Empirical Orthogonal Maps (EOM) and distance between empirical spatial distributions. Application to Mauritanian octopus distribution over the period 1987-2017

Nicolas Bez, Didier Renard, Dedah Ahmed-Babou

Empirical Orthogonal Maps (EOM) and distance between empirical spatial distributions. Application to Mauritanian octopus distribution over the period 1987-2017.

BEZ Nicolas[1], RENARD Didier[4], AHMED-BABOU Dedah[1,2,3]

1 MARBEC, IRD, Univ Montpellier, Ifremer, CNRS, Sète, France

2 IMROP, Nouadhibou, Mauritania

3 OESP, Nouackchott, Mauritania

4 PSL, Centre de Géosciences, Fontainebleau, France

Correspondence author: nicolas.bez@ird.fr

Running headline: Factorization of spatio-temporal data using EOM.

Short recommendation:

This paper tackles the recurrent problem of analysing time series of spatial observations with spatially explicit methods. The approach is deliberately data based. We clarify and transform the method called Min-max Autocorrelation Factor to address ecological questions. EOMs are coined as a spatially explicit improvement of EOFs.

Abstract

Point 1. Exploratory data analysis of spatio-temporal ecological observations often leads to decompose the spatio-temporal problem into a spatial part multiplied by a temporal part (factorization). Empirically, factorization can be done by principal component analyses (PCA). However, this produces factors that are spatially correlated leading to incomplete factorization. The purpose of this work is twofold: first, to detail the Min-max Autocorrelation Factor (MAF) developed a long time ago and to highlight its limitations and assumptions and, second, to adapt the method to ecological applications leading to Empirical Orthogonal Maps (EOM).

Point 2. Empirical Orthogonal Maps (EOM) owe their name to the fact that they correspond to Empirical Orthogonal Functions (EOF) enhanced by a spatially explicit PCA. The method allows to extract the spatial patterns that explain most of the variability of a set of spatio-temporal observations indexed by time and to quantify their relative importance through time.

Point 3. Application on a time series of sixty-one scientific monitoring surveys targeting octopus distribution off the Mauritanian coast indicates that tens basic maps are sufficient to recover 68% of the total variability, and that the first two EOMs explain 38.4% of this variability. It also demonstrates the existence of a clear seasonality of the spatial patterns that are then estimated by weighted means of the EOMs.

Point 4. This manuscript clarifies the concept of orthogonality between factors in a spatial context (whether real for some distances and assumed for others). This provides the conditions for using Euclidean distance between spatial distributions, which, in turn, founds the reduction of a large set of spatial distributions into a small subset of basic spatial distributions explaining most of the variability within the set of input maps.

Key-words: Spatio-temporal data, factorization, decomposition, distance between maps, spatial distribution, dimension reduction.

1. Introduction

Several approaches, stochastic or deterministic, parametrical or empirical, can be explored to analyse data with a dual spatial and temporal dimension. The complexity of the landscape of possible approaches is complicated by the semantic that, sometimes, introduces ambiguity. For instance, it is common to say that some methods are spatial while, in fact, they are not. Our criterion is to consider that a method is spatial (respectively temporal), if the interchange of two observations in space (respectively in time) modifies the result. Spatial representation of outputs is not sufficient to pretend that the method is spatial. For instance, Empirical Orthogonal Functions - EOFs (Lorenz, 1956) produce outputs that can be represented on a geographical scale. However, EOFs do not account for spatial structures and spatial autocorrelation that exist in the observations and are not, per se, spatial statistics. This was precisely the improvement brought in by Switzer and Green (1984) with the Min-max Autocorrelation Factors - MAF that explicitly account for spatial correlation.

When dealing with spatio-temporal observations, a common objective of a large part of the different approaches is to factorize the spatio-temporal problem into a spatial part multiplied by a temporal part. Based on the Stochastic Partial Derivative Equation - SPDE framework recently developed by Lindgren et al. (2011), Thorson et al. (2015) proposed a Spatial Factor Analysis – SFA. This approach uses a reduced number of orthogonal random fields with SPDE characteristics. Each latent random field is indeed a stationary random Gaussian field characterized by a Matérn spatial covariance function (related to the distance between observations). In such a model, spatial independence between factors is guaranteed by construction and their number is constrained to be small for inference and parsimony considerations. This framework is thus quite appealing in that it allows identifying few factorial maps summarising the input spatio-temporal signal. A possible drawback of this approach is that the factors, both their numbers and their shapes, are model based. Once the model types

are chosen (i.e. Matérn spatial autocorrelations functions), the parameters are inferred using Integrated Nested Laplace Nested Approximation - INLA algorithms making the overall approach efficient.

Switzer & Green (1984) proposed a factorization based on so-called min-max autocorrelation factors – MAFs to filter out the noise of a series of multichannel spatial imagery data. Even though their approach recourses to the computation of variograms, no variogram model is required and their approach is model-free. The MAF procedures is a sequence of two Principal Component Analyses – PCA, and can be fully developed empirically, that is, without any parametric assumption. Since their seminal work, several studies have been developed using MAF decompositions (Shapiro & Switzer, 1989) notably in marine ecology (Fujiwara, 2008; Woillez et al., 2009; Petitgas et al., 2020).

