UN MÉTODO ESTADÍSTICO PARA ENCONTRAR LAS RELACIONES DE DEPENDENCIA ENTRE UN CONJUNTO DE VARIABLES IRREGULARMENTE MUESTREADAS EN EL TIEMPO

A STATISTICAL METHOD FOR FINDING THE DEPENDENCY RELATIONSHIPS BETWEEN A SET OF VARIABLES IRREGULARLY SAMPLED IN TIME

Sonia Yamile Roa Velandia[1], Gladys Elena Salcedo Echeverry[1], Fernando Roberto Momo[2].

[1] Grupo de Investigación y Asesoría en Estadística, Universidad del Quindío, Colombia. sonyaroa@gmail.com, gsalcedo@uniquindio.edu.co

[2] Instituto de Ciencias, Universidad Nacional General Sarmiento, Buenos Aires, Argentina. fmomo@ungs.edu.ar

Recibido: Mayo 15 de 2012
Aceptado: Junio 12 de 2012
*Correspondencia del autor. Grupo de Investigación y Asesoría en Estadística, Universidad del Quindío, Carrera 15 Calle 12 Norte, Armenia, Quindío, Colombia. gsalcedo@uniquindio.edu.co.

RESUMEN

El siguiente artículo propone un método que permite reconstruir una serie de tiempo igualmente espaciada a partir de otra serie que ha sido registrada en tiempos irregulares. Mediante un modelo autorregresivo irregular con parámetros variando en el tiempo, para la serie de tiempo desigualmente espaciada, se pueden interpolar nuevos valores para encontrar una serie equiespaciada similar a la original. Luego, el modelo es aplicado a seis series de tiempo irregulares provenientes de un estudio ecológico hecho sobre la columna de agua en la estación de monitoreo Paraná del Canal de Beagle, Tierra de Fuego (Argentina), en el cual se desea conocer las relaciones de interdependencia entre las variables de estudio para interpretar la dinámica del sistema.

Palabras clave: Causalidad, modelos autorregresivos, parámetros funcionales, series de tiempo, Tierra de Fuego.

ABSTRACT

This paper proposes a method that allows the reconstruction of an equally-spaced time series from another series recorded with irregular times. By using an irregular autoregressive model with parameters varying in time for the unequally spaced time series, new values can be interpolated to find an equispaced series similar to the original series. Then, the model is applied to a set of six irregular time series obtained from an ecological study conducted on the water column of the monitoring station Paraná, located in the Beagle Channel, Tierra de Fuego (Argentina). The study was designed to know the dependence relationships among the study variables which allow interpreting the temporal dynamics of the system.

Keywords: Causality, autoregressive models, functional parameters, time series, Tierra de fuego.

INTRODUCCIÓN

En estadística una *serie de tiempo* es un conjunto de observaciones registradas en el tiempo o el espacio. La mayoría de los análisis de las series de tiempo están basados en dos supuestos a veces difíciles de verificar en la práctica, la estacionariedad y la regularidad en el espaciamiento de las observaciones. La estacionariedad supone que las observaciones siempre oscilan alrededor de una recta constante que representaría la media, y que además la dispersión de los datos permanece dentro de una banda de amplitud constante. La regularidad de las observaciones tiene que ver con el equiespaciamiento de las mismas.

Las series de tiempo se pueden analizar desde un enfoque temporal o desde un enfoque espectral. En el primer caso, se utilizan los modelos estadísticos y en el segundo caso, las transformadas de Fourier, o actualmente, transformadas wavelet (1). Los procesos Autoregressivos (AR) forman una clase de modelos muy útil e importante debido al hecho que ellos pueden explicar una amplia variedad de fenómenos, son más fáciles de estimar y sus propiedades estadísticas asintóticas son más fáciles de estudiar. Por todas sus ventajas, los modelos AR han sido extendidos desde series estacionarias hasta series localmente estacionarias (2) y de dominios discretos a dominios continuos (3, 4), sin embargo la mayoría de estos modelos son desarrollados para series igualmente espaciadas.

En cuanto al supuesto de estacionariedad, pensar en modelos para procesos no estacionarios es casi imposible porque la no estacionariedad como tal no permite el desarrollo de una teoría asintótica. Los procesos localmente estacionarios presentados por Dahlhaus (5) constituyen una clase particular de procesos no estacionarios y son caracterizados especialmente por admitir una representación espectral variando en el tiempo. En los últimos años, ha evolucionado el desarrollo y aplicación de una familia de modelos utilizada para series no estacionarias, se trata de los modelos con coeficientes variando en el tiempo (6), y una clase particular de estos modelos son los autorregresivos con parámetros variando en el tiempo (2). Es de esta forma como se incorpora la no estacionariedad en los modelos, en lugar de someter los datos a transformaciones que pueden dificultar posteriormente la interpretación de los resultados.

En la literatura de las series de tiempo, no existe un modelo que incorpore a la vez tanto la no estacionariedad como la irregularidad del muestreo. Salcedo et al.(7) proponen *el modelo autorregresivo variando en el tiempo para series no estacionarias irregularmente espaciadas*. Un modelo como tal, no sólo permite interpretar la dinámica de una serie irregular no estacionaria, sino que también posibilita la predicción de una serie igualmente espaciada utilizando los valores de la serie irregular.

Por otra parte, cuando se desean conocer las interrelaciones entre varias series de tiempo, una manera de hacerlo con series equiespaciadas, es utilizando los modelos vectoriales autorregresivos (8, 9). No obstante, dichos modelos tampoco han sido adaptados al diseño irregularmente espaciado.

