Estimación de datos faltantes de precipitación de la Estación Meteorológica ISER Pamplona, Colombia

/Application of methodologies to estimate missing data of precipitation of the weather station ISER

Jeisson Leal R [1], María Esther Rivera [2]

[1] Joven investigador de último semestre de Ingeniería Ambiental. Facultada de Ingenierías y Arquitectura. Universidad de Pamplona, Jeison.leal@unipamplona.edu.co.
[2] Lic. En Matemáticas y física, PhD en Hidrología. Grupo de Investigaciones Ambientales Agua, Aire y Suelo (GIAAS). Facultad de Ingenierías y Arquitecturas. Programa de Ingeniería Ambiental; Universidad de Pamplona, maes@unipamplona.edu.co.

FECHA DE RECEPCIÓN DEL ARTÍCULO: 08/08/2016 FECHA DE ACEPTACIÓN DE ARTÍCULO: 08/11/2016

RESUMEN

En meteorología y climatología, comúnmente no se registran la totalidad de datos meteorológicos debido a factores antrópicos, naturales, como instrumentales. Con base a este problema y gracias a los datos de precipitación 1973-2015 aportados por el Instituto de Hidrología, Meteorología y Estudios Ambientales de Colombia (IDEAM), se estimó mediante diferentes metodologías de datos faltantes los valores que no fueron registrados en la serie mensual multianual de precipitación de la estación meteorológica ISER PAMPLONA, ubicada en el municipio de Pamplona, Norte de Santander, Colombia. Para ello, se utilizó el software SOLAS for missing data analysis versión 5.0, aplicando ocho metodologías de imputación: Group Means, Hot Deck, Predicted Mean, Predictive Model Based, Propensity Score, Predictive Mean Matching, Mahalanobis Distance y Propensity Score/Predictive Mean/Mahalanobis Distance Combination Method, determinando cual o cuales de estas metodologías resultaron ser aceptables al momento de usar los datos de la serie mensual multianual de precipitación de la estación ISER PAMPLONA para posteriores análisis de cualquier tipo. Se obtuvieron y compararon las gráficas de precipitación anual y mensual acumulada para cada una de las metodologías utilizadas. Las gráficas no mostraron diferencias notables entre sí. Los datos faltantes de cada metodología fueron promediados obteniéndose los valores de 8.89, 25.52, 25.69 y 25.80 mm correspondientes a febrero de 1984, marzo de 1976, marzo de 1987 y noviembre de 1995, respectivamente. Para la serie mensual multianual de precipitación de la estación meteorológica ISER PAMPLONA es posible realizar la aplicación de cualquier metodología mencionada, esto se debe a que la cantidad de datos faltantes es menor al 1%.

PALABRAS CLAVE

Dato faltante, imputación, Iser Pamplona, hidrología, SOLAS

ABSTRACT

In meteorology and climatology, normally not all the meteorological data are registered due to both anthropogenic and natural factors, like instruments. On the base of this problem and thanks to the precipitation data from 1973-2013 contributed by IDEAM, the missing data values that were not recorded in the multiannual monthly precipitation series of the ISER PAMPLONA weather station (located in Pamplona, Norte de Santander - Colombia) were estimated using different methodologies. To achieve this, the software SOLAS for missing data analysis software version 4.02 was used, where the value of the missing data was estimated by eight different methodologies: Group means, Hot Deck, Predicted mean imputation, Predictive model based, propensity score method, Predictive mean matching method, Mahalanobis distance and Propensity score/predictive mean/mahalanobis distance combination method. It was determined which of these methodologies resulted to be acceptable in the moment of using them for later analysis of any kind. The graphs of annual precipitation and monthly accumulated for each methodology were obtained and compared, the graphs did not show significant differences between each other. The missing data calculated by each methodology were averaged, obtaining values of 8.89, 25.52, 25.69 and 25.80 mm corresponding to February 1984, March 1976, March 1987 and November 1995, respectively.

Para citar este artículo / To cite this article
J. Leal, M. E. Rivera Estimación de datos faltantes de precipitación de la Estación Meteorológica ISER Pamplona, Colombia. Revista Ingenieros Militares, No 11, pp. 83, 2016.

For multiyear monthly precipitation series of the weather station PAMPLONA ISER is possible the application of any methodology mentioned, this is because the amount of missing data is less than 1%.

KEYWORDS

Imputation, Iser Pamplona, hydrology, missing data, SOLAS.

