L. Barba y N. Rodríguez, 2018. Singular Spectrum Analysis and Autoregressive models for Ecuadorian shrimp catch forecasting. *Boletín Geológico y Minero,* 129 (3): 535-548
ISSN: 0366-0176
DOI: 10.21701/bolgeomin.129.3.005

Singular Spectrum Analysis and Autoregressive models for Ecuadorian shrimp catch forecasting

L. Barba[1,2] y N. Rodríguez[1]

(1) Escuela de Ingeniería Informática, Pontificia Universidad Católica de Valparaíso, 2362807 Valparaíso, Chile.
(2) Facultad de Ingeniería. Universidad Nacional de Chimborazo, 060102 Riobamba, Ecuador.
lbarba@unach.edu.ec

ABSTRACT

Shrimp trawl fishing is a relevant activity that contributes significantly to the generation of economic resources for Ecuador through the creation of jobs and the commercialization of the shrimps. Unfortunately the high variability of the marine ecosystem signals are complex making forecasting a difficult task. This paper evaluates the use of singular spectrum analysis (SSA) for improving forecasting by conventional methods. SSA is implemented in four steps, embedding, decomposition, grouping, and diagonal averaging. The effective length of the embedding window is selected by means of the information provided by the differential energy of the eigenvalues. The SSA decomposition is used to extract two types of components, one of low frequency which represents the inter-annual component, and the other of high frequency which represents the annual component. Once the components have been obtained, three models based on SSA are implemented, the first is based on the autoregressive model (SSA-AR), the second is an artificial neural network (ANN) based on the Levenberg-Marquardt algorithm (SSA-ANN-LM), and the third is another ANN based on the Cuckoo Search algorithm (SSA-ANN-CS). The historical data that was used to evaluate the models are the shrimp monthly catches in the Gulf of Guayaquil from 2000 to 2012. The empirical results show the superiority of the proposed models with respect to the conventional models. The best approximation was achieved with the enhanced model SSA-AR, with a MAPE of 1.0%, a R^2 of 99% and a RMSE of 1.5%.

Key words: fishery forecasting, singular spectrum analysis, Levenberg-Marquardt, Cuckoo Search.

Análisis de Espectro Singular y modelos autorregresivos para el pronóstico de stock de camarón ecuatoriano

RESUMEN

La pesca por arrastre de camarón es una actividad relevante que contribuye de manera significativa con la generación de recursos económicos para Ecuador a través de la empleabilidad y la comercialización. Por desgracia, la alta variabilidad de las señales de los ecosistemas marinos presenta complejidad, tornando el pronóstico de stock en una tarea difícil. En este trabajo se evalúa Análisis de Espectro Singular (SSA) para mejorar la predicción de los métodos convencionales. SSA se implementa en cuatro etapas, embebido, descomposición, agrupación y promedio de las diagonales. La longitud efectiva de la ventana de embebido se selecciona por medio de la información proporcionada por la energía diferencial de los valores propios. La descomposición SSA se utiliza para extraer dos tipos de componentes, una de baja frecuencia que representa la componente interanual, y otra de alta frecuencia que representa la componente anual. Una vez obtenidas las componentes se implementan tres modelos de pronóstico. El primer modelo es lineal autorregresivo (SSA-AR), el segundo es una Red Neuronal Artificial (ANN) basada en Levenberg-Marquardt (SSA-ANN-LM) y el tercero es otra ANN basada en Cuckoo Search (SSA-ANN-CS). Los datos históricos utilizados para evaluar los modelos son las capturas mensuales de camarón en el Golfo de Guayaquil desde el año 2000 hasta el 2012. Los resultados empíricos muestran la superioridad de los modelos propuestos con respecto a los modelos convencionales. La mejor aproximación se logra con el modelo SSA-AR, con un MAPE de 1.0%, un R^2 de 99.9% y un RMSE de 1.5%.

Palabras clave: pronóstico de pesca, análisis de espectro singular, Levenberg-Marquardt, Cuckoo Search.

VERSIÓN ABREVIADA EN CASTELLANO

Introducción

La plataforma continental ecuatoriana es rica en recursos pesqueros, su adecuada explotación maximiza la economía de las ciudades costeras. La flota pesquera de arrastre de camarón en el Ecuador cuenta con 57 años de funcionamiento. Durante los últimos años el Instituto Nacional de Pesca (INP) promueve la pesquería racional con el fin de evaluar su potencial, diversificar la producción, promover el desarrollo y lograr la utilización óptima y eficiente. El golfo de Guayaquil es el punto central de esta actividad, por tanto las capturas mensuales de camarón durante los años 2000 al 2013 serán modeladas en este trabajo.

Los métodos convencionales difícilmente presentan resultados precisos cuando se está tratando con sistemas dinámicos como es el caso del ecosistema marino. A continuación se describen brevemente aquellos métodos que han logrado modelar señales con alta variabilidad y que serán parte de los modelos a proponer.

Herramientas de inteligencia computacional han logrado buena aproximación. Una red neuronal artificial con aprendizaje de Levenberg-Marquardt supera a los métodos convencionales en una amplia variedad de problemas (Kermani et al., 2005; Yetilmezsoy y Demirel, 2008; Mukherjee y Routroy, 2012).

Los métodos basados en metaheurísticas también han demostrado óptimos resultados en pronóstico. Cuckoo Search (CS) es una metaheurística relativamente nueva que ha sido aplicada en diversos ámbitos, tales como medio ambiente (Jiang et al., 2014) y electricidad (Wang et al., 2014). CS es un algoritmo de optimización libre del cálculo del gradiente, esta característica ha influenciado para su crecimiento en popularidad (Walton et al., 2013).

Pronóstico basado en la extracción de componentes es una estrategia eficiente en el dominio de pesca (Rodríguez et al., 2014). Análisis de Espectro Singular es una técnica potente para el análisis de series de tiempo, la cual está principalmente enfocada en la extracción de componentes intrínsecas. La aplicación de SSA frecuentemente se asocia con las publicaciones de Broomhead D. y King G.P. (1986), una ventaja principal de SSA es que no requiere de conocimiento a priori sobre el número de períodos relevantes y su duración. Tendencia, estacionalidad y ruido son comúnmente extraídos por medio de SSA (Xiao et al., 2014; Marques et al., 2006; Hassani et al., 2015; Abdollahzade et al., 2015; Telesca et al., 2013; Chen et al., 2013; Viljoen y Nel, 2010).

En este trabajo son evaluados tres modelos de pronóstico, dos basados en inteligencia computacional y uno basado en el modelo lineal autorregresivo. Los tres modelos de pronóstico son mejorados por medio de la etapa de preprocesamiento basada en SSA, la cual es usada para extraer las componentes anual e interanual de la serie observada. El artículo está estructurado de la siguiente manera. La sección 2 describe el preprocesamiento por medio de SSA. La sección 3 presenta la predicción basada en las Redes Neuronales Artificiales y el modelo Autorregresivo. La sección 4 describe las métricas de rendimiento. En la sección 5 se presentan los Resultados y Discusión. Finalmente las conclusiones se muestran en la Sección 6.

Metodología

Preprocesamiento de los datos por medio de SSA

La técnica Análisis de Espectro Singular (SSA) es implementada en la etapa de preprocesamiento de los datos para extraer las componentes de la serie de tiempo; SSA se describe en cuatro pasos (Golyandina and Stepanov, 2005), embebido, descomposición de valores singulares (SVD), agrupación y promedio de las diagonales.

