VI Congreso Galego de Estatística e Investigación de Operacións
Vigo 5-7 de Novembro de 2003

METODOLOGÍAS DE ANÁLISIS Y IMPUTACIÓN DE DATOS FALTANTES EN SERIES DE VELOCIDAD DEL VIENTO

Armando Mateus Ferreira [1]

[1] Escola Superior Agrária
Instituto Politécnico de Castelo Branco, Portugal

RESUMO

En este trabajo se pretende estudiar diversos métodos de análisis y imputación de datos no completos, bajo el enfoque de su aplicación para completar los valores faltantes en series de velocidad del viento. Se ha procedido a la imputación de los datos faltantes con un procedimiento basado en la metodología hot-deck secuencial.

Palabras chave: datos faltantes, EM, hot-deck, Weibull

1 - INTRODUCCIÓN

Los datos faltantes son un problema familiar en todas las áreas de investigación. Si los datos son recolectados y almacenados por alguno sistema automático, como es el caso de la recogida de datos en las actuales estaciones meteorológicas, pueden ocurrir fallos en el automatismo, ocasionados por la falta de energía, ó por algún tipo de mala configuración del mecanismo. Estos tipos de falta de datos que suelen escapar al controlo del investigador, si bien que inevitables, no son deseables.

A fin de definir los años-tipo, se necesitan de muestras completas de datos. Así, hay que de algún modo, intentar completar las bases de datos, si no con los datos reales, pero al menos con datos generados por alguna metodología estadística que garantice la aleatoriedad de los datos imputados, y que se ajuste a la estructura del proceso estocástico subyacente.

Kline y Milligan (1998) han simulado series incompletas con la eliminación aleatoria de 10% y 20% de los datos de velocidad del viento, en rachas de 4, 8, 12, 16, 20 y 24 datos, y han estimado los datos faltantes por diversas metodologías (matrices de transición de Markov, análisis de la tendencia – aumento ó disminución de la velocidad de una hora a la siguiente, y modelo autorregresivo), y han concluido que las técnicas ensayadas han conducido a resultados muy parecidos a la serie original. Los errores en las estimaciones de la velocidad media han sido inferiores a unos 5%, y los errores en las estimaciones teóricas de la energía producida por un aerogenerador han sido aún menores, de la orden de unos 2% o menos.

2 - GENERALIDADES

En general se considera una matriz de datos $\mathbf{X}$, que en las bases de datos del viento contiene las observaciones de *K* años y *N* casos (que suelen ser los registros horarios, de 3 en 3 horas, ó registros diarios):

$$\mathbf{X} = \left[x_{ij} \right] \quad i = 1, 2, ..., N \quad j = 1, 2, ..., K$$

Cuando se trabaja con datos no completos es importante saber si el mecanismo que genera la ausencia de datos es aleatorio o no. Este mecanismo se clasifica en tres tipos (Little y Rubin, 1987; Rubin, 1987; Schafer, 2000; Schafer y Olsen, 1998):

- Completamente aleatorio (MCAR: "*Missing Completely at Random*"): si la probabilidad que el valor de una variable X_j sea observado para un caso *i* no depende del valor de esa variable, x_{ij}, ni del valor de las demás variables consideradas, x_{ik}, $k \neq j$.
- Aleatorio (MAR: "*Missing at Random*"): si la probabilidad que el valor de una variable X_j sea observado para un caso *i* no depende del valor de esa variable, x_{ij}, pero quizá sí del que toma alguna otra variable observada x_{ik}, $k \neq j$.
- No aleatorio: si la probabilidad de que un valor x_{ij} sea observado depende del propio valor x_{ij}.

En general, los métodos de análisis de datos no completos suponen al menos que el mecanismo es MAR. Este mecanismo es el presupuesto formal que permite estimar la relación entre variables a partir de dos datos observados, y enseguida usar esta relación para obtener predicciones no sesgadas para los valores faltantes a partir de los datos observados .

