In press in *Spatial Statistics*

Application of Moran Eigenvector Maps (MEM) to irregular sampling designs

Anik Brind'Amour[1]*, Stéphanie Mahévas[1], Pierre Legendre[2] and Lise Bellanger[3]

[1] Unité Écologie et Modèles pour l'Halieutique, IFREMER, Rue de l'île d'Yeu, B.P. 21105, 44311 Nantes Cedex 03, France. *Corresponding author: Anik.Brindamour@ifremer.fr

[2] Département de Sciences Biologiques, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Québec, Canada H3C 3J7

[3] Laboratoire de Mathématiques Jean Leray UMR CNRS 6629, Université de Nantes, Nantes, France

Abstract

Moran's eigenvectors maps (MEM) are attractive mathematical objects as they are fairly simple to calculate and can be used in most studies of spatially-explicit data. There is, however, an aspect of MEM analysis that still requires some investigation: the effect of irregular sampling on their modeling performance. This study investigates empirically the behavior of MEMs under different irregularity schemes. It is focusing on simulated scenarios representing sampling designs frequently encountered in ecology. We advocate that MEMs can be computed and correctly used with data coming from irregularly designed sampling surveys, given some precautions. We suggest that when the sampling sites are equally spaced but do not cover the entire study area, the MEMs can be computed directly on the coordinates of the sampling sites without any important loss of information. Whereas, when the phenomenon of interest is tackled using randomly stratified sampling designs, the MEMs should be computed on a reconstructed space of regular sampling sites followed by removal of the missing sites, before analysis. This solution of rebuilding a (regular) sampling space guarantees to capture the underlying process under study, improves the modeling results and relaxes the impact of the choice of the weighting matrix on the computation of MEMs.

Key words: autocorrelation, irregular sampling, sampling schemes, spatial analyses, statistical methods

1. INTRODUCTION

Ecosystems in general and marine communities in particular are complex systems composed of a large number of entities interacting with one another at various spatial and temporal scales. Characterization of these scales is an essential step to understand and predict the effects of changes in the processes governing these systems. It relies on mathematical and statistical methods that allow the quantitative description of the spatial and temporal complexity and are sufficiently robust to handle any type of sampling designs. Marine ecological surveys are often irregular (i.e. unevenly spaced) in space or time. Irregularity seems to be the rule rather than the exception. Sampling irregularity may have different causes and consequently display different patterns. In this study, we are dealing with two types of irregularity: i) a "random" irregularity encountered when the phenomenon of interest is tackled using randomly stratified sampling designs or the dataset contains missing sites or time points, and ii) a "constrained" irregularity, when the sampling sites are equally spaced but do not cover the entire study area because topography or other constraints prevent sampling in some sections (i.e. partial coverage).

Observed spatial distributions of species may arise from a plurality of endogenous and exogenous processes (e.g. species interactions, growth, population dynamics, physical forcing) occurring at multiple spatial scales (Vaclavik et al. 2012). This mixture of processes and scales clearly calls for mathematical tools capable of accounting for or modeling such patterns (Dray et al. 2012). Among the statistical methods, the Moran Eigenvector Maps (MEM, Dray et al. 2006) and its original form, the Principal Coordinates of Neighbor Matrices (PCNM, Borcard and Legendre 2002), are good candidates for analyzing such patterns. MEMs are derived from spectral graph theory and characterize a wide range of autocorrelation structures based on the survey design, i.e. the distances between the *n* sampling sites or times (Dray et al. 2006). It is thus a spectral decomposition of the spatial (or

61 temporal) relationships among the sampling sites (or dates). This decomposition generates
62 ($n-1$) *eigenfunctions*, which are new orthogonal variables that can be used in statistical
63 models as explanatory variables representing the spatial or temporal relationships among the
64 study sites.

65 The MEMs and derived approaches have proved very helpful in studying the spatial and
66 temporal distributions of ecological communities (Bellchambers et al. 2011; Brind'Amour et
67 al. 2005; Mikulyuk et al. 2011). However, these studies have been conducted almost
68 exclusively in a context of regular sampling. Although no technical reason prevents the
69 spectral decomposition in a context of irregular sampling sites (i.e. unequally-spaced
70 sampling sites), little is actually known regarding the behavior of the MEMs in such a
71 context. When Borcard and Legendre (2002) first introduced the PCNM approach, they
72 mentioned that irregular sampling schemes affected the amplitude, the phase and the periods
73 of the sine waves generated by the PCNM. They suggested that PCNM developed with
74 irregular sampling sites are suitable descriptors but likely consist of multiple spatial scales,
75 making the interpretation of the spatial descriptors more difficult. Dray et al. (2006)
76 elaborated a little more on the sensitivity of the connectivity matrix in the case of irregular
77 distribution of sampling sites and illustrated the consequences of the sampling irregularity on
78 the number of positive/negative eigenvalues and the spatial structures described by the
79 associated eigenvectors. They came to the conclusion that sampling irregularity, defined
80 through the spatial relationships among neighboring sites, may have substantial impact on the
81 behavior and interpretation of the MEMs. Ecological surveys are designed to study processes
82 overriding the spatial and temporal distribution of species. However, in some cases it is
83 difficult or impossible to sample over the whole area where a specific ecological process
84 occurs. For instance, recruitment of a species may take place in shallow coastal areas where
85 the draft of the boat prevents access. In that case, the observation scale (study extent) at which

sampling is conducted is smaller than the ecological scale at which the process occurs. Such a mismatch between the observation and process scales may also have a significant effect on the interpretation of the MEM.

In cases of regular sampling with some points of the grid missing, it is common practice to develop the MEMs on the original matrix of sampling coordinates of the irregularly distributed sites (Fuentes-Rodriguez et al. 2013; Jombart et al. 2008, Sattler et al. 2010, Mikulyuk et al. 2011, Sharma et al. 2011, Vaclavik et al. 2012,). However, as suggested by Borcard and Legendre (2002), one can develop the MEMs on a transformed matrix of coordinates that has been filled with supplementary sampling sites to make it regular; by construction MEMs are orthogonal to one another. The added sites are then removed after the MEM have been computed (i.e. rows of the eigenvector matrix; Blanchet et al. 2013 and Borcard et al. 2004) . This procedure presents however the disadvantage of loosing practical mathematical properties of the MEMs: the orthogonality among the MEMs and the maximization of spatial autocorrelation (Moran's I). The choice of filling or not the voids with supplementary sites, prior to the computation of the MEMs, and the number of supplementary sites needed to attain a sufficiently fine resolution without major loss of orthogonality, remain key questions in the development and interpretation of MEMs.

This study aims at empirically shedding light on some key questions about the development and interpretation of MEMs: Is the MEM approach relevant with irregular sampling designs? Do the MEMs developed with irregular sampling sites capture the spatial scales they are supposed to capture? Is there an irregularity threshold beyond which MEMs cannot be safely used? Is there a solution to counteract the problems caused by sampling irregularity? Should we compute MEMs using supplementary sites, or not? Using simulations, we empirically investigated the impact of various irregular sampling schemes (irregular distribution of sites and partial coverage of a study are a) in the development and

111 interpretation of the MEMs. We focused on a selection of scenarios characterizing sampling
112 designs frequently encountered in ecological studies.

113 ## 2. METHODS AND DATA

114 ### *2.1 Simulation approach*

115 Investigation of the MEM behavior with regard to sampling irregularity was done by
116 testing the ability of the MEMs to correctly detect spatial structures commonly observed in
117 ecology under three different scenarios of sampling irregularity. The simulations were
118 designed to mimic sampling strategies frequently encountered in ecological studies. They
119 were based on three components that may affect the computation and interpretation of the
120 MEMs: i) irregularity of the sampling sites produced by sub-sampling the sampling zone (i.e.
121 random sampling), ii) irregularity of the sampling sites generated by partial coverage of the
122 sampling zone (i.e. blocks of unsampled sites), and iii) the process-observation mismatch, i.e.
123 whether the observations match or not the scale of the ecological process under study. We
124 focused on three combinations of these components that we called scenarios (see below).
125 Each scenario is thus answering a specific question regarding irregularity in sampling
126 strategies. All simulations were replicated 100 times by modifying the spatial coordinates of
127 the samples. Precisely, we randomly sampled over the complete grid the number of cells
128 corresponding to the subsampling thresholds.

129 In each scenario, we compared two approaches for computing MEMs (Fig. 1):

130 - The complete-grid approach (MEM_{comp}): compute the MEM from the coordinates of
131 all sites in the full grid, including the unsampled points, then remove from the MEM
132 matrix the rows that correspond to the unsampled sites.
133 - The reduced-grid approach (MEM_{red}): use only the geographic coordinates of the
134 sampled sites to construct the MEM data matrix. That approach is commonly used in
135 the literature.

