Tratamiento Estadístico a la Pérdida e Inconsistencia de Datos del Módulo de Registro Histórico del Sistema de Manejo de Energía del Ecuador del Centro Nacional de Control de Energía - CENACE

A. J. Pacheco † H. Capa ‡

† *Corporación Centro Nacional de Control de Energía - CENACE*
‡ *Escuela Politécnica Nacional*

Resumen-- Se revisa el marco conceptual para el análisis de datos faltantes sustentado en métodos de inferencia estadística, se analizan las estimaciones de los métodos de imputación tradicionales (Hot deck), imputación simple, imputación múltiple e interpolación óptima por series de tiempo, cuidando siempre de mantener el tamaño de la muestra y que no condicione la potencia estadística del estudio y a la vez permita controlar posibles sesgos en las series de datos. Además se aprovecha las ventajas que incorpora el programa estadístico -STATA, para conseguir las estimaciones de los métodos señalados y resolver el problema de los datos faltantes en distintos conceptos de potencia activa instantánea de las barras de carga del Sistema Nacional Interconectados del Ecuador. El análisis se fundamenta en datos reales por aplicación de seis métodos de imputación para estudiar qué 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 los datos perdidos pueden ser reemplazados en un 66% a través de métodos de imputación múltiple o simple y 39% por otros métodos como hot deck con Regresión Condicionada, además los datos reemplazados no subestiman la varianza. Se propone que el CENACE estime sus datos a través de métodos de Imputación múltiple, simple e interpolación óptima de series, ya que la matriz de datos refleja su semejanza entre las observaciones de las series de tiempo reales.

Palabras Clave-- **Datos perdidos, Imputación Simple, Múltiple, Series de Tiempo, Sistema Eléctrico del Ecuador, STATA.**

1. INTRODUCCIÓN

En el desarrollo teórico de la mayoría de técnicas y modelos estadísticos se parte de supuestos que no se satisfacen en la práctica. Uno de los más comunes que con seguridad ha enfrentado cualquier analista es el de los datos faltantes, también denominados perdidos o incompletos. Disponer de un archivo de datos completos es obligatorio para cierto tipo de modelaciones (la regresión, por ejemplo); pero al 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 (eliminación de los datos, el método de las medias y el *hot-deck*, por ejemplo). Rubin (1976), propuso un marco conceptual para el análisis de datos faltantes sustentado en métodos de inferencia estadística, posteriormente, la aparición de los métodos de máxima verosimilitud permitió generar estimadores robustos en donde las observaciones faltantes se asumen como variables aleatorias y los datos imputados se generan sin necesidad de ajustar modelos. En el año de 1987, Rubin, introdujo el concepto de imputación múltiple en el que la premisa sustentaba de que cada dato faltante debe ser reemplazado a partir de $m>1$ simulaciones, la aplicación de esta técnica se facilita por los avances computacionales y el desarrollo de métodos bayesianos de simulación (Schafer, 1997) los que se pueden aplicar utilizando paquetes comerciales. Además se revisa estimaciones de datos faltantes a través de series de tiempo.

2. DEFINICIÓN DE IMPUTACIÓN

Imputar significa sustituir observaciones, ya sea porque se carece de información (valores perdidos) o porque se detecta que algunos de los valores recolectados no corresponden con el comportamiento esperado. En esta situación es común que se desee reponer las observaciones y se decida aplicar algún método de sustitución de datos y de imputación.

Para algunos procesos estadísticos como la regresión lineal, análisis de componentes principales, análisis de varianza, etc. se requieren de datos completos y producir algoritmos para estos modelos con datos faltantes puede ser demasiado complicado y costoso.

A continuación se revisarán el patrón de datos perdidos y los métodos de imputación.

3. PÉRDIDA DE DATOS – PATRONES[3]

Interpretando a la base de datos como una matriz, en donde las filas son las unidades de observación y las columnas corresponden a las variables de interés, la elección del método de imputación debería considerar el comportamiento de los datos faltantes, de acuerdo al análisis visual que permite identificar los patrones como se muestra en el Fig. 1.

