Efecto de diversos métodos de preprocesamiento matemático al completar datos faltantes en los monitoreos del complejo lagunar Ciénaga Grande de Santa Marta, mediante el enfoque de atípicos aditivos

Effect of various preprocessing methods in filling missing observations by additive outliers approach, during the monitoring of Complejo Lagunar Ciénaga Grande de Santa Marta

Recibido para evaluación: 16 de Febrero de 2009
Aceptación: 04 de Noviembre de 2010
Recibido versión final: 23 de Noviembre de 2010

Marcos A. Carvajalino Fernández*
Walberto A. Troncoso Olivo**

RESUMEN

Se presentan los resultados de la aplicación de diversas técnicas de preprocesamiento de datos con la finalidad de adecuar las series de tiempo de las variables fisicoquímicas medidas en el complejo lagunar Ciénaga Grande de Santa Marta a los requerimientos mínimos de la metodología de determinación de observaciones faltantes mediante atípicos aditivos a través del software Tramo and Seats for Windows (TSW: TRAMO and SEATS for WINDOWS, Gómez & Maravall, 2009). Las series utilizadas corresponden a los registros históricos de un proyecto de monitoreo para la rehabilitación de la zona entre 1993 y 2008. Se evaluaron cuatro enfoques de preprocesamiento: Interpolación histórica, escalamiento, división y división con interpolación histórica. Los cuatro métodos fueron comparados aplicándolos a 4 variables en 3 estaciones de monitoreo dentro del sistema lagunar.

Los métodos de interpolación histórica y escalamiento mostraron resultados favorables al preparar los datos de las series históricas para el uso del método de atípicos aditivos, con valores de cuadrado medio del error (MSE) de los residuales entre 0.172 x 10-2 y 19.28 para la interpolación histórica y entre 0.232 x 10-2 y 17.818 para el escalamiento. Se recomienda el uso de interpolación histórica en casos de series cortas con vacíos distribuidos aleatoriamente o en estudios donde la frecuencia original del muestreo sea un factor decisivo, mientras que en casos de series con vacíos agrupados y mayor longitud, deben ser tratados mediante escalamiento de la frecuencia de muestreo.

Palabras Clave: CGSM, enfoque de atípicos aditivos, MSE, observaciones faltantes, series de tiempo.

ABSTRACT

Results on application of various data preprocessing techniques in order to adapt time series of physicochemical variables from Complejo Lagunar Ciénaga Grande de Santa Marta to the minimal requirements for the additive outliers approach for determining missing observations through *Tramo and Seats for Windows (TSW)* software are presented. The series used are historical surveys from a monitoring project for the rehabilitation of the area between 1993 and 2008. Four preprocessing approaches were evaluated: Historical interpolation, scaling, division and division plus historical interpolation. These four methods were compared through test on 4 variables in 3 monitoring stations inside the lagoon complex.

Historical interpolation and scaling methods showed favorable results preparing historical series data for the application of additive outliers approach showing mean square errors (*MSE*) for the residuals ranging from 0.172 x 10^{-2} to 19.28 for historical interpolation and from 0.232 x 10^{-2} to 17.818 for scaling. Use of historical interpolation is recommended in short series cases with randomly distributed holes or studies were original monitoring frequency is a decisive factor, while cases of longer series with grouped holes should be treated through scaling of the monitoring frequency.

Key Words: Additive outliers approach, CGSM, missing observations, MSE, time series.

***1. Ing. Ambiental y Sanitario, Cand. Esp. Gerencia de Proyectos.* Pasante Línea de Monitoreo y Evaluación de Efectos, Programa de Calidad Ambiental Marina (CAM).**

2. Mg. Biología Marina. Coordinador Línea Monitoreo y Evaluación de Efectos, Programa de Calidad Ambiental Marina (CAM).

Instituto de Investigaciones Marinas y Costeras «José Benito Vives de Adréis» (INVEMAR). Colombia

marcos.carvajalino@gmail.com

maancafe240@gmail.com

1. INTRODUCCIÓN

Monitorear el comportamiento de las variables fisicoquímicas en el complejo lagunar Ciénaga Grande de Santa Marta (CGSM) ha sido labor constante del Instituto de Investigaciones Marinas y Costeras (INVEMAR) durante más de 10 años, con apoyo de la Corporación Autónoma Regional del Magdalena (CORPAMAG) y de la Unidad de Parques Nacionales Naturales, debido a la gran importancia natural que reviste dicho ecosistema para el desarrollo y protección de los recursos naturales tanto a nivel local, regional, nacional e internacional. El complejo lagunar ha sido objeto de diversas investigaciones y declarado como sitio de importancia según la convención RAMSAR para el fomento y la protección de los humedales y las zonas de migración y conservación de aves y peces estuarinos en 1998 (Rivera & Caicedo, 1998).

La labor de monitoreo incluye las actividades de planificación, toma de muestras, análisis de datos y publicación de resultados a través del Sistema de Información Ambiental Marina (SIAM), para las cuales se cuenta con recursos y metodologías concertadas y validadas. Pese a lo anterior, diversos inconvenientes de índole técnica, humana, de movilidad o de seguridad han causado que la continuidad del monitoreo se haya visto interrumpida en varias oportunidades, generando un número importante de observaciones faltantes en las series de tiempo almacenadas y por ende vacíos de información que hacen difícil el análisis de datos históricos.

