The application of a general time series model to floodplain fisheries in the Amazon

Ronny O. Vallejos\textsuperscript{a1}, Nidia Noemi Fabré\textsuperscript{b}, Vandick da Silva Batista\textsuperscript{b}, Jonathan Acosta\textsuperscript{a}

\textsuperscript{a} Departamento de Matemática, Universidad Técnica Federico Santa María
Casilla 110-V, Valparaíso, Chile

\textsuperscript{b} Universidade Federal de Alagoas, Instituto de Ciências Biológicas e da Saúde,
Campus A. C. Simões, Av. Lourival Melo Mota, Tabuleiro do Martins, Maceió, AL, 57072-970, Brazil

Abstract

Time series analysis is a common tool in environmental and ecological studies to construct models to explain and forecast serially correlated data. There are several statistical techniques that are used to deal with univariate and multivariate (more than one series) chronological patterns of fisheries data. In this paper, an additive stochastic model is proposed with explicative and predictive features to capture the main seasonal patterns and trends of a fisheries system in the Amazon. Under the assumption that the multivariate fish yield is a linear function of its past values, other explanatory variables, an exogenous environmental variable, and seasonal variables, a parametric model is constructed to describe and predict the change of existing fish species living in a floodplain environment in the Central Amazon Basin. The estimation procedure is carried out via maximum likelihood estimation. The model explained, on average, 78\% of the variability in yield of the study species. The model represents the optimal solution (minimum mean square mean error) among the class of all multivariate autoregressive processes with exogenous and seasonal variables. Predictions for one period ahead are provided to illustrate how the model works in practice.

Keywords: Time series, Common features, VAR model, Forecast, Manaus Fishery

1. Introduction

The use of statistical techniques for predictive purposes is common in ecology and fisheries (e.g. McCleary and Hassan, 2008). However, although serial autocorrelation may be a common feature in nature, it has rarely been addressed in
these studies leading to biased parameter estimates, misinterpretation of significance levels and, consequently, erroneous inferences concerning the magnitude and direction of effects (Ostrom, 1990). An important aspect of time series analyses relates to periodic variations caused by biological, physical or other environmental phenomena (Shumway and Stoffer, 2000). Dealing with these influences is of fundamental importance when making forecasts, particularly when there is some form of environmental cycle. Seasonal changes are common in natural systems, generating oscillations in populations and affecting community structure. Such seasonality drive changes in biomass and ecosystem structure, modifying upward or downward the use of energy and materials in the system (Holling, 1973; Winemiller, 1990).

The development of models, which summarize serial variations, is an important strategy to draw ecological inference (Austin, 2002). This is particularly the case in tropical ecosystems (Bayley, 1995). The analysis of time series based on frequencies, commonly known as spectral analysis, is particularly relevant in ecology where series are frequently characterized by seasonal components and cycles that cannot be easily estimated by simple inspection (Arnade et al., 2005). This includes the analysis of the spectral density associated with linear processes, a mathematical transformation of the correlation function. The estimation of the spectral density, called periodogram, can be used via hypothesis testing to determine whether a particular frequency is significant. Spectral analysis can be combined with graphical tools and previous knowledge (usually fitting naive processes) to provide more sophisticated models that are either better able to describe some patterns of the series, or are able to provide precise forecasting for unobserved future values of the process (Allende et al. 2002).

Many tropical aquatic ecosystems show cyclical patterns of flooding (Hamilton et al., 1996). Cycles regularly observed over time have already been described in the Amazon region, and their correlation with flooded areas has been used to reconstruct regional inundation patterns (Sippel et al. 1998). Moreover, the ecological effects of regular flood cycles are the main basis for the development of Junk’s influential flood-pulse concept (Junk et al. 1989). Flood cycles have also been related to patterns of fish landings and primary production outputs which also follow cycles regularly observed over time (e.g. Arnade and Gehlhar, 2005; Gates, 2000; Waugh and Miller, 1970). If these cycles are not rigorously accounted with during analysis, they can introduce statistical noise into temporal and spatial distribution analysis of abundance. Consequently, fisheries managers may make decisions based on biased information.

Various types of mathematical or statistical models have been used for the analysis of fisheries data (multivariate regression models, exponential smoothing, and autoregressive integrated moving average (ARIMA) processes, Stergiou et al., 1997). In the context of fisheries, ARIMA models are considered efficient tools for the analysis of the fishery for pilchard (Koutroumanidis et al. 2006; Stergiou, 1989; Stergiou and Christou, 1996; Stergiou et al. 1997), and are most commonly used to make a few number (short-term) of future predictions (Stergiou, 1989; Lloret et al. 2000; Parsons and Colbourne, 2000).
Extensions of ARIMA models for series whose correlation function decays slowly as a function of the lag distance (long memory time series) have been used to study monthly river flows (Ooms and Franses, 2001). More recently, recursive algorithms that produce optimal estimates (e.g. the Kalman filter) have been used when there is large natural variability and measurement error in fisheries data (Peterman et al. 2003). Neurofuzzy networks have been developed to improve wastewater flow-rate forecasting (Fernandez et al. 2009). Multivariate time series have also been used to complement the univariate analyses, providing cross-correlation computations and information on stability, common features and causality (Engel and Granger, 1987; Engel and Kozichi, 1993). The vector autoregressive (VAR) model is one of the most common tools to describe multivariate characteristics and interactions of a time series vector, considering the serial correlation inside and across the series (Lutkephol, 2005; Gan 2005). Some of the models described above can be combined in a way to analyze time series with more complex patterns.

In this paper, an additive (combined) model is proposed with explicative and predictive properties to capture the key characteristics and changes of existing fish species that are part of a cyclically complex fisheries system in the Amazon. The model was constructed on the assumption that the multivariate response variable (fish yield) can be decomposed into three terms: an autoregression of the response variables, an exogenous environmental variable, and a seasonal component. The model was used to describe and predict the tropical floodplain environmental seasonal change using yield of fish species (r2 species) that show seasonal fluctuations (Barletta et al. 2010; Winemiller and Rose, 1992) - the main fishery resource guild in the Central Amazon Basin.