136 *2.1.1 Scenario 1: Random sampling design* (S1)

137 The first scenario tested the ability of the MEM to correctly capture the spatial structure
138 in the case of a random sampling strategy. Random sampling was assessed at eight different
139 thresholds (25, 50, 60, 70, 80, 90, 100%), representing the percentage of the studied area
140 covered by the sampled sites. Although the lower thresholds (25%) may seem quite small,
141 they are representative of what ecologists use in randomly stratified scientific surveys
142 (Brind'Amour et al. 2014). In this scenario, irregularity was random and created unequal
143 distances between neighboring sampled sites (Fig. 2). The process under study matched the
144 observation scale and occurred at "global" scale, that is, over the entire zone under study. In
145 that scenario, the distances between neighboring sites increased in irregularity as the threshold
146 decreased.

147 *2.1.2 Scenario 2: Blocks of missing data* (S2)

148 The second scenario was developed to test the ability of MEMs to correctly capture the
149 spatial structure when the survey includes blocks of missing observations, such as
150 inaccessible areas (Fig. 2; Jones et al. 2008). In this scenario, the sampled sites covered 50%
151 of the whole area. They were regularly spaced, but as the area was greatly reduced, the
152 distances between sites were irregular. The process under study occurred at "global" scale and
153 did not match the observation scale.

154 *2.1.3 Scenario 3: Random sampling design and blocks of missing data on a global structure*
155 (S3)

156 This scenario tested the capacity of MEMs to capture the spatial structure when the
157 survey was randomly designed and included blocks of missing observations (as in the second
158 scenario). In that scenario, the sampled area covered 50% of the total area and we tested the
159 effect of different sampling thresholds on the MEMs (Fig. 2). The process occurred at
160 "global" scale and did not match the observation scale.

161 *2.2 MEM computation*

162 There are various ways of computing MEMs (see Dray et al. 2006 and Legendre and
163 Legendre 2012 for details). In our simulations, we computed both MEMs and db-MEMs
164 following the steps using the packages *spacemake R* (Dray 2013) and *spdep* (Bivand et al.
165 2013, 2015) in R (R Core Team 2014). It is worth mentioning that the package *adespatial*
166 recently developed by Dray et al. (2016) can now compute the MEMs and db-MEMs.
167 Differences between the two types of MEMs are summarized in Appendix A1. The two types
168 were calculated on a matrix of 20 x 20 = 400 sites for the three scenarios. They were
169 computed using a distance matrix transformed into a similarity matrix (Legendre and
170 Legendre 2012, p.861) weighed by a connectivity matrix (see details in Appendix A1).

171 *2.3 Predefined spatial structures*

172 Different ways can be used to simulate spatial structures. For instance, one can use the
173 MEMs themselves to generate response values or simulate independent geostatistical
174 distributions. In here, we simulated the spatial structures using the MEM themselves and
175 using empirical variograms with various ranges to modify the degree of spatial
176 autocorrelation. The use of the MEMs themselves was done as an "experiment" to verify if we
177 could correctly capture the predefined MEM as a spatial structure. With that approach we
178 were expecting to capture perfectly the modeled spatial structure with the MEM_{comp} given that
179 the same MEM served as the response and it was also included in the set of explanatory
180 variables. The predefined spatial structures were created from MEMs computed on a 2D
181 regular grid of 20 by 20 cells (n = 400 cells). For the three scenarios, we selected four MEMs
182 from the entire set produced (MEM #1 called MEM01, 10, 150, 350; Fig. 3). In the three
183 scenarios, the selected MEMs were analyzed as separate response variables (i.e. single
184 variable), corresponding to a gradient of spatial structures varying from coarse to very fine
185 spatial scales. Coarse spatial scale was characterized by large positive eigenvalues, medium

186 spatial scale by intermediate positive eigenvalues, fine spatial scale by small negative
187 eigenvalues, and very fine spatial scale by large negative eigenvalues. For each
188 studied j^{th} MEM, called Y^j, representing a selected predefined spatial structure, a random
189 noise was added using values sampled from a normal distribution with mean of 0 and σ of
190 0.05, N(0,σ).

191 The second type of predefined spatial structures were developed using geostatistical
192 distributions. The four predefined spatial structures modeled corresponded to a gradient of
193 spatial scales varying from coarse to very fine spatial scales (Fig. 3). Empirical variograms
194 were developed using spatially correlated random fields computed on a 2D regular grid of 20
195 by 20 cells (n = 400 cells) followed by unconditional Gaussian simulations (Pebesma 2004).
196 The variogram (γ) was modeled using a spherical model (Cressie 1993, for more details):

197 $$\gamma(h) = \frac{c}{2}\left(\frac{3h}{a} - \frac{h^3}{a^3}\right) \qquad h \leq a \qquad (1)$$
198 $$\gamma(h) = c \qquad h > a$$
199

200 where γ(h) is the variogram value at distance h, a is the range, i.e. the distance, h, beyond
201 which the autocorrelation is presumably zero, and c is the sill, i.e. the value at which the
202 variogram levels off. For the simulations we arbitrarily fixed c at a value equal to 5 and varied
203 the range from a coarse (a = 20 cell length), to medium (a = 5), to fine (a = 3), and very fine
204 spatial autocorrelation structure (a= 1). To simulate the four spatial structures we fixed the
205 three beta coefficients to zero (i.e. no linear trend) and as mentioned above we varied the
206 range. The simulations were repeated 100 times per spatial structure using R package *gstat*
207 (Pebesma 2004).

208 *2.4 MEM evaluation*

209 The predefined spatial structures were used as response variables Y in linear regression
210 models to evaluate the MEM behavior. We compared the two following:

$$Y_i^{(j)} = m + \left(\mathrm{MEM}_{comp}\right)_i \times \beta_{comp}^{(j)} + \varepsilon_i^{(j)} \text{and,} \quad (2)$$

$$Y_i^{(j)} = m + (\mathrm{MEM}_{red})_i \times \beta_{red}^{(j)} + \varepsilon_i^{(j)} \quad (3)$$

where $\varepsilon_i^{(j)} \sim \mathrm{N}(0; \sigma)$ and $Y_i^{(j)}$ is fixed to be either one of the four MEMS or variogram simulations, in the three scenarios. n is the number of observations and depends on the number of missing observations. $\mathrm{MEM_{comp}}$ $(\mathrm{resp.\,MEM_{red}})$ is the matrix of predictor variables, $(MEM)_i$ is a line vector containing the i^{th} line of matrix MEM, β is the column vector of parameters (regression coefficients) to be estimated, and m is the intercept.

The ability of MEM analysis to correctly capture the predefined spatial structures was evaluated using five criteria: the number of significant $\mathrm{MEM_{comp}}$ and $\mathrm{MEM_{red}}$ in each scenario, the adjusted R^2 of the fitted models, and the collinearity (estimated by Pearson correlation coefficients) between the members of the subset of $\mathrm{MEM_{comp}}$ considered as predictors. The first criterion (i.e. number significant MEM) was obtained from multiple regression analyses after forward selection between the response variable Y and the explanatory MEM (spatial descriptors). Given that (n–1) MEMs are generated, the choice of a method to correctly select significant MEMs with regard to the overestimation of the variance explained is an important issue that has been discussed (Blanchet et al. 2008). In here, we used a forward selection analysis based on a permutation procedure (*forward.sel* function) developed by S. Dray in the R package *packfor*. It follows the recommendations of Blanchet et al. (2008) and Munoz (2009) to split the regression model in different parts (or submodels) to circumvent the problem of over parameterization. In here we divided MEMs into four submodels corresponding to a gradient of spatial structures varying from coarse to very fine spatial scales. Coarse spatial scale was mainly characterized by large positive eigenvalues, medium spatial scale by intermediate positive eigenvalues, fine spatial scale by small negative eigenvalues, and very fine spatial scale by large negative eigenvalues. The significance of regressions was tested as suggested in Blanchet et al. (2008), by applying a forward selection

on each submodel with a double stopping rule (i.e. α threshold and a maximum threshold for the global model). This procedure controls for type I error inflation. The predicted values were estimated using linear regression models by fitting the significant MEM (previously identified) to the response variable *Y* (i.e. predefined spatial structure).

3. RESULTS

3.1 Scenarios

The simulations were conducted using MEMs and db-MEMs. As no difference was found between the two types of MEMs, only the results with MEM, generalization of db-MEM, are presented here for two extreme cases of the predefined spatial structures (Fig. 3): the coarser and finer spatial scales using the MEMs (MEM01 and MEM350) and the variogram simulations (ranges = 20 and 3). The Appendix B contains results for all the spatial scales (including those presented here) calculated for the two types of spatial structures (MEMs and variograms) and the three scenarios.