Una de las iniciativas del programa de Calidad Ambiental Marina (CAM) propone establecer una metodología para el análisis de datos históricos correspondientes a las variables fisicoquímicas del complejo lagunar CGSM utilizando herramientas estadísticas conocidas como análisis de series de tiempo, las cuales incluyen entre otras, descomposición en factores estacionarios y tendencias generales, descripción de comportamientos acumulativos por periodo de tiempo, métodos de alisamiento exponencial, modelos ARIMA y predicción de valores futuros.

El presente trabajo reúne los resultados de la fase previa a la aplicación de las técnicas mencionadas, consistente en proponer una solución a las observaciones faltantes en las series de tiempo mediante el preprocesamiento de los datos existentes y la interpolación final, utilizando la metodología de atípicos aditivos a través del software *TSW*.

Ninguna técnica de preprocesamiento de datos puede reemplazar la verosimilitud que entrega un correcto esquema de monitoreo y su adecuada implantación y cumplimiento; sin embargo en casos en los cuales el problema de las observaciones faltantes ya exista, se puede utilizar algunas técnicas que mantengan una lógica consistente con las bases matemáticas del proceso.

En el pasado, el INVEMAR en convenio con la Universidad del Quindío realizó una serie de estudios (Hurtado, 1997) que buscaban analizar y aplicar la metodología de atípicos aditivos ideada por Peña y Maravall (1990) como una forma de interpolar las observaciones faltantes de las antiguas bases de datos CIENAGA I y CIENAGA II (Suárez & Torres, 1997) para estimar modelos matemáticos que interrelacionan las variables fisicoquímicas monitoreadas. El éxito de dichas iniciativas radicó principalmente en la existencia de una base de datos homogénea y con un número bajo de observaciones faltantes respecto al número total de datos existentes, lo cual debería ser la condición normal de un programa de monitoreo eficaz que permita utilizar el enfoque de atípicos aditivos para completar los datos por medio de una metodología estadística eficaz.

Actualmente, el INVEMAR cuenta con bases de datos de variables fisicoquímicas para la CGSM con registros desde el año 1993 hasta la fecha a través del proyecto de monitoreo de las condiciones ambientales y las condiciones estructurales y funcionales de las comunidades vegetales y de los recursos pesqueros durante la rehabilitación de la Ciénaga Grande de Santa Marta (INVEMAR, 2009). Sin embargo, como ha sido mencionado, dichas bases de datos presentan gran cantidad de observaciones faltantes (entre 35 y 51% del total de datos en algunas estaciones), haciéndose necesario un tratamiento previo en busca de reducir el número de las mismas, de manera que se cumplan los requisitos mínimos para utilizar el programa *TSW* e interpolar de manera fiable la mayor cantidad posible de datos ausentes.

Se evaluaron los siguientes pretratamientos para encontrar el que menores inconvenientes presentara para ajustar las series a los requerimientos del software:

Volumen 13 - No. 3, Diciembre de 2010, Medellín ISSN 0124.177X.

1. Aplicar el criterio de media aritmética si se encontraba una observación faltante entre dos observaciones reales, asumiendo un comportamiento correlativo entre tres meses continuos (*interpolación histórica*).
2. Reducir los datos desde la frecuencia actual (mensual) a una de menor frecuencia (bimestral, trimestral, etc…) utilizando reglas de decisión y promedios (*Escalamiento*).
3. Separar la serie de tiempo en conjuntos más pequeños a fin de aplicar el proceso de atípicos aditivos a varios conjuntos y después unificarlos (*División*).
4. Separar la serie de tiempo en conjuntos más pequeños y aplicarles el criterio de interpolación histórica para tener mayor cantidad de datos (*División con interpolación histórica*).

La mayoría de estos enfoques son utilizados de forma empírica en instituciones dedicadas al análisis de información como el INVEMAR; estos métodos hacen parte del principio de «*imputación*» y son adaptados de Enders (2010, Cap. 2), una explicación más detallada de los mismos se encuentra en la sección 4.

2. ÁREA DE ESTUDIO

El acopio de datos correspondiente a la calidad fisicoquímica de las aguas de la CGSM se realiza a través de múltiples estaciones de monitoreo distribuidas dentro del complejo lagunar en intervalos de tiempo definidos por el diseño de la red. Dichos intervalos han sufrido variaciones en su frecuencia entre 1993 y 2008 y esta irregularidad es uno de los factores que incide en la presencia de observaciones faltantes para la serie mensual; se utilizaron los datos correspondientes a la zona del espejo lagunar CGSM (11º0' – 10º43' N, 74º16' – 74º31' O) por presentar la mayor cantidad de observaciones: la zona se distribuye en 9 estaciones perimetrales y 1 estación de monitoreo central al cuerpo de agua con registros de multiplicidad de variables (Figura 1) .

