Analysing and forecasting fisheries time series: purse seine in Indian Ocean as a case study

Gianpaolo Coro\textsuperscript{1}, Scott Large\textsuperscript{2}, Chiara Magliozi\textsuperscript{1}, and Pasquale Pagano\textsuperscript{1}

\textsuperscript{1}Istituto di Scienza e Tecnologie dell’Informazione “Alessandro Faedo” – CNR, Via Moruzzi 1, 56124 Pisa, Italy

\textsuperscript{2}International Council for the Exploration of the Sea (ICES), H. C. Andersens Boulevard 44-46, 1553 Copenhagen Denmark

Abstract

Catch and fishing effort time series are used by managers to safeguard the availability of resources in the future. Fisheries organisations estimate the status of a stock and the levels for sustainable harvest. Based on these indicators, regulations are developed to guarantee the availability of food and sustain economic growth. For example, tuna stocks are important not only in terms of nutrition, but also for the welfare, culture, revenue and employment of the countries that rely upon them. However, fisheries time-series are the result of many natural and non-natural factors and are usually difficult to predict. Here, we propose a method to aggregate fisheries time series in space and time, to detect hidden periodicities, to predict the time series in the future and to identify the most stressed locations in the fishing area. We apply our method to tuna fisheries data in the Indian Ocean and compare it with respect to other methods. We demonstrate that the method is able to highlight periodic patterns in this high fishing activity area and to forecast the fisheries time series in the future. We use a research e-Infrastructure to comply with modern approaches supporting reproducibility, repeatability and results sharing.

\textsuperscript{*}Corresponding Author: tel: +39 050 315 2978; fax: +39 050 621 3464; e-mail: gianpaolo.coro@isti.cnr.it
Keywords: Time Series Forecasting, Short-Time Fourier Transform, Signal processing, Singular Spectrum Analysis, Fisheries, Indian Ocean, Purse seine, Tuna

1 Introduction

Fisheries time series are important to guarantee availability of new recruits and consequently to sustain human food security and countries economic growth (Farmer and Froeschke, 2015). These data include reports of catch, fishing effort, fish abundance, length, maturity and mortality. They are the most important information used to estimate the biomass of a stock, its status and the associated maximum sustainable yield and are usually compiled by fisheries management organizations. Forecasting fisheries time series is integral for fisheries management, because it allows policy makers to develop strategy and enact management decisions that can achieve goals in light of uncontrollable events (Stergiou and Christou, 1996).

Fisheries management typically uses a suite of methods to quantify stock status. Many stocks are assessed using population dynamics models (Pikitch et al., 2004) built on a simplified mechanistic understanding of a stocks biology and the dynamics of associated fleets (e.g., catch-at-age, length-weight models, etc), informed with survey data quantifying stock and catch statistics (Hilborn et al., 1992). However, these models are complex, require significant expertise to be accurately parametrised (Hinton and Maunder, 2004, Maunder et al., 2006, Erzini et al., 2005, Hsieh et al., 2009, O’Neill, 2005, Butterworth, 2007) and are data hungry (Ward et al., 2014). In data-limited situations or when quantitative mechanistic models are not able to be developed, statistical time-series models are often used (Ward et al., 2014). These methods apply a non-mechanistic framework (Stergiou and Christou, 1996, Froese and Coro, 2014) that is able to provide forecasts of stock status for the management process (Farmer and Froeschke, 2015, Ye et al., 2015).

Generally, forecasting time series of fisheries catch statistics requires modelling all the factors that influenced the catch (Ward et al., 2014). For example, catches can decrease because of overfishing (Shono, 2008) and this influences the catch-per-unit of effort (CPUE), a measure of fish abundance largely used in fisheries management (Maunder et al., 2006).
Further, these time series are usually non-stationary, i.e. their statistical properties change over time, and require powerful techniques to extract information, detect seasonality patterns and to be forecast. Nevertheless, the catch and effort series in a certain year are correlated with the ones in the previous year and this allows applying appropriate techniques from the signal processing domain, e.g. Kalman filters (Gudmundsson, 1994). Indeed, several methods use signal processing techniques to identify changes in a stock size by detecting variability patterns at different spatial and temporal scales. For example, patterns in population time series have been detected using correlation statistics (Waluda et al., 2004), autocorrelation and Lowess smoothing (Collie and Spencer, 1994, Spencer and Collie, 1997). Spectral analysis has been used to estimate indicators of stock exploitation and productivity in the North Sea (Ravier and Fromentin, 2004, Rice and Gislason, 1996, Gislason and Rice, 1998). In particular, spectral analysis has been demonstrated to highlight correlation between the exploitation level of a stock and the slope of the population size spectrum (Bianchi et al., 2000). Frequency analysis has also been used to study overall changes in demersal fish populations structures in several ocean areas (Bianchi et al., 2000). The main limitation of most time series approaches (although counterexamples exist, e.g. Patterson et al. (2001)) is the assumption that the time series is stationary, i.e. its statistical properties (mean, variance etc.) are constant over time. This influences the correct estimation of the spectral characteristics of the time series and of its evolution in time (Hsieh et al., 2009). Few examples of fisheries statistics forecasting are available (Ménard et al., 2007, Fablet and Le Josse, 2005), e.g. predictive models that use growth equations, climatic variables, life history traits and/or estimations of population dynamics to project a catch series in the future (Stergiou et al., 1997, Yamanaka, 1988, Stergiou, 1991, Hollowed et al., 2011). These models usually do not take into account other non-ecological factors that could modify a catch trend, such as piracy activity, overfishing and regulations, which are “hidden” in the structure of the time series.

In this context, we focus on tuna fishing activities. Tuna are important not only in terms of nutrition, but also for the welfare, culture, revenue and employment of the countries that rely upon them (Gillett, 1997). In several areas (e.g. the Pacific and Indian Oceans), tuna
fisheries produce several times the incomes of all other fisheries combined (Lymer et al., 2008, Gillett, 1997). Thus, regulating tuna fisheries is crucial to safeguard the many sectors depending on them (Gillett, 1997).