Con el fin de conocer las relaciones de causalidad entre un conjunto de series no estacionarias e irregularmente espaciadas, se propone en este artículo la realización de dos etapas: La primera consiste en utilizar el modelo de Salcedo et al.(7) para ajustar modelos irregulares a las series irregulares, y a través de estos, encontrar las series equiespaciadas más cercanas a las originales. La segunda etapa consiste en ajustar un modelo vectorial autorregresivo a las series regulares.

Este trabajo fue motivado a partir de un conjunto de datos obtenidos de un estudio ecológico hecho sobre la columna de agua en la estación Paraná, ubicada en una zona costera de Tierra de Fuego, extremo sur de Argentina y Chile (Ver Figura 1), en donde se requiere conocer las relaciones de interdependencia entre las variables Nitritos, Temperatura del agua, Fosfatos, pH, Silicatos y Clorofila, medidas en dicho sistema ecológico. Debido a las dificultades de acceso, bien sea por factores de tipo climático o del mismo terreno, las observaciones han sido registradas a intervalos de tiempo irregulares lo que imposibilita la aplicación de los modelos tradicionales de series de tiempo. La Figura 2 muestra las seis series de tiempo que representan las observaciones, las cuales van de Marzo de 2005 a Diciembre de 2006, con un espaciamiento irregular aproximado de 15 dias.

EL MODELO AUTORREGRESIVO (AR) IRREGULAR

En esta sección se presenta el modelo de Salcedo et al.(7) con $p = 1$ para analizar una serie de tiempo no estacionaria e irregularmente espaciada. Se trata del

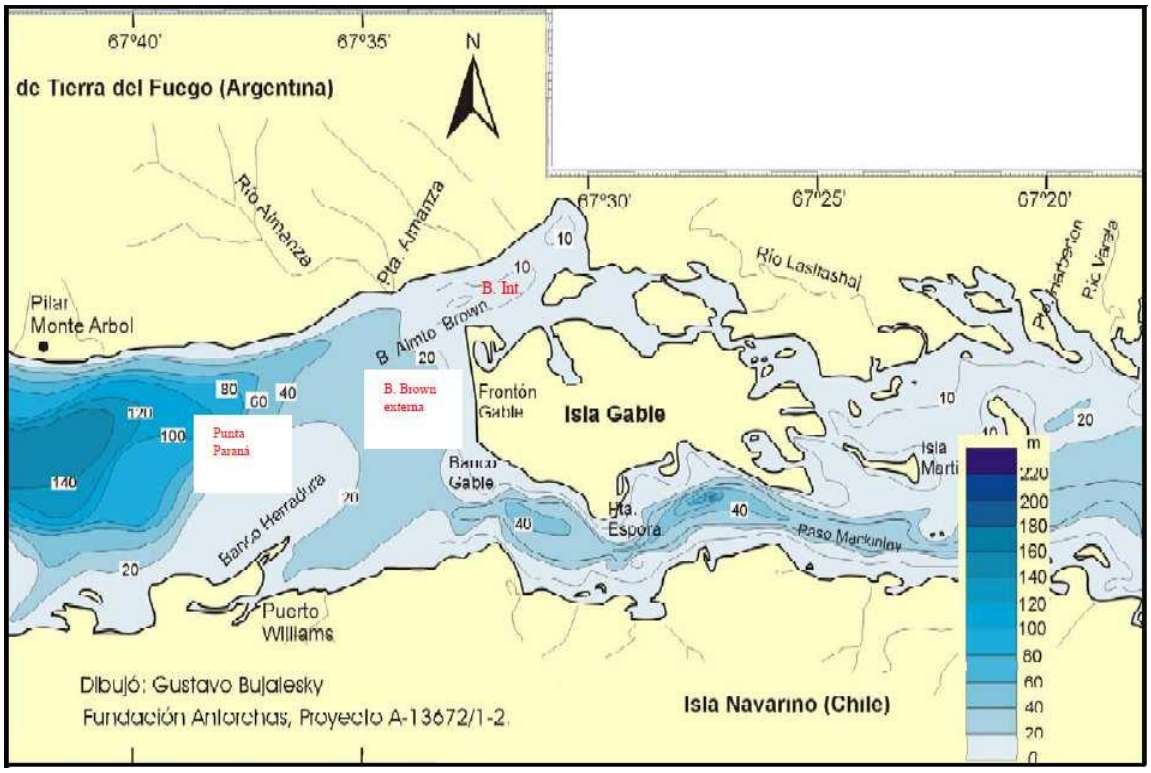

Figura 1. Estación de Investigación Paraná en Tierra de Fuego

modelo autorregresivo irregular de orden 1 con parámetro autorregresivo funcional y errores normales no correlacionados. Más concretamente, suponga que se dispone de un conjunto de datos no equiespaciados

$$\{X_{t_1}, X_{t_2}, ..., X_{t_T}\},$$

no necesariamente estacionarios. El modelo AR irregular de orden 1 es dado por

$$X_{t_i} = f(t_i) X_{t_{i-1}} + \epsilon_{t_i} \qquad [1]$$

donde $i = 1, 2, ..., T = 2^N$, $0 < t_1 < t_2 < ... < t_n = 1$, y los errores ϵ_{t_i} se suponen no correlacionados, normales e idénticamente distribuidos con media cero y varianza σ_ϵ^2.