The remainder of the paper proceeds as follows. The next section presents a description of the study area, data set, and two statistical techniques: the notion of common features and multivariate autoregressive processes with seasonal and exogenous components. In Section 3 the main results are reported, including a brief univariate analysis of the times series, the estimation of the most relevant frequencies, the estimation of the parameters of the proposed model, residual analysis, forecasting, and cross-validation. Finally, a discussion including possible topics for further research is presented in Section 5.

2. Materials and Methods

The following section is divided in four parts. First, a description of data set, study area, data collection, and the variables included in the study is given. Second, the common features framework are explained. This procedure is a decision rule to decide if a feature (trend, cycle or association) is common to a set of variables. Third, a motivation of the statistical model used for predictive purposes in an ecological context is given. Fourth, the multivariate autoregressive processes (VAR) are introduced as a useful tool to obtain predictions. Then an additive VAR process, that is a combination of a multivariate autoregressive process, a seasonal variable, and an exogenous variable, is described. For those readers in-
interested in the mathematical details of the methods, supplementary information is provided in Appendix B.

2.1. Data Sources

2.1.1. Study area

The study was carried out in an Amazon ecoregion and uses fisheries data from the Amazon basin (Figure 1) which covers about 700,000 km² (Santos and Ferreira, 1999). Within the basin there is seasonal variation that recur regularly governed by the alternation of dry and wet seasons (although always in a warm, humid tropical climate). The basin is delimited to the north by the Guiana Shield, to the south by the Central-Brazil Shield, and to the west by the Sub-Andean foreland (Sioli, 1984). Geochemically, the basin can be divided into three subsystems: Western, North-South, and Central Amazon (Fittkau et al., 1975). This last region contains a large floodplain area enriched by Andean sediments through the main channel and effluents (the Purus, the Juruá and the Japurá - Figure 1) with nutrients that indirectly sustain the aquatic fauna. Commercial fishing landings in Central Amazon are centralized in Manaus (Batista and Petrere, 2003; Petrere, 1978). The Manaus small-scale fishery is typically based on fish caught from canoes using purse seine nets (Batista et al. 2004). The fish are then transferred to larger vessels which transport the catch to Manaus.

2.1.2. Data set

Landing data series were collected in the main fishing harbor of Central Amazon (Manaus city). The data were collected on a daily basis from January 1994 to December 2004 by the Federal University of Amazonas. Fishery yield time series for the years 1994-2004 were extracted from a central database. The following time series data were considered: 1. total monthly catch of commercial fishing boats (i.e. all fishing yields combined); 2. total monthly commercial catches of curimatá (Prochilodus nigricans), jaraqui-fina (Semaprochilodus taenirurus), jaraqui-grossa (Semaprochilodus insignis), matrinxã (Brycon amazonicus), pacu-comum (Mylossoma duriventris), sardinhas (Triportheus spp), and tambaqui (Colossoma macropomum). These species are the most important fish species landed in Manaus in terms of yield, totaling more than 70% of the total Central Amazon catches (Batista and Petrere, 2003).

The daily river level series in Manaus was supplied by the Brazilian Geological Survey (CPRM) and was available for most days over the study period. However, as the lack of data for several days could bias the monthly average, we used average values for the 14th, 15th and 16th days each month. The independent variables used for the development of the models are listed in Table 1.

2.2. Common Features in Time Series

Engle and Kozichii (1993) introduced a class of statistical tests for the hypothesis that some feature of a data set is common to several variables. Examples are serial correlation, trends, seasonality. They developed the necessary conditions to have a the following definition: A feature that is present in each of a group
Figure 1: Study region in the North of Brazil.

Table 1: Variables and species under study.

<table>
<thead>
<tr>
<th>$x_{1,t}$</th>
<th>Total yield</th>
<th>$x_{6,t}$</th>
<th>Pacu-comum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_{2,t}$</td>
<td>Curimatá</td>
<td>$x_{7,t}$</td>
<td>Sardinhas</td>
</tr>
<tr>
<td>$x_{3,t}$</td>
<td>Jaraqui-Fina</td>
<td>$x_{8,t}$</td>
<td>Tambaqui</td>
</tr>
<tr>
<td>$x_{4,t}$</td>
<td>Jaraqui-Grossa</td>
<td>$x_{9,t}$</td>
<td>River Level</td>
</tr>
<tr>
<td>$x_{5,t}$</td>
<td>Matrinxã</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

of series is said to be common to those series if there exists a non-zero linear combination of the series that does not have the feature. To formalize these ideas the authors restricted the analysis to regression models so that the presence or absence of a common feature can be addressed considering an hypothesis testing
procedure for a suitable parameter of interest. Engle and Kozichi (1993) used a regression model of the form

\[ y_t = x_t \beta + z_t \gamma + \varepsilon_t, \quad t = 1, 2, \ldots, T, \]  

(1)

where \( H_0 : \gamma = 0 \), represents no feature and \( H_1 : \gamma \neq 0 \) represents the inclusion or presence of a feature. About the parameter \( \gamma = 0 \) is necessary to assume some distributional assumptions for the vector \( \{y, x, z\} \). This ensures that a Lagrange Multiplier (LM) statistic has a \( \chi^2 \) limiting distribution. The reader is referred to Appendix A for the mathematical derivations associated to the common feature test.

In Section 3.1 the common feature procedure will be used to test if an annual cycle is common to each series. Also, the same test will be considered to test if the first and second order autocorrelations are significantly common to all series.

2.3. Modelling Fish Yield Production

The conceptual basis of the model we propose in Equation (3) is based on the empirically based assumption that variation of the river level affects the fish population dynamics as much as changes in fishing pressure. The model also incorporates a historical signal based on the time series of the past landings, i.e. the present value of landings can be broadly predicted from recent past values. Such a model with high correlations between river level oscillation and fishery yield could realistically describe temporal continuities or synchronicities (Moran, 1953) within cycles of large scale events. The proposed model also incorporates the uncertainty related to the ecological system through stochastic variables that take into account the variation of the system due to those variables that are not considered in the model. The suggested model is a variant of the widely used decomposition models in time series. A remarkable good feature that the model preserves is the linear structure allowing to estimate the unknown parameters in the same way as in a regression model.