In this paper, we particularly focus on data of tuna catch publicly distributed by the Indian Ocean Tuna Commission (IOTC) that include time series of catch, effort and locations. This intergovernmental fisheries management organisation is responsible for the management of tuna and tuna-like species in the Indian Ocean. IOTC periodically publishes monthly catch statistics (IOTC, 2015), in order to report about the exploitation of commercial stocks like tuna. IOTC principal aims are to avoid stock depletion and to inform fisheries regulatory international organisations (e.g. the Food and Agriculture Organisation of the United Nations). For these data, we propose a novel linkage of two techniques common in other fields, Singular Spectrum Analysis (SSA) and Short-Time Fourier Transform, to aggregate statistics on the exploitation of a stock and to spatially predict discrete catch per unit of area. Our method spatially aggregates multiple streams of data (e.g., catch reports, fishing effort, or fishing activity) and this aggregation for an area produces a single time series of the barycentres, or coordinates of the most exploited 0.5 degree square areas of the fishing effort. Additionally, the method forecasts these values into the future. To aid users who are not experts in signal processing and forecasting, we automated the parametrisation of this approach and published it as-a-service through a research e-Infrastructure endowed with cloud computing.

2 Materials and methods

In this section, a general overview of our method is given and then the details of each step of the process are explained. All the forecasting and analysis techniques referred in this section, are described in the Appendix. An overall view of our process to analyse fisheries time series is depicted in Figure 1. In summary, our method produces the following information: (i) estimation of the time series of catch, effort, CPUE, longitude and latitude in the most exploited locations, (ii) aggregated statistics on the exploitation of a stock, (iii) highlight of periodic segments in the time series of the catch, fishing effort, catch-per-unit
of effort (CPUE) and coordinates of the most exploited 0.5° square areas, (iv) forecast of all the fisheries time series in the future (in the next year by default).

2.1 Data

Our method requires the input dataset (a table) to contain at least information on: longitude (decimal degrees), latitude (decimal degrees), fishing hours and catch quantity. This information is common in the fishing activity records produced by fisheries management organizations during a certain time frame, usually reporting the quantity of caught fish (and stocks), the effort spent fishing (in hours) and the fishing locations. Our method also adds catch-per-unit of effort (CPUE) information as the ratio between catch and fishing hours, if not already present in the input.

To evaluate the effectiveness of our approach to predict fishing effort and catch and detect inner periodicities, we used IOTC reports of purse seine fishing of yellowfin (Thunnus albacares) and skipjack (Katsuwonus pelamis) tuna. This data is publicly available for downloading and consultation on the IOTC website (IOTC, 2015). Fishing hours and the locations of purse seine fishing activity between 1983 and 2013 are reported for both species in Figure 2.

The choice of using IOTC data is due to the fact that this kind of data is of interest to international organisations that monitor marine and food resources availability as well as fishery activity (e.g. the Food and Agriculture Organization of the United Nations - FAO). These organisations may rely on data provided by Regional Fisheries Management Organisations (RFMOs), international organisations formed by countries with fishing interests in an area (e.g. IOTC), to assess the status of the stocks and plan conservation strategies. Thus, we assumed that presenting a direct application of our method to IOTC tuna data would have been interesting in this context.

1Also available, in a simplified version, on the D4Science e-Infrastructure: http://data.d4science.org/uri-resolver/id?fileName=PURSE_SEINE_YFT_-_1983_-_2012.csv&mp-id=564a4402e4b094f1f0348bb0&contentType=text%2Fcsv
2.2 Time series aggregation

Since catch reports overlap in time and space, as first step, our method aggregates input data by 0.5 degree squares in the fishing area. Aggregation is made by summing time-overlapping reports at 0.5° spatial resolution. At the end of this process, a new dataset is produced, reporting information for each 0.5° cell in time. In the rest of the paper, we will name “fishing dimensions” the numerical information attached to the cells, i.e. the coordinates of the cell, the catch effort, the fishing hours, the CPUE and the time.

At this stage, there will be several 0.5° locations associated to each reporting time \( t \). Our process simplifies this scenario in order to work with one time series for each dimension. To such aim, for each time \( t \) it sums the values of the fishing dimensions (separately) associated to the 0.5° cells. Furthermore, it takes the barycentre of the fishing effort as representative of all the fishing locations at time \( t \). This barycentre is calculated by weighting each 0.5° cell for the number of fishing hours spent by the vessels in the cell; this aims at having the coordinates of the barycentre close to the highest fishing effort locations, i.e. to the locations where the vessels spent most of the time. In summary, calculating the fishing effort barycentre for each time \( t \), allows obtaining a sequence of 0.5° cells that represent the highest fishing effort zones. For each fishing dimension, the sum of the values at time \( t \) is allocated to the barycentre at time \( t \). In this way, one time series for each dimension is created. The barycentres allow detecting the most exploited zones and visualising general spatial trends of fishing activity (Section 3).

The coordinates of the fishing barycentre at time \( t \) are:

\[
\text{BarycenterLatitude}_t = \frac{\sum_{i=1}^{n} \text{latitude}_{i,t} \times \text{effort}_{i,t}}{\sum_{i=1}^{n} \text{effort}_{i,t}}
\]
\[
\text{BarycenterLongitude}_t = \frac{\sum_{i=1}^{n} \text{longitude}_{i,t} \times \text{effort}_{i,t}}{\sum_{i=1}^{n} \text{effort}_{i,t}}
\]

where \( n \) is the number of 0.5° cells reporting fishing activity at time \( t \); \( \text{latitude}_{i,t} \), \( \text{longitude}_{i,t} \) and \( \text{effort}_{i,t} \) are the coordinates and the fishing effort of the \( i \)-th involved 0.5° cell at time \( t \).
The output of this processing step is a dataset containing the summed time series of catch, fishing hours, CPUE and the time series of the barycentres coordinates. As final step, these statistics are possibly time-aggregated (by further summing the values) daily, monthly and annually, depending on the time granularity of the original data set.

