programar el algoritmo que se desea aplicaar	x	x	
Imputación Simple	Patrón de datos faltantes (MCAR) Se requiere especificar un modelo, en donde las covariables están altamente correlacionadas con la variable a		Utiliza procedimeintos estadisticamente robustos	Requiere que los datos faltantes sigan el patrón MAR. Ocurren situaciones en que los supuestos del métodos no se cumplen. No siempre es fácil encontrar un moedelo adecuada para la variable de interés.			x
Imputación Múltiple	Patrón de datos faltantes (MCAR) Se requiere especificar un modelo, en donde las covariables están altamente correlacionadas con la variable a imputar. El modelo que ser requiere para imputar debe ser el mismo o muy similar al que se utilizará en el análisis secundario de datos.	Se generan m subconjuntos de datos imputados por medio de simulaciones. Se combina en forma apropiada a fin de obtener estimadores robustos.	Muy intuitivo y fácil de entender. Utiliza procedimientos estadísticos robustos. Genera distintas opciones de datos imputados y las combina en forma adecuada. Permite calcualr el error estándar de los estimadores.	Requiere que los datos faltantes sigan el patrón MAR. Ocurren situaciones en que los supuestos del métodos no se cumplen. No siempre es fácil encontrar un moedelo adecuada para la variable de interés. Se requiere de paquetes estadisiticos de computo que contengan algoritmos de cálculo para este propósito. Si no se utilizan con precaución pueden funcionar como cajas negras. Se le deja la resposabilidad a un procedimiento estadístico.		x	x

4.5 RESULTADOS

4.5.1 TASA DE NO RESPUESTA

Como fase previa a la aplicación de los procedimientos de imputación se emplea el programa STATA para verificar que la no respuesta no estuviera asociada con algún patrón atípico (NMAR), con el propósito de estar en condiciones de afirmar que los datos que hacen falta se generaron por un proceso aleatorio (MAR) o completamente aleatorio (MCAR), situación que es deseable para la aplicación de los algoritmos de imputación que se comparan.

Antes de iniciar el proceso de imputación, se vio importante conocer las estadísticas descriptivas y la forma que asume la distribución de frecuencias, por lo que en el Anexo No. 2 se presenta estadísticas descriptivas e histograma de la variable con valores perdidos o *missing value*.

Para determinar si la información omitida sigue un patrón aleatorio, se emplea la información generada por STATA a través del comando ***mvpatterns***, el que lista los patrones de valores perdidos de las variables y su frecuencia, se denota por "+" a los valores no perdidos y por "." a los valores perdidos.

A continuación se presenta el proceso de análisis efectuado para la Empresa Eléctrica Ambato, barra Totoras, que incluirá el comando utilizado en STATA y sus resultados.

COMANDO STATA: *mvpatterns MW_MON MW_AMB MW_BAN*

CUADRO 4.3

RESULTADOS DEL PATRÓN DE DATOS DE POTENCIA ACTIVA DE LA POSICIÓN MONTALVO PROCESADOS POR STATA

Variable	type	obs	mv	variable label
mw_mon	float	2919	9	MW_MON
mw_amb	float	2928	0	MW_AMB
mw_ban	float	2928	0	MW_BAN

_pattern	_mv	_freq
+++	0	2919
.++	1	9

De los resultados del programa estadístico se observa que los datos presentan un patrón de datos perdidos aleatorio (.++ indican que la información perdida sigue un patrón aleatorio). De igual forma, todas las posiciones de las diferentes barras de las Empresas Eléctricas, presentan un patrón aleatorio de datos perdidos.

4.5.2 MODELOS AJUSTADOS

Como se había mencionado, como requisito previo para la aplicación de los algoritmos de imputación que utilizan modelos de regresión, es necesario ajustar los modelos propuestos y verificar la significancia estadística de los parámetros asociados a las covariables incorporadas en las ecuaciones propuestas. Por lo tanto, previamente se encontrará cómo se relaciona la variable de interés con un grupo de covariables correlacionadas y posteriormente se realizarán las respectivas regresiones.

De acuerdo a lo indicado, se realiza el análisis detallado para la Empresa Eléctrica Ambato, y se presentan los resultados para las demás Empresas.

1. EMPRESA: E.E.AMBATO

BARRA: Totoras

CUADRO 4.4

RESULTADOS DEL PROGRAMA STATA DEL ANÁLISIS DE CORRELACIÓN DE LAS VARIABLES CORRESPONDIENTES A LA BARRA DE TOTORAS

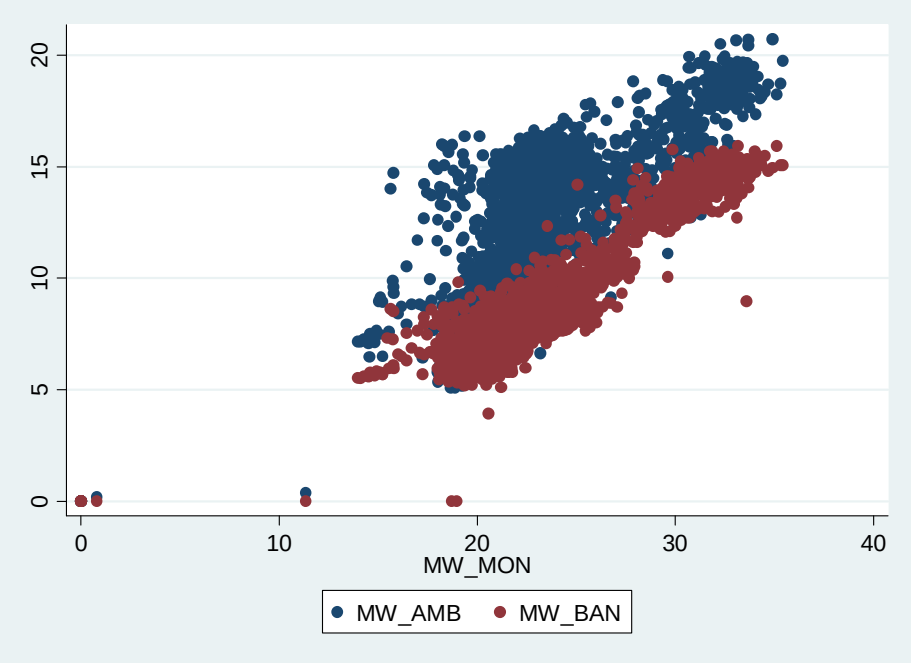
	mw_mon	mw_ban	mw_amb	dia	hora	n_dia_~m
mw_mon	1.0000					
mw_ban	0.9223	1.0000				
mw_amb	0.7888	0.8104	1.0000			
dia	-0.0222	-0.0304	0.0444	1.0000		
hora	0.5399	0.6091	0.5281	-0.0002	1.0000	
n_dia_sem	0.0651	0.0447	0.1308	0.0376	-0.0003	1.0000

Debido a que el dato perdido a imputar corresponde a la potencia activa de la barra de Totoras de la Empresa Eléctrica Ambato de la posición Montalvo, se emplearán para modelar su comportamiento las variables con mayor correlación, las que corresponde a potencia activa de la posición Baños (mw_ban) y la potencia activa de la posición Ambato (mw_amb). Cabe señalar que las variables mw_ban y mw_amb no presentan datos perdidos, tal como se puede observar en el punto 4.4.1.

Se presenta el gráfico de dispersión y los resultados de la regresión:

GRÁFICO 4.1

ANÁLISIS DE DISPERSIÓN DE LA POSICIÓN MONTALVO CONSIDERANDO LAS POSICIONES AMBATO Y BAÑOS.



CUADRO 4.5

RESULTADOS DE STATA RESPECTO A LA REGRESIÓN DE LA POTENCIA ACTIVA DE LA POSICIÓN MONTALVO

Source	SS	df	MS	Number of obs	= 2919
				F(2, 2916)	= 8637.44
Model	37928.2853	2	18964.1427	Prob > F	= 0.0000
Residual	6402.29445	2916	2.19557423	-squared	= 0.8556
				Adj R-squared	= 0.8555
Total	44330.5798	2918	15.192111	Root MSE	= 1.4817

mw_mon	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mw_ban	1.296361	.0188879	68.63	0.000	1.259326	1.333396
mw_amb	.1284507	.012783	10.05	0.000	.1033861	.1535154
_cons	10.43846	.0990043	105.43	0.000	10.24433	10.63258

Los resultados muestran que los parámetros resultaron estadísticamente significativos al 1% y el coeficiente de determinación (R^2) arroja valores adecuados.

2. CATEG-SD

a) Barra: **Pascuales,** Posición: **Cervecería**

El dato perdido a imputar corresponde a la potencia activa de la barra de Pascuales de la Empresa CATEG-SD de la posición Cervecería; por tanto, para el proceso de imputación se emplearán la potencia reactiva de la posición Cervecería (mvar_cer). No se emplea la variable potencia activa de la posición Vergeles (mw_ver) pues esta variable también tiene datos perdidos.

CUADRO 4.6
RESULTADOS DE STATA RESPECTO A LA REGRESIÓN DE LA POTENCIA ACTIVA DE LA POSICIÓN CERVECERÍA

Source	SS	df	MS	Number of obs	= 2924
Model	82022.964	1	82022.964	F(1, 2922)	= 4714.14
Residual	50840.8875	2922	17.3993455	Prob > F	= 0.0000
				R-squared	= 0.6173
				Adj R-squared	= 0.6172
Total	132863.851	2923	45.454619	Root MSE	= 4.1713

Los resultados muestran que los parámetros resultaron estadísticamente significativos al 1% y el coeficiente de determinación (R^2) arroja valores aceptables.

b) Barra: **Pascuales,** Posición: **Vergeles**

El dato por imputar es la potencia activa de la posición Vergeles de la Empresa CATEG, para este fin se emplearán las variables potencia reactiva de la posición Vergeles (mvar_ver) y la potencia reactiva de la posición Cervecería (mvar_cer).

CUADRO 4.7
RESULTADOS DE LA REGRESIÓN DE LA POTENCIA ACTIVA DE LA POSICIÓN VERGELES

Source	SS	df	MS	Number of obs	= 2925
Model	137622.962	2	68811.4812	F(2, 2922)	= 4228.90
Residual	47545.9882	2922	16.2717276	Prob > F	= 0.0000
				R-squared	= 0.7432
				Adj R-squared	= 0.7431
Total	185168.951	2924	63.3272745	Root MSE	= 4.0338

La revisión de los resultados muestra que los parámetros son estadísticamente significativos al 1% y el coeficiente de determinación (R^2) arroja valores aceptables.

c) BARRA: **Trinitaria**, Posición: **Padre Canals**

La potencia activa (MW) de esta barra no posee datos perdidos, por lo tanto, no se aplican los procesos de imputación.

d) BARRA: **Trinitaria**, Posición: **Guasmo**

La variable potencia activa de la posición Guasmo se imputará primero en función de la variable potencia reactiva de la misma posición (mvar_gua) y también una conjunción de la variable potencia reactiva (mvar_ gua) y de la variable potencia activa de la posición Padre Canales (mw_pca)

d.1) Regresión Lineal con la variable potencia reactiva (mvar:gua)

CUADRO 4.8
RESULTADOS DE LA REGRESIÓN DE LA POTENCIA REACACTIVA DE LA POSICIÓN GUASMO

Source	SS	df	MS	Number of obs	= 2905
Model	30571.8984	1	30571.8984	F(1, 2903)	= 1584.98
Residual	55994.4403	2903	19.2884741	Prob > F	= 0.0000
Total	86566.3387	2904	29.8093453	R-squared	= 0.3532
				Adj R-squared	= 0.3529
				Root MSE	= 4.3919

Los resultados muestran que los parámetros no resultaron estadísticamente significativos al 1% y el coeficiente de determinación (R^2) no arroja valores aceptables, por esta razón se analizara el otro caso.

d.2) Regresión lineal con las variables potencia reactiva (mvar_gua) y potencia activa (mw_pca)

CUADRO 4.9
RESULTADOS DE LA REGRESIÓN DE LA POTENCIA ACTIVA Y REACTIVA DE LA POSICIÓN GUASMO

Source	SS	df	MS	Number of obs	= 2904
Model	78130.4658	2	39065.2329	F(2, 2901)	= 13487.61
Residual	8402.39627	2901	2.89637927	Prob > F	= 0.0000
Total	86532.8621	2903	29.808082	R-squared	= 0.9029
				Adj R-squared	= 0.9028
				Root MSE	= 1.7019

Los resultados muestran que los parámetros resultaron estadísticamente significativos al 1% y el coeficiente de determinación (R^2) arroja valores aceptable y con mayor (R^2). Por lo tanto se empleará para la modelación esta relación.

e) BARRA: **Trinitaria**, Posición: **Pradera**

La variable potencia activa de la posición Pradera se imputará en función de la variable potencia reactiva de la misma posición (mvar_pra) y también a través de las dos variables potencia reactiva de la posición Pradera (mvar_pra) y de la variable potencia activa de la posición Padre Canales (mw_pca).

e.1) Regresión lineal con la variable potencia reactiva (mvar_pra)

CUADRO 4.10
RESULTADOS DE LA REGRESIÓN DE LA POTENCIA ACTIVA DE LA POSICIÓN PRADERA

Source	SS	df	MS	Number of obs	= 2699
				F(1, 2697)	= 1809.78
Model	30427.2715	1	30427.2715	Prob > F	= 0.0000
Residual	45343.7159	2697	16.8126496	R-squared	= 0.4016
				Adj R-squared	= 0.4013
Total	75770.9873	2698	28.0841317	Root MSE	= 4.1003

Los resultados muestran que los parámetros no resultaron estadísticamente significativos al 1% y el coeficiente de determinación (R^2) no arroja valores aceptables, por esta razón se analizara el otro caso.

e.2) Regresión lineal con las variables potencia reactiva (mvar_pra) y potencia activa (mw_pca)

CUADRO 4.11
RESULTADOS DE LA REGRESIÓN DE LA POTENCIA ACTIVA Y REACTIVA DE LA POSICIÓN PRADERA

Source	SS	df	MS	Number of obs	= 2698
				F(2, 2695)	= 2771.60
Model	50983.6664	2	25491.8332	Prob > F	= 0.0000
idual	24787.3143	2695	9.19751922	R-squared	= 0.6729
				Adj R-squared	= 0.6726
Total	75770.9807	2697	28.0945423	Root MSE	= 3.0327

Los resultados muestran que los parámetros resultaron estadísticamente significativos al 1% y el coeficiente de determinación (R^2) arroja valores adecuados con mayor (R^2).

Sobre la base de lo anterior, la modelación debe realizarse con las variables mvar_pra y mw_pca.

f) BARRA: **Policentro**

En esta posición existe solamente un dato perdido de potencia activa, el cual será imputado a través de la variable potencia reactiva del mismo punto (mvar_pol) dado que corresponde a la variable de más fuerte correlación.

CUADRO 4.12
RESULTADOS DE LA REGRESIÓN DE LA POTENCIA ACTIVA DE LA POSICIÓN POLICENTRO

Source	SS	df	MS	Number of obs	= 2927
				F(1, 2925)	= 11431.32
Model	799603.252	1	799603.252	Prob > F	= 0.0000
Residual	204599.282	2925	69.9484726	R-squared	= 0.7963
				Adj R-squared	= 0.7962
Total	1004202.53	2926	343.199772	Root MSE	= 8.3635

Se observa del cuadro anterior que los parámetros resultaron estadísticamente significativos al 1% y el coeficiente de determinación (R^2) arroja valores adecuados.

3. EMPRESA ELÉCTRICA QUITO (EEQ)

BARRA: **Pomasqui**

Debido a que el dato perdido a imputar corresponde a la potencia activa de la barra de Pomasqui de la Empresa Eléctrica Quito de la posición Quito 1, se empleó para su modelación la variable con mayor correlación que corresponde a potencia activa de la posición Quito 2 (mvar_qui2).

CUADRO 4.13
RESULTADOS DE LA REGRESIÓN DE LA POTENCIA ACTIVA DE LA POSICIÓN POMASQUI

Source	SS	df	MS	Number of obs = 2921
				F(1, 2919) =13119.30
Model	561500.458	1	561500.458	Prob > F = 0.0000
Residual	124931.981	2919	42.7995823	R-squared = 0.8180
				Adj R-squared = 0.8179
Total	686432.438	2920	235.079602	Root MSE = 6.5421

Los resultados son estadísticamente significativos al 1% y el coeficiente de determinación (R^2) arroja valores adecuados.

4. EMPRESA ELÉCTRICA COTOPAXI (ELEPCO)

BARRA: **Mulaló;** Posición**: Ambato**

La variable que posee datos perdidos corresponde a potencia activa de la posición Ambato (MW) de la Empresa ELEPCO, para lo que se emplea la variable de mayor correlación, en este caso, la potencia reactiva (MVAR).

CUADRO 4.14
RESULTADOS DE LA REGRESIÓN DE LA POTENCIA ACTIVA DE LA POSICIÓN AMBATO DE LA BARRA MULALÓ

Source	SS	df	MS		
				Number of obs	= 2921
				F(1, 2919)	=13173.73
Model	9137.01002	1	9137.01002	Prob > F	= 0.0000
Residual	2024.55497	2919	.693578269	R-squared	= 0.8186
				Adj R-squared	= 0.8186
Total	11161.565	2920	3.82245376	Root MSE	= .83281

Los resultados arrojan que los parámetros son estadísticamente significativos al 1% y el coeficiente de determinación (R^2) arroja valores aceptables.

5. EMPRESA ELÉCTRICA GUAYAS – LOS RÍOS (EMELGUR)

a) BARRA: **Dos Cerritos**

La variable a imputar corresponde a potencia activa (MW) de la barra Dos Cerritos de la Empresa EMELGUR para lo que se utiliza la variable de mayor correlación que en este caso corresponde a la potencia reactiva de la misma barra (MVAR).

CUADRO 4.15
RESULTADOS DE LA REGRESIÓN DE LA POTENCIA ACTIVA DE LA POSICIÓN DOS CERRITOS

Source	SS	df	MS		
				Number of obs	= 2871
				F(1, 2869)	= 7810.57
Model	155481.789	1	155481.789	Prob > F	= 0.0000
Residual	57111.986	2869	19.9065828	R-squared	= 0.7314
				Adj R-squared	= 0.7313
Total	212593.775	2870	74.074486	Root MSE	= 4.4617

Los resultados resultaron estadísticamente significativos al 1% y el coeficiente de determinación (R^2) arroja valores adecuados.

BARRA: **Pascuales**

La potencia activa (MW) de esta barra no posee datos perdidos, por lo tanto, no se aplica los procesos de regresión lineal.

6. EMPRESA ELÉCTRICA REGIONAL NORTE (EMELNORTE)

BARRA: **Ibarra;** Posición: **Otavalo**

La variable que posee datos perdidos corresponde a potencia activa de la posición Otavalo de la barra Ibarra de la Empresa EMELNORTE; por lo tanto, para realizar las imputaciones se emplea la variable con mayor correlación que en este caso corresponde a la potencia reactiva de la posición Cotacachi (mvar_cot).

CUADRO 4.16
RESULTADOS DE LA REGRESIÓN DE LA POTENCIA ACTIVA DE LA POSICIÓN OTAVALO

Source	SS	df	MS	Number of obs = 2926
				F(1, 2924) = 1606.37
Model	16444.9587	1	16444.9587	Prob > F = 0.0000
Residual	29933.9614	2924	10.2373329	R-squared = 0.6546
				Adj R-squared = 0.6544
Total	46378.9201	2925	15.8560411	Root MSE = 3.1996

Los resultados muestran que los parámetros resultaron estadísticamente significativos al 1% y el coeficiente de determinación (R^2) arroja valores adecuados.

7. Empresa Eléctrica Santo Domingo (EMELSAD)

En esta empresa no se registraron valores perdidos para ninguna de las posiciones Santo Domingo 1 y Santo Domingo 2, por lo tanto, no se aplican los procesos de regresión lineal.

4.5.3 PROCEDIMIENTO HOT DECK

Debido a que se realiza la imputación de las observaciones de la variable potencia activa, que tiene datos perdidos, es posible aplicar el procedimiento hot-deck utilizando la sintaxis en STATA. Mediante la opción impute (#) se indica el número de simulaciones que se desean efectuar; para el caso de este análisis se emplearán 100, en concordancia con lo establecido por la ley de los grandes números, en el Anexo 4, se realiza una comparación gráfica que sustenta lo señalado; además, se señala que esta comparación se describe para esta Empresa únicamente, y para la demás se sigue el mismo esquema.

Para la primera empresa “E.E. Ambato” se presenta la sintaxis empleada en el programa estadístico STATA y su salida. Para las demás empresas, se exponen los resultados.

1. EMPRESA ELÉCTRICA AMBATO

BARRA: **Totoras;** Posición: **Montalvo**

Como se requiere efectuar la imputación de la variable mw_mon, que cumplía el requisito de significancia estadística, es posible aplicar el procedimiento hot-deck.

Comando de STATA: *hotdeck mw_mon using impmon, store by(dia hora) impute(5) keep (anio mes dia hora n_dia_sem mw_amb mw_ban)*

CUADRO 4.17

RESULTADO DEL PROGRAMA STATA AL EJECUTAR EL COMANDO HOTDECK

DELETING all matrices....
Table of the Missing data patterns
* signifies missing and - is not missing
Varlist order: mw_mon

pattern	Freq.	Percent	Cum.
*	9	0.31	0.31
-	2,919	99.69	100.00
Total	2,928	100.00	

Luego de concluidas la cien simulaciones solicitadas al programa STATA, los resultados arrojados de los aplicación son mostrados en el Cuadro 4.17 e indican que las simulaciones se ejecutaron sobre el 0.31% de datos perdidos de una base de datos de 2.919 registros.

Al aplicar el comando impute(#) se generan simulaciones en concordancia con el valor colocado en #. Debido a que en el estudio se colocó 100 imputaciones, entonces se generan 100 simulaciones que son guardadas por el programa STATA en un directorio y archivo distinto al que se trabaja. En este caso, se crearon y guardaron los archivo impmon1_dia-hora,……., impmon100_dia-hora.

Se debe considerar que los archivos mencionados se respaldan en el directorio donde el usuario de STATA mantiene en forma regular sus datos, en caso de problemas es posible utilizar el comando “cd” que permite observar en la pantalla la ruta de acceso y el nombre del directorio donde se encuentran alojados los archivos creados.

Con los cien archivos creados se procede a encontrar el promedio de las imputaciones realizadas por el programa. A continuación se presentan los resultados obtenidos luego de extraer el promedio de las cien imputaciones:

CUADRO 4.18
RESULTADO DE APLICAR EL MÉTODO HOT DECK A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN MONTALVO

Nº	ANIO	MES	DIA	HORA	N_DIA_SEM	mw_hd
1	2007	9	3	8	2	21.725
2	2007	9	3	12	2	22.895
3	2007	9	5	12	4	20.761
4	2007	9	5	13	4	19.888
5	2007	9	5	14	4	20.379
6	2007	9	12	13	4	21.040
7	2007	9	12	14	4	21.588
8	2007	9	13	13	5	21.872
9	2007	9	13	16	5	22.678

2. CATEG-SD

a) BARRA: **Pascuales;** Posición: **Vergeles**

Se presentan los resultados obtenidos luego de extraer el promedio de las cien imputaciones ejecutadas por el programa estadístico STATA:.

CUADRO 4.19
RESULTADO DE APLICAR EL MÉTODO HOT DECK A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN VERGELES.

ANIO	MES	DIA	HORA	N_DIA_SEM	mw_hd
2007	9	3	17	2	39.753
2007	9	3	18	2	38.178
2007	9	22	9	7	41.360

b) BARRA: **Pascuales;** Posición: **Cervecería**

Los resultados obtenidos luego de extraer el promedio de las cien imputaciones ejecutadas por el programa estadístico son:

CUADRO 4.20
RESULTADO DE APLICAR EL MÉTODO HOT DECK A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN CERVECERÍA.

ANIO	MES	DIA	HORA	N_DIA_SEM	mw_hd
2007	9	23	7	1	35.049
2007	9	23	8	1	34.106
2007	9	23	9	1	38.291
2007	9	23	10	1	39.735

b) BARRA: **Trinitaria;** Posición: **Padre Canales**

No se ejecuta el procedimiento debido a que esta barra no tiene datos perdidos

c) BARRA: **Trinitaria;** Posición: **Guasmo**

Se presentan los resultados obtenidos luego de calcular el promedio de las cien imputaciones ejecutadas por el programa estadístico:

CUADRO 4.21
RESULTADO DE APLICAR EL MÉTODO HOT DECK A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN GUASMO.

ANIO	MES	DIA	HORA	N_DIA_SEM	mw_hd
2007	9	3	3	2	22.796
2007	9	3	6	2	21.657
2007	9	3	10	3	21.660
2007	9	3	20	2	36.623
2007	9	5	9	4	25.206
2007	9	5	10	4	24.323
2007	9	5	20	4	35.880

2007	9	10	10	2	21.453
2007	9	10	11	2	22.546
2007	9	10	12	2	24.117
2007	9	10	13	2	22.194
2007	9	11	9	9	23.326
2007	9	11	14	3	27.153
2007	9	11	15	3	26.868
2007	9	21	10	6	23.085
2007	9	24	6	2	23.189
2007	9	24	7	2	22.244
2007	9	24	9	2	21.596
2007	9	25	10	3	24.522
2007	9	30	7	1	22.384
2007	9	30	8	1	22.474
2007	9	30	9	1	24.310
2007	9	30	10	1	23.995

d) BARRA: **Trinitaria;** Posición: **Pradera**

Dada la magnitud de la información faltante se presenta un grupo de datos obtenidos por este método.

CUADRO 4.22

Resultado de aplicar el método hot deck a la variable Potencia Activa de la posición Pradera.

ANIO	MES	DIA	HORA	N_DIA_SEM	mw_hd
2007	9	1	3	7	19.793
2007	9	1	6	7	12.049
2007	9	1	7	7	19.766
2007	9	1	10	7	23.631

2007	9	1	11	7	25.725
2007	9	1	13	7	18.946
2007	9	1	18	7	22.901
2007	9	1	21	7	27.523
2007	9	1	22	7	26.934
2007	9	2	1	1	22.754
2007	9	2	2	1	20.020
2007	9	2	4	1	16.853
2007	9	2	7	1	19.265
2007	9	2	9	1	15.992
2007	9	2	12	1	24.234
2007	9	2	19	1	30.235
2007	9	2	21	1	30.547
2007	9	2	24	1	17.804
2007	9	3	4	2	21.087
2007	9	3	5	2	17.791
2007	9	4	7	3	18.617
2007	9	4	11	3	23.569
2007	9	4	12	3	20.270
2007	9	4	18	3	24.818
2007	9	4	21	3	31.017
2007	9	4	22	3	25.436
2007	9	5	1	4	22.175
2007	9	5	2	4	20.878
2007	9	5	4	4	20.465
2007	9	5	5	4	16.862
2007	9	5	8	4	21.108

e) BARRA: **Policentro**

Se presentan los resultados obtenidos luego de calcular el promedio de las cien imputaciones ejecutadas por el programa estadístico:

CUADRO 4.23

RESULTADO DE APLICAR EL MÉTODO HOT DECK A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN POLICENTRO.

ANIO	MES	DIA	HORA	N_DIA_SEM	mw_hd
2007	9	3	18	2	89.518

3. EMPRESA ELÉCTRICA QUITO

BARRA: **Pomasqui**

Los resultados arrojados luego de calcular el promedio de las cien imputaciones se presentan a continuación:

CUADRO 4.24

Resultado de aplicar el método Hot Deck a la variable Potencia Activa de la posición Pomasqui

ANIO	MES	DIA	HORA	N_DIA_SEM	mw_hd
2007	9	18	10	3	60.374
2007	9	18	13	3	60.814
2007	9	18	15	3	57.568
2007	9	18	17	3	53.610
2007	9	18	18	3	58.995
2007	9	20	17	5	74.719
2007	9	20	18	5	73.968

4. EMPRESA ELÉCTRICA COTOPAXI

BARRA: **Mulaló;** Posición: **Ambato**

Los resultados obtenidos luego de extraer el promedio de las cien imputaciones ejecutadas por el programa estadístico son:

CUADRO 4.25
RESULTADO DE APLICAR EL MÉTODO HOT DECK A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN AMBATO

ANIO	MES	DIA	HORA	N_DIA_SEM	mw_hd
2007	9	4	17	3	6.795
2007	9	4	18	3	8.751
2007	9	12	3	4	3.481
2007	9	12	4	4	4.391
2007	9	24	11	2	7.635
2007	9	25	6	3	7.915
2007	9	27	24	5	4.603

5. EMPRESA ELÉCTRICA REGIONAL GUAYAS – LOS RÍOS

a) BARRA: **Dos Cerritos ATR**

Se presentan los resultados obtenidos luego de calcular el promedio de las cien imputaciones ejecutadas por el programa estadístico:

CUADRO 4.26

RESULTADO DE APLICAR EL MÉTODO HOT DECK A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN DOS CERRITOS ATR

ANIO	MES	DIA	HORA	N_DIA_SEM	mw_hd
2007	9	1	2	7	42.507
2007	9	1	3	7	40.728
2007	9	1	4	7	41.257
2007	9	1	5	7	39.972
2007	9	1	6	7	37.729
2007	9	1	7	7	39.509
2007	9	23	8	1	38.283
2007	9	23	9	1	43.303
2007	9	23	10	1	45.123
2007	9	23	11	1	48.675
2007	9	23	12	1	49.251
2007	9	23	13	1	47.754
2007	9	23	14	1	50.332
2007	9	23	15	1	49.922
2007	9	24	6	2	38.703
2007	9	24	7	2	39.194
2007	9	24	8	2	36.645
2007	9	24	9	2	43.585
2007	9	24	10	2	45.623
2007	9	24	11	2	46.565
2007	9	24	12	2	47.269
2007	9	24	13	2	51.067
2007	9	24	14	2	49.297
2007	9	24	15	2	51.409
2007	9	24	16	2	50.580
2007	9	24	17	2	48.246

2007	9	24	18	2	47.922
2007	9	26	6	4	41.972
2007	9	26	7	4	34.562
2007	9	26	8	4	39.009
2007	9	26	9	4	39.132
2007	9	26	10	4	39.246
2007	9	26	11	4	36.101
2007	9	26	12	4	42.581
2007	9	26	13	4	42.417
2007	9	26	14	4	49.378
2007	9	26	15	4	41.320
2007	9	26	16	4	42.775
2007	9	27	7	5	35.493
2007	9	27	8	5	39.295
2007	9	27	9	5	42.068
2007	9	27	10	5	46.485
2007	9	27	11	5	49.442
2007	9	27	12	5	48.913
2007	9	27	13	5	50.010
2007	9	27	14	5	50.853
2007	9	27	15	5	50.513
2007	9	27	16	5	49.722
2007	9	28	7	6	38.247
2007	9	28	8	6	37.998
2007	9	28	9	6	42.803
2007	9	28	10	6	45.443
2007	9	28	11	6	46.046
2007	9	28	12	6	47.621
2007	9	28	13	6	48.391
2007	9	28	14	6	47.958
2007	9	28	15	6	48.395

b) BARRA: **Pascuales**

No se ejecuta el procedimiento debido a que esta barra no tiene datos perdidos.

6. EMPRESA ELÉCTRICA REGIONAL NORTE

BARRA: **Ibarra;** Posición: **Otavalo**

Los resultados obtenidos luego de extraer el promedio de las cien imputaciones ejecutadas por el programa estadístico son:

CUADRO 4.27
RESULTADO DE APLICAR EL MÉTODO HOT DECK A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN OTAVALO

ANIO	MES	DIA	HORA	N_DIA_SEM	mw_hd
2007	9	3	8	2	19.058
2007	9	3	10	2	20.516

7. EMPRESA ELÉCTRICA SANTO DOMINGO

BARRA: **Santo Domingo**

No se ejecuta el procedimiento debido a que esta barra no tiene datos perdidos

4.5.4 PROCEDIMIENTO HOT DECK CON REGRESIÓN

Si se conoce que existe una relación de dependencia entre la variable de interés y un vector de covariables, el comando hot-deck efectúa la imputación a partir de los métodos de regresión e igual que en caso anterior se ejecuta la simulación para 100 imputaciones. Para la primera empresa “E.E. Ambato” se presenta la

sintaxis empleada en el programa estadístico STATA y su salida. Para las demás empresas, se exponen los resultados.

1. EMPRESA ELÉCTRICA AMBATO

BARRA: **Totoras;** Posición: **Montalvo**

Comando de STATA: *hotdeck mw_mon using imphdrmon, store by (hora n_dia_sem) impute (100) keep (anio mes dia hora n_dia_sem mw_amb mw_ban) command (regress mw_mon mw_amb mw_ban) parms (mw_amb mw_ban)*

CUADRO 4.28
RESULTADO DEL PROGRAMA STATA AL EJECUTAR EL COMANDO HOTDECK CON REGRESIÓN

```
  pattern         Freq.     Percent       Cum.
------------------------------------------------------------
      *               9            0.31          0.31
      -           2,919           99.69        100.00

              Total           2,928      100.00
2928

          Number of Obs.              =        2928
          No. of Imputations          =        100
          % Lines of Missing Data     =    .30737705       %
          F( 773.403 ,2)              =        8633.9992
          Prob > F                    = 0.0000

Variable        Average   Between   Within    Total       df          t        p-value
                   Coef.    Imp. SE   Imp. SE     SE

mw_amb          0.1279     0.001     0.013     0.013        1.1e+07          9.994   0.000
mw_ban          1.2968     0.001     0.019     0.019        2.9e+07          68.598  0.000

Variable        [95% Conf. Interval]

mw_amb          0.0992    0.1566
mw_ban          1.2544    1.3392
```

Al igual que en el caso anterior, en este caso se generan tanto archivos como simulaciones sean requeridas. En el estudio se generaron cien imputaciones y los resultados serán almacenados en los archivos impmon1_dia-

ndsemana,……., impmon100_dia-ndsemana. Posteriormente con estos archivos se generan los promedios de los parámetros estimados de las cien imputaciones efectuadas, mismas que son empleadas para imputar los datos faltantes.

Al observar la tabla que arroja STATA, se advierte la existencia de la columna Average Coef., que representa los valores de los coeficientes de regresión que se deben aplicar para generar los valores imputados, los cuales se obtuvieron como un promedio simple de las cien simulaciones efectuadas.

De la misma forma que en el método hot deck se presentan resultados obtenidos luego de extraer el promedio de las cien imputaciones ejecutadas por el programa estadístico son:

CUADRO 4.29
RESULTADO DE APLICAR EL MÉTODO HOT DECK CON REGRESIÓN A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN MONTALVO

ANIO	MES	DIA	HORA	mw_hdr
2007	9	3	8	21.653
2007	9	3	12	22.351
2007	9	5	12	24.189
2007	9	5	13	22.693
2007	9	5	14	24.056
2007	9	12	13	23.057
2007	9	12	14	24.049
2007	9	13	13	22.675
2007	9	13	16	23.568

2. CATEG - SD

a) BARRA: **Pascuales;** Posición: Vergeles

Los resultados obtenidos luego de extraer el promedio de las cien imputaciones ejecutadas por el programa estadístico son:

CUADRO 4.30
RESULTADO DE APLICAR EL MÉTODO HOT DECK CON REGRESIÓN A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN VERGELES.

ANIO	MES	DIA	HORA	mw_hdr
2007	9	3	17	45.854
2007	9	3	18	47.553
2007	9	22	9	34.269

b) BARRA: **Pascuales;** Posición: Cervecería

Se presentan los resultados obtenidos luego de extraer el promedio de las cien imputaciones ejecutadas por el programa estadístico:

CUADRO 4.31
RESULTADO DE APLICAR EL MÉTODO HOT DECK CON REGRESIÓN A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN CERVECERÍA.