When applied to multivariate time series, it is often the case that the different variables are systematically measured over time (preferably regular but not necessarily). In these cases MAFs can be readily applied, as in Solow (1994) or in Woillez et al. (2009). The situation is more problematic considering time series of spatial distributions. Precisely because it is a model-free approach, the main drawback of MAF is that they need isotopic observations, i.e. data observed at the same spatial points over time (Wackernagel, 2003). This is seldom achieved for spatial survey following a scientific random design. In these cases, observations must be transformed into isotopic sets of data prior to factorization by ad hoc means. For instance, Petitgas et al. (2020) used migration to the nearest point of a common grid prior to the computation of MAFs. However, when the sampling protocol varies from one survey to the other, or when the number of samples fluctuates between surveys, heterotopy cannot be solved by simple migration; instead, a spatial interpolation (e.g. kriging) is required to return to isotopy.

Particularly suitable situations embrace the regular grids. In these cases, as we will see in this manuscript, the first PCA of the MAFs is nothing but an EOF. This paves the way for a new

nomenclature that is further justified by the following consideration. As it is underlined in this paper, MAFs are ordered by descending spatial autocorrelation intensity (the first MAFs are, by construction, those with the strongest autocorrelation at small geographical distance). This is the reason why the first MAFs suppress any noise component. However, in the ecological context, we defend the idea that what matters is rather the relative importance of the different spatial factors in the variability the observations, which corresponds to a different perspective. The objective is no longer to filter out the noise but to build the spatio-temporal patterns that best explain the time series of distribution maps, with the possible consequence that the most important part of the spatio-temporal signal could be the noise (nevertheless, this is important to know from an ecological point of view). We thus reformulate the MAF so that the first factors of the decomposition, hereafter called Empirical Orthogonal Maps – EOMs, make it possible to reproduce a reasonable portion of the variability of the initial data, which is not necessarily ensured by MAF.

The objectives of this manuscript are then threefold. First, we detail the steps of calculation of EOMs and clarify the standardization step which is important for the interpretation of the outputs. Second, we indicate how to choose the factorial maps that explain most of the variability. Third, we emphasize that the orthogonality only concerns the short spatial scales and discuss the possibility to compute Euclidean distance between spatial distributions.

The developments are illustrated by an application on a time series of thirty years of scientific monitoring surveys of the octopus spatial distribution off the Mauritanian coast.

2. Method

The Min-Max Autocorrelation Factor (MAF) was developed by Switzer and Green (1984) to eliminate noise in a family of hyperspectral images. When the image series is indexed by time, MAFs allow the spatio-temporal information present in the data set to be represented as a

product of two terms. The first term (purely spatial) expresses spatial patterns from the data which, once recombined with weights fixed by the second term (purely temporal) give back the initial set of observations. In statistics, the factorisation brings the concept of independence. However, as mentioned below, this indeed only refers to non-correlation. The above-mentioned factorisation amounts to an augmented Empirical Orthogonal Factorization (EOF), which we therefore call an Empirical Orthogonal Mapping (EOM). Each term is important and meaningful: "Empirical" indicates that no parametric assumption is required, "Orthogonal" refers to non correlation, and "Map" specifies that the factors of the decomposition are maps or spatial distributions (contrary to PCA where the factors are variables). The EOM is thus a min-max autocorrelation factorization of a set of spatially regular distributions indexed by time.

2.1. Construction of EOMs

The whole process is developed outside of any probabilistic framework even though the use of random function formalism would be relevant. In this regard, capital letters are used to denote matrices. The EOMs are obtained by linear combination of a set of spatially regular distribution maps represented by the matrix Z of dimension $S \times T$ where S indicates the spatial dimension (e.g. the number of geographical positions sampled each time) and T the temporal dimension (e.g. the number of times the spatial distribution is observed). Without loss of generality, we consider that Z has been centred according to $Z - 1 * m^t$ where m^t is the transpose of m, a $1 \times T$ vector containing the mean of each map, and where 1 is a $S \times 1$ vector of ones. The matrix notation provides a connexion with the continuous space-time framework where $Z[s,t]$ represent the value at site $s,\ s = 1, \dots, S$ and at time $t,\ t = 1, \dots, T$.

EOMs consists of a double PCA where the first PCA is nothing but an EOF. It is based on the eigen elements of ρ_0, the correlation matrix of Z, that is the variance-covariance matrix of $Z_s = Z * D_{s^{-1}}$ where $D_{s^{-1}}$ is the $T \times T$ diagonal matrix of the inverse standard deviations.

Standardizing the observations (or not) is key for the final interpretation of the output. The correlation considered here is the correlation between observations made at the same site but at different time, that is for a 0 distance in the geographical space. If we denote Λ_0 the vector of the eigenvalues and P_0 the $T \times T$ matrix of the corresponding eigenvectors, the EOF are written as

$$F = Z_s * P_0 = Z * D_{s^{-1}} * P_0$$

where F is a $S \times T$ matrix with the same dimension as Z. The variances of the factors are given by the corresponding eigenvalues traditionally organised by decreasing values. Their sum equals T the number of input spatial distributions. This first PCA is a projection of the standardized raw data in an orthogonal (but not orthonormal) base formed by the empirical orthogonal factors. In the spatio-temporal representation, the orthogonality between the factors refers to non-correlation of the factor values at the same geographical points. However, these factors may exhibit spatial correlations in the sense that their spatial covariances may be different from zero. The factors of an EOF are statistically, but not spatially, orthogonal.