2.3 Periodic patterns detection

Catch, effort, and CPUE appear non-stationary but can present periodic sections (Figures 3 and 4). Therefore, as explained in the Appendix, Short-Time Fourier Transform (STFT, Allen (1997)) can be used to identify hidden periodicities. STFT is applied to the signals and the shift of the sliding analysis window is half of the window length. In the end, the process reports the found periodicities, along with the strength of their associated power spectrum (i.e. the summed squares of the Fourier Transform values). In Signal Processing, a value of a time series is named sample, whereas the sampling frequency is the number of samples per second. This terminology will be used throughout the description of our method. In our method, if the distance between two samples is not constant (non-uniform sampling), the time series is uniformly sampled before applying STFT in order to have the same time distance between two points. This step uses the time series missing values replenishment algorithm of the Weka framework (Holmes et al., 1994). The algorithm accounts for missing values in the series using linear interpolation, which limits the applicability of this method to time series with short gaps. The sampling period is selected as the minimum time distance between two subsequent samples. One example of output of this step of our method is reported in Table 1, where the normalised power spectrum $P$ is categorised either as “High” ($P > 0.6$) or “Moderate” ($0.5 < P \leq 0.6$). Also “Weak” ($0.3 < P \leq 0.5$) and “None” ($P \leq 0.3$) categories are possible, even if not reported in the table. The method also produces the spectrogram of the signal (Appendix).

Since it cannot make prior assumptions about the periodicities of a signal, our method iteratively applies STFT using all the possible windows lengths divisible by two (in order to maximise the performance of STFT). Our process uses cloud computing to execute these processes fast and concurrently (Section 2.6). In the end, our method selects only the peri-
odicities with “Moderate” and “High” strength and reports the corresponding time frames. Uncertainty is associated to each periodicity based on the length of the sliding window.

In summary, this step detects periodicities by means of a multi-time analysis. The aim is to provide indications to fisheries scientists about time ranges in which periodic phenomena (e.g. monsoons, seasons change etc.) could have influenced the fisheries.

2.4 Time series forecasting

Catch statistics often contain periodicities, because fishing modality and locations change according to the season or to the abundance of a stock. Singular Spectrum Analysis (SSA) is a forecasting technique that is generally more accurate on signals containing periodicities, because the periodic portions contain hidden structures that can be captured by the SSA eigenvectors (Appendix). Hidden phenomena behind the time series could be somehow related to the SSA eigenvectors, although it is not easy to assess a direct association. However, the effectiveness of the algorithm depends on the selected number of eigenvectors and on the length of the chunking window (Appendix).

The forecasting step of our method requires two input parameters: the length of the chunking window (named “SSA samples”) and a threshold for the number of hidden components to consider (named “eigenvalues threshold”). Our method applies SSA several times, using different combinations of these two parameters. In particular, the method increases the chunking window length of one unit per each experimental run, and for each of these steps it executes SSA using thresholds of 30th, 40th, 50th and 60th percentiles to filter the hidden components. These percentiles were selected after analysing the performance of SSA on a large public repository of time series of catch statistics (Food and Agriculture Organization of the United Nations, 2015). Thus, if the signal has $N$ points, the method performs $N \times 4$ analyses (one for each percentile). Each analysis produces a forecast, but only one solution is needed, thus this large space of solutions needs to be filtered to choose the best one. To this aim, our algorithm is calibrated on the time series of the latest year. In particular, it takes the latest year time-series as reference and estimates its basic statistical properties: variance, mean, maximum and minimum values. For exam-
ple, if a time series contains monthly reports and ranges from 2010 to 2012, the algorithm considers only the statistical properties of the monthly reports of 2012. At this point, the method finds the forecast time series having statistical properties that are most similar to the ones of the latest year. The rationale behind this choice is that fishing activity usually does not change abruptly from one year to another, unless population collapses or regulations suddenly change (Stergiou and Christou, 1996).

The algorithm assesses the similarity between the last year time series and the SSA forecasts by using basic similarity measures (Ding et al., 2008). In particular, similarity is calculated based on the relative differences of the statistical properties: two time series are assumed to be “similar” if the relative difference of each statistical property is less than 50%. This non-strict empirical threshold accounts for moderate discrepancy between two time series. Additionally, the algorithm calculates an “overall similarity” as the product of the relative differences; high “overall similarity” corresponds to low product values and exact match corresponds to 0%. Based on this measurement, the best SSA forecast is selected as the forecast with the highest “overall similarity” with respect to the latest year time series.

2.5 Comparison with other forecasting approaches

The forecast produced by our approach was evaluated with respect to the forecasts produced by ARIMA, Exponential Smoothing and Artificial Neural Networks (described in the Appendix). These models were selected because they have demonstrated high accuracy in different domains, e.g. fisheries (Stock and Watson, 1998), conservation biology (Carroll and Pearson, 2000) and financial analysis (De Gooijer and Hyndman, 2006). One major distinction between them, is that ARIMA and Exponential Smoothing are parametric models, whereas Artificial Neural Networks and SSA are non-parametric models. The difference between these two classes of models has been extensively studied and their accuracy may vary depending on the application field. For example, although ARIMA has demonstrated high performance in population abundance forecasting (Ward et al., 2014), it has shown worse performance than non-parametric models on catch time series (Stergiou and Christou, 1996). One major issue of fisheries time series is that they reflect a complex
combination of factors like population abundance, market prices, and behaviour of fishers and may require complex modelling of the correlation between the time series values (Ward et al., 2014). This is the case of Artificial Neural Networks, which realise a complex combination of the values but are not tied directly to the last observation (Thrush et al., 2008). Thus, these models have demonstrated higher performance on rainfall prediction (Toth et al., 2000) than on species population prediction (Ward et al., 2014), although their performance in this case increases when periodic phenomena are present.

Generally, assuming a linear dependency between the values of a fisheries time series is not sufficient, in fact performance increases as soon as non-linear combinations are used. Yearly catch data are not characterised by strong autocorrelation especially in longer-term trends (Stergiou and Christou, 1996), which is the case of macroeconomic time series (Stock and Watson, 1998). In this complex scenario, SSA shows higher performance than parametric approaches, for both short and long time series (Silva and Hassani, 2015), due to indirect modelling of the dependency between the values and noise filtering features (Zokaei et al., 2001). For these characteristics, SSA has shown to be more robust to abrupt changes than ARIMA, Exponential Smoothing and Neural Networks, e.g. in predicting the United States import/export time series after the recession (Silva and Hassani, 2015).