ANIO	MES	DIA	HORA	mw_hdr
2007	9	23	7	29.745
2007	9	23	8	28.553
2007	9	23	9	30.474
2007	9	23	10	30.891

c) BARRA: **Pascuales;** Posición: **Padre Canals**

No se ejecuta el procedimiento debido a que esta barra no tiene datos perdidos

d) BARRA: **Pascuales;** Posición: **Guasmo**

Los resultados obtenidos del cálculo del promedio de las cien imputaciones ejecutadas por el programa estadístico, se presentan a continuación:

CUADRO 4.32
RESULTADO DE APLICAR EL MÉTODO HOT DECK CON REGRESIÓN A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN GUASMO

ANIO	MES	DIA	HORA	mw_hdr
2007	9	3	3	21.346
2007	9	3	6	24.122
2007	9	3	10	23.979
2007	9	3	20	37.232
2007	9	5	9	26.599
2007	9	5	10	28.044
2007	9	5	20	39.460
2007	9	10	10	23.203
2007	9	10	11	24.788
2007	9	10	12	28.029
2007	9	10	13	27.171
2007	9	11	9	25.480
2007	9	11	14	27.461
2007	9	11	15	28.635
2007	9	21	10	26.724
2007	9	24	6	28.517
2007	9	24	7	27.403
2007	9	24	9	34.703
2007	9	25	10	39.753
2007	9	30	7	27.979
2007	9	30	8	25.520
2007	9	30	9	25.627
2007	9	30	10	25.226

e) BARRA: **Pascuales;** Posición: **Pradera**

Por la magnitud de la información faltante, se presenta un grupo datos extraídos, por este método.

CUADRO 4.33
RESULTADO DE APLICAR EL MÉTODO HOT DECK CON REGRESIÓN A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN PRADERA

ANIO	MES	DIA	HORA	mw_hdr
2007	9	1	3	20.704
2007	9	1	6	21.016
2007	9	1	7	18.640
2007	9	1	10	17.141
2007	9	1	11	16.627
2007	9	1	13	24.350
2007	9	1	18	26.205
2007	9	1	21	28.191
2007	9	1	22	27.114
2007	9	2	1	19.952
2007	9	2	2	21.176
2007	9	2	4	16.931
2007	9	2	7	18.490
2007	9	2	9	19.473
2007	9	2	12	22.515
2007	9	2	19	29.002
2007	9	2	21	28.411
2007	9	2	24	18.608
2007	9	3	4	18.891
2007	9	3	5	21.405
2007	9	4	7	22.459
2007	9	4	11	18.175

2007	9	4	12	20.586
2007	9	4	18	27.445
2007	9	4	21	28.785
2007	9	4	22	27.734
2007	9	5	1	18.825
2007	9	5	2	19.527
2007	9	5	4	20.747
2007	9	5	5	17.496

f) BARRA: **Policentro**

A continuación se exponen los resultados obtenidos luego de extraer el promedio de las cien imputaciones ejecutadas por el programa estadístico:

CUADRO 4.34
RESULTADO DE APLICAR EL MÉTODO HOT DECK CON REGRESIÓN A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN CERVECERÍA.

ANIO	MES	DIA	HORA	mw_hdr
2007	9	3	18	99.312

3. EMPRESA ELÉCTRICA QUITO

BARRA: **Pomasqui**

Se presentan los resultados obtenidos luego de calcular el promedio de las cien imputaciones ejecutadas por el programa estadístico:

CUADRO 4.35
RESULTADO DE APLICAR EL MÉTODO HOT DECK CON REGRESIÓN A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN POMASQUI

ANIO	MES	DIA	HORA	mw_hdr
2007	9	18	10	73.185
2007	9	18	13	73.847
2007	9	18	15	68.902
2007	9	18	17	71.676
2007	9	18	18	64.773
2007	9	20	17	73.210
2007	9	20	18	79.775

4. EMPRESA ELÉCTRICA COTOPAXI

BARRA: **Mulaló;** Posición: **Ambato**

Los resultados obtenidos luego de extraer el promedio de las cien imputaciones ejecutadas por el programa estadístico son:

CUADRO 4.36
RESULTADO DE APLICAR EL MÉTODO HOT DECK CON REGRESIÓN A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN AMBATO.

ANIO	MES	DIA	HORA	mw_hdr
2007	9	4	17	5.128
2007	9	4	18	5.323
2007	9	12	3	6.634
2007	9	12	4	5.684
2007	9	24	11	7.528
2007	9	25	6	5.646
2007	9	27	24	7.078

5. EMPRESA ELÉCTRICA REGIONAL GUAYAS – LOS RÍOS

a) BARRA: **Dos Cerritos ATR**

Los resultados obtenidos del cálculo del promedio de las cien imputaciones ejecutadas por el programa estadístico, se presentan a continuación:

CUADRO 4.37
RESULTADO DE APLICAR EL MÉTODO HOT DECK CON REGRESIÓN A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN DOS CERRITOS ATR

ANIO	MES	DIA	HORA	mw_hdr
2007	9	1	2	40.869
2007	9	1	3	39.750
2007	9	1	4	39.129
2007	9	1	5	39.450
2007	9	1	6	38.979
2007	9	1	7	36.885
2007	9	23	8	33.193
2007	9	23	9	35.007
2007	9	23	10	37.113
2007	9	23	11	42.060
2007	9	23	12	40.634
2007	9	23	13	40.273
2007	9	23	14	41.102
2007	9	23	15	44.151
2007	9	24	6	39.661
2007	9	24	7	37.074
2007	9	24	8	36.864
2007	9	24	9	41.743
2007	9	24	10	47.542
2007	9	24	11	49.276

2007	9	24	12	49.959
2007	9	24	13	50.444
2007	9	24	14	52.276
2007	9	24	15	55.152
2007	9	24	16	51.362
2007	9	24	17	49.104
2007	9	24	18	50.079
2007	9	26	6	41.442
2007	9	26	7	40.927
2007	9	26	8	39.246
2007	9	26	9	45.682
2007	9	26	10	48.751
2007	9	26	11	51.261
2007	9	26	12	50.044
2007	9	26	13	49.431
2007	9	26	14	52.663
2007	9	26	15	54.608
2007	9	26	16	49.838
2007	9	27	7	39.324
2007	9	27	8	40.334
2007	9	27	9	45.066
2007	9	27	10	50.133
2007	9	27	11	50.040
2007	9	27	12	50.851
2007	9	27	13	50.733
2007	9	27	14	53.679
2007	9	27	15	55.971
2007	9	27	16	52.566
2007	9	28	7	39.712
2007	9	28	8	40.628
2007	9	28	9	45.734
2007	9	28	10	47.657

2007	9	28	11	49.999
2007	9	28	12	51.302
2007	9	28	13	52.168
2007	9	28	14	50.459
2007	9	28	15	52.341

b) BARRA: **Pascuales**

No se ejecuta el procedimiento debido a que esta barra no tiene datos perdidos

6. EMPRESA ELÉCTRICA REGIONAL NORTE

BARRA: **Ibarra;** Posición: **Otavalo**

Se presenta los resultados obtenidos a partir de calcular el promedio de las cien imputaciones ejecutadas por el programa estadístico:

CUADRO 4.38
RESULTADO DE APLICAR EL MÉTODO HOT DECK CON REGRESIÓN A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN OTAVALO

ANIO	MES	DIA	HORA	mw_hdr
2007	9	3	8	18.601
2007	9	3	10	22.691

7. EMPRESA ELÉCTRICA SANTO DOMINGO

BARRA: **Santo Domingo**

No se ejecuta el procedimiento debido a que esta barra no tiene datos perdidos

Para todas las Empresas y con este método se generaron 100 simulaciones, lo que equivale a 100 imputaciones, por lo que esta opción se la puede considerar como un procedimiento de imputación múltiple.

4.5.5 IMPUTACIÓN POR REGRESIÓN

El método de regresión imputa valores a partir de ajustar un modelo que vincula la variable de interés con un vector de covariables. La variable imputada se guarda en el archivo de trabajo y en los resultados que se generan se especifica el porcentaje y número de observaciones que fueron imputadas. Para la primera empresa “E.E. Ambato” se presenta la sintaxis empleada en el programa estadístico STATA y su salida. Para las demás empresas, se exponen los resultados.

1. EMPRESA ELÉCTRICA AMBATO

BARRA: **Totoras;** Posición: **Montalvo**

Comando de STATA: *impute mw_mon mw_ban mw_amb, gen(mw_mon1)*

Resultado de STATA

0.31% (9) observations imputed

Los resultados obtenidos al aplicar este comando son los siguientes:

CUADRO 4.39
RESULTADO DE APLICAR EL MÉTODO REGRESIÓN CONDICIONADA A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN MONTALVO

ANIO	MES	DIA	HORA	mw_imp
2007	9	3	8	23.566
2007	9	3	12	23.566
2007	9	5	12	24.217
2007	9	5	13	22.886
2007	9	5	14	23.304
2007	9	12	13	21.528
2007	9	12	14	22.225
2007	9	13	13	23.168
2007	9	13	16	24.579

2. CATEG-SD

a) BARRA: **Pascuales;** Posición: **Vergeles**

Los resultados obtenidos a través de este método se presentan a continuación:

CUADRO 4.40
RESULTADO DE APLICAR EL MÉTODO REGRESIÓN CONDICIONADA A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN VERGELES

ANIO	MES	DIA	HORA	mw_imp
2007	9	3	17	48.368
2007	9	3	18	43.871
2007	9	22	9	40.107

b) BARRA: **Pascuales;** Posición: **Cervecería**

Se presentan los resultados obtenidos a través del método de imputación por regresión:

CUADRO 4.41
RESULTADO DE APLICAR EL MÉTODO REGRESIÓN CONDICIONADA A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN CERVECERÍA

ANIO	MES	DIA	HORA	mw_imp
2007	9	23	7	33.998
2007	9	23	8	33.085
2007	9	23	9	33.815
2007	9	23	10	35.458

c) BARRA: **Trinitaria;** Posición: **Padre Canals**

No se ejecuta el procedimiento debido a que esta barra no tiene datos perdidos

d) BARRA: **Trinitaria;** Posición: **Guasmo**

Los resultados obtenidos a través de este método se presentan a continuación:

CUADRO 4.42
RESULTADO DE APLICAR EL MÉTODO REGRESIÓN CONDICIONADA A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN GUASMO

ANIO	MES	DIA	HORA	mw_imp
2007	9	3	3	20.326
2007	9	3	6	22.518
2007	9	3	10	25.979
2007	9	3	20	35.752
2007	9	5	9	24.063

2007	9	5	10	24.954
2007	9	5	20	39.369
2007	9	10	10	23.055
2007	9	10	11	24.619
2007	9	10	12	25.799
2007	9	10	13	25.899
2007	9	11	9	26.482
2007	9	11	14	31.431
2007	9	11	15	31.861
2007	9	21	10	25.177
2007	9	24	6	21.132
2007	9	24	7	24.939
2007	9	24	9	25.586
2007	9	25	10	25.234
2007	9	30	7	21.483
2007	9	30	8	22.316
2007	9	30	9	24.099
2007	9	30	10	21.761

e) BARRA: **Trinitaria;** Posición: **Pradera**

Los resultados obtenidos con este método son los siguientes:

Debido al volumen de datos perdidos en esta posición se presenta un grupo de los mismos.

CUADRO 4.43

RESULTADO DE APLICAR EL MÉTODO REGRESIÓN CONDICIONADA A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN PRADERA

ANIO	MES	DIA	HORA	mw_imp
2007	9	1	3	12.922
2007	9	1	6	12.922
2007	9	1	7	12.704
2007	9	1	10	16.343
2007	9	1	11	17.217
2007	9	1	13	18.169
2007	9	1	18	20.375
2007	9	1	21	23.449
2007	9	1	22	21.997
2007	9	2	1	14.762
2007	9	2	2	13.941
2007	9	2	4	12.957
2007	9	2	7	10.614
2007	9	2	9	15.199
2007	9	2	12	14.028
2007	9	2	19	23.565
2007	9	2	21	24.738
2007	9	2	24	25.442
2007	9	3	4	12.543
2007	9	3	5	12.715
2007	9	4	7	14.729
2007	9	4	11	19.294
2007	9	4	12	17.989
2007	9	4	18	18.898
2007	9	4	21	23.971
2007	9	4	22	22.306
2007	9	5	1	21.244
2007	9	5	2	23.951

2007	9	5	4	19.856
2007	9	5	5	20.773

f) BARRA: **Policentro**

Se presentan los resultados obtenidos a través del método de imputación por regresión:

CUADRO 4.44
RESULTADO DE APLICAR EL MÉTODO REGRESIÓN CONDICIONADA A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN POLICENTRO

ANIO	MES	DIA	HORA	mw_imp
2007	9	3	18	100.671

3. EMPRESA ELÉCRICA QUITO

BARRA: **Pomasqui**

Se presentan los resultados obtenidos a través del método de imputación por regresión:

CUADRO 4.45
RESULTADO DE APLICAR EL MÉTODO REGRESIÓN CONDICIONADA A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN QUITO 1

ANIO	MES	DIA	HORA	mw_imp
2007	9	18	10	91.098
2007	9	18	13	90.577
2007	9	18	15	92.356
2007	9	18	17	84.548

2007	9	18	18	87.107
2007	9	20	17	86.413
2007	9	20	18	86.153

4. EMPRESA ELÉCTRICA COTOPAXI

BARRA: **Mulaló;** Posición: **Ambato**

Los resultados obtenidos con este método son los siguientes:

CUADRO 4.46
RESULTADO DE APLICAR EL MÉTODO REGRESIÓN CONDICIONADA A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN AMBATO

ANIO	MES	DIA	HORA	mw_imp
2007	9	4	17	5.6745
2007	9	4	18	5.4847
2007	9	12	3	4.8204
2007	9	12	4	5.0102
2007	9	24	11	6.9081
2007	9	25	6	8.2366
2007	9	27	24	7.3826

5. EMPRESA ELÉCTRICA REGIONAL GUAYAS – LOS RÍOS

a) BARRA: **Dos Cerritos**

Los resultados obtenidos a través de este método se presentan a continuación:

CUADRO 4.47

RESULTADO DE APLICAR EL MÉTODO REGRESIÓN CONDICIONADA A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN DOS CERRITOS

ANIO	MES	DIA	HORA	mw_imp
2007	9	1	2	42.066
2007	9	1	3	41.479
2007	9	1	4	41.049
2007	9	1	5	40.425
2007	9	1	6	39.917
2007	9	1	7	37.768
2007	9	23	8	37.847
2007	9	23	9	40.498
2007	9	23	10	40.511
2007	9	23	11	42.998
2007	9	23	12	43.784
2007	9	23	13	45.138
2007	9	23	14	46.076
2007	9	23	15	44.746
2007	9	24	6	37.321
2007	9	24	7	35.978
2007	9	24	8	39.061
2007	9	24	9	48.125
2007	9	24	10	51.925
2007	9	24	11	55.216
2007	9	24	12	50.639
2007	9	24	13	54.466
2007	9	24	14	57.0466
2007	9	24	15	57.490
2007	9	24	16	57.120
2007	9	24	17	54.002
2007	9	24	18	54.086

2007	9	26	6	39.878
2007	9	26	7	39.721
2007	9	26	8	41.961
2007	9	26	9	49.409
2007	9	26	10	53.160
2007	9	26	11	55.087
2007	9	26	12	55.712
2007	9	26	13	54.574
2007	9	26	14	56.702
2007	9	26	15	58.212
2007	9	26	16	57.431
2007	9	27	7	37.378
2007	9	27	8	37.378
2007	9	27	9	50.035
2007	9	27	10	53.056
2007	9	27	11	55.035
2007	9	27	12	54.879
2007	9	27	13	55.712
2007	9	27	14	57.952
2007	9	27	15	59.098
2007	9	27	16	59.202
2007	9	28	7	39.305
2007	9	28	8	40.138
2007	9	28	9	49.410
2007	9	28	10	53.993
2007	9	28	11	53.941
2007	9	28	12	57.171
2007	9	28	13	56.598
2007	9	28	14	58.108
2007	9	28	15	59.411

b) BARRA: **Pascuales**

No se ejecuta el procedimiento debido a que esta barra no tiene datos perdidos

6. EMPRESA ELÉCTRICA REGIONAL NORTE

BARRA: **Ibarra;** Posición: **Otavalo**

Los resultados obtenidos al aplicar el método de imputación en el programa estadístico STATA, son los siguientes:

CUADRO 4.48
RESULTADO DE APLICAR EL MÉTODO REGRESIÓN CONDICIONADA A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN OTAVALO

ANIO	MES	DIA	HORA	mw_imp
2007	9	3	8	16.408
2007	9	3	10	17.180

7. EMPRESA ELÉCTRICA SANTO DOMINGO S.A.

BARRA: **Santo Domingo**

No se ejecuta el procedimiento debido a que esta barra no tiene datos perdidos

4.5.6 IMPUTACIÓN SIMPLE

Este método imputa valores a partir de un modelo de regresión. El algoritmo genera sólo una simulación y al finalizar la ejecución del comando se reporta el número de observaciones que fueron imputadas. Para la primera empresa “E.E. Ambato” se expone la sintaxis empleada en el programa estadístico STATA con su respectiva salida. Para las demás empresas, se exponen los resultados.

1. EMPRESA ELÉCTRICA AMBATO

BARRA: **Totoras;** Posición: **Montalvo**

La variable imputada, en este caso (mw_mon), se guarda en el archivo de trabajo y la salida del programa estadístico especifica el porcentaje y número de observaciones que fueron imputadas. Se expone, para la empresa “E.E.Ambato” la sintaxis utilizada para ejecutar la estimación, posteriormente se presentan los resultados de las demás variables analizadas

Comando de STATA: *uvis reg mw_mon mw_amb mw_ban, gen(mwmonuvis)*

Resultado de STATA:

> [imputing by drawing from conditional distribution without bootstrap]
>
> 9 missing observations on mw_mon imputed from 2919 complete observations.

Al ejecutar el comando, se generan datos dentro del mismo archivo con los valores estimados para los valores faltantes, a continuación se presentan los resultados:

CUADRO 4.49
RESULTADO DE APLICAR EL MÉTODO IMPUTACIÓN SIMPLE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN MONTALVO

ANIO	MES	DIA	HORA	mw_isim
2007	9	3	8	23.241
2007	9	3	12	22.524
2007	9	5	12	23.042

2007	9	5	13	23.595
2007	9	5	14	25.297
2007	9	12	13	21.805
2007	9	12	14	24.040
2007	9	13	13	23.316
2007	9	13	16	23.551

2. CATEG-SD

a) BARRA: **Pascuales;** Posición: **Vergeles**

Los resultados obtenidos al ejecutar este comando en el programa estadístico son:

CUADRO 4.50
RESULTADO DE APLICAR EL MÉTODO IMPUTACIÓN SIMPLE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN VERGELES

ANIO	MES	DIA	HORA	mw_isim
2007	9	3	17	44.166
2007	9	3	18	40.648
2007	9	22	9	39.265

b) BARRA: **Pascuales;** Posición: **Cervecería**

Se presentan los resultados arrojados por el programa STATA:

CUADRO 4.51
RESULTADO DE APLICAR EL MÉTODO IMPUTACIÓN SIMPLE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN CERVECERÍA

ANIO	MES	DIA	HORA	mw_isim
2007	9	23	7	40.061
2007	9	23	8	31.183
2007	9	23	9	44.129
2007	9	23	10	36.688

c) BARRA: **Trinitaria;** Posición: **Padre Canals**

No se ejecuta el procedimiento debido a que esta barra no tiene datos perdidos

d) BARRA: **Trinitaria;** Posición: **Guasmo**

Se presentan los resultados obtenidos por la corrida del programa estadístico:

CUADRO 4.52
RESULTADO DE APLICAR EL MÉTODO IMPUTACIÓN SIMPLE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN GUASMO

ANIO	MES	DIA	HORA	mw_isim
2007	9	3	3	19.803
2007	9	3	6	25.123
2007	9	3	10	27.070
2007	9	3	20	35.332
2007	9	5	9	23.206
2007	9	5	10	25.608
2007	9	5	20	40.451
2007	9	10	10	22.973
2007	9	10	11	26.231

2007	9	10	12	25.505
2007	9	10	13	24.179
2007	9	11	9	24.213
2007	9	11	14	32.238
2007	9	11	15	30.999
2007	9	21	10	25.894
2007	9	24	6	21.331
2007	9	24	7	23.480
2007	9	24	9	28.108
2007	9	25	10	25.607
2007	9	30	7	20.459
2007	9	30	8	21.157
2007	9	30	9	24.108
2007	9	30	10	24.019

e) BARRA: **Trinitaria;** Posición: **Pradera**

Por la cantidad de datos ausentes, se presenta un grupo de datos obtenidos por este método

CUADRO 4.53
RESULTADO DE APLICAR EL MÉTODO IMPUTACIÓN SIMPLE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN PRADERA

ANIO	MES	DIA	HORA	mw_isim
2007	9	1	3	12.757
2007	9	1	6	16.933
2007	9	1	7	13.895
2007	9	1	10	18.959
2007	9	1	11	15.927
2007	9	1	13	17.141
2007	9	1	18	14.952

2007	9	1	21	21.625
2007	9	1	22	21.519
2007	9	2	1	13.406
2007	9	2	2	12.811
2007	9	2	4	13.219
2007	9	2	7	9.466
2007	9	2	9	17.496
2007	9	2	12	14.422
2007	9	2	19	21.113
2007	9	2	21	24.685
2007	9	2	24	28.003
2007	9	3	4	13.837
2007	9	3	5	15.550
2007	9	4	7	16.052
2007	9	4	11	18.889
2007	9	4	12	20.252
2007	9	4	18	18.737
2007	9	4	21	17.916
2007	9	4	22	18.295
2007	9	5	1	22.607
2007	9	5	2	24.239
2007	9	5	4	14.078
2007	9	5	5	25.480

f) BARRA: **Policentro**

Los resultados obtenidos al ejecutar este comando en el programa estadístico son:

CUADRO 4.54

RESULTADO DE APLICAR EL MÉTODO IMPUTACIÓN SIMPLE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN POLICENTRO

ANIO	MES	DIA	HORA	mw_isim
2007	9	3	18	99.505

3. EMPRESA ELÉCTRICA QUITO

BARRA: **Pomasqui**

Los resultados obtenidos por ejecución de comandos para imputación simple existente en STATA, arrojan los siguientes datos:

CUADRO 4.55
RESULTADO DE APLICAR EL MÉTODO IMPUTACIÓN SIMPLE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN QUITO 1

ANIO	MES	DIA	HORA	mw_isim
2007	9	18	10	104.110
2007	9	18	13	101.85
2007	9	18	15	86.901
2007	9	18	17	83.595
2007	9	18	18	90.784
2007	9	20	17	87.934
2007	9	20	18	96.943

4. EMPRESA ELÉCTRICA COTOPAXI

BARRA: **Mulaló;** Posición: **Ambato**

Los resultados obtenidos al ejecutar este comando en el programa estadístico son:

CUADRO 4.56
RESULTADO DE APLICAR EL MÉTODO IMPUTACIÓN SIMPLE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN AMBATO

ANIO	MES	DIA	HORA	mw_isim
2007	9	4	17	3.527
2007	9	4	18	4.864
2007	9	12	3	4.166
2007	9	12	4	5.925
2007	9	24	11	6.572
2007	9	25	6	8.914
2007	9	27	24	6.920

5. EMPRESA ELÉCTRICA REGIONAL GUAYAS – LOS RÍOS

a) BARRA: **Dos Cerritos**

Los resultados obtenidos al ejecutar este comando en el programa estadístico son:

CUADRO 4.57

RESULTADO DE APLICAR EL MÉTODO IMPUTACIÓN SIMPLE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN DOS CERRITOS

ANIO	MES	DIA	HORA	mw_isim
2007	9	1	2	34.332
2007	9	1	3	41.946
2007	9	1	4	41.839
2007	9	1	5	39.028
2007	9	1	6	44.472
2007	9	1	7	33.341
2007	9	23	8	36.002
2007	9	23	9	40.319
2007	9	23	10	46.785
2007	9	23	11	35.159
2007	9	23	12	46.586
2007	9	23	13	49.645
2007	9	23	14	41.052
2007	9	23	15	42.330
2007	9	24	6	41.408
2007	9	24	7	40.024
2007	9	24	8	30.067
2007	9	24	9	44.125
2007	9	24	10	48.046
2007	9	24	11	51.638
2007	9	24	12	50.343
2007	9	24	13	61.277
2007	9	24	14	54.059
2007	9	24	15	66.112
2007	9	24	16	55.509
2007	9	24	17	58.123
2007	9	24	18	55.964

2007	9	26	6	38.857
2007	9	26	7	44.717
2007	9	26	8	48.971
2007	9	26	9	56.478
2007	9	26	10	45.983
2007	9	26	11	62.166
2007	9	26	12	54.559
2007	9	26	13	62.003
2007	9	26	14	51.550
2007	9	26	15	62.576
2007	9	26	16	63.524
2007	9	27	7	35.260
2007	9	27	8	39.386
2007	9	27	9	44.883
2007	9	27	10	44.688
2007	9	27	11	61.914
2007	9	27	12	53.047
2007	9	27	13	59.234
2007	9	27	14	64.159
2007	9	27	15	65.089
2007	9	27	16	48.848
2007	9	28	7	47.866
2007	9	28	8	41.883
2007	9	28	9	43.308
2007	9	28	10	62.695
2007	9	28	11	57.231
2007	9	28	12	54.138
2007	9	28	13	50.078
2007	9	28	14	52.680
2007	9	28	15	63.112

b) BARRA: **Pascuales**

No se ejecuta el procedimiento debido a que esta barra no tiene datos perdidos

6) EMPRESA ELÉCTRICA REGIONAL NORTE

BARRA: **Ibarra;** Posición: **Otavalo**

Los resultados obtenidos por ejecución de comandos para imputación simple existente en STATA, arrojan los siguientes datos:

CUADRO 4.58
RESULTADO DE APLICAR EL MÉTODO IMPUTACIÓN SIMPLE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN OTAVALO

ANIO	MES	DIA	HORA	mw_isim
2007	9	3	8	15.964
2007	9	3	10	19.727

7. EMPRESA ELÉCTRICA SANTO DOMINGO

BARRA: **Santo Domingo**

No se ejecuta el procedimiento debido a que esta barra no tiene datos perdidos

4.5.7 IMPUTACIÓN MÚLTIPLE

La imputación múltiple puede ser realizada en el programa STATA, en cuyos comandos se debe especificar el número de simulaciones requeridas, para el caso del análisis presente se especifican 50 en la opción m(#) y los datos imputados se guardan en 50 archivos independientes que son generados de acuerdo al nombre que el usuario requerirá. No se procede como los métodos

anteriores, es decir generación de 100 simulaciones ya que se requiere de un computador con mayores recursos que el actual (1 Mbytes de memoria). Nuevamente se ha procedido en como en los métodos anteriores, se describe para la primera empresa eléctrica los comandos que se utilizan para obtener los datos imputados. Para las demás empresas se presentan los resultados.

Adicionalmente, para combinar las imputaciones en un solo archivo de trabajo en el programa STATA se dispone de un comando “micombine” que se utiliza para mezclar archivos y generar estimadores combinados de los cincuenta modelos de regresión ajustados.

1. EMPRESA ELÉCTRICA AMBATO

a) BARRA: **Totoras;** Posición: **Montalvo**

Como se especificaron 50 simulaciones, los datos imputados se guardan en cincuenta archivos, impmmvismon1…….impmmvismon50.
Comando de STATA para calcular imputación múltiple de las variables requeridas: *ice mw_mon mw_amb mw_ban using imputmicemon, genmiss(imput) id(sort) m(100) cmd(regress mw_mon mw_amb mw_ban) boot (mw_mon)*

CUADRO 4.59
RESULTADO DEL PROGRAMA STATA AL EJECUTAR EL COMANDO PARA IMPUTACIÓN MÚLTIPLE

#missing values	Freq.	Percent	Cum.
0	2,919	99.69	99.69
1	9	0.31	100.00
Total	2,928	100.00	

Variable	Command	Prediction equation
mw_mon	regress mw_mon mw_amb mw_ban	mw_amb mw_ban
mw_amb	regress mw_mon mw_amb mw_ban	[No missing data in estimation sample]
mw_ban	regress mw_mon mw_amb mw_ban	[No missing data in estimation sample]

```
Imputing
[Only 1 variable to be imputed, therefore no cycling needed.]
1..2..3..4..5..6..7..8..9..10..11..12..13..14..15..16..17..18..19..20..21..22..
> 23..24..25..26..27..28..29..30..31..32..33..34..35..36..37..38..39..40..41..4
> 2..43..44..45..46..47..48..49..50..file imputmicemon.dta saved
```

Debido a que es necesario combinar archivos con las simulaciones generadas, se debe ejecutar el comando “micombine” en el nuevo archivo generado por el comando anterior, esto con el objetivo disponer de un vector combinado de parámetros estimados. Entonces los resultados son integrados en el archivo de trabajo base. A continuación se presenta como es la estructura del comando de STATA y sus despliegues de salida.

Comando de STATA: *micombine regress mw_mon mw_amb mw_ban, gen(monmice) impid(_j)*

CUADRO 4.60
RESULTADO DEL PROGRAMA STATA AL EJECUTAR EL COMANDO PARA COMBINAR ARCHIVOS DE SIMULACIONES

Multiple imputation parameter estimates (50 imputations)

mw_mon	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mw_amb	.1283209	.0129696	9.89	0.000	.1028905	.1537514
mw_ban	1.296189	.0191336	67.74	0.000	1.258673	1.333706
_cons	10.44097	.0998958	104.52	0.000	10.24509	10.63684

2928 observations.

A continuación se presenta los resultados obtenidos de la corrida efectuada por el programa estadístico:

CUADRO 4.61
RESULTADO DE APLICAR EL MÉTODO IMPUTACIÓN MÚLTIPLE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN MONTALVO

ANIO	MES	DIA	HORA	mw_imul
2007	9	3	8	23.566
2007	9	3	12	23.565
2007	9	5	12	24.216
2007	9	5	13	22.885
2007	9	5	14	23.303
2007	9	12	13	21.528
2007	9	12	14	22.225
2007	9	13	13	23.167
2007	9	13	16	24.571

2. CATEG-SD

a) BARRA: **Pascuales;** Posición: **Vergeles**

Los resultados obtenidos de las simulaciones ejecutadas por STATA son las siguientes:

CUADRO 4.62
RESULTADO DE APLICAR EL MÉTODO IMPUTACIÓN MÚLTIPLE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN VERGELES

ANIO	MES	DIA	HORA	mw_imul
2007	9	3	17	48.374
2007	9	3	18	43.878
2007	9	22	9	40.113

b) BARRA: **Pascuales;** Posición: **Cervecería**

Los resultados obtenidos por ejecución de comandos para imputación simple existente en STATA, arrojan los siguientes datos:

CUADRO 4.63
RESULTADO DE APLICAR EL MÉTODO IMPUTACIÓN MÚLTIPLE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN CERVECERÍA

ANIO	MES	DIA	HORA	mw_imul
2007	9	23	7	34.270
2007	9	23	8	33.243
2007	9	23	9	34.065
2007	9	23	10	35.918

b) BARRA: **Trinitaria;** Posición: **Padre Canals**

No se ejecuta el procedimiento debido a que esta barra no tiene datos perdidos

c) BARRA: **Trinitaria;** Posición: **Guasmo**

Los resultados obtenidos al ejecutar este comando en el programa estadístico son:

CUADRO 4.64
RESULTADO DE APLICAR EL MÉTODO IMPUTACIÓN MÚLTIPLE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN GUASMO

ANIO	MES	DIA	HORA	mw_imul
2007	9	3	3	26.581
2007	9	3	6	27.942
2007	9	3	10	26.363

2007	9	3	20	30.412
2007	9	5	9	25.834
2007	9	5	10	30.173
2007	9	5	20	28.008
2007	9	10	10	27.303
2007	9	10	11	27.926
2007	9	10	12	24.318
2007	9	10	13	31.244
2007	9	11	9	22.855
2007	9	11	14	21.713
2007	9	11	15	27.370
2007	9	21	10	27.461
2007	9	24	6	26.564
2007	9	24	7	23.695
2007	9	24	9	24.515
2007	9	25	10	25.019
2007	9	30	7	27.906
2007	9	30	8	26.442
2007	9	30	9	22.751
2007	9	30	10	22.738

d) BARRA: **Trinitaria;** Posición: **Pradera**

Los resultados obtenidos por ejecución de comandos para imputación múltiple existente en STATA, arrojan los siguientes datos:

(Debido al volumen de datos perdidos en esta posición, se presentan los resultados de un grupo de datos).

CUADRO 4.65

RESULTADO DE APLICAR EL MÉTODO IMPUTACIÓN MÚLTIPLE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN PRADERA

ANIO	MES	DIA	HORA	mw_imul
2007	9	1	3	27.722
2007	9	1	6	18.978
2007	9	1	7	25.133
2007	9	1	10	26.216
2007	9	1	11	25.249
2007	9	1	13	20.775
2007	9	1	18	17.512
2007	9	1	21	22.471
2007	9	1	22	20.759
2007	9	2	1	24.287
2007	9	2	2	20.588
2007	9	2	4	22.976
2007	9	2	7	22.157
2007	9	2	9	22.123
2007	9	2	12	21.808
2007	9	2	19	27.343
2007	9	2	21	18.332
2007	9	2	24	20.412
2007	9	3	4	21.529
2007	9	3	5	22.249
2007	9	4	7	19.234
2007	9	4	11	22.035
2007	9	4	12	24.446
2007	9	4	18	24.659
2007	9	4	21	21.634
2007	9	4	22	21.730

2007	9	5	1	22.045
2007	9	5	2	21.752
2007	9	5	4	25.377
2007	9	5	5	23.947

e) BARRA: **Policentro**

Los resultados obtenidos al ejecutar este comando en el programa estadístico son:

CUADRO 4.66
RESULTADO DE APLICAR EL MÉTODO IMPUTACIÓN MÚLTIPLE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN POLICENTRO

ANIO	MES	DIA	HORA	mw_imul
2007	9	3	18	100.662

3. EMPRESA ELÉCTRICA QUITO

BARRA: **Pomasqui**

Los resultados obtenidos por ejecución de comandos para imputación múltiple existente en STATA, arrojan los siguientes datos:

CUADRO 4.67
RESULTADO DE APLICAR EL MÉTODO IMPUTACIÓN MÚLTIPLE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN QUITO 1

ANIO	MES	DIA	HORA	mw_imul
2007	9	18	10	90.867
2007	9	18	13	90.349
2007	9	18	15	92.113
2007	9	18	17	84.363
2007	9	18	18	86.904
2007	9	20	17	86.215
2007	9	20	18	85.955

4. EMPRESA ELÉCTRICA COTOPAXI

BARRA: **Mulaló;** Posición: **Ambato**

Los resultados obtenidos por ejecución de comandos para imputación múltiple existente en STATA, arrojan los siguientes datos:

CUADRO 4.68
RESULTADO DE APLICAR EL MÉTODO IMPUTACIÓN MÚLTIPLE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN AMBATO

ANIO	MES	DIA	HORA	mw_imul
2007	9	4	17	5.685
2007	9	4	18	5.495
2007	9	12	3	4.830
2007	9	12	4	5.023
2007	9	24	11	6.917
2007	9	25	6	8.244
2007	9	27	24	7.389

5. EMPRESA ELÉCTRICA REGIONAL GUAYAS – LOS RÍOS

a) BARRA: **Dos Cerritos**

Los resultados obtenidos al ejecutar este comando en el programa estadístico son:

CUADRO 4.69
RESULTADO DE APLICAR EL MÉTODO IMPUTACIÓN MÚLTIPLE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN DOS CERRITOS

ANIO	MES	DIA	HORA	mw_imul
2007	9	1	2	42.169
2007	9	1	3	41.601
2007	9	1	4	41.179
2007	9	1	5	40.569
2007	9	1	6	40.071
2007	9	1	7	37.977
2007	9	23	8	38.053
2007	9	23	9	40.641
2007	9	23	10	40.652
2007	9	23	11	43.079
2007	9	23	12	43.850
2007	9	23	13	45.170
2007	9	23	14	46.084
2007	9	23	15	44.784
2007	9	24	6	37.541
2007	9	24	7	36.229
2007	9	24	8	39.239
2007	9	24	9	48.084
2007	9	24	10	51.793
2007	9	24	11	55.007

2007	9	24	12	50.539
2007	9	24	13	54.273
2007	9	24	14	56.789
2007	9	24	15	57.226
2007	9	24	16	56.864
2007	9	24	17	53.822
2007	9	24	18	53.901
2007	9	26	6	40.034
2007	9	26	7	39.885
2007	9	26	8	42.069
2007	9	26	9	49.338
2007	9	26	10	52.997
2007	9	26	11	54.880
2007	9	26	12	55.490
2007	9	26	13	54.380
2007	9	26	14	56.455
2007	9	26	15	57.929
2007	9	26	16	57.169
2007	9	27	7	37.595
2007	9	27	8	37.595
2007	9	27	9	49.946
2007	9	27	10	52.898
2007	9	27	11	54.829
2007	9	27	12	54.677
2007	9	27	13	55.489
2007	9	27	14	57.675
2007	9	27	15	58.898
2007	9	27	16	58.896
2007	9	28	7	39.476
2007	9	28	8	40.287
2007	9	28	9	49.339
2007	9	28	10	53.812

2007	9	28	11	53.762
2007	9	28	12	56.912
2007	9	28	13	56.353
2007	9	28	14	57.830
2007	9	28	15	59.099

b) BARRA: **Pascuales**

No se ejecuta el procedimiento debido a que esta barra no tiene datos perdidos

6. EMPRESA ELÉCTRICA REGIONAL NORTE

BARRA: **Ibarra;** Posición: **Otavalo**

Los resultados obtenidos por ejecución de comandos para imputación múltiple existente en STATA, arrojan los siguientes datos:

CUADRO 4.70
RESULTADO DE APLICAR EL MÉTODO IMPUTACIÓN MÚLTIPLE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN OTAVALO

ANIO	MES	DIA	HORA	mw_imul
2007	9	3	8	16.413
2007	9	3	10	17.184

7. EMPRESA ELÉCTRICA SANTO DOMINGO

BARRA: **Santo Domingo**

No se ejecuta el procedimiento debido a que esta barra no tiene datos perdidos

4.5.8 INTERPOLACION OPTIMA-SERIES DE TIEMPO

La interpolación óptima es posible realizar a través del programa STATA. Para el efecto, es necesario convertir la variable tiempo en una serie histórica y darle el formato del tipo de datos a ser analizados, para finalmente emplear el comando de interpolación de los datos de la variable en la que se encuentran los datos perdidos. Continuando con el proceso establecido en los métodos anteriores, se expondrá para la primera empresa eléctrica los comandos que se utilizan para obtener los datos imputados. Para las demás empresas se presentan los resultados.