- REVISTA INGENIEROS MILITARES número 11, Año 2016, ISSN 2145-3144 / BOGOTÁ-COLOMBIA

INTRODUCCIÓN

La importancia que representa la precipitación en el diseño de obras civiles, en la agricultura y en la seguridad de la sociedad frente a algunos desastres naturales hace imprescindible contar con la totalidad de los datos de precipitación para obtener resultados ajustados a la realidad del fenómeno que se estudia; desafortunadamente no siempre se registran todos los datos de precipitación, lo cual genera inconvenientes a la hora de estudiar el fenómeno. Dado este inconveniente, varios autores mencionan diferentes métodos para estimar los datos faltantes [1]-[4], algunos modelos se basan en la utilización de una estación o estaciones vecinas para dar respuesta a los datos faltantes como es el caso de curvas de doble masa o interpolación con otras estaciones, en donde están estaciones vecinas presentan características geoclimatológicas muy similares a la estación de estudio. También, existen otras metodologías que se basan en modelos de Regresión Múltiple Lineal, Cadenas de Markov, Metodos Bayesianos o técnicas de Monte Carlo para dar solución a los datos faltantes sin necesidad de contar con una estación vecina logrando una buena estimación.

Los datos faltantes son aquellos datos que no son registrados debido a cualquier acontecimiento. Dentro de la imputación de datos se encuentran los métodos de imputación simple y múltiple. La imputación simple consiste en asignar un valor por cada dato faltante basándose en el valor de la propia variable o de otras variables, generando una base de datos completa, mientras que la imputación múltiple consiste en asignar a cada dato faltante varios valores (m), generando m conjuntos de datos completos, en cada conjunto de datos completo se estiman los parámetros de interés y posteriormente se combinan los resultados obtenidos [5].

Los datos faltantes se clasifican de la siguiente forma [6]:

1. MCAR Missing Completely At Random (Completamente aleatorio): Se da este tipo cuando la probabilidad de que el valor de una variable sea observado para un individuo no dependiente ni del valor de esa variable, ni del valor de las variables consideradas. Es decir, la ausencia de información no está originada por ninguna variable presente en la matriz de datos.
2. MAR Missing At Random (Aleatorio): Se da este tipo si la probabilidad de que el valor de una variable sea observado por un individuo no dependiente del valor de esa variable, pero quizá sí del que toma alguna otra variable observada. Es decir, la ausencia de datos está asociada a variables presentes en la matriz de datos.
3. NMAR Not Missing At Random: Se produce este tipo de mecanismo en el caso en el cual la probabilidad de que un valor sea observado depende del propio valor, siendo este un valor desconocido.

Imputaciones simples

Group Means: Propuesta por primera vez por [7] y posiblemente uno de los métodos más sencillos, consiste simplemente en imputar el dato o datos faltantes como el promedio aritmético de los datos que se tienen.

Hot Deck: proceso de duplicación. Cuando existe un dato faltante, se duplica un valor ya existente en la muestra para reemplazarlo [5]. Una ventaja de la imputación Hot Deck es que los valores imputados no sufren la pérdida de variabilidad. 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 [8].

Predicted Mean: Este método de imputación de datos faltantes usa la imputación Hot Deck basada en la imputación mediante regresión lineal por mínimos cuadrados, propuesto por primera vez por [9].

En su forma más simple, es más cercana a la imputación por el vecino más cercano en donde la distancia se define basándose en los valores predichos yi del modelo de imputación [10]

Imputaciones múltiples

Rubin considera que el número mínimo de imputaciones para proporcionar estimaciones validas es, en general, tres y Schafer no aconseja utilizar más de 10 [5]. Las imputaciones múltiples usadas se dividen en:

Predictive Model Based: Esta imputación se desarrolla utilizando una regresión de mínimos cuadrados ordinaria o un análisis de discriminantes. La información de un conjunto especificado por el usuario de covariables se utiliza para imputar los valores perdidos en las variables a ser imputadas, el uso de este modelo estima nuevos parámetros de regresión lineal que son extraídos aleatoriamente de una distribución bayesiana [11].

