Tratamiento para Datos Faltantes en Series de Radiación Solar

Adaro J.A. , Marchesi J.O. y Zizzias J.H.

Grupo de Energía Solar - Facultad de Ingeniería - Universidad Nacional de Río Cuarto
Ruta 36 Km 601, (5800) Río Cuarto, Córdoba - Argentina, Tel. y Fax: 0358-4676246
e-mail: aadaro@ing.unrc.edu.ar

Resumen *- Para poder realizar un aprovechamiento racional de la energía solar es indispensable tener un conocimiento apropiado de la distribución espacio-temporal del recurso a nivel de la superficie terrestre. Por ello el Grupo de Energía Solar está realizando mediciones de radiación global y directa. En la toma de datos existen muchas posibilidades de cometer error de diferentes tipos, pero lo más significativo es la falta de datos. Lo que se busca es tener una metodología que indique qué hacer ante esta situación por lo que en este trabajo se realiza un estudio en el tratamiento de los datos obtenidos para inferir valores que permitan ser incorporados a las series en las que, por diferentes motivos, falta de información. La metodología que se utilizó para la incorporación de datos faltantes es el análisis de series de tiempo basado en modelos de espacio de estado.*

Palabras-clave: *Radiación Solar, Tratamiento de Datos, Serie de Tiempo*

Missing Data Treatment in Series Solar Radiation

Abstract *- In order to make a rational use of solar energy is essential to have a proper knowledge of the spatio-temporal resource level the surface. Therefore Solar Energy Group is conducting measurements of global radiation and direct. In the data collection there are many possibilities for error of different types, but the most significant is the lack of data. What is sought is to have a methodology to indicate what to do in this situation, so in this work we make a study on the treatment of the data to infer values that will be incorporated into the series who for various reasons are lack of information. The methodology used for the incorporation of missing data is the time series analysis based on state space models.*

Key words: *Solar Radiation, Data Processing, Time Series*

INTRODUCCIÓN

Para toda planificación que se pretenda iniciar destinada al aprovechamiento de la energía solar se deberá contar con la información básica, esto es, cantidad y variabilidad temporal de la radiación solar. Cuando se trata de diseñar instalaciones solares o simular sistemas de aprovechamiento energético se utiliza software que simula el funcionamiento de los mismos requiriéndose que los datos de entrada sean lo más confiables posible. Si bien existen muchas publicaciones destinadas a la generación de series sintéticas de radiación utilizando diversos métodos estadísticos (Aguiar et al., 1992, 1988) es central la decisión de realizar mediciones sistemáticas tanto de radiación directa como de radiación global.

El Grupo de Energía Solar, entre sus objetivos (Adaro et al., 2000), apunta a mejorar la información de los datos meteorológicos con diferentes propósitos, entre los que pueden citarse disponer de información para cálculos de diseño y obtener datos

confiables que permitan realizar trabajos de simulación detallados que requieren de datos diarios como horarios. Otro objetivo es la utilización de series existentes que tienen poco desarrollo temporal o por la presencia de datos atípicos (outliers) y a grandes cambios estructurales.

La idea básica de una serie de tiempo es muy simple. Consiste en el registro de cualquier cantidad fluctuante medida en diferentes momentos. Podemos tener, por ejemplo, un registro en un período de varios años, un registro en un período de varios días, o un registro de la variación en intensidad de una señal de varias horas. La característica común de todos los registros que pertenecen al dominio de las "series de tiempo" es que están influenciados, aunque sea parcialmente, por fuentes de variación aleatoria. Entonces, si deseamos explicar la estructura de las fluctuaciones en una serie de tiempo, debemos recurrir a lo que llamamos el estudio de las series de tiempo.

Hay dos aspectos en el estudio de las series de tiempo: el análisis y el modelado. El objetivo del análisis es resumir las propiedades de una serie y remarcar sus características principales. Esto puede hacerse ya sea en el dominio del tiempo o en el dominio de las frecuencias. En el primero se concentra la atención en las relaciones entre las observaciones en puntos diferentes del tiempo, mientras que en el segundo se estudia la periodicidad de las propias mediciones y variables relacionadas con ellas. Estas dos formas de análisis no son competitivas sino que, muy por el contrario, son complementarias. La misma información es procesada en diferentes formas dando distintas visiones de la naturaleza de la serie de tiempo.