3.1.1 Scenario 1: Random sampling design (S1)

MEM spatial structures – Comparison of the MEM_{comp} and the MEM_{red} suggests that the first approach globally outperformed the second (Fig. B1, Appendix B). The MEM_{comp} always captured the predefined spatial structure and showed adjusted R^2 always above those of MEM_{red} (Fig. 4a and 4d). That result stands for the sub-sampling thresholds that could be tested (i.e. > 50%) independently of the spatial scale. On the other hand, when the MEMs are computed directly from the reduced matrix of coordinates (the reduced approach), it takes between 2 to 7 MEMs to capture the predefined spatial structure (regardless of the sub-sampling threshold). That approach of computing MEMs succeeds in modeling the coarse predefined spatial structure with adjusted R^2 comparable to those obtained for the MEM_{comp} but fails in capturing fine spatial structures with adjusted R^2 varying between 0 and 0.5.

With the MEM_{red} the property of orthogonality is preserved and the MEMs are thus uncorrelated to one another. This property also holds for all sub-sampling thresholds. This is not the case for the MEM_{comp} where orthogonality is lost when the missing sampling sites are removed. Nevertheless, the correlation coefficients among the MEM_{comp} are very low as they never reach values higher than 0.12 (Fig. B1, Appendix B).

Variogram simulations – Results of the modeling of the variogram simulations using the MEM_{comp} and the MEM_{red} are in line with the previous results using the MEM as spatial structures (Fig. B4, Appendix B): the MEM_{comp} slightly outperformed the MEM_{red}. The MEM_{comp} always captured the predefined spatial structure and showed adjusted R^2 almost always above those of MEM_{red} (Fig. 5a). However, as shown previously, at low sampling thresholds (i.e. < 50%) when the number of explanatory variables equals or exceeds the number of sites, the selection procedure stops and no MEM_{comp} is included in the models (Fig. 5d), thereby lowering the adjusted R^2 to 0. On the other hand, when the MEMs are computed directly from the reduced matrix of coordinates (the *reduced* approach), it takes between 10 to 40 of the generated MEMs to capture the predefined spatial structure. That approach of computing MEMs succeeds in capturing the predefined spatial structure below 50% but displays higher variability at low sampling thresholds. For the two approaches, the effect of sub-sampling on the global fit of the predefined spatial structure grows worst as the scale of the spatial structure decreases. For instance, when the structures are characterized by coarse spatial scales, the adjusted R^2 stabilizes at ~ 0.85 for all the sub-sampling thresholds above 50% of the sampled area (Fig. 5a). When spatial structures are defined at fine spatial scales, MEM_{comp} shows on average a better fit of ~ 10%. The two approaches fail in capturing the spatial structures at low sampling thresholds.

283 *3.1.2 Scenario 2: Blocks of missing data (S2)*

284 *MEM spatial structures.* – When the spatial structures occur at coarse (MEM01 and MEM10)
285 spatial scales, the MEM_{red} and MEM_{comp} give similar results (Fig. B2,Appendix B). The two
286 approaches slightly differ when the spatial structures are at fine scales (MEM ≤150). In these
287 cases, it takes on average 5 to 15 MEM_{red} to detect the predefined spatial structure but it never
288 succeeds in modeling the spatial structure as efficiently as with MEM_{comp} (Fig. 4b). Indeed,
289 when the MEM_{comp} are used, they systematically captured the modeled MEM and showed
290 adjusted R^2 values 5 to 15% higher than when using the MEM_{red}. In these cases, the
291 collinearity induced by removing the supplementary sites is always well below 0.1% (Fig. B2,
292 Appendix B).

293 *Variogram simulations.* – When the predefined spatial structures are developed using
294 variogram simulations, the results are very similar to those presented above, i.e. the MEM_{red}
295 and MEM_{comp} give similar results notwithstanding the simulated spatial scales (Fig. 5b and
296 Fig. B5, Appendix B). They both show decreasing adjusted R^2 and increasing uncertainty at
297 medium and fine spatial structures.

298 *3.1.3 Scenario 3: Random sampling design and blocks of missing data on a global structure*
299 *(S3)*

300 *MEM spatial structures.* – Results for that scenario are similar to those obtained in S1: the
301 MEMcomp clearly outperformed the MEMred (Fig. 4c and 4f). In contrast to S1, the results
302 in S3indicate a stronger impact of increasing irregularities (i.e. increasing sub-sampling) on
303 the capacity of the MEM to correctly detect the predefined spatial structure (Fig. B3,
304 Appendix B). The MEM_{comp} systematically selected the predefined spatial structure (Fig. 4f)
305 and reached better fits than the MEM_{red}, regardless of the level of sub-sampling and the nature
306 of the predefined spatial structure (i.e. coarse to very fine structures; Fig. B3 in Appendix B).
307 The effect of irregularity is most obvious using the MEM_{red} and notably when modeling fine

spatial structures. In that case, the adjusted R^2 drops more rapidly than it did in S1 and never reaches values above 0.3 (Fig. 4c). The counterpart of using MEM_{comp} is emphasized by the collinearity, which sometimes reaches values equal to 0.1 (Fig. B3, Appendix B).

Variogram simulations.–Results for that scenario are very similar to those obtained in *S1*: the MEM_{comp} slightly outperformed the MEM_{red} (Fig. 5c and 5f). In contrast to *S1*, the results in *S3* indicate a stronger difference between the two approaches on the capacity of the MEM to correctly detect the spatial structure (Fig. B6, Appendix B). The MEM_{comp} selected the simulated spatial structure (Fig. 5c) and reached better fits between 10 to 40% in comparison to the MEM_{red}, at levels of sub-sampling above 50% (Fig. B6 in Appendix B). As in *S1*, at low sampling thresholds (i.e. < 50%), very few MEM_{comp} are included in the models whereas 5 to 10 MEM_{red} are selected to reach Adjusted R^2 varying between 0.70 (coarse spatial scales) and 0.1 (fine spatial scales).

4. DISCUSSION

The number of studies using the MEM approach and its derivatives has more than doubled in recent years (Bellchambers et al. 2011; Blanchet et al. 2013; Fuentes-Rodriguez et al. 2013; Jombart et al. 2008; Mikulyuk et al. 2011;Sattler et al. 2010; Sharma et al. 2011; Sharma et al. 2012; Vaclavik et al. 2012); the original papers describing the method received hundreds of citations listed on Web of Science and Google Scholars. Most of these applications used irregular sampling designs. However, very few of them have actually discussed the effect of irregularity on the development and interpretation of the MEM (Blanchet et al. 2013 and Borcard et al. 2004). Our study aimed at investigating empirically the capacity of MEM analysis to correctly identify predefined spatial structures at various spatial scales, under different scenarios of irregularity. This was done to help ecologists use the full potential of the MEM approach in ecological modeling. We suggest to develop the MEMs on a regular sampling grid, followed by removal of the missing sites. We also warn

against sampling irregularity when the sampling sites cover a low proportion of the studied area and when one wishes to model ecological processes occurring at very fine spatial scales.

4.1 MEM_{red} or the common way of computing the MEMs

Our study tested the performance of two ways of computing MEMs (MEM_{comp} and MEM_{red}) to correctly captured different predefined spatial structures. This was done with the objective of comparing the commonest way of computing the MEM (MEM_{red}) with another less common approach (MEMcomp). We used two types of predefined spatial structures, one based on the MEM themselves, that can seen as tautological (or dependent) with the MEM_{comp}, and another one using independent spatial structures. We considered the primer predefined spatial structure as a "controlled situation" where we were expecting to capture perfectly the pattern using the MEM_{comp}. In that sense, the MEMcomp responded as expected and gave almost a perfect fit notwithstanding the scenario and the spatial scales. It was more a less a test for the MEM_{red} as most of the studies using the MEMs are developing the MEMs directly on the sample sites without filling the voids (e.g. Mikulyuk et al. 2011; Sattler et al. 2010). For that type of MEMs construction, results were generally considered good or lukewarm at broader and finer spatial scales respectively.

The independent predefined spatial structures (i.e. variograms) were used as comparison between the two types of MEMs. When the selection procedure allowed the MEM_{comp} to be computed (threshold > 50%), they MEMs showed between 10 to 15% better fit than the MEM_{red}. At low thresholds, i.e. when the proportion of sites sampled is low given the surface of the studied zone (< 50% or less) and broad spatial structures are expected, we suggest that the MEM_{red} can be safely used.