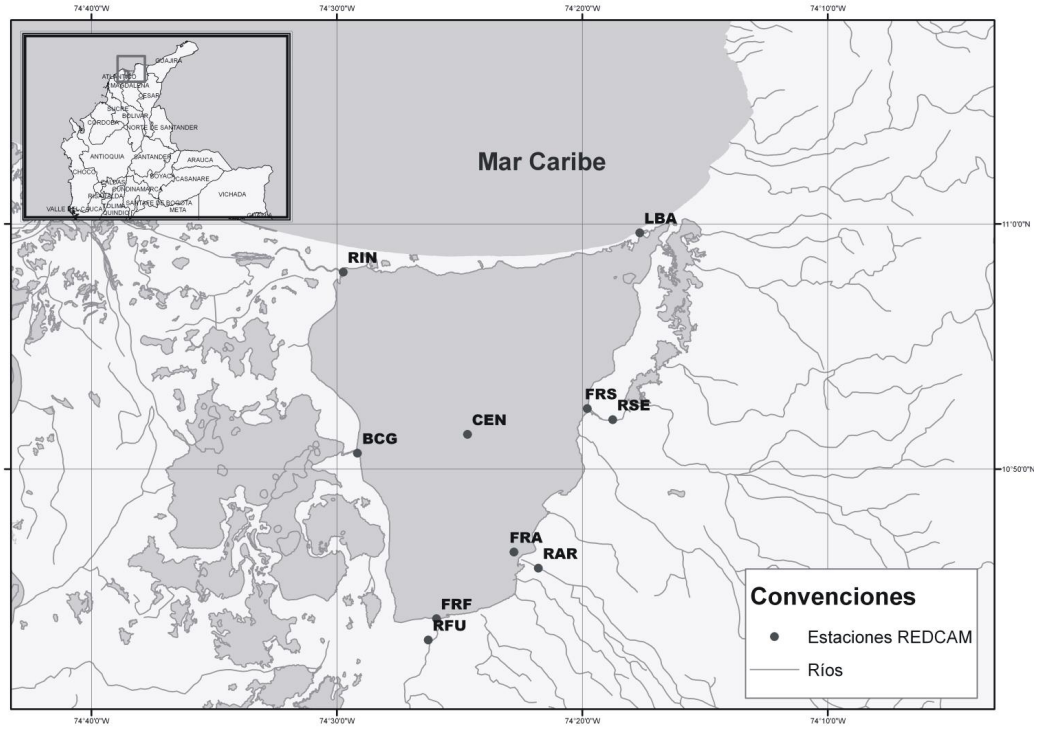

Figura 1. Área de estudio del complejo lagunar CGSM y estaciones de muestreo (Ver. Tabla 1 para nombres in extenso)

3. MARCO REFERENCIAL

La metodología de estimación de faltantes por *atípicos aditivos* fue abordada por Peña y Maravall (1990) para procesos estocásticos lineales posiblemente no estacionarios en virtud de que la mayoría de técnicas para analizar series de tiempo requieren que éstas sean *regulares* y completas en su secuencia de observaciones (no existan observaciones faltantes), lo cual es muy poco frecuente.

El método de atípicos aditivos depende únicamente de las observaciones y la función de autocorrelación inversa o dual en la serie (Gómez *et al.,* 1993); el proceso se basa en que la expresión de la esperanza condicional de la observación faltante (valor esperado, dados los datos existentes) es la misma que resulta de agregar un número arbitrario al faltante, considerar y manejar la observación como un atípico aditivo (observación que no sigue el comportamiento normal de la serie o su modelo) y finalmente retirar el efecto del atípico por medio de un análisis de intervención a la observación propuesta (Peña y Maravall, 1990).

Una explicación rigurosa del trasfondo matemático del método ha sido determinado por Gómez *et al.* (1993), Maravall y Peña (1992), Peña y Maravall (1990), Suárez y Torres (1997), el método considera que sobre una secuencia de *n* periodos se observa la serie:

$$Z=\left(Z_{t_1},\ldots,Z_{t_m}\right)\rightarrow\left(t_1\le t_2\le\cdots\le t_m\right)$$

Donde $m<n$ y por ende existen $h=m-n$ datos faltantes, ahora se considera que la serie completa $\tilde{Z}=\left(Z_1,\ldots,Z_n\right)$ con todos los datos (faltantes y existentes) sigue el modelo general ARIMA:

$$\Phi(B)\tilde{Z}_t=\Theta(B)a_t$$

donde $\Phi(B)$ y $\Theta(B)$ son polinomios finitos del operador de retraso B y a_t es un proceso de ruido blanco.

Se rellenan los h vacíos con números arbitrarios y se construye una nueva serie «observada» mediante:

$$Z_t=\begin{cases} z_t+w_t & ,t_1,\ldots,t_h \\ z_t & De\ lo\ contrario \end{cases}$$

De lo anterior se identifica w_t como un elemento del vector W de parámetros desconocidos que representa los desfases entre los valores esperados y los números arbitrarios escogidos; el tratamiento de los datos arbitrarios como atípicos aditivos continúa con una serie de complejas demostraciones que incluyen mínimos cuadrados generalizados, factorización de Cholesky, algebra matricial, etc… que no serán abordadas en el presente estudio. Finalmente se llega a la conclusión de que un estimador apropiado para W sería:

$$\hat{w}=R_D^{-1}\rho_D(B)Z^k$$

Donde R_D es la matriz simétrica de orden $(k\times k)$ cuyos elementos son las autocorrelaciones duales del proceso y Z^k es el vector de valores arbitrarios. Finalmente se retira el efecto del atípico aditivo obteniendo los estimadores de las observaciones faltantes como:

$$\hat{z}^k=Z^k-\hat{w}$$

Volumen 13 - No. 3, Diciembre de 2010, Medellín ISSN 0124.177X.

Todo este proceso se encuentra automatizado y disponible gracias al desarrollo del software libre *TSW* (Gómez & Maravall, 1996), el cual posee una interfaz gráfica amigable al usuario final e ingreso de datos en formatos de entrada predefinidos para el mismo. Este programa permite determinar automáticamente los parámetros del modelo con el cual se realizará la interpolación o ingresarlos de forma manual (caso de series con comportamientos conocidos). Pese a su gran flexibilidad el programa *TSW* tiene algunas limitaciones para su uso, entre ellas está el hecho de que no puede generar más de 31 regresores (interpolar más de 30 datos), adicionalmente la relación entre cantidad de observaciones reales disponibles y faltantes en la serie que se desee interpolar debe ser suficiente para realizar el modelado por métodos ARIMA (en el presente estudio series mayores a 60 datos y con faltantes <45% del total generaron resultados positivos).