El primer paso es implementado para mapear la serie de tiempo de longitud N en una matriz de L filas y K=N-L+1 columnas. Los L vectores desfasados son las filas de la matriz de trayectoria H (matriz de Hankel),

$$H = \begin{pmatrix} x_1 & x_2 & \cdots & x_K \\ x_2 & x_3 & \cdots & x_{K+1} \\ \vdots & \vdots & \vdots & \vdots \\ x_L & x_{L+1} & \vdots & x_N \end{pmatrix}$$

El segundo paso consiste en la descomposición de H. La SVD de H tiene la forma,

$$H = \sum_{i=1}^{L} \sqrt{\lambda_i U_i V_i^T}$$

La terminología estándar identifica a $\sqrt{\lambda_i}$ como valores singulares de H, mientras que U y V son los vecto-

res singulares izquierdo y derecho de H respectivamente. Además el conjunto $\sqrt{\lambda_i}U_iV_i$ es llamado iésimo eigentriple de H.

Las matrices elementales pueden ser representadas por $H_i=\sqrt{\lambda_i}U_iV_i^T$.

Las matrices elementales son agrupadas para obtener las componentes

$$H_1 = A_1$$
$$H_2 = \sum_{i=2}^{L} A_i$$

A partir de las matrices H_1 y H_2 por medio del paso promedio de las diagonales serán obtenidas dos componentes aditivas, interanual y anual respectivamente, ambas de longitud N.

La predicción de captura de camarón $\hat{x}$, es obtenida a partir de la adición de las componentes estimadas interanual $\hat{x}_{ia}$ y anual $\hat{x}_a$, como se indica a continuación, $\hat{x}(n+1)=\hat{x}_{ia}(n+1) + \hat{x}_a(n+1)$, donde n representa el instante de tiempo. La predicción de las componentes se realiza por medio de los tres modelos propuestos: Autorregresivo Lineal, una Red Neuronal basada en Levenberg-Marquard y una Red Neuronal basada en Cuckoo Search.

SSA combinado con el modelo Autorregresivo

La predicción de cada componente es implementada a partir de las siguientes ecuaciones,

$$\hat{x}_a(n+1) = \sum_{i=0}^{m-1}\alpha_i x_{ia}(n-1) + \sum_{i=0}^{m-1}\beta_i x_a(n-1)$$

$$\hat{x}_a(n+1) = \beta_i x_{ia}(n-1)$$

donde m es el tamaño de la ventana, α_i y β_i son los coeficientes de x_{ia} y x_a respectivamente. Los coeficientes son encontrados por medio del método de mínimos cuadrados lineales (LSM),

$$x_{ia}=\alpha\ Z_{ia}$$
$$x_a=\beta\ Z_a$$

donde Z_{ia} es la matriz de regresores interanuales y Z_a es la matriz de regresores anuales (e interanuales); los coeficientes se obtienen a partir del cálculo de la matriz pseudoinversa

$$\alpha=Z_{ia}^{\dagger}x_{ia}$$
$$\beta=Z_a^{\dagger}x_a$$

SSA combinado con Redes Neuronales

La predicción no lineal está basada en redes neuronales artificiales, cuya salida es

$$\hat{x}(n+1) = \sum_{j=1}^{Q} b_j h_j$$

$$h_j = \sum_{i=1}^{m} w_{ji} Z_i$$

donde $\hat{x}$ es el valor estimado, n es el instante de tiempo, Q es el número de nodos ocultos, b_j y w_{ji} son los pesos lineales y no lineales de las conexiones de la red. A la salida de los nodos ocultos se aplica la función de activación sigmoidal.

$$f(x) = \frac{1}{1+e^{-x}}$$

El aprendizaje de la red es evaluado por medio de dos algoritmos, uno basado en el descenso del gradiente (Levenberg-Marquardt) y otro basado en una metaheurística (Cuckoo Search).

Métricas de Rendimiento

La exactitud del pronóstico es evaluada por medio de las métricas: Mean Absolute Percentage Error (MAPE),

Coeficiente de Determinación (R^2), Root Mean Squared Error (RMSE) y Error Relativo (RE). El número óptimo de regresores es encontrado por medio de la métrica Generalized Cross Validation (GCV).

Resultados y discusión

La serie de tiempo contiene muestras mensuales de la pesca de arrastre de camarón desde el año 2000 al 2012, en el golfo de Guayaquil (Ecuador), las mayores capturas se concentran en este golfo con respecto al resto del país. La señal muestra alta variabilidad, periodos de abundancia así como de escases debido a los periodos de veda.

Las componentes de alta y baja frecuencia, anual e interanual son extraídas por medio de SSA. El preprocesamiento de los datos por medio de SSA ha sido implementado a partir de una longitud inicial de ventana L=N/2, donde N es el tamaño de la muestra. La energía de los valores propios obtenidos en el segundo paso de SSA se representa de forma gráfica para encontrar el tamaño de ventana efectiva (Barba L. et al. 2014), en este caso L=6. El embebido vuelve a realizarse con este valor efectivo, la componente extraída muestra las oscilaciones de baja frecuencia; mientras que la componente x_a muestra el comportamiento periódico de alta frecuencia, como se muestra en la Figura 2a.

Para la predicción, la muestra se dividió en dos grupos, entrenamiento (75%) y validación (25%). El orden del modelo Autorregresivo es calibrado mediante training (Figura 2b), los resultados de las métricas de rendimiento se muestran en la Tabla 1.

Las Tablas 2, 3 y 4, presentan la evaluación de los tres modelos mejorados por medio de SSA, SSA-AR, SSA-ANN-LM, y SSA-ANN-CS. El mejor modelo es SSA-AR en base a las métricas calculadas, MAPE de 1.0%, R^2 de 99.9% y RMSE de 0.48%. El pronóstico de la componente interanual presenta valores eficientes por medio de los tres modelos; mientras que el pronóstico de la componente anual presenta menor exactitud por medio de los métodos basados en la red neuronal. El pronóstico convencional lineal y no lineal obtiene resultados deficientes, así como el pronóstico basado en SSA con un tamaño de ventana convencional.

El rendimiento de los modelos para el pronóstico one-step ahead, es evaluado a través del test de correlación de Pitman (Tabla 4). El test confirma la superioridad de SSA-AR con respecto a los demás.

Conclusiones

En este trabajo se presentaron tres modelos autorregresivos mejorados por medio de Análisis de Espectro Singular para el pronóstico one-step ahead de stock de camarón en el Golfo de Guayaquil.

Se evaluaron tres modelos SSA-AR, SSA-ANN-LM y SSA-ANN-CS, todos basados en las componentes interanual y anual que fueron extraídas de la serie de tiempo por medio de SSA.

Los tres modelos permitieron mejorar los modelos convencionales. La mayor ganancia fue obtenida por medio del modelo SSA-AR con un MAPE de 1.0%, un RMSE de 0.48%, y R^2 de 99.9%.

Debido a los resultados prometedores, este modelo será evaluado con otras series temporales relacionadas con pesca y otras áreas de conocimiento.

Introduction

The Ecuadorian continental shelf is rich in fishery resources and their proper exploitation maximizes the economy of coastal towns. The shrimp trawl fishing fleet in Ecuador has been in operation for 57 years; during recent years the National Fishery Institute (Instituto Nacional de Pesca - INP, 2015) has been promoting rational fishing in order to assess its potential, to diversify production and to promote its development and efficient use. The data published by INP shows that the highest shrimp catches are concentrated in the Gulf of Guayaquil, the monthly catches in the gulf from 2000 to 2012 will be used in this study to forecast the stock.

Conventional methods present inaccurate results when dealing with dynamical systems. Tools of computational intelligence reach good approximations, for instance an artificial neural network with Levenberg-Marquardt learning outperforms the conventional Newton method and gradient based methods for a wide variety of problems (Kermani *et al.*, 2005; Yetilmezsoy and Demirel, 2008; Mukherjee and Routroy, 2012).

Metaheuristic-based methods have also shown optimal results in forecasting. Cuckoo Search (CS) is relatively a new metaheuristic that has been applied in diverse fields such as the environment (Jiang *et al.,* 2014) and electricity (Wang *et al.,* 2014). CS is a gradient free method, this feature has influenced its growing popularity (Walton *et al.,* 2013).