Propensity Score o Índice de Propensión: Definido por [12] como la probabilidad condicional de recibir un tratamiento dadas las características de un pre-tratamiento [13]. El índice de Propensión para la una unidad i, $e(x_i)$, puede ser estimado a partir de regresiones logísticas de las condiciones del tratamiento z_i en un vector de covarianza x_i y viene dado por la ecuación 1 :

$$ln\left(\frac{e(x_i)}{1-e(x_i)}\right) = \beta x_i \qquad (1)$$

Donde β es el vector de los coeficientes de regresión. El índice de propensión estima la probabilidad de un tratamiento de ser asignado basándose en los pretratamientos de las covariaves ya observadas, es común que este método utiliza modelación paramétrica, particularmente en las regresiones logísticas, aunque métodos no paramétricos como arboles de regresión o modelos boosted pueden ofrecer, en algunos casos, mejores resultados [14], [15], [16] Citado por [17].

Predictive Mean Matching: Es uno de los métodos más comúnmente usados [18]. El método Predictive Mean Matching (PMM) se puede considerar como un método de imputación de Nearest Neigtbor Donor (NND) particular [19]. En este método utiliza regresiones lineales y se basa en un bootstrap bayesiano aproximado [11]. La idea básica del modelo de imputación PMM es utilizar métodos de regresión para llegar a una estimación del dato faltante de la variable x. Sin embargo, en lugar de utilizar dicha estimación, identifica uno o más vecinos que posean valores estimados similares, el valor observado del vecino más próximo es utilizado como el valor imputado para el dato faltante de la variable x [20].

Mahalanobis Distance: Es una medida que se puede utilizar para medir la similitud entre dos vectores. Los vectores serán los casos del conjunto de datos los cuales se componen de los valores de las covariables especificadas para el cálculo [11]. La Distancia de Mahalanobis viene dada por la ecuación 2:

$$d(\vec{x_i}, \vec{y}) = \sqrt{(\vec{x_i}, \vec{y})^T \; S^{-1}(\vec{x_i}, \vec{y})} \qquad (2)$$

Donde: (x_i)representa el vector con los datos completos, y representa el vector con datos faltantes y S^{-1} representa la matriz de covarianzas. Cada valor de dato faltante a imputar, es extraído al azar del subconjunto de valores observados que poseen una menor distancia de Mahalanobis.

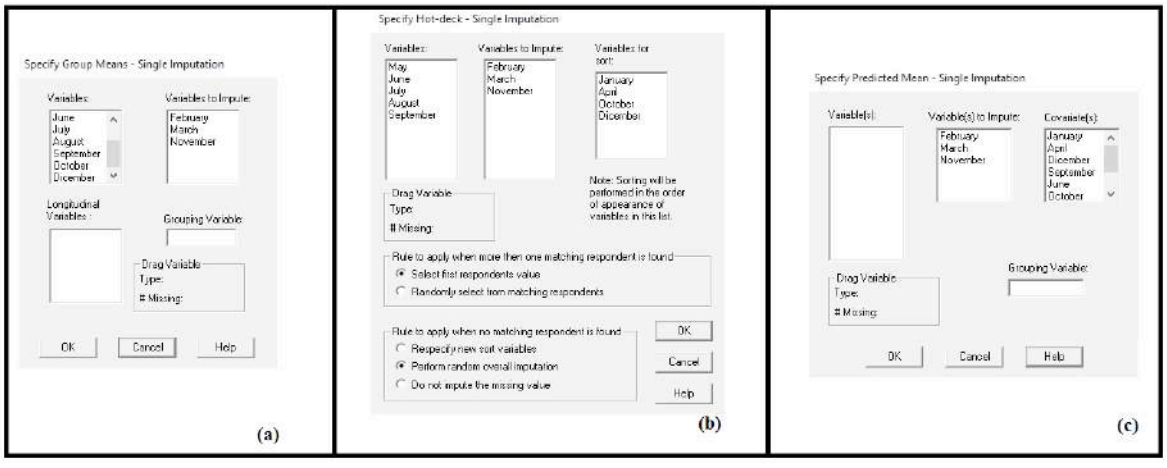

Figura 2. Proceso de imputación simple.

- REVISTA INGENIEROS MILITARES
número 11, Año 2016, ISSN 2145-3144 / BOGOTÁ-COLOMBIA

Propensity Score/Predictive Mean/Mahalanobis Distance Combination Method: Es una combinación de los métodos de imputación múltiple Propensity score, Predictive Mean Matching y la Distancia de Mahalanobis. Utiliza el índice de Propensión y Predictive Mean Matching al conjunto de datos. Los resultados son utilizados como covariables y se aplica el método de la Distancia de Mahalanobis para encontrar casos que se puedan utilizar para imputar los datos faltantes [11].