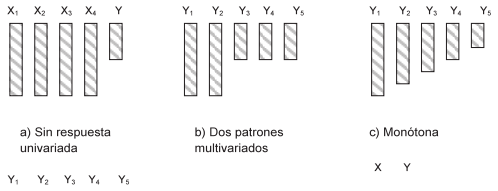

Figura 1: Patrones de ausencia de datos

3.1. MCAR (missing completely at random) – Perdidos completamente al azar[8]

Los datos están perdidos completamente al azar cuando la probabilidad de que el valor de una variable Yj, sea observado para un individuo i no depende ni del valor de esa variable, yij, ni del valor de las demás variables consideradas yij, i≠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))

3.2. MAR (missing at random) – Pérdidos al azar[8]

La ausencia de datos está asociada a variables presentes en la matriz de datos. Por ejemplo, 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)

3.3. NMAR (not missing at random) – Pérdidos no al azar[8]

La hipótesis de datos perdidos no al azar (NMAR) es general y se produce cuando la probabilidad de que un valor yij sea observado depende del propio valor yij, 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 y 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].

4. METODOS DE IMPUTACIÓN

4.1. Imputación por el método de las medias no condicionadas[3]

Se asume que los datos faltantes siguen un patrón MCAR y consiste en la sustitución de los datos utilizando promedios. Su aplicación afecta la distribución de probabilidad de la variable imputada, atenúa la correlación con el resto de variables y subestima la varianza.

En este procedimiento de imputación, el valor medio de la variable se preserva, pero los estadísticos que definen la forma de la distribución como la varianza, covarianza, quantiles, sesgo, kurtosis, etc. se ven afectados.

4.2. Imputación por medias condicionadas para datos agrupados[3]

Una variante del caso anterior consiste en formar categorías a partir de covariables correlacionadas con la variable de interés, y posteriormente imputar los datos faltantes con observaciones provenientes de una submuestra que comparte características comunes. En este procedimiento también se asume que el patrón de datos es MCAR. Además se debe considerar que existirán tantos promedios como categorías se formen los que contribuye a atenuar el sesgo en cada celda pero de ninguna manera a eliminarlo.

4.3. Imputación Hot Deck[3],[8]

Este método tiene como objetivo llenar los registros vacíos (receptores) con información completa (donantes); los datos faltantes se reemplazan a partir de una selección aleatoria de valores observados, los cuales no introducen sesgos a la varianza del estimador. Además el propósito de este método es preservar la distribución de probabilidad de las variables con datos incompletos.

El algoritmo ubica los registros completos e incompletos, identifica características comunes de los donantes y receptores y decide los valores que se emplearán para imputar los datos omitidos. Es

fundamental para la aplicación del procedimiento generar agrupaciones que garanticen que la imputación se llevará a cabo entre observaciones con características comunes, y la selección de los donantes se realiza en forma aleatoria evitando que se introduzca sesgo en el estimador de la varianza.

Existen variantes del procedimiento Hot Deck y una de ellas es el algoritmo secuencial, el que consiste en que éste parte de un proceso de ordenación de los datos en cada subgrupo y selecciona donantes en la medida que recorre el archivo de datos. Otra variante de *Hot Deck*, es el método aleatorio, el que consiste en identificar los registros que no poseen datos y elige en forma estocástica al donante. Además existe la posibilidad que el donante sea el "vecino más cercano" al registro de datos y la selección se efectúa a partir de la definición de criterios de distancia.

4.4. Imputación por Regresión[3]

Si la presencia de los datos faltantes es MCAR, es factible emplear modelos de regresión para imputar información en la variable Y, a partir de covariables (X1,X2,……..Xp) correlacionadas.

Este procedimiento consiste en eliminar las observaciones con datos incompletos y ajustar la ecuación de la regresión para predecir los valores de $\hat{y}$ que permitirá sustituir los valores faltantes, de modo que el valor de $\hat{y}$ se construye como una media condicionada de las covariables X`s.