3 - MÉTODOS DE IMPUTACIÓN DE DATOS FALTANTES

- **Imputación mediante cadenas de Markov**

La matriz de transición de Markov es fundamentada en el principio de que de una hora (ó otra periodicidad de recogida de datos) a la siguiente, la velocidad del viento tendrá un comportamiento de transición regular y determinado por un proceso probabilístico subyacente. Esta metodología implica la creación de una matriz cuadrada $n \times n$, siendo *n* el número de categorías en que se clasifica ó agrupan las velocidades del viento, que define las probabilidades de transición de velocidad del viento de una a otra categoría. Cuando se encuentra un dado faltante, este es estimado mediante las probabilidades de transición entre estados ó categorías de velocidad, definidas en la matriz referida. Esta metodología, frecuentemente usada en estudios de energía solar (Aguiar *et al*., 1988; Aguiar y Collares-Pereira, 1992; Fernández-Alvares, 2001) también es referida en diversos estudios de energía eólica (Castino *et al.*, 1998; Kline y Milligan, 1998; Sparis *et al.*, 1995).

- **Imputación mediante medias incondicionales**

Consiste en rellenar un dato no observado x_{ij} por la media aritmética simple de los valores observados de la variable, $\overline{x}_j$. Este método, aunque es extensamente utilizada, tiene diversas desventajas: invalida las estimaciones de la varianza derivadas de las fórmulas estándar de la varianza para conocer la verdadera varianza de los datos; la distribución real de los datos resulta distorsionada por la sustitución de los datos ausentes por la media; y modifica la correlación observada porque todos los datos ausentes tendrán un único valor constante.

- **Procedimientos de regresión**

Se incluyen en este conjunto de procedimientos de imputación aquellos que asignan a los campos a imputar el valor de predicción de modelos del tipo:

$$x_m = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

donde x_m es la variable que contiene campos o casos a imputar y las variables $\{x_j; j \neq m\}$ son los regresores. Estos regresores son las variables de control u otras variables altamente correlacionadas con la dependiente. En este tipo de modelos, además de los supuestos de aleatoriedad *MAR* descritos con anterioridad, se añaden los supuestos sobre el término de error aleatorio ε (García-Lopez, 2000).

- **Imputación por el método de Buck**

Consiste en estimar los datos faltantes en una muestra *K*-variante, mediante técnicas de regresión y calcular la matriz de covarianzas usando los datos así completados (Bárcena, 2000; Buck, 1960; Beale y Little, 1975). Sea *m* el número de casos para los cuales hay datos completos de las *K* variables; estos datos pueden reagruparse de modo que los casos completos formen las primeras *m* filas de la matriz de datos **X**, y los *N-m* casos en los cuales hay datos faltantes para algunas de las variables formen las *N-m* filas restantes. El conjunto de valores para un individuo *i* cualquiera de los *n-m* restantes, $\mathbf{x}_i$, puede reordenarse formando dos grupos, donde $\mathbf{x}_{i(1)}$ y $\mathbf{x}_{i(2)}$ denotan el conjunto de valores observados e ausentes, respectivamente:

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{x}_{i(1)} \\ \mathbf{x}_{i(2)} \end{pmatrix}$$

Este método consiste en usar el vector de medias y la matriz de covarianzas calculadas con las *m* filas completas para obtener una estimación por regresión lineal por mínimos cuadrados de los valores ausentes $\mathbf{x}_{i(2)}$ en función de los valores observados $\mathbf{x}_{i(1)}$.