Debido a las limitaciones ya mencionadas y la naturaleza de las bases de datos utilizadas en el presente trabajo, ninguna serie de tiempo disponible se ajusta a los requerimientos del software debido principalmente a la gran cantidad de faltantes (Figura 2) y por ende se hace necesario aplicar algún método de preprocesamiento con el fin de cumplir dichos requerimientos mínimos.

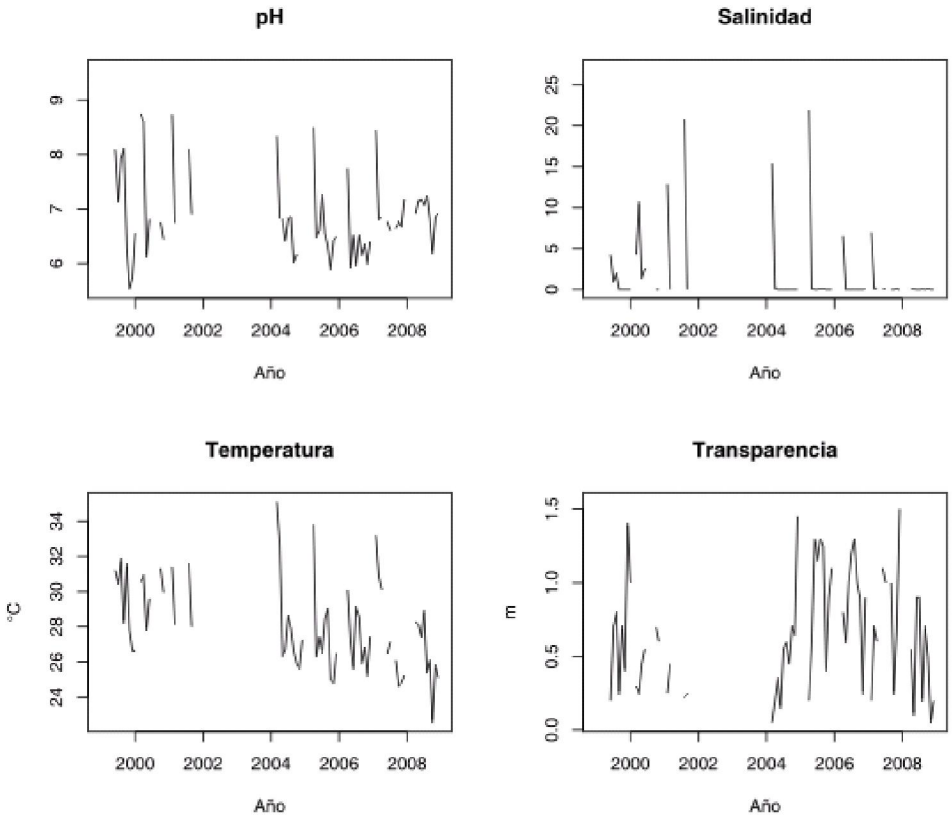

Figura 2. Estado general de las series de tiempo para cuatro variables fisicoquímicas tomadas del monitoreo del complejo lagunar CGSM (Estación: Río Aracatáca), se pueden visualizar la gran cantidad de observaciones faltantes en diferentes periodos

4. Materiales y Métodos

4.1 Preprocesamiento

Se realizó una comparación entre cuatro métodos de preprocesamiento y adecuación de datos previo a la utilización del software *TSW*, seleccionados de acuerdo a una revisión bibliográfica de diversas investigaciones en las que se ha hecho necesario completar datos (Astel *et al.*, 2004, Simeonov *et al.*, 2003, Divino & McAleer, 2009) y consultando experiencia y criterios de

personas asociadas al presente trabajo. A continuación se enumeran y explican las bases y técnicas empleadas:

Interpolación histórica

Se da el nombre de *interpolación histórica* en el marco del presente trabajo al procedimiento de usar la *media aritmética* como criterio de reemplazo en observaciones faltantes cuyos datos más próximos (observación previa y posterior) sean conocidos, es decir, considerar una observación faltante como el promedio de su valor en el monitoreo previo y en el monitoreo subsiguiente siempre y cuando estos sean medidas reales (Figura 3). Esta es la medida más común que se adopta al encontrar una observación faltante en un conjunto de datos y se puede encontrar multitud de ejemplos en la literatura científica (e.g. Simeonov *et al*., 2003, Zhou *et al*., 2007a, Zhou *et al,* 2007b).

El objetivo de este método radica en disminuir el número de observaciones faltantes sin aumentar demasiado la incertidumbre de la serie y sin recaer en propagación del error al no utilizar observaciones interpoladas como base para nuevos cálculos.

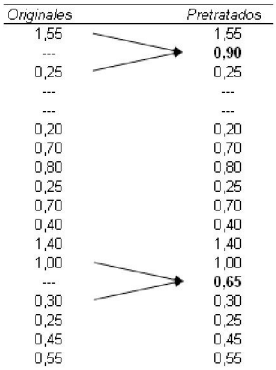

Figura 3. Método de interpolación histórica para series de tiempo, las flechas indican casos viables para la aplicación del método.

Escalamiento

Esta técnica se basa en redimensionar la frecuencia de muestreo de un valor mayor (meses) a un valor menor (i.e. trimestres, semestres,…); de esta manera podría usarse una regla de decisión de acuerdo a la cantidad de información contenida en la frecuencia mayor dentro de cada intervalo de la frecuencia menor para elegir si la observación será una medida de tendencia de las observaciones existentes o se considerará un vacío en la información (Figura 4). Para el presente trabajo se realizó escalamiento de meses a trimestres, usando la media aritmética como medida de tendencia central.