SSA was included in our method for fisheries analysis because of the above features. A general report of the performances of several parametric and non-parametric forecasting models on fisheries time series is left to Stergiou and Christou (1996). Instead, in this section the accuracies of the forecasts of the IOTC data are compared. In this comparison, only techniques that correlate the time series of a certain year with the one of the previous years were taken into account, because this dependency is crucial to produce a good forecast of the kind of time series our method manages (Stergiou and Christou, 1996). Thus, methods like Generalised Additive Models (GAMs) were excluded because they are less sensitive to recent trends (Farmer and Froeschke, 2015). Further, surplus-yield models were discarded too, because they have demonstrated to be less suited for catch series forecasting (Stergiou and Christou, 1996).

The barycentres of the fishing effort signal in 2012 of the IOTC dataset described in
Section 2.1 were used to calibrate the models, whereas the 2013 data were used to compare the models forecasts. Thus, our SSA-based forecasting process used the 2012 data statistical properties as reference. As implementation of ARIMA and Exponential Smoothing, the “forecast” R package (Hyndman and Khandakar, 2007) was used to train the models on the 2012 data. The software automatically estimated ARIMA(1,1,2) as the best ARIMA model and suggested a 20 months window length for Exponential Smoothing. As for Artificial Neural Networks, the “nnet” R package (Ripley and Venables, 2015) was used and a Feed-Forward Neural Network was trained on 2012 data using a growing strategy (Bryson et al., 1979) to detect the best number of layers and and neurons. The best topology had 20 input nodes and 2 hidden layers with 20 neurons in each layer.

Similarity indicators for each forecast are reported in Table 2 and the forecasts are visually compared in Figure 5. Apart from standard measurements (R square and mean error, (Farmer and Froeschke, 2015)), a number of indicators were added to appreciate the performance of the SSA forecast with respect to the other techniques. In particular, Table 2 reports (i) the relative variation of the statistical properties of the forecasts with respect to the true signal (i.e. the 2013 observations), (ii) the average logarithm of the absolute scaled error (ASE, Hyndman and Koehler (2006)), (iii) the Bayesian information criterion (BIC, Kass and Wasserman (1995)) (considering the approaches as models fitting the same data with different parametrizations) and (iv) the mean squared error. It is notable that our SSA-based method is the best punctual approximation of the true signal, whereas ARIMA is the one that best approximates the true signal mean. This is important information when high accuracy of the prediction is required, for example to precisely identify possible high fishing pressure locations.

Further, an important overall difference in the quality of the forecasts is given by the ASE values. Since ARIMA(1,1,2) is a random-walk process, lower ASE for a non-random walk model (e.g. SSA) indicates that there is structure in the data (Ward et al., 2014). ASE is often used as a measure of the overall forecast quality, because higher ASE with respect to a random-walk process indicates either data over-fitting or low-accuracy (Ward et al., 2014). The selection criterion here proposed for SSA, generates the only forecast having
lower ASE than ARIMA. This is an indicator of the overall good quality of our produced forecast. Finally, since ARIMA(1,1,2) was estimated to be the best ARIMA model, this indicates that there is correlation between adjacent values (because the $p$ and $d$ parameters are equal to 1). This correlation may be due to inner 1-2 years periodicities in the fishing effort (Stergiou and Christou, 1996), which is compliant with the indications produced by the Fourier analysis reported in the next section.

### 2.6 Software availability

Our method is open-source and available as a JAVA program\(^2\). It is part of the gCube open-source framework (Coro and Candela, 2014, Coro, 2015a). Our implementation automatically performs SSA and STFT analyses using the described parameters combinations, and automatically selects the best set of parameters. It is also published as-a-service on the D4Science computational e-Infrastructure (Candela et al., 2009, Coro and Candela, 2014, Coro et al., 2014, Coro, 2015c,b)\(^3\). D4Science is a research e-Infrastructure endowed with cloud computing facilities to process data, which also speeds up our parameters selection process. Additionally, it stores the computational output on a high-availability distributed storage system and allows retrieving the history of the executed experiments, along with a summary of the used parameters and of the output. It also allows sharing parameters and results with colleagues and thus makes experiments repeatable.

### 3 Results

In this section, we report the application of our method to IOTC tuna fisheries data and we demonstrate that the method helps understanding how fishing activity in a certain area has changed in time.

\(^2\)On a public SVN repository: https://svn.research-infrastructures.eu/public/d4science/gcube/trunk/data-analysis/FisheriesTimeSeriesAnalysis/

\(^3\)A Web interface is available at https://i-marine.d4science.org/group/biodiversitylab/processing-tools
3.1 Estimation of the barycentres

We applied the set of techniques constituting our method to the IOTC dataset described in Section 2.1 (Figure 2), in the time frame between 1983 and 2012 with monthly aggregation, whereas we withhold the 2013 data to test the accuracy of the forecast for 2013. The time series aggregation process extracted the barycentres of the fishing effort in time. The spatial distribution of these locations, when time is neglected, is depicted in Figure 6. As expected, locations are concentrated in the region of highest effort. The spatial distribution of the monthly barycentres in time are compared in Figure 9 with respect to the overall distribution of the real fishing locations. It is notable that the fishing activity area moves eastward and is more widespread in 2012 with respect to 1983. However, in 1983 the barycentres were all concentrated in a narrow area and fishing effort was lower (more yellow points) in each month, whereas in 2012 high effort was spent along a horizontal line. The monthly time series of the dimensions are depicted in Figures 3 and 4 and the annual ones are in Figures 7 and 8. Since the native aggregation of the IOTC dataset is per month, daily series were not calculated. In the charts, the “c-square” (or “Csquare”) index is reported to give one single representation of longitude and latitude pairs: a c-square code (concise spatial query and representation system, Rees (2003)) is a geographical code that provides a simple spatial indexing of geographic square areas. The notation system represents the latitude, the longitude and the size of a square traced around a coordinate pair. The length of the code depends on the square size. For example, a 0.5°x 0.5° cell in the range -0.5° W to 0° W, 51.5° N to 52.0° N has code 7500:110:3, whereas if the size is set to 1°x 1° the code becomes 7500:110.