Forecasting based on components is an efficient strategy in the fisheries domain (Rodriguez *et al.,* 2014). Singular spectrum analysis (SSA) is a potent technique for time series analysis, which is mainly focused on the intrinsic component extraction. SSA application is associated with the publications of Broomhead and King (1986), one main advantage of SSA is that it does not require a priori knowledge about the number of periodicities and their duration. Trend, seasonality, and noise have been commonly extracted with SSA (Xiao *et al.,* 2014; Marques *et al.,* 2006; Hassani *et al.*, 2015; Abdollahzade et al., 2015; Telesca *et al.,* 2013; Chen *et al.*, 2013; Viljoen and Nel, 2010).

Three forecasting models are evaluated in this study, two based on computational intelligence and one based on a conventional auto-regressive linear model. The three forecasting models are enhanced with a preprocessing stage based on SSA to extract the annual and inter-annual components. The paper is structured as follows. Section 2 describes the preprocessing with SSA. Section 3 presents the prediction with the autoregressive neural networks and the autoregressive model. Section 4 describes the performance metrics. Section 5 presents the Results and Discussions. Finally the conclusions are shown in Section 6.

Methods

Data preprocessing based on SSA

The SSA is described in four steps: embedding, decomposition, grouping and diagonal averaging (Golyandina and Stepanov, 2005).

The embedding step maps the time series x of length N, to a sequence of multi-dimensional lagged vectors. The window length L is an integer with values $1<L<N$, the embedding creates $K=N-L+1$ lagged vectors. The L-lagged vectors are the rows of the trajectory matrix (Hankel matrix),

$$H = \begin{pmatrix} x_1 & x_2 & \cdots & x_K \\ x_2 & x_3 & \cdots & x_{K+1} \\ \vdots & \vdots & \vdots & \vdots \\ x_L & x_{L+1} & \vdots & x_N \end{pmatrix}$$

The elements $H_{ij}=x_{i+j-1}$, and the anti-diagonals consist of equal elements.

The second step in SSA is the singular value decomposition of the trajectory matrix, it has the form

$$H = \sum_{i=1}^{L} \sqrt{\lambda_i U_i V_i^T}$$

where $\lambda\iota$ is the i-th eigenvalue of $S=H\ H^T$. The standard SVD terminology calls $\sqrt{\ddot{e}_i}$ singular values of H, whereas U and V are left and right singular vectors of H respectively. Besides, the collection $\sqrt{\lambda_i}U_iV_i$ is called the ith eigentriple of H.

Elementary matrices can be represented with $H_i=\sqrt{\lambda_i}U_iV_i^T$.

From the elementary matrices, the grouping step is executed to create partitions of the set of indices $(1, \ldots, d)$ into disjoint subsets $(I_1, \ldots, I_j)$.

The diagonal averaging is the four and last step of SSA. This process transform the matrices $H_{I1}, \ldots, H_{Ij}$, into new j time series named components, all of length N (as the observed time series). The diagonal averaging process is illustrated as follows,

$$c_i = \begin{cases} \dfrac{1}{k-1}\displaystyle\sum_{l=1}^{k} H(l,k-l) & 2 \le k \le L \\ \dfrac{1}{L}\displaystyle\sum_{l=1}^{L} H(l,k-l) & L < k \le K+1 \\ \dfrac{1}{K+L-k+1}\displaystyle\sum_{l=k-K}^{L} H(l,k-l) & K+2 \le k \le K+L \end{cases}$$

where c_i is the i-th element of the extracted component.

Prediction based on components

The prediction of the shrimp trawl fishing $\hat{x}$ is obtained from the components extracted in the preprocessing stage based on SSA. The inter-annual component x_{ia} and the annual component x_a are used in the prediction equation

$$\hat{x}\,(n+1)= \hat{x}_{ia}(n+1) + \hat{x}_a(n+1)$$

where n represents the time instant.

SSA combined with the Autoregressive model

The equations used to compute $\hat{x}_a$ and $\hat{x}_{ia}$ are,

$$\hat{x}_a(n+1) = \sum_{i=0}^{m-1} \alpha_i x_{ia}(n-i) + \sum_{i=0}^{m-1} \beta_i x_a(n-i)$$

$$\hat{x}_{ia}(n+1) = \beta_i x_{ia}(n-i)$$

where m is the time window size (number of lagged values), α_i and β_i are the ith coefficients of x_{ia} and x_a respectively. The annual component x_a has influence of the inter-annual component, therefore x_{ia} is used as external input.

The coefficient estimation is based on the linear least square method (LSM), by using the linear relationship expressed in the equations

$$x_{ia}=\alpha \;\; Z_{ia}$$
$$x_{a}=\beta \;\; Z_{a}$$

where Z_{ia} is the matrix of inter-annual regressors with N rows and m columns, α is a vector of m rows and one column, Z_a is the matrix of annual and inter-annual regressors with N rows and 2m columns, and β is a vector of 2m rows and one column. The coefficients are computed by using the Moore-Penrose pseudoinverse $Z_{ia}^{\dagger}$ and $Z_a^{\dagger}$ as follows

$$\alpha=Z_{ia}{}^{\dagger}x_{ia}$$
$$\beta=Z_{a}{}^{\dagger}x_{a}$$

SSA combined with neural networks

The auto-regressive neural networks implemented have a common structure of three layers (Freeman and Skapura, 1991), the inputs are the lagged terms, which are contained in the matrix of regressors Z (Z_{ia} or Z_a). The sigmoid transfer function is applied at the hidden layer, and the predicted value is obtained at the output layer. The ANN output is:

$$\hat{x}(n+1)=\sum_{j=1}^{Q} b_j h_j$$

$$h_j=\sum_{i=1}^{m} w_{ji} Z_i$$

where $\hat{x}$ is the estimated value, n is the time instant, Q is the number of hidden nodes, b_j and w_{ji} are the linear and nonlinear weights of the ANN connections respectively; the sigmoid transfer function is applied at the hidden level with

$$f(x)=\frac{1}{1+e^{-x}}$$

The ANN is denoted with ANN(m,Q,1), with inputs, Q hidden nodes, and 1 output. The ANN weights b and w are updated with the application of the learning algorithm.

Levenberg-Marquardt algorithm

Levenberg-Marquardt is an optimization algorithm of high application due to the accuracy. The scalar u is a parameter used in LM to determine the behavior. If u increases its value, the algorithm works as the steepest descent algorithm with low learning rate; whereas

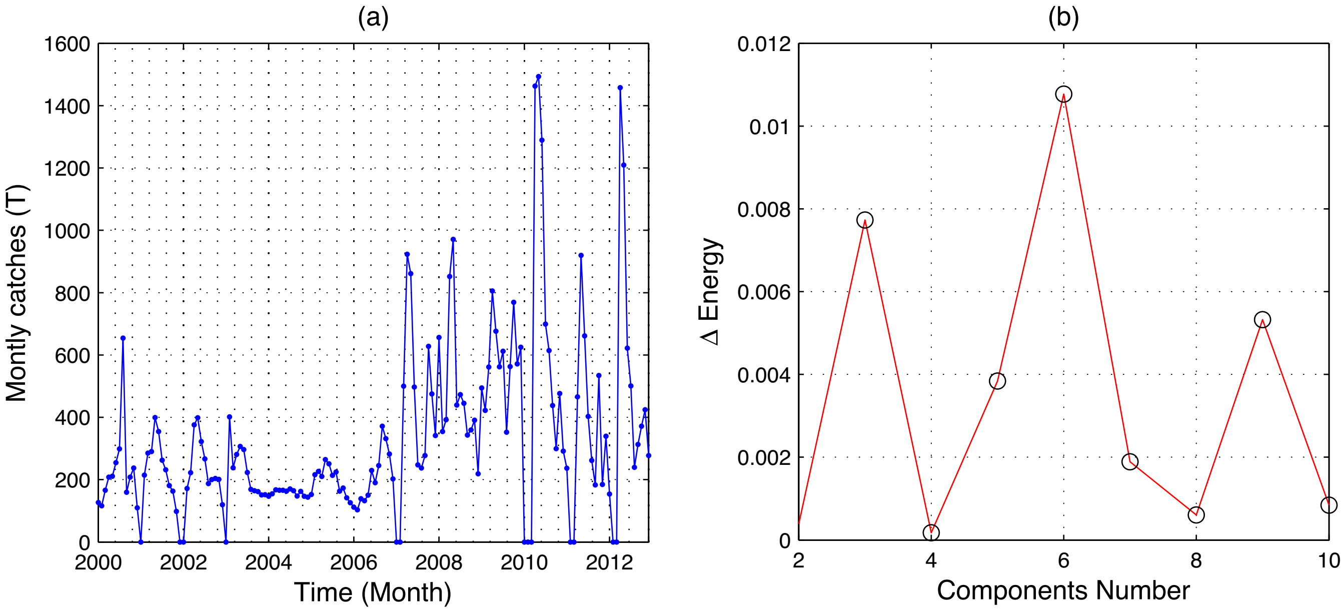

Figure 1. (a) Shrimp catches (b) Differential energy of eigenvalues.
Figura 1. *(a) Captura de camarón (b) Energía diferencial de los valores propios.*