- **Imputación hot-deck**

El procedimiento hot-deck consiste en imputar el valor no observado x_{ij} de una variable X_j para el caso ó individuo *i* mediante el valor que esa variable toma en otro caso semejante a él, escogido de entre aquellos para los que se dispone de valores de X_j (Bailar y Bailar, 1978, 1983; Ford, 1983; Little y Rubin, 1976; Rubin, 1987; Sande, 1983). Una ventaja de la imputación hot-deck es que los valores imputados no sufren la pérdida de variabilidad que tienen los valores imputados los demás métodos. Bajo la hipótesis que el mecanismo que genera la ausencia de datos es MAR ó MCAR, las estimaciones de la media e varianza son no sesgadas. Otra ventaja es que el método no necesita de fuertes presupuestos matemáticos para la estimación de los valores faltantes. El mayor inconveniente de algunos métodos de imputación hot-deck es que requieren tomar decisiones subjetivas (por ejemplo, cuáles son las variables más relevantes para determinar la semejanza a fin de formar los grupos homogéneos) que afectan la calidad de los resultados. La mayoría de los métodos de imputación hot-deck que se usan en la actualidad se fundamentan más en el sentido común que en el desarrollo teórico riguroso, donde la efectividad des estos procedimientos es a menudo desconocida.

- **El algoritmo EM**

El algoritmo EM es una técnica general para ajustar modelos en situaciones de datos ausentes. El algoritmo conjuga la relación entre los datos ausentes y los parámetros desconocidos del modelo. Si uno conoce los datos faltantes, la estimación de los parámetros del modelo sería inmediata; de modo similar, si se conocen los parámetros del modelo, entonces es posible obtener preediciones no sesgadas de los datos ausentes. Esta interdependencia entre los parámetros del modelo y los datos no observados sugiere un método iterativo donde se empezará por remplazar los datos ausentes por unos valores estimados y a continuación proceder a una primera estimación de los parámetros; enseguida se usarán los parámetros así estimados para proceder a una nueva estimación de los datos faltantes, que serán posteriormente utilizados para obtener unas estimaciones de los parámetros, y así sucesivamente hasta que la convergencia de los parámetros sea aceptable (Little, 1982; Little y Rubin, 1987; Schafer y Olsen, 1998). La designación EM ha sido introducida por Dempster *et al.* (1977), en un trabajo donde los autores exponen el desarrollo teórico del algoritmo, sus propiedades, tales como que en cada iteración incrementa el valor de la función de verosimilitud $l\left(\theta \mid X_{obs}\right)$ y en que presentan algunos ejemplos de aplicación.

4 - APLICACIÓN DEL PROCEDIMIENTO HOT-DECK

En este apartado hemos seguido la siguiente orientación: a partir de una serie completa de velocidad y dirección del viento, hemos retirado datos, a fin de simular las situaciones de datos faltantes, como ocurren en la mayoría de las series obtenidas en las estaciones meteorológicas. Sobre estas series en que se han producido artificialmente datos faltantes, serán a continuación aplicados algunos de los métodos de análisis de datos faltantes anteriormente descritos, a fin de intentar reconstruirlas (Ferreira, 2003). La comparación entre la serie original completa y las series reconstruidas ha tenido en cuenta a los siguientes aspectos:

- Análisis de los estadísticos elementales (media, mediana, varianza, asimetria, apuntamiento) de las series imputadas y su comparación con las de la serie original.
- Reconstrucción de la distribución de probabilidad.
- Reconstrucción de las funciones de autocorrelación y autocorrelación parcial.
- Un estadístico del tipo $\sum\left(x_{i_{original}} - x_{i_{imputado}}\right)^2$; si el valor *i*-ésimo de la serie original no ha sido eliminado, entonces en la serie imputada $x_i = x_{i_{original}}$ es un valor observado y luego dicho valor no contribuye para la suma de cuadrados; si el *i*-ésimo valor de la serie original ha sido eliminado y enseguida imputado, probablemente $x_i \neq x_{i_{original}}$, y entonces este valor producirá un incremento en la suma de cuadrados.
- Un estadístico del tipo $\sum\left(x_i - \overline{x}_o\right)^2$ como una medida de bondad de ajuste entre los valores imputados x_i y la media de la serie original $\overline{x}_o$. Si el *i*-ésimo valor de la serie original no ha sido eliminado, entonces en la serie imputada $x_i = x_{i_{original}}$ es un valor observado, y $\left(x_i - \overline{x}_o\right)$ contribuye para $\sum\left(x_i - \overline{x}_o\right)^2$ con el mismo peso con que ha contribuido para la varianza de la serie original; si el *i*-ésimo valor ha sido imputado, entonces probablemente $x_i \neq x_{i_{original}}$ y $\left(x_i - \overline{x}_o\right)$ contribuye para $\sum\left(x_i - \overline{x}_o\right)^2$ con el un peso distinto de aquél con que ha contribuido para la varianza de la serie original. Luego, la imputación será tanto más perfecta cuanto más semejantes sean $\sum\left(x_{i_{original}} - \overline{x}_o\right)^2$ y $\sum\left(x_{i_{imputado}} - \overline{x}_o\right)^2$.
- La comparación de las series temporales usando la metodología de cointegración. Se analizan en particular los estadísticos del modelo de regresión $Y_t = \beta_1 + \beta_2 X_t + u_t$, en que Y_t es la serie original, completa, y X_t es cada una de las series con diferentes porcentajes de valores imputados por las diversas metodologías de imputación. En este modelo san analizados los coeficientes del vector de cointegración, así como algunos estadísticos de ajuste del modelo: el coeficiente de determinación R^2 , la función de máxima verosimilitud (MV), los criterios de información de Akaike (AIC) y de Schwarz (BIC). Se utiliza el test aumentado de Dickey-Fuller (ADF) sobre los residuos del modelo de regresión, donde se analizarán los estadísticos de ajuste referidos.