The second PCA thus aims to construct new factors with no spatial correlation for a given spatial distance (which usually corresponds to the distance to the nearest neighbour in the case of systematic spatial sampling or to the average distance to the nearest neighbour in irregular cases). The EOFs are first standardised, that is divided by the square root of their eigenvalues:

$$F_s = F * D_{\Lambda_0^{-1/2}}$$

The second PCA is then based on the diagonalization of the variance-covariance matrix of their spatial increments for a given distance lag h, that is on the eigen decomposition of (twice) the matrix of variogram and cross-variogram values between standardized EOFs. If we denote Γ_h the $T \times T$ matrix of (twice) the variogram and cross-variogram, Λ_h its eigen values and P_h its eigen vectors, the EOMs are obtained by:

$$X = F_s * P_h = Z * D_{s^{-1}} * P_0 * D_{\Lambda_0^{-\frac{1}{2}}} * P_h = Z * P$$

With the operator P

$$P = D_{s^{-1}} * P_0 * D_{\Lambda_0^{-1/2}} * P_h$$

In this expression, the matrix P explicitly describes the sequence of operations required to build EOMs. First, the data are standardized (diagonal matrix $D_{s^{-1}}$); then, their statistical variance-covariance matrix is diagonalized and the data are projected in this orthogonal space (P_0); then, the factors constituting this orthogonal space are normalized (diagonal matrix $D_{\Lambda_0^{-1/2}}$); and, finally, (twice) the variogram-cross-variogram matrix at distance h of the projected data is diagonalized and the projected data are projected once more in this new orthogonal space (P_h). X is an $S \times T$ matrix.

By construction, the final factors X, have a unit variance, and are uncorrelated locally and at distance h (the proof is given in Switzer and Green, 1984). The eigenvalues Λ_h equal (twice) the variogram of the final factors at distance h and are generally organised in increasing order so that the first factors correspond to the strongest spatial structures. For two EOMs ranked i and j, which are up to now similar to MAFs, we get:

$$\gamma_{i,j}(h) = \begin{cases} \frac{1}{2}\Lambda_{\mathrm{h}}[i] & \text{if } i = j \text{ (simple variogram)} \\ 0 & \text{if } i \neq j \text{ (cross-variogram)} \end{cases}$$

Each EOM ($\mathrm{X}[.,t], t = 1, \ldots, T$) represents a spatial distribution, which is a linear combination of the input spatial distributions. As EOMs are obtained by linear combinations of the input spatial distributions, this can be reversed (given that P is invertible; see the section "Practical considerations"). Each input spatial distribution can, in turn, be expressed as a linear combination of the full set of the EOMs without any approximation (back-transformation):

$$\mathrm{Z} = \mathrm{X} * \mathrm{P}^{-1} = X * Q$$

Approximation of the input spatial distributions are obtained by the linear combination eventually reduced to the r-first EOMs:

195 $$\hat{Z}_1^r = \mathrm{X}[.,1{:}\,\mathrm{r}] * Q[1{:}\,r,.]$$

196 $$(S \times T) = (S \times r) * (r \times T)$$

197 Obviously, the larger the r parameter, the better the approximation. While the primer objective
198 of Switzer and Green (1984) was to remove the noisy part of a set of images, they ordered the
199 factors by increasing variograms values at distance h in order to favour the most spatially
200 regular factors. However, the objective here is to select as few factors as possible that explain
201 as much variability as possible. There is a priori no reason for the two arrangements to converge.
202 The EOMs are thus re-arranged by their percentage of explained variance as follows.
203 In PCA, the variance or the total inertia I is equal to the sum of the variances of the input
204 variables, that is to the trace of the variance-covariance matrix of $I = tr(C_Z)$, and the
205 percentage of variance explained by the first factors is equal to the sum of their eigenvalues
206 over the inertia. This cannot be directly transposed to the EOMs.
207 Therefore we define the percentage of variance explained by each EOM as:

208 $$\frac{tr(C_{\hat{Z}_\mathrm{r}^r})}{tr(C_Z)} = \frac{tr(Q[r,.]^t * Q[r,.])}{tr(C_Z)} = \frac{\sum Q[r,.]^2}{tr(C_Z)}$$

209 This makes it possible to rank and re-arrange the EOMs according to the percentage of variance
210 they explain and select the ones that explain the largest part of the input variance.
211
212 2.2. Mathematical orthogonality and statistical independence
213 In spatial statistics, the orthogonality implies the absence of correlation between factors at any
214 given possible distance (i.e. not only locally for zero distance). Strictly speaking, this means
215 that the set of EOMs form an orthogonal base if and only if they are spatially uncorrelated.
216 However, the definition of the EOMs insures zero covariance only at distances 0 and at distance
217 h. While it is a step ahead towards spatial non-correlation compared to traditional EOFs, the
218 EOMs are not fully spatially orthogonal and do not form a basis *sensus stricto*.

However, a weak definition of non-correlation is envisaged here. The EOMs are said to be weakly spatially uncorrelated if the mean value of their cross-variogram values equal 0 at any given distance:

$$\overline{\gamma_{i,j}(h)} = 0 \quad \forall h, \text{ if } i \neq j$$

This can be applied to the full set of the EOMs or to the r-first ones in order to qualitatively diagnose if the space where the input spatial distributions are projected is more or less spatially orthogonal.