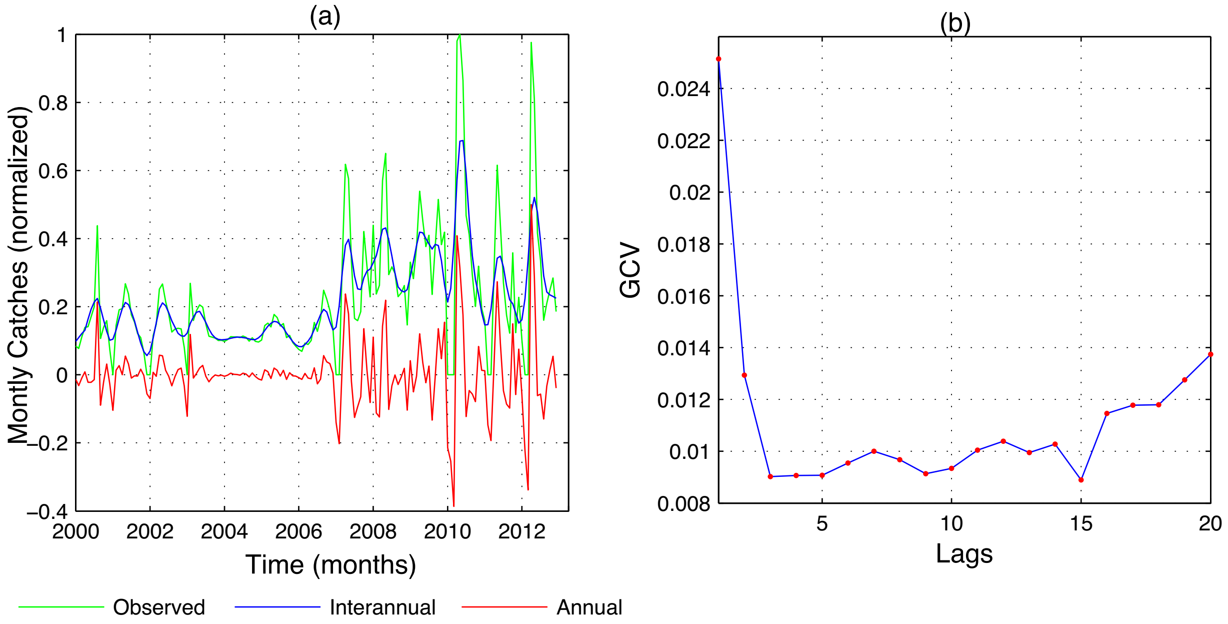

Figure 2. (a) Normalized shrimp catches, interannual and annual components (b) Lagged values calibration.
Figura 2. *(a) Capturas de camarón normalizadas, componentes interanual y anual (b) Calibración de la ventana de tiempo.*

if *u* decreases the value until zero, the algorithm works as the Gauss-Newton method (Hagan *et al.*, 2002).

The weights updating is computed with,

$$\omega_{n+1}=\omega_n+\Delta(\omega_n)$$
$$\Delta(\omega_n)= -[J^T(\omega_n)\ J\ (\omega)+u_n\ I]^{-1}J^T(\omega_n)e(\omega_n)$$

where ω_n is the weight vector composed by w_{ji}, ..., b_j, Δ(ù$_n$) is the weight increment, *j* is the Jacobian matrix, *I* is the identity matrix, *T* means transposed, and e is the error vector.

The general fitness function used in *LM* is mean squared error (MSE), the elements of the Jacobian matrix corresponds to the partial derivative of the fitness function regarding each weight, as follows

$$J(\omega_n)=\frac{\partial e(\omega_{j,i})}{\partial \omega_{j,i}}$$

where the order of the matrix *j* is *Nxk*, *N* is the sample size, and *k* the number of weights (of all the ANN connections).

Cuckoo Search algorithm

Cuckoo Search was developed by Yang and Deb (2009). CS is a metaheuristic algorithm, inspired on the brood parasitism of some cuckoo species. Additionally CS can be considered as an extension of Lévy-flights algorithm; CS is based on these three rules (Yang, 2010):

- Each cuckoo lays one egg at a time and places it in a random nest.
- Best nests (if high quality) remain for generations.
- The number of nests available is fixed, and the discovery rate of an egg deposited by a bird host is $\rho_a \in 0,1$. In this case the bird host chooses one of two options: gets rid of the nest egg or builds a new nest (new random solution).

The generation of new solutions, in this case the ANN weights set is

$$\omega(n+1)= \omega(n)+\alpha \oplus L\acute{e}vy(\lambda), \qquad \alpha>0$$

where ω is the weights vector, α is the step size, which should be related to the scale of the problem, $\oplus$ means entrywise multiplication, and $L\acute{e}vy(\lambda)$ flight provides a random walk, whose steps correspond to a $L\acute{e}vy$ distribution with infinite mean and variance; the random walk process obeys a power-law-step-length distribution with a heavy tail (Yang and Deb, 2009).

Performance Metrics

The prediction accuracy was evaluated with residual metrics. The mean absolute percentage error (MAPE), coefficient of determination (R^2), root mean squared

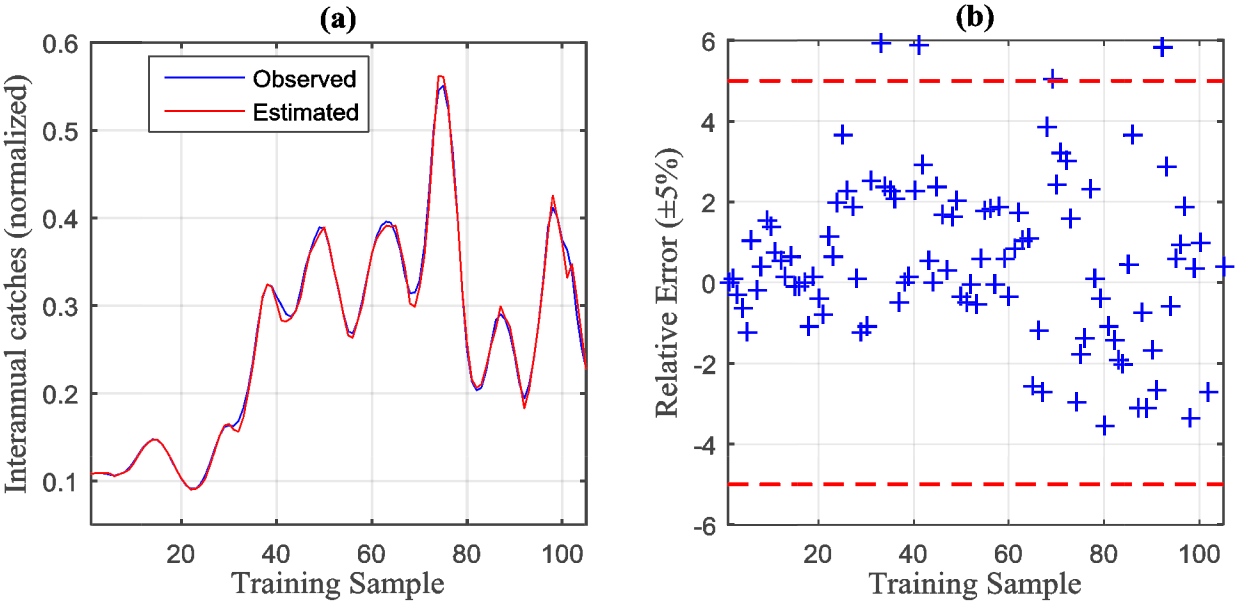

Figure 3. Model Order Calibration for SSA-AR through the training sample (a) Interannual Component Observed versus Predicted (b) Relative Error.
Figura 3. *Calibración del modelo SSA-AR por medio de la muestra de training (a) Componente Interanual Observada versus Pronosticada (b) Error Relativo.*

error (RMSE), and relative error (RE). The number of optimal lagged values was found with the generalized cross validation (GCV) metric.