La principal razón para modelar una serie de tiempo es permitir la predicción de sus valores futuros. La característica distintiva de un modelo de este tipo es que no se realiza ningún intento para formular una relación de comportamiento entre la serie de tiempo considerada y otras variables. Los movimientos de la serie son explicados solamente en términos de su propio pasado o por su posición en relación al tiempo. Las predicciones se realizan mediante extrapolación. Muchas series de tiempo ocurren en las ciencias marinas, geofísica, físicas en especial en meteorología, disciplina en la cual las mediciones de radiación solar constituyen un caso particular.

Uno de los más importantes problemas en el análisis de las series es el de predicción de valores futuros de la serie dado algunos datos sobre los valores pasados. Supongamos que tenemos datos que se extienden en la historia pasada y remota de la serie, o sea que tenemos todas las observaciones hasta el momento actual, y deseamos predecir el valor de algún momento futuro. Podemos considerar una predicción basada en una combinación lineal de todos los valores pasados. El problema allí presentado implica lograr una predicción "óptima" en algún sentido. Suponiendo que las propiedades estadísticas de la serie son totalmente conocidas, ésto se transforma en un problema puramente matemático y su solución fue dada en forma independiente por Wiener (1949) y Kolmogorov (1941) usando el error medio cuadrático como el criterio de optimalidad. El método de Wiener está basado en un enfoque en el dominio de las frecuencias e involucra técnicas matemáticas complicadas conocidas como "factorización espectral". Este método es muy difícil de aplicar a menos que la serie tenga una densidad espectral muy simple.

Box y Jenkins (1976) propusieron un método mucho más simple basado en ajustar un modelo Autorregresivos de Medias Móviles (ARMA en ingles) y luego calcular las predicciones directamente del modelo ajustado. Las predicciones son simplemente calculadas usando un algoritmo recursivo. El método es extremadamente fácil de aplicar pero requiere, por supuesto, que la serie sea bien ajustada por un modelo ARMA. El método de Wiener-Kolmogorov no requiere que la serie sea conformable con un modelo con un número finito de parámetros pero es mucho más difícil de aplicar a los datos.

Wiener extendió su método para tratar el problema asociado del filtrado cuando no se puede observar directamente la serie si no que, en su lugar, observamos una serie que consiste en la serie de interés más un ruido que corrompe a la primera. El problema aquí es construir predicciones de la serie de interés basándose en la observación de la serie corrompida. La solución de Wiener al problema de filtrado envuelve esencialmente las mismas herramientas matemáticas usadas en su solución al problema de predicción.

Una nueva y poderosa solución al problema

de filtrado fue ideado por Kalman (1960) usando la llamada representación de espacio de estado de una serie de tiempo. Esto provee una descripción muy compacta del modelo y está basado en el resultado conocido que dice que cualquier ecuación en diferencias (o diferencial) lineal de orden finito puede ser escrita como una ecuación vectorial en diferencias (o diferencial) lineal de primer orden. La ventaja de esta última representación es que involucra solamente dependencia de un paso, o sea que posee la propiedad de Markov, lo cual conduce a un algoritmo simple y elegante para calcular las predicciones de valores futuros de la serie conocida como el algoritmo del filtro y suavizador de Kalman. Por otro lado esa idea conduce a la representación de espacio de estado en donde las matrices de sistema que contienen los parámetros dependen del tiempo. Luego, mediante una aplicación adecuada del algoritmo del filtro y suavizador de Kalrnan, se obtienen las estimaciones y las predicciones de la serie. Un trabajo pionero dentro de esta área es el de Harvey y Durbin (1986). Un estudio profundo de este enfoque es el trabajo enciclopédico de Harvey (1989). Un tratamiento moderno del tema puede verse en Abril (1999).

METODOLOGÍA

El objetivo propuesto en el trabajo es contribuir al mejoramiento de la base de datos en construcción a través de la sistemática medición de radiación solar global y el análisis de los datos como la búsqueda de una metodología para su tratamiento. En la toma de datos existen muchas posibilidades de cometer errores de diferentes tipos. La falta de datos es un fallo significativo que se puede producir por diferentes motivos. Lo que se debe tener en claro es qué hacer ante la falta de datos. Ningún software realiza un tratamiento adecuado de los datos faltantes. Están los que los remplazan por valores promedios en determinados intervalos prefijados o los que simplemente eliminan en la estructura de los datos aquellos espacios en donde no se tiene la información. Por ello en esta ocasión la propuesta de trabajo fue la de remplazar los datos faltantes por valores inferidos de las series de datos precedentes.