El modelo [1] muestra la dinámica o la dependencia de una serie de tiempo irregular no estacionaria, donde $f(t_i)$ es el parámetro autorregresivo funcional y representa la incorporación de la no estacionariedad de la serie en el modelo. Por su parte, el subíndice i en el argumento de la función incorpora la irregularidad de las observaciones en el modelo. Los parámetros de interés son los valores funcionales $f(t_i)$ y la varianza de los errores σ_ϵ^2, los cuales deben ser estimados a partir de los datos. La estimación de este modelo se hace por la vía de los mínimos cuadrados ordinarios con una previa expansión en series tipo *wavelet* (1) del parámetro funcional $f(t_i)$, lo cual viabiliza la estimación, y para lo cual se hace necesario (sin pérdida de generalidad) que $T = 2^N$, $N \in \mathbb{N}$ donde $\mathbb{N}$ representa el conjunto de los números naturales. Por otra parte, la estimación de σ_ϵ^2 se hace calculando el error cuadrado medio de los residuales del modelo. Para más detalles estadísticos sobre la estimación del modelo [1] consultar (7).

RESULTADOS Y DISCUSIÓN

El modelo [1] es ajustado a cada una de las series de tiempo de la Figura 2. Inicialmente se estima el parámetro funcional $f(t_i)$ y a partir del modelo estimado dado por

$$\hat{X}_{t_i} = \hat{f}(t_i)\, X_{t_{i-1}} \qquad [2]$$

se puede ir reconstruyendo una serie regularmente espaciada en puntos t_i distintos a los observados. $\hat{X}_{t_i}$ representa el valor estimado y X_{t_i} representa el valor observado, ambos en el tiempo t_i.

Las Figuras 3, 4, 5, 6, 7, 8 muestran la estimación de las funciones autorregresivas $f(t_i)$, y la calidad del ajuste para cada una de las seis series de tiempo observadas. Más específicamente, en las figuras (a) aparece el parámetro AR funcional estimado $\hat{f}(t_i)$ para la respectiva serie de tiempo. En las figuras (b) aparecen las series de tiempo observadas (líneas continuas) y el correspon-