1. EMPRESA ELÉCTRICA AMBATO

a) BARRA: **Totoras;** Posición: **Montalvo**

Se genera inicialmente la variable tiempo con su respectivo rezago, para lo cual se emplea el comando de STATA *gen newdia_t=d(1jun2007)+ dia_t-1.* Se debe considerar en la variable el inicio de la serie histórica, es decir, la fecha en la que se inicia la serie.

A continuación se da formato a la variable creada, de tal forma de convertir el número registrado en STATA a formato fecha en la expresión de tiempo que requiera el análisis; para este caso se emplea el mínimo formato disponible en STATA 9.0 (diario) esto es, *format newt %td.* Debido que es necesario declarar a la variable como datos históricos se emplea además el comando *tsset newt.* Se aclara que, en el programa disponible (STATA9.0) no cuenta con un formato de horas ni segundos por lo que se realiza la aproximación al formato diario.

Finalmente para encontrar los datos perdidos a través de interpolación, se emplea el comando *ipolate mw_mon newt, gen(internw_mon)*

Al ejecutar el comando, se generan datos dentro del mismo archivo con los valores estimados para los valores faltantes, a continuación se presentan los resultados:

CUADRO 4.71
RESULTADO DE APLICAR INTERPOLACIÓN A LA SERIE HISTÓRICA CORRESPONDIENTE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN MONTALVO

ANIO	MES	DIA	HORA	mw_series
2007	9	3	8	22.594
2007	9	3	12	22.656
2007	9	5	12	24.438
2007	9	5	13	24.500
2007	9	5	14	24.563
2007	9	12	13	25.979
2007	9	12	14	26.521
2007	9	13	13	23.688
2007	9	13	16	23.469

2. CATEG-SD

a) BARRA: **Pascuales;** Posición: **Vergeles**

Los resultados obtenidos al ejecutar este comando en el programa estadístico son:

CUADRO 4.72
RESULTADO DE APLICAR INTERPOLACIÓN A LA SERIE HISTÓRICA CORRESPONDIENTE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN VERGELES

ANIO	MES	DIA	HORA	mw_series
2007	9	3	17	51.598
2007	9	3	18	52.044
2007	9	22	9	37.766

b) BARRA: **Pascuales;** Posición: **Cervecería**

Se presentan los resultados arrojados por el programa STATA:

CUADRO 4.73
RESULTADO DE APLICAR INTERPOLACIÓN A LA SERIE HISTÓRICA CORRESPONDIENTE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN CERVECERÍA

ANIO	MES	DIA	HORA	mw_series
2007	9	23	7	27.956
2007	9	23	8	27.899
2007	9	23	9	27.842
2007	9	23	10	27.784

c) BARRA: **Trinitaria;** Posición: **Padre Canals**

No se ejecuta el procedimiento debido a que esta barra no tiene datos perdidos

d) BARRA: **Trinitaria;** Posición: **Guasmo**

Se presentan los resultados obtenidos por la corrida del programa estadístico:

CUADRO 4.74
RESULTADO DE APLICAR INTERPOLACIÓN A LA SERIE HISTÓRICA CORRESPONDIENTE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN GUASMO

ANIO	MES	DIA	HORA	mw_series
2007	9	3	3	19.313
2007	9	3	6	22.125
2007	9	3	10	24.863
2007	9	3	20	39.713
2007	9	5	9	27.525
2007	9	5	10	28.125

2007	9	5	20	40.913
2007	9	10	10	25.290
2007	9	10	11	25.680
2007	9	10	12	26.070
2007	9	10	13	26.460
2007	9	11	9	21.675
2007	9	11	14	21.675
2007	9	11	15	28.700
2007	9	21	10	29.138
2007	9	24	6	20.225
2007	9	24	7	21.100
2007	9	24	9	22.688
2007	9	25	10	26.363
2007	9	30	7	17.235
2007	9	30	8	17.670
2007	9	30	9	18.105
2007	9	30	10	18.540

e) BARRA: **Trinitaria;** Posición: **Pradera**

Por la cantidad de datos ausentes, se presenta un grupo de datos obtenidos por este método

CUADRO 4.75
RESULTADO DE APLICAR INTERPOLACIÓN A LA SERIE HISTÓRICA CORRESPONDIENTE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN PRADERA

ANIO	MES	DIA	HORA	mw_series
2007	9	1	3	23.213
2007	9	1	6	20.425
2007	9	1	7	19.500
2007	9	1	10	25.150
2007	9	1	11	25.700
2007	9	1	13	24.900
2007	9	1	18	30.638
2007	9	1	21	30.650
2007	9	1	22	29.725
2007	9	2	1	21.650
2007	9	2	2	20.425

2007	9	2	4	19.125
2007	9	2	7	23.888
2007	9	2	9	24.563
2007	9	2	12	23.775
2007	9	2	19	27.825
2007	9	2	21	29.888
2007	9	2	24	23.363
2007	9	3	4	22.950
2007	9	3	5	22.950
2007	9	4	7	24.450
2007	9	4	11	26.275
2007	9	4	12	26.225
2007	9	4	18	28.275
2007	9	4	21	32.683
2007	9	4	22	31.042
2007	9	5	1	14.525
2007	9	5	2	12.400
2007	9	5	4	10.775
2007	9	5	5	11.275
2007	9	5	8	12.600
2007	9	5	10	13.313
2007	9	5	13	14.400
2007	9	5	17	31.113
2007	9	5	19	34.575
2007	9	5	21	32.325

f) BARRA: **Policentro**

Los resultados obtenidos al ejecutar este comando en el programa estadístico son:

CUADRO 4.76
RESULTADO DE APLICAR INTERPOLACIÓN A LA SERIE HISTÓRICA CORRESPONDIENTE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN POLICENTRO

ANIO	MES	DIA	HORA	mw_series
2007	9	3	18	104.778

3. EMPRESA ELÉCTRICA QUITO

BARRA: **Pomasqui**

Los resultados obtenidos por ejecución de comandos para imputación simple existente en STATA, arrojan los siguientes datos:

CUADRO 4.77
RESULTADO DE APLICAR INTERPOLACIÓN A LA SERIE HISTÓRICA CORRESPONDIENTE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN QUITO 1

ANIO	MES	DIA	HORA	mw_series
2007	9	18	10	89.373
2007	9	18	13	88.900
2007	9	18	15	87.970
2007	9	18	17	89.735
2007	9	18	18	90.996
2007	9	20	17	90.996
2007	9	20	18	92.359

4. EMPRESA ELÉCTRICA COTOPAXI

BARRA: **Mulaló;** Posición: **Ambato**

Los resultados obtenidos al ejecutar este comando en el programa estadístico son:

CUADRO 4.78

RESULTADO DE APLICAR INTERPOLACIÓN A LA SERIE HISTÓRICA CORRESPONDIENTE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN AMBATO

ANIO	MES	DIA	HORA	mw_series
2007	9	4	17	8.305
2007	9	4	18	7.994
2007	9	12	3	6.618
2007	9	12	4	6.845
2007	9	24	11	7.862
2007	9	25	6	6.283
2007	9	27	24	6.067

5. EMPRESA ELÉCTRICA REGIONAL GUAYAS – LOS RÍOS

a) BARRA: **Dos Cerritos**

Los resultados obtenidos al ejecutar este comando en el programa estadístico son:

CUADRO 4.79

RESULTADO DE APLICAR INTERPOLACIÓN A LA SERIE HISTÓRICA CORRESPONDIENTE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN DOS CERRITOS

ANIO	MES	DIA	HORA	mw_series
2007	9	1	2	41.336
2007	9	1	3	40.671
2007	9	1	4	40.007
2007	9	1	5	39.343
2007	9	1	6	38.679
2007	9	1	7	38.014
2007	9	23	8	34.700
2007	9	23	9	35.800
2007	9	23	10	36.900
2007	9	23	11	38.000
2007	9	23	12	39.100
2007	9	23	13	40.200

2007	9	23	14	41.300
2007	9	23	15	42.400
2007	9	24	6	37.446
2007	9	24	7	39.493
2007	9	24	8	41.539
2007	9	24	9	43.586
2007	9	24	10	45.632
2007	9	24	11	47.679
2007	9	24	12	49.725
2007	9	24	13	51.771
2007	9	24	14	53.818
2007	9	24	15	55.864
2007	9	24	16	57.911
2007	9	24	17	59.957
2007	9	24	18	62.004
2007	9	26	6	40.425
2007	9	26	7	41.550
2007	9	26	8	42.675
2007	9	26	9	43.800
2007	9	26	10	44.925
2007	9	26	11	46.050
2007	9	26	12	47.175
2007	9	26	13	48.300
2007	9	26	14	49.425
2007	9	26	15	50.550
2007	9	26	16	51.675
2007	9	27	7	42.941
2007	9	27	8	44.032
2007	9	27	9	45.123
2007	9	27	10	46.214
2007	9	27	11	47.305
2007	9	27	12	48.395
2007	9	27	13	49.486
2007	9	27	14	50.577
2007	9	27	15	51.668
2007	9	27	16	52.759
2007	9	28	7	42.795
2007	9	28	8	44.340
2007	9	28	9	45.885
2007	9	28	10	47.430
2007	9	28	11	48.975
2007	9	28	12	50.520
2007	9	28	13	52.065
2007	9	28	14	53.610

2007	9	28	15	55.155

b) BARRA: **Pascuales**

No se ejecuta el procedimiento debido a que esta barra no tiene datos perdidos

6) EMPRESA ELÉCTRICA REGIONAL NORTE

BARRA: **Ibarra;** Posición: **Otavalo**

Los resultados obtenidos por ejecución de comandos para imputación simple existente en STATA, arrojan los siguientes datos:

CUADRO 4.80
RESULTADO DE APLICAR INTERPOLACIÓN A LA SERIE HISTÓRICA CORRESPONDIENTE A LA VARIABLE POTENCIA ACTIVA DE LA POSICIÓN OTAVALO

ANIO	MES	DIA	HORA	mw_series
2007	9	3	8	18.264
2007	9	3	10	18.736

7. EMPRESA ELÉCTRICA SANTO DOMINGO

BARRA: **Santo Domingo**

No se ejecuta el procedimiento debido a que esta barra no tiene datos perdidos

4.5.9 ANÁLISIS DE RESULTADOS

Los cinco métodos de imputación propuestos se aplicaron para sustituir a los datos perdidos en la variable potencia activa instantánea en las barras de carga del Sistema Nacional Interconectado; estos datos estimados se comparan con los datos reales que fueron registrados por el Área de Análisis de la Operación en los

procesos de validación de la información y almacenados en la base de datos del Sistema de Adquisición y Datos y Reportes (SADYR), base de datos paralela a la analizada de Histórico del EMS, para observar cuál es la mejor estimación.

A continuación se presentan los análisis de los errores presentados por los distintos métodos de imputación, para tres datos perdidos en la base de datos de histórico del RANGER:

1. EMPRESA ELÉCTRICA AMBATO

BARRA: **Totoras;** Posición: **Montalvo**

CUADRO 4.81
RESULTADO DE ERRORES PRESENTADOS EN A VARIABLE POTENCIA ACTIVA DE LA POSICIÓN MONTALVO

Método de Imputación	*POSICIÓN*					
	MONTALVO					
	VALOR1	*VALOR2*	*VALOR3*	*% ERROR 1*	*% ERROR 2*	*% ERROR 3*
Hot Deck	20.379	22.895	21.872	12.537	2.798	5.557
Hot Deck con Regresión	24.056	22.351	22.675	3.243	5.108	2.091
Regresión Condicionada	23.304	23.566	23.168	0.015	0.048	0.040
Imputación Simple	25.297	22.524	23.316	8.569	4.374	0.677
Imputación Múltiple	23.303	23.565	23.167	0.012	0.044	0.037
Series de Tiempo	24.563	22.656	23.688	5.418	3.813	2.283
Dato Original	23.300	23.554	23.159			

Se observa que el menor error en la estimación con relación al dato real se presenta en el método de Imputación Múltiple para los tres casos analizados. En el método de Imputación Múltiple los valores se sustituyeron de manera aleatoria y no se generaron sesgos en la asignación del valor imputado.

Adicionalmente se observa un menor error cuando los datos son imputados por Regresión Condicionada. Por tanto, cualquiera de los dos métodos podría ser utilizado para estimar la potencia activa de la posición Montalvo.

2. CATEG-SD

a) BARRA: **Pascuales;** Posición: **Cervecería**

CUADRO 4.82
POTENCIA ACTIVA DE LA POSICIÓN CERVECERÍA POR DISTINTOS MÉTODOS DE IMPUTACIÓN

Método de Imputación	*POSICIÓN*					
	CERVECERÍA					
	VALOR1	*VALOR2*	*VALOR3*	*% ERROR 1*	*% ERROR 2*	*% ERROR 3*
Hot Deck	35.049	34.106	38.291	2.333	2.584	12.415
Hot Deck con Regresión	29.745	28.553	30.474	13.154	14.118	10.535
Regresión Condicionada	33.998	33.085	33.815	0.737	0.487	0.726
Imputación Simple	40.061	31.183	44.130	16.967	6.207	29.555
Imputación Múltiple	34.270	33.243	34.065	0.058	0.011	0.008
Series de Tiempo	27.956	27.899	27.842	18.375	16.085	18.263
Dato Original	34.250	33.247	34.062			

Para esta posición también se observa que el menor error se da cuando se estiman los datos perdidos a través del método imputación múltiple y regresión condicionada.

No arrojan buenas estimaciones los métodos de series de tiempo, hot deck, regresión condicionada e imputación simple.

b) BARRA: **Pascuales;** Posición: **Vergeles**

CUADRO 4.83
POTENCIA ACTIVA DE LA POSICIÓN VERGELES POR DISTINTOS MÉTODOS DE IMPUTACIÓN

Método de Imputación	*POSICIÓN*					
	VERGELES					
	VALOR1	*VALOR2*	*VALOR3*	*% ERROR 1*	*% ERROR 2*	*% ERROR 3*
Hot Deck	39.753	38.178	41.360	17.794	12.962	3.127
Hot Deck con Regresión	45.854	47.553	34.269	5.176	8.413	14.554
Regresión Condicionada	48.368	43.871	40.107	0.022	0.018	0.004
Imputación Simple	44.166	40.648	39.265	8.666	7.330	2.097
Imputación Múltiple	48.374	43.878	40.113	0.036	0.034	0.017
Series de Tiempo	51.598	52.044	37.766	6.701	18.650	5.835
Dato Original	48.357	43.863	40.106			

Al igual que en los casos anteriores la estimación de los datos perdidos pueden ser realizadas por los métodos imputación múltiple o por regresión condicionada.

De la misma forma que en el caso anterior, no arrojan buenas estimaciones los métodos de hot deck, regresión condicionada e imputación simple.

c) BARRA: **Trinitaria;** Posición: **Guasmo**

CUADRO 4.84
POTENCIA ACTIVA DE LA POSICIÓN GUASMO POR DISTINTOS MÉTODOS DE IMPUTACIÓN

Método de Imputación	*POSICIÓN*					
	GUASMO					
	VALOR1	*VALOR2*	*VALOR3*	*% ERROR 1*	*% ERROR 2*	*% ERROR 3*
Hot Deck	21.660	26.868	22.474	19.911	13.279	6.220
Hot Deck con Regresión	23.979	28.635	25.520	11.337	7.577	20.617
Regresión Condicionada	25.979	31.861	22.316	3.940	2.835	5.474
Imputación Simple	27.070	31.000	21.157	0.092	0.055	0.006
Imputación Múltiple	26.363	27.370	26.442	2.522	11.660	24.974
Series de Tiempo	24.863	28.700	17.670	8.070	7.367	16.486
Dato Original	27.045	30.983	21.158			

Para el caso de la posición Guasmo se observa que el menor error en la imputación se obtiene cuando se aplica el método imputación simple. El método de imputación simple sustituye los valores perdidos en forma aleatoria, por lo que no se generaron sesgos en la asignación de los valores imputados.

La peor estimación la realiza los procedimientos de hot deck, hot deck con regresión e imputación múltiple.

d) BARRA: **Trinitaria;** Posición: **Pradera**

CUADRO 4.85
POTENCIA ACTIVA DE LA POSICIÓN PRADERA POR DISTINTOS MÉTODOS DE IMPUTACIÓN

Método de Imputación	*POSICIÓN*					
	PRADERA					
	VALOR1	*VALOR2*	*VALOR3*	*% ERROR 1*	*% ERROR 2*	*% ERROR 3*
Hot Deck	29.930	18.998	21.591	20.517	19.425	9.347
Hot Deck con Regresión	28.559	18.105	12.681	14.994	23.212	35.776
Regresión Condicionada	24.840	23.586	19.746	0.019	0.035	0.003
Imputación Simple	31.080	24.583	20.109	25.148	4.261	1.846
Imputación Múltiple	21.961	21.612	21.107	11.571	8.337	6.896
Series de Tiempo	32.663	13.313	28.450	31.518	43.538	44.087
Dato Original	24.835	23.578	19.745			

La potencia activa de la posición Pradera es la variable que mayores datos perdidos presentan, por lo tanto, al realizar las estimaciones con los distintos métodos de imputación el procedimiento que presenta menor error es a través de regresión condicionada, puesto que a través del comando de STATA los datos se organizan por el patrón de datos perdidos .

No realizan una buena estimación los métodos series de tiempo, imputación simple, hot deck y hot deck con regresión.

e) BARRA: **Policentro**

CUADRO 4.86

POTENCIA ACTIVA DE LA POSICIÓN POLICENTRO POR DISTINTOS MÉTODOS DE IMPUTACIÓN

Método de Imputación	*POSICIÓN*	
	POLICENTRO ATR	
	VALORES	*ERROR %*
Hot Deck	89.518	9.910
Hot Deck con Regresión	99.312	0.054
Regresión Condicionada	100.671	1.315
Imputación Simple	99.505	0.141
Imputación Múltiple	100.662	1.305
Series de Tiempo	104.778	5.447
Dato Original	99.365	

En esta posición existe una solo dato perdido el cual fue estimado y de los resultados, el método de imputación con menor error correspondió al procedimiento hot deck con regresión. La peor estimación la realizó hot deck.

3. EMPRESA ELÉCTRICA QUITO

BARRA: **Pomasqui**

CUADRO 4.87

POTENCIA ACTIVA DE LA POSICIÓN QUITO 1 POR DISTINTOS MÉTODOS DE IMPUTACIÓN

Método de Imputación	*POSICIÓN*					
	QUITO 1					
	VALOR1	*VALOR2*	*VALOR3*	*ERROR 1%*	*ERROR 2%*	*ERROR 3%*
Hot Deck	57.568	60.374	74.719	15.823	17.505	1.617
Hot Deck con Regresión	68.902	73.185	73.210	0.750	0.000	0.435
Regresión Condicionada	92.356	91.098	86.413	35.044	24.476	17.521
Imputación Simple	86.901	104.112	87.934	27.068	42.258	19.589
Imputación Múltiple	92.113	90.867	86.215	34.690	24.160	17.251
Series de Tiempo	87.970	89.373	90.996	28.631	22.120	23.753
Dato Original	68.389	73.185	73.530			

Para el caso de la posición Quito 1, se observa que el menor error en la imputación se obtiene cuando se aplica el método imputación hot deck con regresión, lo demás métodos presenta error muy altos superiores al 10%.

4. EMPRESA ELÉCTRICA PROVINCIAL COTOPAXI

BARRA: **Mulaló;** Posición: **Ambato**

CUADRO 4.88
POTENCIA ACTIVA DE LA POSICIÓN AMBATO POR DISTINTOS MÉTODOS DE IMPUTACIÓN

Método de Imputación	*POSICIÓN*					
	AMBATO					
	VALOR1	*VALOR2*	*VALOR3*	*ERROR 1%*	*ERROR 2%*	*ERROR 3%*
Hot Deck	4.603	6.795	7.635	33.457	92.553	-16.222
Hot Deck con Regresión	7.078	5.128	7.528	2.311	45.316	-14.586
Regresión Condicionada	7.383	5.675	6.908	6.716	60.808	-5.152
Imputación Simple	6.920	3.527	6.572	0.029	0.061	-0.037
Imputación Múltiple	7.389	5.685	6.917	6.813	61.112	-5.286
Series de Tiempo	6.067	8.305	7.862	12.300	135.349	-19.672
Dato Original	6.918	3.529	6.570			

Se observa que para la posición Ambato el menor error en la imputación se obtiene cuando se aplica el método imputación simple. El método de imputación simple sustituye los valores perdidos en forma aleatoria, por lo que no se generaron sesgos en la asignación de los valores imputados.

Por otro, lado se observa que el método de hot deck arroja la peor estimación del dato.

5. EMPRESA ELÉCTRICA REGIONAL GUAYAS – LOS RÍOS

BARRA: **Dos Cerritos**

CUADRO 4.89

POTENCIA ACTIVA DE LA POSICIÓN DOS CERRITOS POR DISTINTOS MÉTODOS DE IMPUTACIÓN

Método de Imputación	POSICIÓN DOS CERRITOS ATR					
	VALOR1	VALOR2	VALOR3	ERROR 1%	ERROR 2%	ERROR 3%
Hot Deck	41.257	47.754	49.722	1.276	3.916	1.764
Hot Deck con Regresión	39.129	40.273	52.566	6.367	18.969	7.586
Regresión Condicionada	41.050	45.139	59.202	1.771	9.178	21.167
Imputación Simple	41.839	49.645	48.848	0.118	0.111	0.025
Imputación Múltiple	41.179	45.170	58.896	1.461	9.115	20.540
Series de Tiempo	40.007	40.200	52.759	4.266	19.115	7.980
Dato Original	41.790	49.700	48.860			

En esta posición el método de imputación con menor error correspondió al del procedimiento imputación simple.

Los métodos que introducen errores altos son regresión condicionada, imputación múltiple, series de tiempo y hot deck con regresión.

6. EMPRESA ELÉCTRICA REGIONAL NORTE

BARRA: Ibarra; Posición: **Otavalo**

CUADRO 4.90
POTENCIA ACTIVA DE LA POSICIÓN OTAVALO POR DISTINTOS MÉTODOS DE IMPUTACIÓN

Método de Imputación	POSICIÓN OTAVALO			
	VALOR1	VALOR2	ERROR 1%	ERROR 2%
Hot Deck	20.516	19.058	19.519	16.162
Hot Deck con Regresión	22.691	18.601	32.195	13.378
Regresión Condicionada	17.180	16.408	0.087	0.010
Imputación Simple	19.727	15.964	14.923	2.694
Imputación Múltiple	17.184	16.413	0.109	0.045
Series de Tiempo	18.736	18.264	9.152	11.322
Dato Orginal	17.165	16.406		

Se observa que el menor error en la estimación con relación al dato real se presenta en el método de Imputación Múltiple y Regresión Condicionada para los tres casos analizados. En el método de Imputación Múltiple los valores se sustituyeron de manera aleatoria y no se generaron sesgos en la asignación del valor imputado.

Por tanto, cualquiera de los dos métodos podrían ser empleados para estimar la potencia activa de la posición Otavalo.

CAPÍTULO 5

CONCLUSIONES Y RECOMENDACIONES

5.1. CONCLUSIONES

1. En el presente trabajo se han analizado los fundamentos teóricos de distintos procedimientos de imputación dando cuenta de sus bondades y limitaciones. Asimismo, los métodos se han aplicado con un propósito muy preciso, observar su impacto en los datos de potencia activa instantánea de las barras de carga del Sistema Nacional Interconectado del Ecuador.

2. Todos los métodos de imputación estudiados tienen limitaciones y su correcta aplicación depende de la manera en que se comporten los datos faltantes. En la medida en que la falta de respuesta no muestre un patrón aleatorio, la eficacia de todas las metodologías se debilita, aún en los procedimientos de imputación múltiple.

3. Al culminar este trabajo, es posible aseverar que es factible aplicar los métodos de imputación estadística a los datos de los registros de potencia activa instantánea provenientes del EMS, de las barras de carga del Sistema Nacional Interconectado del Ecuador y que, como fruto del análisis, se puede determinar que los métodos a aplicarse en las barra de carga corresponde a:

EMPRESA	BARRA	POSICIÓN	METODO IMPUTACIÓN
E.E.Ambato	Totoras	Montalvo	Imputación múltiple Regresión Condicionada
CATEG-SD	Pascuales	Cervecería	Imputación múltiple Regresión Condicionada
	Trinitaria	Guasmo	Imputación Simple
		Pradera	Regresión Condicionada
	Policentro	Policentro ATR	Hot Deck con Regresión
E.E.Quito	Pomasqui	Quito 1	Hot Deck con Regresión
E.E.Cotopaxi	Mulaló	Ambato	Imputación Simple
EMELGUR	Dos Cerritos	Dos Cerritos ATR	Imputación Simple
EMELNORTE	Ibarra 69 kV	Otavalo	Imputación múltiple Regresión Condicionada

4. La teoría recomienda que cuando las tasas de no respuesta son muy elevadas en las variables de interés (25% o más), se debe mantener la idea de desechar la base de datos. En el presente estudio, se observa que la mayor tasa de pérdida de datos corresponde al 7,82% asociado a los datos de potencia activa instantánea de la posición Pradera de la Subestación Trinitaria, siendo factible por tanto aplicar los métodos de imputación estadística.

5. El proceso de imputación debe preservar el valor real, es decir el valor imputado debe ser lo más cercano posible al valor real. En el presente trabajo, en las imputaciones se logró un error inferior al 1%, y en la mayoría de los casos incluso un error cercano a 0%, cumpliendo con el criterio descrito.

6. Luego de la imputación, y para cada uno de los métodos con menor error en cada barra, se verifica que los nuevos conjuntos de datos, preservaron las distribuciones reales así como sus estadísticos media y varianza, por tanto, no se ha afectado el sesgo ni se ha subestimado la varianza de los datos.

7. Del análisis de los datos se observa que el método Hot Deck presentó la estimación menos favorable de la variable potencia activa instantánea para todas las barras, debido a que se verificó sesgo en los estadísticos.

8. Se observa que los datos perdidos de la variable potencia activa instantánea de las barras de carga del Sistema Nacional Interconectado del Ecuador pueden ser estimadas en un porcentaje aproximado al 66% mediante procedimientos de imputación simple y múltiple, ya que estos métodos reemplazan los datos perdidos en forma estocástica y esta es una característica del comportamiento de la variable analizada. Además, el empleo de estos métodos para la variable de estudio garantiza que no se introduzcan sesgos de asignación en los datos, ni se subestime o sobreestime la varianza.

9. La variable con mayor tasa de no respuesta (7,82%) corresponde a la potencia activa de la posición Pradera de la subestación Trinitaria, para lo cual el método de estimación que mejor la caracteriza es el de regresión condicionada, por cuanto este método garantiza variabilidad en los valores imputados y contribuye a reducir el sesgo en la varianza .

10. El objetivo de la imputación es obtener una base de datos completa y consistente para que posteriormente estos datos puedan ser analizados mediante técnicas estadísticas estándares.

11. La imputación múltiple permitió hacer uso eficiente de los datos, obtener estimadores no sesgados y reflejar adecuadamente la incertidumbre que la no respuesta parcial introduce en la estimación de parámetros.

12. No se sugiere aplicar el método de imputación por regresión cuando el análisis secundario de datos involucra técnicas de análisis de covarianza o de correlación, ya que sobreestima la asociación entre las

variables y sus modelos de regresión múltiple pueden sobredimensionar el valor del coeficiente de determinación R^2.

13. El método de sustitución de datos perdidos a través de la media tiene implicaciones negativas en la varianza del estimador e introduce distorsiones en el patrón de correlación de los datos.

14. No existe el mejor método de imputación. Cada situación es diferente y la elección del procedimiento de sustitución de datos depende de la variable de estudio, del porcentaje de datos faltantes y del uso que se hará de la información imputada.

15. El análisis exploratorio y la consistencia de información, dan la pauta para elegir el método que genera mejores estimaciones. Se debe considerar que los aspectos que funcionaron en este análisis, no necesariamente generan buenos resultados para otras investigaciones de este tipo.

16. Los métodos de imputación estadística pueden ser aplicados a las demás variables eléctricas (voltajes, frecuencia, etc) del Sistema Eléctrico del Ecuador, pero deberán seguir el procedimiento señalado en este trabajo de tesis.

5.2 RECOMENDACIONES

1. El análisis realizado en este trabajo, determina que los métodos de imputación son factibles de aplicarse a la variable potencia activa instantánea proveniente del histórico del EMS; por lo tanto, se recomienda que este procedimiento establecido en esta tesis sea implantado en el proceso de Validación de Información del Área de Análisis de la Operación de la Dirección de Operaciones del Centro de Control de Energía “CENACE”.

2. En este trabajo de tesis se ha enfrentado el problema de los datos faltantes; sin embargo, se recomienda una segunda fase que debería enfocarse al análisis de valores extremos (outlier), pues se observa que cuando por mantenimientos o fallas, estos valores llegan a ser cero o con valores mayores a los promedios de las potencias activas instantáneas de las barras de carga del Sistema Nacional Interconectado, influyen en la forma de la distribución de esta variable.

3. Contar con una base de datos completa del Histórico del RANGE es difícil, pues se encuentran presentes muchos elementos que pueden fallar en la cadena de transmitir el dato al Centro de Control de Energía CENACE; por lo tanto, se recomienda que se realice un seguimiento más estrecho a los datos, de tal forma que se puedan ejecutar acciones inmediatas de reposición del equipo fallado y con ello se consiga minimizar el tiempo en que el dato no es transmitido a la Corporación.

4. El análisis de datos realizados a través de interpolación por series temporales se enfrentó a limitaciones con relación al formato, pues la mínima unidad de análisis del programa STATA 9.0 era diaria. Esta limitación implicó tomar esta unidad en los análisis de los datos, con lo cual se estima que se introdujeron sesgos en la estimación del dato y por este motivo en los análisis comparativos de los métodos resultó no ser una de las mejores estimaciones; por lo tanto, se sugiere considerar esta opción con la nueva versión del programa estadístico que considere resoluciones horarias.

5. Al realizar el análisis de los datos por comandos de STATA para imputación múltiple con simulación de 100 variables se encontraron limitaciones en el equipo de computación, por este motivo se recomienda que para ejecutar las simulaciones mayores a 50 imputaciones se deba emplear un equipo o probar con un equipo de mayor capacidad en memoria RAM (mayor a 1 Giga).

REFERENCIAS BIBLIOGRÁFICAS

1. ABB Inc. (2003). *Proyecto de Modernización de los Sectores Eléctrico, de Telecomunicaciones y Servicios Rurales*. Quito.

2. Badler, Clara; Clara Alsina; Cristina Puigsubirá y María Vitelleschi. (2004). "Tratamiento de Bases de Datos con Información Faltante según Análisis de las Pérdidas con SPSS". Universidad Nacional de Rosario. Novenas Jornadas "Investigaciones en la Facultad" de Ciencias Económicas y Estadísticas".

3. Carlin, Bradley P., y Thomas A. Louis. (2000). Bayes and Empirical Bayes Methods for Data Analysis. . Segunda Edición. Florida. Chapman & Hall./CRC.

4. Capa, Holger. (2007). Modelación de Series Temporales. Primera Edición. Quito. Escuela Politécnica Nacional.

5. Chatterjee, Samprit, y Bertram Price. (1991). Regression Analysis by Example. Segunda Edición. New Yersey. John Wiley & Sons, Inc.

6. Congdon, Peter. (2001). Bayesian Statistical Modelling. Primera Edición. Wilthshire. John Wiley & Sons, Ltd.

7. Crawley, Michael J. (2005). Satistics An Introduction using R. Primera Edición. Wilthshire. John Wiley & Sons, Ltd.

8. Dobson, Annette J. (1990). An Introduction to Generalized Linear Models. Primera Edición. Londres. Chapman & Hall.

9. Ezzati-Rice, Trena M.; Mansour Fahimi, David Judkins y Meena Khare. (1992). “Serial Imputation of Nhanes III With Mixed Regression and Hot-Deck Techniques”. Westat, Inc. MD 20850.

10. Gentle, James E. (2003). Random Number Generation and Monte Carlo Methods. Segunda Edición. Virginia. Springer Science + Bussines Media, Inc.

11. Gill, Jeff. (2002). Bayesian Methods A Social and Behavioral Sciences Approach. Primera Edición. USA. Chapman & Hall./CRC.

12. Gómez García J., y J. Palarea Albaladejo. (2003). “Inferencia Basada en imputación múltiple en problemas con información incompleta”. IX Conferencia Española de Biometría.

13. Gujarati, Damodar N. (2003). Econometría. Cuarta Edición. México. McGraw-Hill/Irwin.

14. Horton, Nicholas J., y Stuart R. Lipsitz. (2001). “Multiple Imputation in Practice: Comparision of Software Packages for Regression Models With Missing Variables”. American Statistical Association. Vol 55 No. 3.

15. Horton, Nicholas J., y Ken P. Kleinman. (2007). “Much Ado About Nothing: A Comparision of Missing data Methods and Software to Fit Incomplete Data Regression Models”. American Statistical Association. Vol 61 No. 1.

16. ISTAC / ULL: Proyecto “Depuración de encuestas estadísticas”. (2004). “Técnicas de Edición e Imputación de Datos Estadísticos”. http: //webpages.ull.es/users/Isaac.

17. Juárez Alonso Carlos Alberto. (2004). Fusión de Datos: Imputación y Validación. Tesis Doctoral. Universidad Politécnica de Cataluña. Barcelona España.

18. Kedem, Benjamin y Fokianos Konstantinos. (2002). Regression Models for Time Series Analyisis. Primera Edición. New Yersey. John Wiley & Sons, Inc.

19. Kenett, Ron S., y Shelemyahu Zacks. (2000). Estadística Industrial Moderna. Primera Edición. México. International Thomson Editores.

20. Little, Roderick J.A., y Donald B. Rubin. (2002). Statistical Analysis With Missing Data. Segunda Edición. New Yersey. John Wiley & Sons, Inc.

21. Longford, Nicholas T. (2005). Missing Data and Small – Area Estimation. Segunda Edición. New York. Springer Science + Bussines Media, Inc.

22. Morgan, Byron J.T. (1984). Elements of Simulation. Primera Edición. Cambridge. Chapman & Hall.

23. Puerta Goicochea, Aitor. (2002). “Imputación Basada en Árboles de Clasificación”. EUSTAT.

24. Rodríguez, Gioconda, y Juan Vallecilla. (2008). “Adquisición de Datos en el Sistema de Manejo de Energía, Network Manager”. Revista Técnica “energía” Edición No. 4.

25. Rubin, Donald B. (2004). Multiple Imputation for Nonresponse in Surveys. Segunda Edición. New Yersey. John Wiley & Sons, Inc.