2.4. Multivariate Autoregressive Models with Seasonal and Exogenous Components

Let us consider a multivariate time series with \( K \) components, say

\[ y_t = (x_{1,t}, x_{2,t}, \ldots, x_{K,t})'. \]

The random vector \( y_t \) follows a multivariate autoregressive (VAR) model of order \( p \) if

\[ y_t = \nu + A_1 y_{t-1} + A_2 y_{t-2} + \ldots + A_p y_{t-p} + u_t, \quad t = 0, \pm 1, \pm 2, \ldots \]  

(2)

where \( \nu = (\nu_1, \nu_2, \ldots, \nu_K)' \) is a fixed \( K \times 1 \) vector of intercept terms, \( u_t = (u_{1,t}, u_{2,t}, \ldots, u_{K,t})' \) is a multivariate white noise, i.e. \( \mathbb{E}(u_t) = 0, \mathbb{E}(u_s, u_t') = 0 \), for all \( s \neq t \), \( \mathbb{E}(u_t, u_t') = \Sigma_u \), and \( A_i, i = 1, 2, \ldots, p \) are fixed \( K \times K \) coefficient matrices. Similar to the one-dimensional case, the process (2) is stable if its reverse characteristic polynomial has no roots in and on the complex unit circle (Reinsel, 2003).
Now, let us assume that model (2) holds and that the observed values are $y_1, y_2, \ldots, y_T$. There are several ways to estimate the parameters of the model such as (2). The least squares estimates, the Yule-Walker estimates and the maximum likelihood estimates. In Appendix B there is a description of the maximum likelihood estimates assuming that the distribution of the process is Gaussian, following the notation of Luthkephol (2005).

When there is clear seasonality in each variable, one approach to described the process $y_t$ is to consider the seasonal part as a deterministic component that can be estimated independently using spectral methods. Then, we suggest a model in which the response variable can be decomposed into three parts: seasonal variables ($W_t$) that captures the periods of each variable, exogenous variables ($V_t$) that are relevant to predict or explain the response variable, and an autoregressive term $Z_t$ that describes present information as a linear combination of the past until a certain fixed lag. In addition the model contains a white noise sequence that summarizes all uncertainty due to the fact that variables could exist that help to explain the response variable, but these variables are not present in the model. Hence the additive model for the multivariate response $y_t$ is

$$y_t = \nu + W_t + V_t + Z_t + u_t,$$

where $\nu = (\nu_1, \nu_2, \ldots, \nu_K)'$ is a fixed $K \times 1$ vector of intercept terms, $W_t = (w_{1,t}, w_{2,t}, \ldots, w_{K,t})'$, $\omega_{it}$ represents the seasonal behavior of the $i$-th species that is modeled by $w_{i,t} = \sum_{j=0}^{k_i} C_{ij} \cos(2\pi \omega_i j t / n) + D_{ij} \sin(2\pi \omega_i j t / n)$, $i = 1, 2, \ldots, K$, $n$ is the sample size, $V_t = BE_t$, $E_t = (E_{1,t}, E_{2,t}, \ldots, E_{L,t})'$ is a set of environmental variables that are modeled as exogen variables (river level and its past values), $B$ is a matrix of size $K \times L$, $Z_t$ represents the linear autoregression of the species, $Z_t = A_1 y_{t-1} + \ldots + A_p y_{t-p}$, and $u_t$ is a white noise sequence with covariance matrix $\Sigma_u$. Notice that the coefficients $C_{ij}, D_{ij}, A_i, \nu$ and $B$ can be estimated using maximum likelihood or least squares estimation as an extension of the method described in the Appendix B. The additive model (3) is used to explain and predict the variables $x_{1,t}, x_{2,t}, \ldots, x_{K,t}$.

The computational routines constructed to implement model (3), to compute the predictions, and to generate all figures, were developed in the programming language R. These routines are available upon request.

3. Results

3.1. Univariate Analysis and Testing for Common Features

Considering each variable as a single component, our goal was to describe the cycles and the correlation structure inside and across the series. Formula (A.3) was used to test for common cycles, and for first and second order autocorrelation of each series. There are clear seasonal components in each series (Figure C.4). Applying moving average filters and then fitting simple regression models to each smoothed variable we observed no significant trends in the series. The average river showed a clear periodicity and the highest values for
The total yield were associated with the highest values of the river level. Irregular patterns are apparent over short periods of time in some series while some series have an inverse relationship with others. The Pearson correlation between all possible pairs of variables was computed obtaining the correlation matrix (4). Highest values were highlighted: $\text{cor}[\text{Pacu-comum}(x_{6,t}), \text{Sardinhas}(x_{7,t})] = 0.722$ and $\text{cor}[\text{Pacu-comum}(x_{6,t}), \text{Total yield}(x_{1,t})] = 0.690$.

\[
\begin{pmatrix}
1.000 \\
0.507 \\
0.331 \\
0.357 \\
0.248 \\
0.690 \\
0.418 \\
0.150 \\
0.165
\end{pmatrix}
\begin{pmatrix}
1.000 \\
1.000 \\
1.000 \\
1.000 \\
1.000 \\
1.000 \\
1.000 \\
1.000 \\
1.000
\end{pmatrix}
\]

The autocorrelation and cross correlation functions were computed for all variables. First and second order significant correlations were found inside and across the variables. Also, the cross correlation functions sometimes generated a seasonal behavior (not shown here). Similar patterns for the partial correlation function were observed for other variables. Moreover, because all variables exhibit a stable behavior for the mean and variance and there are significant first or second order serial autocorrelation, standard approaches such as the Box-Jenkins methods are appropriate.