La metodología utilizada es la de los modelos estructurales de series de tiempo. Recordemos que la idea básica de los modelos estructurales de series de tiempo es que ellos pueden ser puestos como modelos de regresión en donde las variables explicativas son funciones del tiempo con coeficientes que pueden cambiar a través del tiempo. La estimación actual de los coeficientes ó filtrada se logra poniendo al modelo en forma de espacio de estado y aplicándole luego el denominado Filtro de Kalman (Abril, 1999; Harvey et al., 1993). La representación matemática de los modelos de espacio de estado o también conocido como modelo lineal gaussiano de espacio de estado tiene la forma:

$$y_t = Z_t\,\alpha_t + \varepsilon_t \qquad \varepsilon_t \approx N(0, H_t) \tag{1}$$

$$\alpha_t = T_t\,\alpha_{t-1} + R_t\eta_t,\ \eta_t \approx N(0, Q_t)\ t = 1,...n, \tag{2}$$

donde y_t es un vector de orden px1 de observaciones y α_t es un vector de orden mx1 inobservable llamado vector de estado. La idea central es que el desarrollo del modelo esta determinado por α_t de acuerdo a la Ec. (2) presentada anteriormente, pero como α_t no puede ser observado directamente se deben basar el análisis en las y_t observaciones. Las matrices Z_t, T_t, R_t, H_t y Q_t se suponen inicialmente conocidas y los términos de ε_t y η_t errores que se supone son serialmente independientes e independientes entre sí en todo momento del tiempo.

El tratamiento estadístico de los modelos estructurales de series de tiempo está basado en la forma de espacio de estado, el filtro de Kalman y el suaviazador asociado. La función de verosimilitud se construye a partir del filtro de Kalman en términos de la predicción un paso hacia adelante y se maximiza con respecto a los hiperparámetros por optimización numérica. El vector marcador ("score") de los parámetros puede obtenerse a través de un algoritmo de suavizado asociado al filtro de Kalman. Una vez que los hiperparámetros fueron estimados el filtro se usa para conseguir predicciones de los residuos un paso adelante, lo que nos permite calcular los estadísticos para probar normalidad, correlación serial y bondad de ajuste. De esta manera el filtro analiza el sistema cada vez que hay una predicción pudiendo repetirse el proceso paso a paso

(Avila Blas et al., 1999; 2000).

Se obtuvieron mediciones utilizando pirheliómetro y piranómetro en los que se observa por diferentes causas la ausencia de datos en instancias en las cuales de no haber sido por los fallos se deberían haber registrado medidas. El análisis se centró en buscar una metodología que permitiera completar la información con la intención de disponer datos completos por períodos.

RESULTADOS

Los modelos estructurales de serie de tiempos estudiados se aplicaron a las series de radiación obtenidas por el sistema de medición a los fines analizar la posibilidad de completar los datos faltantes. Se puede afirmar como resultado general que la metodología de sustitución de datos faltantes es satisfactoria cuando la ausencia de los mismos no es importante en relación al total a los efectos de completar las series permitiendo encontrar valores estadísticos representativos de los valores de radiación solar incidente. Nuestros datos básicos son la radiación solar global desde el día 17 al 23 de marzo, tomados por un piranómetro de alta precisión marca EKO modelo SBP 801.

Como ejemplo de los resultados se tomó la radiación global de Río Cuarto del mes de marzo de 2003 y se registró el valor de radiación con un intervalo entre mediciones de un minuto. A partir de estos datos se calcularon los valores integrados horarios. Del total de 744 valores horarios que se dispone sólo se tomaron 724 debido a que se suprimieron valores en el día 19/03 simulando fallos a los efectos de corroborar la hipótesis de que usar valores inferidos por medio del tratamiento estadístico de serie de tiempo basado en modelos de espacio de estado resulta una metodología apropiada para completar datos faltantes.