26. StataCorp LP. (2005). Stata Documentation Version 9 – Data Management. New York.

27. StataCorp LP. (2005). Stata Documentation Version 9 – Graphics. New York.

28. StataCorp LP. (2005). Stata Documentation Version 9 – Getting Started With Stata for Windows. New York.

29. StataCorp LP. (2005). Stata Documentation Version 9 – Quick Reference and Index. New York.

30. StataCorp LP. (2005). Stata Documentation Version 9 – User´s Guide. New York.

31. Morales Manchego, Argemiro. (2007). Entrenamiento Especializado STATA 1 y 2. Colombia. SOFTWARE shop Inc.

32. Tsiatis, Anastasios A. (2006). Semiparametric Theory and Missing Data. Raleigh. Springer Science + Bussines Media, Inc.

33. Tusell Palmer, Fernando. (2005). “Multiple imputation of time series: an application to the construction of historical price indexes”. Universidad del País Vasco, Departamento de Economía Aplicada.

34. Useche, Lelly y Dulce Mesa. (2006). “Una Introducción a la Imputación de Valores Perdidos”. Universidad Central de Venezuela. Terra Nueva Etapa. XXII, número 031.

35. Von BHippel, Paul T. (2004). Biases in SPSS 12.0 Missing Value Analysis. American Statistical Association. Vol 58 No. 2.

36. Wayman, Jeffrey C. (2003). “Multiple Imputation For Missing Data: What Is It And How Can I Use It”. Annual Meeting of the American Educational Research Association. Chicago IL.

37. Yang C. Yuan. “Multiple Imputation for Missing Data: Concepts and New Development”. SAS Institute Inc.

38. Peña Daniel, Tiao George, Tsay Ruey. (2001) “A course in Time Series Analysis”. John Wiley&Sons. Canadá

ANEXOS

ANEXO No. 1
VALORES DE m PARA QUE EL NIVEL DE D_m SEA EXACTO

Niveles para muestras grandes (en%) de Dm con distribución de referencia Fk,v como una función de los niveles nominales ∪

∪número de componentes a ser probados; m número de imputaciones propias

γ_0 fracción de información perdida. Precisión de resultados =5000 simulaciones

$\rho_0 = \gamma_0 / 1 - \gamma_0$

			$\alpha = 1\%$				$\alpha = 5\%$				$\alpha = 10\%$			
k	m	γ_0 =	0.1	0.2	0.3	0.5	0.1	0.2	0.3	0.5	0.1	0.2	0.3	0.5
2	2		1.0	1.2	1.6	2.5	4.9	5.3	5.9	7.5	9.9	10.3	11.0	12.9
	3		1.0	1.0	1.0	1.3	4.9	4.9	5.0	5.5	9.9	9.8	10.0	10.9
	5		1.0	1.0	1.1	1.2	5.0	5.0	5.1	5.6	10.0	10.0	10.2	10.9
	10		1.0	1.0	1.1	1.2	5.0	5.1	5.3	5.7	10.1	10.2	10.4	11.0
	25		1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.0	10.0	9.9	9.9	10.0
	50		1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.0	10.0	9.9	9.9	10.0
	100		1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.0	10.0	10.0	10.0	10.1
3	2		1.0	1.1	1.3	1.7	5.1	5.3	5.6	6.3	10.3	10.6	11.1	12.0
	3		1.0	1.0	1.0	1.0	5.1	5.2	5.3	5.7	10.2	10.5	10.9	12.3
	5		1.0	1.0	1.1	1.3	5.0	5.2	5.4	6.2	10.1	10.3	10.8	12.2
	10		1.0	1.0	1.1	1.2	5.0	5.2	5.3	5.9	10.1	10.3	10.6	11.6
	25		1.0	1.0	1.1	1.2	5.0	5.1	5.2	5.6	10.1	10.2	10.4	10.9
	50		1.0	1.0	1.0	1.0	5.0	5.0	5.0	5.1	10.0	10.0	10.0	10.2
	100		1.0	1.0	1.0	1.0	5.0	5.0	5.1	5.1	10.0	10.0	10.1	10.2
5	2		0.9	0.8	0.8	0.9	5.1	4.8	4.5	4.0	10.5	10.4	10.1	9.2
	3		1.0	1.0	1.0	0.9	5.2	5.5	5.7	6.1	10.5	11.3	12.1	14.4
	5		1.1	1.1	1.2	1.4	5.2	5.6	6.1	7.7	10.4	11.1	12.2	15.4
	10		1.0	1.1	1.2	1.5	5.1	5.3	5.6	6.9	10.1	10.4	11.1	13.1
	25		1.0	1.0	1.1	1.3	5.0	5.2	5.3	6.0	10.1	10.3	10.6	11.5
	50		1.0	1.0	1.0	1.1	5.0	5.1	5.1	5.4	10.0	10.1	10.2	10.7
	100		1.0	1.0	1.0	1.1	5.0	5.0	5.1	5.2	10.0	10.1	10.1	10.4
10	2		0.8	0.5	0.3	0.1	5.1	4.0	2.9	1.5	10.8	10.1	8.5	5.4
	3		1.1	0.9	0.6	0.3	5.6	5.9	5.7	4.9	11.3	12.7	13.8	16.2
	5		1.1	1.2	1.3	1.4	5.4	6.3	7.4	11.0	10.7	12.4	14.8	22.7
	10		1.1	1.2	1.4	2.2	5.2	5.8	6.8	10.3	10.4	11.4	13.1	19.0
	25		1.0	1.1	1.2	1.6	5.0	5.2	5.6	7.1	10.0	10.4	11.0	13.4
	50		1.0	1.0	1.1	1.3	5.0	5.1	5.4	6.1	10.0	10.2	10.6	11.8
	100		1.0	1.0	1.1	1.2	5.0	5.2	5.3	5.8	10.1	10.2	10.5	11.3

ANEXO No. 2
ESTADISTICAS DESCRIPTIVAS E HISTROGRAMA DE LA VARIABLE CON VALORES PEDIDOS

BASE DE DATOS CON INCLUSION DE DATOS PERDIDOS DEL MES DE SEPTIEMBRE AÑO 2007

E.E.AMBATO -TOTORAS

anio ANIO

---

type: numeric (int)

range: [2007,2007] units: 1
unique values: 1 missing .: 0/2928

tabulation: Freq. Value
2928 2007

---

mes MES

---

type: numeric (byte)

range: [6,9] units: 1
unique values: 4 missing .: 0/2928

tabulation: Freq. Value
720 6
744 7
744 8
720 9

---

dia DIA

---

type: numeric (byte)

range: [1,31] units: 1
unique values: 31 missing .: 0/2928

mean: 15.7541
std. dev: 8.80846

percentiles: 10% 25% 50% 75% 90%
4 8 16 23 28

---

hora HORA

---

type: numeric (byte)

range: [1,24] units: 1
unique values: 24 missing .: 0/2928

mean: 12.5
std. dev: 6.92337

percentiles: 10% 25% 50% 75% 90%
3 6.5 12.5 18.5 22

---

n_dia_sem N_DIA_SEM

---

type: numeric (byte)

range: [1,7] units: 1
unique values: 7 missing .: 0/2928

```
      tabulation:  Freq.  Value
                    432  1
                    408  2
                    408  3
                    408  4
                    408  5
                    432  6
                    432  7

-------------------------------------------------------------------------------------------------------------
feriado                                                                                        FERIADO
-------------------------------------------------------------------------------------------------------------

            type:  numeric (byte)

           range:  [0,1]                        units:  1
      unique values:  2                         missing .:  0/2928

      tabulation:  Freq.  Value
                  2904  0
                    24  1

-------------------------------------------------------------------------------------------------------------
empresa                                                                                        EMPRESA
-------------------------------------------------------------------------------------------------------------

            type:  string (str6)

      unique values:  1                         missing "":  0/2928

      tabulation:  Freq.  Value
                  2928  "AMBATO"

-------------------------------------------------------------------------------------------------------------
mw_amb                                                                                          MW_AMB
-------------------------------------------------------------------------------------------------------------

            type:  numeric (float)

           range:  [0,20.713188]                units:  1.000e-08
      unique values:  2872                      missing .:  0/2928

            mean:   11.6311
       std. dev:   3.65892

      percentiles:        10%       25%       50%       75%       90%
                     7.1584   8.15641   11.8359   14.5682   16.0253

-------------------------------------------------------------------------------------------------------------
mw_mon                                                                                          MW_MON
-------------------------------------------------------------------------------------------------------------

            type:  numeric (float)

           range:  [0,35.4375]                  units:  1.000e-07
      unique values:  2288                      missing .:  9/2928

            mean:   22.8286
       std. dev:   3.89771

      percentiles:        10%       25%       50%       75%       90%
                     19.461   20.3125   22.1099        24   29.1875

-------------------------------------------------------------------------------------------------------------
mw_ban                                                                                          MW_BAN
-------------------------------------------------------------------------------------------------------------

            type:  numeric (float)

           range:  [0,15.9375]                  units:  1.000e-07
      unique values:  2201                      missing .:  0/2928

            mean:   8.40609
```

```
        std. dev:   2.47526

     percentiles:       10%       25%       50%       75%       90%
                          6    6.6875   7.99907   8.84917    12.978
```

CATEG-PASCUALES

```
anio                                                              ANIO
-------------------------------------------------------------------------------------------------------------------

            type:  numeric (int)

           range:  [2007,2007]          units:  1
     unique values:  1                  missing .:  0/2928

       tabulation:  Freq.  Value
                    2928  2007

-------------------------------------------------------------------------------------------------------------------
mes                                                               MES
-------------------------------------------------------------------------------------------------------------------

            type:  numeric (byte)

           range:  [6,9]                units:  1
     unique values:  4                  missing .:  0/2928

       tabulation:  Freq.  Value
                    720  6
                    744  7
                    744  8
                    720  9

-------------------------------------------------------------------------------------------------------------------
dia                                                               DIA
-------------------------------------------------------------------------------------------------------------------

            type:  numeric (byte)

           range:  [1,31]               units:  1
     unique values:  31                 missing .:  0/2928

            mean:   15.7541
       std. dev:   8.80846

     percentiles:       10%       25%       50%       75%       90%
                          4         8        16        23        28

-------------------------------------------------------------------------------------------------------------------
hora                                                              HORA
-------------------------------------------------------------------------------------------------------------------

            type:  numeric (byte)

           range:  [1,24]               units:  1
     unique values:  24                 missing .:  0/2928

            mean:      12.5
       std. dev:   6.92337

     percentiles:       10%       25%       50%       75%       90%
                          3       6.5      12.5      18.5        22

-------------------------------------------------------------------------------------------------------------------
n_dia_sem                                                         N_DIA_SEM
-------------------------------------------------------------------------------------------------------------------

            type:  numeric (byte)

           range:  [1,7]                units:  1
```

unique values: 7 missing .: 0/2928

tabulation: Freq. Value
432 1
408 2
408 3
408 4
408 5
432 6
432 7

---

feriado FERIADO

---

type: numeric (byte)

range: [0,1] units: 1
unique values: 2 missing .: 0/2928

tabulation: Freq. Value
2880 0
48 1

---

empresa EMPRESA

---

type: string (str8)

unique values: 1 missing "": 0/2928

tabulation: Freq. Value
2928 "CATEG-SD"

---

mw_cer MW_CER

---

type: numeric (float)

range: [23.711281,60.321564] units: 1.000e-06
unique values: 2342 missing .: 4/2928

mean: 39.143
std. dev: 6.742

percentiles: 10% 25% 50% 75% 90%
30.7864 34.36 38.5813 42.7162 49.3643

---

mvar_cer MVAR_CER

---

type: numeric (float)

range: [0,14.80427] units: 1.000e-08
unique values: 2211 missing .: 0/2928

mean: 7.18545
std. dev: 2.77487

percentiles: 10% 25% 50% 75% 90%
3.55053 5.06733 7.00925 9.27417 10.9951

---

mw_ver MW_VER

---

type: numeric (float)

range: [20.058313,56.027462] units: 1.000e-06
unique values: 2367 missing .: 3/2928

```
            mean:   40.0583
       std. dev:   7.95784

      percentiles:       10%       25%       50%       75%       90%
                    28.5158   34.7582   39.4843   47.3198   50.5161
```

---

mvar_ver MVAR_VER

```
            type:  numeric (float)

           range:  [3.34635,17.209801]         units:  1.000e-07
    unique values:  2211                   missing .:  0/2928

            mean:   10.8137
       std. dev:   2.65604

      percentiles:       10%       25%       50%       75%       90%
                    6.84333   8.98734   11.1754   12.7975   14.1278
```

CORRELACIONES

	mvar_ver	mw_ver	mvar_cer	mw_cer	n_dia_~m	hora	dia
mvar_ver	1.0000						
mw_ver	0.7958	1.0000					
mvar_cer	0.6158	0.7508	1.0000				
mw_cer	0.3718	0.7107	0.7859	1.0000			
n_dia_sem	0.2786	0.2161	0.2281	0.2035	1.0000		
hora	0.1304	0.4348	0.3537	0.5342	-0.0004	1.0000	
dia	0.0752	0.0463	0.0143	-0.0375	0.0389	0.0015	1.0000

CATEG-POLICENTRO

anio ANIO

```
            type:  numeric (int)

           range:  [2007,2007]                 units:  1
    unique values:  1                      missing .:  0/2928

        tabulation:  Freq.  Value
                     2928  2007
```

---

mes MES

```
            type:  numeric (byte)

           range:  [6,9]                       units:  1
    unique values:  4                      missing .:  0/2928

        tabulation:  Freq.  Value
                      720  6
                      744  7
                      744  8
                      720  9
```

---

dia DIA

```
            type:  numeric (byte)

           range:  [1,31]                      units:  1
    unique values:  31                     missing .:  0/2928

            mean:   15.7541
```

```
        std. dev:   8.80846

       percentiles:       10%       25%       50%       75%       90%
                          4         8         16        23        28

-------------------------------------------------------------------------------------------------------------
hora                                                                                          HORA
-------------------------------------------------------------------------------------------------------------

            type:  numeric (byte)

           range:  [1,24]                       units:  1
     unique values:  24                         missing .:  0/2928

            mean:     12.5
       std. dev:   6.92337

       percentiles:       10%       25%       50%       75%       90%
                          3         6.5       12.5      18.5      22

-------------------------------------------------------------------------------------------------------------
n_dia_sem                                                                                     N_DIA_SEM
-------------------------------------------------------------------------------------------------------------

            type:  numeric (byte)

           range:  [1,7]                        units:  1
     unique values:  7                          missing .:  0/2928

       tabulation:  Freq.  Value
                    432  1
                    408  2
                    408  3
                    408  4
                    408  5
                    432  6
                    432  7

-------------------------------------------------------------------------------------------------------------
feriado                                                                                       FERIADO
-------------------------------------------------------------------------------------------------------------

            type:  numeric (byte)

           range:  [0,1]                        units:  1
     unique values:  2                          missing .:  0/2928

       tabulation:  Freq.  Value
                   2880  0
                     48  1

-------------------------------------------------------------------------------------------------------------
empresa                                                                                       EMPRESA
-------------------------------------------------------------------------------------------------------------

            type:  string (str8)

     unique values:  1                          missing "":  0/2928

       tabulation:  Freq.  Value
                   2928  "CATEG-SD"

-------------------------------------------------------------------------------------------------------------
mw                                                                                            MW
-------------------------------------------------------------------------------------------------------------

            type:  numeric (float)

           range:  [0,112.09945]                units:  1.000e-06
     unique values:  2380                       missing .:  1/2928

            mean:   77.0066
       std. dev:   18.5257
```

```
      percentiles:       10%       25%       50%       75%       90%
                      53.7272   58.6019   77.8184   93.8214   100.656

-------------------------------------------------------------------------------------------------------------
mvar                                                                                                  MVAR
-------------------------------------------------------------------------------------------------------------

             type:  numeric (float)

            range:  [0,29.636]                   units:  1.000e-07
      unique values:  2209                       missing .:  0/2928

             mean:   16.3721
          std. dev:   4.9941

      percentiles:       10%       25%       50%       75%       90%
                      10.1336   12.2936   15.9898   20.3453   23.3153

CORRELACIONES
                     |      mw     mvar n_dia_~m     hora      dia      mes
          -----------+---------------------------------------------------
                mw |   1.0000
              mvar |   0.8923   1.0000
        n_dia_sem |   0.0607   0.0402   1.0000
              hora |   0.6353   0.5170   0.0003   1.0000
               dia |   0.0449   0.0880   0.0380   0.0004   1.0000
              mes |  -0.0682  -0.1724  -0.0105  -0.0004   0.0007   1.0000
```

CATEG-TRINITARIA

```
anio                                                                                                  ANIO
-------------------------------------------------------------------------------------------------------------

             type:  numeric (int)

            range:  [2007,2007]                  units:  1
      unique values:  1                          missing .:  0/2928

        tabulation:  Freq.  Value
                     2928  2007

-------------------------------------------------------------------------------------------------------------
mes                                                                                                    MES
-------------------------------------------------------------------------------------------------------------

             type:  numeric (byte)

            range:  [6,9]                        units:  1
      unique values:  4                          missing .:  0/2928

        tabulation:  Freq.  Value
                     720  6
                     744  7
                     744  8
                     720  9

-------------------------------------------------------------------------------------------------------------
dia                                                                                                    DIA
-------------------------------------------------------------------------------------------------------------

             type:  numeric (byte)

            range:  [1,31]                       units:  1
      unique values:  31                         missing .:  0/2928

             mean:   15.7541
          std. dev:   8.80846

      percentiles:       10%       25%       50%       75%       90%
                           4         8        16        23        28

-------------------------------------------------------------------------------------------------------------
hora                                                                                                  HORA
```

---

type: numeric (byte)

range: [1,24] units: 1
unique values: 24 missing .: 0/2928

mean: 12.5
std. dev: 6.92337

percentiles: 10% 25% 50% 75% 90%
3 6.5 12.5 18.5 22

---

n_dia_sem N_DIA_SEM

---

type: numeric (byte)

range: [1,7] units: 1
unique values: 7 missing .: 0/2928

tabulation: Freq. Value
432 1
408 2
408 3
408 4
408 5
432 6
432 7

---

feriado FERIADO

---

type: numeric (byte)

range: [0,1] units: 1
unique values: 2 missing .: 0/2928

tabulation: Freq. Value
2880 0
48 1

---

empresa EMPRESA

---

type: string (str8)

unique values: 1 missing "": 0/2928

tabulation: Freq. Value
2928 "CATEG-SD"

---

subestacion SUBESTACION

---

type: string (str10)

unique values: 1 missing "": 0/2928

tabulation: Freq. Value
2928 "TRINITARIA"

---

pent PENT

---

type: string (str6)

unique values: 1 missing "": 0/2928

tabulation: Freq. Value
2928 "GUASMO"

---
mw_gua MW_GUA
---

type: numeric (float)

range: [0,42.225002] units: 1.000e-06
unique values: 2257 missing .: 23/2928

mean: 26.7099
std. dev: 5.45979

percentiles: 10% 25% 50% 75% 90%
21.0996 22.8554 25.376 28.8719 36.1284

---
mvar_gua MVAR_GUA
---

type: numeric (float)

range: [1.7506982,12.6] units: 1.000e-07
unique values: 2218 missing .: 0/2928

mean: 6.85277
std. dev: 1.77211

percentiles: 10% 25% 50% 75% 90%
4.63951 5.5732 6.78473 8.09683 9.26248

---
mw_pca MW_PCA
---

type: numeric (float)

range: [10.204885,47.925003] units: 1.000e-06
unique values: 1929 missing .: 1/2928

mean: 29.7192
std. dev: 7.09169

percentiles: 10% 25% 50% 75% 90%
21.45 24.075 29.0722 32.775 41.8058

---
mvar_pca MVAR_PCA
---

type: numeric (float)

range: [1.4499913,16.903774] units: 1.000e-07
unique values: 2102 missing .: 0/2928

mean: 10.1605
std. dev: 2.34705

percentiles: 10% 25% 50% 75% 90%
7.04611 8.37462 10.275 11.775 13.05

---
mw_pra MW_PRA
---

type: numeric (float)

range: [0,36.900002] units: 1.000e-07
unique values: 2287 missing .: 229/2928

mean: 22.2684
std. dev: 5.29945

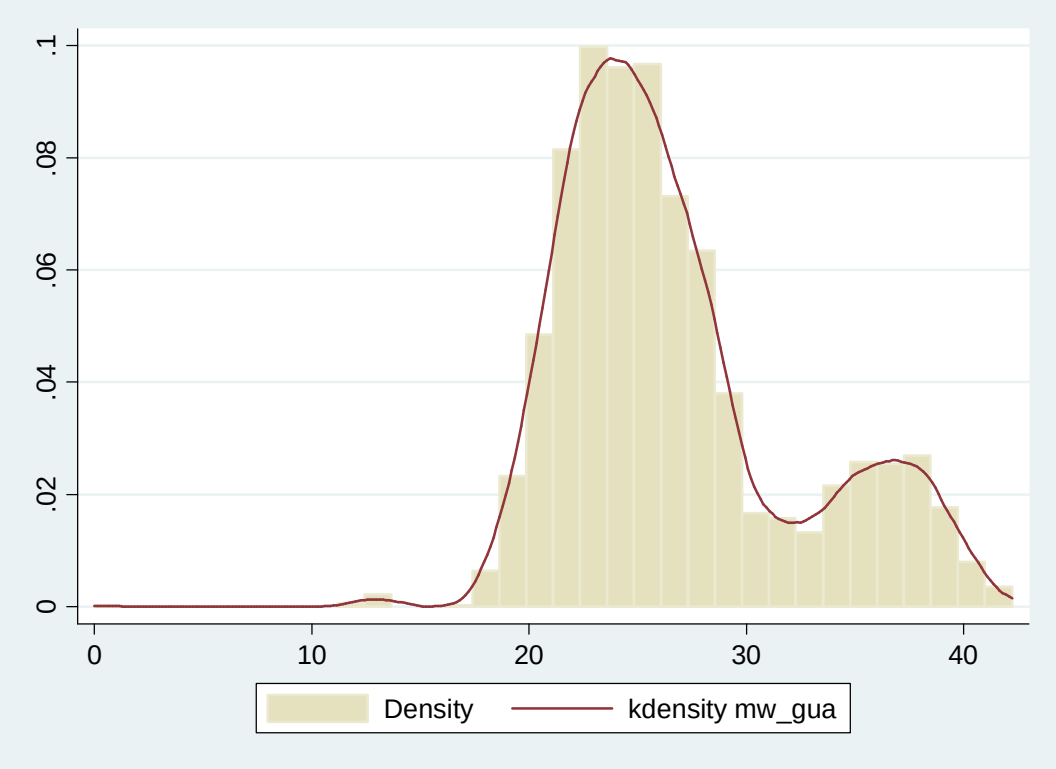
```
      percentiles:       10%       25%       50%       75%       90%
                       14.85   19.1212   22.2172   25.5781    29.475

-------------------------------------------------------------------------------------------------------
mvar_pra                                                                                MVAR_PRA
-------------------------------------------------------------------------------------------------------

            type:  numeric (float)

           range:  [-2.7,18.237612]          units:  1.000e-09
     unique values:  2246                    missing .:  0/2928

            mean:    5.2249
         std. dev:   3.50386

      percentiles:       10%       25%       50%       75%       90%
                          .3   2.99927   5.35852   7.00197      8.85
```

CORRELACIONES

	mvar_pca	mw_pca	mvar_gua	mw_gua	n_dia_~m	hora	dia	mes
mvar_pca	1.0000							
mw_pca	0.6891	1.0000						
mvar_gua	0.1961	0.3923	1.0000					
mw_gua	0.4985	0.9151	0.5943	1.0000				
n_dia_sem	0.2011	0.0591	0.0341	0.0883	1.0000			
hora	0.5358	0.7738	0.3463	0.7120	-0.0025	1.0000		
dia	-0.0049	0.0193	0.1781	0.0296	0.0391	0.0015	1.0000	
mes	-0.2350	-0.0599	0.2234	-0.0129	-0.0019	0.0033	0.0020	1.0000

COMANDO STATA:

twoway histogram mw_gua, color(*.5) || kdensity mw_gua



COMANDO STATA:

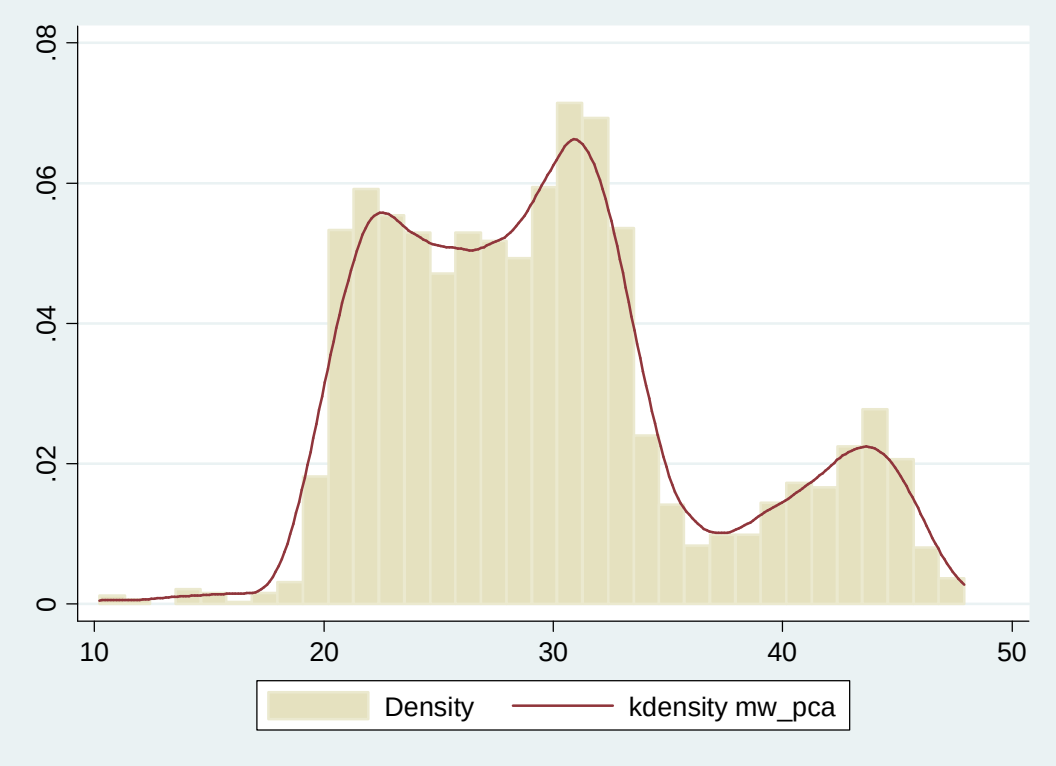
summarize mw_gua,detail

MW_GUA

Percentiles		Smallest		
1%	18.56952	0		
5%	20.06362	12.03433		
10%	21.09963	12.075	Obs	2905
25%	22.85539	12.6	Sum of Wgt.	2905
50%	25.37597		Mean	26.7099
		Largest	Std. Dev.	5.459794
75%	28.87185	41.4		
90%	36.12843	41.625	Variance	29.80935
95%	37.97672	41.85	Skewness	.8058512
99%	39.975	42.225	Kurtosis	3.177933

COMANDO STATA:

twoway histogram mw_pca, color(*.5) || kdensity mw_pca



COMANDO STATA:

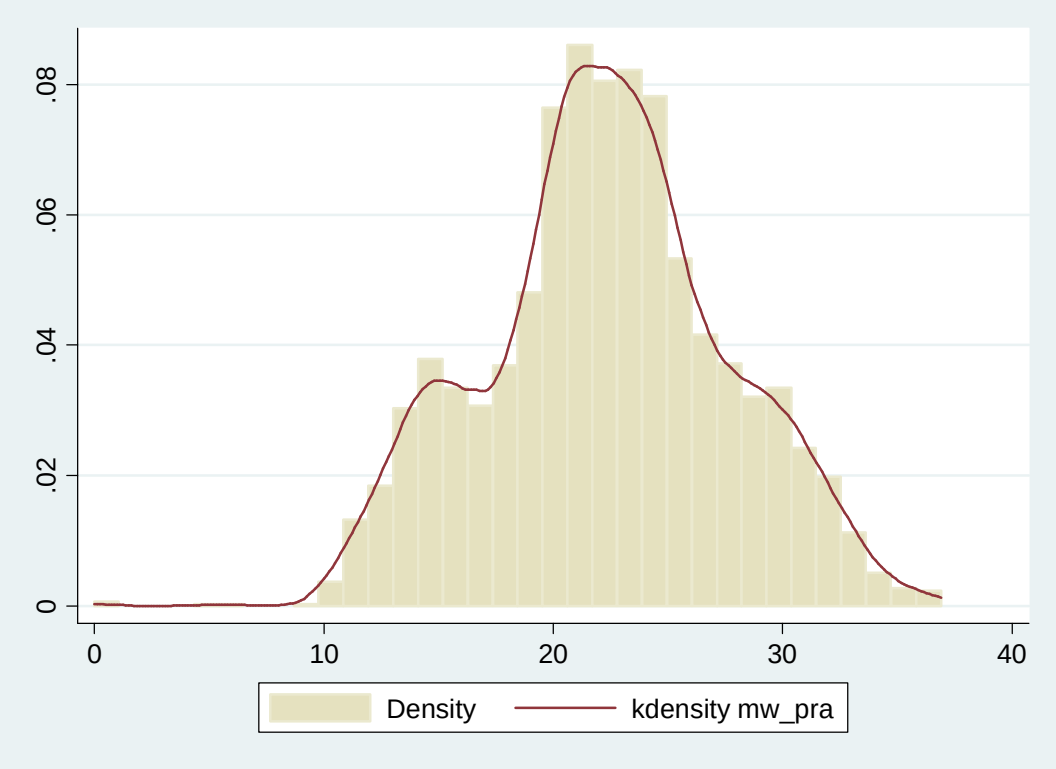
summarize mw_pca ,detail

MW_PCA

Percentiles		Smallest		
1%	18.525	10.20488		
5%	20.56407	10.23503		
10%	21.45	10.26499	Obs	2927
25%	24.075	11.12923	Sum of Wgt.	2927
50%	29.07219		Mean	29.71915
		Largest	Std. Dev.	7.091689
75%	32.775	47.39772		
90%	41.80583	47.475	Variance	50.29205
95%	44.04406	47.59046	Skewness	.6169428
99%	45.975	47.925	Kurtosis	2.771432

COMANDO STATA:

twoway histogram mw_pra, color(*.5) || kdensity mw_pra



COMANDO STATA:

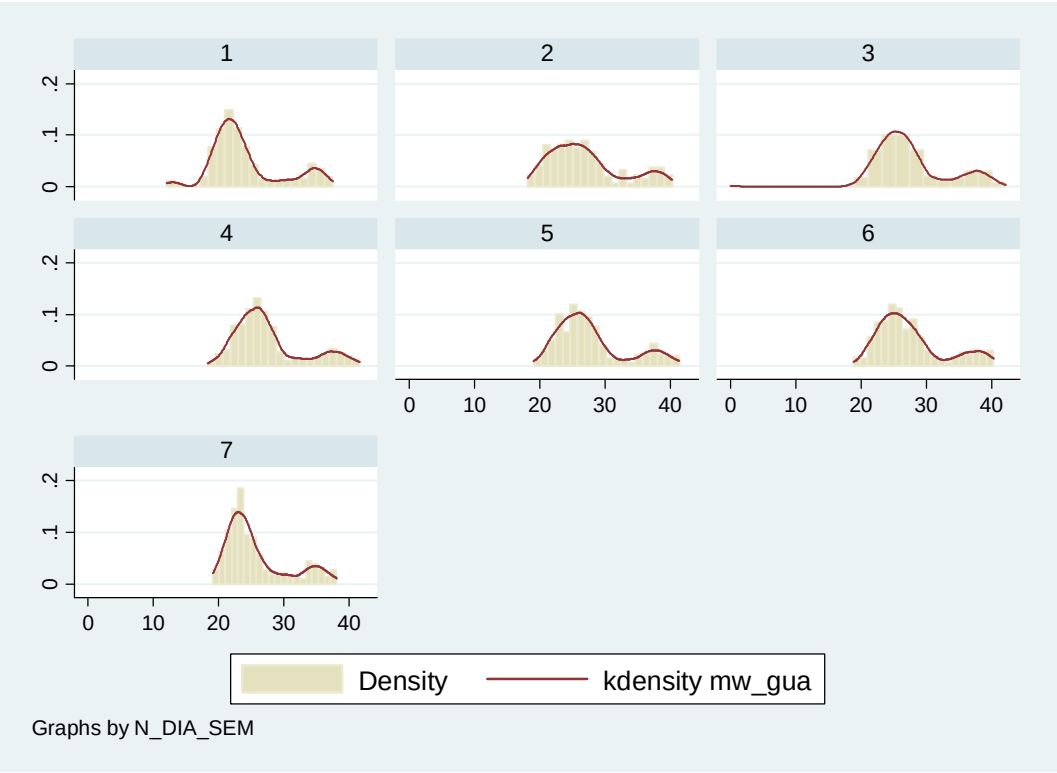
summarize mw_pra ,detail

MW_PRA

Percentiles		Smallest		
1%	11.175	0		
5%	13.35	0		
10%	14.85	4.824888	Obs	2699
25%	19.12121	6.345084	Sum of Wgt.	2699
50%	22.21718		Mean	22.26837
		Largest	Std. Dev.	5.299446
75%	25.57807	36.15		
90%	29.475	36.225	Variance	28.08413
95%	31.275	36.45	Skewness	-.0266059
99%	33.97919	36.9	Kurtosis	2.825087

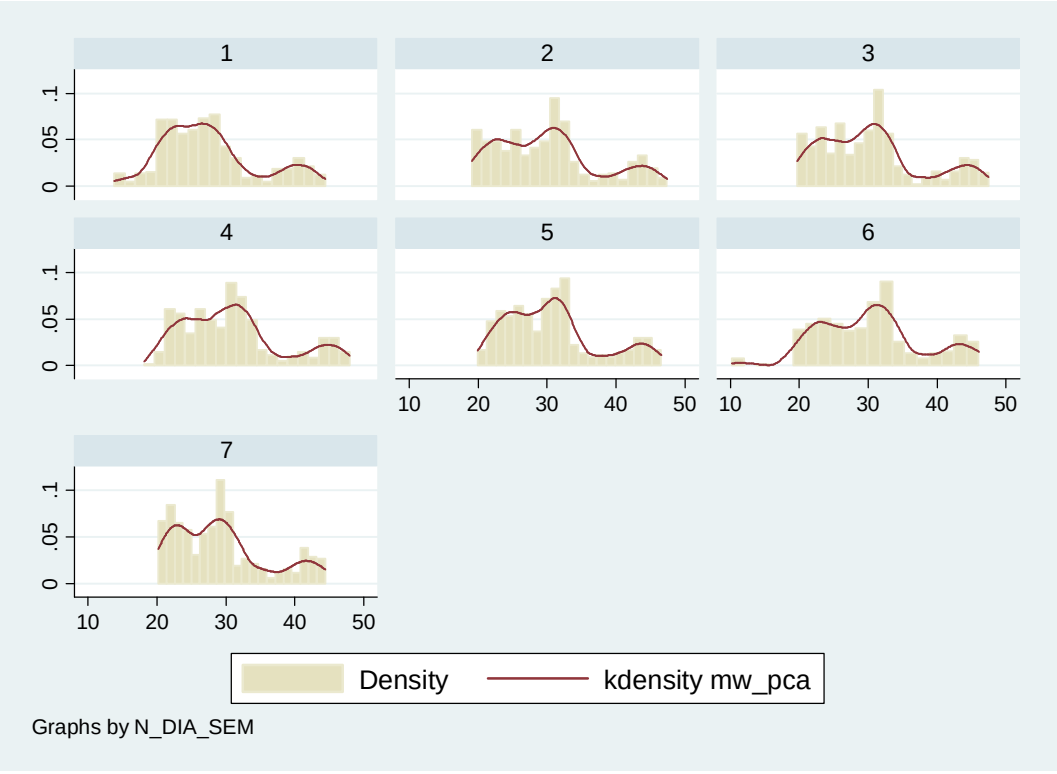
COMANDO STATA:

twoway histogram mw_gua, color(*.5) || kdensity mw_gua||, by (n_dia_sem)