A basic spectral analysis of all variables, revealed a common peak in the periodogram at $w_0 = 11/132 = 0.0833$. This value corresponds to a period of 12 months. In order to confirm that the frequency $w_0$ was present in each series we used the test function \(A.3\) with a significance level of 0.05. The regression model

\[x_{i,t} = \beta_i + \gamma_{i1} \sin(2\pi w_0 t) + \gamma_{i2} \cos(2\pi w_0 t),\]

for all \(i = 1, \ldots, 9\), \(t = 1, \ldots, 132\) was used, where \(\gamma_{i1}\) and \(\gamma_{i2}\) are the parameters to be tested. In all cases we compared the test function \(s(y_i)\) with the quantile 95 of a

<table>
<thead>
<tr>
<th>Total yield</th>
<th>$\hat{\gamma}_{i1}$</th>
<th>$\hat{\gamma}_{i2}$</th>
<th>$s(y_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>x_{1,t}</td>
<td>-269.96</td>
<td>-551.62</td>
<td>25.13</td>
</tr>
<tr>
<td>Curimatá</td>
<td>x_{2,t}</td>
<td>-132.94</td>
<td>-101.84</td>
</tr>
<tr>
<td>Jaraqui-Fina</td>
<td>x_{3,t}</td>
<td>-073.54</td>
<td>002.04</td>
</tr>
<tr>
<td>Jaraqui-Grossa</td>
<td>x_{4,t}</td>
<td>365.89</td>
<td>067.65</td>
</tr>
<tr>
<td>Matrinxã</td>
<td>x_{5,t}</td>
<td>124.89</td>
<td>-098.43</td>
</tr>
<tr>
<td>Pacu-comum</td>
<td>x_{6,t}</td>
<td>-174.41</td>
<td>-223.36</td>
</tr>
<tr>
<td>Sardinhas</td>
<td>x_{7,t}</td>
<td>-145.29</td>
<td>-069.35</td>
</tr>
<tr>
<td>Tambaqui</td>
<td>x_{8,t}</td>
<td>-040.93</td>
<td>-058.07</td>
</tr>
<tr>
<td>River Level</td>
<td>x_{9,t}</td>
<td>298.10</td>
<td>-359.78</td>
</tr>
</tbody>
</table>
chi-square distribution with 2 degrees of freedom (Table 2). Notice that for all \( i = 1, 2, \ldots, 9 \), \( s(y_i) > \chi^2_{0.95}(2) = 5.99 \). Thus the frequency \( w_0 \) was present in each one of the variables \( x_{1,t}, \ldots, x_{9,t} \). We tested if the first order autocorrelation is present in each series when we assume that the variables follow a first order multivariate autoregressive (VAR(1)) process. We computed the test function (A.3) for all possible pair of variables. In all cases the test function was greater than \( \chi^2_{0.95}(2) = 5.99 \) (Table 3), thus the first order autocorrelation was present in all individual series. The same conclusion was obtained for second order autocorrelation when assuming a VAR(2) process for the series. Collating the information provided by simple and partial autocorrelation functions, it was adequate to assume a VAR(2) model for the vector of the study series. Moreover the Akaike’s criterion (Lutkepohl, 2005) indicated that among the VAR(\( p \)) processes with \( p \leq 10 \) the VAR(2) model has associated the smallest coefficient.

### 3.2. Fitting a Multivariate Autoregressive Model

The additive model (3) was fitted to the 8 catch series considering the river level and its lag values as exogenous variables. The estimated frequencies were obtained for each variable, using the periodogram to preserve the linearity of the model with respect to the unknown parameters (Table 4). To obtain a suitable number of frequencies, the \( R^2 \) coefficient associated with an harmonic regression model was computed. In each case, the most relevant frequencies were kept. Alternatively, statistical inference for the frequencies can be carried out by using the fact that twice the ratio between the periodogram and the spectral density associated to a stationary process at frequency \( \omega \) has a limiting chi-square distribution with two degrees of freedom. This distributional result can then be used to derive approximate confidence intervals for the spectrum (Fuller, 1995).

Now, assuming that all frequencies are known, we particularized the additive model (3) for the 8 monthly catch series as a response variable. That is

\[
y_t = \nu + W_t + BE_t + A_1 y_{t-1} + A_2 y_{t-2} + u_t, \quad t = 1, \ldots, 132,
\]
Table 4: Estimated frequencies $\tilde{w}_{ij}$ amplified by the sample size, for the 8 monthly catch series $x_{1t}, \ldots, x_{8t}$. $\hat{w}_{ij} = \tilde{w}_{ij}/132.$

<table>
<thead>
<tr>
<th></th>
<th>$\hat{w}_{i,0}$</th>
<th>$\hat{w}_{i,1}$</th>
<th>$\hat{w}_{i,2}$</th>
<th>$\hat{w}_{i,3}$</th>
<th>$\hat{w}_{i,4}$</th>
<th>$\hat{w}_{i,5}$</th>
<th>$\hat{w}_{i,6}$</th>
<th>$\hat{w}_{i,7}$</th>
<th>$\hat{w}_{i,8}$</th>
<th>$\hat{w}_{i,9}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_{1t}$</td>
<td>11</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_{2t}$</td>
<td>11</td>
<td>4</td>
<td>7</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_{3t}$</td>
<td>22</td>
<td>1</td>
<td>11</td>
<td>2</td>
<td>12</td>
<td>44</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_{4t}$</td>
<td>11</td>
<td>1</td>
<td>22</td>
<td>2</td>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_{5t}$</td>
<td>11</td>
<td>7</td>
<td>4</td>
<td>22</td>
<td>1</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_{6t}$</td>
<td>11</td>
<td>1</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_{7t}$</td>
<td>11</td>
<td>22</td>
<td>10</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_{8t}$</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>11</td>
<td>9</td>
<td>7</td>
<td>6</td>
<td>8</td>
</tr>
</tbody>
</table>

where $y_t = (x_{1t}, x_{2t}, \ldots, x_{8t})'$, $B$ is a $8 \times 13$ matrix of coefficients, and $E_t$ represents the exogenous vector defined as $E_t = (x_{9,t}, x_{9,t-1}, \ldots, x_{9,t-12})'$, $p = 2$ and $u_t$ is a multivariate white noise as in (3). By maximum likelihood, the estimated model was

$$\hat{y}_t = \hat{\nu} + \hat{W}_t + \hat{B}E_t + \hat{A}_1 y_{t-1} + \hat{A}_2 y_{t-2}$$

where $\hat{\nu}, \hat{W}_t, \hat{B}, \hat{A}_1$ and $\hat{A}_2$ are given in the Appendix C.