4.2 Irregularity: effect of random design vs blocks of missing data

In this study we showed that removing blocks of sampling sites (e.g. scenario *S2*) was less harmful to the conclusions than randomly removing the same number of sites (50%

threshold in *S1*) in a regularly-spaced design. This can be explained by the fact that the proportion of regular distances among the sites in the first case is kept relatively high in comparison to the second case where any distance can be eliminated. Recent studies using the MEMs (Bellchambers et al. 2011; Fuentes-Rodriguez et al. 2013; Mikulyuk et al. 2011; Sattler et al. 2010; Sharma et al. 2011; Vaclavik et al. 2012) fell in our *S2* and *S3* scenarios with varying sub-sampling thresholds (all below 40%). They developed the MEMs directly on the coordinates of the sampling sites without adding supplementary sites. Given the results of our simulations, these studies might have missed some spatial scales of variability and presumably underestimated the importance of the predictors in terms of their contributions to the overall goodness of fit of their models. While this shortcoming most likely did not affect the spatial patterns they observed, it might have had some influence on the relative contributions of the spatial components they estimated in their variance partitioning analyses (Fuentes-Rodriguez et al. 2013; Mikulyuk et al. 2011; Sattler et al. 2010).

Munoz (2009) developed and tested a smoothing model to select significant distance-based eigenvector maps (DBEM, a particular case of MEM), on a regular and irregular sampling designs. He found no differences between the two designs and concluded that the DBEM approach was highly suitable for analyzing ecological surveys. Munoz results cannot be compared directly with ours, as his smoothing model does not keep, by definition, individual elements (i.e. MEMs) but rather combines them in sub-models using smoothing windows. Notwithstanding this difference, our results showed that at a low sub-sampling threshold (>90% of the sites kept) the models developed using the MEM_{red} produced similar but not as good results as those developed with the MEM_{comp} approach. In a sense, this is in agreement with what Munoz observed in his work as his regular and irregular sampling schemes were composed of the same number of sites (2500 points) and only differed in their

spatial positions. Whether the conclusions of Munoz (2009) would still hold under different sub-sampling thresholds and other irregular schemes remains an open question.

4.3 Rebuilding regularity: an efficient solution

When they first introduced the PCNM method, Borcard and Legendre (2002) suggested to outwit the problem of missing observations by adding geographic coordinates in the dataset prior to MEM computation. The solution used in here slightly differs from theirs, as we are filling the voids as they suggested, but we are adding supplementary sites to mimic a regular sampling scheme. As advocated by Dray et al. (2006, p.487), the choice of a spatial weighting matrix **W** is crucial in the computation of MEMs and in the case of regular sampling, the structures defined by the eigenvectors (i.e. MEMs) are less sensitive to the choice of **W**. Therefore, recreating a regular sampling matrix offers the advantage of allowing the computation of the MEMs using any neighboring relationships in **W**, in addition to keeping a fine spatial resolution among the sites. In here we rebuilt a "rectangular cuboid " grid by using the maximum and minimum values on the X and Y axes. This way of recreating a complete and regular sampling grid may not always be the optimal technique, particularly when the sampling zone has the shape of a "rectangular parallelepiped". In that case, our technique may artificially expand the sampling zone and thus the number of MEM. We suggest that special care should be taken when developing the complete sampling grid (i.e. cell size and shape of the total extent). On the other hand, if one decide to use the MEMs that are computed directly on the matrix with missing observations, the choice of the weighting matrix should be optimized (Dray et al. 2006).

Building the MEMs on a reconstructed matrix of regular sampling has two drawbacks. First, it introduces correlations among the MEM, thus losing, to a certain degree, the orthogonality property of the MEM. This was already pointed out by Borcard and Legendre (2002) and Borcard et al. (2004, p. 1828). In here we confirm that statement: with irregular sampling

surveys, one has to accept the compromise of losing the appealing property of orthogonality in the modeling process. Second, the selection procedure (if correctly applied) stops when the number of variables equals or exceeds the number of sites. This situation can be circumvented by maintaining the number of variables lower than the number of sampling sites by, for instance, dividing the MEMs selection in different submodels and correcting accordingly for type 1 error. In This situation cannot happen with the MEM_{red}, because their number will always be less than the total number of sites and they in our study, they may captured a spatial structure notwithstanding the sub-sampling level. In cases where sampling irregularity is very high and induces strong correlations between MEM_{comp}, one may use MEM_{red}. However, in such cases the MEM approach may not be the most appropriate modeling technique, although alternatives are scarce (*e.g.* Empirical Orthogonal Functions, Kutzbach 1967).

In this study, we explored the efficiency of rebuilding a regular sampling grid prior to calculation of the MEMs by testing the ability of the "reconstructed" MEMs to correctly capture the different predefined spatial structures. Application of such a solution indicated that the predefined spatial structure was identified and showed very good adjustment for all the scenarios, using an appropriate statistical selection procedure. That solution also succeeded well in modeling the various spatial structures tested, from coarse to medium spatial scales. Rebuilding a regular sampling grid has an interesting advantage of assessing the inter-annual comparison of spatial structures using randomly stratified sampling designs. Given that the reconstructed grid is common to all the sampling years, the spatial analyses can thus be done on the same basis thereby allowing direct comparison of the spatial scales among years.

4.4 Strong effect of irregularity with fine spatial structures.—Whether very small and negative eigenvalues should be included in a modeling process remains an open debate (Munoz 2009) and is beyond the scope of this study.

Nevertheless, our simulations and other authors suggest that it would be probably safer to discard them in highly irregular sampling surveys. Indeed, the MEMred were less efficient in capturing the finer spatial scales, as showed their lower R^2. The failure of the two MEMs to capture the fine spatial scales using the variogram simulations (i.e. range equal to 1) can be explained by the choice of the grid size that we used (distance between two sites equals 1). This underlines that when the scale of the pattern is smaller than the sampling design, we are not able to detect a signal.

5. CONCLUSION

Our simulations dealt with relatively simple spatial structures (i.e. simulations with various ranges or predefined MEMs), while in most ecological studies, the spatial distribution of species is more complex and varies over a wide range of spatial scales. We tested two ways of computing the MEMs and conclude that both approaches can be used. Nevertheless some precautions must be taken to prevent their misuse. When the MEMs are computed on the "*complete*" or reconstructed space of the sampling sites prior to analysis, we suggest that the sets of eigenvectors with positive and negative eigenvalues can be used safely together in further analyses, given that a relevant selection procedure of significant variables is used. In our simulations no difference was found between the computation of MEMs or db-MEMs. In all scenarios, developing MEM over the complete area and subsequently reducing them to fit the sampling design created correlations among the MEMs (i.e. non-orthogonal eigenvectors), however, the values of the correlations were low (maximum of 0.10 in absolute value) and did not preclude the use of the MEM_{comp} as spatial descriptors in (partial) regression or canonical analyses. The importance of the correlations among the MEMs in highly complex ecosystems remains to be tested. For that particular aspect we call upon mathematicians to study the properties of the MEM_{comp} in a reduced sampling design and particularly the loss of equivalence between eigenvalues and Moran's *I*. We also showed that MEMs can be

computed directly on the coordinates when blocks of sites are missing, without any significant loss of information, and correctly interpreted if the process under study matches the scale of observation, which is generally the case. However, when the MEMs are computed directly using the spatial coordinates, special care should be taken in defining a relevant connectivity matrix and thus choosing appropriate neighboring relationships. The developments in here were applied in a spatial context, although similar conclusions could likely be drawn for temporal analyses.

Acknowledgments

This work was carried out under the project COSELMAR funded by the Regional Council of the Pays de la Loire. The authors would like to thank the scientists and crews who participated in the NURSE surveys in the Bay of Vilaine nursery grounds between 2008 and 2010.The authors acknowledge Stéphane Dray for his review and his comments that greatly improved the manuscript.

REFERENCES

Bellchambers, L.M., Meeuwig, J.J., Evans, S.N., Legendre, P., 2011. Modelling habitat associations of the common spider conch in the Cocos (Keeling) Islands. Mar. Ecol. Prog.Ser. 432, 83-90

Blanchet, F.G., Bergeron, J.A.C., Spence, J.R., He, F., 2013. Landscape effetcs of disturbance, habitat heterogeneity and spatial autocorrelation for a ground beetle (Carabidae) assemblage in mature boreal forest. Ecography 36, 636-647

Blanchet F. G., Legendre, P., Borcard D., 2008. Forward selection of explanatory variables. Ecology 89, 2623-2632

Borcard, D., Legendre, P., 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. Ecol. Model. 153, 51-68

Borcard, D., Legendre, P., Avois-Jacquet, C., Tuomisto, H., 2004. Dissecting the spatial structure of ecological data at multiple scales. Ecology 85, 1826-1832

Brind'Amour, A., Boisclair, D., Legendre, P., Borcard, D., 2005. Multiscale spatial distribution of a littoral fish community in relation to environmental variables. Limnol. Oceanogr. 50, 465-479

Brind'Amour, A., P., L., J., M., S., V., Fovau, A., Le Bris, H., 2014. Morphospecies and taxonomic sufficiency of benthic megafauna in scientific bottom trawl surveys. Cont. Shelf Res. 72, 1-9

Dray, S., Legendre, P., Peres-Neto, P.R., 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices. Ecol. Model. 196, 483-493