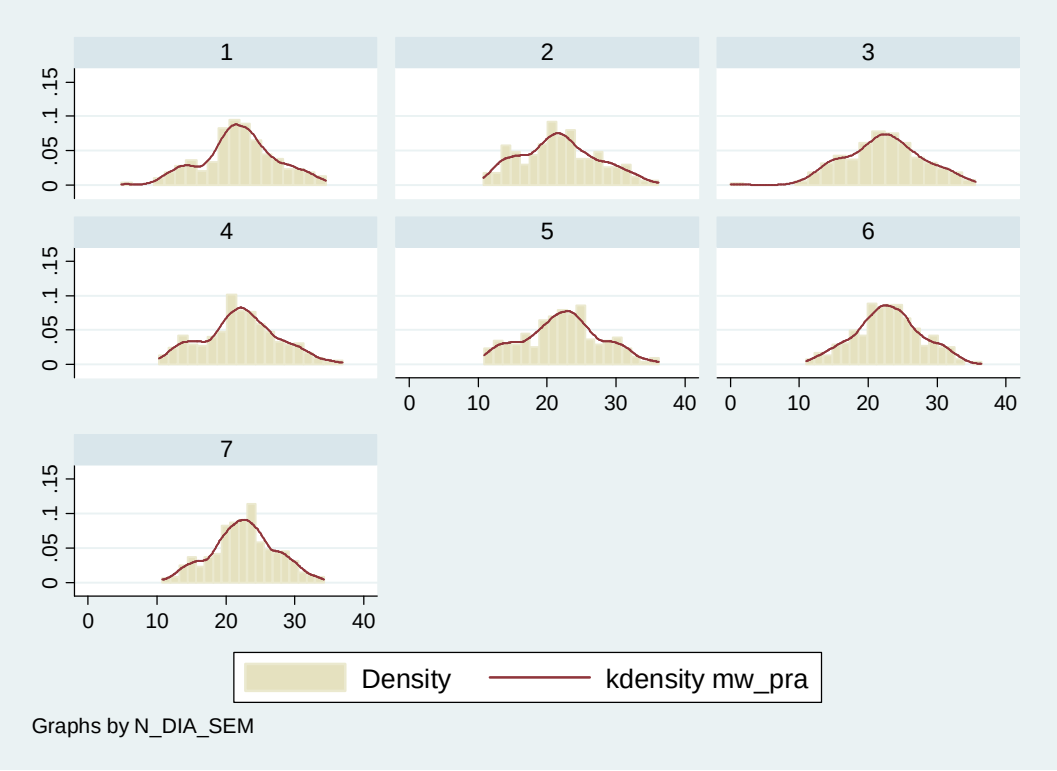
COMANDO STATA:

twoway histogram mw_pca, color(*.5) || kdensity mw_pca||, by (n_dia_sem)



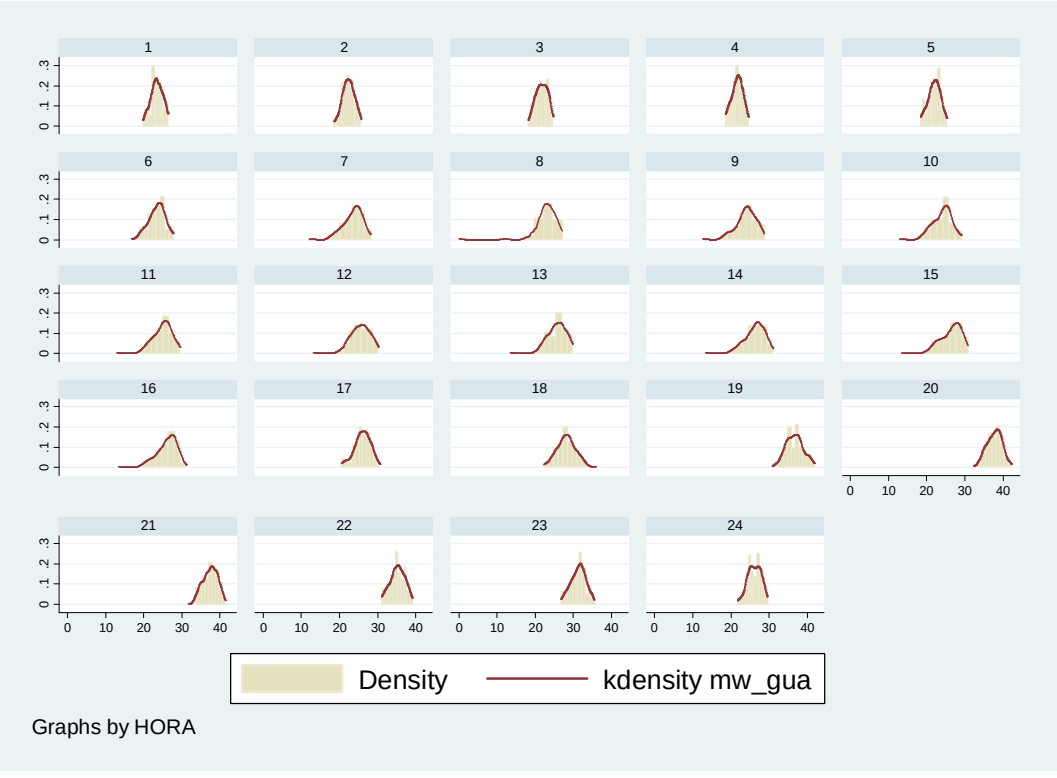
COMANDO STATA:

twoway histogram mw_pra, color(*.5) || kdensity mw_pra||, by (n_dia_sem)



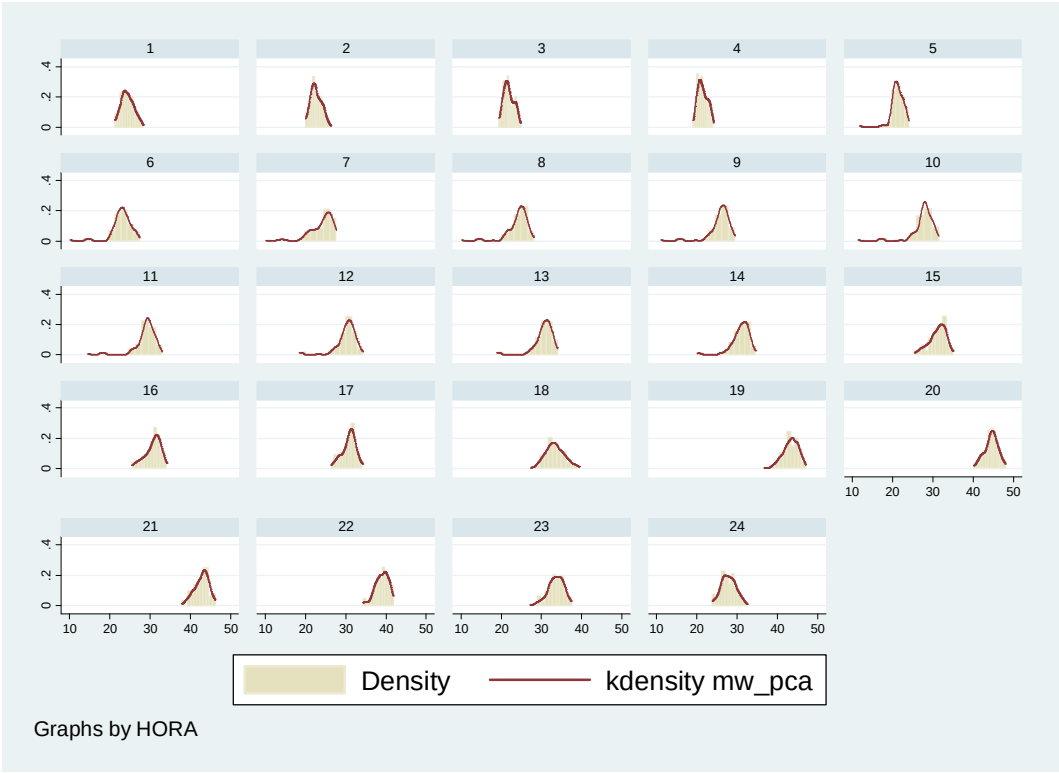
COMANDO STATA:

twoway histogram mw_gua, color(*.5) || kdensity mw_gua||, by (hora)



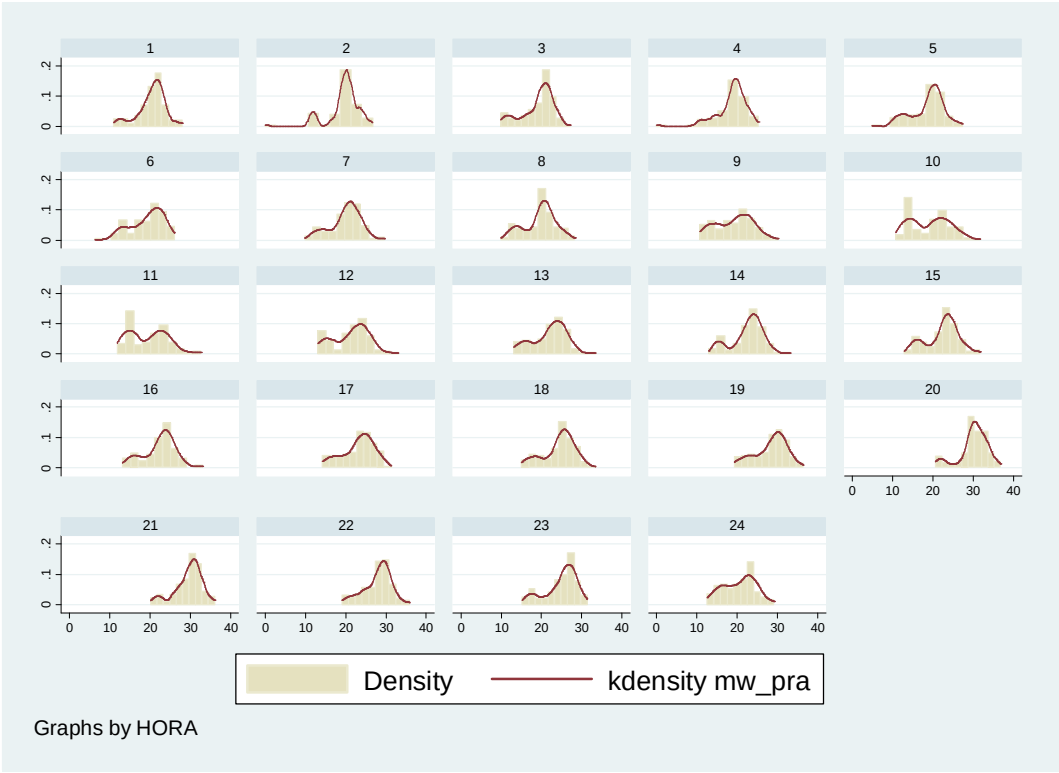
COMANDO STATA:

twoway histogram mw_pca, color(*.5) || kdensity mw_pca||, by (hora)



COMANDO STATA:

twoway histogram mw_pra, color(*.5) || kdensity mw_pra||, by (hora)



E.E. QUITO- POMASQUI

anio ANIO

---

type: numeric (int)

range: [2007,2007] units: 1
unique values: 1 missing .: 0/2928

tabulation: Freq. Value
2928 2007

---

mes MES

---

type: numeric (byte)

range: [6,9] units: 1
unique values: 4 missing .: 0/2928

tabulation: Freq. Value
720 6
744 7
744 8
720 9

---

dia DIA

---

type: numeric (byte)

range: [1,31] units: 1
unique values: 31 missing .: 0/2928

mean: 15.7541
std. dev: 8.80846

percentiles: 10% 25% 50% 75% 90%
4 8 16 23 28

---

hora HORA

---

type: numeric (byte)

range: [1,24] units: 1
unique values: 24 missing .: 0/2928

mean: 12.5
std. dev: 6.92337

percentiles: 10% 25% 50% 75% 90%
3 6.5 12.5 18.5 22

---

n_dia_sem N_DIA_SEM

---

type: numeric (byte)

range: [1,7] units: 1
unique values: 7 missing .: 0/2928

tabulation: Freq. Value
432 1
408 2
408 3
408 4
408 5
432 6
432 7

---

feriado FERIADO

---

type: numeric (byte)

range: [0,1] units: 1
unique values: 2 missing .: 0/2928

tabulation: Freq. Value
2904 0
24 1

---

empresa EMPRESA

---

type: string (str7)

unique values: 1 missing "": 0/2928

tabulation: Freq. Value
2928 "EEQUITO"

---

subestacion SUBESTACION

---

type: string (str8)

unique values: 1 missing "": 0/2928

tabulation: Freq. Value
2928 "POMASQUI"

---

pent PENT

---

type: string (str7)

unique values: 1 missing "": 0/2928

tabulation: Freq. Value
2928 "QUITO 1"

warning: variable has embedded blanks

---

mw_qui1 MW_QUI1

---

type: numeric (float)

range: [14.618927,111.156] units: 1.000e-07
unique values: 2710 missing .: 7/2928

mean: 64.6284
std. dev: 15.3323

percentiles: 10% 25% 50% 75% 90%
42.8716 54.2663 64.7405 75.9301 85.5434

---

mvar_qui1 MVAR_QUI1

---

type: numeric (float)

range: [6.6566253,44.041115] units: 1.000e-07
unique values: 2543 missing .: 0/2928

mean: 21.4269
std. dev: 4.58282

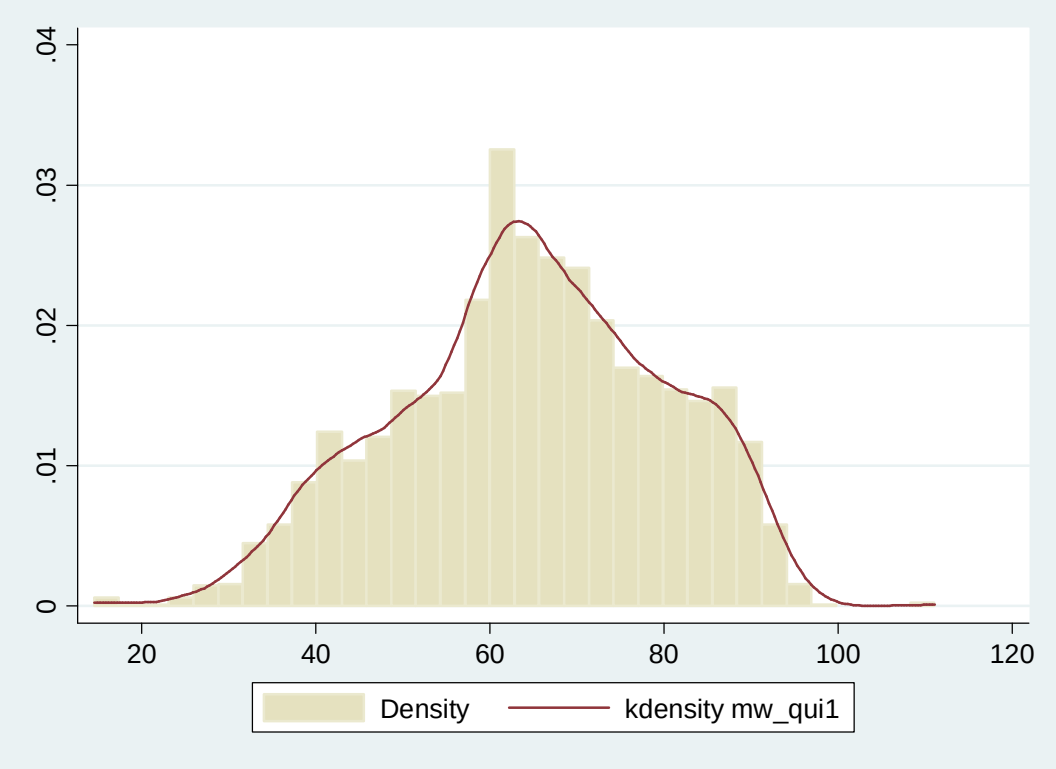
```
      percentiles:       10%       25%       50%       75%       90%
                      15.7402   18.3951   21.3829   24.4508    27.131

-------------------------------------------------------------------------------------------------------
mw_qui2                                                                                MW_QUI2
-------------------------------------------------------------------------------------------------------

            type:  numeric (float)

           range:  [-3.7997334,53.547203]       units:  1.000e-08
     unique values:  2625                       missing .:  7/2928

            mean:   30.1137
       std. dev:   9.75709

      percentiles:       10%       25%       50%       75%       90%
                      17.4687   23.9265   30.7016   37.2275   43.0384

-------------------------------------------------------------------------------------------------------
mvar_qui2                                                                              MVAR_QUI2
-------------------------------------------------------------------------------------------------------

            type:  numeric (float)

           range:  [-.16072673,28.534807]       units:  1.000e-08
     unique values:  2510                       missing .:  0/2928

            mean:   12.7285
       std. dev:   3.93521

      percentiles:       10%       25%       50%       75%       90%
                      7.99671   10.2036   12.5192   15.0153   17.7007
```

CORRELACIONES

	mw_qui1	mvar_q~1	mw_qui2	mvar_q~2	mes	dia	hora	n_dia_~m
mw_qui1	1.0000							
mvar_qui1	0.1164	1.0000						
mw_qui2	0.9044	-0.1076	1.0000					
mvar_qui2	0.2902	0.6528	0.1661	1.0000				
mes	0.2779	0.3027	0.3370	0.3635	1.0000			
dia	0.0699	0.0242	0.0491	0.0083	-0.0010	1.0000		
hora	0.3657	0.1834	0.2596	0.4513	-0.0014	-0.0004	1.0000	
n_dia_sem	0.1157	0.1941	0.0589	0.0629	-0.0103	0.0385	0.0000	1.0000

COMANDO STATA:

twoway histogram mw_qui1, color(*.5) || kdensity mw_qui1



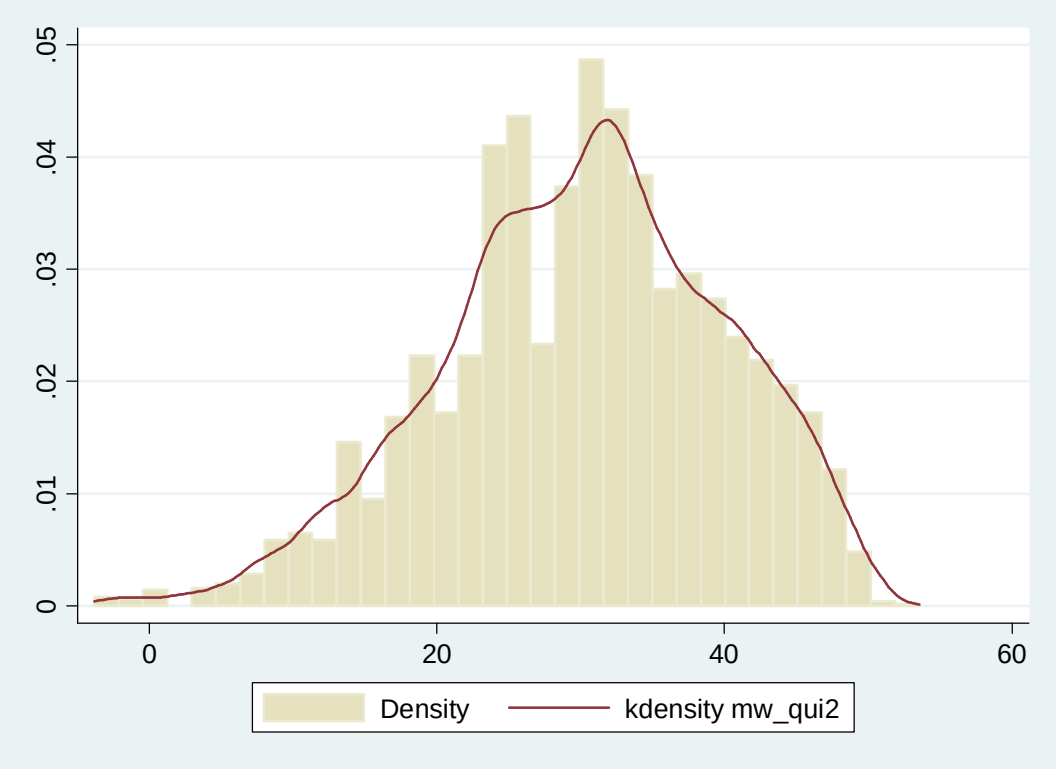
COMANDO STATA:

summarize mw_qui1,detail

MW_QUI1

	Percentiles	Smallest		
1%	29.87268	14.61893		
5%	38.47696	14.71448		
10%	42.87163	15.12862	Obs	2921
25%	54.2663	16.02822	Sum of Wgt.	2921
50%	64.74049		Mean	64.6284
	Largest		Std. Dev.	15.33231
75%	75.93005	96.68264		
90%	85.54338	98.2696	Variance	235.0796
95%	88.65627	110.6549	Skewness	-.2022683
99%	93.03078	111.156	Kurtosis	2.544995

twoway histogram mw_qui2, color(*.5) || kdensity mw_qui2



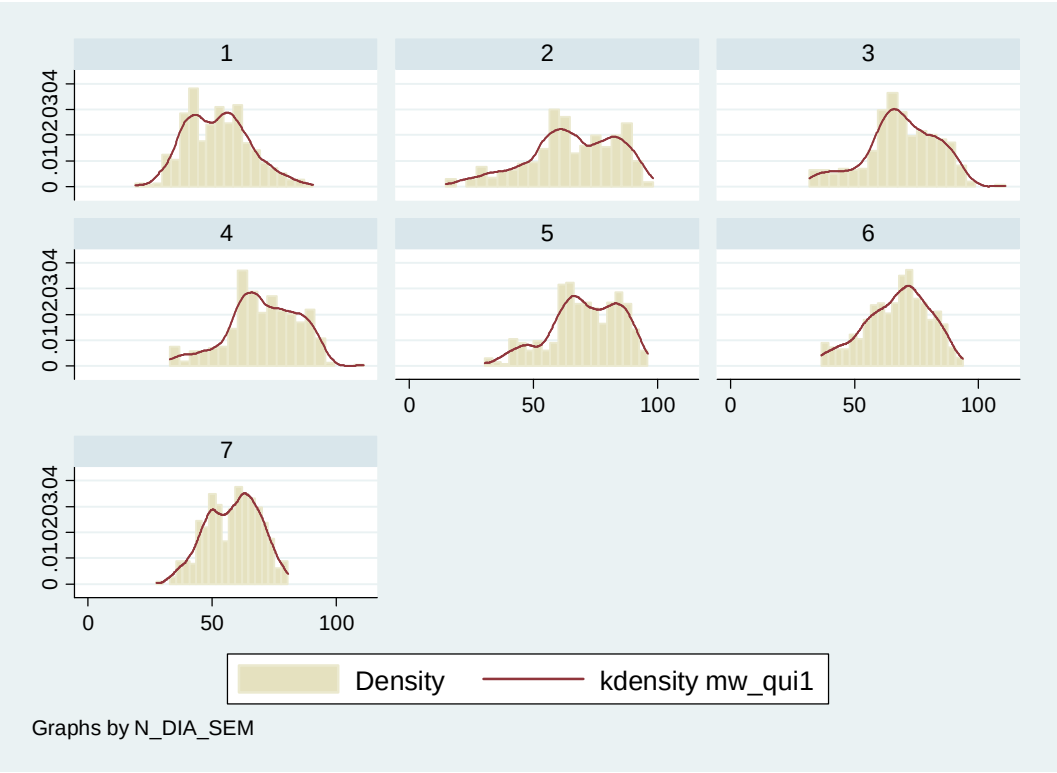
COMANDO STATA:

MW_QUI2

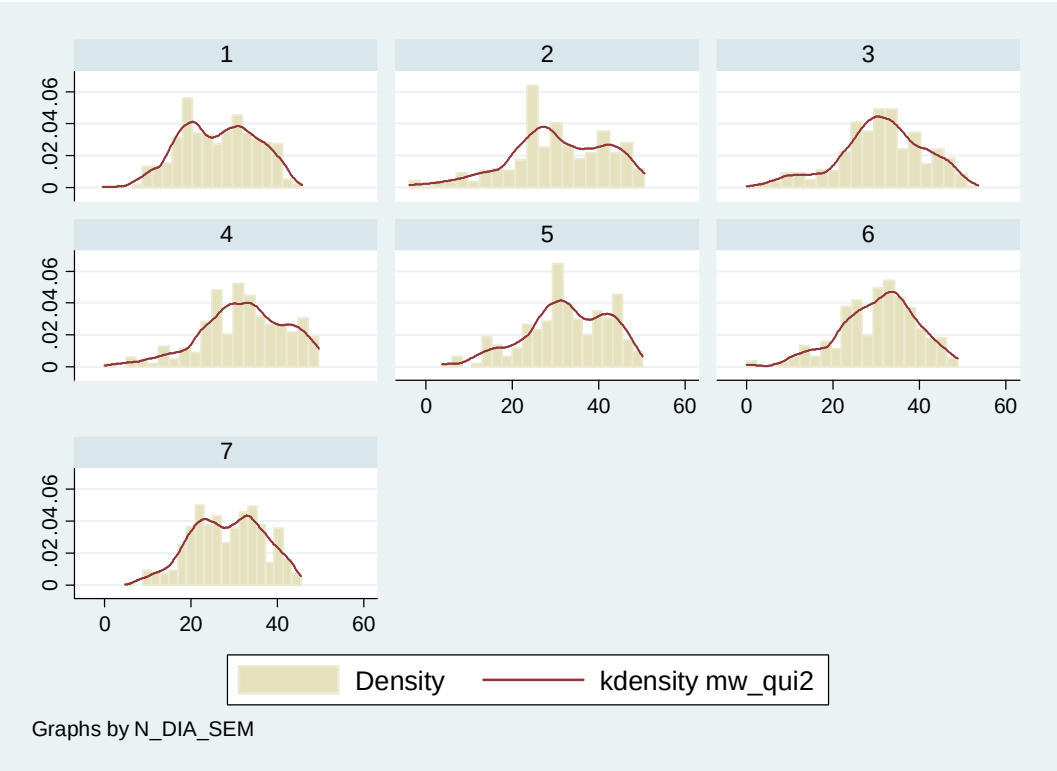
	Percentiles	Smallest		
1%	5.809041	-3.799733		
5%	13.36806	-3.619953		
10%	17.46868	-3.361758	Obs	2921
25%	23.92651	-2.614144	Sum of Wgt.	2921
50%	30.70162		Mean	30.11374
		Largest	Std. Dev.	9.757094
75%	37.22745	49.9588		
90%	43.03835	50.26398	Variance	95.20088
95%	45.62517	50.66072	Skewness	-.3200775
99%	48.37184	53.5472	Kurtosis	2.855984

COMANDO STATA:

twoway histogram mw_qui1, color(*.5) || kdensity mw_qui1||, by (n_dia_sem)

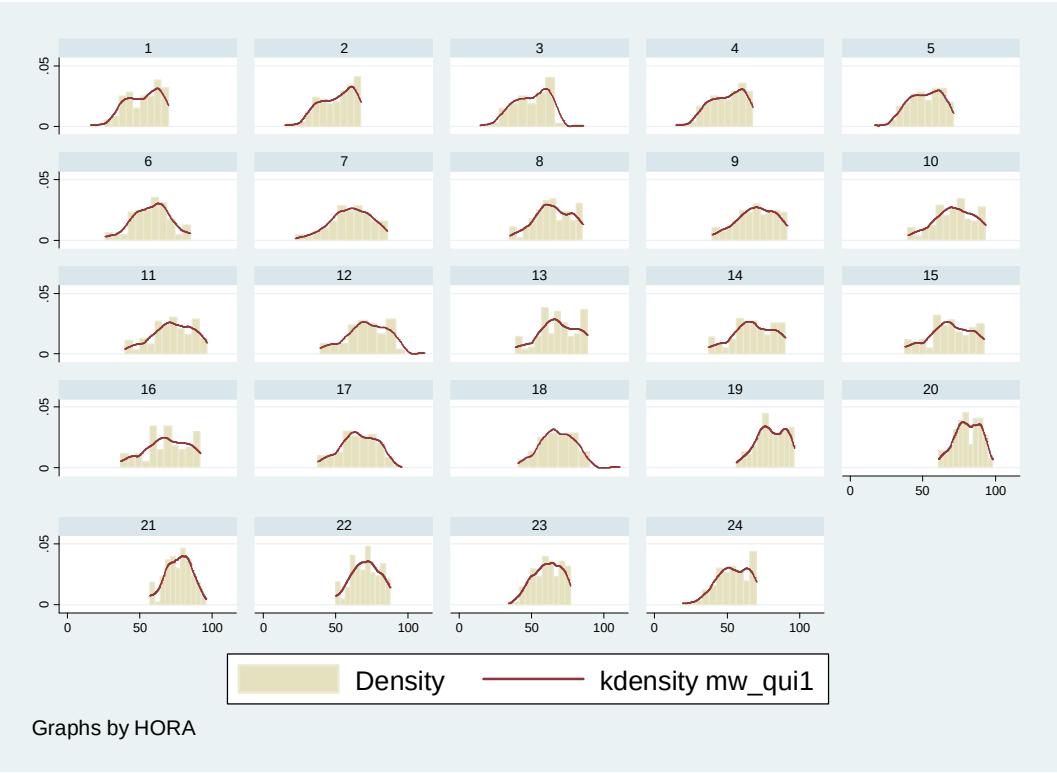


twoway histogram mw_qui2, color(*.5) || kdensity mw_qui2 ||, by (n_dia_sem)

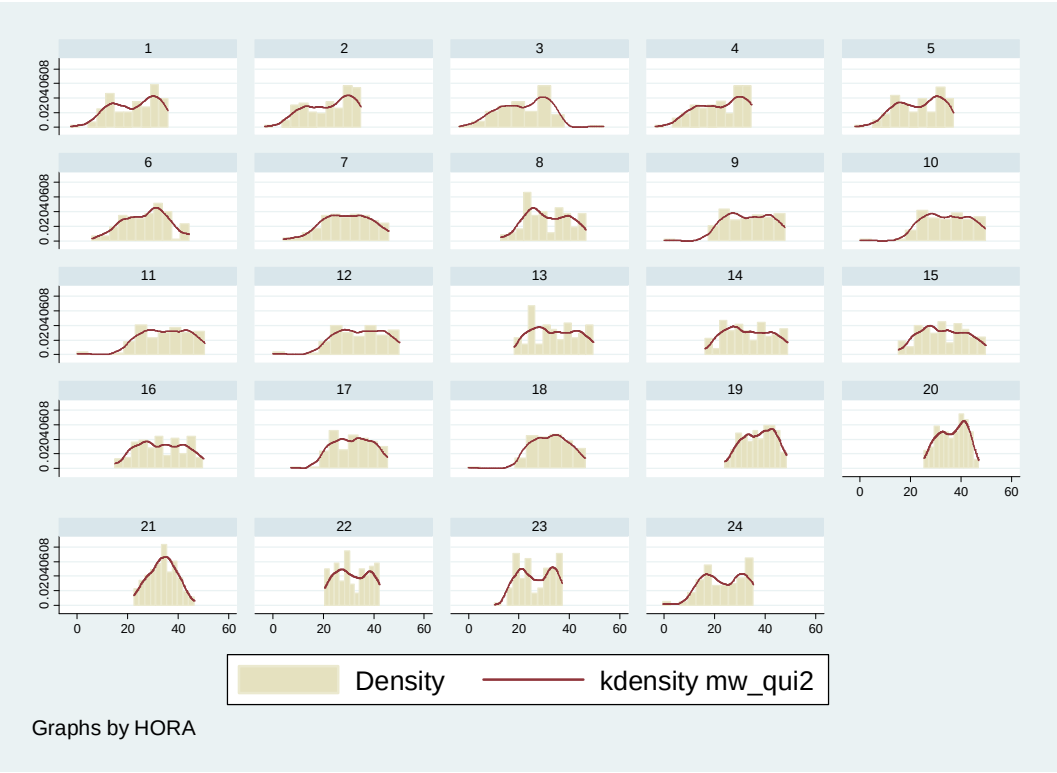


COMANDO STATA:

twoway histogram mw_qui1, color(*.5) || kdensity mw_qui1||, by (hora)



twoway histogram mw_qui2, color(*.5) || kdensity mw_qui2 ||, by (hora)



ELEPCOSA - AMBATO

anio ANIO

---

type: numeric (int)

range: [2007,2007] units: 1
unique values: 1 missing .: 0/2928

tabulation: Freq. Value
2928 2007

---

mes MES

---

type: numeric (byte)

range: [6,9] units: 1
unique values: 4 missing .: 0/2928

tabulation: Freq. Value
720 6
744 7
744 8
720 9

---

dia DIA

---

type: numeric (byte)

range: [1,31] units: 1
unique values: 31 missing .: 0/2928

mean: 15.7541
std. dev: 8.80846

percentiles: 10% 25% 50% 75% 90%
4 8 16 23 28

---

hora HORA

---

type: numeric (byte)

range: [1,24] units: 1
unique values: 24 missing .: 0/2928

mean: 12.5
std. dev: 6.92337

percentiles: 10% 25% 50% 75% 90%
3 6.5 12.5 18.5 22

---

n_dia_sem N_DIA_SEM

---

type: numeric (byte)

range: [1,7] units: 1
unique values: 7 missing .: 0/2928

tabulation: Freq. Value
432 1
408 2
408 3
408 4
408 5
432 6
432 7

---

feriado FERIADO

---

type: numeric (byte)

range: [0,1] units: 1
unique values: 2 missing .: 0/2928

tabulation: Freq. Value
2904 0
24 1

---

empresa EMPRESA

---

type: string (str6)

unique values: 1 missing "": 0/2928

tabulation: Freq. Value
2928 "ELEPCO"

---

subestacion SUBESTACION

---

type: string (str6)

unique values: 1 missing "": 0/2928

tabulation: Freq. Value
2928 "AMBATO"

---

pent PENT

---

type: string (str8)

unique values: 1 missing "": 0/2928

tabulation: Freq. Value
2928 "AMBATO_E"

---

mw MW

---

type: numeric (float)

range: [0,11.363799] units: 1.000e-07
unique values: 2251 missing .: 7/2928

```
                mean:   6.90664
            std. dev:   1.95511

          percentiles:       10%       25%       50%       75%       90%
                          3.93626   5.09014   7.34285    8.2211   8.79547
```

---

mvar MVAR

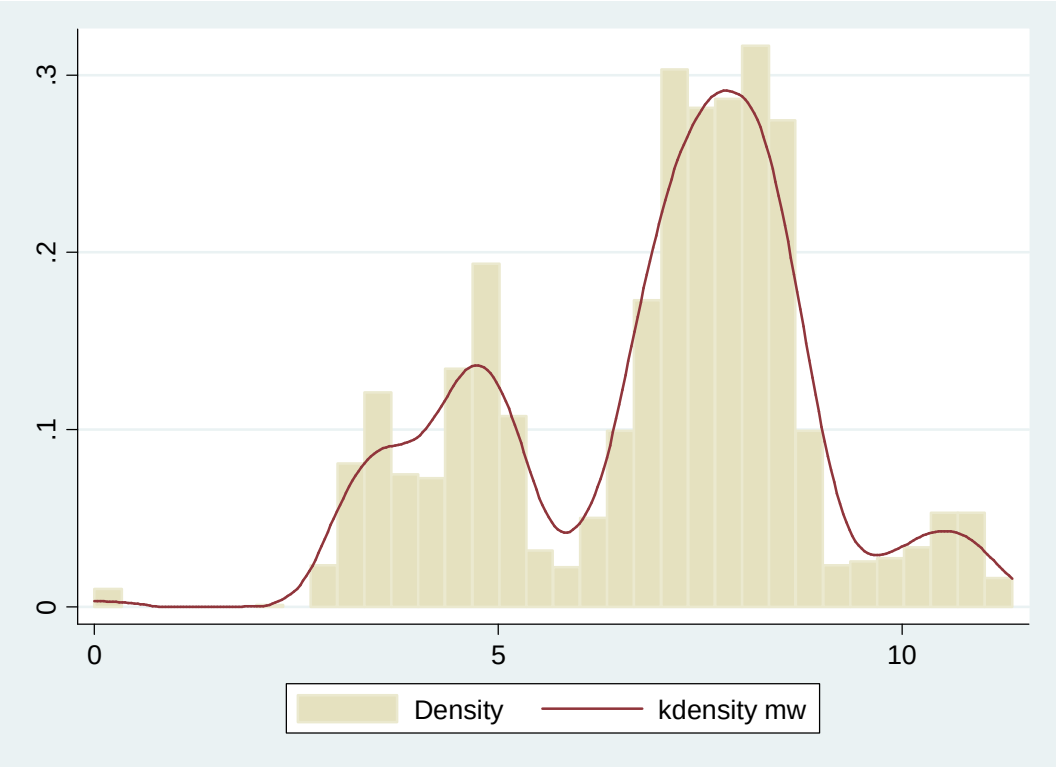
```
                type:  numeric (float)

               range:  [0,3.8918183]              units:  1.000e-08
         unique values:  1911                     missing .:  0/2928

                mean:    2.5118
            std. dev:   .668895

          percentiles:       10%       25%       50%       75%       90%
                           1.436    2.0822    2.6208   3.01492    3.3028
```