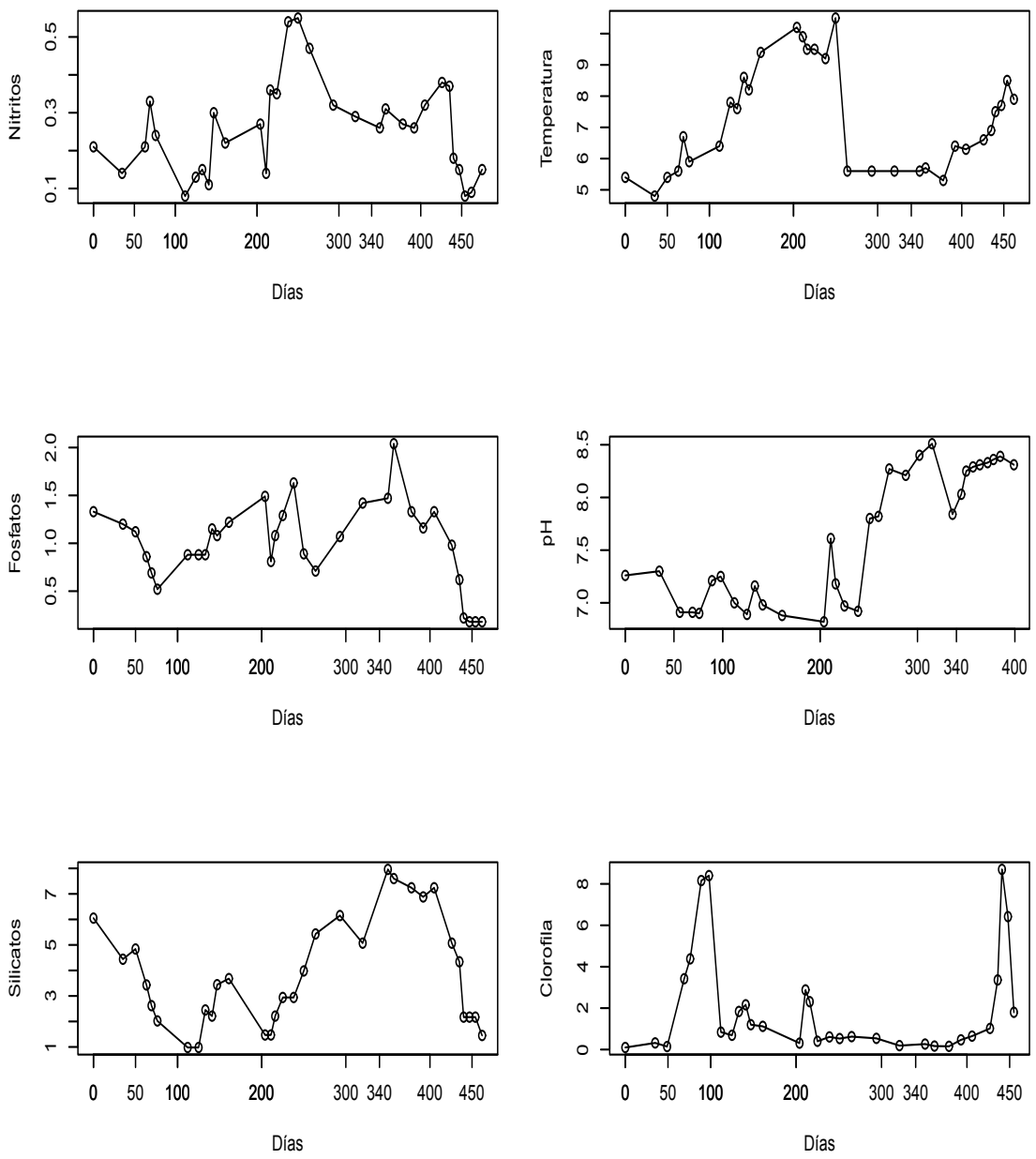

Figura 2. Series irregulares medidas en la Estación Paraná

diente ajuste después de estimado el modelo (líneas punteadas). En las figuras (c) aparecen los residuales estimados después del ajuste. Dichos residuales se calculan mediante la fórmula $\hat{\epsilon}_{t_i} = X_{t_i} - \hat{f}(t_i)\, X_{t_i-1}$ en los puntos t_i observados. En las figuras (d) aparece un gráfico tipo ANOVA (no parámetrico) para las desviaciones estandar de los coeficientes de la transformada wavelet de los residuales para algunas escalas de resolución. Se trata de un intervalo de confianza del 95 % para la mediana de dichas desviaciones estandar en las tres escalas de resolución más gruesas, a saber 3, 4 y 5 (ver (1) y (7)). La superposición de dichos intervalos de confianza indican la no correlación temporal de los residuales, lo cual sugiere que el modelo AR irregular de orden 1 estimado para las series irregulares de la Figura 2, realmente elimina la autocorrelación temporal. Por otra parte, dado que el modelo [1] asume normalidad de los residuales, se aplicó la prueba de Kolmogorov-Smirnov para los residuales de las figuras (c). Todos los *p*-valores dieron por encima de 0.06, lo cual sugiere que a un nivel de significancia del 5 % no se pueden rechazar las hipótesis que los errores siguen una distribución normal. Con lo anterior se deduce que el ajuste del modelo [1] es apropiado para las series de análisis. Finalmente, la estimación de σ^2_{ϵ} para cada serie se hace mediante el cálculo del error cuadrado medio de las

series residuales de las figuras (c), tales estimaciones fueron 0.0088, 0.9982, 0.0948, 0.0878, 0.939, 4.22 para los residuales de Nitritos, Temperatura del agua, Fosfatos, pH, Silicatos y Clorofila, respectivamente. Note que estas varianzas son relativamente pequeñas en comparación con la variabilidad de las series observadas.

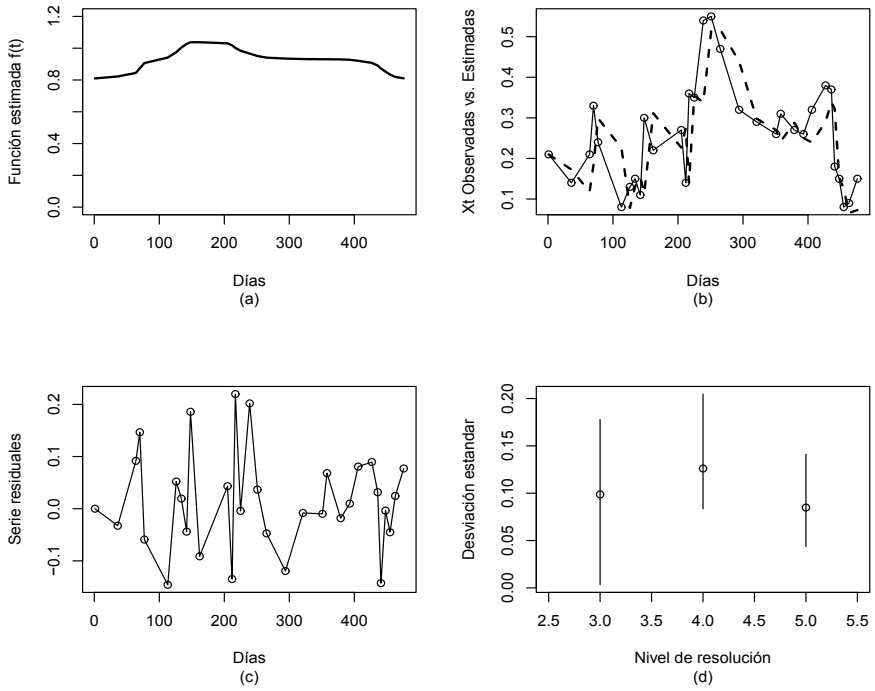

Figura 3. Modelo ajustado y función AR estimada para Nitritos.