El presente método no es factible aplicarlo cuando el análisis secundario de datos involucra técnicas de análisis de datos o de correlaciones, pues sobreestima la asociación entre variables, y en modelos de regresión múltiple puede sobredimensionar el valor del coeficiente de determinación R^2.

Una modificación a este procedimiento es la imputación por "regresión estocástica" en donde los datos faltantes se obtienen con un modelo de regresión más un valor aleatorio asociado al término de error, garantizando así la variabilidad de los valores imputados y contribuyendo a reducir el sesgo en la varianza y en el coeficiente de determinación del modelo.

4.5. Imputación por Máxima Verosimilitud[3]

Los métodos de imputación por máxima verosimilitud tienen como objetivo realizar estimaciones máximo verosímiles de los parámetros de una distribución cuando existen datos faltantes. Se asume que los datos faltantes siguen un patrón MAR y que la distribución marginal de los registros observados esta asociada a una función de verosimilitud para un parámetro θ desconocido, siempre que el modelo sea adecuado para el conjunto de datos completos.

Se resume el procedimiento para estimar los parámetros de un modelo utilizando una muestra de datos faltantes, de la siguiente manera:

1. Estimar los parámetros del modelo con los datos completos con la función de máxima verosimilitud.
2. Utilizar los parámetros estimados para predecir los valores omitidos.
3. Sustituir los datos por las predicciones y obtener nuevos valores de los parámetros maximizando la verosimilitud de la muestra completa.
4. Aplicar el algoritmo hasta lograr la convergencia, la que se obtiene cuando el valor de los parámetros no cambia entre dos iteraciones sucesivas.

Un procedimiento eficiente para maximizar la verosimilitud cuando existen datos faltantes es el algoritmo EM, que fue proporcionado por Dempster, Laird y Rubin(1977).

4.6. Algoritmo EM[3]

Se supone una muestra de tamaño n de una variable aleatoria, en la que alguna de la variable tiene datos faltantes. Se asume que los datos faltantes se generan al azar. Las dos situaciones de valores faltantes son: i) algunos elementos de la muestra están completos (x1,……,xn) y otros no tienen datos (x n+1,……,xm) y ii) algunas variables no tienen datos.

Se supone que se trabaja con una matriz de datos Y=(y1,…..,yn), donde yi es un vector de dimensión p1*1, y un conjunto de datos ausentes Z=(z1,……,zm), con zi un vector de dimensión p2*1 y el problema consiste en estimar el vector de parámetros q con la información disponible.

La función de distribución conjunta de las variables (Y,Z) se escribe como: $f(Y,Z|\theta=f(Z|Y,\theta)f(Y|\theta)$, por lo que se tiene que el $\log f(Y|\theta)=\log f(Y,Z|\theta)-\log f(Z|Y,\theta)$. En el procedimiento de máxima verosimilitud, el primer miembro de la expresión $\log(f(Y|\theta)$ corresponde a la función de datos observados, cuya maximización en θ genera el estimador de máxima verosimilitud, en cambio el término $\log f(Y,Z|\theta)$ es la función que se hubiese observado con la muestra completa y $f(Z|Y,\theta)$ proporciona la densidad de los datos ausentes, siempre que se conozca la muestra y el vector de parámetros θ.

Por tanto, la función de verosimilitud es $L(\theta|Y)=Q(\theta|Y,Z) - \log f(Z|Y,\theta)$. El algoritmo EM es un

procedimiento iterativo para encontrar el estimador de máxima verosimilitud de q, utilizando la función Q(Θ|Y,Z). La aplicación del algoritmo se logra ejecutando los siguientes pasos:

1. Paso E (predicción) del algoritmo EM, Se calcula $\hat{\theta}^{(i)}$ cuando i=1, a través de la esperanza matemática de las funciones de los valores perdidos que aparecen en la función de verosimilitud completa, por su esperanza condicionada Q(Θ|Y,Z) con respecto a la distribución Z dados los valores de $\hat{\theta}^{(i)}$ y los datos observados Y.
2. El paso M (maximización), Maximiza la función Q(Θ|Y,Z) con respecto a q. Este paso equivale a maximizar la verosimilitud completa donde se han sustituido las observaciones faltantes por estimadores.