CORRELACIONES

```
             |       mw     mvar     hora n_dia_~m
-------------+------------------------------------
          mw |   1.0000
        mvar |   0.9048   1.0000
        hora |   0.2325   0.1478   1.0000
   _dia_sem |   0.1671   0.2033  -0.0003   1.0000
```

COMANDO STATA:

twoway histogram mw, color(*.5) || kdensity mw



COMANDO STATA:

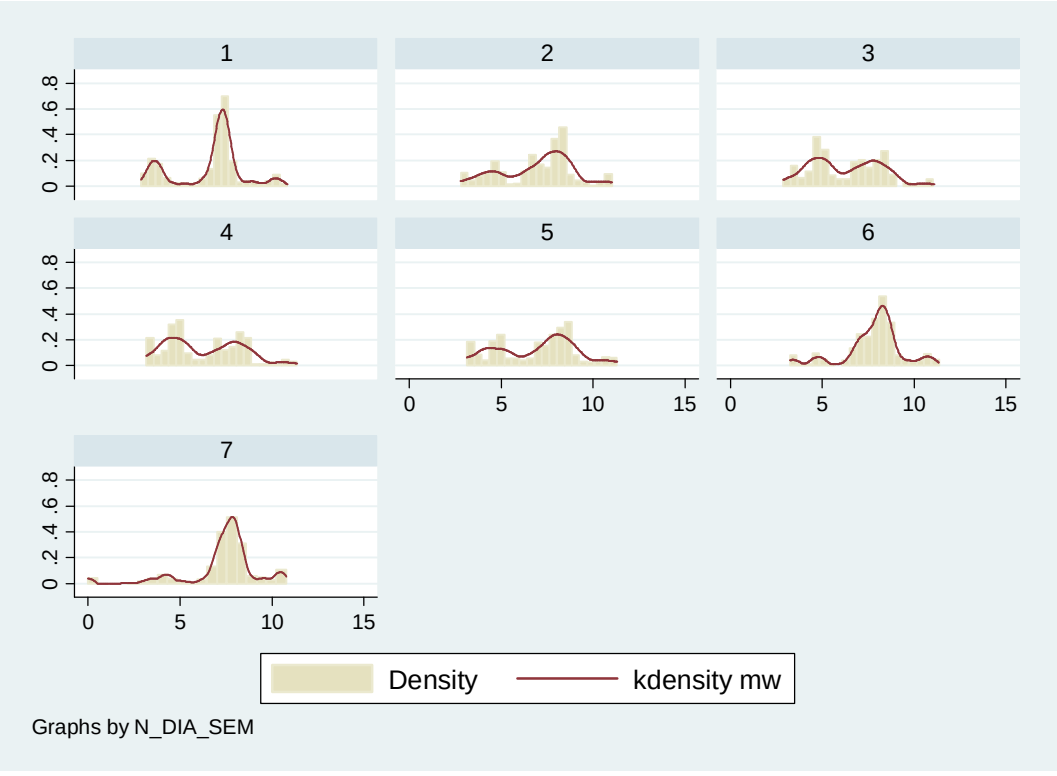
summarize mw ,detail

MW

Percentiles	Smallest		
1%	2.961926	0	
5%	3.447509	0	
10%	3.936261	0	Obs 2921
25%	5.090142	0	Sum of Wgt. 2921
50%	7.342853		Mean 6.906642
	Largest		Std. Dev. 1.95511
75%	8.2211	11.3085	
90%	8.795466	11.31397	Variance 3.822454
95%	10.1238	11.3444	Skewness -.3994619
99%	10.94087	11.3638	Kurtosis 2.767622

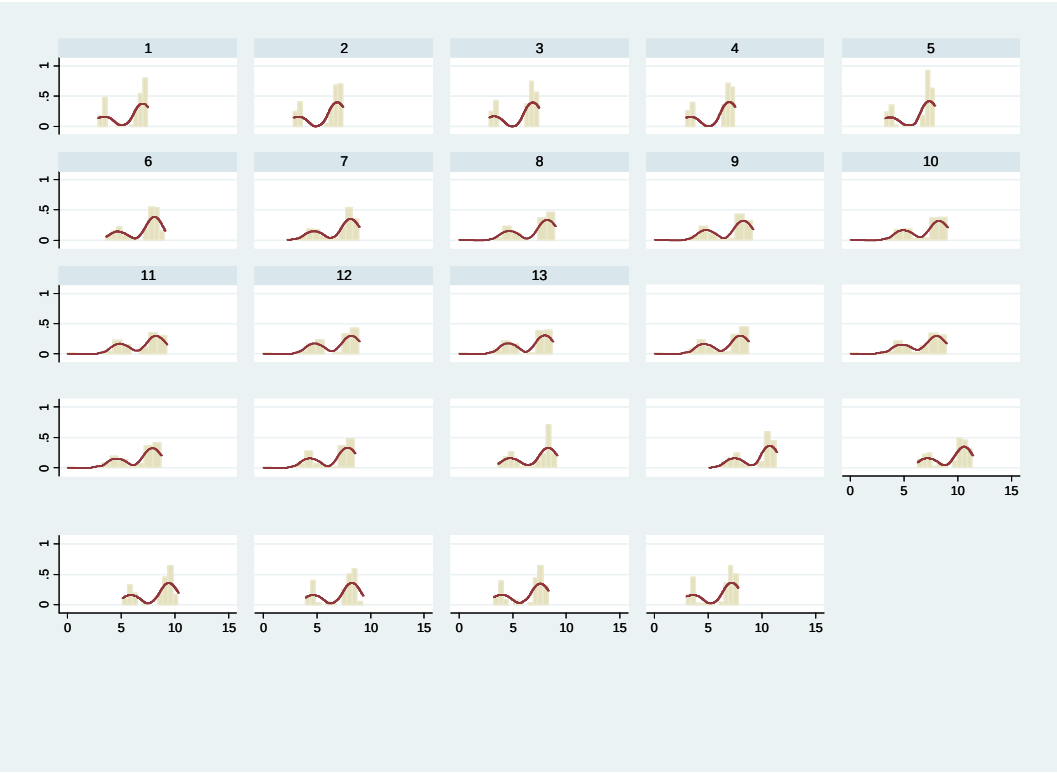
COMANDO STATA:

twoway histogram mw, color(*.5) || kdensity mw||, by (n_dia_sem)



COMANDO STATA:

twoway histogram mw, color(*.5) || kdensity mw||, by (hora)



EMELGUR-MILAGRO

anio ANIO

---

type: numeric (int)

range: [2007,2007] units: 1
unique values: 1 missing .: 0/2928

tabulation: Freq. Value
2928 2007

---

mes MES

---

type: numeric (byte)

range: [6,9] units: 1
unique values: 4 missing .: 0/2928

tabulation: Freq. Value
720 6
744 7
744 8
720 9

---

dia DIA

---

type: numeric (byte)

range: [1,31] units: 1
unique values: 31 missing .: 0/2928

mean: 15.7541
std. dev: 8.80846

percentiles: 10% 25% 50% 75% 90%
4 8 16 23 28

---

hora HORA

---

type: numeric (byte)

range: [1,24] units: 1
unique values: 24 missing .: 0/2928

mean: 12.5
std. dev: 6.92337

percentiles: 10% 25% 50% 75% 90%
3 6.5 12.5 18.5 22

---

n_dia_sem N_DIA_SEM

---

type: numeric (byte)

range: [1,7] units: 1
unique values: 7 missing .: 0/2928

tabulation: Freq. Value
432 1
408 2
408 3
408 4
408 5
432 6
432 7

---

feriado FERIADO

---

type: numeric (byte)

range: [0,1] units: 1
unique values: 2 missing .: 0/2928

tabulation: Freq. Value
2904 0
24 1

---

empresa EMPRESA

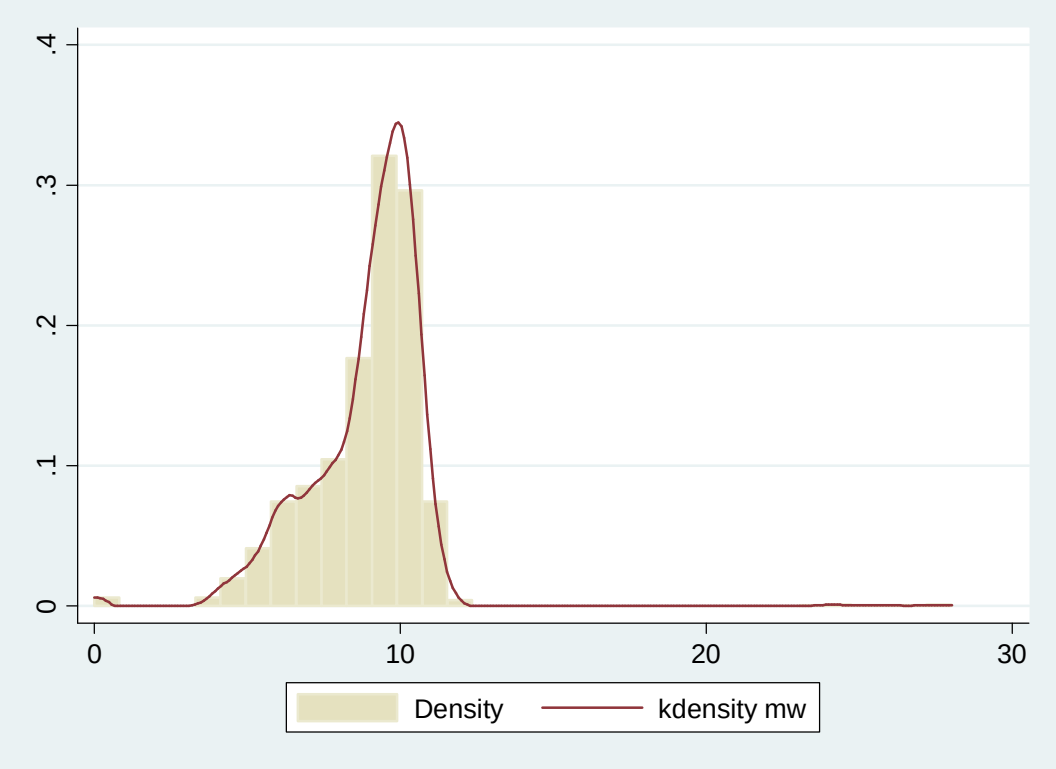
```
------------------------------------------------------------------------------

            type:  string (str7)

      unique values:  1                       missing "":  0/2928

         tabulation:  Freq.  Value
                     2928  "EMELGUR"

------------------------------------------------------------------------------
mw                                                                        MW
------------------------------------------------------------------------------

            type:  numeric (float)

           range:  [0,28.0347]                  units:  1.000e-07
      unique values:  2188                      missing .:  6/2928

            mean:   8.91155
       std. dev:   1.93247

      percentiles:        10%       25%       50%       75%       90%
                     6.29754   8.04236   9.33112   10.0937    10.554

------------------------------------------------------------------------------
mvar                                                                    MVAR
------------------------------------------------------------------------------

            type:  numeric (float)

           range:  [-4.2217422,9.6794996]       units:  1.000e-09
      unique values:  2073                      missing .:  0/2928

            mean:   .646166
       std. dev:   2.14497

      percentiles:        10%       25%       50%       75%       90%
                    -2.66246    -1.706   1.61136    2.3661    2.7548
```

CORRELACIONES

	mw	mvar	n_dia_~m	hora
mw	1.0000			
mvar	0.1148	1.0000		
n_dia_sem	0.3427	0.0893	1.0000	
hora	0.0818	0.0058	0.0035	1.0000

COMANDO STATA:

twoway histogram mw, color(*.5) || kdensity mw



COMANDO STATA:

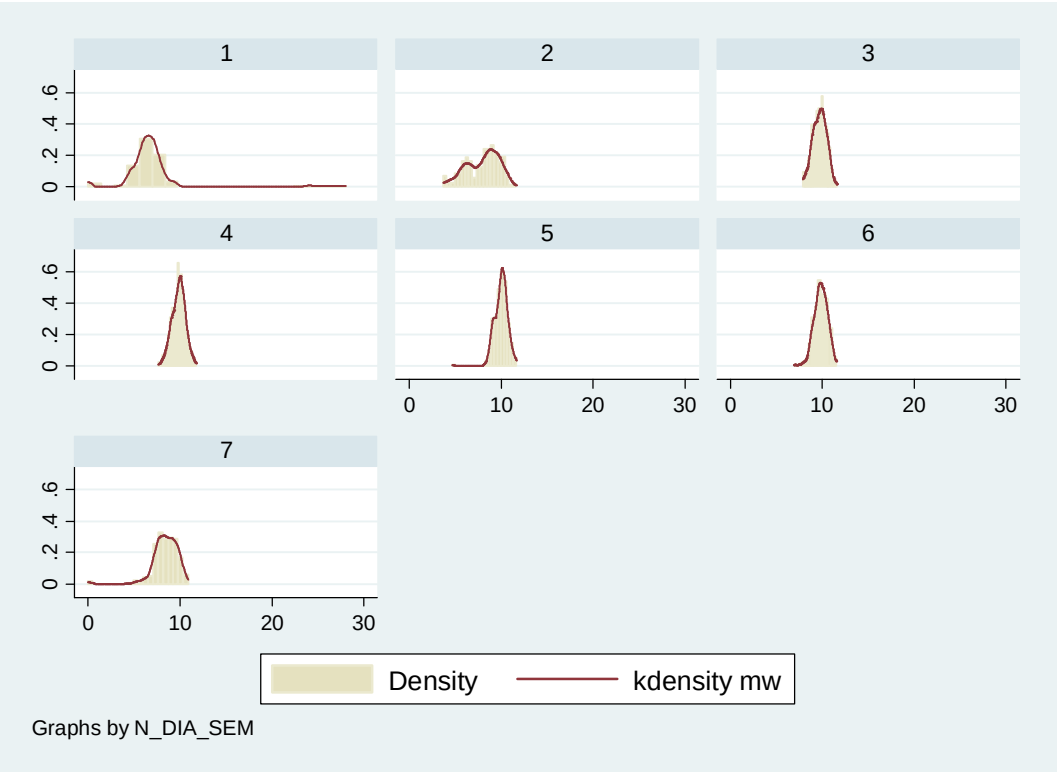
summarize mw ,detail

MW

Percentiles		Smallest		
1%	4.094181	0		
5%	5.59311	0		
10%	6.297544	0	Obs	2922
25%	8.042356	0	Sum of Wgt.	2922
50%	9.331115		Mean	8.911545
	Largest		Std. Dev.	1.932468
75%	10.09373	26.0988		
90%	10.55402	27.1026	Variance	3.734434
95%	10.8267	27.5328	Skewness	.8206209
99%	11.4003	28.0347	Kurtosis	20.61241

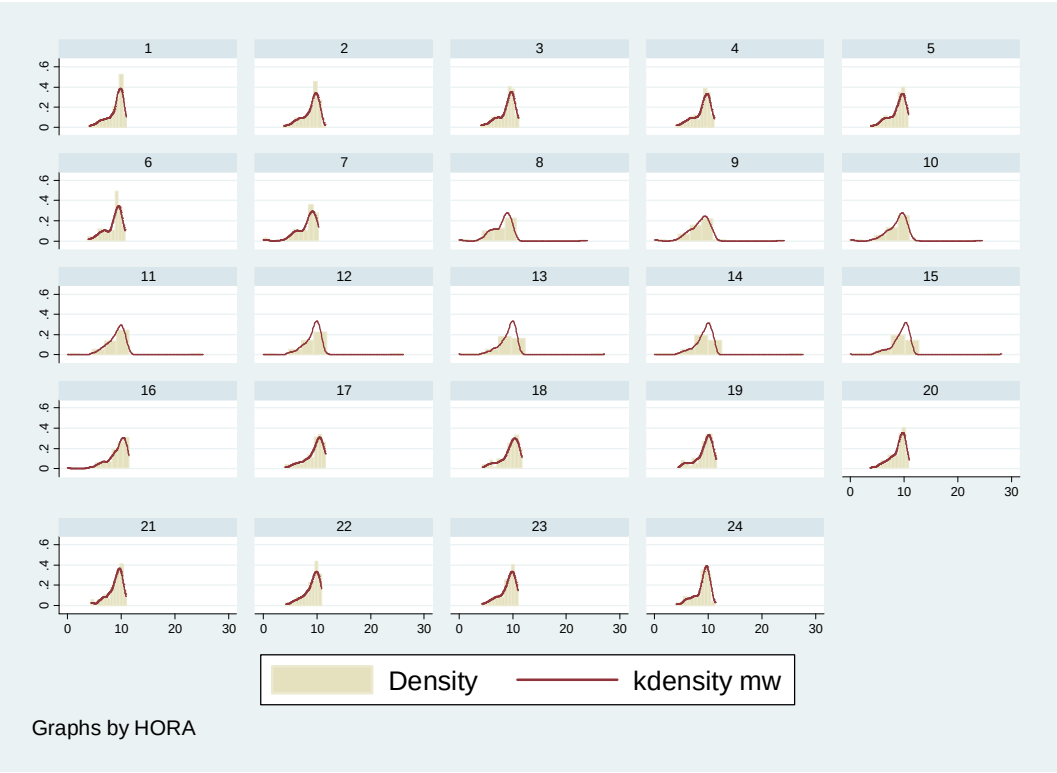
COMANDO STATA:

twoway histogram mw, color(*.5) || kdensity mw||, by (n_dia_sem)



COMANDO STATA:

twoway histogram mw, color(*.5) || kdensity mw||, by (hora)



EMELGUR-DOS CERRITOS

anio ANIO

---

type: numeric (int)

range: [2007,2007] units: 1
unique values: 1 missing .: 0/2928

tabulation: Freq. Value
2928 2007

---

mes MES

---

type: numeric (byte)

range: [6,9] units: 1
unique values: 4 missing .: 0/2928

tabulation: Freq. Value
720 6
744 7
744 8
720 9

---

dia DIA

---

type: numeric (byte)

range: [1,31] units: 1
unique values: 31 missing .: 0/2928

mean: 15.7541
std. dev: 8.80846

percentiles: 10% 25% 50% 75% 90%
4 8 16 23 28

---

hora HORA

---

type: numeric (byte)

range: [1,24] units: 1
unique values: 24 missing .: 0/2928

mean: 12.5
std. dev: 6.92337

percentiles: 10% 25% 50% 75% 90%
3 6.5 12.5 18.5 22

---

n_dia_sem N_DIA_SEM

```
            type:  numeric (byte)

           range:  [1,7]                units:  1
      unique values:  7                 missing .:  0/2928

         tabulation:  Freq.  Value
                      432  1
                      408  2
                      408  3
                      408  4
                      408  5
                      432  6
                      432  7

-------------------------------------------------------------------------------
feriado                                                         FERIADO
-------------------------------------------------------------------------------

            type:  numeric (byte)

           range:  [0,1]                units:  1
      unique values:  2                 missing .:  0/2928

         tabulation:  Freq.  Value
                     2904  0
                       24  1

-------------------------------------------------------------------------------
empresa                                                         EMPRESA
-------------------------------------------------------------------------------

            type:  string (str7)

      unique values:  1                 missing "":  0/2928

         tabulation:  Freq.  Value
                     2928  "EMELGUR"

-------------------------------------------------------------------------------
mw                                                                MW
-------------------------------------------------------------------------------

            type:  numeric (float)

           range:  [27.691456,71.537796]        units:  1.000e-06
      unique values:  2257                  missing .:  57/2928

            mean:   47.2707
         std. dev:   8.61046

        percentiles:        10%       25%       50%       75%       90%
                          37.2     40.05   46.5684   52.6119      60.6

-------------------------------------------------------------------------------
mvar                                                              MVAR
-------------------------------------------------------------------------------

            type:  numeric (float)
```

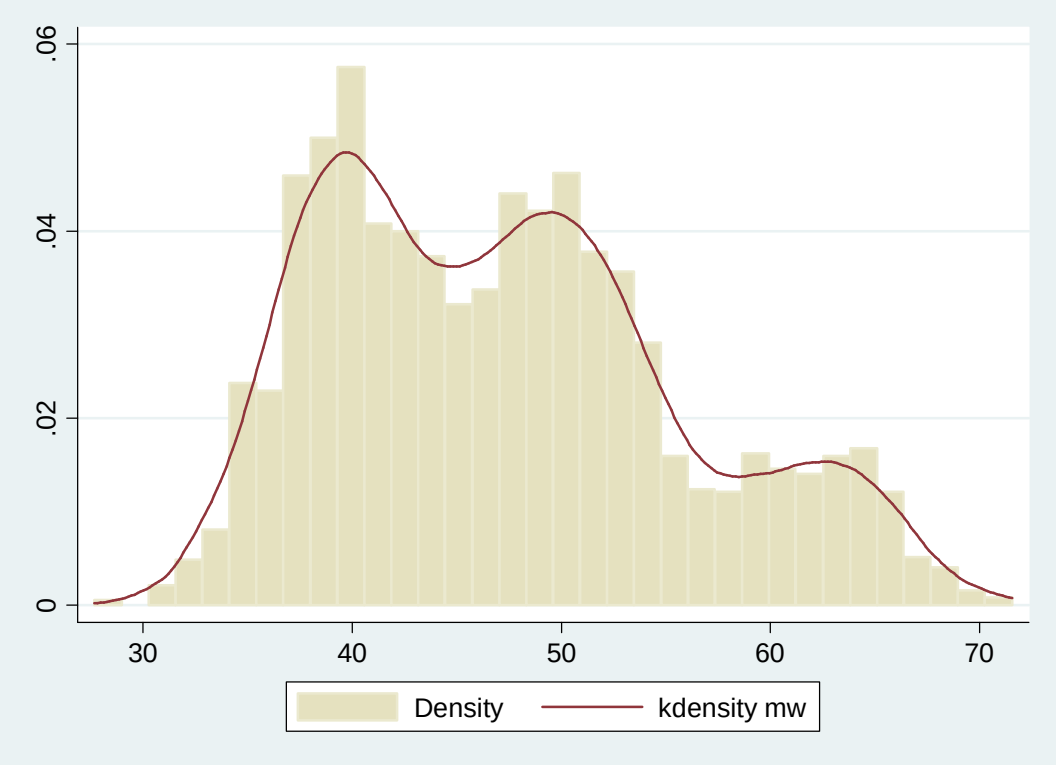
```
        range:  [6.4500003,26.380415]        units:  1.000e-07
unique values:  2212                     missing .:  0/2928

        mean:     18.35
     std. dev:   3.63325

  percentiles:       10%       25%       50%       75%       90%
                  13.8597   15.2433   18.4272   21.5088   23.0543
```

CORRELACIONES

```
             |       mw     mvar n_dia_~m     hora
-------------+------------------------------------
          mw |   1.0000
        mvar |   0.8557   1.0000
  n_dia_sem |   0.1085   0.1518   1.0000
        hora |   0.7208   0.5755   0.0035   1.0000
```

COMANDO STATA:

twoway histogram mw, color(*.5) || kdensity mw



COMANDO STATA:

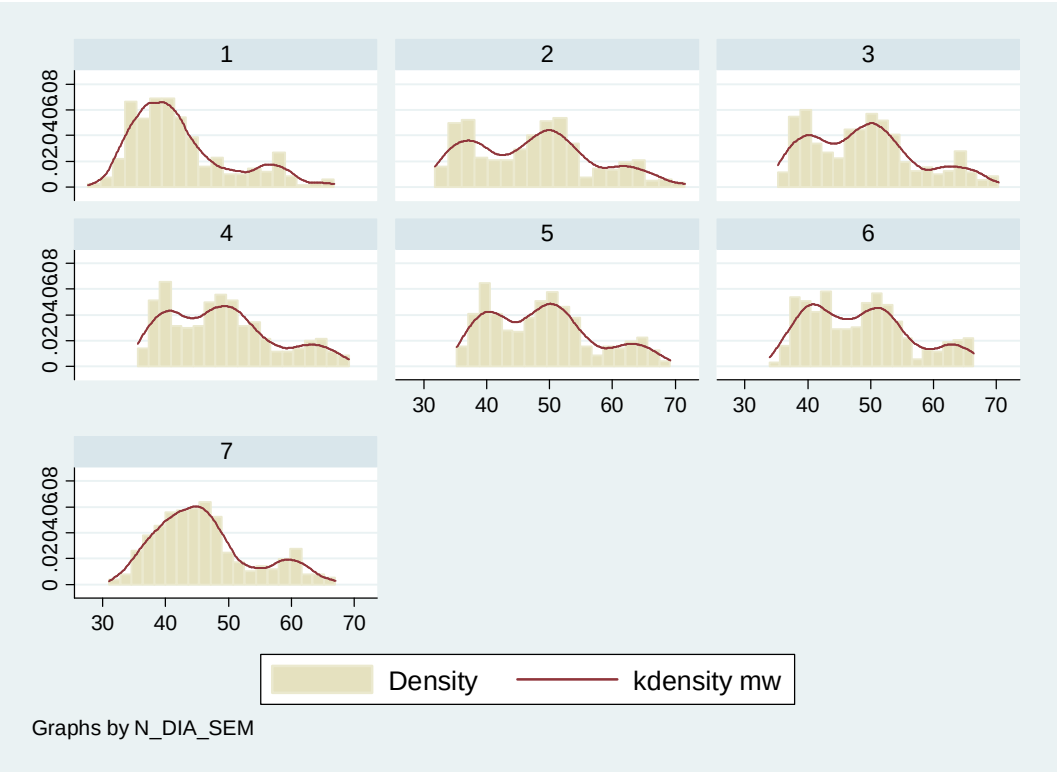
summarize mw ,detail

```
                          MW
-------------------------------------------------------------
      Percentiles      Smallest
```

1%	32.85	27.69146		
5%	35.4	28.11694		
10%	37.2	30.87942	Obs	2871
25%	40.05	30.92434	Sum of Wgt.	2871
50%	46.56841		Mean	47.27065
		Largest	Std. Dev.	8.610461
75%	52.61187	69.76545		
90%	60.6	70.40741	Variance	74.14004
95%	63.97351	71.037	Skewness	.4994232
99%	67.1226	71.5378	Kurtosis	2.491961

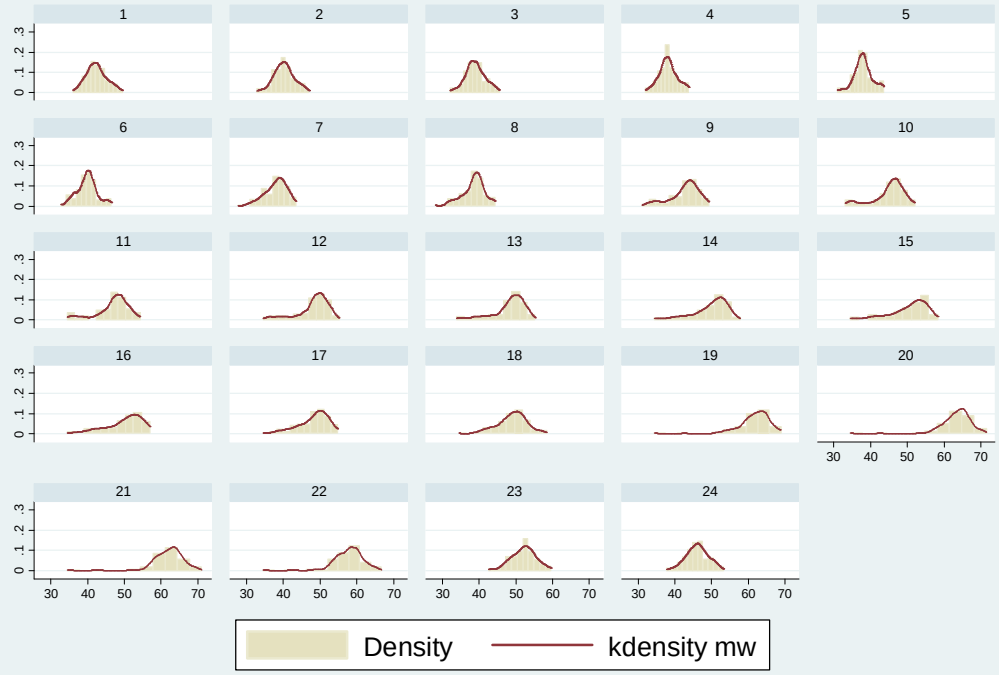
COMANDO STATA:

twoway histogram mw, color(*.5) || kdensity mw||, by (n_dia_sem)



COMANDO STATA:

twoway histogram mw, color(*.5) || kdensity mw||, by (hora)



EMELGUR-PASCUALES

anio ANIO

---

type: numeric (int)

range: [2007,2007] units: 1
unique values: 1 missing .: 0/2928

tabulation: Freq. Value
2928 2007

---

mes MES

---

type: numeric (byte)

range: [6,9] units: 1
unique values: 4 missing .: 0/2928

tabulation: Freq. Value
720 6
744 7
744 8
720 9

---

dia DIA

---

type: numeric (byte)

range: [1,31] units: 1
unique values: 31 missing .: 0/2928

mean: 15.7541
std. dev: 8.80846

percentiles: 10% 25% 50% 75% 90%
4 8 16 23 28

---

hora HORA

---

type: numeric (byte)

range: [1,24] units: 1
unique values: 24 missing .: 0/2928

mean: 12.5
std. dev: 6.92337

percentiles: 10% 25% 50% 75% 90%
3 6.5 12.5 18.5 22

---

n_dia_sem N_DIA_SEM

---

type: numeric (byte)

range: [1,7] units: 1
unique values: 7 missing .: 0/2928

tabulation: Freq. Value
432 1
408 2
408 3
408 4
408 5
432 6
432 7

---

feriado FERIADO

---

type: numeric (byte)

range: [0,1] units: 1
unique values: 2 missing .: 0/2928

tabulation: Freq. Value
2904 0
24 1

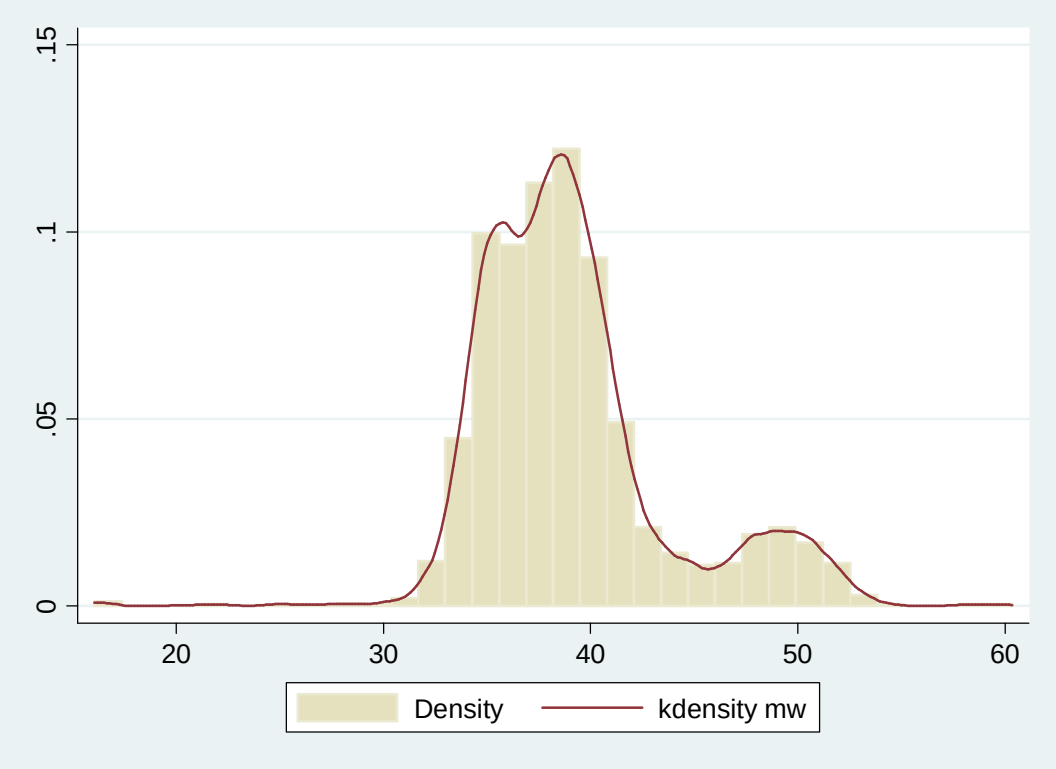
```
empresa                                                 EMPRESA
-------------------------------------------------------------------------------

             type:  string (str7)

      unique values:  1                  missing "":  0/2928

         tabulation:  Freq.  Value
                      2928  "EMELGUR"

-------------------------------------------------------------------------------
mw                                                          MW
-------------------------------------------------------------------------------

             type:  numeric (float)

            range:  [16.063231,60.329906]        units:  1.000e-06
      unique values:  2339                   missing .:  0/2928

             mean:   39.1121
          std. dev:   4.69912

       percentiles:       10%       25%       50%       75%       90%
                       34.5542   35.9494   38.3396   40.6342   47.0689

-------------------------------------------------------------------------------
mvar                                                        MVAR
-------------------------------------------------------------------------------

             type:  numeric (float)

            range:  [6.2963262,26.7708]          units:  1.000e-07
      unique values:  2224                   missing .:  0/2928

             mean:   16.8527
          std. dev:   2.10739

       percentiles:       10%       25%       50%       75%       90%
                       14.1797   15.2615   16.7401   18.2615     19.6
```

CORRELACIONES

```
             |      mw     mvar n_dia_~m     hora
-------------+------------------------------------
          mw |   1.0000
        mvar |   0.8277   1.0000
  n_dia_sem |   0.1136   0.1437   1.0000
        hora |   0.5703   0.4655   0.0000   1.0000
```

COMANDO STATA:

twoway histogram mw, color(*.5) || kdensity mw



COMANDO STATA:

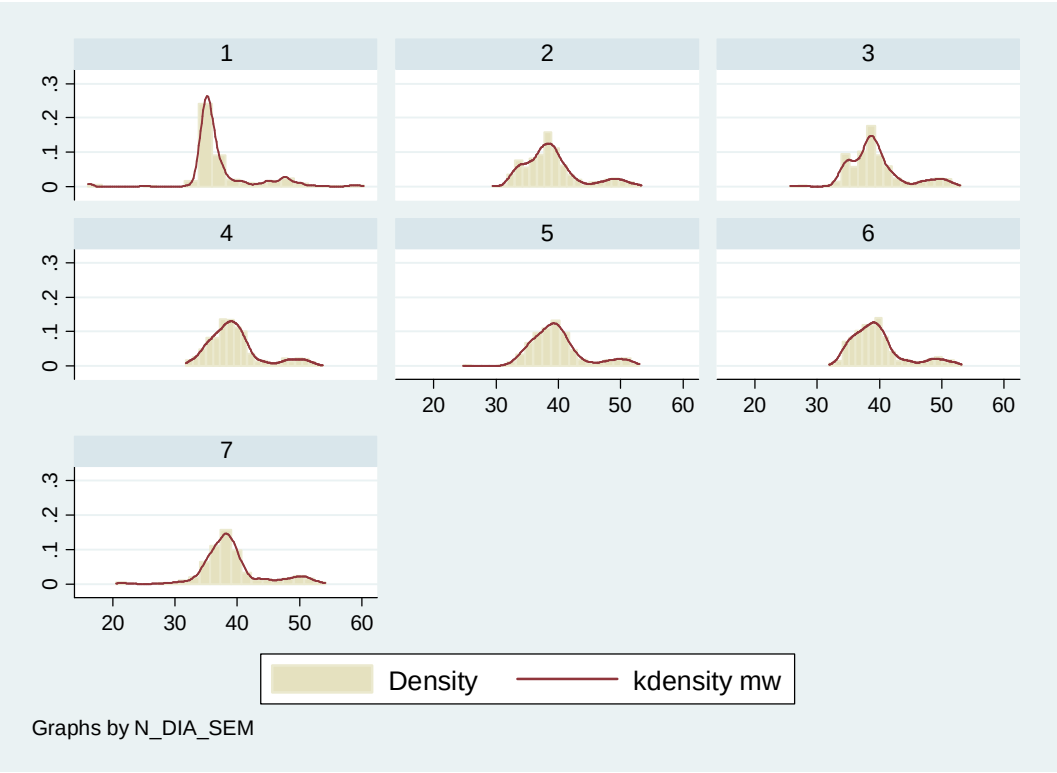
summarize mw ,detail

MW

Percentiles		Smallest		
1%	32.04462	16.06323		
5%	33.85318	16.12116		
10%	34.5542	16.17469	Obs	2928
25%	35.94936	16.17588	Sum of Wgt.	2928
50%	38.33961		Mean	39.11213
	Largest		Std. Dev.	4.699119
75%	40.63425	57.93966		
90%	47.06889	58.60893	Variance	22.08172
95%	49.67982	59.37381	Skewness	.8544613
99%	52.01184	60.32991	Kurtosis	5.095019