METODOLOGÍA

Por medio del IDEAM se obtuvieron los datos de la serie mensual multianual de precipitación de la estación meteorológica ISER PAMPLONA; luego se importaron los datos al software SOLAS for missing data nalysis versión 5.0 (figura 1). Una vez importados los datos se inició la imputación de estos mediante diferentes metodologías.

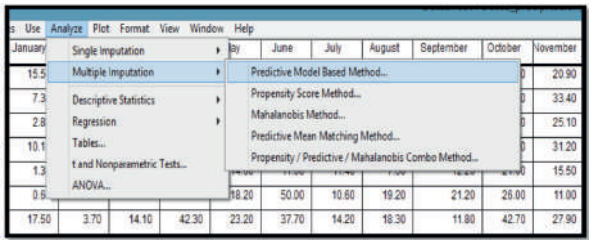

Figura 3. Metodologías de imputación múltiple.

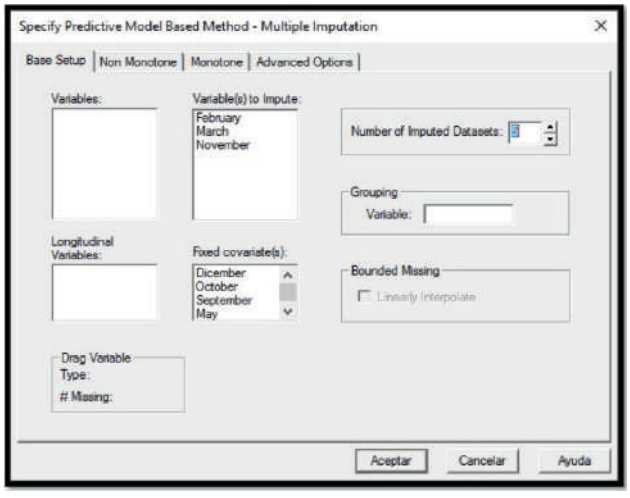

Figura 4. Proceso de imputación múltiple.

Group Means

AÑO Y MES	PRECIPITACION(mm)
1984 - FEBRERO	16.3
1976 - MARZO	19.82
1987 - MARZO	19.82
1995 - NOVIEMBRE	25.51

Hot Deck

AÑO Y MES	PRECIPITACION(mm)
1984 - FEBRERO	6.00
1976 - MARZO	22.40
1987 - MARZO	22.40
1995 - NOVIEMBRE	18.00

Predicted Mean

AÑO Y MES	PRECIPITACION(mm)
1984 - FEBRERO	10.19
1976 - MARZO	30.21
1987 - MARZO	23.33
1995 - NOVIEMBRE	33.09

Propensity Score

AÑO Y MES	PRECIPITACION(mm)
1984 - FEBRERO	9.22
1976 - MARZO	40.64
1987 - MARZO	22.8
1995 - NOVIEMBRE	26.83

Predictive Model Based

AÑO Y MES	PRECIPITACION(mm)
1984 - FEBRERO	8.60
1976 - MARZO	26.15
1987 - MARZO	23.80
1995 - NOVIEMBRE	28.10

Predictive Mean Matching

AÑO Y MES	PRECIPITACION(mm)
1984 - FEBRERO	7.67
1976 - MARZO	23.06
1987 - MARZO	38.84
1995 - NOVIEMBRE	20.53

Mahalanobis Distance Combination Method

AÑO Y MES	PRECIPITACION(mm)
1984 - FEBRERO	8.29
1976 - MARZO	24.6
1987 - MARZO	28.42
1995 - NOVIEMBRE	20.11

Propensity Score/Predictive Mean/Mahalanobis Distance Combination Method

AÑO Y MES	PRECIPITACION(mm)
1984 - FEBRERO	8.97
1976 - MARZO	17.28
1987 - MARZO	26.08
1995 - NOVIEMBRE	34.21

Figura 5. Datos faltantes imputados para cada metodología

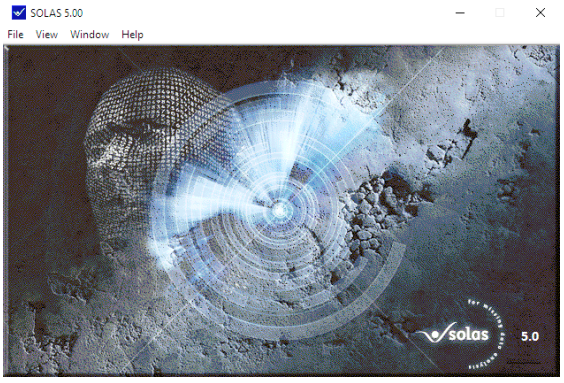

Figura 1. Software SOLAS for missing data analysis version 5.0.