2.3. Distance between spatial distributions

In the EOMs framework, each input spatial distribution corresponds to a vector of coefficients of their EOMs decompositions. If the r-first EOMs are uncorrelated, these coefficients give the coordinates of the input spatial distributions in the orthonormal space defined by the r-first EOMs. Each dimension of this space is defined by an EOM, that is, a basic spatial distribution. The first factorial plan is thus the 2D space defined by the two spatial distributions that explain most of the variability of the input spatial distributions that is the first two EOMs. Each input distribution can be represented by a point in this 2D space, thanks to its first two coordinates. This can be generalised to any r-dimensional EOMs space and opens the use of a distance-based metric to compare distribution maps. For instance, hierarchical ascending classification or k-means, to group spatial distributions whose decompositions in the base of EOMs are similar. Given the above discussion on orthogonality (in the mathematical and statistical senses), the recourse to Euclidean distance between spatial distributions in an r-dimensional EOM space is as relevant as the orthogonality of the EOMs is effective.

There is a strict equivalence in considering:

- the coefficients of the decomposition of an input spatial distribution on r basic spatial distributions/EOMs

- the coefficients of the approximation of an input spatial distribution by the linear combination of r basic spatial distributions/EOMs
- the coordinates of an input spatial distribution in an orthogonal space made of r basic spatial distributions/EOMs

2.4. Standardised and non-standardised EOMs

As indicated by the back-transformation equations, the coefficients of the decomposition of a given spatial distribution, say $Z[.,t]$, in an r-dimensional EOM space correspond to $P^{-1}[1{:}r,t]$, that is the r-first lines of the t^{th} column of the rebuilding matrix P^{-1}.

Back-transformation, dimension reduction and approximation can refer to the standardized input data:

$$\mathrm{Z_s} = \mathrm{X} * \mathrm{Q_s}$$

with

$$P_s = P_0 * D_{\Lambda_0^{-\frac{1}{2}}} * P_h \quad \text{and} \quad Q_s = P_s^{-1}$$

$$\widehat{\mathrm{Z}_{s_1}^{\mathrm{r}}} = \mathrm{X}[.,1{:}\mathrm{r}] * \mathrm{Q_s}[1{:}r,.]$$

This allows analysing the shape of the input spatial patterns irrespective to their level of variability. In this case, spatial distributions are similar if they have the same patterns, up to a multiplicative value. The corresponding percentage of correlation explained by the r-first EOMs is given by

$$\frac{tr(C_{\widehat{\mathrm{Z_s^r}}})}{T} = \frac{\sum Q_s[r,.]^2}{T}$$

which may differ from the percentage of variance explained.

2.5. Practical considerations

EOMs are built for a given reference distance h. In practice, one often needs to use some tolerance around the reference distance either to account for sampling sites that are irregularly spaced and/or to account for a large distance interval for which the cross-variograms will be null on average to ensure weak independence for instance.

The signs of the eigen-elements are purely conventional but coherent between eigenvectors and their eigenvalues. Therefore, their interpretation must be established jointly. An EOM whose coefficients are all negative is interpreted according to its symmetric values with respect to 0.

Empirical variance-covariance matrices are not always positive definite. In particular, when $T \geq S$, that is when the number of surveys is larger than or equal to the number of sampling sites per survey, the variance-covariance matrices is singular and cannot be inverted. In this case, its eigen-elements do not exist and the EOM decomposition is not possible.

The time series of input spatial distributions may not be regular in time. It is also worth mentioning that indexes t could be interchanged without modifying the EOM outputs. EOM is not a method that is explicitly temporal. However, unlike EOF that are not a spatial method (sites could be interchanged without modifying the EOF outputs), EOM are spatially explicit.

All the computations were performed under R using the package RGeostats, which can be freely downloaded from http://cg.ensmp.fr/rgeostats. The scripts are available in the archive associated to this manuscript.

3. Data

EOMs are used to analyse the time series of sixty-one ($T = 61$) octopus surveys on board the research vessels N'Diago and AL-Awam during the periods 1987-1996 and 1997-2017 respectively. Each survey follows a stratified random sampling based on three latitudinal strata (Fig. 1). The average number of samples is 102 per survey with some surveys having only few

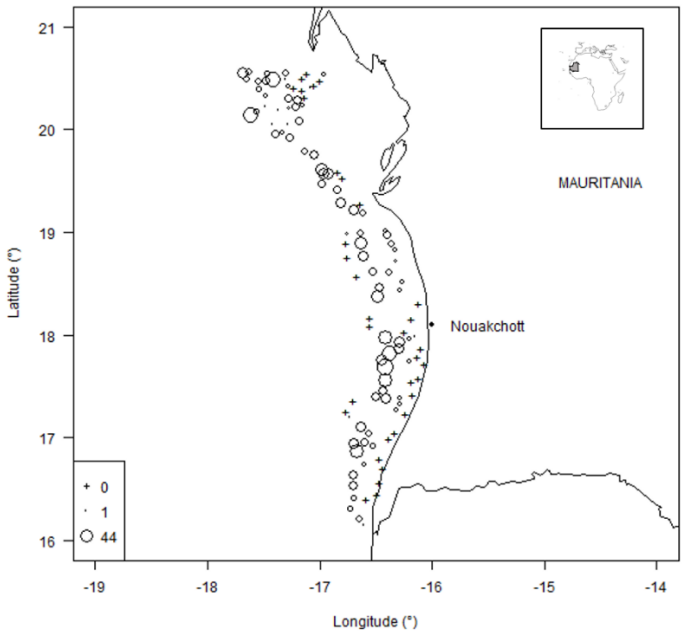
tens of samples. In each sampling site, the density of octopus is provided in number of individuals per swept area (on average 0.055 km²). An inter-calibration experiment between the two research vessels was carried out for the period 1987-89 in order to make the data series homogeneous by taking into account the change of fishing gear that took place in 1989 (Gascuel et al., 2007). The timing of the surveys is not regular but there is at least one survey per year. The position of the sampling sites of each survey are drawn at random and are thus different from one survey to the others, with surveys with a low spatial coverage. So, prior to EOMs computations, survey data are interpolated by ordinary block kriging (Chilès & Delfiner, 2012) on a regular 0.1° x 0.1° grid restricted to the polygon of presence of octopus. The number of active grid cells is $S = 341$. As $T < S$, it is possible to compute EOMs.

