Spatial modelling of plant diversity from high-throughput environmental DNA sequence data

A. Studeny [1] & F. Forbes [1] & E. Coissac [2] & A. Viari [1] & C. Mercier [2] & L. Zinger [2] & A. Bonin [2] & F. Boyer [2] & P. Taberlet [2]

[1] *INRIA Grenoble – Rhône-Alpes, 655 avenue de l'Europe, Montbonnot, 38334 Saint Ismier Cedex*
[2] *Laboratoire d'Ecologie Alpine, BP 53, 2233 rue de la piscine, 38041 Grenoble Cedex 9*

angelika.studeny@inria.fr

Résumé. Cet article présente une approche statistique pour modéliser les corrélations spatiales entre espèces dans un écosystème. L'originalité réside dans la particularité des données, génerées par des séquençages à haut-débit de l'ADN environnemental d'échantillons de sol. Les données utilisées dans cet étude étaient recueillis à la station biologique CNRS des Nouragues, en Guyane Française. L'étude décrit les relations spatiales bivariées de ces données par un modèle linéaire de co-régionalisation séparable où l'on estime un paramètre de cross-corrélation. Sur la base de cette estimation, nous visualisons le modèle de co-occurrences sous forme de graphes d'interactions. Les limites de cette approche sont discutées ainsi que les alternatives possibles.

Mots-clés. Correlation spatiale, champs aléatoires Gaussiens, modèles linéaires de coregionalisation, ADN environnemental, données de haute dimension

Abstract. This paper considers a statistical modelling approach to investigate spatial cross-correlations between species in an ecosystem. A special feature is the origin of the data from high-troughput environmental DNA sequencing of soil samples. Here we use data collected at the Nouragues CNRS Field Station in French Guiana. We describe bivariate spatial relationships in these data by a separable linear model of coregionalisation and estimate a cross-correlation parameter. Based on this estimate, we visualise plant taxa co-occurrence pattern in form of 'interaction graphs' which can be interpreted in terms of ecological interactions. Limitations of this approach are discussed along with possible alternatives.

Keywords. Spatial correlation, Gaussian random fields, linear coregionalization, environmental DNA, high dimensional data

1 Motivation

Recent technological and computational advances have confronted applied statistics with data sets of steadily growing size and complexity. In ecology, the introduction of molecular methods is currently revolutionizing traditional approaches of assessing species diversity, the composition of ecological communities and the co-occurrence of species [15]. The idea of creating species inventories based on genetic 'barcodes' [5, 12] has rapidly gained popularity and multi-taxonimic DNA 'fingerprints' are now starting to be used for assessing diversity and composition of ecological communities.

Meta-barcoding consists in extracting and identifying DNA-fragments from environmental samples with the help of taxon-specific genetic markers (see details below) [2]. In contrast to other methods, metabarcoding data contains indirect information on the presence of a target species and is thus able to trace multiple taxa simultaneously In analogy to DNA-barcoding, the genetic 'fingerprints' deposited in an environmental sample, such as in soil, are expressed and amplified applying PCR protocols (*polymerase chain reaction*) and subsequently identified as MOTUs (*molecular operational taxonomic units*) up to the lowest possible taxonomic resolution with the help of DNA-sequence reference data bases [12], established on purpose or extracted from standard DNA libraries, such as *GenBank* (www.ncbi.nlm.nih.gov/genbank). This provides a novel set of tools to revisit traditional questions in community ecology about the coexistence of species.

Relatively low-cost and easily applied in the field, these new sampling techniques are likely to be widely used in near future. In order to be effective for biodiversity assessment, these tools have to be combined with statistical methodology and analysis tailored towards the specifics of these data: On one hand, we are faced with a large quantity of data, where the dimension is larger than the sample size, on the other their taxonomic classification is often incomplete in the sense that species identification is not always possible. The output sequences contain noise as well as systematic error caused by the PCR protocol. Sequence abundance is highly variable between MOTUs, ranging between hundreds or less per sample for rare sequences up to several thousands for the most abundant, and can also vary considerably within the same MOTU across samples.

