Empirical Orthogonal Maps (EOM) and Principal Spatial Patterns: Illustration for Octopus Distribution Off Mauritania Over the Period 1987–2017

Nicolas Bez, Didier Renard, Dedah Ahmed-Babou

▶ **To cite this version:**

Nicolas Bez, Didier Renard, Dedah Ahmed-Babou. Empirical Orthogonal Maps (EOM) and Principal Spatial Patterns: Illustration for Octopus Distribution Off Mauritania Over the Period 1987–2017. Mathematical Geosciences, 2023, 55 (1), pp.113-128. ⟨10.1007/s11004-022-10018-w⟩. ⟨hal-03892478⟩

HAL Id: hal-03892478

https://hal.science/hal-03892478v1

Submitted on 22 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

Mathematical Geosciences
January 2023, Volume 55 Issue 1 Pages 113-128
https://doi.org/10.1007/s11004-022-10018-w
https://archimer.ifremer.fr/doc/00795/90703/

Archimer
https://archimer.ifremer.fr

Empirical Orthogonal Maps (EOM) and Principal Spatial Patterns: Illustration for Octopus Distribution Off Mauritania Over the Period 1987–2017

Bez Nicolas [1, *], Renard Didier [2], Ahmed-Babou Dedah [1, 3]

[1] MARBEC, IRD, Univ Montpellier, Ifremer, CNRS, INRAE, Séte, France
[2] PSL, Centre de Géosciences, Fontainebleau, France
[3] IMROP, Nouadhibou, Mauritania

* Corresponding author : Nicolas Bez, email address : nicolas.bez@ird.fr

didier.renard@mines-paristech.fr ; dedah.ahmed-babou@ird.fr

Abstract :

Analysis of spatiotemporal observations often leads to decomposition of the problem into a spatial part multiplied by a temporal part (factorization). Principal component analyses produce factors that are temporally uncorrelated but that remain spatially correlated, leading to incomplete factorization. Min–max autocorrelation factors developed many years ago are adapted here to ecological applications, leading to empirical orthogonal maps (EOMs). EOMs owe their name to the fact that they are indeed an enhancement of empirical orthogonal functions which extract the spatial patterns that explain most of the variability of a set of spatiotemporal observations indexed by time. Application on a time series of 61 scientific monitoring surveys targeting octopus distribution off the Mauritanian coast indicates that ten basic maps are sufficient to recover 68% of the total variability, and that the first two EOMs explain 38.4% of this variability. This manuscript clarifies the concept of orthogonality between factors in a spatial context. This provides the conditions for using Euclidean distance between spatial distributions, which in turn supports 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.

Keywords : Spatiotemporal data, Factorization, Principal maps, Distance between maps, Dimension reduction

Please note that this is an author-produced PDF of an article accepted for publication following peer review. The definitive publisher-authenticated version is available on the publisher Web site.

1 INTRODUCTION

Several approaches, stochastic or deterministic, parametric or empirical, can be explored to analyze data with dual spatial and temporal dimensions. 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 exchange of two observations in space (respectively in time) modifies the result. Spatial representation of outputs is not sufficient to claim that the method that generated the outputs 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 auto-correlation 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 auto-correlation factors - MAF that explicitly account for spatial correlation.

When dealing with spatio-temporal observations, an objective shared by 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) conducted a spatial factor analysis – SFA. This approach uses a reduced number of orthogonal random fields with SPDE characteristics. Each latent random field is 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 summarizing 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 auto-correlations functions), the parameters are inferred using integrated nested Laplace approximation - INLA algorithms making the overall approach efficient.

Switzer and Green (1984) proposed a factorization based on so-called min-max auto-correlation factors – MAFs to filter out the noise of a series of multi-channel spatial imagery data. Even though their approach uses variogram computations, 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 (e.g., Shapiro and Switzer 1989) notably in marine ecology (Fujiwara 2008; Woillez et al. 2009; Petitgas et al. 2020), and several papers have contributed to further describe the connection between MAF and the linear model of coregionalization (LMC) (Desbarats

and Dimitrakopoulos 2000; Vargas-Guzman and Dimitrakopoulos 2003). These latter references established also that for particular situations like the intrinsic correlation models, the absence of cross-correlation at some particular distance propagates to all spatial distances insuring full orthogonality between factors.