As all regression techniques, kriging is a smoothing. This means that the kriging maps do not have the same level of variability than the raw input observations. This is in favour of the use of standardized EOMs that allow comparing and grouping the surveys only based on the shapes of their spatial distributions.

4. Results

The first EOM alone explain 28.5% of the overall variability (Fig. 2) and four basic EOMs were enough to recover more than half of the input variability. The ranking based on the percentage of explained correlation is not fully consistent with the ranking based on the variogram value (see for instance, the sixth EOM whose spatial regularity is not intermediate between that of the fifth and the seventh ones). The local variability, i.e. the variogram value at the reference distance of 0.1°, starts to increase after the tenth EOMs. Meanwhile, the ten first EOMs carry 68% of the overall variability. This is considered as a good compromise for dimension reduction ($r = 10$).

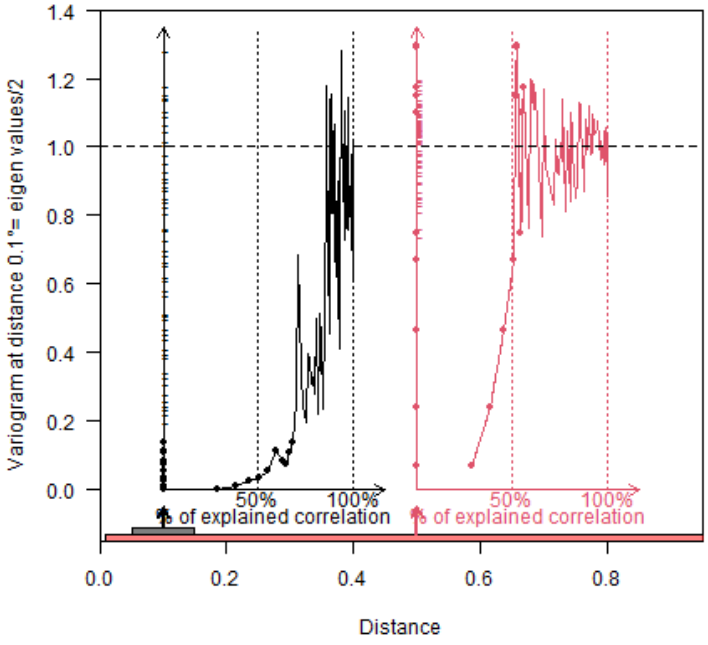
317 The cross-variogram values of the ten selected EOMs are equal to 0 for the reference distance
318 (i.e. $h = 0.1°$) and are, on average, equal to zero when mixing all pairs of EOMs together (Fig.
319 3a) which is consistent with a weak non-correlation. However, their fluctuations clearly
320 increase for larger distances. When EOMs are computed for a large interval of reference
321 distance, say for distances between 0° and 1°, their cross-variogram values are strongly reduced
322 around 0 on average for all distance classes between 0° and 1° (Fig. 3b) indicating that they
323 form an orthogonal base in the strong sense. However, this produces EOMs whose spatial
324 structures are rapidly pure noise with no spatial structure and low descriptive power (Fig. 2).
325 The first factorial space, i.e. the space formed by the two first EOMs, explains 38.4% of the
326 overall input correlation. In this factorial space, the surveys display two clear groups with little
327 overlap between them (Fig. 4). This is further investigated through a hierarchical ascending
328 classification (HAC) based on the 10 first EOMs using the Ward distance. The HAC underlines
329 the existence of two groups of surveys with similar spatial patterns (Fig. 5) that strongly
330 matches the climatic seasons (Fig. 5; accuracy = 68%). While the first three EOMs get clearly
331 and statistically (ANOVA with $p.value < 10^{-3}$; Fig. 6) different scores for the two clusters,
332 ANOVA diagnostics are also statistically significant for EOM ranked 6, 8, 9 and 10. Finally,
333 the vectors of the average scores per cluster are used to estimate the mean spatial distribution
334 of each cluster (Fig. 4).



335

336 Figure 1. Sampling protocol. Typical survey data (March, 2015). Octopus densities are in

337 number of individuals per swept area (on average 0.055 km²).