- **Descripción da le base de datos original**

Los datos que iremos utilizar enseguida es una base de datos de velocidad y dirección del viento, de la estación meteorológica de Castelo Branco (Portugal), para el período de 1988 a 1999. Las coordenadas de la estación son 7º 30’ W y 39º 50’ N; altitud sobre el nivel del mar 413 m;

altura del anemómetro y veleta: 10 m sobre el nivel de suelo. El anemómetro es de cazoletas. La estación se sitúa cerca de 2 Km al este de la ciudad, una zona llana, sin obstáculos en las cercanías. La serie contiene las mediciones trihorarias (registros a las 0:00, 3:00, 6:00, 9:00, 12:00, 15:00, 18:00, 21:00 horas locales) de la velocidad media del viento (Km/hora) y la dirección media (en grados), referentes al período de 1 de Enero de 1988 al 31 de Diciembre de 1999. La base de datos está completa, no existiendo ningún dato faltante. A fin de entrenar el mecanismo de imputación de datos faltantes, se han creado a partir de esta base de datos unas otras bases de datos en las cuales se han eliminada al azar unos 5%, 10%, 15%, 20%, 25%, 30%, 35% y 40% de las observaciones. Se ha testado la aleatoriedad de los mecanismos de datos faltantes, y se ha verificado que son MCAR los mecanismos de ausencia de datos en las series derivadas de la serie original por eliminación de diversos porcentajes de datos.

- **Análisis de las series imputadas con metodología hot-deck**

Se ha procedido (Ferreira, 2003) a la imputación de los datos faltantes con un procedimiento basado en la metodología semejante al hot-deck secuencial, referida en Ford (1983), Sandle (1983) y Little y Rubin (1987), pero con algunas modificaciones descritas a continuación. Tras identificar las rachas de longitud superior a 40 datos ausentes, se han rellenado estas observaciones faltantes con los datos de las flechas equivalentes en uno de los años anteriores o siguientes de la serie, seleccionando el año donante por el criterio de mayor similitud de los datos inmediatamente antes y después de la racha ausente, sea en el trozo de la serie en el cual se verifica la racha ausente, sea en las flechas equivalentes en los años cercanos. Para imputar los restantes datos faltantes, en rachas de longitud inferior a 40 datos, tras intentar identificar algunos segmentos con igual número de datos y que, por el análisis de los datos inmediatamente anteriores y siguientes al segmento, sea lo más similar posible, en valor medio y varianza, así como en la tendencia creciente ó decreciente, a los valores inmediatamente anteriores y siguientes al hueco, se ha rellenado el hueco con este segmento donante.