La regla de decisión a aplicar se puede resumir como: Considérese una serie de tiempo $X=\left(X_{t_1},\ldots,X_{t_n}\right)\rightarrow\left\{t_1\leq t_2\leq\cdots\leq t_n\right\}$ y muestreo anual f_x (f_x= 12 para meses) que contiene observaciones faltantes, y además considérese una nueva serie escalada $Y=\left(Y_{t_1},\ldots,Y_{t_m}\right)\rightarrow\left\{t_1\leq t_2\leq\cdots\leq t_m\right\}$ con muestreo anual $f_y<f_x$ (f_y = 4 para trimestres), los valores de la serie Y siguen el criterio:

$$Y_t=\begin{cases}\dfrac{\sum_{i=1}^{l}X_{it}}{l} & ,vacios<0{,}51\\ \text{"}vacio\text{"} & De\ lo\ contrario\end{cases}$$

Donde $l=\dfrac{f_x}{f_y}$, de esta forma se logra una nueva serie de promedios trimestrales con menor cantidad de vacíos.

Volumen 13 - No. 3, Diciembre de 2010, Medellín ISSN 0124.177X.

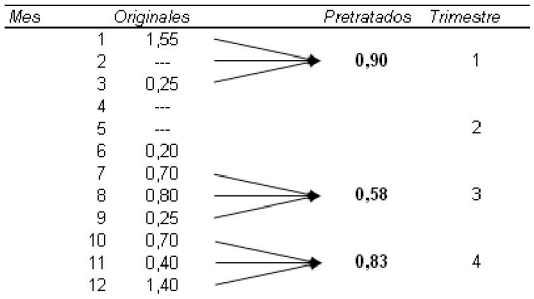

Figura 4. Método de interpolación por escalamiento, los valores al final de las flechas corresponden a promedios trimestrales de los datos originales, periodos vacíos indican no cumplimiento de la regla de decisión.

División

El método de división consiste en repartir ordenadamente la serie de tiempo en un número *n* de subseries de forma que se pueda aplicar el procedimiento de atípicos aditivos por aparte a cada subserie y posteriormente unir los resultados; Martínez *et al.* (1996) realiza una aplicación de este método para evitar una limitante en el paquete estadístico *SCA*[2].

Para permitir la aplicación del programa *TSW* a pesar de sus limitaciones se seleccionaron las subseries manteniendo los criterios de menos de 30 observaciones faltantes y menos del 45% del total de observaciones por serie como vacíos de información.

División con interpolación histórica

Este método se aplica como una alternativa al enfoque de división pues se puede presentar que la cantidad de observaciones faltantes impida el uso eficiente del mismo. Consiste en utilizar inicialmente sobre la serie un método de interpolación histórica y posteriormente dividirla en rangos adecuados.

2. Scientific Computing Associates Software

4.1 Procedimiento general

Del banco de datos del espejo lagunar CGSM, se escogieron 12 variables de importancia ambiental para realizar las pruebas de los pretratamientos, estas fueron: *Amonio, clorofila a, nitratos, nitritos, nitrógeno total, ortofosfatos, oxígeno disuelto, pH, salinidad, sólidos suspendidos totales, temperatura y transparencia*. Con dichas variables se realizó un proceso de selección inicial, en el cual se retiraron del estudio los meses con medidas incompletas al inicio de las series de tiempo o monitoreos muy aislados, esto debido a que meses o periodos con datos faltantes ubicados al inicio de las series no poseen un trasfondo histórico registrado del cual se puedan hacer inferencias o aproximaciones, además estos datos faltantes al inicio incrementarían la incertidumbre al aproximar datos futuros dentro de la serie original.

Las estaciones escogidas al igual que sus periodos de datos disponibles, se encuentran registradas en la Tabla 1.

		Inicio		Fin	
Estación	**Abrev.**	**Año**	**Mes**	**Año**	**Mes**
Centro de la Ciénaga	CEN	1993	2	2008	12
Boca del Caño grande	BCG	1993	2	2008	12
Boca de la Barra	LBA	1993	2	2008	12
Frente al Río Aracatáca	FRA	1995	1	2008	12
Frente al Río Fundación	FRF	1993	2	2008	12
Frente al Río Sevilla	FRS	1993	2	2008	12
Río Aracatáca	RAR	1999	1	2008	12
Río Fundación	RFU	1999	1	2008	12
Río Sevilla	RSE	1999	1	2008	12
Rinconada	RIN	1993	2	2008	12

Tabla 1. Estaciones de muestreo, abreviaturas estandarizadas y periodos a evaluar

Posteriormente se aplican los preprocesamientos a evaluar en el trabajo a través de la construcción de macros en *VBA*[3] y el uso de *Microsoft Office Excel*, tras la aplicación del método de interpolación histórica se encontró que las únicas series que cumplían el requisito del programa de máximo 30 observaciones faltantes eran *pH, salinidad, temperatura y transparencia* correspondientes a las estaciones *RAR, RFU y RSE* (12 series de tiempo). El objetivo de las corridas fue la comparación de los efectos de los métodos de preprocesamiento para encontrar el que menos error generara con los datos originales, por lo anterior se tomó la decisión utilizar solo estas variables y estaciones.

Una vez obtenidas las series preprocesadas, se utilizó el software *TSW* configurado con determinación automática de parámetros para el modelo ARIMA, con el fin de verificar si se cumplen todos los supuestos y analizar los resultados en función de los cuadrados medios del error (*MSE*) entre los residuales del modelo y los parámetros interpolados como atípicos aditivos para el mismo.