En la Fig. 1 se muestran los horarios de radiación (en Joule/metros cuadrados) desde el día 17 al 23 de marzo de 2003 en donde se puede observar la ausencia de valores desde 16:00 hs. del día 19/03 a las 11:00 hs. del día 20/03. Se realiza el estudio y predicción de los valores faltantes utilizando para ello un software específico obteniendo un reporte o pronósticos de valores (Koopman et al, 1995).

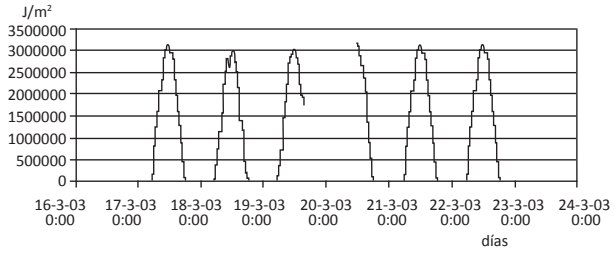

***Figura 1** - Valores horarios de radiación (en J/m²) desde el día 17 al 23 de marzo de 2003*

En la Fig. 2 se puede ver la descomposición de la serie de valores previos a los datos ausentes mostrando su tendencia estimada, estacionalidad estimada y componente irregular.

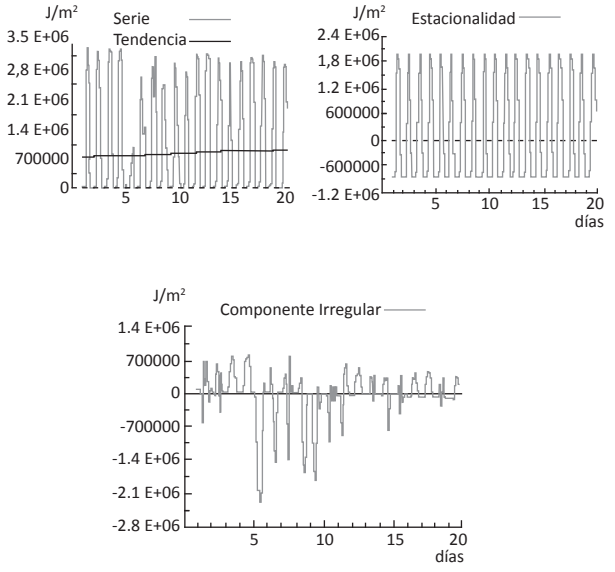

***Figura 2** - Serie original con tendencia estimada, estacionalidad estimada y con componente irregular*

En la Fig. 3 se muestran los residuos y su correlograma, residuos auxiliares irregulares con intervalo de confianza del 95%, distribución de los residuos auxiliares irregulares de los valores horarios de radiación global, residuos auxiliares del nivel con intervalo de confianza del 95% y distribución de los residuos auxiliares del nivel.

En la Fig. 4 se muestran los valores predichos y sus respectivos residuos, las primeras diferencias de los valores medidos y valores predichos reen-

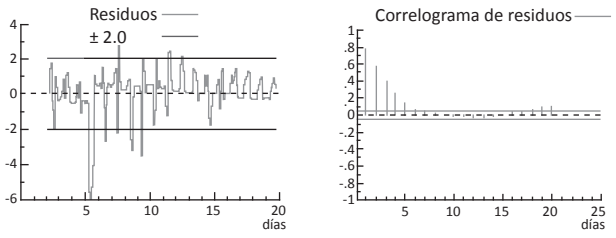

***Figura 3** - Residuos y correlograma de los residuos con intervalos de confianza del 95%*

cuentran en el interior de los intervalos de confianza al igual que los residuos de predicción.

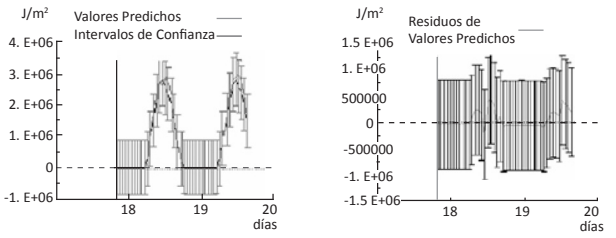

***Figura 4** - Valores predichos con intervalos de confianza y residuos de valores predichos*

Finalmente en la Fig. 5 se muestra la serie completa con la sustitución de los valores faltantes por los valores pronosticados.