The c-square view allows having a compact representation and understanding how often exploitation involves the same locations. In order to build the representations in Figure 4 and 8, the distinct 0.5° c-squares of the barycentres in time were enumerated, assigning an increasing integer number to each distinct c-square. In the enumeration process, equal c-squares in the sequence had the same integer number assigned. These numbers are reported in the figures with respect to time, and the downwards peaks indicate that a c-square is equal to a previous one. In other words, downwards peaks highlight c-square repetitions.
As a result, the annual time series always presents new barycentres, whereas the monthly
time series presents repetitions. The overall increasing trend of the index agrees with the
locations dispersion shown in Figure 9.

It is also notable that the annual time series aggregation hides most of the periodic prop-
erties of the signals. For example, it flattens the longitude time series, whereas it highlights
that latitude strongly changes after 1996, which could be an effect of the El Niño-Southern
Oscillation.

3.2 Analysis of the most exploited locations

Based on the barycentres time series, Figure 10 highlights the c-square locations having
highest fishing activity, CPUE and catch at 1° aggregation. The most exploited location
(1005:206) is the same for all the dimensions, whereas the lower exploitation locations
are different among the dimensions. This means that in these locations high fishing effort
does not correspond to high catch. To further explore this aspect, the time series of each
dimension was analysed in the locations with highest fishing effort (Figure 11). It is notable
that in each dimension the highest effort locations are mostly present up to 1995 and show
continuous segments concentrated in few years.

3.3 Periodicities detection

Short-Time Fourier analysis was applied to all the time series with monthly granularity,
in order to detect periodic patterns in the fishing dimensions (Table 1). The (uniformly
sampled) signals had 360 values (samples), thus STFT was applied using window lengths
ranging up to 256 samples, with a slide of the window equal to half the window length.
In particular, the following windows were used: 16 samples (~1 year), 32 samples (~2
years and half), 64 samples (~4 years and 9 months), 128 samples (~9 years and half)
and 256 samples (~19 years and half). The different window lengths emphasize different
periodic components. Fishing dimensions signals mostly have periodicities of ~3 months (4
samples), but also present periodicities of ~4 months (5 samples), ~6 months (8 samples),
~9 months (10 samples) and ~11 months (12 samples). Fishing effort, catch and CPUE
have also periodic components of $\sim 60$ months (64 samples). Latitude and longitude present
eriodicities between $\sim 3$ and $\sim 9$ months in several time frames. One example of Short-Time
Fourier analysis applied to the fishing effort time series is reported in the Appendix.

Although it is difficult to give precise interpretation to all the periodicities reported in
Table 1, several known properties of fisheries time series can be recognized. Short period-
icities in fisheries time series may be related to short-term ocean-atmosphere interactions
(e.g. surface heat-exchange phenomena) (Stergiou and Christou, 1996, Zupanovic, 1968,
Colebrook and Taylor, 1984, Kort, 1970). In particular, periods of 4-6 years highlighted
by the 64 samples window have been indicated for other fisheries time series as depending
on environmental parameters variation, e.g. air temperature (Stergiou and Pollard, 1994)
or marine populations variations (Kort, 1970, Shuntov et al., 1981, Colebrook and Taylor,
1984). Further, the 10-11 years periodicity highlighted by the 128 samples window may be
related to cycles in the sunspots number (Love and Westphal, 1981, Stergiou and Christou,
1996, Vasilkov et al., 1981, Regner and Gačić, 1974, Hathaway et al., 2002). Finally, pe-
eriodicities of 1-2 years are consistent with the forecasts reported in Section 2.5. However,
it is not straightforward to correlate the above phenomena with the periodicities found in
the IOTC data, because fluctuations may depend on other factors than abundance or climate
change, e.g. change in regulations or variation of the fish price.

The El Niño-Southern Oscillation (ENSO) can be an important cause of interannual
variability of climate and tuna catch. It is generally related to short-term periodicities
(Mysak, 1986). In particular, ENSO has been associated with irregular 4-5 year periodicities
in the tuna catch and climate time series (Ménard et al., 2007), but also with 30-60 months
periodic phenomena (Yasunari, 1987). Further, 3-5 years periodicities have been associated
to the global propagating wave of ENSO in the Indian Ocean (White and Cayan, 2000) that
consequently influence tuna catch (Murtugudde et al., 2000, Lehodey et al., 1997).

### 3.4 Forecasting

The SSA-based time series forecasting method of Section 2.4 was applied to all the monthly
fishing dimensions signals. The process found the SSA parameters combination that max-
imised the similarity with respect to the monthly time series of 2012. Forecasts of the fishing effort using several parameters combinations are reported in Figure 12, compared with the true time series of 2013. The relative similarity scores associated to the forecasts displayed in Figure 12 are reported in Table 3. The scores highlight that the SSA using a 126 samples window and a 0.07 eigenvalues threshold is the best according to our selection criterion. The analysis of all the fishing dimensions time series is reported in Table 4. This table shows the similarity of the best forecasts with respect to both the 2012 and the 2013 time series. For sake of completeness, the table also reports several standard similarity measurements: mean squared error (MSE) (Owen et al., 2008), Bayesian information criterion (BIC) (Kass and Wasserman, 1995), R-squared measure of goodness of fit (R sqr.) (Cheung and Rensvold, 2002) and mean error (Armstrong, 2001).

In Figure 13, the forecasts (in blue) and the time series (in red) of each dimension are displayed. On the right-hand side, the forecast is highlighted and compared with the true 2013 time series. The charts highlight that the forecast of the fishing effort is the most similar to the real signal, whereas the other forecasts show similar trends (i.e. they agree on the first derivative), sometimes coinciding with the real values (e.g. in the case of latitude and longitude). According to Table 4, especially looking at the R-squared values, the quality of the forecasts is overall very good. One particular case is the longitude series, which shows good agreement at the beginning of 2013 but only trend agreement in the end. This may indicate that a number of unknown factors have suddenly influenced and changed the fishing effort barycentre longitudes either in 2012 or in 2013.