$$MAPE = \frac{1}{N_V}\left|\sum_{i=1}^{N_V}\frac{(x_i - \hat{x}_i)}{x_i}\right|$$

$$R^2 = \left[1 - \frac{\sigma^2(x-\hat{x})}{\sigma^2(x)}\right]$$

$$RMSE = \sqrt{\sum_{i=1}^{N_V}(x_i - \hat{x}_i)^2}$$

$$RE = \frac{(x_i - \hat{x}_i)}{x_i}$$

$$GCV = \frac{RMSE}{\left(1 - \frac{P}{N_v}\right)^2}$$

where N_v is the validation sample size, x is the observed value, $\hat{x}$ is the forecasted value, σ^2 is the variance, and P is the number of model parameters.

Results and discussion

The results of the implementation of the forecasting model during the stages of preprocessing and prediction are presented in this section.

	MAPE	R^2	RMSE	RE
Models	(%)	(%)	(%)	(%)
SSA-AR ()	1.7	99.6	0.7	92.4 (±5%)

Table 1. Results of the Interannual Component Prediction through the training sample for calibration purpose.
Tabla 1. *Resultados del Pronóstico de la Componente Interanual a través de la muestra de training con propósito de calibración.*

The time series contains 156 monthly samples of the shrimp trawl fishing from 2000 to 2012 in the Gulf of Guayaquil (Ecuador), the highest shrimp catches are concentrated in this gulf with respect to the rest of the country. Figure 1a shows the monthly shrimp catch in metric tons, a complex nonlinear behaviour can be observed in the figure, with some null values due to the shrimp ban. The observed signal is decomposed into two components, annual and inter-annual. The components are predicted via linear and non-linear models, the sample is split into two groups, training and validation. The training sample represents 75% of the data, and consequently the validation sample represents 25%.

Preprocessing with singular spectrum analysis

The time series was preprocessed with SSA. The ini-

	MAPE	R^2	RMSE	RE
Models	(%)	(%)	(%)	(%)
SSA-AR ()	2.3	99.1	0.99	91.4 (±7%)
SSA-ANN-LM	3.1	98.4	1.3	94.3 (±7%)
SSA-ANN-CS	3.7	98.1	1.5	88.6 (±7%)

Table 2. Results of the Interannual Component Prediction through the validation sample.
Tabla 2. *Resultados del Pronóstico de la Componente Interanual a través de la muestra de validación.*

	MAPE	R^2	RMSE	RE
Models	(%)	(%)	(%)	(%)
SSA-AR ()	8.6	99.8	0.84	74.3 (±15%)
SSA-ANN-LM	19.5	99.5	1.5	68.6 (±15%)
SSA-ANN-CS	88.1	83.6	9.6	17.1 (±15%)

Table 3. Results of the Annual Component Prediction through the validation sample.
Tabla 3. *Resultados del Pronóstico de la Componente Anual a través de la muestra de validación.*

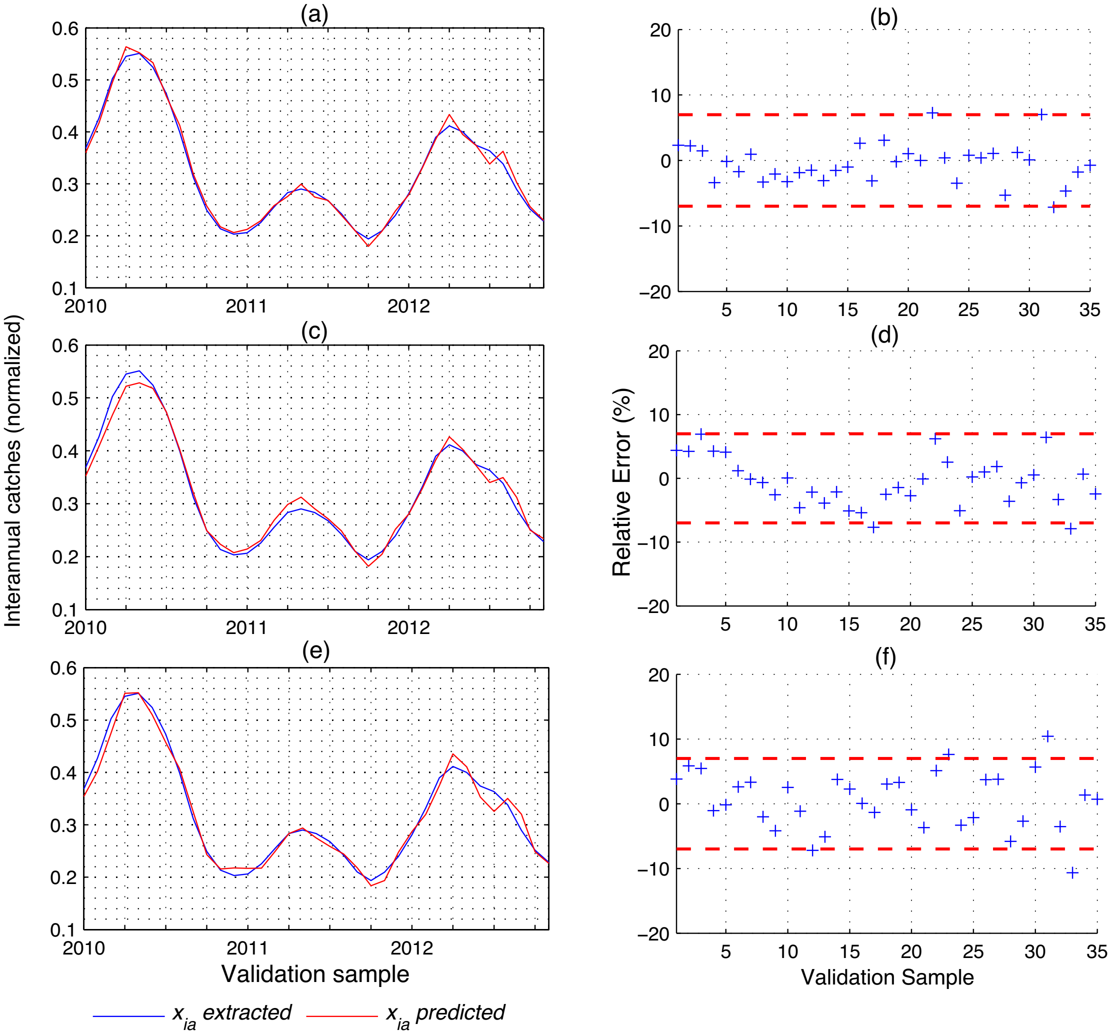

Figure 4. Interannual Component Prediction through the validation sample (a) SSA-AR (b) Relative Error SSA-AR (c) SSA-ANN-LM (d) Relative Error SSA-ANN-LM (e) SSA-ANN-CS (f) Relative Error SSA-ANN-CS.
Figura 4. *Predicción de la Componente Interanual a través de la muestra de validación (a) SSA-AR (b) Error Relativo SSA-AR (c) SSA-ANN-LM (d) Error Relativo SSA-ANN-LM (e) SSA-ANN-CS (f) Error Relativo SSA-ANN-CS.*

tial window length used was computed with the general window length value *L=N/2*; where *N* is the sample size (N=156). The differential energy of the eigenvalues obtained in the second step of SSA was plotted to find the highest relative energy concentration (Barba L. *et al.*, 2014); this is shown in Figure 1b. From