Imputación simple. Se accedió al menú, después analyze y simple imputation en donde se seleccionó el método de imputación simple a usar. Para el caso del método Group Means se seleccionaron los meses de Febrero, Marzo, y Noviembre como variables a imputar (figura 2a). Para la metodología Hot Deck, además de seleccionar los meses de Febrero, Marzo, y Noviembre como variables a imputar, se seleccionaron los meses de Enero, abril, Octubre y Diciembre como variables de clasificación dada su proximidad con los meses de datos faltantes (figura 2b). Para finalizar, en la metodología Predictive Mean se seleccionaron las variables a imputar correspondientes a los meses de Febrero, Marzo y Noviembre, los demás meses se asignaron como covariables (figura 2c).

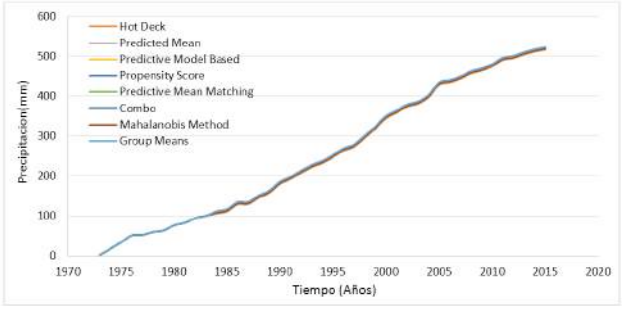

Figura 5. Datos faltantes imputados para mes de febrero.

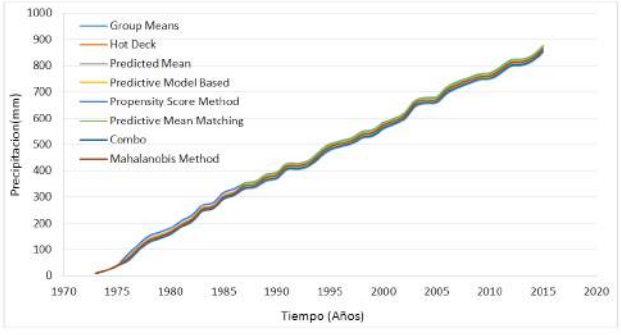

Figura 6. Datos faltantes imputados para mes de marzo

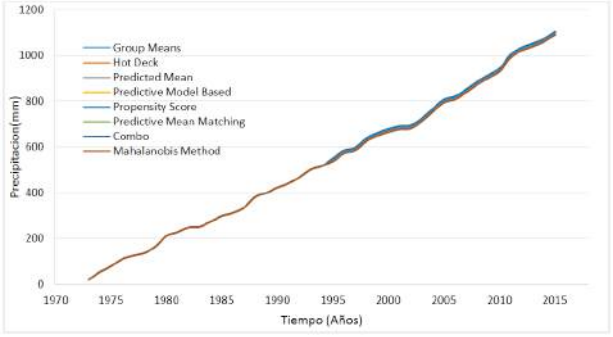

Figura 7. Datos faltantes imputados para mes de noviembre

Imputación múltiple. Se accedió al menú, analyze y multiple Imputation en donde se seleccionó el método de imputación múltiple a usar (figura 3). Posteriormente se seleccionaron las variables a imputar correspondientes a los meses de Febrero, Marzo, y Noviembre, los demás meses se seleccionaron como covariables fijas. Se seleccionó 5 como el número de imputaciones a realizar. Este procedimiento se realizó para todas las metodologías de imputación múltiple (figura 4).

RESULTADOS Y DISCUSIÓN

En la figura 4 se muestra cada valor de dato faltante hallado para cada mes y año del registro mensual multianual de precipitación de la estación meteorológica ISER PAMPLONA mediante las diferentes metodologías de imputación.

El promedio de las metodologías de datos faltantes fue de 8.89, 25.52, 25.69 y 25.80 mm correspondientes a febrero de 1984, marzo de 1976, marzo de 1987 y noviembre de 1995. Se cuenta con un registro de 42 años de precipitaciones mensuales multianuales de la estación meteorológica ISER PAMPLONA, en donde los datos faltantes corresponden al 0,79% del 100% de datos.

La figura 5, 6 y 7 representan las acumulaciones mensuales de precipitación (1973-2015) de los datos imputados por cada metodología para de los meses de Febrero, Marzo y Noviembre, respectivamente de datos faltantes se observan los datos obtenidos de precipitación anual acumulada para todos los meses y para cada mes con dato faltante.