Dray, S., Pélissier, R., Couteron, P., Fortin, M.-J., Legendre, P., Peres-Neto, P.R., Bellier, E., Bivand, R., Blanchet, F.G., De Cáceres, M., Dufour, A.-B., Heegaard, E., Jombart, T., Munoz, F., Oksanen, J., Thioulouse, J., Wagner, H.H., 2012. Community ecology in the age of multivariate multiscale spatial analysis. Ecol. Monogr. 82, 257-275

Dray, S., Blanchet, G., Borcard, D., Guenard, G., Jombart, T., Larocque, G., Legendre, P., Madi, N, Wagner, H.H., 2016. adespatial: Multivariate Multiscale Spatial Analysis. R package version 0.0-7. http://CRAN.R-project.org/package=adespatial

Fuentes-Rodriguez, F., Juan, M., Gallego, I., Lusi, M., Fenoy, E., Leon, D., Penalver, P., Toja, J., Casas, J.J., 2013. Diversity in Mediterranean farm ponds: trade-offs and synergies between irrigation modernisation and biodiversity conservation. Fresh. Biol. 58, 63-78

Jombart, T., Devillard, S., Dufour, A.-B., Pontier, D., 2008. Revealing cryptic spatial patterns in genetic variability by a new multivariate method. Heredity 101, 92-103

Jones, M.M., Tuomisto, H., Borcard, D., Legendre, P., Clark, D.B., Olivas, P.C., 2008. Explaining variation in tropical plant community composition: influence of environmental and spatial data quality. Oecologia 155, 593-604

Kutzbach, J.E., 1967. Empirical eigenvectors of sea-level pressure, surface temperature and precipitation complexes over North America. Journal of Applied Meteorology 6, 791-802

Legendre, P., Legendre, L., 2012. Numerical Ecology, 3rd English edition. Elsevier Science BV, Amsterdam.

Mikulyuk, A., Sharma, S., Van Egeren, S., Erdmann, E., Nault, M.E., Hauxwell, J., 2011. The relative role ef environmental, spatial, and land-use patterns in explaining aquatic macrophyte community composition. Can. J. Fish. Aquat. Sci. 68, 1778-1789

Munoz, F., 2009. Distance-based eigenvector maps (DBEM) to analyse metapopulation structure with irregular smapling. Ecol. Model. 220, 2683-2689

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Comp.& Geosc. 30, 683-691

522 R Development Core Team, 2014. R: A language and environment for statistical computing.
523 R Foundation for Statistical Computing, Vienna, Austria

524 Sattler, T., Borcard, D., Arlettaz, R., Bontadina, F., Legendre, P., Obrist, M.K., Moretti, M.,
525 2010. Spider, bee, and bird communities in cities are shaped by environmental control
526 and high stochasticity. Ecol. 91, 3343-3353

527 Sharma, S., Legendre, P., Boisclair, D., Gauthier, S., 2012. Effects of spatial scale and choice
528 of statistical model (linear versus tree-based) on determining species-habitat
529 relationships. Can. J. Fish. Aquat. Sci. 69, 2095-2111

530 Sharma, S., Legendre, P., De Caceres, M., Boisclair, D., 2011. The role of environmental and
531 spatial processes in structuring native and non-native fish communities across
532 thousands of lakes. Ecography 34, 762-771

533 Vaclavik, T., Kupfer, J.A., Meentemeyer, R.K., 2012. Accounting for multi-scale spatial
534 autocorrelation improves performance of invasive species distribution modelling
535 (ISDM). J. Biogeogr. 39, 42-55

536

Figure captions

Fig. 1. Flowchart of the methodology used to develop the *reduced* and the *complete* MEM sets in this study. Texts in bold indicate the final dimensions of the MEM matrix. The full grid in this example includes 72 sites whereas in the real study, we used 200 sites (see Methods for details).

Fig. 2. (a) Schematic 2D illustration of the four scenarios tested in this study. X and Y axes represent longitude and latitude respectively. The gray scale corresponds to a predefined spatial structure used as an example of a response variable ***Y*** in this study. Detailed description is found in text with the corresponding scenario.

Fig. 3. Illustration of the two types of predefined spatial structures used in the study, corresponding to a gradient of spatial scales varying from coarse to very fine spatial scales. The spatial structures (a) matched four selected eigenvectors or MEMs, or (b) were simulated using geostatistical distributions with varying ranges (see Methods for details).

Fig. 4. MEM simulations. Comparison of the two approaches to compute MEMs in the three scenarios. By column, (a-d) display the adjusted R^2 and (e-h) the number of MEM_{comp} or percentage of MEM_{red} required to model the various predefined spatial structures at different sampling thresholds (abscissa for S1 and S3: 10, 25, 50, 60, 70, 80, 90, 100%). Results are presented for the two most contrasted spatial structures (coarser spatial scale: filled circles, black; finer spatial scale: empty squares, grey). Solid lines correspond to MEM_{comp} while the dashed lines correspond to MEM_{red}. Intervals correspond to plus and minus the standard deviation estimated from the 100 replicated simulations. Results for all simulations are presented in Appendix B (Fig. B1 to B3).

562

563 Fig. 5. Variogram simulations. Comparison of the two approaches to compute MEMs in the
564 three scenarios. By column, (a-c) display the adjusted R^2 and (d-f) the number of MEM
565 required to model the various predefined spatial structures at different sampling thresholds
566 (abscissa: 25, 50, 60, 70, 80, 90, 100%). Results are presented for the two most contrasted
567 spatial structures (coarser spatial scale: filled circles, black; finer spatial scale: empty squares,
568 grey). Solid lines correspond to MEM_{comp} while the dashed lines correspond to MEM_{red}.
569 Intervals correspond to plus and minus the standard deviation estimated from the 100
570 replicated simulations. Results for all simulations are presented in Appendix B (Fig. B4 to
571 B6).

APPENDIX A: Methodological issues

A.1 MEM computation

Dray et al. (2006) and Legendre & Legendre (2012) thoroughly described the MEM and their relation with distance-based MEM or db-MEM (called PCNM in the first papers describing the method). In this section, we summarize their relationship and describe how the MEMs were computed. According to the aforementioned authors, MEMs can be generally defined as follow:

$$[\boldsymbol{W}]_{ij} = [\boldsymbol{S}]_{ij} \circ [\boldsymbol{B}]_{ij} \quad (1)$$

where**W** is a spatial weighting matrix computed as the Hadamard product (element-wise product) of a binary connectivity matrix **B** by a similarity matrix **S**:$\mathbf{W} = [w_{ij}] = [s_{ij} \times b_{ij}]$. The main difference between the db-MEM and MEM is how those two matrices (**S**and **B**) are defined. In our simulations, db-MEM and a second type of MEM were defined as follows:

db-MEM

$$[S]_{ij} = 1 - \left(\frac{d_{ij}}{4\tau}\right)^2$$

$$[B]_{ij} = \begin{matrix} 1 \ if \ \ d_{ij} \leq \tau \\ 0 \ if \ \ d_{ij} > \tau \ or \ i = j \end{matrix}$$

where τ is the largest d_{ij} along the minimum spanning tree (MST).

MEM

$$[S]_{ij} = 1 - \left(\frac{d_{ij}}{max(d_{ij})}\right)^2$$

$$[B]_{ij} = \begin{matrix} 1 \ if \ \ d_{ij} \leq \sqrt{2d^2} \\ 0 \ if \ \ d_{ij} > \sqrt{2d^2} \ or \ i = j \end{matrix}$$

Where d is min(d_{ij}) excluding the diagonal values, and $\sqrt{2d^2}$ is the hypotenuse of a square with edges equal to d[1].

[1]In the case where longitude and latitude are not equally spaced (i.e. a rectangle instead of a square) we used $\sqrt{(d1^2 + d2^2)}$ instead of $\sqrt{2d^2}$, where $d1$ and $d2$ is the minimum distance between two sites for the longitude and latitude respectively. This formula was notably used in the first case study with the Fishing times of the French bottom trawlers where the sample unit is a statistical rectangle.