Con el valor obtenido en el paso M ($\hat{\theta}^{(i+1)}$), se vuelve a ejecutar el paso E, y se itera hasta lograr la convergencia, es decir, hasta que la diferencia $\left\|\hat{\theta}^{(i+1)} - \hat{\theta}^{(i)}\right\|$ sea suficientemente pequeña.

4.7. Imputación Múltiple[2],[3],[6]

La imputación múltiple utiliza métodos de Monte Carlo y sustituye los datos faltantes a partir de un número (m>1) de simulaciones. La metodología consta de varias etapas, y en cada simulación se analiza la matriz de datos completos a partir de métodos estadístico convencionales y posteriormente se combinan los resultados para generar estimadores robustos, su error estándar e intervalos de confianza.

El procedimiento de imputación múltiple propuesto por Rubin [2] se describe a continuación y se lo representa en el Fig. 2.

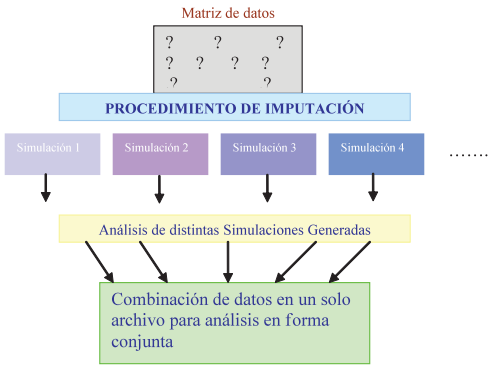

Figura 2: Esquema de la imputación múltiple

Sea Q una variable aleatoria y se supone que se desea estimar la media, la varianza o su coeficiente de correlación con otras variables. Además se considera a X la matriz de los datos disponibles que se encuentra compuesta como Y=(Yobs,Yaus), (Yobs,Yaus) es valor del estimador Q que se genera a partir de los datos y U=U(Yobs,Yaus) el error estándar de $\hat{e}$. Para el conjunto de datos completos se los estandariza, es decir.

$$(\hat{Q} - Q)/\sqrt{U} \approx N(0,1)$$

Cuando no existen datos faltantes, y considerando que se dispone de m>1 simulaciones independientes de datos imputados Y(1)aus,……., Y(m)aus, entonces se calcula el valor de estimador $\hat{Q}^{(i)} = \hat{Q}(Y_{obs}, Y_{aus}^{(i)})$ y sus respectivos errores $U^{(i)} = U(Y_{obs}, Y_{aus}^{(i)})$ i=1,……m. El estimador Q es el promedio de los estimadores $\overline{Q} = m^{-1} \sum \hat{Q}^{(i)}$.

El error estándar de $\overline{Q}$ se calcula a partir de la varianza entre las distintas imputaciones $B = (m-1)^{-1} \sum (\hat{Q}^{(i)} - \overline{Q})^2$ y debido a que la varianza de cada una de las imputaciones es $\overline{U} = m^{-1} \sum U^{(i)}$, el estimador de la varianza total sería

$$T = (1 + m^{-1})B + \overline{U}.$$

La prueba de hipótesis y los intervalos de confianza se construyen a partir de una aproximación a la t de student por medio de $(\overline{Q} - Q)/\sqrt{T} \approx t_v$, donde los grados de libertad se determina por medio de

$$v = (m-1)\left[\overline{U}/(1 + m^{-1})B\right]^2.$$

El incremento relativo de la varianza debido a la presencia de datos faltantes a través de $r = (1 - m^{-1})B/\overline{U}$, y la tasa de datos faltante se aproxima a

$$\lambda = r/(1+r)$$

ó

$$\lambda = \left[(r+2)/(v+3)\right]/(1+r).$$

El procedimiento de Rubin cuenta con los siguientes supuestos principales:

1. El patrón de datos faltantes es MAR, es decir que la probabilidad de que existan datos faltantes en la variable Y depende de otros variables pero no de Y.
2. El modelo (estadístico o econométrico) empleado para generar los datos imputados debe ser apropiado, es decir que, exista correlación alta entre la variable a imputar y el vector de covariables que se utiliza para modelar los datos que se utilizarán reemplazando los faltantes.