4 Discussion and conclusions

In this paper, we have reported an approach to forecast spatially-aggregated fisheries time series and to discover their hidden periodicities. By applying our method to purse seine fisheries data between 1983 and 2013 published by the Indian Ocean Tuna Commission fisheries management organizations, we have highlighted general patterns in the locations that were most exploited by the fisheries. Furthermore, our method detected long and short periodicities hidden in the fisheries time series. Finally, the process was able to produce a
good quality forecast of the time series in 2013 based on data up to 2012, as demonstrated by
a number of statistical similarity measurements with respect to the real data. We have shown
a complete example of application of our method to official reports from the main Indian
Ocean fisheries management organisation (IOTC), using several visual representations of
the output to highlight properties and movements of the fishing activity. We have also
reported the properties of our forecasting method and compared them with the ones of other
state-of-the-art approaches.

The reported use case is limited to Indian Ocean purse seine fisheries and it is difficult
to infer general conclusions based on it. Overall, fisheries time series refer to differently
aggregated nationwide data. Thus, a generally applicable method should manage multiple
temporal and spatial scales. General issues with time series (also valid for the fisheries and
biological ones), are that (i) the future may not be described by the past series, (ii) there are
too many variables to take into account that may introduce noise or that are unknown, (iii)
the time series can be biased (Goodrich, 1992). Thus, the complexity of the fisheries time
series requires using powerful techniques that have shown high accuracy in other domains.
Factors like micro- and macro-economics, climate change etc. could increase forecast ac-
curacy if explicitly modelled (Dement’Eva, 1987). Since this operation can be hard, models
should try to automatically discover these hidden phenomena (Stergiou and Christou, 1996).
However, abrupt variations from one year to another are rare in fisheries time series, e.g.
catches usually tend to remain high for few successive years if the environmental condi-
tions, the fish population and the law regulations do not change very much (Stergiou and
Christou, 1996).

We designed our method to meet the above data-related requirements. By default con-
figuration, our method applies a 0.5° x 0.5° geospatial aggregation. However, it can be
configured to work with other aggregations because the method workflow remains valid.
Further, it can be applied to time granularities other than monthly. In fact, daily and annual
time series are automatically managed by the aggregation process when available (Sec-
tion 2.2), and SSA and STFT are theoretically applicable also to these time scale, without
changing the overall method workflow. Support to this last statement comes from general
assessments made by other studied on the SSA forecasting properties (Section 2.5). Further, the forecast can be extended for more than one year as far as the latest year for which data are available is reasonably influential on the forecast (Section 2.4).

Thus, in principle our method could be suited for time series having properties of non-stationarity, no abrupt changes, short gaps and long/short-spanning hidden periodicities, e.g. many biological time series (Stergiou et al., 1997). It could be directly used on more recent IOTC data and generally to the information produced and distributed by most of the Regional Fisheries Management Organisations for their areas of competence. For example, the produced information could be useful to predict the seasonal movements of fishing vessels, to study their response to external natural or non-natural phenomena (e.g. the El Niño-Southern Oscillation) and thus to build advices for best practices in the future. Our method could be also tested in the context of fishing pressure trends prediction, most exploited areas detection and stock yield reduction prevention.

Finally, the novelty of our method is also that the processes are published by a research e-Infrastructure (Coro, 2015c, b). This instance offers cloud computing to execute the large amount of processing required by the periodicities detection and forecasting steps in our method. Without these facilities, finding the best parameters for the analysis would have been much more time consuming and difficult. The e-Infrastructure also offers sharing facilities, recording of experimental configuration history and output storage. Thus, our method allows for experiments repeatability and reproducibility, in compliance with new Science 2.0 requirements (Waldrop, 2008). One of the advantages of our method is that it relies on these technologies to combine complex techniques coming from different domains, perform large calculations, reduce the complexity of the experimental setup and publish the results.

Supplementary material

The following supplementary material is available at ICESJMS online: Appendix containing a report of the used time series forecasting and analysis techniques.
Acknowledgments

The reported work has been partially supported by the BlueBRIDGE project (H2020 framework of the European Commission, H2020-EINFRA-2015-1, grant agreement No. 675680).

References


ing measurement invariance”, *Structural equation modeling*, Vol. 9, Taylor & Francis, pp. 233–255.


fish assemblages to changes in exploitation”, 


Hyndman, R. J. and Khandakar, Y. (2007), Automatic time series for forecasting: the fore-


O’Neill, M. F. (2005), *Reference point management and the role of catch-per-unit effort in prawn and scallop fisheries*, Department of Primary Industries and Fisheries and Fisheries Research and Development Corporation.


White, W. B. and Cayan, D. R. (2000), “A global el niño-southern oscillation wave in surface temperature and pressure and its interdecadal modulation from 1900 to 1997”,


**URL:** http://2011.isiproceedings.org/index.php

Figure 1: Overview of our method:

1. Producing
   - long. lat.
   - Fishing hours
   - catch
   - CPUE

2. Building time series

3. Detecting periodicities

4. Forecasting

Figure 1: Overview of our method: 1. provide an input dataset in the form of a table, containing space and time reports of catch quantity, fishing effort (in hours) for a certain stock in a fishing area. Build a CPUE time series by dividing catch by fishing effort; 2. apply a spatio-temporal aggregation process: produce a time series of the fishing locations barycentres by weighting locations with respect to fishing effort. Report the overall quantity in time of each fishing dimension to analyse (catch, fishing effort, CPUE, barycentres latitude and longitude). Aggregate the time series according to several time granularities, if possible: daily, monthly, annually; 3. detect periodicities in each dimension time series and for each time aggregation, using a multi-windowed Short-Time Fourier Transform. Report only the periodic time frames having “Moderate” or “High” strength; 4. Forecast each time series, using several parametrisations of the SSA process, reporting the forecast that is most similar to the time series of the latest reporting year.
Figure 2: Distribution of purse seine fishing effort (panel a; in hours per 0.5 degree square) in the Indian Ocean between 1983 and 2012, where high (red; emphasized in panel 2e), moderate (orange; emphasized in panel 1d) and less exploited (yellow; emphasized in panel 1c) areas overlap. Squares represent 0.5° squares to recognize the uncertainty present on the reported locations. Bounding boxes identify the regions magnified in the other figures, which depict the effort ranges separately. Higher effort locations are concentrated in region 2, whereas lower effort is distributed across the whole Indian Ocean.
Figure 3: Monthly time series of the overall fishing effort, catch and CPUE from IOTC data.
Figure 4: Monthly time series of the latitudes and longitudes of the fishing effort barycentres of IOTC data, produced by the first step of our method. The “Csquare Index” is an enumeration of the 0.5° c-square areas involved in the fishing activity. The repetition of these indexes in the chart indicates that the barycentres repeat in time and indicate that heavy fishing effort occurs at the same place over time.
Figure 5: Comparison of different forecasting approaches for Indian Ocean monthly fishing effort in 2013.