585 In the present study, all MEM (i.e. MEM_{comp} and MEM_{red}) and db-MEM (i.e. db-MEM_{comp}
586 and db-MEM_{red}) were computed following the steps described in Dray et al. (2006) using the
587 packages *spacemakeR* and *spdep* (Bivand 2011) in R (R Core Team 2014). First, we
588 calculated a matrix of pairwise Euclidean distances **D** among the 400 sampling sites for
589 scenarios S1 to S3 (**D** = $[d_{ij}]$; 20 x 20 grid cells) or 212 sampling sites for S4 (i.e. 53% of 400
590 sites). We then transformed that matrix **D** into a similarity matrix using the appropriate
591 formula (above) to produce db-MEM or MEM. We computed a binary matrix of connectivity
592 among sites **B** using a distance-based approach with arguments setting the lower and upper
593 distance bounds to define neighbors (see above). This was done using the function *dnearneigh*
594 in the *spdep* package in R (Bivand2011).The MEM and db-MEM eigenfunctions are found by
595 eigenanalysis of the doubly centered matrix $= \left(\mathbf{I} - \frac{1}{n}\mathbf{1}_n{}^{\mathrm{T}}(\mathbf{1}_n)\right)\mathbf{W}\left(\mathbf{I} - \frac{1}{n}\mathbf{1}_n{}^{\mathrm{T}}(\mathbf{1}_n)\right)$**.** It can
596 produce positive and negative eigenvalues corresponding to orthogonal eigenvectors. In the
597 present work, we kept all the eigenvalues different from zero and used them all as the set of
598 spatial descriptors (i.e. MEM or db-MEM). Most studies in the literature that are computing
599 MEMs are in fact computing db-MEM (e.g. Bellchambers et al. 2011, Sharma et al. 2011).
600

601 **References**

602 Bellchambers, L. M., J. J. Meeuwig, S. N. Evans, and P. Legendre. 2011. Modelling habitat
603 associations of the common spider conch in the Cocos (Keeling) Islands. Marine Ecology
604 Progress Series**432**:83-90.

605 Bivand, R. with contributions by with contributions by Micah Altman, Luc Anselin, Renato
606 Assunção, Olaf Berke, Andrew Bernat, Guillaume Blanchet, Eric Blankmeyer, Marilia
607 Carvalho,Bjarke Christensen, Yong- wan Chun, CarstenDormann, Stéphane Dray,
608 Rein Halbersma, Elias Krainski, Pierre Legendre, Nicholas Lewin-Koh, Hongfei Li,
609 Jielai Ma, Giovanni Millo, Werner Mueller, Hisaji Ono, Pedro Peres-Neto, Gianfranco
610 Piras, Markus Reder, Michael Tiefelsdorf, and Danlin Yu. (2011). spdep: Spatial

611 Dray, S., P. Legendre, and P. R. Peres-Neto. 2006. Spatial modelling: a comprehensive
612 framework for principal coordinate analysis of neighbour matrices. Ecological
613 Modelling**196**:483-493.
614 Legendre, P. and L. Legendre. 2012. Numerical Ecology. Third English edition. Elsevier
615 Science BV, Amsterdam.
616 R Core Team.2014. R: A language and environment for statistical computing. R Foundation
617 for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.
618 Sharma, S., P. Legendre, M. De Caceres, and D. Boisclair.2011. The role of environmental
619 and spatial processes in structuring native and non-native fish communities across
620 thousands of lakes. Ecography**34**:762-771.
621

APPENDIX B: Detailed results from the simulations using the MEM[2]

Collinearity is measured by computing the mean and the standard error of the Pearson correlation coefficients among all the MEMs two by two.

MEM simulations

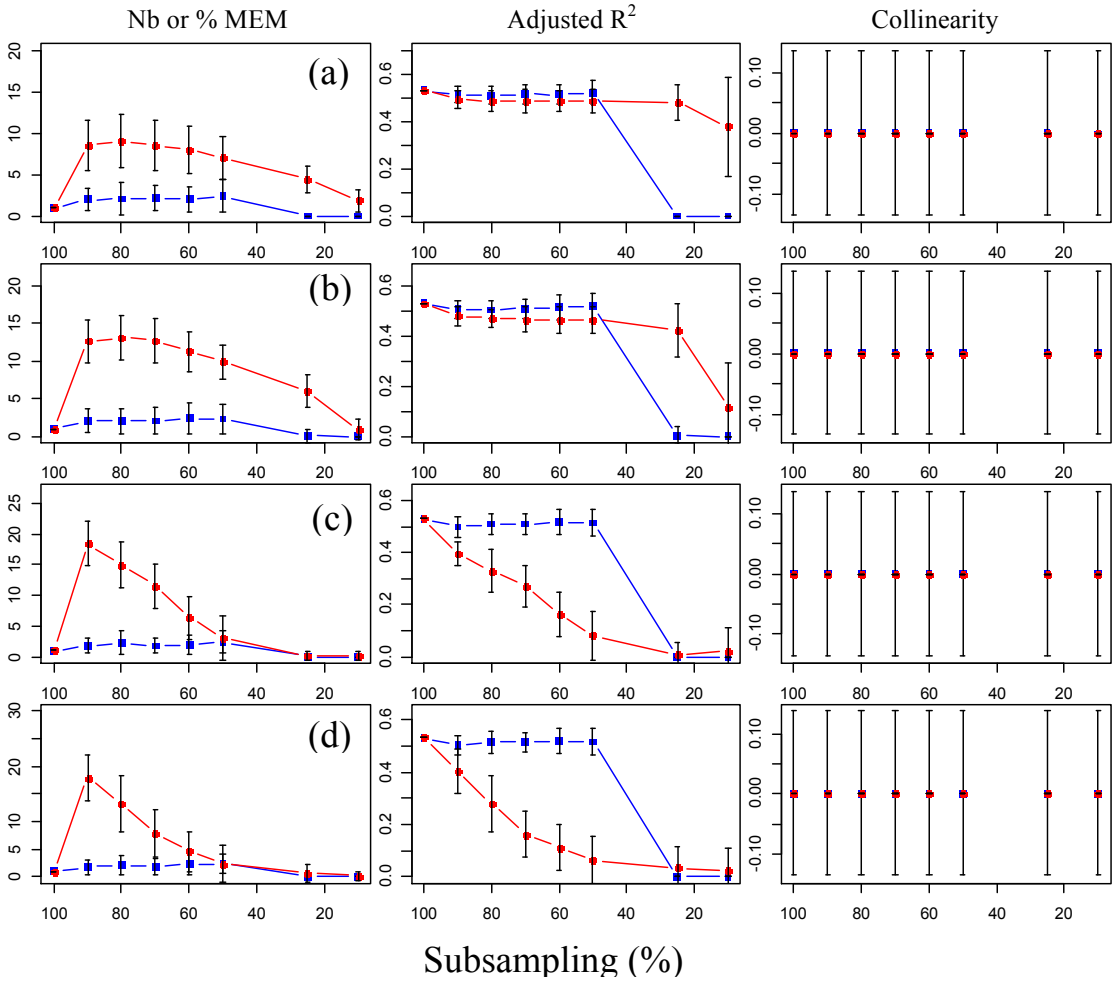
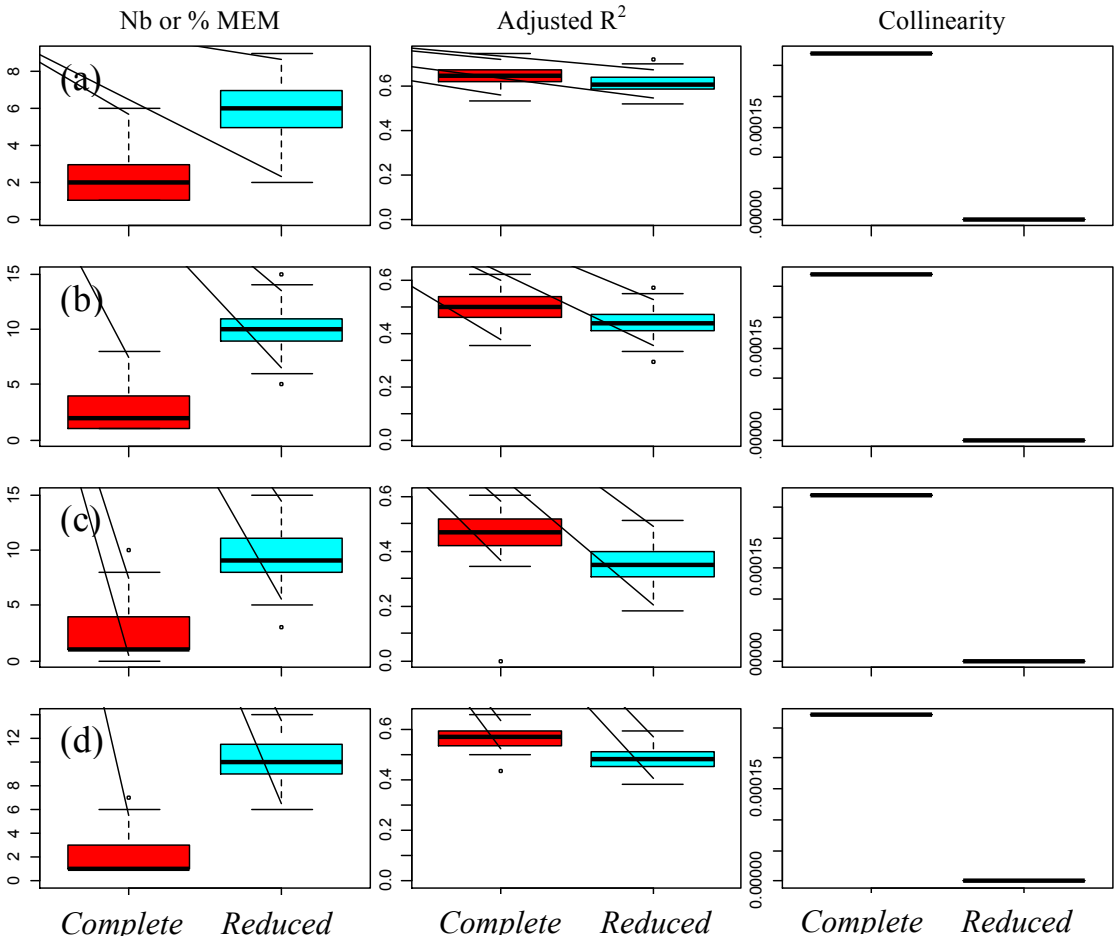