4.8. Imputación Múltiple Markov Chain Monte Carlo (MCMC)[5]

El procedimiento MCMC es una colección de de

procesos de simulación generados por métodos de selección aleatoria mediante cadenas de Markov, y es uno de los procedimientos que se consideran más adecuados para generar imputaciones cuando se está en presencia de problemas de estimación no triviales.

El método MCMC se aplica para explorar la distribución posterior en inferencia bayesiana. 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:

1. 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.
2. Distribución Posterior: Concluida la simulación del primer paso, se obtiene el vector de la media de la población y de la matriz de covarianza de la muestra completa. Entonces estas nuevas estimaciones son usadas en el primer paso.

Finalmente se realizan 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-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: $\left(Y_{per}^{(1)}, \theta^{(1)}\right), (Y_{per}^{(2)}, \theta^{(2)}), \ldots\ldots\ldots\ldots\right)$
La que converge a la distribución $p\left(Y_{per}, \theta \mid Y_{obs}\right)$

El resultado de la estimación EM provee un buen valor inicial para comenzar el proceso MCMC.

4.9. Interpolación óptima y función de auto correlación inversa para series de tiempo de una variable con valores perdidos[9]

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:

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

donde δi 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 como un proceso ARMA:

$$\theta(B) z_t = \phi(B) \nabla^d a_t \quad (2)$$

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

$$z_t = \sum_{i=1}^{\infty} \pi_i z_{t-i} + a_t \quad (3)$$

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

$$\hat{z}_T^{(0)} = \sum_{i=1}^{\infty} \pi_i z_{t-i} \quad (4)$$

y su estimado, el que se construye con 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 todo j, tal que pj ¹ 0:

$$z_T = \pi_j^{-1} (z_{T+j} - \sum_{\substack{i=j \\ i \neq j}}^{\infty} \pi_i z_{T+t-i}) - \frac{a_{T+j}}{\pi_j} \quad (5)$$

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

$$\hat{z}_T^{(j)} = \pi_j^{-1} (z_{T+j} - \sum_{i \neq j} \pi_i z_{T+t-i}) \quad (6)$$

con varianza σ_a^2 / π_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:

$$\hat{z}_T = \sum_{j=0}^{\infty} \left(\pi_j^2 \Big/ \sum \pi_j^2 \right) \hat{z}_T^{(j)} \quad (7)$$

donde π_0 =-1.

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

$$\hat{z}_{T,F} = \sum_{j=0}^{n-T} \frac{\pi_j^2}{\sum_0^n \pi_i^2} \quad (8)$$

Los procedimientos para cálculo de los valores perdidos en series de tiempo se lo resumen así:

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.

El procedimiento es iterativo hasta cuando las series hayan sido completadas por el interpolador óptimo. 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.

5. APLICACIÓN AL SISTEMA NACIONAL INTERCONECTADO DEL ECUADOR

CENACE cuenta con un sistema de última generación que le permite realizar la supervisión en tiempo real del sistema eléctrico del Ecuador a través del EMS, el cual adquiere la información proveniente desde el campo a través de las unidades terminales remotas (RTU) de las subestaciones de transmisión y generación del país y que posteriormente por el sistema de comunicaciones se transmiten datos hacia los Centros de Control de CENACE y CELEC-TRANSELECTRIC.

Este procedimiento de recopilación de los datos es vulnerable a sufrir daños que pueden ocurrir tanto en las RTUs, sistema de comunicaciones o 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 y ello implique contar con matrices de datos incompletas especialmente potencias activas instantáneas de las barras de carga del Sistema Nacional Interconectado del Ecuador, lo que dificulta la preparación de la información para los procesos técnicos y comerciales que ejecuta CENACE.