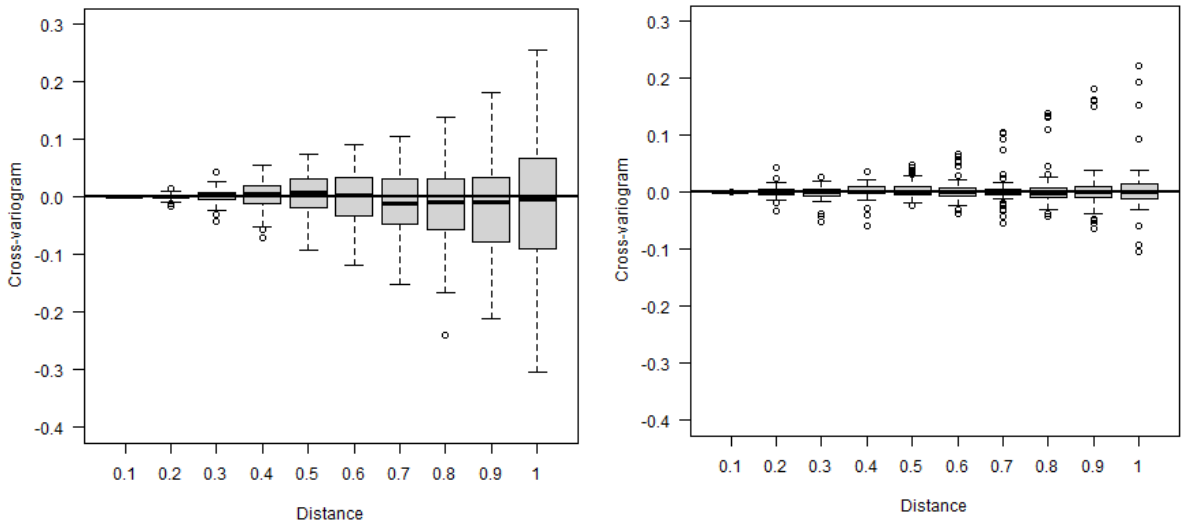
338

339 Figure 2. Values of the omnidirectional variogram of the EOMs for the grid cell distance (0.1°)

340 as a function of the percentage of correlation explained by the r-first EOMs. In black, the EOMs

341 corresponding to a reference distance equal to the grid cell size ($h = 0.1 +/-0.05$). In red, the

342 EOMs obtained when $h = 0.5 +/-0.5$. The first ten EOMs are depicted by plain circles.



343 Figure 3. Cross variogram values for the first distance lags for the first 10 EOMs. Left: EOMs
344 used in the study corresponding to a reference distance equal to the grid cell size ($h = 0.1 +$
345 $/-0.05$). Right, EOMs obtained when using a large interval reference distance ($h = 0.5 +$
346 $/-0.5$).