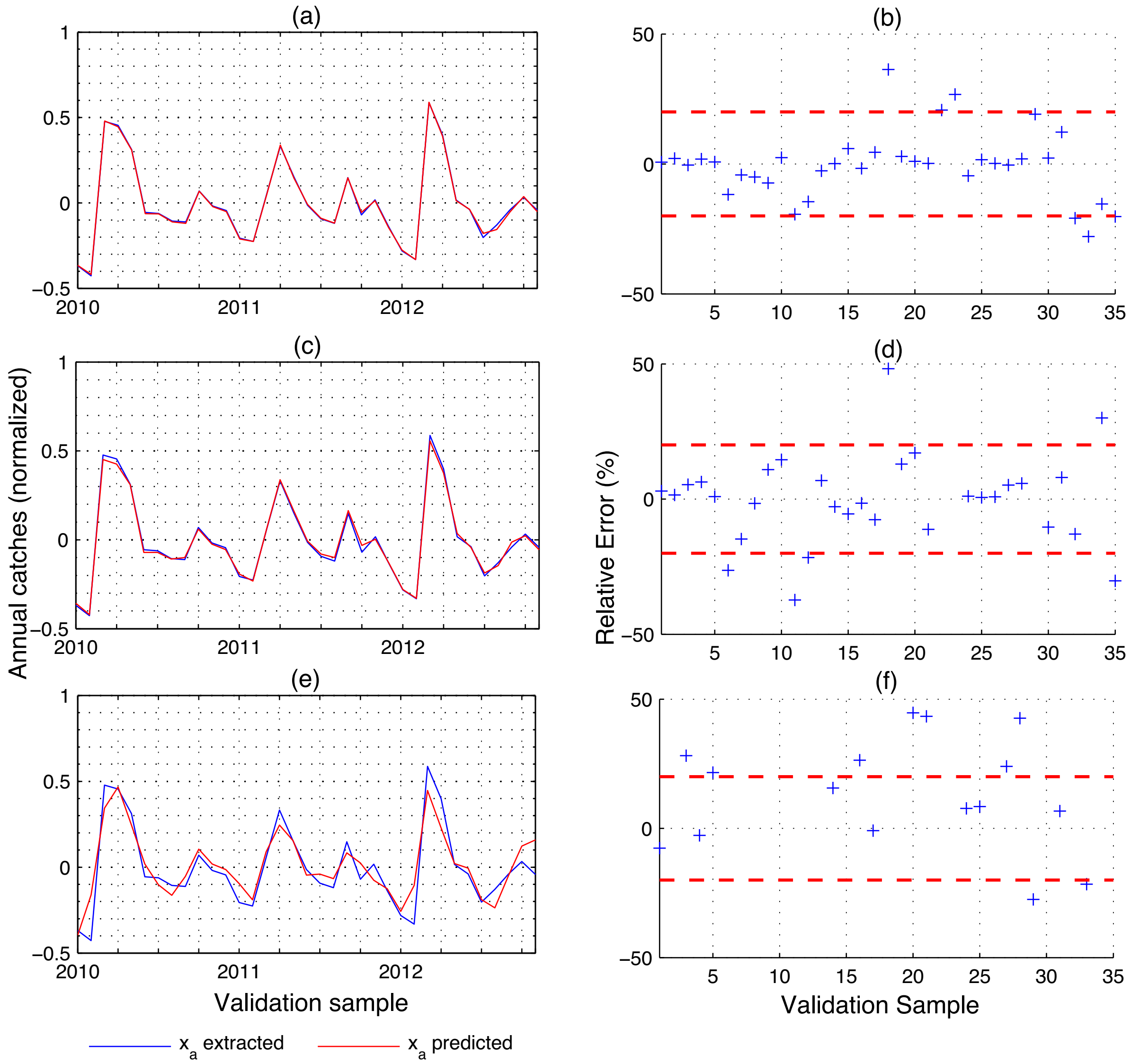

Figure 5. Annual Component Prediction through the validation sample (a) SSA-AR (b) Relative Error SSA-AR (c) SSA-ANN-LM (d) Relative Error SSA-ANN-LM (e) SSA-ANN-CS (f) Relative Error SSA-ANN-CS.

Figura 5. *Predicción de la Componente Anual a través de la muestra de validación (a) SSA-AR (b) Error Relativo SSA-AR (c) SSA-ANN-LM (d) Error Relativo SSA-ANN-LM (e) SSA-ANN-CS (f) Error Relativo SSA-ANN-CS.*

	MAPE	R^2	RMSE	RE
Models	(%)	(%)	(%)	(%)
Conventional AR	36.7	69.2	12	25.7 (±15%)
Conventional ANN	29.4	79.1	9.9	34.4 (±15%)
SSA-AR ()	1.0	99.9	0.48	94.3 (±3 %)
SSA-ANN-LM	5.3	99.4	2.3	91.4 (±15%)
SSA-ANN-CS	19.2	89.3	9.4	62.8 (±15%)

Table 4. Results of the Shrimp Catches Prediction in Guayaquil Gulf.

Tabla 4. *Resultados del Pronóstico de Pesca de Camarones en el Golfo de Guayaquil.*

Figure 1b, the highest peak *L=6* was used as the effective window length. The preprocessing restarts with this effective value.

The components extracted with SSA are shown in Figure 2a. The component $\hat{x}_{ia}$ shows the fluctuations of low frequency, whereas the $\hat{x}_{a}$ component shows the fluctuations of high frequency.

Prediction via SSA-AR model

The prediction with the autoregressive model is implemented with the components extracted through