Las estimaciones de la media y mediana parecen resultar poco afectadas: los mayores desvíos, con respecto a los parámetros de la serie original, son de orden de 0,4% . El valor de la varianza se mantiene prácticamente igual a la varianza de la serie original completa, al paso que en todos los restantes métodos se observa una ligera tendencia de disminución. Los coeficientes de asimetría se mantienen estables. Los coeficientes de apuntamiento varían ligeramente entorno del coeficiente de la serie original, sin ninguna tendencia asociada al porcentaje de valores imputados.

Cuadro 4.1: Variación de las estimaciones de los parámetros estadísticos de las series imputadas con respecto a los parámetros de la serie original

	$\Delta_{o,i_{5\%}}$	$\Delta_{o,i_{10\%}}$	$\Delta_{o,i_{15\%}}$	$\Delta_{o,i_{20\%}}$	$\Delta_{o,i_{25\%}}$	$\Delta_{o,i_{30\%}}$	$\Delta_{o,i_{35\%}}$	$\Delta_{o,i_{40\%}}$
Media	-0,0012	-0,0005	0,0017	0,0001	0,0016	0,0042	-0,0001	0,0004
Mediana	-0,0001	-0,0001	-0,0001	-0,0001	-0,0001	-0,0001	-0,0001	-0,0001
Desviación típica	-0,0016	-0,001	-0,0014	-0,0029	-0,0015	0,0022	0,0055	0,0042
Varianza	-0,0031	-0,002	-0,0027	-0,0057	-0,003	0,0045	0,011	0,0083
Asimetria	-0,0042	0,0216	-0,0161	0,0008	-0,0048	0,0051	0,0356	0,0071
Apuntamiento	-0,0137	0,1123	-0,082	0,0348	-0,0351	0,0633	0,1538	-0,0086
Máximo	0	0	0	0	-0,0588	0	0	-0,0196

Las funciones de autocorrelación de las series imputadas tienen un patrón de comportamiento semejante a la serie original, pese todavía que los valores de los coeficientes de autocorrelación se van disminuyendo a la medida que se aumenta el porcentaje de valores imputados.

Hay una tendencia nítida para el aumento del parámetro de forma β del modelo Weibull (cuadro 4.3) que todavía se mantiene casi constante para las diversas series con diferentes porcentajes de valores imputados; el parámetro de escala η sufre un ligero aumento con respecto a la serie original, pero se mantiene casi constante para las diversas series.

Cuadro 4.2: Parámetros de la distribución Weibull para las series imputadas con hot-deck

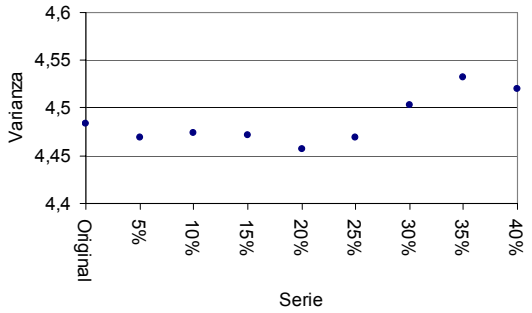
	β	η	% non zero	Máx. verosim.
Serie completa	1,8310	3,8961	0,9317	-65376,2358
5%	1,9217	4,0400	0,9315	-65259,4727
10%	1,9173	4,0373	0,9328	-65362,8521
15%	1,9198	4,0430	0,9330	-65429,8061
20%	1,9246	4,0428	0,9320	-65273,9740
25%	1,9225	4,0445	0,9328	-65380,4774
30%	1,9247	4,0615	0,9316	-65389,0642
35%	1,9114	4,0473	0,9306	-65357,0852
40%	1,9138	4,0479	0,9314	-65394,6848

La sumas de cuadrados $\sum\left(x_{i_{original}} - x_{i_{imputado}}\right)^2$ y $\sum\left(x_{i_{imputado}} - \overline{x}_o\right)^2$ se presentan en el cuadro 4.3. La suma de cuadrados y varianza de va aumentando con el porcentaje de datos imputados. Los valores de las sumas de cuadrados y varianzas de de las diversas series imputadas son muy semejantes a los valores de la serie original, con excepción con 30% o más de valores imputados, donde estos valores son ligeramente superiores a los de la serie original.