Figure B.1 Results for the first scenario (S1) for the (a) coarse, (b) medium (c) fine, and (d) very fine spatial structures using the MEM_{comp} (blue squares) and MEM_{red} (red circles). The left side panels present the number of MEM_{comp} and MEM_{red} required to model the various predefined spatial structures, the middle panels give the adjusted R^2, and the right side panels

[2] Analyses using the DB-MEMs gave exactly the same results. To avoid duplication, we decided not to present them.

describe the collinearity between the MEMs at different sampling thresholds in percent (25, 50, 60, 70, 80, 90, 100).

634



635

636 Figure B.2 Results for the second scenario (S2) for the (a) coarse, (b) medium (c) fine, and
637 (d) very fine spatial structures using the MEM_{comp} (*Complete*) and MEM_{red} (*Reduced*). The
638 left side panels present the number of MEM_{comp} or percentage of MEM_{red} required to model
639 the various predefined spatial structures, the middle panels give the adjusted R^2, and the right
640 side panels describe the collinearity between the MEMs.

641

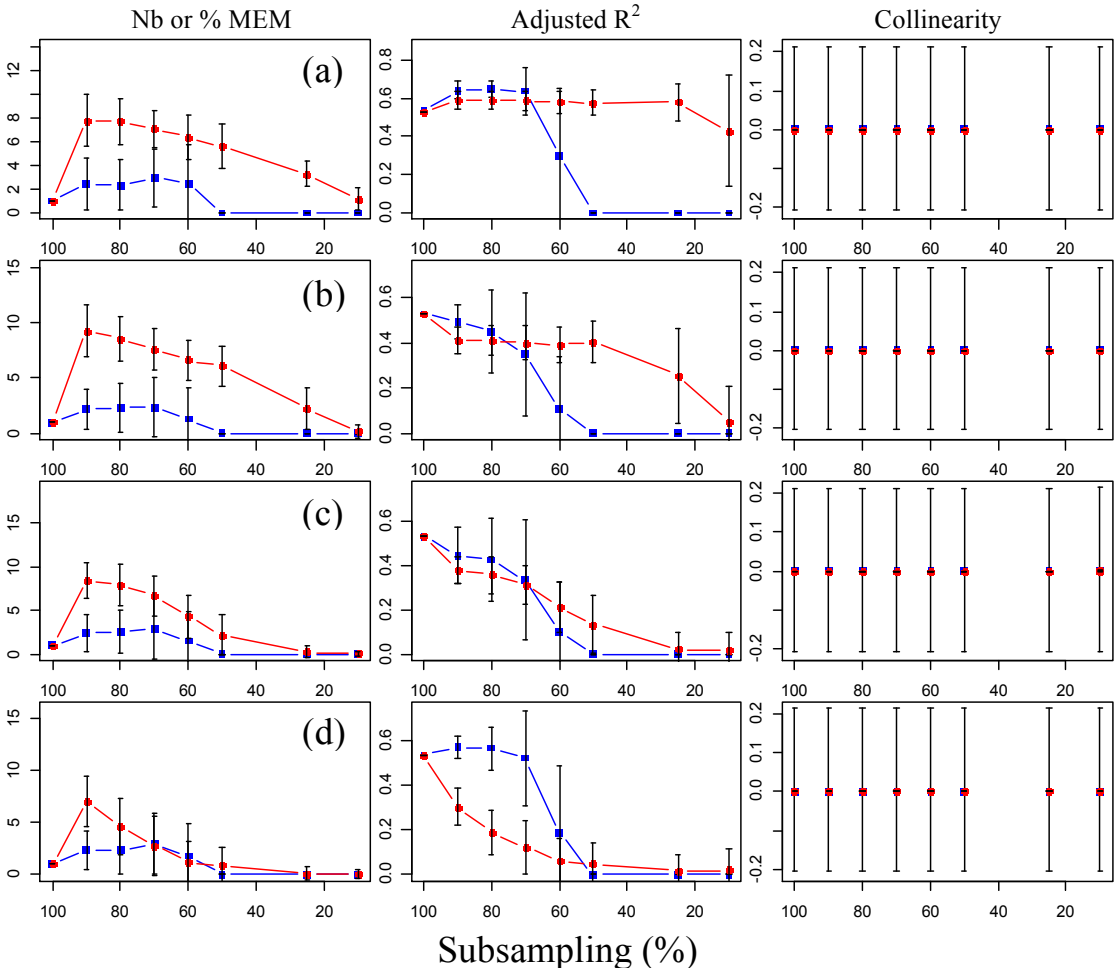


Figure B.3 Results for the third scenario (S3) for the (a) coarse, (b) medium (c) fine, and (d) very fine spatial structures the MEM_{comp} (blue squares) and MEM_{red} (red circles). The left side panels present the number of MEM_{comp} or percentage of MEM_{red} required to model the various predefined spatial structures, the middle panels give the adjusted R^2, and the right side panels describe the collinearity between the MEMs at different sampling thresholds in percent (25, 50, 60, 70, 80, 90, 100).

Variogram simulations

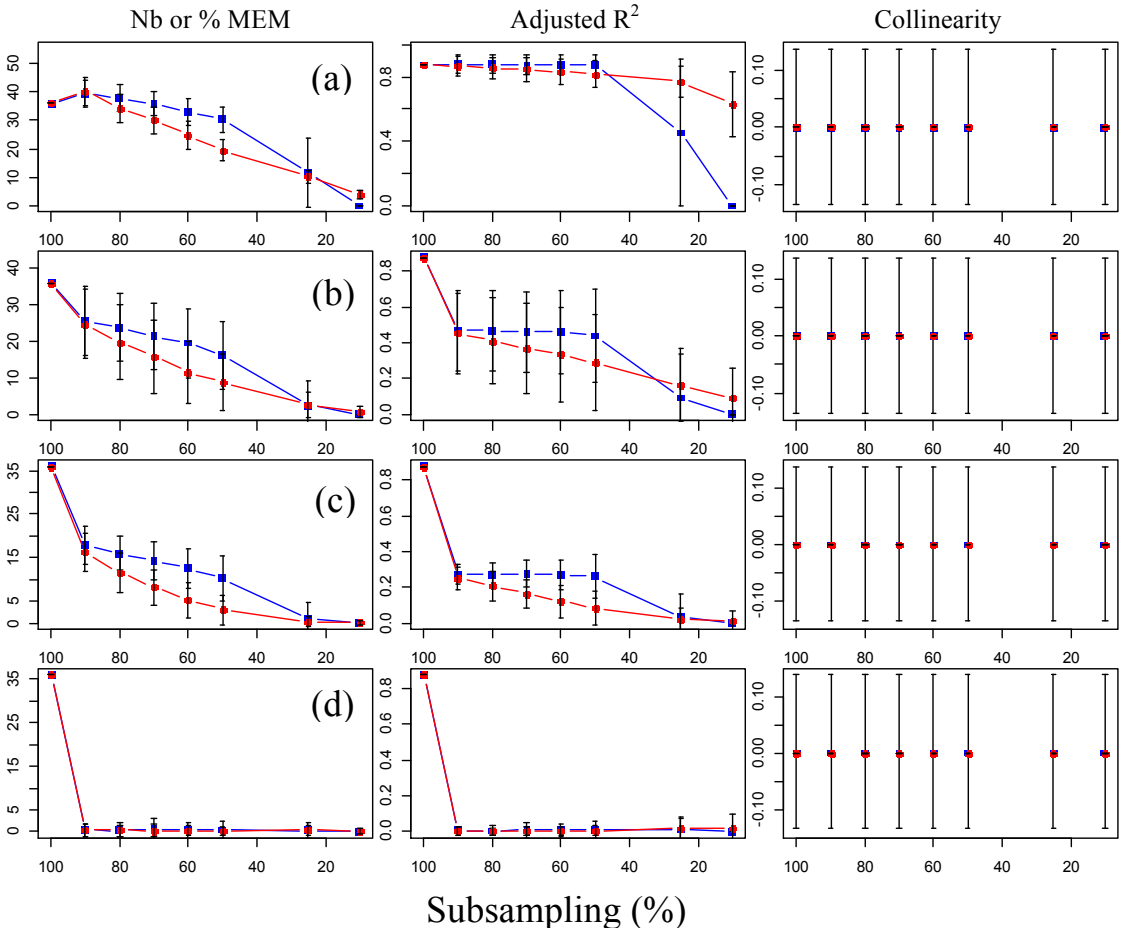


Figure B.4 Results for first scenario (S1) for the (a) coarse, (b) medium, and (c) fine spatial structures using the MEM_{comp} (blue squares) and MEM_{red} (red circles). The left side panels present the number of MEM_{comp} and MEM_{red} required to model the various predefined spatial structures, the middle panels give the adjusted R^2, and the right side panels describe the collinearity between the MEMs at different sampling thresholds in percent (25, 50, 60, 70, 80, 90, 100).