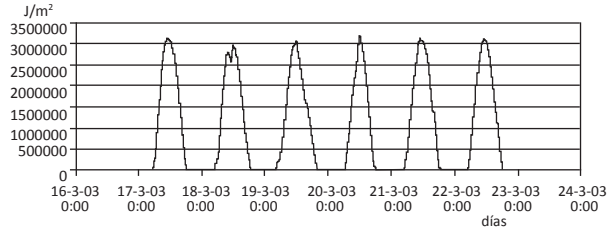

***Figura 5** - Valores horarios de radiación (en J/m²) con valores predichos*

Se observan los residuos del modelo ajustado con su intervalo de confianza del 95%, el correlograma también con su intervalo de confianza del 95%, el periodograma, a la densidad espectral estimada y la distribución de esos residuos. Por todo lo dicho anteriormente y de la observación de esta figura se puede concluir que el ajuste es altamente satisfactorio. Análisis similares a los realizados con los residuos del modelo ajustado se realizaron con los residuos del componente irregular y con los de la tendencia. En ambos casos se aceptan las respectivas hipótesis de normalidad.

Al comparar los valores inferidos por el modelo estructural de serie de tiempo con los valores medidos se observa que el coeficiente de correlación o coeficiente de Pearson da igual a 0.962629, mostrando que esta metodología de inferir valores faltantes da una muy buena aproximación.

CONCLUSIONES

Si bien sólo se tomaron datos de un día de alta radiación solar se puede concluir que los modelos estructurales de series de tiempo son una metodología valiosa y útil para inferir valores faltantes en las mediciones obtenidas de radiación solar global. Para dar validez general al método se debería analizar en series de radiación solar global para días de bajo índice de claridad k_t como así también en series de radiación solar directa y difusa.

REFERENCIAS

Aguiar y Collares Pereira, "Tag: A time dependent, autoregressive, gaussian model for generating synthetichourly radiation". Solar Energy 49, 3, 167-174, (1992).

Aguiar, Collares-Pereira and Conde, "Simple procedure for generating sequences of daily radiation values using a library of Markov transitions matrices". Solar Energy vol. 40 No. 3, pp. 269-279, (1988).

Adaro, Césari, Lema, Galimberti y Barral, "Estudio comparativo de serie de radiación solar". Proceedings of the Millenium Solar Forum, México, (2000).

Wiener, "The Extrapolation, Interpolation an Smoothing of Stationary Time Series with Engineering Aplications". Wiley: New York, (1949).

Kolmogorov, "Interpolation and extrapolation von stationaren Zufalligen Folgen". Bull. Acad. Sci (Nauk). USSR, Ser. Math., 5, 3-14, (1941).

Box and Jenkins, "Time Series Analysis: Forecasting an Control" (Revised ed) Holden-Day, Inc: San Francisco, (1976).

Kalman, "A new approach to linear filtering an prediction problems". Trans. ASME, J. Basic Eng., 82D, 35-45, (1960).

Harvey and Durbin, "The effects of seat belt le-

gislation on Britis road road casualties: a case study in structural time series modeling". J. Roy. Statis. Soc., A, 149, 187-227, (1986).

Harvey, "Forecasting, Structural Time Series Models and the Kalman Filter". Cambrid University Press: Cambridge, (1989).

Abril, "Análisis de serie de tiempo basado en modelos de espacio de estado", Eudeba, Argentina, (1999).

Harvey y Shepard, "Structural Time Series Models", Handbook of Statistics 11, 621-302, (1993).

Avila Blas, Abril y Lesino, "Análisis estadístico estructural de series de radiación diarias". Avances en Energías Renovables y Medio Ambiente 3, 2, 11.17-11.20, (1999).

Avila Blas, Abril y Lesino, "Radiación y temperaturas diarias: Un modelo de correlación estructural". Avances en Energías Renovables y Medio Ambiente 4, 2, 11.31-11.36, (2000).

Koopman, Harvey, Doornik y Shepard, "Stamp 5.0 Structural Time Series Analyser, Modeller and Predictor". 1a. edición. Chapman and Hall, London, (1995).