Por esta razón, se analiza el problema de los 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.

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 meses anteriores; es decir, la base de datos completa corresponde a los meses de julio hasta agosto del 2007.

En la primera fase de este trabajo se cuantifica la falta de respuesta 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, un algoritmo de imputación múltiple y reemplazo de dato por series históricas.

A continuación se presenta el proceso de análisis efectuado para la Empresa Eléctrica Ambato y la barra Totoras, a través del programa STATA y sus resultados.

Tabla 1: 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 aleatoriamente.

Debido a que en los métodos hot deck, hot deck con regresión, imputación múltiple se requiere incluir el número de imputaciones, se realiza un grupo de simulaciones que permiten obtener el número de simulaciones que son requeridas para la estimación del valor. A continuación se presentan las figuras:

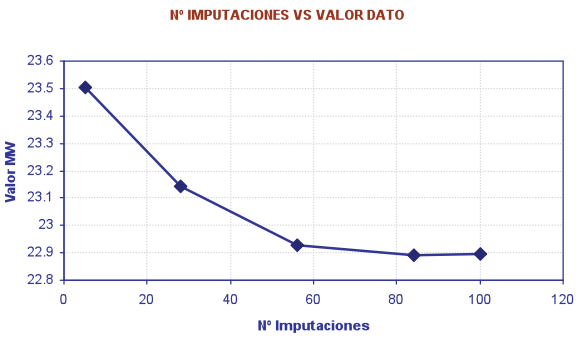

Figura 3: Número de imputaciones por método Hot Deck

Tabla 2: Imputaciones método de hot deck a la variable potencia activa de la posición Montalvo

# imput	Simulación	Valor Original
5	23,502	23,554
28	23,144	23,554
56	22,926	23,554
84	22,891	23,554
100	22,895	23,554

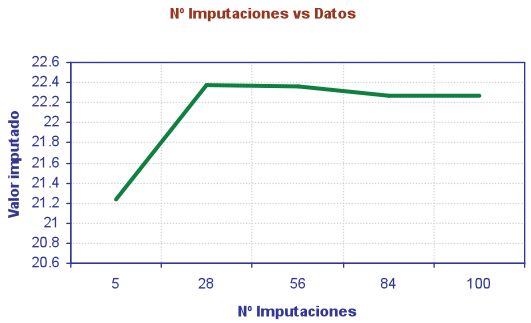

Figura 4: Número de imputaciones por método Hot Deck con regresión

Tabla 3: Imputaciones método de hot deck con regresión a la variable potencia activa de la posición Montalvo

# imput	dato	valor original
5	21,239	23,553
28	22,379	23,553
56	22,365	23,553
84	22,271	23,553
100	22,271	23,553

Los gráficos y Tablas presentados permiten concluir que las 100 simulaciones son apropiadas para estimación del dato, pues a partir de este número el valor estimado es constante y no existen variaciones significativas; cumpliendo por tanto, con lo establecido en la Ley de los Grandes Números.

A continuación se presentan los datos arrojados por el programa STATA con los diferentes métodos:

Tabla 4: Aplicación del método hot deck a la variable potencia activa de la posición Montalvo

AÑO	MES	DIA	HORA	mw_hd
2007	9	3	8	21,725
2007	9	3	12	22,895
2007	9	5	12	20,761
2007	9	5	13	19,888
2007	9	5	14	20,379
2007	9	12	13	21,040
2007	9	12	14	21,588
2007	9	13	13	21,872
2007	9	13	16	22,678

Tabla 5: Aplicación del método hot deck con regresión a la variable potencia activa de la posición Montalvo

AÑO	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

Tabla 6: Aplicación del método regresión condicionada a la variable potencia activa de la posición Montalvo

AÑO	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

Tabla 7: Aplicación del método imputación simple a la variable potencia activa de la posición Montalvo

AÑO	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

Tabla 8: Aplicación del método imputación múltiple a la variable potencia activa de la posición Montalvo

AÑO	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

Tabla 9: Aplicación del método interpolación a series de tiempo a la variable potencia activa de la posición Montalvo

AÑO	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

5.1. Análisis de los Resultados

Los seis métodos de imputación propuestos se aplicaron para sustituir a los datos perdidos en la variable potencia activa instantánea de la posición Montalvo en la Empresa Eléctrica Ambato. Estos datos estimados se comparan con los datos reales que fueron registrados por CENACE 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.