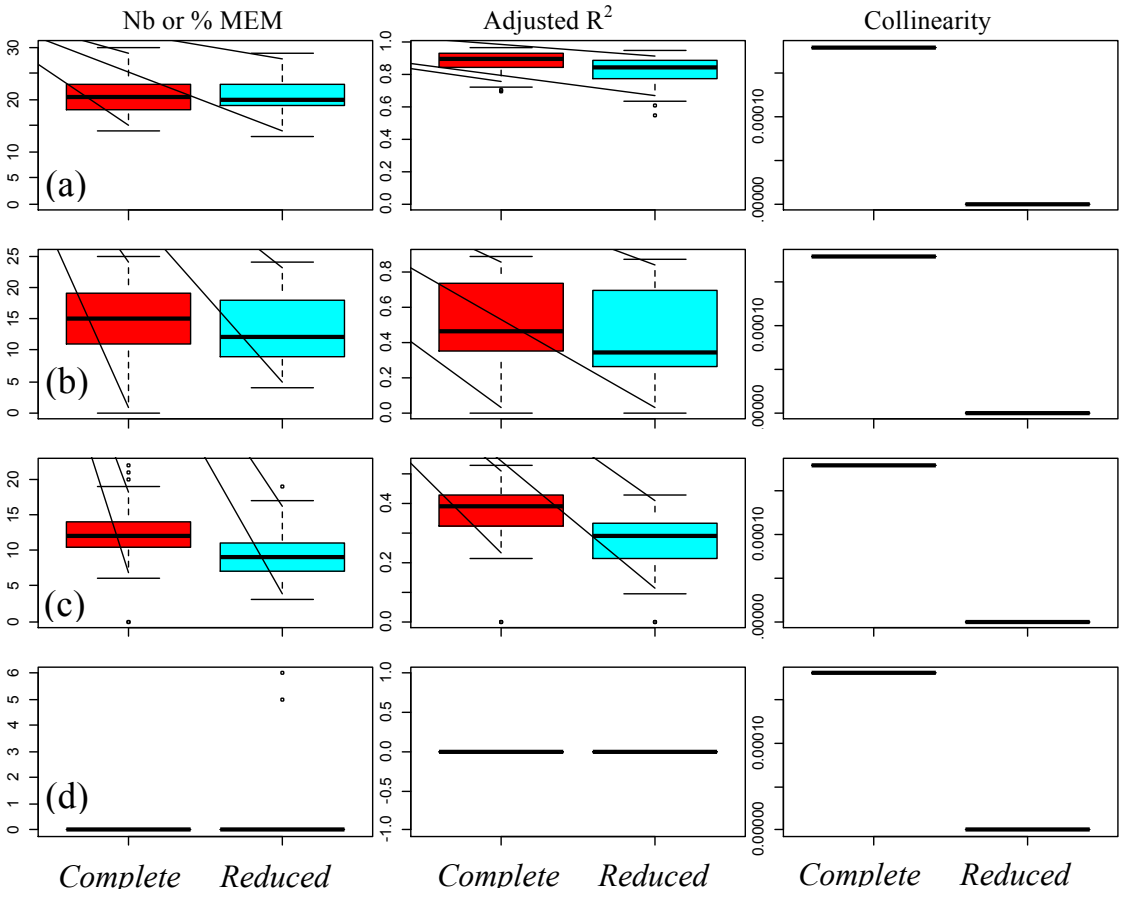


Figure B.5 Results for second scenario (S2) for the (a) coarse, (b) medium, and (c) fine spatial structures using the MEM_{comp} (*Complete*) and MEM_{red} (*Reduced*). The left side panels present the number of MEM_{comp} or percentage of MEM_{red} required to model the various predefined spatial structures, the middle panels give the adjusted R^2, and the right side panels describe the collinearity between the MEMs.

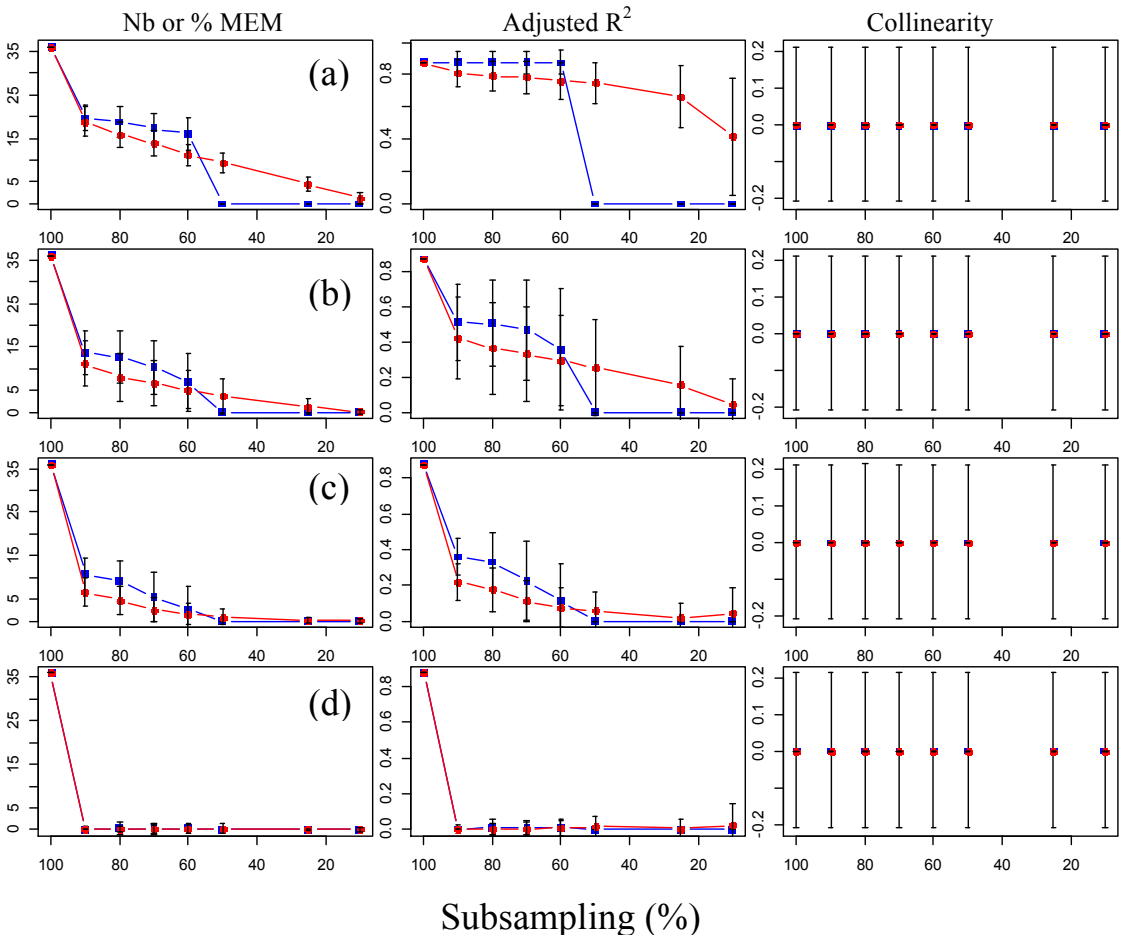


Figure B.6 Results for third scenario (S3) for the (a) coarse, (b) medium, and (c) fine spatial structures using the MEM_{comp} (blue squares) and MEM_{red} (red circles). The left side panels present the number of MEM_{comp} or percentage of MEM_{red} required to model the various predefined spatial structures, the middle panels give the adjusted R^2, and the right side panels describe the collinearity between the MEMs at different sampling thresholds in percent (25, 50, 60, 70, 80, 90, 100).