Figure 6: Distribution of the 0.5° cells of the barycentres of fishing effort in the Indian Ocean between 1983 and 2012. Red colour refers to highest fishing hours. The colours are separated to better visualise the distribution of the pressure. Panel a displays the overall effort, which is contained in region 2. Panel 1c and 1d display the spatial distribution of low and medium effort locations in region 1. Panel 2e reports the highest effort locations, which span region 2.
Figure 7: Annual time series of the overall fishing hours, catch and CPUE from IOTC data.
Figure 8: Annual time series of the longitudes and latitudes of the fishing effort barycentres of IOTC data. The “Csquare Index” is an enumeration of the 0.5° c-square areas involved in the fishing activity. The low repetition rate of these indexes in the chart indicates that the annual barycentres almost never repeat in time.
Figure 9: Two sequences of half-degree fishing locations in Indian Ocean between 1983 and 2012: the upper sequence reports the overall real monthly fishing locations; the lower sequence reports the distribution of the monthly fishing effort barycentres. The colour scales are the same for all the images in the sequences.
Figure 10: Histograms of the quantities associated to the most exploited fishing locations aggregated at 1°.
Figure 11: Time series of the most exploited fishing effort locations aggregated at 1°. The charts display the quantities associated in time to these locations in terms of (i) fishing effort, (ii) catch and (iii) CPUE. Lines connect monthly-adjacent points.
Figure 12: Comparison between different forecasts of fishing activity in the Indian Ocean in 2013 using Singular Spectrum Analysis (SSA) with different values of the number of “SSA samples” and the eigenvalues threshold (“Eigen. Thr”).
Figure 13: Forecast of Indian Ocean fisheries time series by our method. The right-hand side charts compare the 2013 real time series with the forecast. The left-hand side charts report both the complete time series up to 2012 and the 2013 forecast.
### Fourier Analysis with 16 samples window (1 year)

<table>
<thead>
<tr>
<th>Time</th>
<th>Fishing Effort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start Time</td>
<td>End Time</td>
</tr>
<tr>
<td>Apr-08</td>
<td>Jul-09</td>
</tr>
<tr>
<td>May-09</td>
<td>Aug-02</td>
</tr>
<tr>
<td>Nov-03</td>
<td>Feb-06</td>
</tr>
<tr>
<td>Feb-06</td>
<td>May-09</td>
</tr>
<tr>
<td>Mar-11</td>
<td>Jun-12</td>
</tr>
</tbody>
</table>

### Fourier Analysis with 32 samples window (2 years and half)

<table>
<thead>
<tr>
<th>Time</th>
<th>Fishing Effort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start Time</td>
<td>End Time</td>
</tr>
<tr>
<td>Sep-86</td>
<td>Feb-89</td>
</tr>
<tr>
<td>Sep-86</td>
<td>Feb-89</td>
</tr>
<tr>
<td>Nov-87</td>
<td>May-09</td>
</tr>
<tr>
<td>Aug-95</td>
<td>Nov-03</td>
</tr>
<tr>
<td>Mar-03</td>
<td>Aug-02</td>
</tr>
<tr>
<td>Nov-03</td>
<td>Feb-06</td>
</tr>
<tr>
<td>Feb-06</td>
<td>May-09</td>
</tr>
<tr>
<td>Mar-11</td>
<td>Jun-12</td>
</tr>
</tbody>
</table>

### Fourier Analysis with 64 samples window (4 years and 9 months)

<table>
<thead>
<tr>
<th>Time</th>
<th>Fishing Effort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start Time</td>
<td>End Time</td>
</tr>
<tr>
<td>Oct-92</td>
<td>Jun-12</td>
</tr>
<tr>
<td>Oct-92</td>
<td>Jun-12</td>
</tr>
</tbody>
</table>

### Fourier Analysis with 128 samples window (9 years and half)

<table>
<thead>
<tr>
<th>Time</th>
<th>Fishing Effort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start Time</td>
<td>End Time</td>
</tr>
<tr>
<td>Oct-92</td>
<td>Jun-12</td>
</tr>
<tr>
<td>Oct-92</td>
<td>Jun-12</td>
</tr>
</tbody>
</table>

### Fourier Analysis with 256 samples window (19 years and half)

<table>
<thead>
<tr>
<th>Time</th>
<th>Fishing Effort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start Time</td>
<td>End Time</td>
</tr>
<tr>
<td>Oct-92</td>
<td>Jun-12</td>
</tr>
<tr>
<td>Oct-92</td>
<td>Jun-12</td>
</tr>
</tbody>
</table>

### Table 1: Report of all the periodic components found by the Short-Time Fourier Transform in the fishing dimensions of the 1983-2012 IOTC reports.