5. RESULTADOS Y DISCUSIÓN

De los cuatro métodos planteados solo mostraron resultados el método de *interpolación histórica* y el de *escalamiento*; las observaciones faltantes con respecto al tamaño de la serie causan un efecto adverso en el método de división, dado que al seleccionar la serie tratando de evitar las limitantes del programa se generaban nuevos errores por tamaño insuficiente de las divisiones utilizadas en el programa *TSW*, incluso al utilizar la división con interpolación histórica previa para la generación de mayor cantidad de observaciones, se seguía presentando dicho error.

Se elaboraron modelos ARIMA para cada una de las series de variables en estudio (Tabla 2) utilizando la opción de identificación automática de parámetros en *TSW.* La diferencia entre el parámetro estacional de los modelos (*s* en la Tabla 2) hace infructuosa una comparación entre los mismos, sin embargo, se puede resaltar el hecho de que no existe una homogeneidad aparente entre los comportamientos de las estaciones que se puede explicar por la compleja dinámica de los ecosistemas estuarinos. Por ejemplo se nota un orden de autoregresión entre 3 y 0 en la salinidad dependiendo de la cercanía al mar de la estación evaluada.

Tabla 2. Modelos ARIMA de algunas variables fisicoquímicas del complejo lagunar CGSM, s=12 indica estacionalidad anual y s=4 indica estacionalidad trimestral en el estudio.

Estación	Variable	*ARIMA (p,d,q)(P,D,Q)$_s$*	
		Int. Histórica	**Escalamiento**
RAR	pH	$(0,1,1)(0,1,1)_{12}$	$(0,0,0)(1,0,0)_4$
	Salinidad	$(0,0,0)(0,1,1)_{12}$	$(0,1,0)(0,1,1)_4$
	Temperatura	$(1,0,0)(1,0,0)_{12}$	$(0,0,1)(0,1,0)_4$
	Transparencia	$(0,1,1)(0,1,1)_{12}$	$(0,1,1)(0,1,0)_4$
RFU	pH	$(1,0,0)(0,1,1)_{12}$	$(3,0,1)(0,0,1)_4$
	Salinidad	$(0,1,0)(0,1,1)_{12}$	$(0,1,1)(0,1,1)_4$
	Temperatura	$(0,1,1)(0,1,1)_{12}$	$(0,1,1)(0,1,1)_4$
	Transparencia	$(0,0,1)(1,0,0)_{12}$	$(1,0,0)(0,1,0)_4$
RSE	pH	$(1,0,0)(0,1,1)_{12}$	$(0,0,0)(1,0,0)_4$
	Salinidad	$(3,0,0)(0,1,1)_{12}$	$(0,0,2)(0,1,0)_4$
	Temperatura	$(0,1,1)(0,1,1)_{12}$	$(0,0,0)(0,1,0)_4$
	Transparencia	$(0,1,1)(0,1,1)_{12}$	$(0,0,2)(0,0,0)_4$

Tabla 3. Cuadrados medios del error (MSE) de los residuales correspondientes a los datos interpolados por modelos ARIMA, un valor menor de MSE indica un mejor ajuste de la serie original con la serie interpolada.

Estación	Variable	*MSE*	
		Int. Histórica	**Escalamiento**
RAR	pH	0.967×10^{-2}	0.453×10^{-2}
	Salinidad	19.28	9.117
	Temperatura	0.355×10^{-2}	1.104
	Transparencia	0.105	0.582×10^{-1}
RFU	pH	0.470×10^{-2}	0.162
	Salinidad	9.124	17.818
	Temperatura	0.172×10^{-2}	0.711
	Transparencia	0.470×10^{-1}	0.792×10^{-1}
RSE	pH	0.553×10^{-2}	0.232×10^{-2}
	Salinidad	6.925	8.915
	Temperatura	0.287×10^{-2}	0.871
	Transparencia	0.321	0.619×10^{-1}

3. Visual Basic para Aplicaciones

Se utiliza el *MSE* de los residuales (Tabla 3) como una medida de identificación de la bondad de ajuste entre los valores obtenidos en el modelo ARIMA y los datos de la serie preprocesada e interpolada por *TSW* a través de los diversos métodos ya analizados, de acuerdo a Hurtado y Salcedo (1994) valores pequeños de *MSE* sugieren un buen ajuste de los datos al modelo.

Finalmente se verifica el ajuste de las interpolaciones resultantes a través de métodos gráficos como los observados en la Figura 5, se puede observar en la mayoría de los casos un buen ajuste de los valores interpolados con la estructura general de los datos originales de las series de tiempo.