On average, 78% of the variability of the response variable was accounted for by the statistical model (3). For $x_{1t}$ (Total Landings), the 83% of the variability was accounted by model (3). For $x_{2t}$ (Curimatá), $x_{3t}$ (Jaraqui-Fina), $x_{4t}$ (Jaraqui-Grossa), $x_{5t}$ (Matrinxã), $x_{6t}$ (Pacu-comum), $x_{7t}$ (Sardinhas), $x_{8t}$ (Tambaqui), the percentages of variability explained by model (3) were 63%, 72%, 76%, 78%, 85%, 82%, 86%, respectively. If we disregard the seasonal effect ($V_t$), the variability of the response variables explained by the model decreases to approximately 43%, indicating the importance of the seasonal cycles in the modeling of the landings.

A portmanteau test (Ljung and Box, 1978, Lutkephol, 2005) for the residual autocorrelations was not significant up to lag $H = 20$ ($p \approx 0.53$). Thus, there was not enough evidence to reject the hypothesis of no autocorrelation among the residuals. The Jarque-Bera test for multivariate normality of a white noise process (Lutkephol, 2005) provided sufficient evidence ($p < 0.0001$) against the normality assumption of the residuals. Although one of the assumptions was not satisfied, the predictive capability of model (3) was remarkable good as will be highlighted in the next section.

### 3.3. Forecasting of Species Yields

To forecast the variables of interest ($x_{1t} - x_{8t}$) it was necessary to propose an initial model for the exogenous variable $x_{9t}$. We proposed two models for $x_{9t}$ (river level). The first one is a combination between a seasonal part and an...
ARMA component, in the form
\[ x_{9,t} = \mu + \sum_{l=0}^{2} [C_l \cos(2\pi \omega_l t/n) + D_l \sin(2\pi \omega_l t/n)] + \sum_{i=1}^{p} \phi_i x_{9,t-i} + \sum_{j=0}^{q} \theta_j \varepsilon_{t-j}, \quad (5) \]

where \( \omega_0 = 11, \omega_1 = 22, \) and \( \varepsilon_t \) is a white noise sequence with variance \( \sigma^2. \)

The second one is a SARIMA(2,0,2) \( \times (1,0,1)_{12} \) model described by the equation
\[ (1 - \phi_1 B - \phi_2 B^2)(1 - \varphi B^{12})x_{9,t} = \alpha + (1 + \theta_1 B + \theta_2 B^2)(1 + \vartheta B^{12})\varepsilon_t. \quad (6) \]

Akaike’s criterion was used to choose the parameters \( p \) and \( q. \) For both models the maximum likelihood method provided estimations for the parameters. To explore the capability of each model to predict future values, a cross-validation experiment was carried out. The original data set was partitioned into two parts (app. 90% and 10%). The \( L_2 \) norm (mean squared error)
\[ ||\hat{\varepsilon}||_{L_2}^2 = \hat{\varepsilon}^t \hat{\varepsilon} = \sum_{i=1}^{2} \hat{\varepsilon}_i^2 \quad (7) \]

was used to summarized deviations of the predicted values from the observed ones for the last 10% of the data. The results are shown in Table 5. The percentage of variability explained by both models was the same, but model (6) had a smaller prediction error. The parameter estimates for model (6) were the following: \( \hat{\phi}_1 = 1.67, \hat{\phi}_2 = -0.70, \hat{\varphi} = -0.62, \hat{\alpha} = 2360.59, \hat{\theta}_1 = -0.37, \hat{\theta}_2 = 0.99, \hat{\vartheta} = -0.82. \)

In Figure 2 we observe the river level, the fitted values and 12 forecasts of the river level for 2005. Model (6) accurately predicted the observed values. The forecasted values for 2005 preserve the original seasonal patterns in the series.

Residual analysis showed that the normality assumption hypothesis and the independence among the residuals of model (6) were satisfied at the 5% significance level. In practice, forecasts provided by model (5) and (6) were quite similar.

In order to predict variables \( x_{1,t} - x_{8,t} \) for 2005, we compared model (3) with four models, VAR(2), VAR(3), VAR(2)+exogenous variable and VAR(2)+seasonal Part. To quantify the discrepancy between the observed and predicted values another cross-validation experiment was carried out. In this case the \( L_2 \) Frobenius
Figure 2: Predicted values of the river level provided by model (6) for 2005. The vertical line separates the observed and predicted values.

The norm for matrices given by

\[ ||A||_F^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 = tr(A^T A), \]

was used. The results are shown in Table 6. Model (3) has associated the smallest value for the \( L_2 \) norm, highlighting its capability of prediction.

Table 6: \( ||\hat{\varepsilon}||_F \) for 5 multivariate models.

| Model              | \( ||\hat{\varepsilon}||_F \) |
|--------------------|-------------------------------|
| VAR(2)             | 8297.890                      |
| VAR(3)             | 9563.656                      |
| VAR(2)+EXOGENOUS   | 7519.741                      |
| VAR(2)+SEASONAL    | 6164.942                      |
| Model (3)          | 5411.081                      |

Predictions for a future year about which nothing is known are shown for the species on different plots in Figure 3. The seasonal behavior is typically preserved by the forecasted values except for \( x_{8,t} \) (Tambaqui), in which the forecasts showed an unexpected decreasing trend not present in the observed values for the period 1994-2004. The series \( x_{8,t} \) has an irregular behavior for the period 1995-1996. This makes it difficult to accurately predict this component. In general, stationary processes do not perform well for series having drastic structural changes - this is
Figure 3: Predicted values of $x_{1t} - x_{8t}$ for January-December of 2005. The vertical line separates the observed and predicted values.
not the case for almost all the series used in this work. However, irregular series (that do not appear to be stationary sequences) can be transformed using the difference operator $\nabla x_t = x_t - x_{t-1}$.