When applied to multivariate time series, the different variables, taken in columns of the database, are often systematically measured over time taken as the rows (preferably regularly, but not always). In these cases, temporal 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. In these cases, each column is a spatial distribution ; the rows corresponding the geographical sites. Precisely because it is a model-free approach, the main drawback of MAF is that they need isotopic observations, that is that data should be 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 methods. 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 a simple migration; instead, a spatial interpolation (e.g., kriging) is required to return to isotopy. Regular and isotopic grids correspond to particularly suitable situations. 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 justified below.

MAFs are ordered by say, descending, spatial auto-correlation intensity (the first MAFs are, by construction, those with the strongest auto-correlation at small geographical distance). This is the reason why the first MAFs suppress noise components. However, in the ecological context for instance, what matters is rather the relative importance of the different spatial factors with regards to the variability of 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). In this paper, MAFs are thus reformulated 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, the calculation of EOMs are detailed, clarifying the standardization steps which is important for the interpretation of the outputs. Second, the percentage of variance explained is defined which gives the basis for the distinction between MAF and EOM. Third, the orthogonality between factors is enlarged to encompass mean orthogonality for spatial scales not restricted to the shortest ones. This opens the possibility to compute Euclidean distance between spatial distributions and to map them into small dimensional spaces.

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

2 METHOD

The whole framework can be developed outside of any probabilistic framework even though the use of random function formalism could be relevant. In this context, capital letters are used to denote matrices. The factors 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 locations sampled each time) and T the temporal dimension (e.g., the number of times the spatial distribution is observed). The matrix notation provides a connection with the continuous space-time framework where $Z[s,t]$ represents the value at site $s, s = 1,\ldots,S$ and at time $t, t = 1,\ldots,T$.

Without loss of generality, Z is supposed to be centered. Its standardized version is denoted $\dot{Z} = Z \cdot D_{s^{-1}}$ where $D_{s^{-1}}$ is the $T \times T$ diagonal matrix of the inverse standard deviations of the column of Z.

2.1 SPATIALLY NON-CORRELATED FACTORS

The factorization consists of a sequence of two PCAs. When dealing with regular gridded data, the first PCA is nothing but an EOF. It is based on the eigen elements of Γ_0, the correlation matrix of Z, that is, the variance-covariance matrix of $\dot{Z}$. 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, hence the notations used (subscript 0 means "related to distance equals to 0"). If we denote λ_0 the vector of the eigenvalues and P_0 the $T \times T$ matrix of the corresponding eigenvectors of Γ_0, the EOFs are given by

$$F_0 = \dot{Z}P_0, \tag{1}$$

where F_0 is a matrix with the same dimension as Z ($S \times T$). The variances of the factors are given by the corresponding eigenvalues. Their sum equals T, that is 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 at constructing new factors with no spatial correlation for a given spatial distance (which usually corresponds to the distance to the nearest neighbor in the case of systematic spatial sampling or to the average distance to the nearest neighbor in irregular cases). The factors of the first PCA are first standardized, that is divided by the square root of their eigenvalues

$$\dot{F}_0 = F_0 D_{\lambda_0^{-1/2}}. \tag{2}$$

The second PCA is then based on the diagonalization of the variance-covariance matrix of their spatial increments for a given reference distance $h = r$, that is on the

eigen decomposition of (twice) the matrix of variogram and cross-variogram values between standardized EOFs for the distance $h = r$. Denoting Γ_r this $T \times T$ matrix, λ_r the vector of its eigenvalues and P_r the matrix of its eigenvectors, the factors associated to the reference distance $h = r$ are

$$F_r = \dot{F}_0 P_r = Z\left(D_{s^{-1}} P_0 D_{\lambda_0^{-1/2}} P_r\right) = Z\Phi_r, \tag{3}$$

where F_r is an $S \times T$ matrix, and where the linear operator Φ_r to transform directly the input raw data into factors is defined by

$$\Phi_r = D_{s^{-1}} P_0 D_{\lambda_0^{-1/2}} P_r. \tag{4}$$

In this expression, the matrix Φ_r explicitly describes the sequence of operations required to build the factors:

1. first, the data are standardized ($D_{s^{-1}}$);
2. then, their statistical variance-covariance matrix is diagonalized and the data are projected in this orthogonal space (P_0);
3. then, the factors constituting this orthogonal space are normalized ($D_{\lambda_0^{-1/2}}$);
4. and, finally, (twice) the variogram-cross-variogram matrix at distance $h = r$ of the projected data is diagonalized and the projected data are projected in this new orthogonal space (P_r).

By construction, the final factors F_r, have a unit variance, and are uncorrelated locally ($h = 0$) and at distance $h = r$ (the proof is given in Switzer and Green, 1984). Each factor $(F_r[.,t], t = 1, \ldots, T)$ represents indeed a spatial distribution, which is a linear combination of the input spatial distributions. As they are obtained by linear combinations of the input spatial distributions, this can be reversed given that Φ_r is invertible with $\Phi_r^{-1} = \Psi_r$ (see the section "Practical considerations" below). Each input spatial distribution can, in turn, be expressed as a linear combination of the full set of the factors without any approximation (back-transformation) by

$$Z = F_r \Psi_r. \tag{5}$$

This can be summarized by the following way and back transformation scheme

$$Z \underset{\Psi_r}{\overset{\Phi_r}{\rightleftharpoons}} F_r. \tag{6}$$

2.2 ORDERING THE FACTORS AND DIMENSION REDUCTION: MAF VERSUS EOM

The back-transformation equation Eq. (5) offers the possibility to use only a subset of the factors when back-computing the spatial distributions from the factors. Approximation or filtration of the input spatial distributions are then obtained by the linear combination of, say, the n-first factors

$$\hat{Z}_{1:n} = F_r[.,1:n]\Psi_r[1:n,.]. \tag{7}$$

The question is thus to order the factors in accordance with the objectives of the dimension reduction.

2.2.1 MAF

While the primary objective of Switzer and Green (1984) was to remove the noise part of a set of images, they ordered the factors, called MAFs, by increasing variograms values at distance $h = r$ in order to favor the spatially most regular factors and to remove the factors with pure or strong nugget structure. The eigenvalues λ_r equal (twice) the variogram of the final factors at distance $h = r$, so that for two factors ranked i and j, one gets

$$\gamma_{i,j}(r) = \begin{cases} \frac{1}{2}\lambda_r[i] & \text{if} \quad i = j \quad \text{(simple variogram)} \\ 0 & \text{if} \quad i \neq j \quad \text{(cross variogram)} \end{cases}. \tag{8}$$

This is the basis for ordering MAF which are organized in increasing order of eigenvalues λ_r so that the first MAF gets the strongest spatial structures.

2.2.2 EOM

An alternative objective is to select as few factors as possible that explain as much variability of the observations as possible, further called EOMs. Obviously, the larger the n parameter, the better the approximation. In standard PCA, the total variance or the total inertia I is equal to the sum of the variances of the input variables, given by the trace of the variance-covariance matrix of $I = tr(C_Z)$, and the percentage of variance explained by the first factors is equal to the sum of their eigenvalues over the total inertia. This cannot be directly transposed to the spatial factors that are produced by a sequence of two PCAs. Alternatively, one can consider, one by one, the sets of T distributions obtained by the back-transformation of a single factor, say factor i

$$\hat{Z}_i = F_r[.,i]\Psi_r[i,.]. \tag{9}$$

The percentage of variance explained by this factor is hereafter defined by

$$p_i = \frac{tr(C_{\hat{Z}_i})}{tr(C_Z)} = \frac{tr(\Psi_r[i,.]^t\Psi_r[i,.])}{tr(C_Z)} = \frac{\sum_{k=1}^{T}\Psi_r[i,k]^2}{tr(C_Z)}, \tag{10}$$

This makes it possible to rank and re-arrange the factors (eigenvalues and eigenvectors) according to the percentage of variance that they explain and select the ones that reproduce the largest part of the input variance. This ordering leads to empirical orthogonal maps - EOMs.