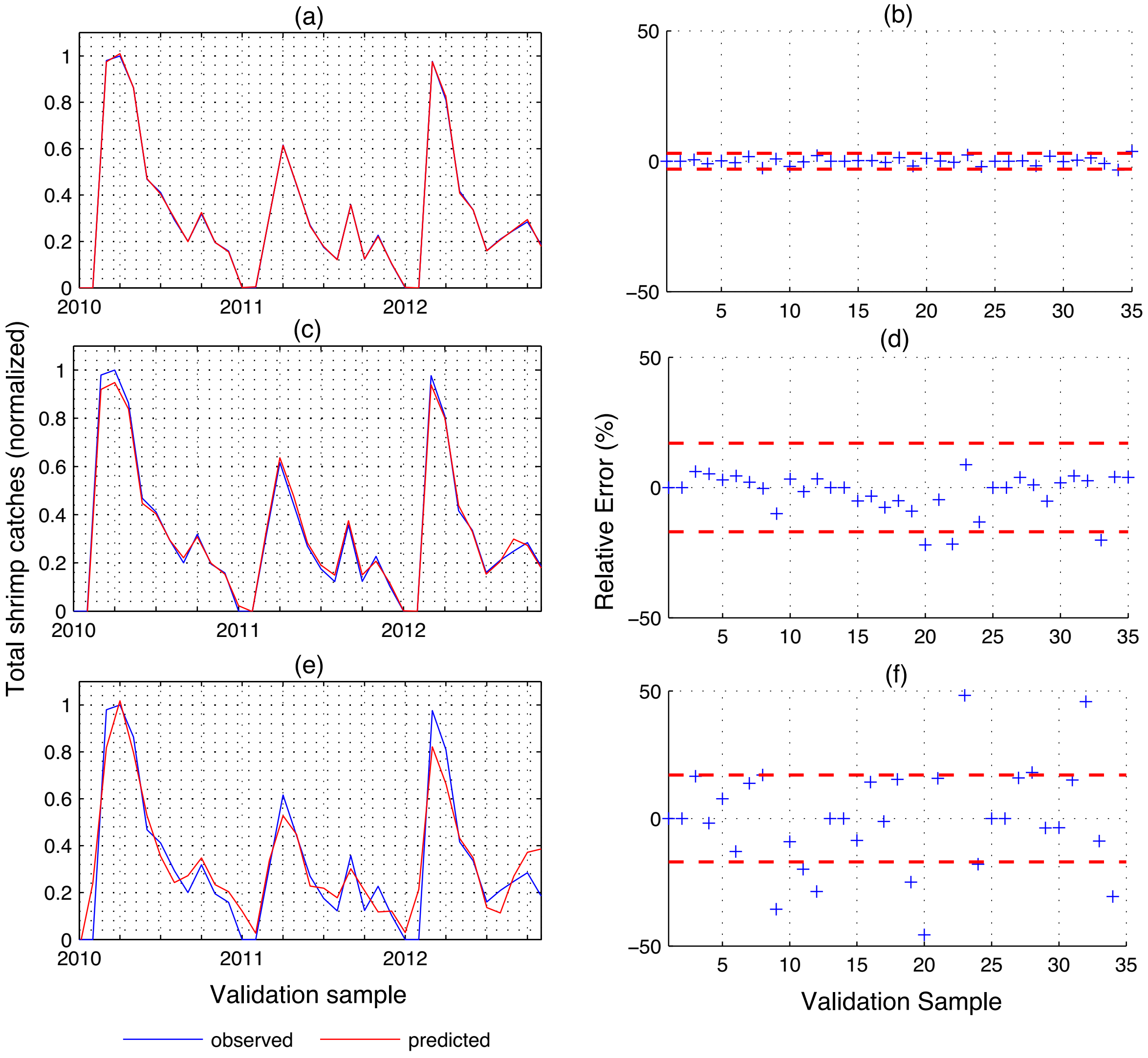

Figure 6. Shrimp catches prediction (a) SSA-AR (b) Relative Error SSA-AR (c) SSA-ANN-LM (d) Relative Error SSA-ANN-LM (e) SSA-ANN-CS (f) Relative Error SSA-ANN-CS

Figura 6. *Predicción de la captura de camarón (a) SSA-AR (b) Error Relativo SSA-AR (c) SSA-ANN-LM (d) Error Relativo SSA-ANN-LM (e) SSA-ANN-CS (f) Error Relativo SSA-ANN-CS*

the SSA method with an effective window length *L=6*. The model has been called SSA-AR.

The number of lagged values was calibrated through the training sample of the inter-annual component by using the GCV metric. The results are presented in Figure 2b, the highest accuracy was achieved for m=15 lags. The observed values versus the predicted values are shown in Figure 3. Table 1 presents the forecasting metric results, a MAPE of 1.7%, a R^2 of 99.6% and a RMSE of 0.7%. On the other hand 92.4 % of the predicted points presented a lower RE than ±5% as can be observed in Figure 3b,

The testing is executed with the validation sample using the calibration values. The results are presented in Figures 4a, 5a, 6a, and Tables 2, 3, and 4. From the figures and residual metrics, high accuracy was

Models Compared	Pitman's correlation
SSA-AR vs SSA-ANN-LM	-0.915
SSA-AR vs SSA-ANN-CS	-0.995
SSA-ANN-LM vs SSA-ANN-CS	-0.897

Table 5. Pitman's correlation, for pairwise comparisons of the models SSA-AR, SSA-ANN-LM, and SSA-ANN-CS.

Tabla 5. *Test de correlación de Pitman , Comparaciones emparejadas para SSA-AR, SSA-ANN-LM, y SSA-ANN-CS.*

observed in the prediction of both components (inter-annual and annual).

The values obtained with the forecasting metrics for the prediction of x_a via SSA-AR model for *L=6* are presented in Table 2. The model reaches a MAPE of 8.6%, a R^2 of 99.8%, and a RMSE of 0.84%. Whereas the prediction results of x_{ia} via SSA-AR model are shown in Table 3, and reach a MAPE of 2.3%, a R^2 of 99.1%, and a RMSE of 0.99%.

The prediction of shrimp catches (x_a+x_{ia}) via SSA-AR from window length *L=6*, are presented in Table 4, it shows a MAPE of 1.0%, a R^2 of 99.9%, and a RMSE of 0.48%. The relative percentage error is presented in Figures 4b, 5b, and 6b. In the prediction of the $\hat{x}_{ia}$ component, 91.4% of the points presents a lower RE than ±7%. In the prediction of the $\hat{x}_a$ component a lower accuracy was obtained, 74.3% of the points present a lower RE than ±15%. The prediction of the monthly catches present 94.3% of the predicted points with a lower error than ±3%.

On the other hand poor results were obtained through the application of SSA-AR based on a conventional window length of *L=N/2=78,* it reaches an MAPE of 46.6%, a R^2 of 35.5%, a RMSE of 23.5% and 28.6% of the predicted points present an error lower than ±15%.

The conventional AR (without the preprocessing stage) presents the poorest results, with an MAPE of 36.7%, a R^2 of 69.2%, a RMSE of 12%, and 25.7% of the predicted points show a relative error lower than ±15%.

The conventional ANN model presents an MAPE of 29.4%, a R^2 of 79.1%, a RMSE of 9.9%, and 34.4% of the predicted points a relative error of ±15%.

Prediction via SSA-ANN-LM model

The prediction of the SSA components through an auto-regressive neural network based on Levenberg-Marquardt is presented in this section. The ANN structure is (*m, Q,* 1), *m* (lags) was calibrated with the GCV metric for SSA-AR, the same value *m*=15 is used here. The number of hidden nodes was computed with $Q=log(N_t)$ (the formula normally used in our experiments), where N_t is the training sample size.

The prediction obtained via enhanced models with the validation sample is shown in Figures 3c, 4c, 5c, and Tables 2, 3, and 4. From the figures and residual metrics, high accuracy was reached in the prediction of both components.

The SSA-ANN-LM model presents an MAPE of 19.5%, a R^2 of 99.5%, and a RMSE of 1.5% for the prediction of x_{ia}. Whereas for x_{ia}, the SSA-ANN-LM model presents an MAPE of 3.1%, a R^2 of 98.4% and a RMSE of 1.3%. The prediction of monthly shrimp catches (x_a+x_{ia}) presents a MAPE of 5.3%, a R^2 of 99.4%, and a RMSE of 2.3%.

The relative percentage error is presented in Figures 3d, 4d, and 5d. In the prediction of x_{ia} 94.3% of the points present a lower RE than ±7%. In the prediction of the x_a 68.6% of the points present a lower RE than ±15%. In the prediction of the total shrimp catches, 91.4% of the points present a lower RE than ±15%.

Prediction via SSA-ANN-CS model

The prediction of the SSA components with the autoregressive neural network based on the Cuckoo Search is presented in this section. The ANN has the same structure as that presented in the SSA-ANN-LM model.

The CS algorithm was presented in the Methods Section, the number of nests used were 25, and the learning rate was set in $\rho a=0.25.$ The prediction executed with the validation sample is shown in Figures 4e, 5e, 6e, and Tables 2, 3, and 4; from the figures and residual values, a high accuracy was reached for the annual component.

The prediction of x_a via SSA-ANN-CS model presents an MAPE of 88.1%, a R^2 of 83.6%, and a RMSE of 9.6%. Whilst the prediction of x_{ia} presents an MAPE of 3.7%, a R^2 of 98.1%, and a RMSE of 1.5%. The prediction of shrimp catches (x_a+x_{ia}) presents a MAPE of 19.2%, a R^2 of 89.3%, and a RMSE of 9.4%.