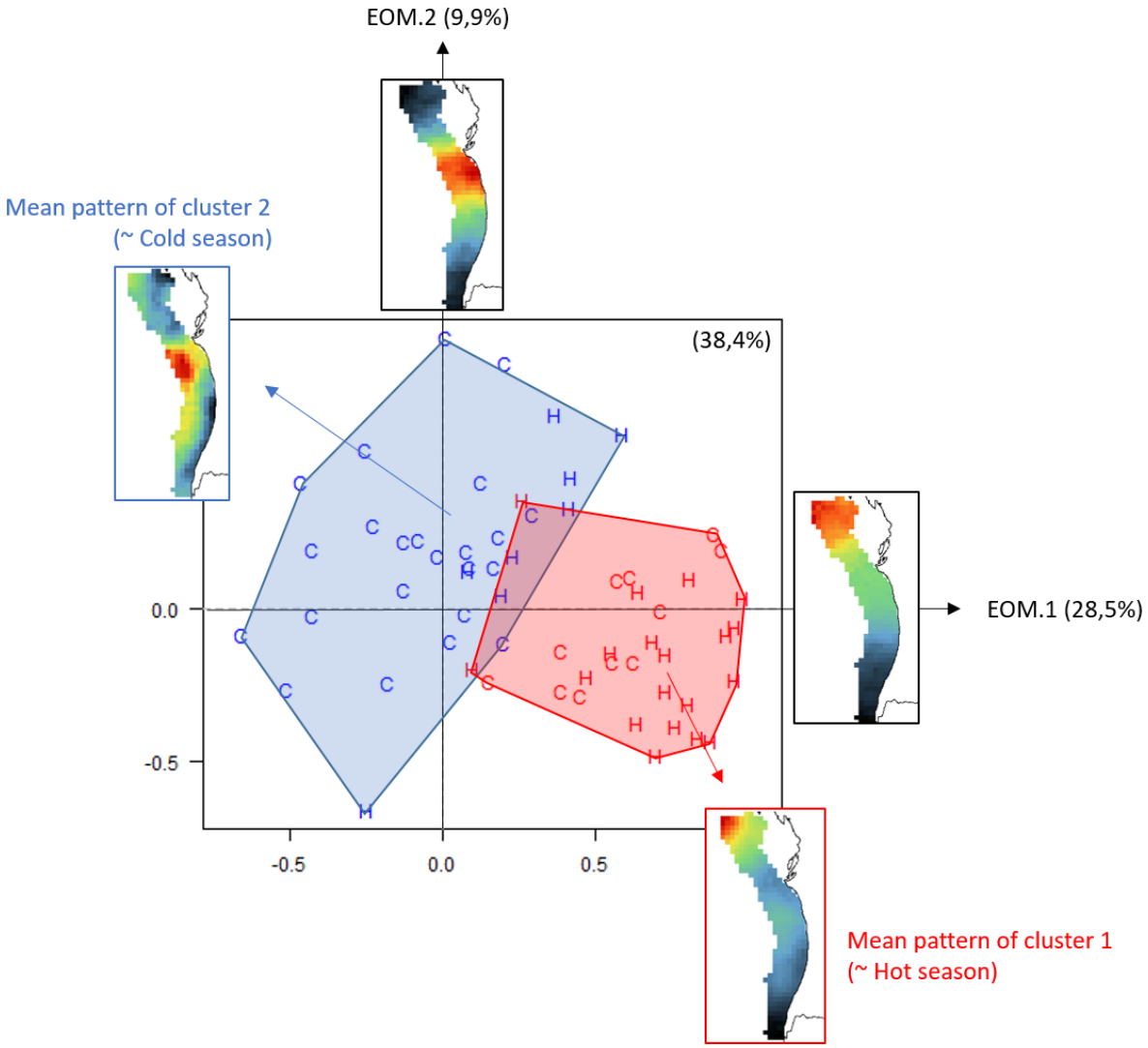


Figure 4. Factorial space made of the first two EOMs. In this space, the survey are located according to their standardised scores in the EOMs factorisation. The two EOMs associated to each dimension of the factorial space are represented. The mean distribution of each group is also represented. Surveys that happened during the hot season are denoted “H” and “C” for the cold season. Polygons are the convex hull encompassing the surveys that belong to the same cluster of the hierarchical ascending analysis.

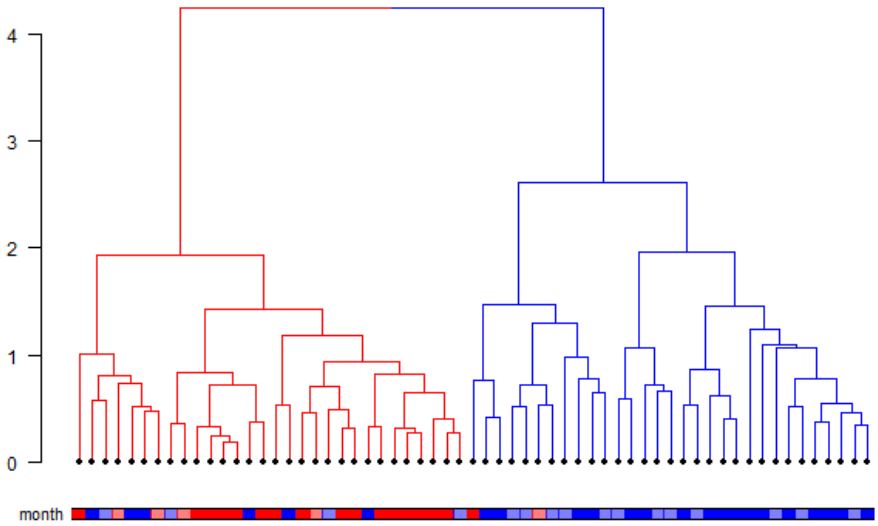


Figure 5. Dendrogram (Ward distance) of the 61 surveys based on the first ten EOMs. The colour bar is defined by the season: in blue, the months corresponding to the cold season (light blue for the two extremal months of the cold season) and, in red, the months corresponding to the hot season (light red for the two extremal months of the hot season).

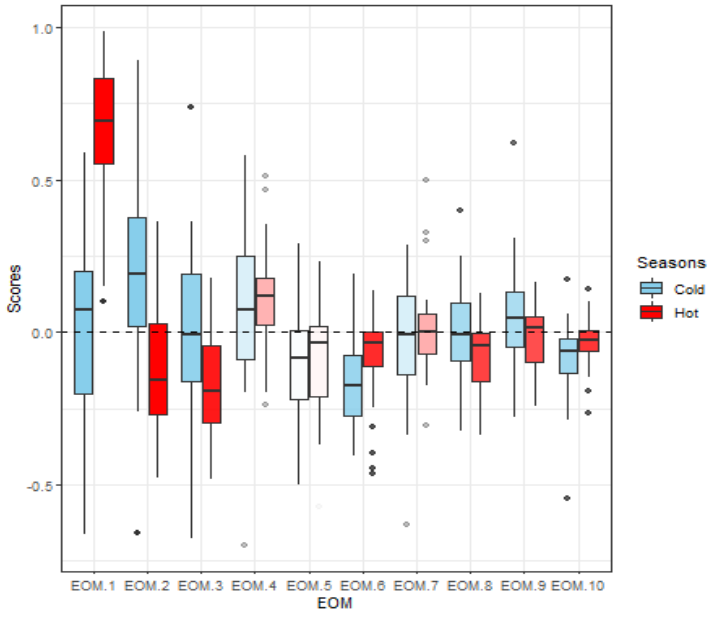


Figure 6. Boxplots per cluster of the scores of the decompositions of the 61 input spatial distributions on the 10 first EOMs. The colour' transparency is proportional to the p-value of the difference between means per cluster: the more transparent, the less significant the difference between the two seasons.

5. Discussion

Being defined by the sequence of two PCAs, EOM is fundamentally an empirical approach. The fact that it refers to variogram in the second PCA does not change this fact. By the way, what is minimized is not the value of a variogram model at a given distance, but the value of empirical covariance between the increments of two EOFs for a given reference geographical distance. The recourse to random function is thus external to the EOM: either before, in order to get values on fixed sampling sites over the study period, or after, to map the EOMs if needed. Being empirical, EOM has the drawback that it can only be performed for sets of sampling sites (spatially regular or not) that are systematically observed over time (isotopy). Factorizing a set of heterotopic variables (i.e. variables that are not observed at the set sets of geographical points) by linear combinations remains an open question. Spatial covariance (simple and cross covariance) can deal with heterotopic variables (Wackernagel, 2003). However, the fundamental nature of factorization being to make a linear combination of the variables at the same points, the problem cannot be fully solved by the recourse to covariance.

Similar to EOFs that are not spatial in their construction but that can be represented spatially, EOMs are not temporal but can be represented temporally. It is thus an abuse of language to say that EOM offers a spatio-temporal approach.