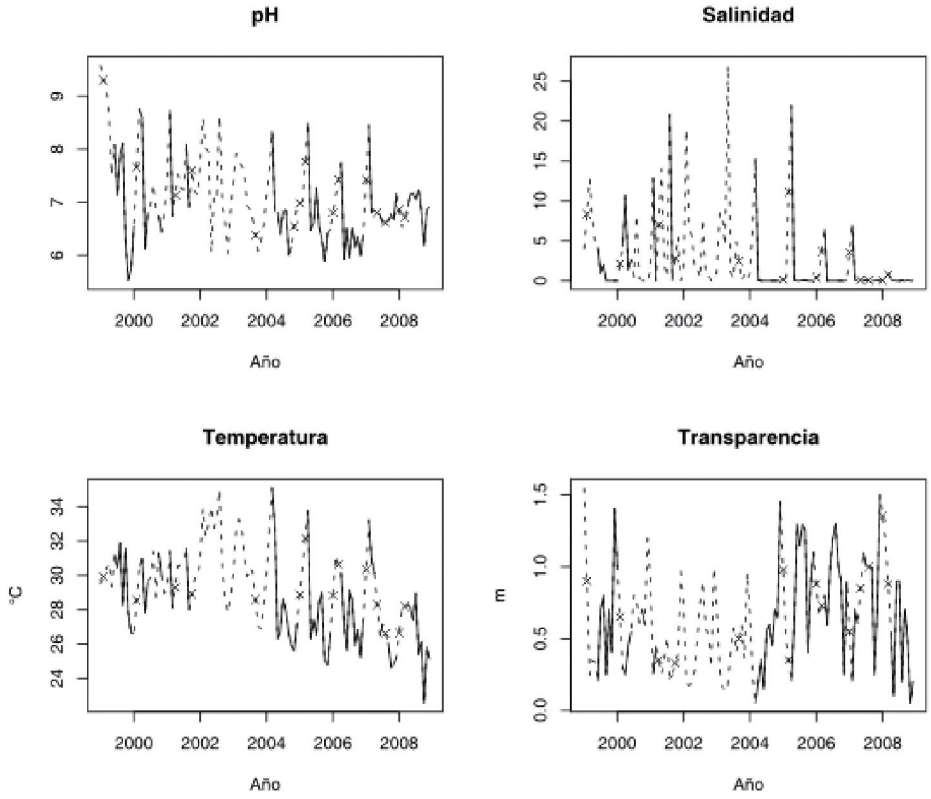

Figura 5. Series de tiempo interpoladas mediante metodología de atípicos aditivos para parámetros fisicoquímicos en el complejo CGSM (e.g. Estación: Río Aracatáca). Las cruces representan valores encontrados por pretratamiento de* interpolación histórica*, las líneas sólidas son los valores originales de la serie y las punteadas los valores interpolados por metodología de atípicos aditivos.

6. CONCLUSIONES

Los métodos de *interpolación histórica* y *escalamiento* mostraron ser eficaces al suplir las carencias de información por datos faltantes en los monitoreos de variables fisicoquímicas del complejo lagunar CGSM. Sin embargo, el hecho de que solo dos de los cuatro métodos evaluados hayan resultado satisfactorios demuestra la dificultad que reviste ajustar series de tiempo en presencia de gran cantidad de observaciones faltantes. Por lo anterior es importante evitar la ausencia o pérdida de información de los ciclos de muestreo en el complejo lagunar CGSM debido a los diversos inconvenientes que se presentan en las salidas de campo o limitaciones presupuestales para los análisis, con el fin de conocer de una manera más verosímil las tendencias y el componente estacional del comportamiento de las variables fisicoquímicas en el sistema.

La *interpolación histórica* fue, según el estudio, el método que mayor ajuste proporciona con la tendencia original de la serie, pues además de mantener la frecuencia mensual propia de la misma, demuestra mayor similitud con los modelos ARIMA identificados para las series de

acuerdo a los *MSE* observados. Hay que resaltar que esta metodología se ve influenciada de forma significativa por el hecho de que las observaciones faltantes se encuentren adecuadamente separadas y distribuidas, los conjuntos continuos de vacíos en la serie (fallas prolongadas en los monitoreos) inutilizan el método. Además, en series con comportamientos muy erráticos el criterio de la media aritmética puede resultar poco adecuado.

El *escalamiento* presentó un ajuste adecuado a los modelos ARIMA identificados y en la mayoría de los casos su diferencia con la *interpolación histórica* no fue muy significativa (Tabla 3). Una ventaja de este método sobre la interpolación histórica radica en que no se ve tan afectado por largos periodos sin toma de datos debido a que estos son reducidos en número durante el escalamiento. Entre las desventajas del método se puede mencionar que la medida de error a la que se expone el usuario es mayor debido al incremento en el uso de promedios, también resaltar que las menores escalas de datos revisten menor resolución en la información para las necesidades del usuario, lo cual puede representar un inconveniente mayor en el caso de requerir una frecuencia específica para un estudio determinado.

Los métodos que se basan en la división de series (*división* y *división con interpolación histórica*) demostraron la menor efectividad durante el estudio debido a sus estrictos requisitos de funcionamiento como son: series muy grandes, vacíos independientes y separados, mayor carga operacional para el usuario y mayor número de posibles combinatorias, lo cual convierte a este método en una opción limitada. Se debe mencionar también que los resultados de utilizar series divididas en la metodología de atípicos aditivos puede llevar a inferencias poco fiables si la serie no es estacionaria siendo que cada subserie puede seguir un modelo distinto y los valores extrapolados diferir mucho de la realidad. La técnica de *división* presenta la ventaja de que es la única técnica entre las evaluadas que no depende de promedios o supuestos de linealidad sino del conjunto original de datos, minimizando de esta manera la propagación de errores.

La comparación de los modelos ARIMA generados por el programa *TSW* para cada preprocesamiento en el presente trabajo es infructuosa, debido a la diferencia en el factor de frecuencia (*s*) entre las dos técnicas.

Se propone la aplicación de optimizaciones en la selección de variables a muestrear dentro del complejo lagunar CGSM, de manera que se recopile la mayor cantidad de información respecto a la variación espacial y temporal de la calidad del agua en esta con un número reducido de variables, lo anterior de acuerdo a investigaciones como las realizadas por Spanos *et al.* (2008) y Simeonov *et al.* (2002) a través del uso de regresiones múltiples y mínimos cuadrados parciales, brindando la oportunidad de estimar emisión desde fuentes de contaminación sin recurrir a medición directa de parámetros con determinación compleja (Simeonov *et al.* 2002).