There is a priori no reason for the two arrangements, MAF and EOM, to coincide.

2.3 MATHEMATICAL ORTHOGONALITY AND STATISTICAL INDEPENDENCE

In spatial statistics, the orthogonality implies the absence of correlation between factors at any given possible distance (i.e., not only for zero distance). Strictly speaking,

this means that the set of EOMs forms an orthogonal basis if and only if they are spatially uncorrelated. However, by construction, EOMs get zero covariance only at distance 0 and at distance $h = r$. While it is a step ahead towards spatial non-correlation compared to traditional EOFs, the EOMs are not fully spatially orthogonal and do not form a basis sensus stricto. In some cases, notably in case of intrinsic correlations models, absence of cross-correlation at some distance between principal components propagates to all distances (Goovaerts 1993; Desbarats and Dimitrakopoulos 2000; Rondon 2012). In these particular cases, statistical non correlation is equivalent to full spatial orthogonality. However, this relies on particular models of coregionalization. In an empirical approach which MAF are in essence, one can not rely on model's characteristics. An empirical alternative is however suggested below.

In their seminal paper, Switzer and Green (1984) computed MAF through a sequence of two PCA, the second one using spatial increments of a fixed spatial distance (e.g., the pixel size for grid data). A natural extension is to consider a range of distances instead of precise distance for this second PCA. A weak definition of orthogonality is thus envisaged here. Hereafter, regionalized variables are said to be weakly orthogonal if the mean value of all their cross-variograms for all possible distances is 0, that is if

$$\overline{\gamma_{i\neq j}(h)} = 0, \quad \forall h. \tag{11}$$

We thus suggest to compute EOMs using a range of spatial distances ($h \in [0,R]$) rather than a reference distance ($h = r$) as in the original paper of Switzer and Green (1984). To do so, the second PCA relies on $\Gamma_{[0,R]}$ the $T \times T$ matrix of (twice) the mean variogram and mean cross-variogram values for all possible distances between 0 and R. Using the above notation, this would lead to the following factorization

$$F_{[0,R]} = Z\Phi_{[0,R]} \quad \text{with} \quad \Phi_{[0,R]} = \left(D_{s^{-1}} P_0 D_{\lambda_0^{-1/2}}\right) P_{[0,R]}. \tag{12}$$

The only difference between this decomposition ($\Phi_{[0,R]}$) and the former one (Φ_r) relies on the second PCA ($P_{[0,R]}$ instead of P_r); the first PCA, non spatial, remains unchanged. Weak non-correlation can be applied to the full set of the EOMs or to the n-first ones in case of dimension reduction, in order to diagnose if the space where the input spatial distributions are projected is more or less spatially orthogonal.

2.4 DISTANCE BETWEEN SPATIAL DISTRIBUTIONS

In the EOMs framework, each input spatial distribution is represented by the vector of coefficients of their EOMs decompositions. If the n-first EOMs are uncorrelated, these coefficients give the coordinates of the input spatial distributions in the orthonormal space defined by the n-first EOMs. Each dimension of this space is defined by an EOM, that is, a basic spatial distribution. The first factorial plan, which corresponds to the first two EOMs, is the two-dimensional space defined by the two principal spatial distributions that explain most of the variability of the input spatial distributions. Each input distribution can be represented by a point in this two-dimensional space.