4. Discussion

Spatial and temporal autocorrelations have frequently been ignored in the analysis of ecological and fisheries data (Bence, 1995; Stockhausen and Fogarty, 2007; Telford and Birks, 2005). When existing spatial or temporal association is ignored, there can be several potential effects on estimations: the variance could be inflated yielding wide confidence intervals for the parameters of interest, spurious association can lead to incorrect inferences about the relationships between independent variables (Lennon, 2000), and, in general, the estimations may lack precision (lack of consistency) when the sample size increases (Schabenberger and Gotway, 2005).

The model developed here successfully captured the effect of seasonal cycles driven by water level changes in according to flood pulse concept proposed by Junk et al. (1989). A flexible general model was developed to account for these autocorrelations that can be used according to researcher needs and data requirements. The model was developed to predict fish species yield in eleven serial landing data sets. It has two main distinguishing characteristics. First, a univariate time series analysis is performed in order to identify what variables should be included in the multivariate model. A common features analysis confirmed that the first and second order serial correlation is common to all individual series. The same behavior was observed for the 12 month period. Second, the model is flexible enough to incorporate other factors that may be important in other ecological systems. The model also allows some factors to be modified to account for different scenarios. For example, if the autoregressive and the cyclical variables explain a high percentage of the variability of each response variable, the introduction of an exogenous variable can be omitted.

The application of this model needs a time series for the response variable of at least two completed periods (24 months), so that the computational routines fit the model run properly (Venables and Ripley, 2010). These data must be stationary (no trends), but in practice the series can be initially detrended or the model can be modified to include a specific trend.

Interpretation of the model output should be initially focused on the analysis of autocorrelation and cross correlation among variables, followed by a spectral analysis. The peaks of the resulting periodogram scan can be used to identify key patterns useful for prediction. In this analysis it is important to use the estimated frequencies as known quantities in model (3) in order to use simple estimation methods for linear models. After the estimation process, we should focus only on the significant estimations in order to reduce the dimension of the estimated model as much as possible. Once the non-significant variables have been disregarded from the final model, the forecast can be computed for future unobserved values in time.
The main drawbacks of model (3) are the distributional assumptions on the residuals (normality) and the independence between them. In practice, it is common to deal with distributions having tails heavier than normal distributions. Also, the residuals associated to model (3) could be serially correlated violating the independence assumption. In general, the choice of a suitable correlation structure for the errors is not an easy task.

Time is explicitly considered within the model as an independent factor that influences ecological systems at an evolutionary scale (the four-dimensional concept of lotic ecosystems - Ward, 1989). Thus, time is an important variable, but one that is hard to evaluate since it typically requires long-term studies that are very rare globally and even more so in small-scale fisheries research. Although an evolutionary time scale was not used here, historical variations may be usefully incorporated into the dynamics of a system, generating a time continuum of responses that maintain the system’s integrity together with lower order variations around the seasonal cycle. At the annual level, there was an inertial effect of the river pulse, generating highly predictable fishery yields. This effect is not peculiar to the Amazon, and can also be observed in fisheries yield data from marine (Stergiou, 1989; Stergiou and Christou, 1996; Stergiou et al. 1997) and other freshwater environments (Welcomme, 1985).

River level was found to be an important variable for predicting the yield of a species with a seasonal strategy and for predicting the total yield including species with an equilibrium strategy. Temporal continuity of river level is important for the continuity of fisheries yield and likely to the stability of production within the aquatic system. It is well known that river level is a powerful explanatory variable related to the biological productivity in tropical flood systems (e.g. Welcomme, 1985). However, the application of the continuity solution to fisheries yield generated by the river pulse (and possibly into biological production) is innovative. This was enough to generate predictions of 12 month cycles with only 2 months of lag independent of other variables. Small time lags indicate that stocks may not quickly return to a stationary distribution following a perturbation (Ives et al. 2010), so present worries about disruption of the system, including catastrophic events or periods must be taken seriously in the management of fishery resources in the Amazon. Such a strong environmental influence may be related to climatic phenomena that change seasonally or interannually and which may drive predictable cycles of different magnitude and length. These small scale temporal autocorrelations generate a temporal continuum that creates a cyclic pattern independent of the statistical ‘noise’ of other fishery or climate variables. Such knowledge gained from a multivariate time series approach can be used to aid fish and environmental managers in the design of policies to further our understanding of the time scale of change of key variables.

Future research should focus on identifying significant new exogenous variables that could eventually be included in the multivariate model. These variables could provide new information and insights that are not presently considered in model - for example, the spatial location where fish are captured or the river surface temperature. It would also be interesting to compare the performance of the
model with other widely used models that are alternatives to classical time series analysis (e.g. neural networks and support vector machines).

5. Acknowledgements

This study was supported by the Brazilian National Research Council (CNPq), the National Fund for the Environment (FNMA) and the United Nations Development Program Project (BRA/00/008). VSB and NNF were supported by CNPQ research grants and ROV was partially supported by Centro Científico y Tecnológico de Valparaiso FB 0821 (grant FB/01RV/11), UTFSM, Chile and Programa Basal CMM U. de Chile. We also wish to thank the Federal University of Amazonas and Integrated Program on Aquatic and Floodplains Resources/PYRA laboratory team for their assistance, and R.J.Ladle a for comments and English revision.