Se puede concluir que el uso de una determinada metodología de preprocesamiento en una serie destinada a interpolación por métodos aditivos depende exclusivamente de la estructura y organización de los vacíos, en casos de vacíos adecuadamente separados y series cortas es mejor la utilización de una *interpolación histórica*, mientras que en situaciones de gran cantidad de vacíos y periodos prolongados sin monitoreo se debe optar por probar la técnica de *escalamiento*. En el caso de distribución de vacíos de manera constante en la serie (i.e. acceso restringido a una zona de muestreo en una determinada época del año) se prefiere utilizar la *división* o *división con interpolación histórica*. Para las series de variables fisicoquímicas del complejo lagunar CGSM se debe optar por utilizar *interpolación histórica* en las 12 series evaluadas en el presente trabajo y por *escalamiento* las demás series existentes antes de interpolarla por la metodología de atípicos aditivos.

AGRADECIMIENTOS

Los autores agradecen al Instituto de Investigaciones Marinas y Costeras «José Benito Vives de Adréis» y al programa de Calidad Ambiental Marina (CAM) por proporcionar los datos y el apoyo logístico en el desarrollo del presente trabajo. Igualmente, agradecen a Janet Vivas- Aguas y a Juan Carvajalino por las correcciones realizadas en el manuscrito final y a Federico Fernández por el apoyo en los análisis estadísticos.

BIBLIOGRAFÍA

Astel, A., Mazerski, J., Polkowska, Z. and Namiesnik, J., 2004. Application of PCA and time series analysis in studies of precipitation in tricity (Poland). Advances in environmental research. 8. pp. 337- 349.

Divino, J. and McAleer, M., 2009. Modeling sustainable international tourism demand to the Brazilian Amazon. Environmental modeling and software. 24. pp. 1411- 1419.

Enders, C., 2010. Applied missing data analysis. 1ª edición. The Guilford Press. New York. ISBN 978-1-60623-639-0.

Gómez, V. y Maravall, A., 2009. TSW (R.136), [Software de cómputo]. Banco de España.

Gómez, V., Maravall, A. and Peña, D., 1993. Computing missing values in time series. Working paper 93-27, Universidad Carlos III de Madrid, división de economía.

Hurtado, L., 1997. Construcción de modelos matemáticos para interpretar relaciones entre las variables de un sistema lagunar – estuarino – tropical. Reporte técnico. Conciencias – Universidad del Quindío – INVEMAR.

Hurtado, L. y Salcedo, G., 1994. Series temporales con aplicaciones a la epidemiología y a la ecología. 1ª Ed. Universidad del Quindío.

INVEMAR, 2009. Base de datos del proyecto «Monitoreo de las condiciones ambientales y las condiciones estructurales y funcionales de las comunidades vegetales y de los recursos pesqueros durante la rehabilitación de la Ciénaga Grande de Santa Marta». [Santa Marta]. Instituto de Investigaciones Marinas y Costeras «José Benito Vives de Adréis». Url: http://siam.invemar.org.co/siam/redcam. [Consultado en: Julio de 2009].

Maravall, A. and Peña, D., 1992. Missing observations and additive outliers in time series models. Working paper 92-40. Universidad Carlos III de Madrid, División de economía.

Martínez, J., Montealegre, E. y Rangel, E., 1996. Estimación de observaciones faltantes en una serie de tiempo usando modelos ARIMA. En: Memorias IV Congreso Colombiano de Meteorología: La variabilidad y el cambio climático y sus impactos socioeconómicos. pp. 319- 323.

Peña, D. and Maravall, A., 1990. Interpolation, outliers and inverse autocorrelations. Working paper 91-08. Universidad Carlos III de Madrid, División de economía.

Rivera, M. and Caicedo, D., 1998. Information sheet on RAMSAR wetlands. URL: http://ramsar.wetlands.org/Database/Searchforsites/tabid/765/language/en-US/Default.aspx. [Consultado en: Septiembre de 2009].

Simeonov, V., Einax, J., Stranimirova, I. and Kraft, J., 2002. Environmetric modeling and interpretation of river water monitoring data. Analytical and bioanalytical chemistry. 374. pp. 898 - 905.

Simeonov, V., Stratis, J., Samara, C., Zachariadis, G., Voutsa, D., Anthemidis, A., Sofoniou, M. and Kouimtzis, T., 2003. Assessment of the surface water quality in northern Greece. Water research. 37. pp. 4119- 4124.

Spanos, T., Simeonov, V., Simeonova, P., Apostolidou, E. and Stratis, J., 2008. Environmetrics to evaluate marine environment quality. Environmental monitoring and assessment. 143. pp. 215- 225.

Suárez, L. y Torres, F., 1997. Estimación de faltantes en las bases de datos Ciénaga – I y Ciénaga – II. Tesis de Maestría. Universidad del Quindío. En (Hurtado, 1997).

Zhou, F., Guo, H. and Liu, L., 2007a. Quantitative identification and source apportionment of anthropogenic heavy metals in marine sediment of Hong Kong. Environmental geology. 53. pp. 295- 305.

Zhou, F., Huang, G., Guo, H., Zhang, W. and Hao, Z., 2007b. Spatio-temporal patterns and source apportionment of coastal water pollution in eastern Hong Kong. Water research. 41. pp. 3429- 3439.

Volumen 13 - No. 3, Diciembre de 2010, Medellín ISSN 0124.177X.