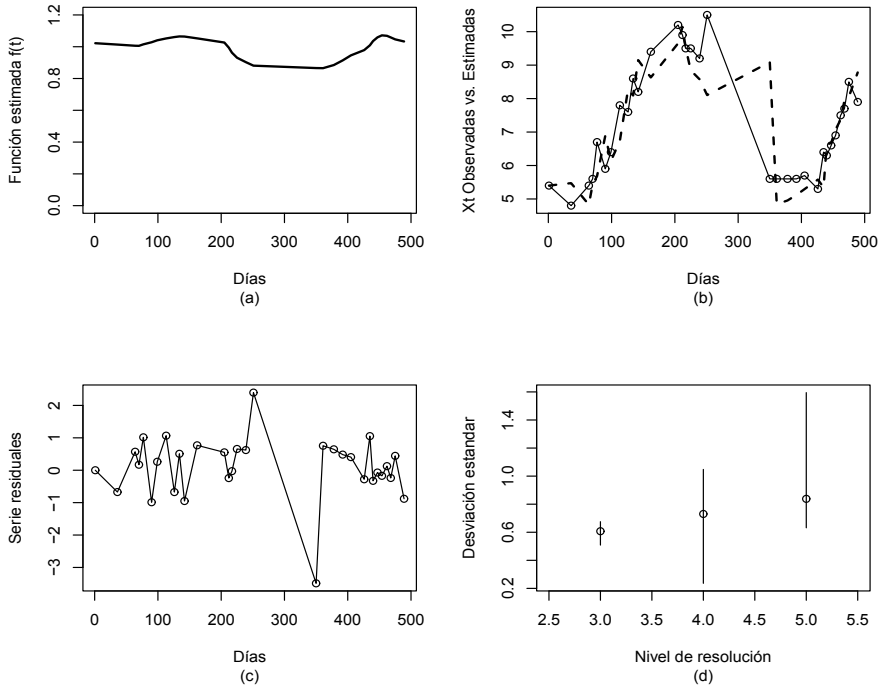

Figura 4. Modelo ajustado y función AR estimada para Temperatura del agua

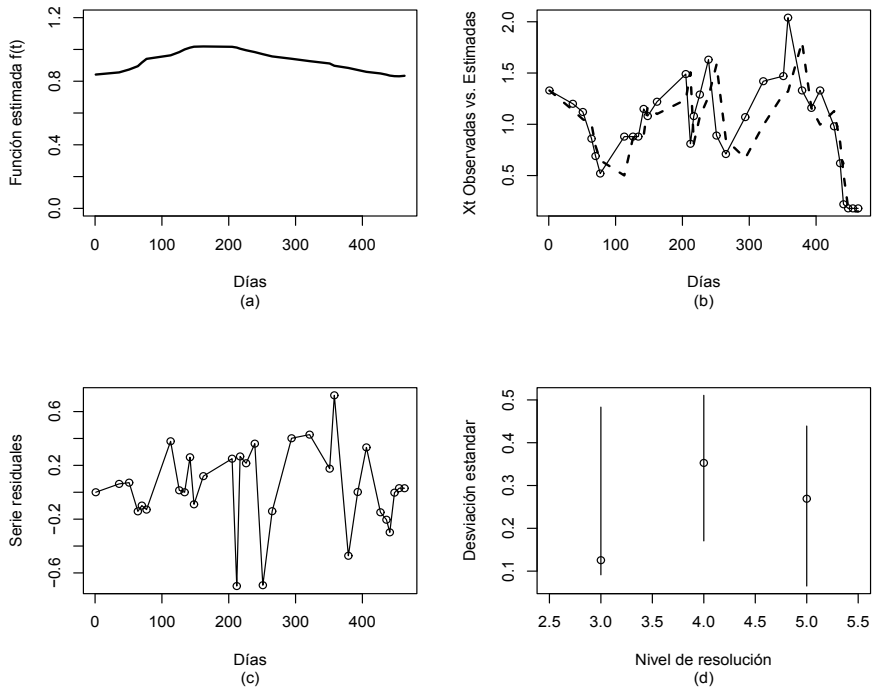

Figura 5. Modelo ajustado y función AR estimada para Fosfatos.

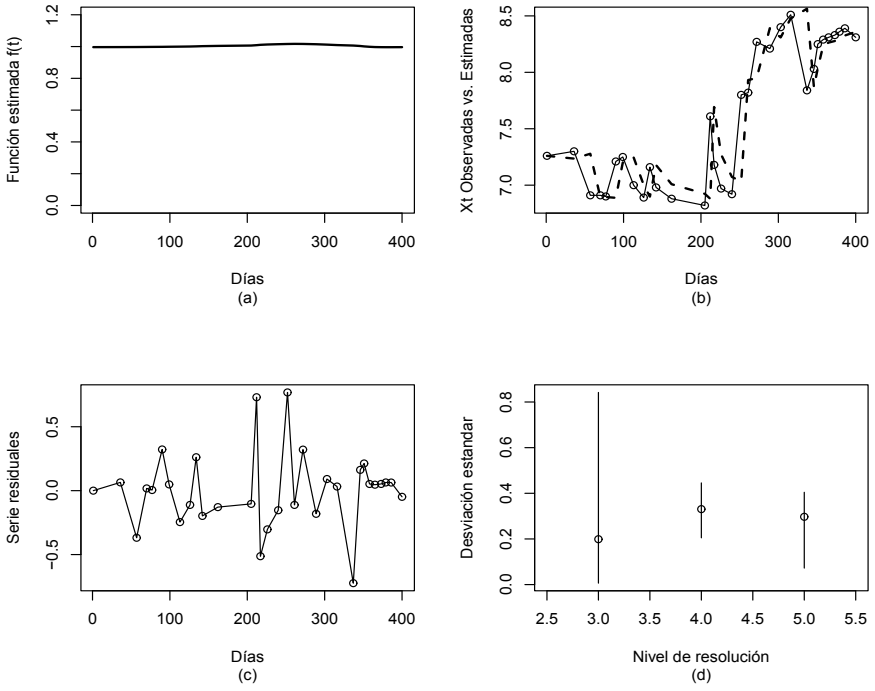

Figura 6. Modelo ajustado y función AR estimada para pH.