Appendix A. Common Features

Under regularity conditions (Engle and Kozichi, 1993) the series \(\{y, x, z\}\) are jointly stationary. Then the least squares estimator and the estimated covariance matrix of \(\gamma\) (\(\gamma\) as in Equation 1) can be written as

\[
\hat{\gamma} = (z'M_xz)^{-1}z'M_xy, \\
\mathbb{V}(\hat{\gamma}) = \sigma_y^2(z'M_xz)^{-1},
\]

where \(x = (x_1, \ldots, x_T)'\), \(M_x = I - x(x'x)^{-1}x'\), and \(I\) is the identity matrix of order \(T\). Then the test statistic is

\[
s(y) = y'M_xz(z'M_xz)^{-1}z'M_xy/\hat{\sigma}_y^2,
\]

where \(\hat{\sigma}_y^2\) is a consistent estimator of the variance of \(y\) given \(\{x, y\}\). Commonly

\[
\hat{\sigma}_y^2 = e'y'e_y/T,
\]

where \(e_yt = y_t - x_t\tilde{\beta}\), and \(\tilde{\beta}\) is the least squares estimate of \(y\) at \(x\). Then

\[
s(y)T(y'z(z')^{-1}z'y/y'y).
\]

Suppose that the two time series \(y_{1t}\) and \(y_{2t}\), are going to be tested for a common feature assuming the models

\[
y_{1t} = x_t\tilde{\beta}_1 + z_t\gamma_1 + \varepsilon_{1t}, \quad (A.1)
\]
\[
y_{2t} = x_t\tilde{\beta}_2 + z_t\gamma_2 + \varepsilon_{2t}, \quad (A.2)
\]

for \(t = 1, 2, \ldots, T\). Then

\[
s(\hat{u}) = \hat{u}'M_xz(z'M_xz)^{-1}z'M_x\hat{u}/\hat{\sigma}_u^2, \quad (A.3)
\]
where \( \hat{u}_t = \hat{y}_{1t} - \delta \hat{y}_{2t} \) and \( \hat{\sigma}_u^2 \) is a consistent estimator of the residual variance. For an LM type test typically
\[
\hat{\sigma}_u^2 = \hat{u}'M_x\hat{u}/T. \tag{A.4}
\]
Since the estimator in (A.3) is obtained by minimizing a function with respect to \( \delta \), requires iterative methods. An asymptotically equivalent estimator is given by
\[
\hat{\delta} = \left[ y_2M_xz(z'M_xz)^{-1}z'M_xz_1 \right]^{-1} y_2' M_x z (z'M_xz)^{-1} z'M_x y_1. \tag{A.5}
\]
Then the test statistics is
\[
s(\hat{u}_n) = \hat{u}' M_x z (z'M_xz)^{-1} z'M_x \hat{u}/\hat{\sigma}^2_u, \tag{A.6}
\]
where again \( \hat{\sigma}_n^2 \) is given by (A.4) replacing \( \hat{\delta} \) by \( \hat{\delta}_n \). Test (A.6) is called regression common feature test. It should be noted that the degrees of freedom of the chi-square distribution depends on the number of variables in \( z \) and the number of variables on the right hand side of equations (A.1) or (A.2).

Appendix B. Estimation and Prediction of VAR Models

Let us write model (2) as a regression model of the form
\[
Y = BZ + U, \tag{B.1}
\]
where
\[
Y = (y_1, y_2, \ldots, y_T), (K \times T)
\]
\[
B = (\nu, A_1, A_2, \ldots, A_p), (K \times (Kp + 1))
\]
\[
Z_t = \begin{bmatrix}
1 \\
y_t \\
\vdots \\
y_{t-p+1}
\end{bmatrix}, ((Kp + 1) \times 1)
\]
\[
Z = (Z_0, Z_1, \ldots, Z_{T-1}), ((Kp + 1) \times T)
\]
\[
U = (u_1, u_2, \ldots, u_T), (K \times T).
\]
In addition, assume that \( u = \text{vec}(U) \sim \mathcal{N}(0, I_K \otimes \Sigma_u) \). Then by the transformation theorem the density of \( y = \text{vec}(Y) \) is derived (See Lutkephol, 2005, p. 88). Then for the mean adjuted VAR model of the form
\[
(y_t - \mu) = A_1(y_{t-1} - \mu) + A_2(y_{t-2} - \mu) + \ldots + A_p(y_{t-p} - \mu) + u_t,
\]
the log likelihood function of \( y \) is given by
\[
\ln(l(\mu, \alpha, \Sigma_u)) = -\frac{KT}{2} \ln(2\pi) - \frac{T}{2} \ln(|\Sigma_u|) - \frac{1}{2} \text{tr}[(Y^0 - AX)\Sigma_u^{-1}((Y^0 - AX))],
\]
17
where \( A = (A_1, A_2, \ldots, A_p) \), \( \alpha = \text{vec}(A) \), and \( Y^0 = (y_1 - \mu, y_2 - \mu, \ldots, y_T - \mu) \).

Differentiating \( \ln(l(\mu, \alpha, \Sigma_u)) \) with respect to \( \mu \), \( \alpha \) and \( \Sigma_u \) and equating to zero gives the system of normal equations that can be solved for \( \mu \), \( \alpha \) and \( \Sigma_u \). One advantage of using the the MI estimates are its properties. Under regularity conditions the consistency and asymptotic normality of the estimations has been established (see Lutkephol, 2005, p. 106-109).

The optimal \( h \)-step (linear minimum mean square error (MSE)) predictor is given by

\[
y_t(h) = \nu + A_1 y_t(h-1) + A_2 y_t(h-2) + \cdots + A_p y_t(h-p),
\]

where \( y_t(j) = y_{t+j} \), for \( j \leq 0 \). Replacing the parameters by the estimations of \( \nu, A_1, A_2, \ldots, A_p \) we get

\[
\hat{y}_t(h) = \hat{\nu} + \hat{A}_1 y_t(h-1) + \hat{A}_2 y_t(h-2) + \cdots + \hat{A}_p y_t(h-p).
\]

**Appendix C. Figures and Estimations**

![Graph of monthly variables](image)