In statistics, the orthogonality of the factors (PCA, EOF, etc) refers to the absence of mutual correlations. In a spatial context, this means the absence of correlation at the same geographical points. After the second PCA, the orthogonality is extended to a given geographical distance but not to all possible distances. So EOMs do not reach a full orthogonality by construction. The only known case were orthogonal factors in the statistical sense are also fully spatially uncorrelated is the case of the proportional covariance also called intrinsic correlation model (Chilès & Delfiner, 2012). This model is very peculiar and very specific. In the present analysis, we have shown that a weak orthogonality can however be considered when mutual correlations

are not null one by one, but on average. We also have indicated that enlarging the distance interval used in the EOM computation can help reaching orthogonality. However, the EOMs that are obtained in this case lack of ecological interest. There is thus a compromise between orthogonality, that makes orthogonal representations and Euclidean distances relevant, and ecological interest when EOMs get meaningful spatial patterns.

MAFs were initially developed to eliminate noise in a set of images and extract their common signal. They are thus naturally ordered by decreasing autocorrelation at the reference distance, that is, by increasing order of the eigenvalue of the second PCA. Considering all MAFs in the analysis and interpretation allows to restore all the initial information and to describe perfectly the spatiotemporal variability of the observed data. However, there is no reason why the most important factors should be the most spatially regular. Indeed, the most important EOMs are rather those that endorse most of the variability. In our study case, the percentages of variance explained by the EOMs indicate that the two rankings are not similar, even though the first ten EOMs are the same.

Depending on the EOMs' algorithm, the approach refers either to the raw data or to their standardized version. The empirical Taylor's power law (Taylor, 1961) has been widely established in ecology. It can be summarized as the relationship between mean and variance (of count data). In this context, looking at spatial patterns relatively to the standard deviations amounts to study the spatial patterns relatively to the biomass. It is thus key to know precisely what kind of PCAs is used and which EOMs are generated. Analyses reported here concern the spatial patterns in relative terms in order to compare and to group the surveys considering only the shapes of their spatial distributions. This also weights down the impact of the kriging made prior to the analyses. This allowed grouping surveys that are characteristic of each ecological season, and allowed building the expected spatial distribution for each ecological season. This means that the climatic season coincided with an ecological season. This was known a priori.

However, this came as an resulting property from the comparison of the spatial distributions and was not included at first in the analysis. This is a key difference that strengthens the status of such a knowledge: it is no longer a priori brought in, but it is deduced from a long series of spatial distributions.

6. Acknowledgement

Authors would like to give deep recognition to the people working at the Institut Mauritanien de Recherche Océanographique et des Pêches - IMROP in Mauritania that organized the surveys, collected the raw data and made the work possible.

7. Conflict of Interest statement

No conflict of interest.

NB conceived the ideas, analysed the data and led the writing of the paper. DAB contributed to the collection the data, their analyses and contributed critically to the draft. DR contributed critically to the draft. All authors gave final approval for publication.

8. Data availability

All the (kriged) input data and all the scripts required to re-run the analysis are available at the following GitHub address : https://github.com/abambad/EOM/

9. References

Chilès, J.-P. & Delfiner, P. (2012). Geostatistics, modeling spatial uncertainties. Second edition, Wiley, pp 699.

Fujiwara, M. (2008). Identifying interactions among salmon populations from observed dynamics. Ecology, 89,1, 4-11.

Gascuel, D., Labrosse, P., Meissa, B., Taleb Sidi, M.O. & Guenette, S. (2007). Decline of demersal resources in North-West Africa: an analysis of Mauritanian trawl-survey data over the past 25 years. African Journal of Marine Science, 29(3): 331–345.

Lindgren, F., Rue, H. & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. Journal of Royal Statistical Society, B, 73, Part 4, 423-498.

Lorenz, E.N. (1956). Empirical orthogonal functions and statistical weather prediction. Massachusetts institute of technology, Departement of meteorology, Scientific report n°1, Cambridge, Massachusetts.

Petitgas, P., Renard, D., Desassis, N., Huret, M., Romagnan, J.-B., Doray, M., Woillez, M. & Rivoirard, J. (2020). Analysing temporal variability in spatial distributions using min–max autocorrelation factors: sardine eggs in the Bay of Biscay. Mathematical Geosciences, 52, 337-354.

RGeostats: The Geostatistical R Package. Version: 12.0.0. Free download from: http://rgeostats.free.fr/.

Shapiro, D.E. & Switzer, P. (1989). Extracting time trends from multiple monitoring sites. Technical report SIMS 132, Department of statistics, Stanford University, California.

Switzer, P., & Green, A.A. (1984). Min/max autocorrelation factors for multivariate spatial imagery. Technical report SWI NFS 06, Department of statistics, Stanford University, California.

Taylor, L.R. (1961). Aggregation, variance and the mean. Nature, 189, 732-735.

Thorson, J.T., Scheuerell, M.D., Shelton, A.O., See, K.E., Skaug, H.J. & Kristensen, K. (2015). Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range. Methods in Ecology and Evolution, 6, 627–637.

Wackernagel, H. (2003). Multivarite Geostatistics – An introduction with applications, 3rd ed. Springer, Berlin.

Woillez, M, Rivoirard, J. & Petitgas, P. (2009). Using min/max autocorrelation factors of survey-based indicators to follow the evolution of fish stocks in time. Aquatic Living Resources, 22, 193–200.