<p>| Table 1: Report of all the periodic components found by the Short-Time Fourier Transform in the fishing dimensions of the 1983-2012 IOTC reports. |
|---|---|---|---|---|---|---|
| <strong>Catch</strong> | <strong>Latitude</strong> | <strong>Longitude</strong> |
| <strong>Start Time</strong> | <strong>End Time</strong> | <strong>Start Sample</strong> | <strong>End Sample</strong> | <strong>Period in samples</strong> | <strong>Uncertainty</strong> | <strong>Strength</strong> |
| <strong>Start Time</strong> | <strong>End Time</strong> | <strong>Start Sample</strong> | <strong>End Sample</strong> | <strong>Period in samples</strong> | <strong>Uncertainty</strong> | <strong>Strength</strong> |
| <strong>Start Time</strong> | <strong>End Time</strong> | <strong>Start Sample</strong> | <strong>End Sample</strong> | <strong>Period in samples</strong> | <strong>Uncertainty</strong> | <strong>Strength</strong> |</p>
<table>
<thead>
<tr>
<th>Effort prediction</th>
<th>SSA</th>
<th>ANN</th>
<th>ARIMA</th>
<th>Exp. Smooth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log MSE</td>
<td>16.61</td>
<td>18.49</td>
<td>16.55</td>
<td>16.62</td>
</tr>
<tr>
<td>Log BIC</td>
<td>16.71</td>
<td>18.60</td>
<td>16.66</td>
<td>16.73</td>
</tr>
<tr>
<td>log avg. ASE</td>
<td>-0.045</td>
<td>0.028</td>
<td>-0.029</td>
<td>0.012</td>
</tr>
<tr>
<td>R sq.</td>
<td>1.08</td>
<td>7.09</td>
<td>1.02</td>
<td>1.09</td>
</tr>
<tr>
<td>Mean Err</td>
<td>-1420.62</td>
<td>3768.65</td>
<td>-362.14</td>
<td>-1254.32</td>
</tr>
<tr>
<td>Relative Variance diff. (%)</td>
<td>89.09</td>
<td>419.23</td>
<td>98.78</td>
<td>100.00</td>
</tr>
<tr>
<td>Relative Mean diff. (%)</td>
<td>6.01</td>
<td>28.30</td>
<td>0.52</td>
<td>5.36</td>
</tr>
<tr>
<td>Relative Min diff. (%)</td>
<td>28.82</td>
<td>69.70</td>
<td>71.08</td>
<td>84.20</td>
</tr>
<tr>
<td>Relative Max diff. (%)</td>
<td>3.69</td>
<td>57.23</td>
<td>29.05</td>
<td>30.52</td>
</tr>
</tbody>
</table>

Table 2: Comparison between different forecasting approaches for the monthly fishing effort reported by IOTC in 2013.
Variation of the SSA forecasts with respect to the signal in 2012

<table>
<thead>
<tr>
<th>SSA parameters</th>
<th>Variance</th>
<th>Mean</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samples: 70 (5 years), Eigen. Thr: 0.088</td>
<td>26.26</td>
<td>7.08</td>
<td>16.84</td>
<td>22.53</td>
</tr>
<tr>
<td>Samples: 98 (7 years), Eigen. Thr: 0.089</td>
<td>0.80</td>
<td>14.66</td>
<td>34.05</td>
<td>2.59</td>
</tr>
<tr>
<td>Samples: 98 (7 years), Eigen. Thr: 0.07</td>
<td>9.58</td>
<td>11.97</td>
<td>45.24</td>
<td>3.04</td>
</tr>
<tr>
<td>Samples: 126 (9 years), Eigen. Thr: 0.07</td>
<td>18.50</td>
<td>3.60</td>
<td>6.90</td>
<td>21.14</td>
</tr>
<tr>
<td>Samples: 156 (11 years), Eigen. Thr: 0.07</td>
<td>5.24</td>
<td>3.94</td>
<td>39.82</td>
<td>20.54</td>
</tr>
<tr>
<td>Samples: 238 (17 years), Eigen. Thr: 0.047</td>
<td>19.38</td>
<td>7.10</td>
<td>20.75</td>
<td>21.36</td>
</tr>
<tr>
<td>Samples: 266 (19 years), Eigen. Thr: 0.035</td>
<td>47.53</td>
<td>2.32</td>
<td>25.05</td>
<td>23.36</td>
</tr>
</tbody>
</table>

Table 3: Similarity measurements of several SSA forecasts with respect to IOTC monthly fishing effort in 2012.
Comparison between time series in 2013 and best forecasts

<table>
<thead>
<tr>
<th>SSA samples</th>
<th>Fishing effort</th>
<th>Catch</th>
<th>CPUE</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>126</td>
<td>322</td>
<td>350</td>
<td>350</td>
<td>378</td>
</tr>
<tr>
<td>Eigen. Thr</td>
<td>0.07</td>
<td>3.2E-16</td>
<td>1.00E-16</td>
<td>2.20E-16</td>
<td>4.80E-17</td>
</tr>
<tr>
<td>Log MSE</td>
<td>16.61</td>
<td>18.90</td>
<td>10.47</td>
<td>1.72</td>
<td>4.31</td>
</tr>
<tr>
<td>Log BIC</td>
<td>16.71</td>
<td>19.01</td>
<td>10.58</td>
<td>1.83</td>
<td>4.42</td>
</tr>
<tr>
<td>R sqr.</td>
<td>1.08</td>
<td>1.56</td>
<td>1.81</td>
<td>1.02</td>
<td>10.91</td>
</tr>
<tr>
<td>Mean Err</td>
<td>-1420.62</td>
<td>1127.25</td>
<td>115.91</td>
<td>0.27</td>
<td>-7.31</td>
</tr>
<tr>
<td>Relative Variance diff. (%)</td>
<td>89.09</td>
<td>47.67</td>
<td>66.47</td>
<td>66.11</td>
<td>246.05</td>
</tr>
<tr>
<td>Relative Mean diff. (%)</td>
<td>6.01</td>
<td>13.33</td>
<td>33.28</td>
<td>6.85</td>
<td>12.02</td>
</tr>
<tr>
<td>Relative Min diff. (%)</td>
<td>28.82</td>
<td>11.10</td>
<td>21.06</td>
<td>6.77</td>
<td>7.30</td>
</tr>
<tr>
<td>Relative Max diff. (%)</td>
<td>3.69</td>
<td>33.37</td>
<td>40.45</td>
<td>199.28</td>
<td>14.79</td>
</tr>
</tbody>
</table>

Similarity between time series in 2012 and best forecasts

| | Variation in the mean (%) | Variation in the minimum (%) | Variation in the maximum (%) | Variation in the variance (%) |
| | 3.60 | 6.90 | 21.14 | 18.50 |
| | 3.69 | 60.46 | 1.76 | 8.77 |
| | 17.20 | 9.92 | 9.39 | 2.42 |
| | 1.59 | 7.43 | 11.17 | 22.46 |
| | 2.17 | 2.42 | 0.11 | 22.02 |

Table 4: Comparison between the best forecasts of the fishing dimensions and the real IOTC monthly-time series in 2013 and in 2012. As for 2013, we report also the following similarity measures: mean squared error (MSE), Bayesian information criterion (BIC), R-squared measure of goodness of fit (R sqr.) and mean error.