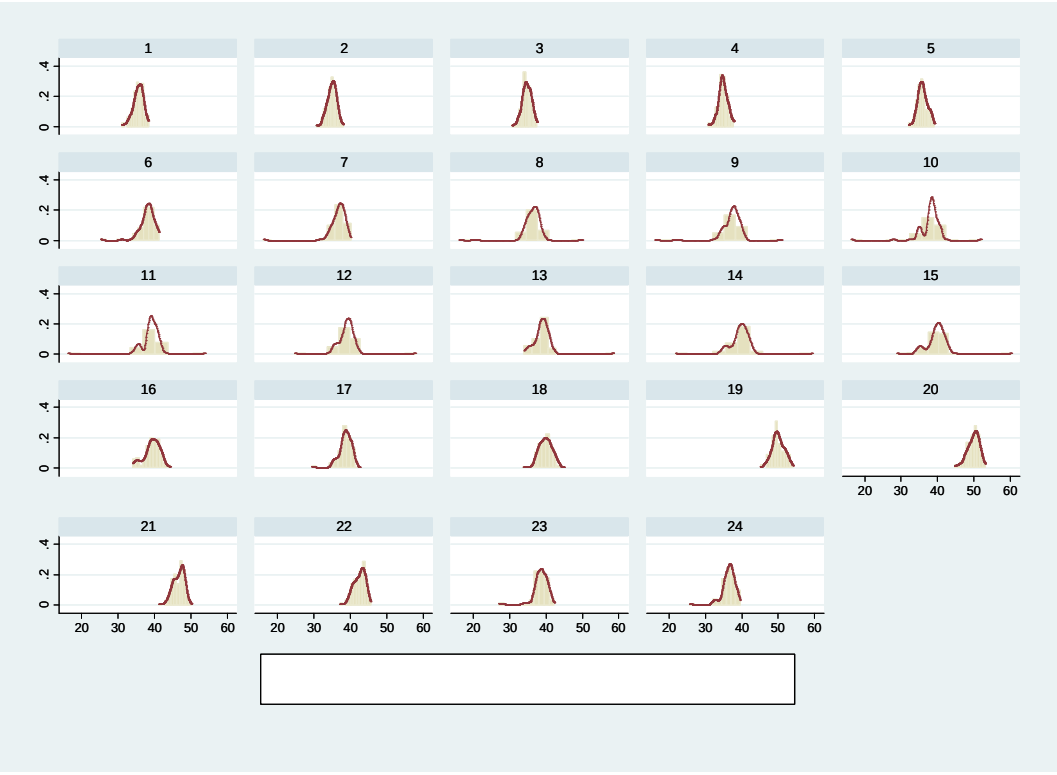
COMANDO STATA:

twoway histogram mw, color(*.5) || kdensity mw||, by (n_dia_sem)



COMANDO STATA:

twoway histogram mw, color(*.5) || kdensity mw||, by (hora)



EMELNORTE-IBARRA

anio ANIO

---

type: numeric (int)

range: [2007,2007] units: 1
unique values: 1 missing .: 0/2928

tabulation: Freq. Value
2928 2007

---

mes MES

---

type: numeric (byte)

range: [6,9] units: 1
unique values: 4 missing .: 0/2928

tabulation: Freq. Value
720 6
744 7
744 8
720 9

---

dia DIA

---

type: numeric (byte)

range: [1,31] units: 1
unique values: 31 missing .: 0/2928

mean: 15.7541
std. dev: 8.80846

percentiles: 10% 25% 50% 75% 90%
4 8 16 23 28

---

hora HORA

---

type: numeric (byte)

range: [1,24] units: 1
unique values: 24 missing .: 0/2928

mean: 12.5
std. dev: 6.92337

percentiles: 10% 25% 50% 75% 90%
3 6.5 12.5 18.5 22

---

n_dia_sem N_DIA_SEM

```
            type:  numeric (byte)

           range:  [1,7]                units:  1
      unique values:  7                 missing .:  0/2928

         tabulation:  Freq.  Value
                      432  1
                      408  2
                      408  3
                      408  4
                      408  5
                      432  6
                      432  7

-------------------------------------------------------------------------------
feriado                                                 FERIADO
-------------------------------------------------------------------------------

            type:  numeric (byte)

           range:  [0,1]                units:  1
      unique values:  2                 missing .:  0/2928

         tabulation:  Freq.  Value
                     2904  0
                       24  1

-------------------------------------------------------------------------------
empresa                                                 EMPRESA
-------------------------------------------------------------------------------

            type:  string (str9)

      unique values:  1                 missing "":  0/2928

         tabulation:  Freq.  Value
                     2928  "EMELNORTE"

-------------------------------------------------------------------------------
mw_cot                                                  MW_COT
-------------------------------------------------------------------------------

            type:  numeric (float)

           range:  [0,4.0391002]          units:  1.000e-07
      unique values:  783               missing .:  0/2928

              mean:   1.94693
         std. dev:   .601032

        percentiles:        10%       25%       50%       75%       90%
                         1.434    1.5774   1.77946   2.04613   3.01639

-------------------------------------------------------------------------------
mvar_cot                                                MVAR_COT
-------------------------------------------------------------------------------

            type:  numeric (float)
```

range: [0,.94405001] units: 1.000e-09
unique values: 381 missing .: 0/2928

mean: .484665
std. dev: .235342

percentiles: 10% 25% 50% 75% 90%
0 .359338 .5019 .6692 .761435

---

mw_ret MW_RET

---

type: numeric (float)

range: [0,15.424478] units: 1.000e-08
unique values: 1861 missing .: 0/2928

mean: 8.19685
std. dev: 3.20091

percentiles: 10% 25% 50% 75% 90%
5.975 6.5247 8.53306 9.35683 12.8861

---

mvar_ret MVAR_RET

---

type: numeric (float)

range: [-5.4016261,5.5750003] units: 1.000e-07
unique values: 1371 missing .: 0/2928

mean: 3.07383
std. dev: 2.15302

percentiles: 10% 25% 50% 75% 90%
0 2.95 3.585 4.37958 4.75

---

mw_ota MW_OTA

---

type: numeric (float)

range: [0,24.4258] units: 1.000e-08
unique values: 2826 missing .: 2/2928

mean: 16.8808
std. dev: 3.98196

percentiles: 10% 25% 50% 75% 90%
11.9334 13.2169 16.8383 19.8776 22.6676

---

mvar_ota MVAR_OTA

---

type: numeric (float)

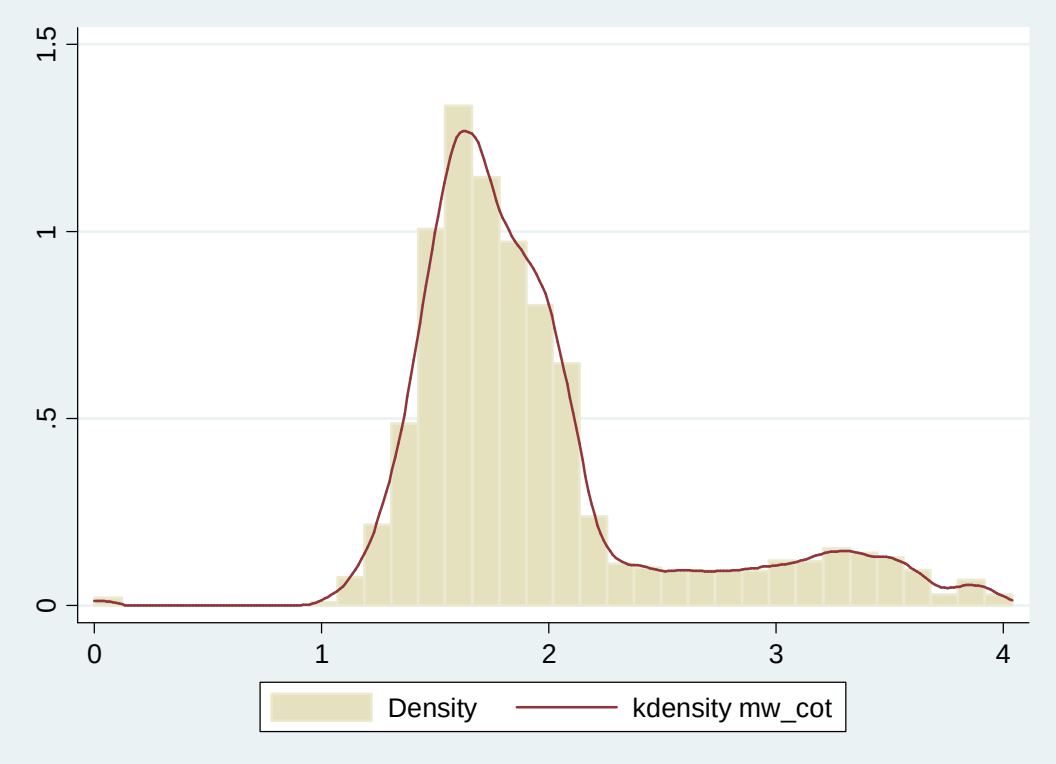
```
        range:  [-.45410001,6.9775]          units:  1.000e-09
unique values:  1438                     missing .:  0/2928

         mean:   1.38592
      std. dev:   1.84039

  percentiles:        10%       25%       50%       75%       90%
                        0         0   .367173    2.5616   4.64173
```

CORRELACIONES

	mw_cot	mvar_cot	mw_ret	mvar_ret	mw_ota	mvar_ota	n_dia_~m	hora
mw_cot	1.0000							
mvar_cot	0.2000	1.0000						
mw_ret	0.5922	0.2503	1.0000					
mvar_ret	0.1336	0.1352	0.4157	1.0000				
mw_ota	0.4732	0.5955	0.4412	0.0507	1.0000			
mvar_ota	0.1292	0.4335	0.2702	0.4554	0.2413	1.0000		
n_dia_sem	0.0953	0.1955	-0.0296	-0.0088	0.1475	0.0979	1.0000	
hora	0.5021	0.3166	0.4333	0.1320	0.3225	0.0689	-0.0003	1.0000

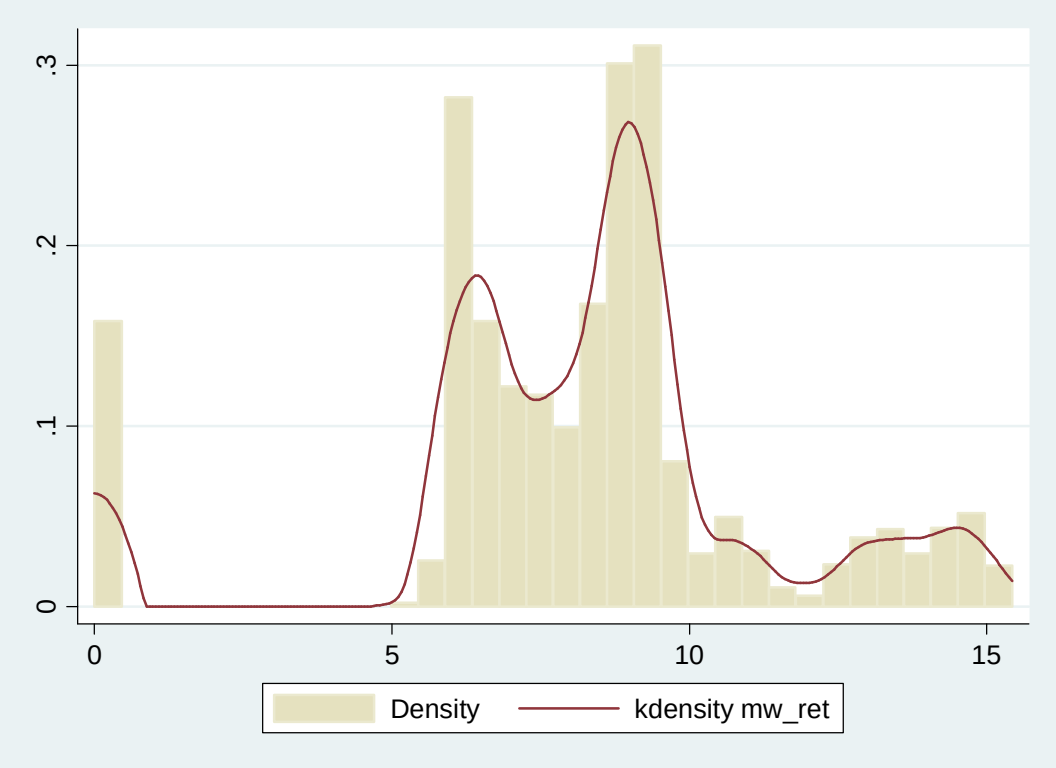
COMANDO STATA:

twoway histogram mw_cot, color(*.5) || kdensity mw_cot

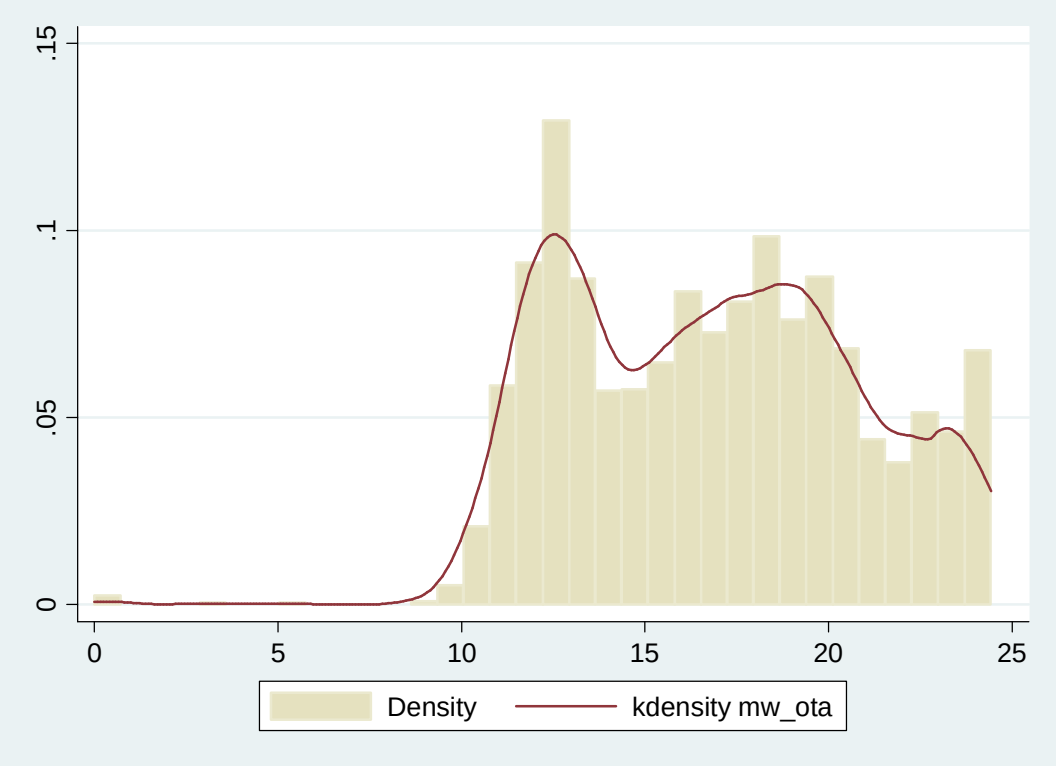


COMANDO STATA:

twoway histogram mw_ret, color(*.5) || kdensity mw_ret



COMANDO STATA:



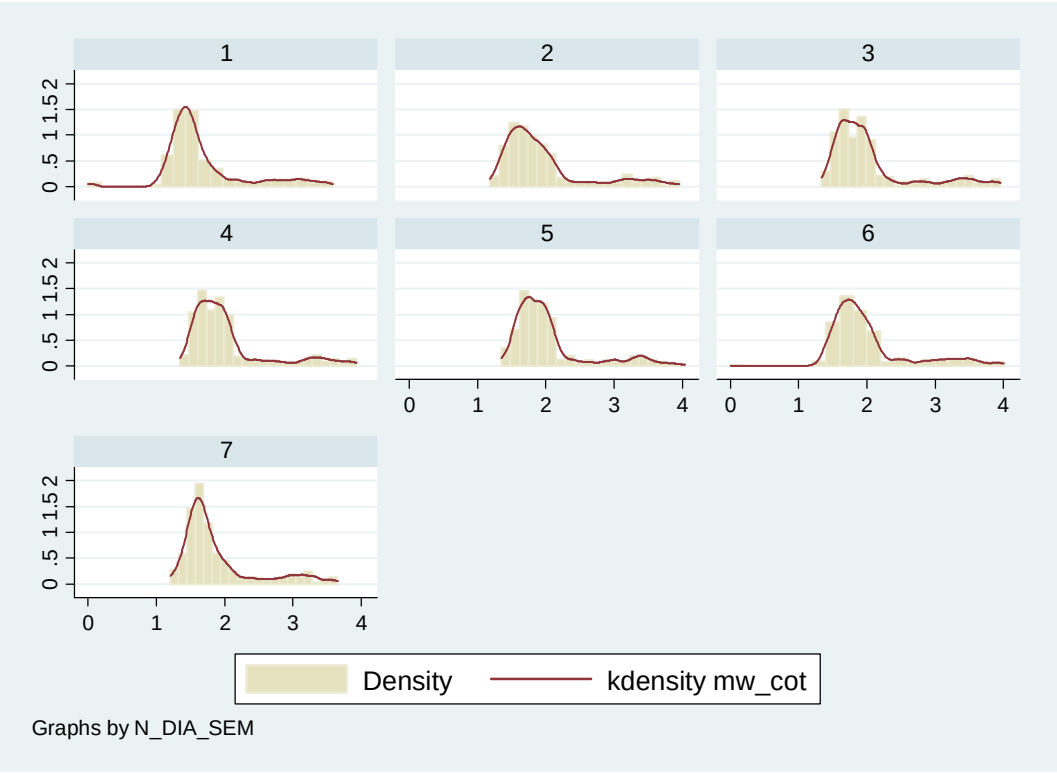
COMANDO STATA:

summarize mw_cot mw_ret mw_ota,detail

```
      MW_COT
-------------------------------------------------------------
     Percentiles      Smallest
 1%      1.1711              0
 5%      1.3384              0
10%       1.434              0       Obs                2928
25%      1.5774              0       Sum of Wgt.        2928

50%    1.779456                      Mean           1.946932
                        Largest       Std. Dev.      .6010321
75%     2.04613        3.95545
90%    3.016392        4.00325       Variance       .3612396
95%    3.383804        4.00325       Skewness        1.45021
99%       3.824        4.0391        Kurtosis        4.89831

                      MW_RET
-------------------------------------------------------------
     Percentiles      Smallest
 1%            0              0
 5%            0              0
10%        5.975              0       Obs                2928
25%       6.5247              0       Sum of Wgt.        2928

50%    8.533062                      Mean           8.196848
                        Largest       Std. Dev.      3.200915
75%     9.35683        15.2721
90%    12.88611        15.2721       Variance       10.24586
95%    14.19521        15.2721       Skewness      -.5304967
99%     14.9838        15.42448      Kurtosis       4.439269

                      MW_OTA
-------------------------------------------------------------
     Percentiles      Smallest
 1%    10.28226              0
 5%    11.2993               0
10%    11.93345              0       Obs                2926
25%    13.21686              0       Sum of Wgt.        2926

50%    16.83834                      Mean           16.88075
                        Largest       Std. Dev.      3.981964
75%    19.87757        24.4258
90%    22.66759        24.4258       Variance       15.85604
95%     23.687         24.4258       Skewness       .0425241
99%    24.41104        24.4258       Kurtosis       2.458278
```

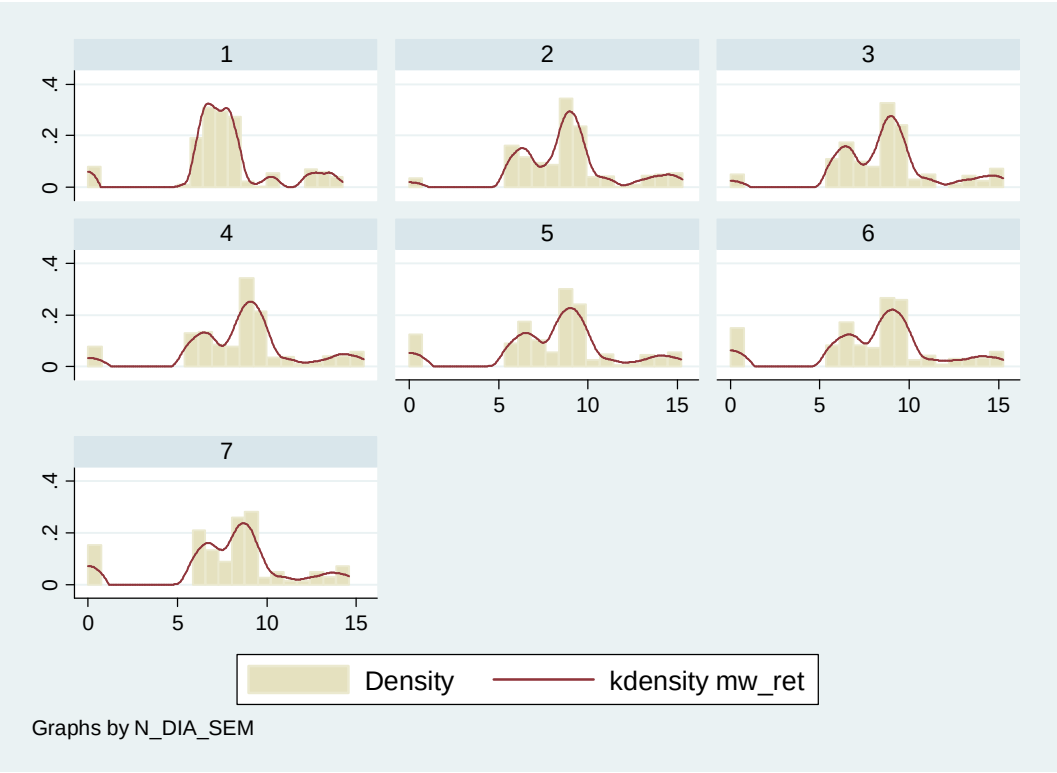
COMANDO STATA:

twoway histogram mw_cot, color(*.5) || kdensity mw_cot||, by (n_dia_sem)



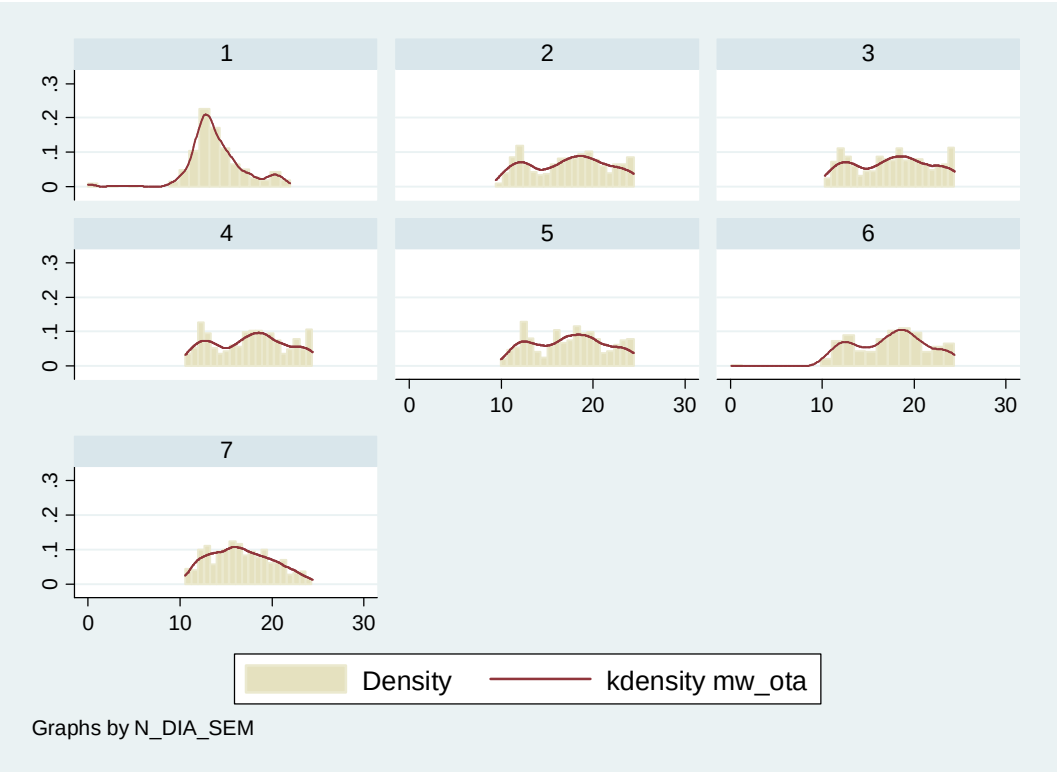
COMANDO STATA:

twoway histogram mw_ret, color(*.5) || kdensity mw_ret ||, by (n_dia_sem)



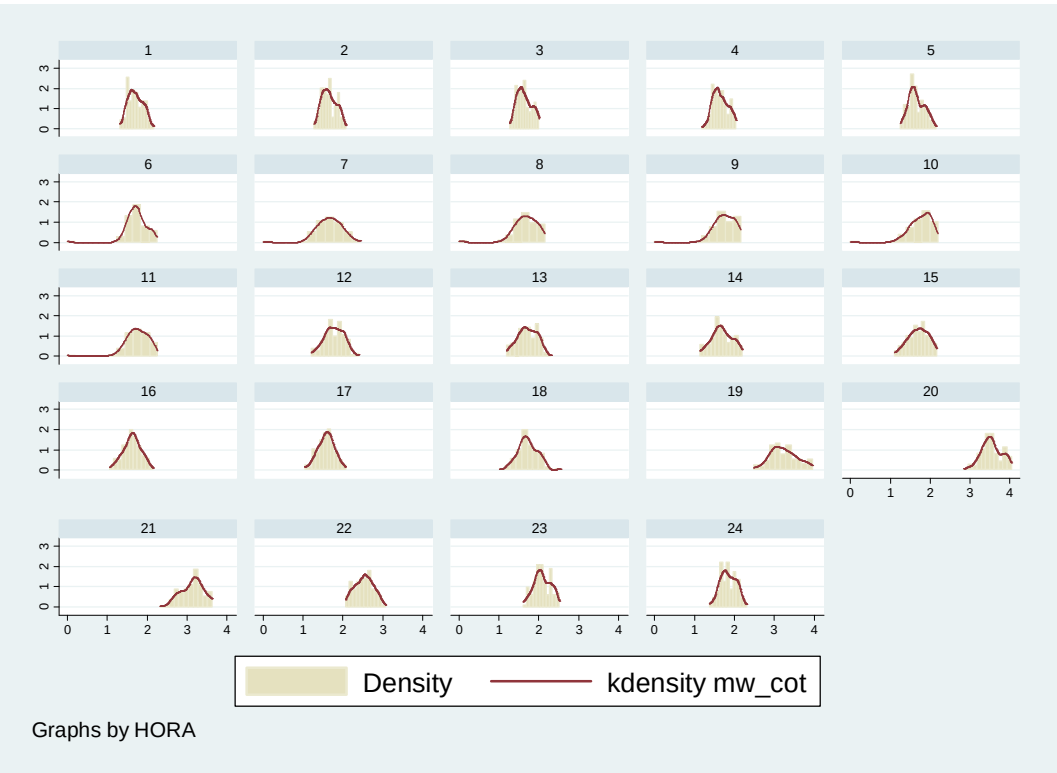
COMANDO STATA:

twoway histogram mw_ota, color(*.5) || kdensity mw_ota ||, by (n_dia_sem)



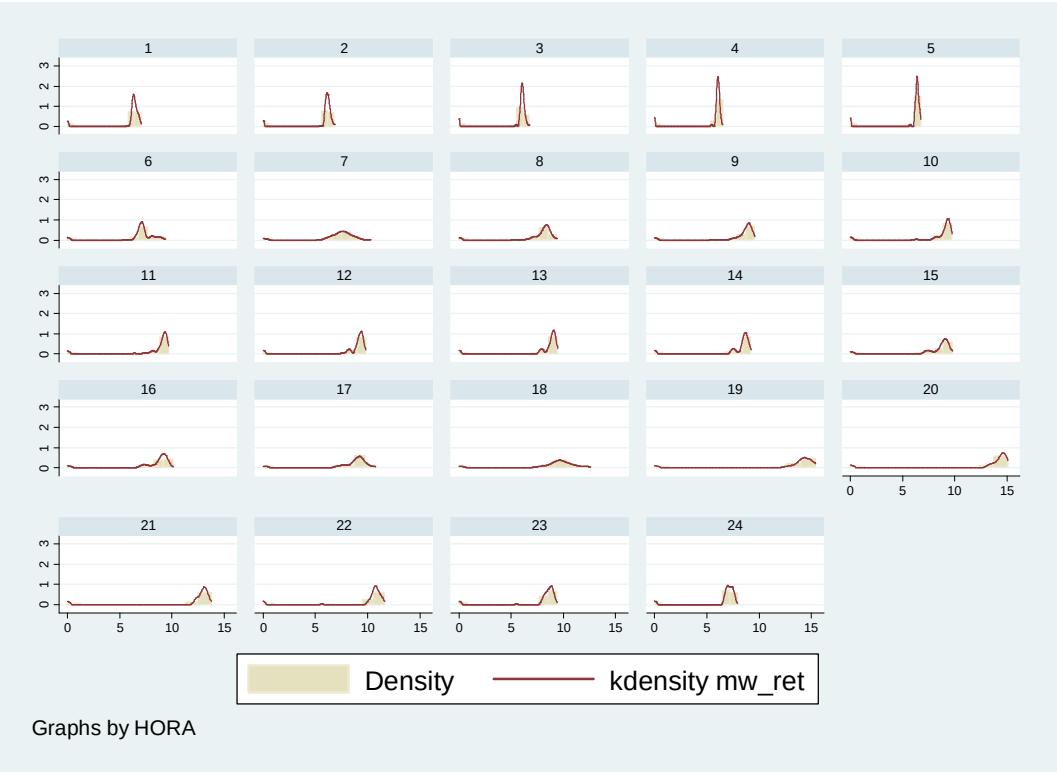
COMANDO STATA:

twoway histogram mw_cot, color(*.5) || kdensity mw_cot||, by (hora)



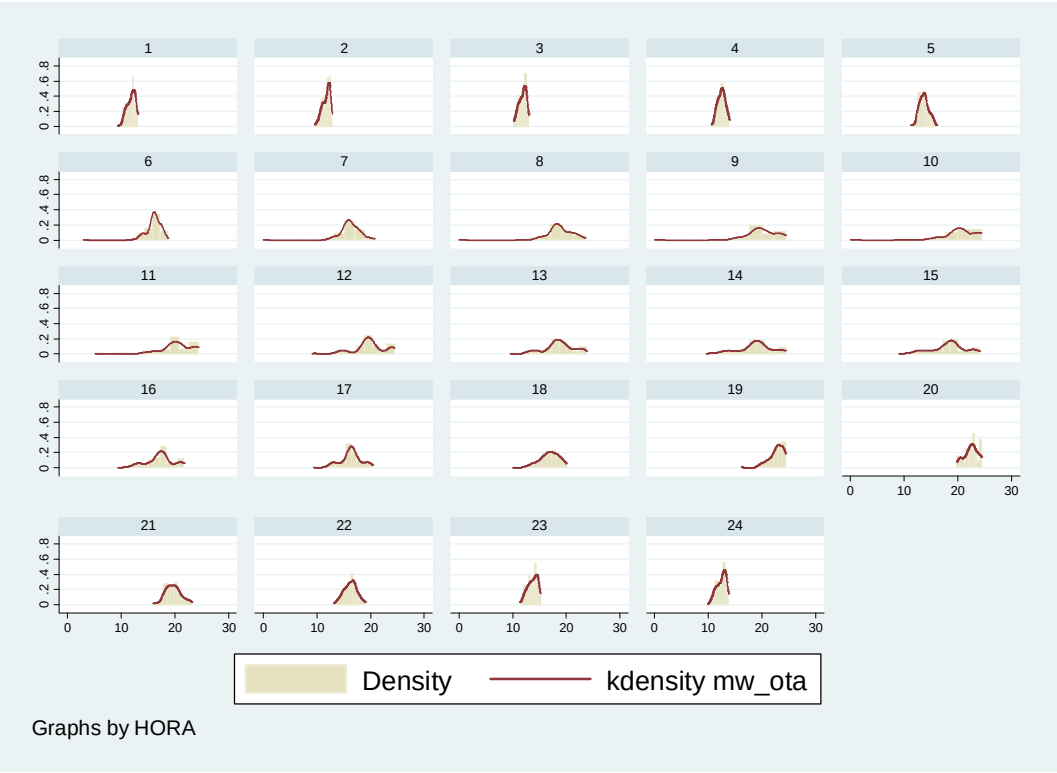
COMANDO STATA:

twoway histogram mw_ret, color(*.5) || kdensity mw_ret ||, by (hora)



COMANDO STATA:

twoway histogram mw_ota, color(*.5) || kdensity mw_ota ||, by (hora)



EMELSAD-S. DOMINGO

anio ANIO

---

type: numeric (int)

range: [2007,2007] units: 1
unique values: 1 missing .: 0/2928

tabulation: Freq. Value
2928 2007

---

mes MES

---

type: numeric (byte)

range: [6,9] units: 1
unique values: 4 missing .: 0/2928

tabulation: Freq. Value
720 6
744 7
744 8
720 9

---

dia DIA

---

type: numeric (byte)

range: [1,31] units: 1
unique values: 31 missing .: 0/2928

mean: 15.7541
std. dev: 8.80846

percentiles: 10% 25% 50% 75% 90%
4 8 16 23 28

---

hora HORA

---

type: numeric (byte)

range: [1,24] units: 1
unique values: 24 missing .: 0/2928

mean: 12.5
std. dev: 6.92337

percentiles: 10% 25% 50% 75% 90%
3 6.5 12.5 18.5 22

---

n_dia_sem N_DIA_SEM

---

type: numeric (byte)

```
            range:  [1,7]                units:  1
      unique values:  7                  missing .:  0/2928

        tabulation:  Freq.  Value
                     432  1
                     408  2
                     408  3
                     408  4
                     408  5
                     432  6
                     432  7

-------------------------------------------------------------------------------
feriado                                                 FERIADO
-------------------------------------------------------------------------------

            type:  numeric (byte)

            range:  [0,1]                units:  1
      unique values:  2                  missing .:  0/2928

        tabulation:  Freq.  Value
                    2904  0
                      24  1

-------------------------------------------------------------------------------
empresa                                                 EMPRESA
-------------------------------------------------------------------------------

            type:  string (str7)

      unique values:  1                  missing "":  0/2928

        tabulation:  Freq.  Value
                    2928  "EMELSAD"

-------------------------------------------------------------------------------
mw_sdo1                                                 MW_SDO1
-------------------------------------------------------------------------------

            type:  numeric (float)

           range:  [8.9386005,22.4182]         units:  1.000e-07
      unique values:  2318                  missing .:  0/2928

            mean:   14.2987
       std. dev:   2.89956

      percentiles:        10%       25%       50%       75%       90%
                     10.6355   11.6688   14.4823   16.1564   18.5541

-------------------------------------------------------------------------------
mvar_sdo1                                               MVAR_SDO1
-------------------------------------------------------------------------------

            type:  numeric (float)

           range:  [2.2704999,7.3899508]       units:  1.000e-07
```

```
       unique values:  1759                 missing .:  0/2928

                  mean:   4.90476
             std. dev:    1.0548

         percentiles:       10%       25%       50%       75%       90%
                         3.5133    4.0391   4.90112   5.81779    6.2857
```