**Figure C.4:** The monthly variables \( x_{1t}, \ldots, x_{8t} \) and the river level \( x_{9t} \) from 1994-2004.

\[
\hat{B} = \begin{pmatrix}
0.043 & -1.264 & 1.067 & -0.516 & -0.532 & 1.693 & -1.300 & -0.717 & 1.578 & 0.502 & 0.167 & 0.137 & 0.905 \\
-0.411 & 0.333 & 0.052 & 0.226 & -0.639 & 0.405 & 0.266 & -0.715 & 0.395 & 0.048 & -0.165 & 0.044 & 0.260 \\
0.158 & -0.292 & 0.171 & -0.311 & 0.269 & -0.079 & -0.350 & 0.320 & -0.032 & -0.144 & 0.129 & -0.190 & -0.073 \\
0.223 & -1.054 & 1.603 & -1.002 & -0.120 & 0.966 & -0.112 & 0.000 & -0.059 & -0.083 & 0.572 & -0.569 & 0.176 \\
-0.108 & 0.153 & -0.061 & 0.026 & -0.020 & 0.537 & -0.426 & -0.192 & -0.061 & 0.275 & -0.409 & 0.334 & -0.167 \\
-0.048 & 0.121 & -0.059 & 0.135 & -0.324 & 0.328 & -0.160 & -0.243 & 0.308 & 0.123 & -0.507 & 0.398 & -0.061 \\
-0.045 & 0.079 & -0.029 & 0.101 & -0.032 & 0.061 & -0.238 & -0.053 & 0.207 & -0.110 & -0.108 & 0.004 & 0.154 \\
-0.023 & 0.133 & -0.018 & 0.336 & 0.103 & 0.170 & 0.183 & 0.400 & 0.260 & 0.245 & 0.366 & 0.071 & 0.080
\end{pmatrix}
\]
\[
\hat{A}_1 = \begin{pmatrix}
0.1193 & -0.0490 & 0.5687 & -0.0117 & 0.0138 & 0.0757 & 0.3375 & -0.1971 \\
-0.0636 & 0.4142 & -0.0821 & -0.0048 & 0.0444 & -0.2368 & 0.2506 & 0.1817 \\
-0.0482 & 0.0077 & 0.1783 & 0.0256 & 0.0407 & 0.3865 & -0.1389 & -0.0707 \\
-0.2886 & 0.3970 & 0.8557 & 0.6469 & -0.1640 & 0.0255 & 0.0265 & 0.3113 \\
-0.0742 & 0.1514 & 0.0573 & 0.0312 & 0.4286 & -0.0208 & 0.0154 & 0.0223 \\
0.0153 & -0.0418 & 0.1541 & -0.0362 & 0.0673 & 0.4301 & 0.1832 & -0.0853 \\
0.3750 & -0.3814 & -0.2918 & -0.4086 & -0.2200 & -0.4150 & 0.2612 & -0.3447 \\
0.0536 & -0.0508 & -0.0717 & -0.0759 & -0.0117 & -0.0588 & -0.1648 & -0.0520 \\
\end{pmatrix}
\]

\[
\hat{A}_2 = \begin{pmatrix}
-0.1153 & -0.0573 & 0.6814 & 0.3310 & -0.1247 & 0.5810 & -1.0172 & -0.5240 \\
-0.0035 & -0.0464 & 0.1887 & 0.1421 & 0.0110 & 0.0562 & 0.0566 & -0.1197 \\
0.0649 & -0.2317 & -0.2150 & -0.0841 & -0.0617 & -0.0297 & 0.3990 & 0.0354 \\
0.5481 & -0.2078 & -0.4333 & -0.5202 & -0.0660 & -1.0777 & -0.5539 & -0.4833 \\
0.0484 & 0.0441 & -0.0562 & 0.0024 & -0.3451 & -0.0719 & -0.0810 & -0.1248 \\
-0.0962 & 0.1724 & 0.3298 & 0.2672 & 0.0929 & -0.0614 & -0.1453 & 0.0270 \\
-0.0381 & -0.0017 & 0.0308 & 0.0886 & -0.0460 & 0.0216 & -0.3514 & 0.1339 \\
-0.0539 & -0.0113 & 0.1766 & 0.0092 & 0.1461 & 0.1771 & -0.0343 & -0.2955 \\
\end{pmatrix}
\]

\[\hat{\nu} = (-2526.9896, 2.6298, 1099.6359, 1135.8473, 436.6437, 60.9658, 5.0562, -5328.2510)\]

\[\hat{W}_t = (\hat{w}_{1,t}, \hat{w}_{2,t}, \ldots, \hat{w}_{8,t})', \hat{w}_{i,t} = \sum_{j=0}^{k_i} \hat{C}_{ij} \cos(2\pi \hat{\omega}_{ij} t/n) + \hat{D}_{ij} \cos(2\pi \hat{\omega}_{ij} t/n),\]

\(i = 1, 2, \ldots, 8, k_1 = 4, k_2 = 3, k_3 = 5, k_4 = 4, k_5 = 5, k_6 = 2, k_7 = 3, k_8 = 9\), and the estimates \(\hat{C} = (\hat{C}_{ij})\) and \(\hat{D} = (\hat{D}_{ij})\) are given by

\[
\hat{C} = \begin{pmatrix}
-294.570 & 345.741 & -466.790 & 390.755 & -140.578 & 0 & 0 & 0 & 0 & 0 \\
-55.019 & -74.384 & 101.657 & 111.915 & 0 & 0 & 0 & 0 & 0 & 0 \\
-15.676 & -3.427 & 26.249 & -98.869 & 2.093 & 59.547 & 0 & 0 & 0 & 0 \\
105.020 & 126.189 & 133.709 & 210.392 & 14.397 & 0 & 0 & 0 & 0 & 0 \\
30.210 & 58.451 & 69.298 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
3.359 & -0.594 & -9.335 & 133.282 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{pmatrix}
\]

\[
\hat{D} = \begin{pmatrix}
265.400 & -400.110 & -285.819 & 80.398 & -1268.232 & 0 & 0 & 0 & 0 & 0 \\
-27.678 & 43.985 & -28.649 & -113.818 & 0 & 0 & 0 & 0 & 0 & 0 \\
219.658 & -111.803 & -149.611 & -226.787 & -8.315 & 0 & 0 & 0 & 0 & 0 \\
-72.250 & 48.083 & -12.819 & -3.463 & 45.737 & 190.743 & 0 & 0 & 0 & 0 \\
-58.376 & 58.013 & -205.048 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-40.445 & 43.382 & 36.818 & -12.016 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{pmatrix}
\]

**References**