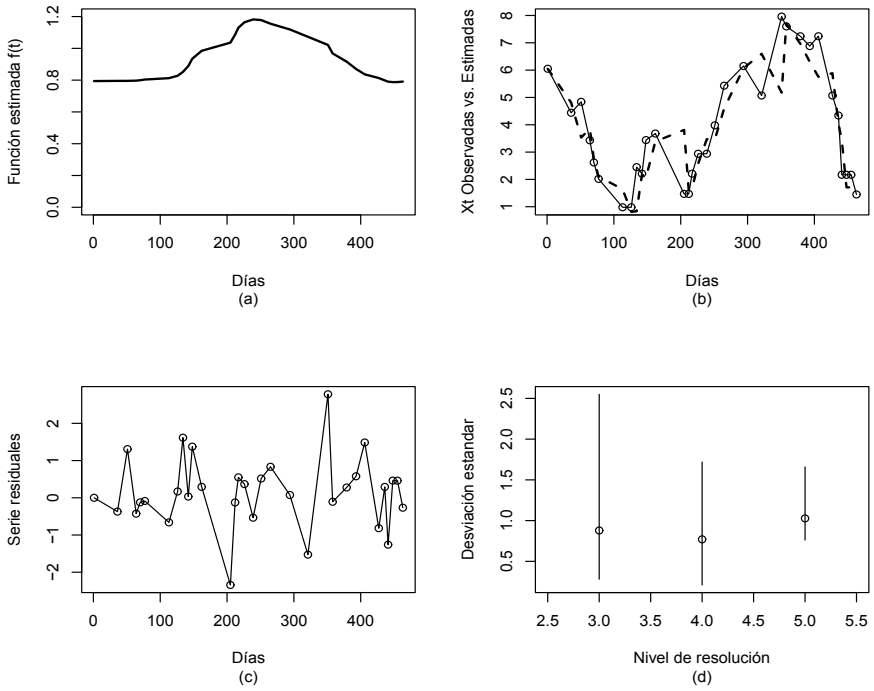

Figura 7. Modelo ajustado y funcón AR estimada para Silicatos.

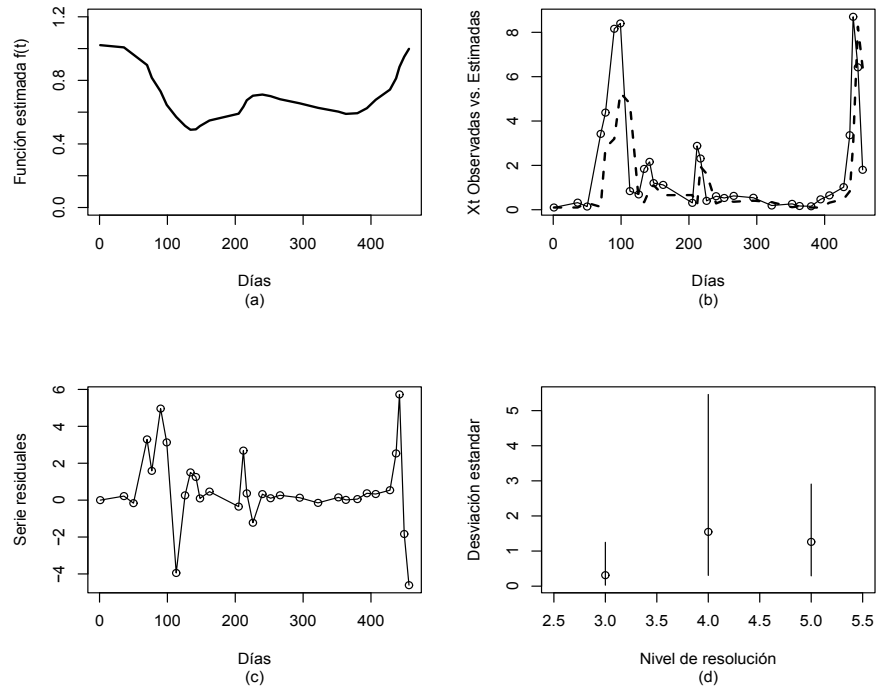

Figura 8. Modelo ajustado y función AR estimada para Clorofila.

El modelo Vectorial Autorregresivo

Para identificar las relaciones de interdependencia entre las seis variables de estudio, se ajusta un modelo vectorial autorregresivo (8, 9) para las seis series de tiempo igualmente espaciadas de la Figura 9, obtenidas a través del modelo [2].
Representando por Z_t el vector de las variables

$$Z_t = (NIT_t,\ TEM_t,\ FOS_t,\ pH_t,\ SIL_t,\ CLO_t)',$$

donde NIT_t, TEM_t, FOS_t, pH_t, SIL_t, CLO_t corresponden a las series Nitritos, Temperatura del agua, Fosfatos, pH, Silicatos y Clorofila, respectivamente, y donde t representa la unidad de tiempo de las series equiespaciadas. En este caso se hizo una interpolación semanal, en el mismo período de muestreo.
Con la ayuda de la Función Matricial de Correlación Parcial y la Función Matricial de Correlación Cruzada a rezagos $k = 1, 2, \ldots 8$ (9), se identificó que el modelo más apropiado para las series es un modelo vectorial AR de orden 1. En forma más explícita el modelo corresponde a

$$Z_t = \Phi_1 Z_{t-1} + \epsilon_t, \qquad [3]$$

donde el vector Z_{t-1} contiene las variables rezagadas una unidad y es dado por

$$Z_{t-1} = (NIT_{t-1}, TEM_{t-1}, FOS_{t-1}, pH_{t-1}, SIL_{t-1}, CLO_{t-1})',$$

y la matriz de parámetros Φ_1 estimada, está dada por

$$\hat{\Phi}_1 = \begin{pmatrix} 0{,}7872 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0{,}8522 & 0 & -0{,}2928 & 0 & 0 \\ -0{,}4666 & 0 & 0{,}7736 & 0 & 0 & 0 \\ 0{,}9641 & 0 & 0 & 0{,}8787 & 0 & 0 \\ 0 & 0 & 0 & 0{,}6791 & 0{,}5802 & -0{,}1633 \\ 3{,}2681 & -0{,}286 & -1{,}195 & -0{,}9128 & 0 & 0{,}5896 \end{pmatrix}.$$