Cuadro 4.3: Suma de cuadrados y varianza de $\left(x_{i_{original}} - x_{i_{imputado}}\right)$ y $\left(x_{i_{imputado}} - \overline{x}_o\right)$

	Serie								
	Original	5%	10%	15%	20%	25%	30%	35%	40%
$\sum(o-i)^2$		5901.9	13542.1	19013.1	24840.2	32352.3	40481.9	48344.2	58837.8
Varianza		0.1684	0.3865	0.5426	0.7089	0.9233	1.1552	1.3797	1.6792
$\sum\left(x_i - \overline{x}_0\right)^2$	157086,3	156597,6	156767,2	156663,4	156185,2	156610,0	157797,2	158820,2	158394,0
Varianza	4,4832	4,4692	4,4741	4,4711	4,4575	4,4696	4,5033	4,5327	4,5205

Gráfico 4.1: Varianza de $\left(x_i - \overline{x}_o\right)$ para las series original y imputadas con el procedimiento hot-deck



Los criterios de análisis de la cointegracion de la serie original con cada una de las series imputadas (*hotXX*, en que *XX* es el porcentaje de valores imputados) se presentan a continuación.

Los indicadores de bondad de ajuste se van empeorando con el aumento del porcentaje de valores imputados. Los vectores de cointegración varían desde (0.0592, 0.9827) en la serie con 5% de valores imputados hasta (0.6087,0.8101) el la serie con 40% de valores imputados, pero así mismo se concluye que la serie original y cada una de las series imputadas siguen siendo cointegradas.

Cuadro 4.4: Análisis de cointegracion de las series imputadas con *hot-deck*.

	Modelo de regresión						Análisis de los resíduos						
Serie	$\hat{\beta}_1$	$\hat{\beta}_2$	R^2	MV	AIC	BIC	ADF	ρ	R^2	DW	MV	AIC	BIC
hot05	0,0592	0,9827	0,9627	-18371,8	1,0487	1,0492	-84,6158	-0,9838	0,4788	2,0002	-18335,8	1,0471	1,0485
hot10	0,1371	0,9578	0,9156	-32298,8	1,8665	1,8670	-83,2579	-0,9280	0,4558	1,9999	-32557,1	1,8589	1,8603
hot15	0,1855	0,9407	0,8825	-38491,9	2,1971	2,1976	-82,3434	-0,8851	0,4330	2,0001	-38148,6	2,1781	2,1795
hot20	0,2459	0,9234	0,8477	-43032,9	2,4563	2,4558	-81,9294	-0,8580	0,4169	1,9999	-42528,3	2,4281	2,4296
hot25	0,3223	0,8982	0,8044	-47419,9	2,7067	2,7072	-82,7978	-0,8488	0,4029	1,9999	-46716,0	2,6672	2,6686
hot30	0,4073	0,8695	0,7594	-51042,7	2,9135	2,9140	-82,5868	-0,8254	0,3883	1,9998	-50099,8	2,8603	2,8618
hot35	0,5066	0,8423	0,7174	-53866,1	3,0747	3,0751	-81,5795	-0,7807	0,3673	2,0000	-52534,3	2,9993	3,0008
hot40	0,6087	0,8101	0,6618	-5701,7	3,2542	3,2547	-79,5912	-0,7229	0,3428	2,0001	-55133,6	3,1477	3,1491

5 - CONCLUSIONES

La metodología hot-deck nos parece que, después de convenientemente refinada y, principalmente, se traducida en alguna aplicación informática sostenible y directamente orientada a series de velocidad del viento, puede ser objeto de interés.