Aunque existen diferencias entre metodologías estas diferencias son poco notables al momento de graficar la precipitación acumulada, esto debido la cantidad de datos faltantes. Para la de precipitación acumulada del mes de marzo se presentó una mayor diferencia entre metodologías de imputación con respecto a las gráficas de febrero y noviembre, dado que el mes de marzo hubo una mayor cantidad de datos faltantes.

CONCLUSIONES

Dada la similitud de los datos obtenidos de la acumulación anual para todos los meses y la acumulación anual para cada mes con dato faltante, se concluye que para el caso de la precipitación mensual multianual de la estación meteorológica ISER PAMPLONA cualquier metodología de imputación de datos es aplicable y realmente no se observan diferencias significativas entre uno u otro método.

Cabe resaltar que la elección del procedimiento para el manejo de datos faltantes resulta una tarea compleja, pues un mismo método en determinadas situaciones produce estimaciones precisas y en otras, no, esto sugiere a los investigadores que, cuando manejen datos faltantes, valoren previamente, el uso de más de una alternativa para tratarlos que les permita una mejor elección del procedimiento a implementar, y no basarse en resultados obtenidos de otras investigaciones donde las condiciones pudieran haber sido diferentes [2].

AGRADECIMIENTOS

A Statistical Solution, por su colaboración y entrega del software Solas versión 5.0.

REFERENCIAS

[1] F.J. Aparicio, *Fundamentos de Hidrología de Superficie.* Editorial Limusa. 1992

[2] M. Cañizares, I. Barroso y Alfonso K. "Datos incompletos: Una mirada crítica para su manejo en estudios sanitarios". *Journal Gaceta Sanitaria, Vol 18, No. 1, pp. 58-63, 2004.*

[3] S. Fattorelli, P. Fernández. "Diseño hidrológico". Associazione Italiana di Idronomia, 2011.

[4] G. S. Monsalve. *"Hidrología en la Ingeniería".* Escuela Colombiana de ingeniería, segunda edición, 1999.

[5] D. G. Otero. Imputación de datos faltantes en un Sistema de Información sobre Conductas de Riesgo. Máster en Técnicas Estadísticas. Universidade De Santiago De Compostela, Universidade Da Coruña y Universidade De Vigo, 2011.

[6] Goicoechea, Aitor Puerta. Imputación basada en árboles de clasificación. En: EUSTAT, [en línea] p.5-19, 2002.

[7] S. S. Wilks. Certain generalizations in the analysis of variance. *Biometrika* 24, 471–94, 1932.

[8] A. M. Ferreira, Metodologías de análisis e imputación de datos faltantes en series

de velocidad del viento. *VI congreso de estadística e investigación de operaciones.* 2003.

[9] S.F. Buck . A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society*, B22, 302-306, 1960.

[10] G. B. Durrant, Imputation Methods for handling ítem-nonresponse in the social sciences: A methodological review. ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute, University of Southampton. 2005.

[11] Statiscal Solutions. Predictive Model Based Multiple Imputation Features. 2014. Disponible en:http://www.statsols.com/products/solas-for-missing-data.

[12] P.R. Rosenbaum, and D.B. Rubin "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika* 70(1), 41-55, 1983.

[13] S. Becker, A. Ichino, "Estimation of Average Treatment Effects Based on Propensity Scores". *The Stata Journal*, vol. 2, no. 4. 2002.

[14] B. Lee, , J. Lessler and E.A. Stuart, Improving propensity score weighting using machine learning. *Statistics in Medicine* 29(3): 337-346 2009.

[15] D. F. McCaffrey, G. Ridgeway, & A. R. Morral, Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychological Methods,* 9, 403-425, 2004.

[16] S. Setoguchi, S. Schneeweiss, M. A. Brookhart, , R. J. Glynn, & E. F. Cook, Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology Drug Safety,* 17(6), 546–555, 2008.

[17] W. Pan y H. Bai, Propensity Score Analysis Fundamental and Developments. 2015.

[18] G. Durrant and C. Skinner, Using missing data methods to correct for meas. Survey Meth, 2006.

[19] M. Zio, U. A Guarnera Semiparametric Predictive Mean Matching: An Empirical Evaluation. Statistical commission and economic commission for Europe. *Conference of European statisticians.* 2006.

[20] R. Williams, Missing Data Part II: Multiple Imputation. University of Notre Dame, 2015.