Se debe mencionar que la base de datos SADYR tuvo vigencia hasta febrero del 2008, y a partir de esta fecha la base del histórico del EMS tiene vigencia como base de datos única en la Corporación CENACE. Adicionalmente se señala, que no es factible mantener las dos bases de datos, debido a que la base de datos SADYR requiere del esfuerzo de los operadores del Centro de Control para realizar el ingreso manual hora a hora, en cambio la nueva base, almacena los datos registrados en campo directamente en el Histórico.

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:

Tabla 10: Resultado de los errores presentados en la 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.

6. CONCLUSIONES

- 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.
- 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 ver Tabla 11.

Tabla 11: Barra de carga del Sistema Nacional Interconectado

EMPRESA	BARRA	POSICIÓN	MÉTODO 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

- 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.

- 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.
- 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 garantiza variabilidad en los valores imputados y contribuye a reducir el sesgo en la varianza.
- 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.
- 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.
- 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.
- 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.
- 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.

7. REFERENCIAS BIBLIOGRÁFICAS

[1] Pacheco, Adriana (2009) "Imputación estadística: Una aplicación al Sistema Nacional Interconectado del Ecuador" Tesis de Maestría Escuela Politécnica Nacional.

[2] Rubin D.(2002), Multiple imputation for nonresponse in surveys, Wiley Classics Library, 2004

[3] Little, Roderick J.A., y Donald B. Rubin. (2002). Statistical Analysis With Missing Data. Segunda Edición. New Yersey. John Wiley & Sons, Inc.

[4] Von Hippel P.(2004), Biases in SPSS 12.0 Missing Value Analysis, The American Statistician, Vol. 58, No. 2.

[5] Carlin, Bradley P., y Thomas A. Louis. (2000). Bayes and Empirical Bayes Methods for Data Analysis. . Segunda Edición. Florida. Chapman & Hall./CRC

[6] http:/www.multiple-imputation.com/

[7] Rodríguez, Gioconda, y Juan Vallecilla. (2008). "Adquisición de Datos en el Sistema de Manejo de Energía, Network Manager". Revista Técnica "energía" Edición No. 4.

[8] Giocoechea, Aitor (2002). "Imputación basada en árboles de clasificación". Eustat.

[9] Peña Daniel, Teno George, Tsay Ruey. (2001). "A course in time series analysis". John Wiley & Sons.

[10] StataCorp LP. (2005). Stata Documentation Version 9 – Data Management. New York.

Hollger Capa Santos.- Nació en Paltas, Ecuador, en 1955. Recibió su título de Matemático (1979) y de Magister en Gerencia Empresarial (MBA,1995) en la Escuela Politécnica Nacional de Quito, Ecuador; y su título de Doctor en Estadística en la Universidad Pierre y Marie Curie de París, Francia. Sus principales campos de investigación están relacionados con la Econometría, el Análisis de Riesgo y el Análisis Bayesiano.

Adriana Janet Pacheco Toscano.- Ingeniera Eléctrica de la Escuela Politécnica Nacional en 1996; y Master en Estadística Aplicada de la Escuela Politécnica Nacional en el 2009. Ha desempeñado sus labores profesionales en la fábrica de transformadores ECUATRAN S.A. como Ingeniera de Investigación y Desarrollo y en el CENACE en el Área de Estudios Eléctricos de la Dirección de Planeamiento. Actualmente se desempeña en la Dirección de Operaciones en el Área de Análisis de la Operación. Sus campos de investigación esta relacionados con el control estadístico de procesos, análisis de datos perdidos, confiabilidad de sistemas eléctricos de potencia, análisis de series de tiempo.