Las ventajas principales de esta metodología (hot-deck) son que la imputación de valores para rellenar los huecos de datos faltantes en la serie se hace a expensas de segmentos de datos observados en la misma serie, donde se preserva al máximo la función de distribución de probabilidades, principalmente de valores cero de velocidad. Tiene el gran inconveniente en la selección de dichos segmentos, cuya elección se hace de manera a respectar la máxima similitud con los datos inmediatamente anteriores y siguientes en la secuencia interrumpida. El principal inconveniente reside en la elección de cual ó cuales deberán ser los criterios estadísticos de similitud, así como en cuantos datos anteriores y siguientes a los huecos se deberá reflejar tal similitud entre el segmento de datos en falta y el segmento donante (con la misma longitud ó numero de datos que el segmento vacío).

6 - REFERENCIAS

Aguiar, R.J., Collares-Pereira, M., Conde, J.P. (1988) Simple procedure for generating sequences of daily radiation values using a library of markov transition matrices. ***Solar Energy***, **40** (3), 269-279.

Aguiar, R.J. y Collares-Pereira (1992) TAG: a time-dependent, autoregressive gaussian model for generating synthetic hourly radiation. ***Solar Energy***, **49** (3), 167-174.

Bailar, B.A. y Bailar III, J.C. (1983) Comparison of the biases of the hot-deck imputation procedure with an "equal-weights" imputation procedure. ***in Incomplete data in sample surveys, Vol. III: Proceedings of the Symposium*** (W.G. Madow y I. Olkin, Eds.). New York: Academic Press, 299-312.

Bárcena, M.J. (2000). ***Técnicas multivariantes para el enlace de encuestas***. Tesis doctoral, Universidad de Bilbao, España.

Beale, E.M.L. y Little, R.J.A. (1975) Missing values in multivariate analysis. ***Journal of the Royal Statistical Society***, ***B***, ***37***, 129-145.

Buck, S.F. (1960) A method of estimation of missing values in multivariate data suitable for use with an electronic computer. ***Journal of the Royal Statistical Society***, ***B***, ***22***, 302-306.

Dempster, A.P., Laird, N.M. y Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. ***J Royal Statistical Society***, ***B***, ***39,*** 1-38.

Dempster, A.P. y Rubin, D.B. (1983) Overview, ***in Incomplete data in sample surveys, Vol. II: Teory and Annotated Bibliography*** (W.G. Madow, I. Olkin y D.B. Rubin, Eds.). New York: Academic Press, 3-10.

Ferreira, Armando Mateus (2003) ***Modelos estadísticos en energía eólica***. Tesis doctoral, Universidad de Córdoba, España.

Ford, B.L. (1983) An overview of hot-deck procedures. ***in Incomplete data in sample surveys, Vol. II: Teory and Annotated Bibliography*** (W.G. Madow, I. Olkin y D.B. Rubin, Eds.). New York: Academic Press, 185-207.

Kline, Jack y Milligan, Michael (1998) An evaluation of hourly average wind-speed estimation techniques. WindPower'98; 27 Abril – 1 Mayo, 1998; Bakersfield, California, USA; NREL/CP-500-24664, National Renewable Energy Laboratory, USA.

Little, R.J.A. (1982) Models for nonresponse in sample surveys. ***J. American Statistical Association***, *77*, 237-250

Little, R.J.A. y Rubin, D.B. (1987) ***Statistical analysis with missing data***. John Wiley and Sons, New York.

Rubin, D.B. (1987) ***Multiple imputation for nonresponse in surveys***. John Wiley and Sons, New York.

Sande, I.G. (1983) Hot-deck imputation procedures. ***in Incomplete data in sample surveys, Vol. III: Proceedings of the Symposium*** (W.G. Madow y I. Olkin, Eds.). New York: Academic Press, 339-349.

Schafer, J.L. (2000) ***Analysis of incomplete multivariate data***. Monographs on Statistics and Applied Probability. Chapman & Hall,/CRC.

Schafer, J.L. y Olsen, M.K. (1998) Multiple imputation for multivariate missing-data problems: a data analyst's perspective. ***Multivariate Behavioural Research***, **33**, 545-571.

Sparis, P., Antonogiannakis, J., Papadopoulos, D. (1995) Markov matrix coupled approach to wind speed and direction simulation. ***Wind Engineering***, **19** (3), 121-133.