Cuando el parámetro matricial Φ_1 se reemplaza por su estimador $\hat{\Phi}_1$ en el modelo [3] y expandiendo el modelo vectorial estimado, se obtiene el siguiente sistema de ecuaciones:

$$\begin{aligned} NIT_t &= 0{,}787 NIT_{t-1} + \hat{\epsilon}_{1,t} \\ TEM_t &= 0{,}852 TEM_{t-1} - 0{,}293 pH_{t-1} + \hat{\epsilon}_{2,t} \\ FOS_t &= -0{,}466 NIT_{t-1} + 0{,}774 FOS_{t-1} + \hat{\epsilon}_{3,t} \\ pH_t &= 0{,}964 NIT_{t-1} + 0{,}879 pH_{t-1} + \hat{\epsilon}_{4,t} \\ SIL_t &= 0{,}679 pH_{t-1} + 0{,}58 SIL_{t-1} - 0{,}163 CLO_{t-1} + \hat{\epsilon}_{5,t} \\ CLO_t &= 3{,}268 NIT_{t-1} - 0{,}286 TEM_{t-1} - 1{,}195 FOS_{t-1} - 0{,}913 pH_{t-1} + 0{,}59 CLO_{t-1} + \hat{\epsilon}_{6,t}. \end{aligned}$$

Así por ejemplo, la primera ecuación indica que los Nitritos en una semana dada dependen de los mismos Nitritos en la semana anterior. La tercera ecuación indica que los Fosfatos en una semana dada dependen de los Nitritos y de los mismos Fosfatos de la semana anterior. La última ecuación indica que la Clorofila en una semana dada depende de los Nitritos, la Temperatura del agua, los Fosfatos, el pH y la misma Clorofila de la semana anterior. Estas relaciones de causalidad se pueden resumir en forma gráfica como en la Figura 10, donde el sentido de cada flecha indica la dirección en que se da la influencia entre las dos variables (círculos) que están en los extremos.