---

mw_sdo2 MW_SDO2

```
                  type:  numeric (float)

                 range:  [9.384074,41.203602]         units:  1.000e-06
       unique values:  2366                 missing .:  0/2928

                  mean:   25.0985
             std. dev:   5.42026

         percentiles:       10%       25%       50%       75%       90%
                         20.0282   21.3188   23.8462   26.2327   35.5122
```

---

mvar_sdo2 MVAR_SDO2

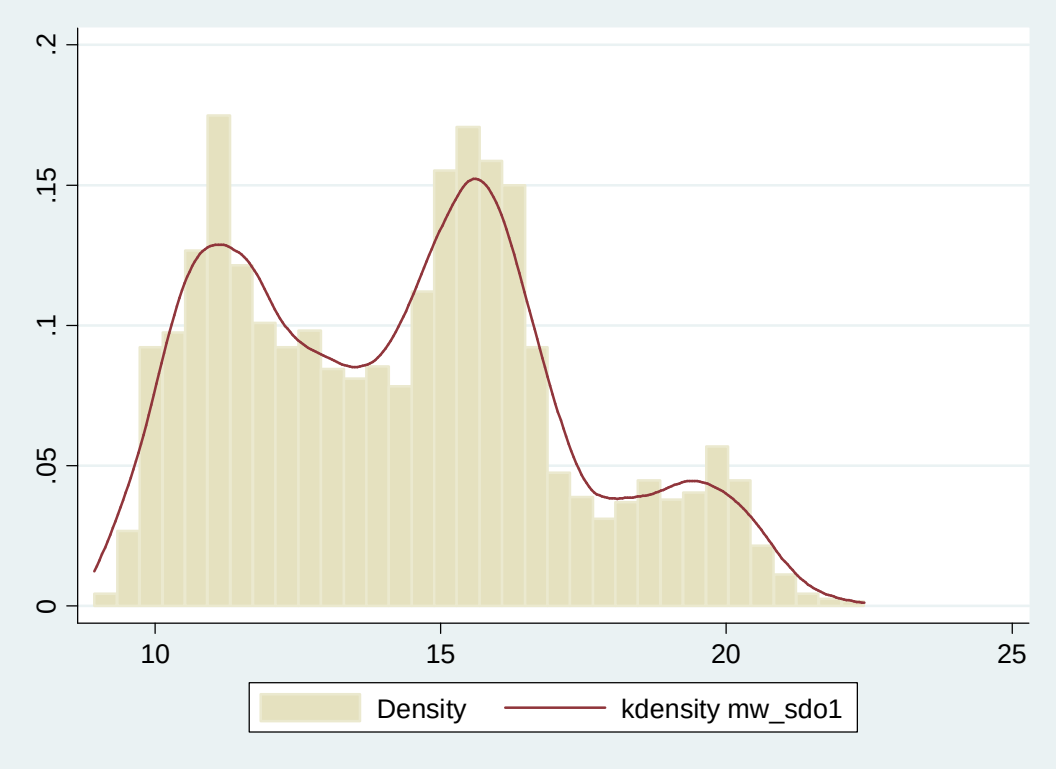
```
                  type:  numeric (float)

                 range:  [-.2868,10.465069]         units:  1.000e-08
       unique values:  2205                 missing .:  0/2928

                  mean:   5.14716
             std. dev:   1.99327

         percentiles:       10%       25%       50%       75%       90%
                         2.84198    3.6806   4.77974   6.73208   7.92297
```

CORRELACIONES

	mw_sdo1	mvar_s~1	mw_sdo2	mvar_s~2	n_dia_~m	hora
mw_sdo1	1.0000					
mvar_sdo1	0.7651	1.0000				
mw_sdo2	0.8724	0.5388	1.0000			
mvar_sdo2	0.7355	0.7951	0.7405	1.0000		
n_dia_sem	0.0701	0.0864	0.1278	0.1991	1.0000	
hora	0.6982	0.5060	0.6530	0.5422	0.0000	1.0000

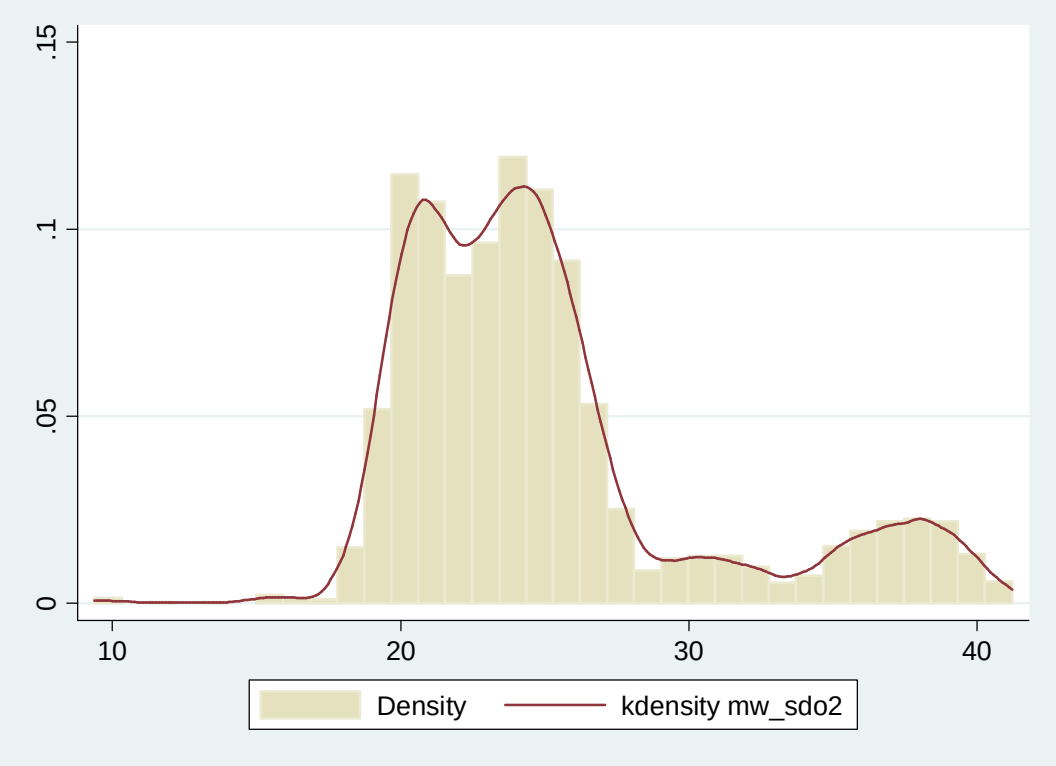
COMANDO STATA:

twoway histogram mw_sdo1, color(*.5) || kdensity mw_sdo1



COMANDO STATA:

twoway histogram mw_sdo2, color(*.5) || kdensity mw_sdo2



COMANDO STATA:

summarize mw_sdo1 mw_sdo2, detail

MW_SDO1

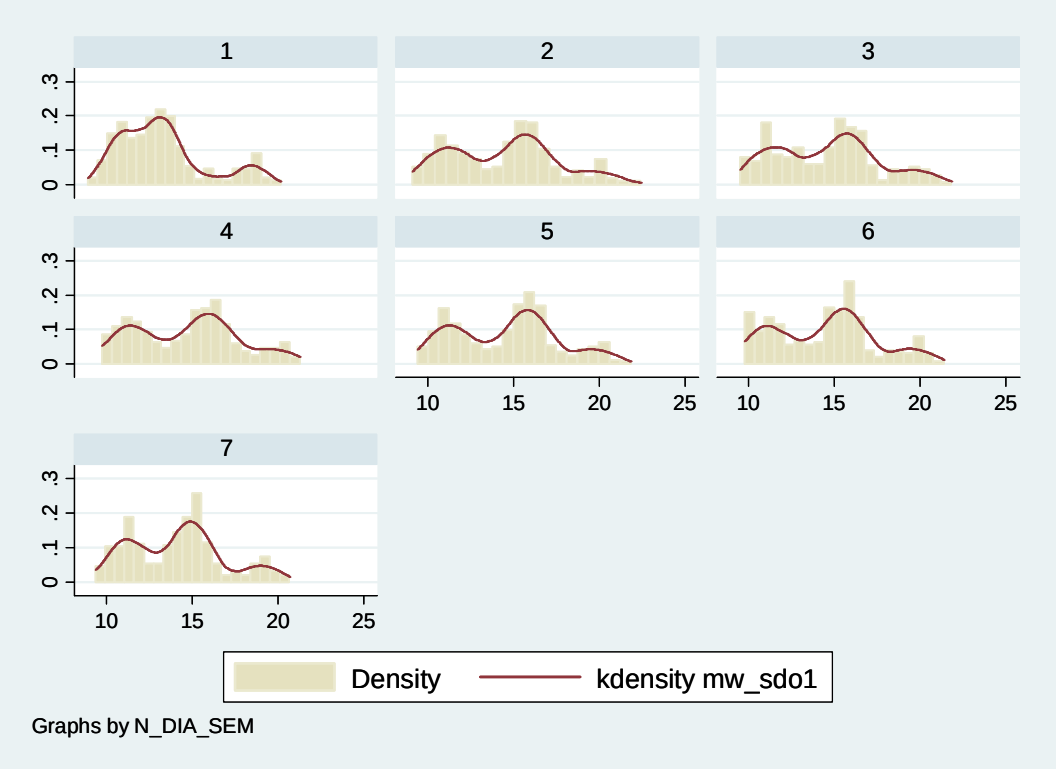
Percentiles		Smallest		
1%	9.679501	8.938601		
5%	10.13347	9.030564		
10%	10.6355	9.1059	Obs	2928
25%	11.66884	9.2493	Sum of Wgt.	2928
50%	14.48228		Mean	14.29873
		Largest	Std. Dev.	2.89956
75%	16.1564	21.8446		
90%	18.55408	21.88853	Variance	8.407447
95%	19.7653	22.0836	Skewness	.3179009
99%	20.66272	22.4182	Kurtosis	2.293675

MW_SDO2

Percentiles		Smallest		
1%	18.30583	9.384074		
5%	19.46461	9.461108		
10%	20.0282	9.509376	Obs	2928
25%	21.3188	9.840275	Sum of Wgt.	2928
50%	23.84623		Mean	25.09846
		Largest	Std. Dev.	5.420262
75%	26.2327	40.82832		
90%	35.51217	40.869	Variance	29.37924
95%	37.9204	41.0124	Skewness	1.274379
99%	39.98156	41.2036	Kurtosis	4.030498

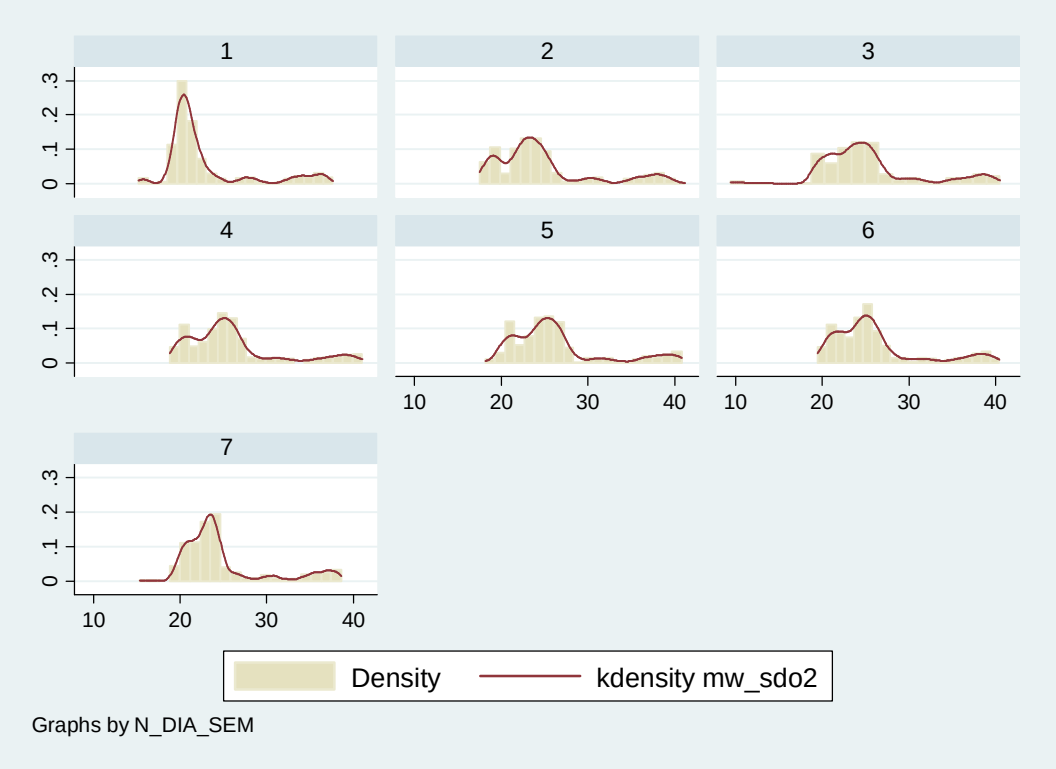
COMANDO STATA:

twoway histogram mw_sdo1, color(*.5) || kdensity mw_sdo1 ||, by (n_dia_sem)



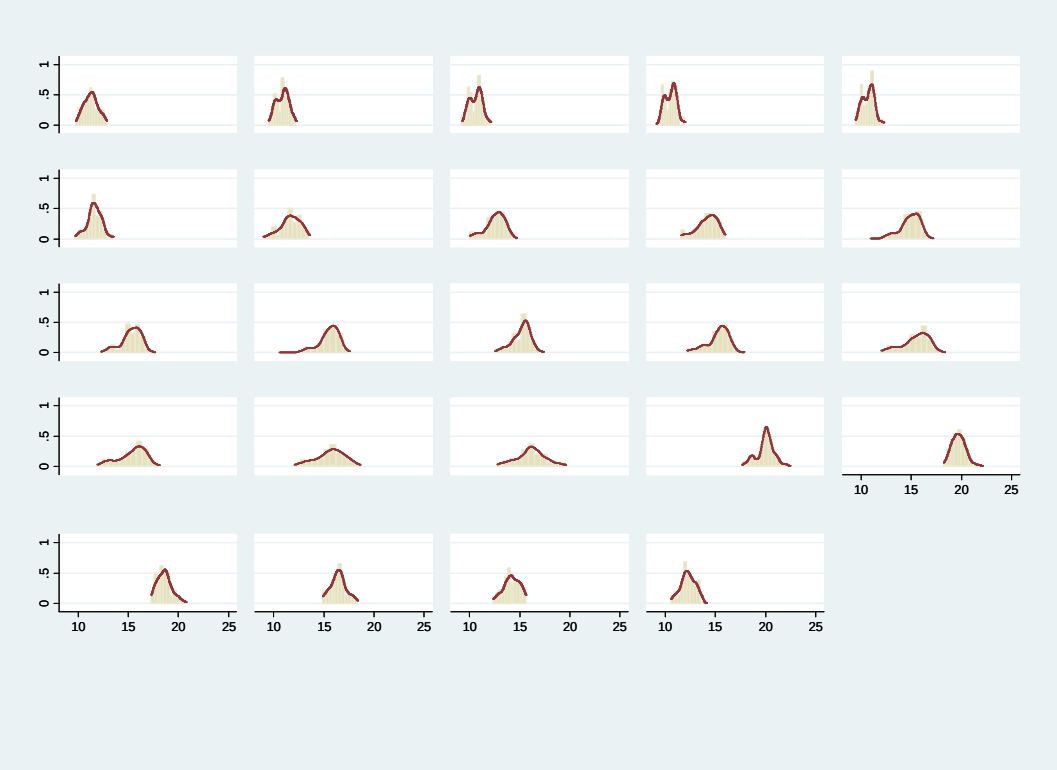
COMANDO STATA:

twoway histogram mw_sdo2, color(*.5) || kdensity mw_sdo2 ||, by (n_dia_sem)



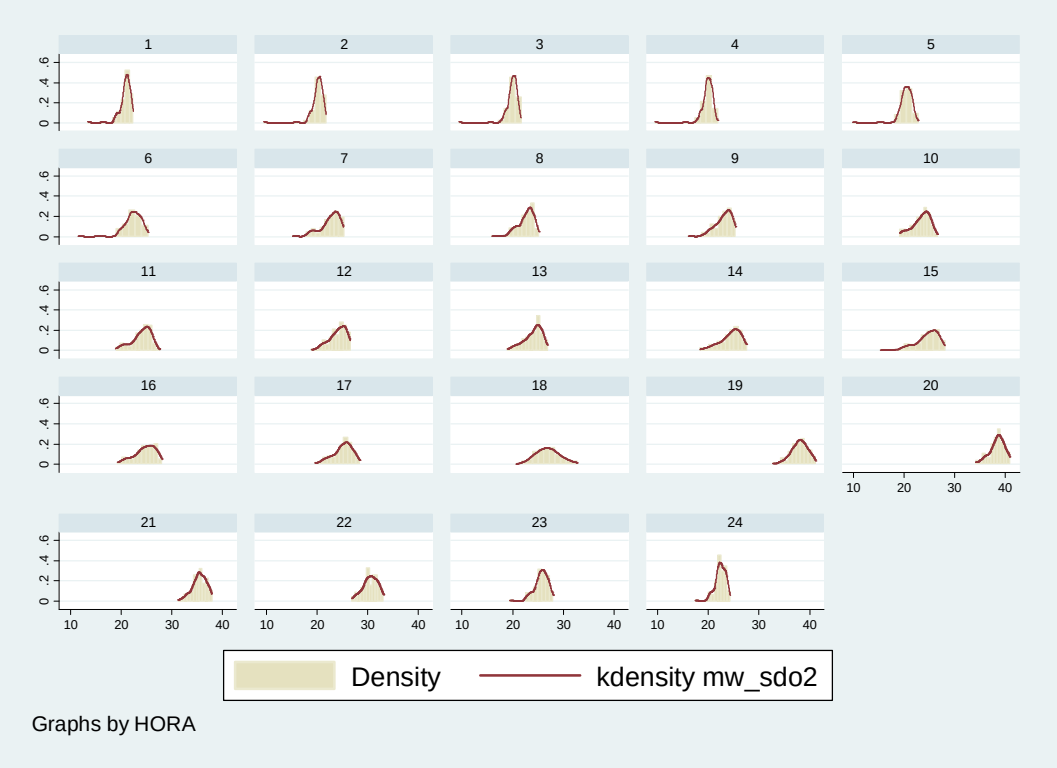
COMANDO STATA:

twoway histogram mw_sdo1, color(*.5) || kdensity mw_sdo1 ||, by (hora)



COMANDO STATA:

twoway histogram mw_sdo2, color(*.5) || kdensity mw_sdo2 ||, by (hora)



ANEXO No. 3
Cálculo de Máxima Verosimilitud para Datos Normales Monótonos a través del Operador de Suavización

El operador de suavización (Beaton,1964) está definido por matrices simétricas de la siguiente manera:

Sea A una matriz simétrica G es decir será el suavizador de la fina y la columna k, esto es reemplazado por otra matriz simétrica H p x p con elementos definidos de la siguiente manera:

$$\begin{aligned} h_{kk} &= -1/g_{kk} \\ h_{jk} &= h_{kj} = g_{jk}/g_{kk} \quad j \neq k \\ h_{jl} &= h_{jl} = g_{jl} - g_{jk}/g_{kk} \quad j \neq kl \neq k \end{aligned} \tag{1}$$

Ilustrando a través de un ejemplo de un caso de una matriz 3 x 3, tenemos:

$$G = \begin{bmatrix} g_{11} & g_{12} & g_{13} \\ g_{12} & g_{22} & g_{23} \\ g_{13} & g_{23} & g_{33} \end{bmatrix}$$

$$H = SWP[1]G = \begin{bmatrix} -1/g_{11} & g_{12}/g_{11} & g_{13}/g_{11} \\ g_{12}/g_{11} & g_{22} - g_{12}^2/g_{11} & g_{23} - g_{13}g_{12}/g_{11} \\ g_{13}/g_{11} & g_{23} - g_{13}g_{12}/g_{11} & g_{33} - g_{13}^2/g_{11} \end{bmatrix}$$

Se usa la notación SWP[k]G para denotar la matriz H definida por (1). Además los resultados de aplicaciones sucesivas del operador sería SWP[k_1], SWP[k_2], ……., SWP[k_t] y para la matriz G sería denotado por SWP[k_1, k_2, ... k_t,]G.

Para efectos de cálculos, el operador de suavización es eficientemente usado si g_{kk} es reemplazo por h_{kk}= -1/$g_{kk.}$ y los elementos g_{jk} y g_{kl} son reemplazados en la fila y columna k por h_{jk}=h_{kj}=-$g_{jk}h_{kk}$, además los g_{jl} se reemplazan por h_{ij} = g_{jl}-$h_{jk}g_{kl.}$

Como propiedades sobresalientes se señala que el operador de suavización:

- Es conmutativo, es decir:

SWP[$_{j,k}$]G=SWP[$_{k,j}$]G

- El orden en el cual un conjunto de suavizadores son colocados, no afecta la respuesta algebraica final, aunque algún orden especifico puede ser computacionalmente más exacto que otro.
- El operador de suavización esta estrechamente ligado a la regresión lineal. Se supone que una muestra de n observaciones y K variables $Y_1,\ldots\ldots\ldots,Y_K$. G es una matriz (K+1) x (K+1), por tanto la matriz G:

$$G=\begin{bmatrix} 1 & \bar{y}_1 & ..... & \bar{y}_j & ..... & \bar{y}_k \\ \bar{y}_1 & n^{-1}\sum y_1^2 & . & . & ..... & n^{-1}\sum y_k y_1 \\ . & . & . & . & . & . \\ . & & & & & \\ \bar{y}_k & . & n^{-1}\sum y_j y_k & . & . & . \\ . & . & . & . & . & . \\ \bar{y}_K & n^{-1}\sum y_1 y_k & . & . & . & n^{-1}\sum y_K^2 \end{bmatrix}$$

donde $\bar{y}_1$.....$\bar{y}_K$ corresponde a la promedio de las muestras, y los sumatorias son sobre las n observaciones. Se aclara que por conveniencia se indexa la fila y la columna desde 0 a K, tanto que la fila y la columna j corresponde a la variable Yj. Suavizaciòn en la finla y columna del campo 0.

$$SWP[0]G=\begin{bmatrix} -1 & \bar{y}_1 & ..... & \bar{y}_j & ..... & \bar{y}_K \\ \bar{y}_1 & s_{11} & ..... & . & ..... & s_{K1} \\ . & . & ..... & . & ..... & . \\ \bar{y}_k & . & ..... & s_{jk} & ..... & . \\ . & . & . & . & . & . \\ \bar{y}_K & s_{K1} & . & . & . & s_{KK} \end{bmatrix} \qquad (2)$$

Donde s_{jk} es la covarianza de la muestra de Y_j y Y_k con factor n^{-1} mayor que $(n-1)^{-1}$. Esta operación corresponde a la corrección del escalar del producto cruz de

Y_1,……Y_K,G para la media de Y_1……Y_k para crear la matriz de covarianza. En términos de regresión, la media en la primera fila y columna de SWP[0]G son los coeficientes de la regresión de Y_1……Y_K en el término constante $Y_o \equiv 1$, y el escalar del producto cruz de la matriz $\{s_{jk}\}$ es la matriz de covarianzas residual desde la regresión. Este punto es llamado como proceso de suavización sobre el término constante. La matriz (2) se denomina matriz de covarianza aumentada de las variables Y_1……Y_K.

La matriz de suavización (2) sobre la fila y columna 1 corresponde a Y1, que es el campo de la matriz simétrica.

SWP[0,1]G=

$$=\begin{bmatrix} -(1+\bar{y}_1^2/s_{11}) & \bar{y}_1/s_{11} & \bar{y}_2-(s_{12}/s_{11})\bar{y}_1 & ..... & \bar{y}_K-(s_{1K}/s_{11})\bar{y}_1 & \bar{y}_K \\ . & -1/s_{11} & s_{12}/s_{11} & ..... & s_{1K}s_{11} & s_{K1} \\ . & . & s_{22}-s_{12}^2/s_{11} & ..... & s_{2K}-s_{1K}s_{12}/s_{11} & . \\ . & . & ..... & ..... & ..... & . \\ . & . & . & . & . & . \\ \bar{y}_K-(s_{1K}/s_{11})\bar{y}_1 & . & . & . & s_{KK}-s_{1K}^2/s_{11} & s_{KK} \end{bmatrix}$$

$$SWP[0,1]G=\begin{bmatrix} -A & B \\ B^T & C \end{bmatrix}$$

donde:

A es una matriz (2 x 2)

B es una matriz 2 x (K – 1)

C es una matriz (K – 1) x (K – 1)

La matriz resultante corresponde a una regresión multivariante de Y_2,……,Y_K sobre Y_1.

La columna j_{th} de B es el intercepto y la pendiente para la regresión de Y_{j+1} sobre Y_1 para j=1…….,K-1.

La matriz C es la matriz de covarianzas residuales de Y_2,……,Y_K dado Y_1.

Los elementos de la matriz A, cuando estos son multiplicados por la varianza residual o covarianza en C y divididos para n, estos dan la varianza y covarianza de los coeficientes de regresión estimados en B.

El termino constante de suavización y los primeros resultados de los q elementos para la regresión multivariante de $Y_{q+1},\ldots\ldots,Y_k$ sobre $Y_1,\ldots\ldots\ldots,Y_q$, sería:

$$SWP[0,1,.....q]G = \begin{bmatrix} -D & E \\ E^T & F \end{bmatrix}$$

donde:

D es una matriz (q +1) x (q+1)

E es una matriz (q + 1) x (K – q)

F es una matriz (K – q) x (K – q)

La columna j_{th} de E es la raíz cuadrada del intercepto y la pendiente de la regresión de Y_{j+q} sobre Y_1 ,……., Y_q para j=1……,K-q.

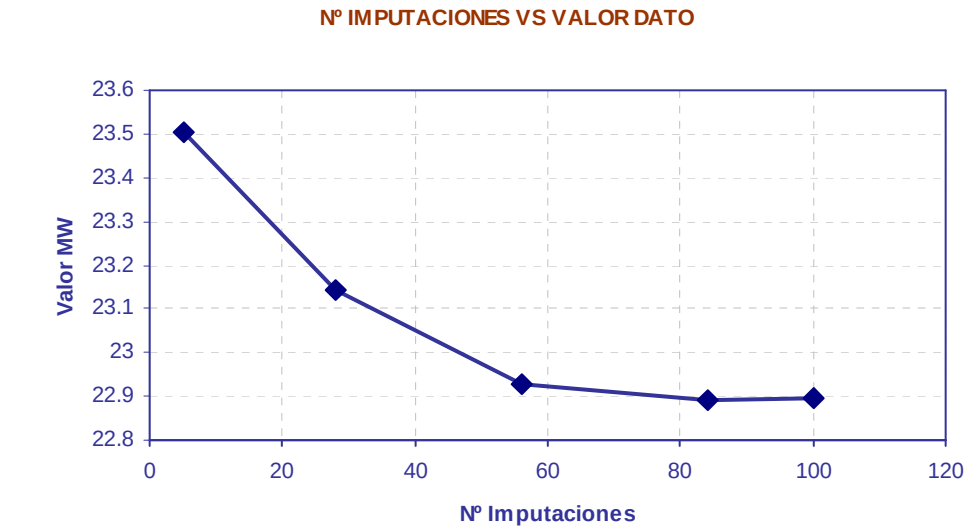
La matriz F es la matriz de covarianzas residuales de $Y_{q+1},\ldots\ldots\ldots\ldots,Y_K$ dado Y_1.

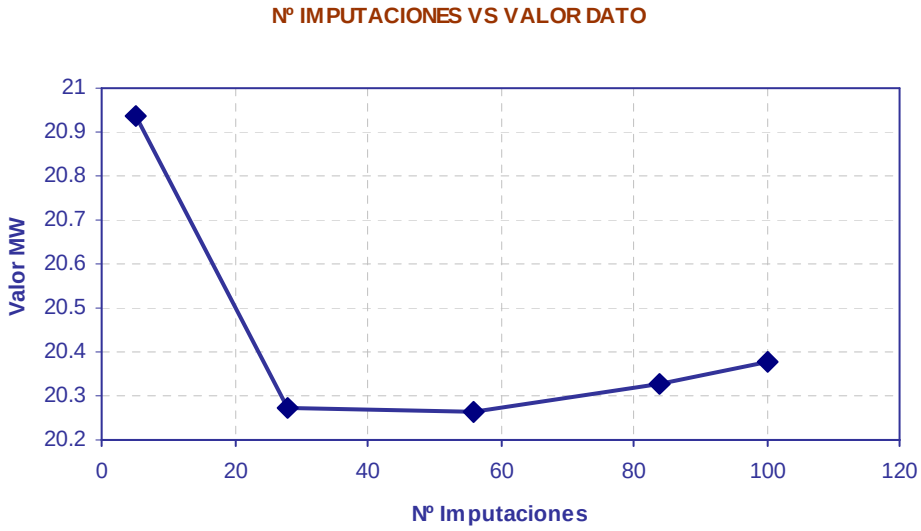
Los elementos de la matriz D, pueden ser empleados para encontrar la covarianza y varianza de los coeficientes de regresiòn estimados en E.

En resumen, el estimador de máxima verosimilitud para regresión lineal multivariante de $Y_{q+1},\ldots\ldots\ldots\ldots,Y_K$ sobre Y_1 ,……., Y_q puede ser encontrado por la suavización de las filas y columnas correspondientes para los términos y las variables del predictor Y_1 ,……., Y_q fuera del producto cruz de la matriz G.

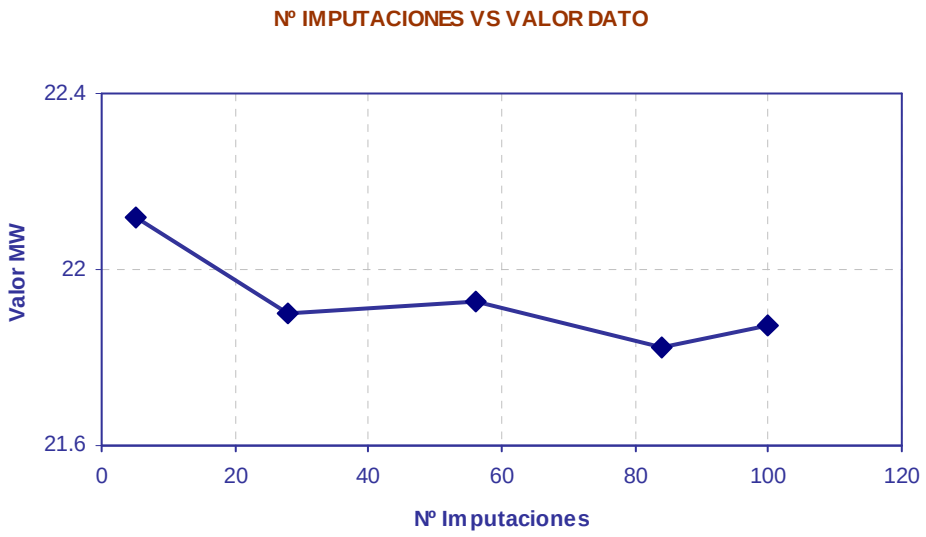
ANEXO No. 4
Comparación gráfica que permite sustentar 100 simulaciones

Por el método de imputación por Hot Deck

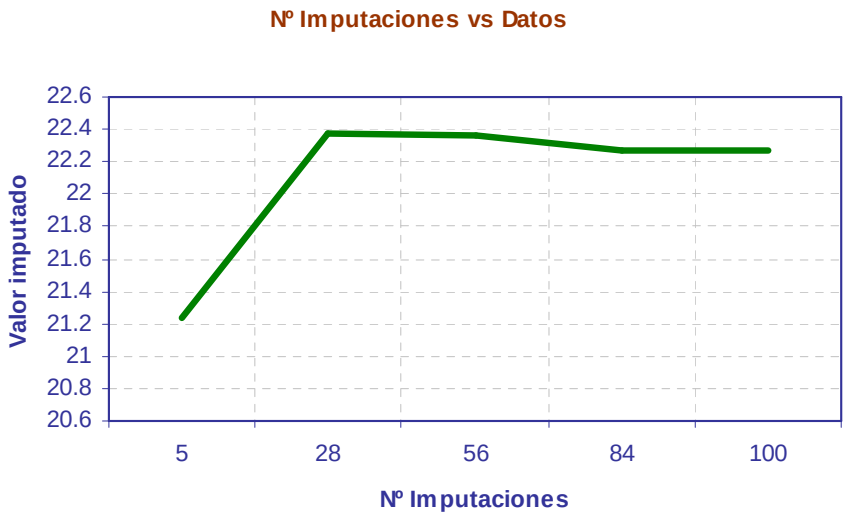


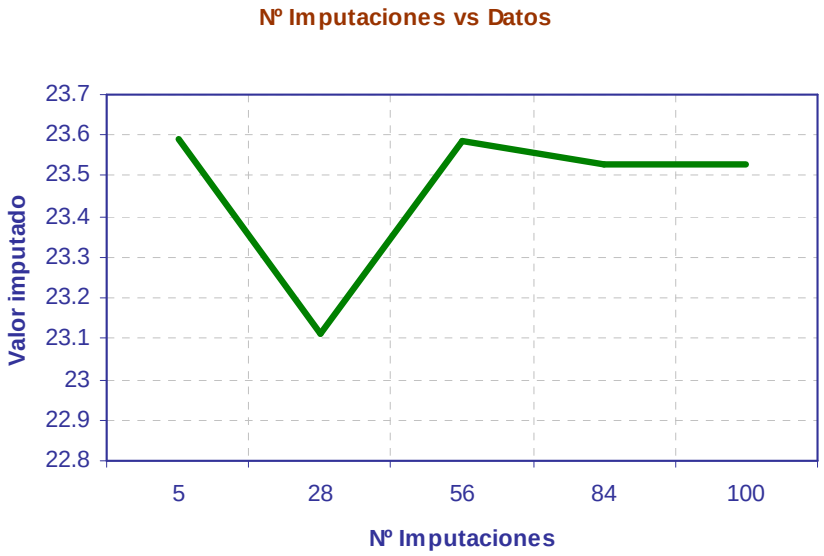


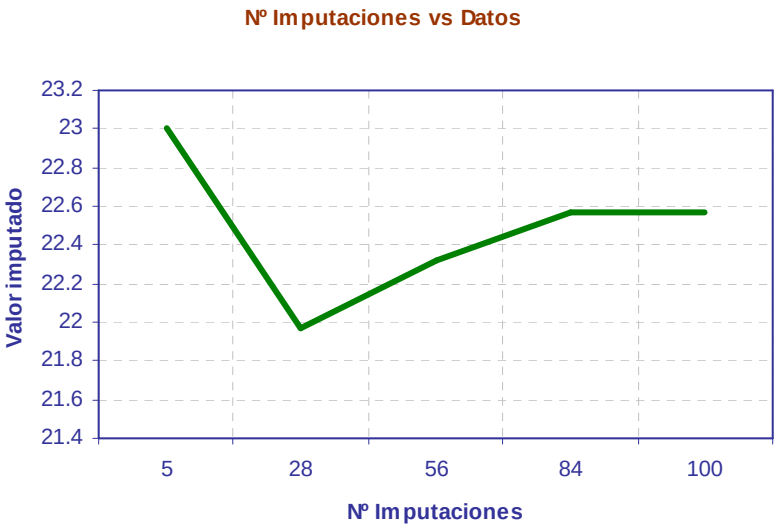
# imput	*dato*	*valor original*
5	20.938	23.30
28	20.271	23.30
56	20.262	23.30
84	20.326	23.30
100	20.379	23.30



Por el método de imputación por Hot Deck con Regresión




Nº Imputaciones vs Datos
Valor imputado
23.7
23.6
23.5
23.4
23.3
23.2
23.1
23
22.9
22.8
5
28
56
84
100
Nº Imputaciones


Nº Imputaciones vs Datos
Valor imputado
23.2
23
22.8
22.6
22.4
22.2
22
21.8
21.6
21.4
5
28
56
84
100
Nº Imputaciones