This can be generalized to any dimensional space and opens the use of a distance-based metric to compare distribution maps or clustering techniques. For instance, hierarchical ascending classification or k-means can be used to group spatial distributions whose decompositions in the basis of EOMs are similar. Given the above discussion on orthogonality (in the mathematical and statistical senses), the use of Euclidean distance between spatial distributions is as relevant as the orthogonality of the EOMs is effective. In this context, there is a strict equivalence in considering:

1. the coefficients of the decomposition of an input spatial distribution on n basic spatial distributions/EOMs,
2. the coefficients of the approximation of an input spatial distribution by the linear combination of n basic spatial distributions/EOMs and,
3. the coordinates of an input spatial distribution in an orthogonal space made of n basic spatial distributions/EOMs.

2.5 STANDARDIZED AND NON-STANDARDIZED EOMS

Back-transformation, dimension reduction and approximation can refer to the standardized input data. This allows analyzing 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 EOMs are the same as their computation is based on standardized data, but the transformation and back-transformation matrices differ slightly as matrix $D_{s^{-1}}$ must be removed.

$$\dot{Z} = F_r \dot{\Psi}_r, \tag{13}$$

with

$$\dot{\Phi}_r = P_0 D_{\Lambda_0^{-1/2}} P_r = \dot{\Psi}_r^{-1}. \tag{14}$$

The percentage of correlation explained by the n-first EOMs may however differ from the percentage of variance explained.

2.6 Practical considerations

EOMs are built for a given reference distance $h = r$. However one often needs to use some tolerance around the reference distance to account for sampling sites that are not regularly spaced, or a given reference distance lag $h \in [0, R]$ to ensure weak non-correlation.

The signs of the eigen elements are purely conventional but coherent between eigenvectors and their eigenvalues. Therefore, their interpretation must be established jointly.

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 indices 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.

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).

All the computations were performed under R using the package RGeostats (MINES-ParisTech/ARMINES 2021). The scripts and the data required to reproduce the analysis are available here: `https://github.com/abambad/EOM_Proj`.

3 APPLICATION

3.1 DATA

EOMs are used to analyze the time series of sixty-one ($T = 61$) octopus surveys made with 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 samples 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 over years but there is at least one survey per year.

The geographical locations of the sampling sites of each survey are drawn at random and are thus different from one survey to another, with surveys with a low spatial coverage. So, prior to EOMs computations, survey data were interpolated by ordinary block kriging (Chilès and Delfiner 2012) on a regular 0.1° x 0.1° grid restricted to the polygon of presence of octopus with a kriging neighborhood of 0.75°. Given the small latitudes of the study area no projection was required (no significant space deformation at these latitudes). The omni-directional experimental variograms got reasonable spatial structures (see Supplementary information). The number of active grid cells is $S = 341$ ($T < S$).

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

3.2 RESULTS

The first EOM alone explains 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, that is the variogram value at the reference distance of 0.1°, starts increasing after the tenth EOMs. Moreover the ten first EOMs carry 68% of the overall variability. This is considered as a good compromise for dimension reduction ($n = 10$).

The cross-variogram values of the ten selected EOMs are equal to 0 for the reference distance (i.e., h=0.1°) and are equal to zero on average when mixing all pairs of EOMs together (Fig. 3a) which is consistent with a weak non-correlation. However, their fluctuations clearly increase for larger distances. When EOMs are computed for a large interval of reference distance, say for distances between 0° and 1°, their cross-variogram values are strongly reduced around 0 on average for all distance classes between 0° and 1° (Fig. 3b) indicating that they form an orthogonal basis in the strong sense. However, this produces EOMs whose spatial structures are rapidly become pure noise with low descriptive power (Fig. 2). The first factorial space, that is the space formed by the two first EOMs, explains 38.4% of the overall input correlation. In this factorial space, the surveys display two clear groups with little overlap (Fig. 4). This is further investigated through a hierarchical ascending classification (HAC) based on the 10 first EOMs using the Ward distance. The HAC underlines the existence of two groups of surveys with similar spatial patterns (Fig. 5) that strongly matches the climatic seasons (Fig. 5; accuracy = 68%). While the first three EOMs get clearly and statistically (ANOVA with $p.value < 10^{-3}$; Fig. 6) different scores for the two clusters, ANOVA diagnostics are also statistically significant for EOMs ranked 6, 8, 9 and 10. Finally, the vectors of the average scores per cluster are used to estimate the mean spatial distribution of each cluster (Fig. 4).