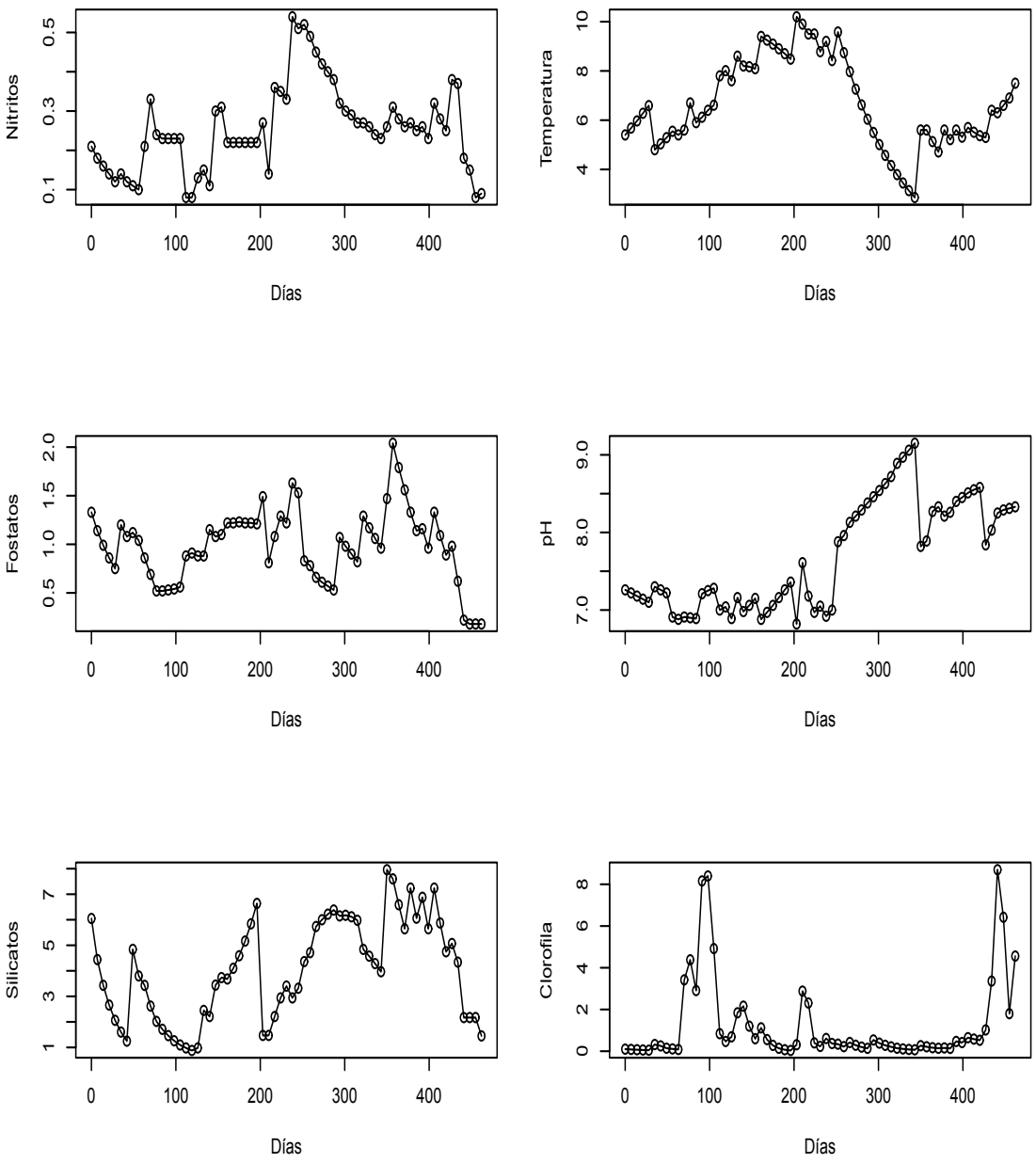

Figura 9. Series regulares obtenidas de interpolación

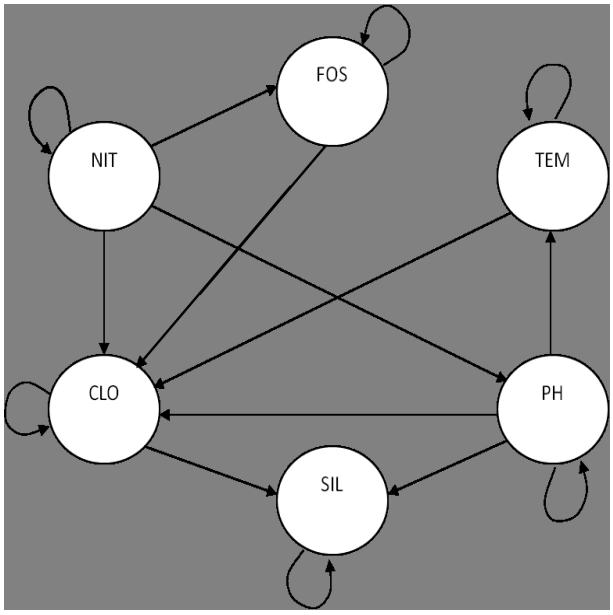

Figura 10. Series regulares obtenidas de interpolación

Note en especial que, la variable más dependiente es la Clorofila, lo cual quiere decir que ante alguna variación de Nitritos, Fosfatos, Temperatura del agua, pH y de la misma Clorofila, ésta se altera en la semana siguiente. Por su parte, los Nitritos son los menos dependientes, o la única variable independiente de las demás. Por el contrario, las variables como el pH y los Nitritos se dice que son las más motrices, una vez que ellas son las que más influyen sobre las demás, indicando que un cambio en ellas se va a reflejar en varias variables. Desde el punto de vista biológico, el modelo refleja acertadamente el comportamiento general de las correlaciones entre las variables en la zona. Así, la ecuación del nitrito es razonable, al depender sólo de sí mismo está indicando que probablemente haya una fuente debida al aporte de este nutriente desde aguas más profundas, básicamente un ingreso de aguas ricas en nitritos, se trata entonces de un nutriente poco oxidado, por lo cual probablemente provenga de aguas del fondo donde la descomposición se realiza con poco oxígeno. La ecuación del fosfato es acertada porque las algas (CLO) incorporan el nitrógeno y el fósforo en una relación particular (básicamente 16 átomos de nitrógeno por cada átomo de fósforo) esto hace que cuando hay mucho de uno, el otro se consume más rápido y entonces eso probablemente genera una autodependencia negativa en el fosfato. Además es correcto que el silicato dependa negativamente de las algas (CLO) de la semana anterior ya que la mayoría de las especies que participan de estas floraciones, son diatomeas, algas que usan silicato para su pared celular. Cabe aclarar que las relaciones de causalidad entre las variables consideradas en este modelo son válidas solamente para el ecosistema en estudio. Para otros ecosistemas dichas relaciones pueden cambiar.

CONCLUSIONES

Cuando los datos son registrados en el tiempo no es apropiado aplicar los modelos tradicionales de regresión sino que se deben utilizar los modelos para series de tiempo. Si además, las series no son estacionarias ni equiespaciadas un modelo AR irregular debe ser considerado. Las relaciones de causalidad entre varias variables de estudio pueden encontrarse mediante un modelo vectorial para series de tiempo equiespaciadas desde que se pueda obtener una muestra equiespaciada a partir de una irregular. En este documento se proporciona una forma de hacerlo.

AGRADECIMIENTOS

Los autores agradecen en especial al profesor Dr. Marcelo Hernando Perez por proporcionar los datos, a la Universidad del Quindío y a COLCIENCIAS por el apoyo financiero al proyecto 111352128221.

BIBLIOGRAFÍA

1. D.B. Percival and A.T. Walden, Wavelet methods for time series analysis, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge, (2000).
2. R. Dahlhaus, M.H. Newmann and R. von Sachs, Nonlinear wavelet estimation of time-varying autoregressive processes, Bernoulli, 5, 5 (1999), 873–906.
3. P.M.T. Broersen and R. Bos, Estimating time series models from irregularly spaced data, MIEEE transactions on Instrumentation and Measurement, (2006).
4. R.H. Jones and Y. Zhang, Models for continuous stationary space-time processes. In Modelling Longitudinal and Spatially Correlated Data: Methods, Applications and Future Directions (Gregoire, T.G. et al., Editors). Lecture Notes in Statistics; 122, Springer-Verlag, New York, 289-298, (1997).
5. R. Dahlhaus, Fitting time series models to non-stationary processes, Annals of Statistics, 25, (1997), 1–37.
6. C. Chiann and P.A.Morettin,Time domain nonlinear estimation of time varying linear systems, Journal of Nonparametric Statistics, 17 (2005), 365–383.
7. G. Salcedo, R.F. Porto, S.Y. Roa and F.R. Momo, A wavelet-based time-varying autoregressive model for non-stationary and irregular time series, Journal of Applied Statistics, 39,11 (2012).
8. G.C. Reinsel, Elements of multivariate time series analysis, Springer Verlag, New York, (1991).
9. W.W. Wei, Time series analysis: Univariate and multivariate methods, Addison-Wesley Publishing Company, California, (1990).