The relative error is presented in Figures 4f, 5f, and 6f. For x_{ia} 88.6% of the points presents a lower RE than ±7%. For x_a 17.1% of the points presents a lower RE than ±15%. The total shrimp catch prediction presents 62.8% of the points with a lower RE than ±15%.

Tables 2, 3, and 4 show the evaluation of the three models enhanced with SSA. The best model is SSA-AR with the lowest residual values, the highest explained variance and the lowest relative error. The inter-annual component was predicted with high accuracy by using the three enhanced models; however in the prediction of the annual component there is a lower accuracy with both models based on the ANNs. Conventional ANN is a little more superior to the conventional AR, however SSA-AR presents a greater superiority compared to the SSA-ANN.

Pitman's Correlation Test for model errors

The performance of the three enhanced auto-regressive models for one-step ahead forecasting is evaluated with the Pitman's correlations test. The Pitman's correlations test is applied to identify the superiority of a model in pairwise comparisons; the test is computed with the correlation corr between Υ and Ψ (Hipel, 1994),

$$\Upsilon=e_1(i)+e_2(i), \quad i=1,\ldots, N$$
$$\Psi=e_1(i)-e_2(i), \quad i=1,\ldots, N$$

where e_1 and e_2 represent the residuals for one-step ahead forecasting obtained with the models 1 and 2 respectively, and *N* is the number of comparisons. The null hypothesis is significant at 5% significance level if corr >1.96/√N.

The evaluated correlations between Υ and Ψ are presented in Table 4.

The results obtained with the application of the Pitman's correlation test present that the model SSA-AR is superior to the models SSA-ANN-LM and SSA-ANN-CS. The complementary SSA-ANN-LM is superior to the SSA-ANN-CS.

Conclusions

In this paper three auto-regressive models enhanced with singular spectrum analysis for one-step ahead forecasting of shrimp trawl fishing are presented. The models were evaluated with the time series of shrimp monthly catches in the Gulf of Guayaquil from 2000 to 2012.

The SSA preprocesses the data to extract the annual and inter-annual components from the observed time series. Once the components have been extracted, they were predicted with three models, one linear and two non-linear; the conventional autoregressive model, an auto-regressive neural network with the Levenberg-Marquardt algorithm, and an artificial neural network with the Cuckoo Search.

The proposed three models SSA-AR, SSA-ANN-LM and SSA-ANN-CS show high accuracy in the prediction of shrimp catches and superiority with respect to conventional AR and ANN models (without preprocessing). The proposed models are also superior to the SSA-based model with a conventional window length. Although conventional ANN is a little superior to conventional AR, it was observed that the high gain was reached with SSA-AR with respect to the SSA-ANN-LM and SSA-ANN-CS.

The Pitman's correlation test confirms the superiority of the SSAS-AR model over the compared models at 5% significance level. The best model presents an MAPE of 1.0%, a RMSE of 0.48% and a R^2 of 99.9%.

Due to the promising results, this model will be evaluated with other time series related with fishery, and other diverse knowledge areas.

References

Abdollahzade, M., Miranian, A, Hassani, H. and Iranmanesh, H. 2015. A new hybrid enhanced local linear neuro-fuzzy model based on the optimized singular spectrum analysis and its application for nonlinear and chaotic time series forecasting. *Information Sciences*, 295, 107-125.

Barba. L., Rodríguez, N. and Montt, C. 2014. Smoothing Strategies Combined with ARIMA and Neural Networks to Improve the Forecasting of Traffic Accidents. *The Scientific World Journal,* vol. 2014, Article ID 152375, 12 pages.

Broomhead, D. and King, G.P. 1986. Extracting qualitative dynamics from experimental data. *Phys D: Nonlinear Phenom*, (20)217-236.

Chen, Q., Van Dam, T., Sneeuw, N., Collilieux, X., Weigelt M., and Rebischung P. 2013. Singular spectrum analysis for modeling seasonal signals from GPS time series. *Journal of Geodynamics*, 72, 25-35.

Freeman, J.A. and Skapura, D.M. 1991. Neural Networks, Algorithms, Applications, and Programming Techniques. Addison-Wesley, California, 401 pp.

Golyandina, N. and Stepanov, D. 2005. SSA-based approaches to analysis and forecast of multidimensional time series. *Proceedings of the Fifth Workshop on Simulation*, 293–298.

Hagan, M., Demuth, H.B. and Bealetitle, M. 2002. Neural Network Design. Hagan Publishing.

Hassani, H., Webster, A., Silva, E. and Heravi, S. 2015. Forecasting U.S. Tourist arrivals using optimal singular spectrum analysis. *Tourism Management*, 46, 322-335.

Instituto Nacional de Pesca (INP), 16/02/2015, http://www.institutopesca.gob.ec/.

Jiang, P. 2014. An Optimized Forecasting Approach Based on Grey Theory and Cuckoo Search Algorithm: A Case Study for Electricity Consumption in New South Wales. *Abstract and Applied Analysis,* Article ID 183095, 13 pages.

Kermani, B.G., Schiffman, S. and Nagle, H. 2005. Performance of the Levenberg Marquardt neural network training method in electronic nose applications. *Sensors and Actuators B*: *Chemical*, 110 (1) 13-22.

Marques, C., Ferreira, J.A., Rocha, A., Castanheira, J.M., Melo-Gonçalves, P., Vaz, N. and Dias, J.M. 2006. Singular spectrum analysis and forecasting of hydrological time series. *Physics and Chemistry of the Earth*, *Parts A/B/C,* 31 (18) 1172-1179.

Mukherjee, I. and Routroy, S. 2012. Comparing the performance of neural networks developed by using Levenberg-Marquardt and Quasi-Newton with the gradient descent algorithm for modelling a multiple response grinding process. *Expert Systems with Applications*, 39 (3) 2397-2407.

Rodríguez, N., Cubillos, C. and Rubio, J.M. 2014. Multi-step-Ahead Forecasting Model for Monthly Anchovy catches

Based on Wavelet Analysis. *Journal of Applied Mathematics,* Article ID 798464, 8 pages.

Telesca, L. Lovallo, M., Shaban, A., Darwich, T. and Amacha, N. 2013, Singular spectrum analysis and _shershannon analysis of spring ow time series: An application to Anjar Spring, Lebanon. *Physica A: Statistical Mechanics and its Applications,* 392 (17) 3789-3797.

Viljoen, H. and Nel, D.G. 2010. Common singular spectrum analysis of several time series. *Journal of Statistical Planning and Inference*, 140 (1) 260-267.

Walton, S., Hassan, O., Morgan, K. and Brown, M.R. 2013. A review of the development and applications of the Cuckoo Search Algorithm. *Swarm Intelligence and Bio-Inspired Computation*. Elsevier, Oxford, 257-271.

Wang, J. Sheng, Z., Zhou, B. and Zhou, S. 2014. Lightning potential forecast over Nanjing with denoised sounding-derived indices based on SSA and CS-BP neural network. *Atmospheric Research*, 137 (0) 245-256.

Xiao, Y., Liu, J., Hu, Y., Wang, Y., Lai, K. and Wang S. 2014. A neuro-fuzzy combination model based on singular spectrum analysis for air transport demand forecasting. *Journal of Air Transport Management,* 39, 1-11.

Yang, X.S. 2010. Nature-Inspired Metaheuristic Algorithms Second Edition. Luniver Press.

Yang, X.S. and Deb, S. 2009. Cuckoo search via Lévy Flights. Proc. of World Congress on Nature & Biologically Inspired Computing (NaBIC 2009), 210-214.

Yetilmezsoy, K. and Demirel, S. 2008. Artificial neural network (ANN) approach for modeling of Pb (II) adsorption from aqueous solution by Antep pistachio (Pistacia Vera L.) shells. *Journal of Hazardous Materials,* 153 (3) 1288-1300.

Recibido: septiembre 2016
Revisado: diciembre 2016
Aceptado: enero 2017
Publicado: septiembre 2018