4 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. Incidentally, 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 estimate or to simulate values on fixed sampling sites over the study period to get isotopic data base, or after, to map the EOMs if needed. The recourse to random functions and to model of coregionalization help understanding when the factors are spatial factors sensus stricto (Goovaerts 1993; Desbarats and Dimitrakopoulos 2000). 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 het-

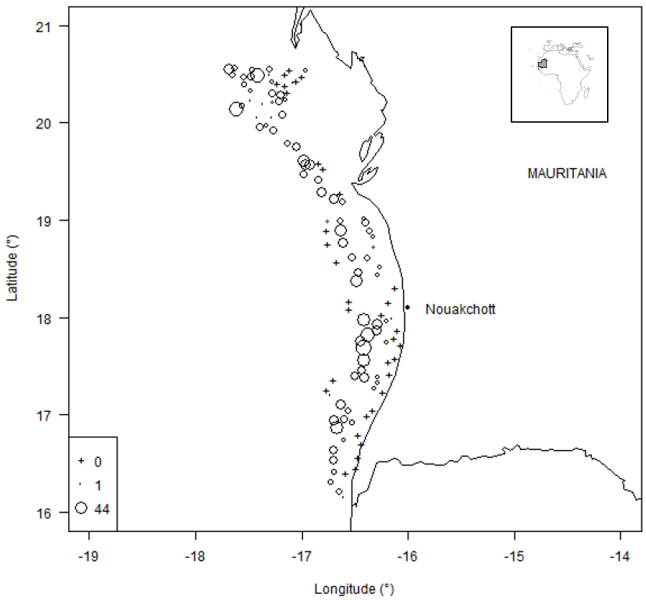

Fig. 1 Sampling protocol. Typical survey data (March, 2015). Octopus densities are in number of individuals per swept area (on average $0.055 km^2$)

erotopic 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 when all covariances are proportional, also called intrinsic correlation model (Chilès and Delfiner 2012; Goovaerts 1993; Desbarats and Dimitrakopoulos 2000). This model is very peculiar and very specific. In the present analysis, we have shown that a weak orthogonality can 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

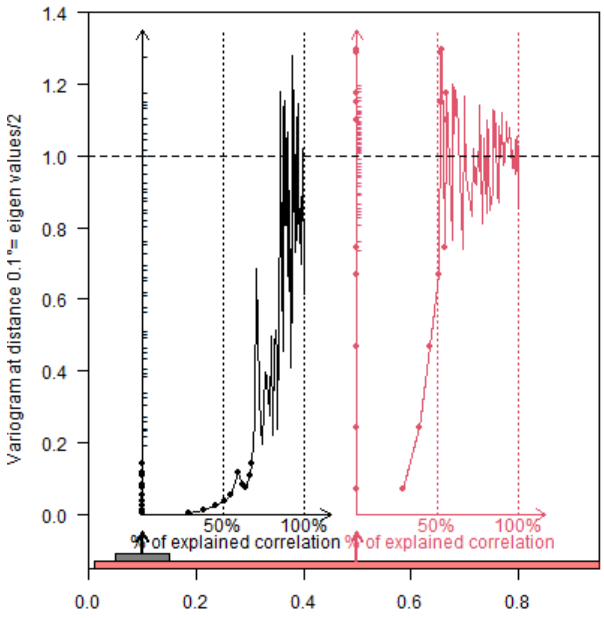

Fig. 2 Values of the omnidirectional variogram of the EOMs for the grid cell distance (0.1°) as a function of the percentage of correlation explained by the r-first EOMs. In black, the EOMs corresponding to a reference distance equal to the grid cell size (h=0.1+/-0.05). In red, the EOMs obtained when h=0.5+/-0.5. The first ten EOMs are depicted by plain circles

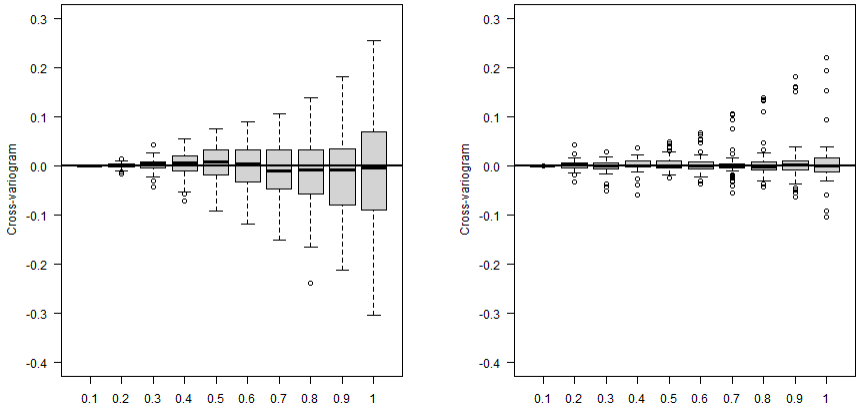

Fig. 3 Cross variogram values for the first distance lags for the first ten EOMs. Left: EOMs used in the study corresponding to a reference distance equal to the grid cell size (h=0.1+/-0.05). Right, EOMs obtained when using a large interval reference distance (h=0.5+/-0.5)

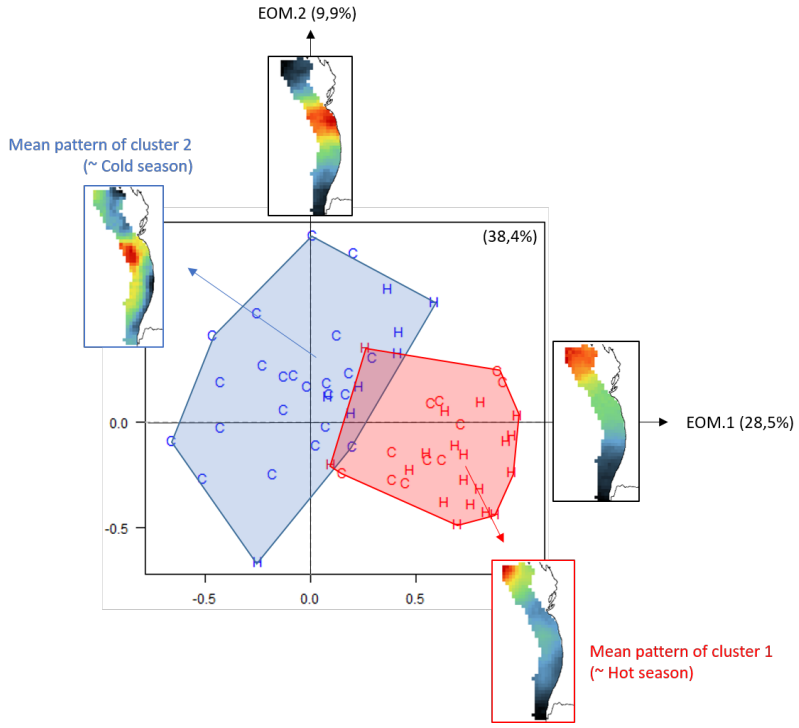

Fig. 4 Factorial space made of the first two EOMs. In this space, the survey are located according to their standardized scores in the EOMs factorization. The two EOMs associated to each dimension of the factorial space are represented. The mean distribution of each group is also represented. Surveys 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

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 auto-correlation 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 spatio-temporal 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 rela-

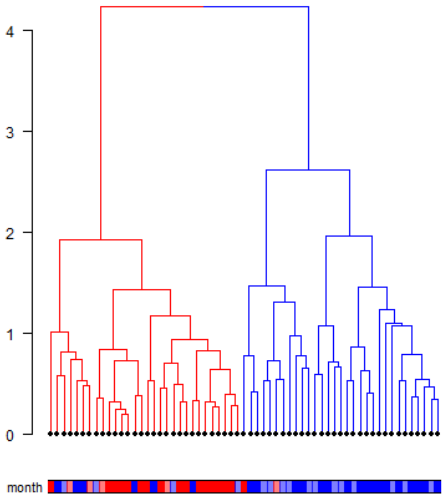

Fig. 5 Dendrogram (Ward distance) of the 61 surveys based on the first ten EOMs. The color bar is defined by the season: in blue, the months corresponding to the cold season (light blue for the two extreme months of the cold season) and, in red, the months corresponding to the hot season (light red for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

tively to the standard deviations amounts to study the spatial patterns relatively to the biomass. Therefore it is crucial 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 performed prior to the analyses. An EOM decomposition together a the dimension reduction make it possible to summarize a large set of distribution maps onto a single factorial plan and to extract meaningful information. When the dispersion of a subgroup of maps is small, this opens the possibility to built a relevant mean distribution map which is of particular importance in ecology. In the present study, 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 corresponds to an a priori knowledge however with no strong empirical foundations. The present comparison and clusering of the spatial distributions strengthens the status of such a knowledge: it is no longer an a priori, but it is now deduced from a long series of spatial distributions.

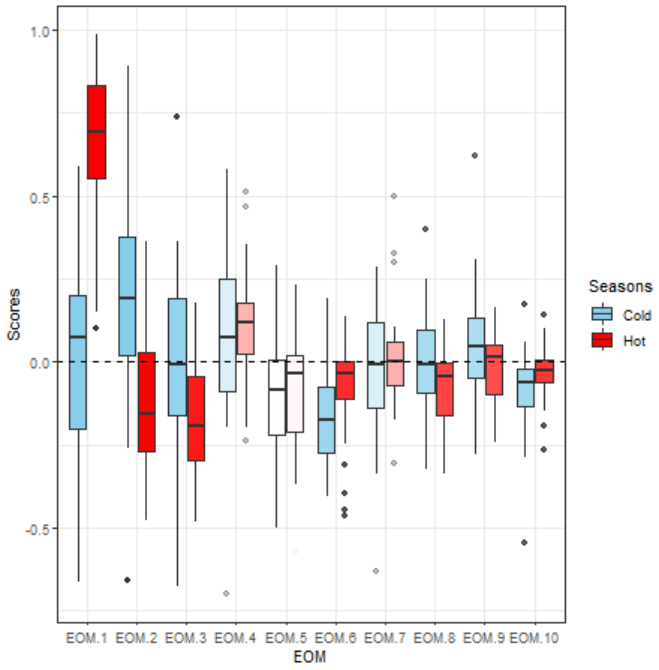

Fig. 6 Boxplots per cluster of the scores of the decompositions of the 61 input spatial distributions on the ten first EOMs. The color's transparency is proportional to the p-value of the difference between means per cluster: the more transparent, the less significant the difference between the two seasons

Acknowledgements 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 application possible.

References

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

Desbarats A, Dimitrakopoulos R (2000) Geostatistical simulation of regionalized pore-size distributions using min/max autocorrelation factors. Mathematical Geology 32(8):919–942

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, 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

Goovaerts P (1993) Spatial orthogonality of the principal components computed from coregionalized variables. Mathematical Geology 25(3):281–302

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(4):423–498

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

MINES-ParisTech/ARMINES (2021) Rgeostats: The geostatistical r package. free download from: http://rgeostats.free.fr/ Version: 12.0.0.

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(4):337–354

Rondon O (2012) Teaching aid: minimum/maximum autocorrelation factors for joint simulation of attributes. Mathematical Geology 44:469–504

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

Solow A (1994) Detecting change in the composition of a multispecies community. Biometrics 50(2):556–565

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

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

Thorson J, Scheuerell M, Shelton A, See K, Skaug H, 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

Vargas-Guzman J, Dimitrakopoulos R (2003) Computational properties of min/max autocorrelation factors. Mathematical Geology 29:715–723

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

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