This paper presents a first attempt at spatial statistical modelling of metabarcoding data. The aim of the modelling approach is to explicitly take into account information on the spatial distribution and the identification of species co-occurrence pattern. After a short overview of the data and the molecular methods used for their extraction, we develop specific research questions, followed by an outline of the statistical analysis. We conclude with a summary of the (preliminary) results and discuss possible generalisations of this model as well as alternatives that we intend to study in the future.

2 Description of the data and aim of the study

The data come from soil samples taken on a regular grid of 19 × 19 points in a 100 ha square area in tropical forest at the Nouragues CNRS Field Station in French Guiana. These data have been collected for preliminary analysis to educate the planning of a subsequent survey with larger spatial coverage in the same region. Soil samples are

processed in the following way:
Extracellular DNA is isolated independently from each soil core. Then a genomic regions (P6 loop of chloroplast *trn*L intron, [11]) is amplified by PCR from the metegenome. This region was chosen because it is variable enough to allow plant species discrimination, but conserved sufficiently at its extremeties to allow the constructions of DNA primers, necessary to run the PCR. After deletion of obvious erroneous sequences from the PCR output, the remaining sequences are distinguished as MOTUs (*molecular operational taxonomic units*). In the so-processed plant data, a total of 601 MOTUs were retained. No covariate data, such as soil pH for example, are available.
The aim is to develop methodology based and tested on this preliminary tropical plant data set which can subsequently be applied as a standard tool kit for biodiversity analysis of high throughput environmental DNA data across diverse taxonomic groups and climate zones. In particular, we aim at inferring spatial co-occurrence pattern between taxonomic groups as well as between different units within the same taxonomic group and at distinguishing such true signals from potential PCR errors.

3 Gaussian random fields for interspecies correlations

We assume for a fixed MOTU that the observed sequence count $Y(x)$ at position $x \in D \subset \mathbb{R}^2$ (where D denotes the 100 m^2 survey plot) follows a Poisson distribution, where the value of the intensity parameter $\lambda(x)$ is a realisation of a (latent) Gaussian random field $\Lambda(x)$ at site x. Spatial correlation in Λ is imposed by a structured additive regression model [7]

$$\log \Lambda(x) = \mu(x) + U(x), \tag{1}$$

where $\mu(x)$ is a local intercept term and $U(x)$ a spatially structured effect, namely a Gaussian random field with mean 0 and covariance function $C(\cdot)$.
(This model can be extended, including covariate information through a linear predictor $\mu(z(x)) = \sum \beta_i z_i(x)$ depending on covariate values $z_i(x)$ at locations x, or even adding more general functional dependencies on the covariate data in terms of smooth regression splines [4]. Model (1) could also incorporate (unstructured) random effects [6], however given that the data comprise only observed sequences counts at the sample sites we chose not to include effects other than the spatial field to avoid overparametrisation).
For an ensemble of MOTUs $\mathbf{Y} = (Y_i(x))_{1 \le i \le N}$, the model is generalised to a multivariate setting taking into account spatial cross-correlation by linking the latent spatial fields as realisations of a multivariate Gaussian process (GP) with cross-covariance matrix $\mathbf{C}$ [8]: for $i = 1, \dots, N$

$$\log \Lambda_i(x) = \mu_i(x) + U_i(x), \text{ where } \mathbf{U} \sim GP(\mathbf{0}, \mathbf{C})$$

the entries of $\mathbf{C}(h) = (C_{ij}(h))_{1 \le i,j \le N}$ with $C_{ij}(h) = \mathbb{E}[U_i(x)U_j(x+h)]$ being the cross-covariances for the spatial fields U_i and U_j, $i \neq j$.

Valid model choices for $\mathbf{C}$ are limited by the condition that, for any finite sample of locations $x_1, \ldots, x_s$, the sample covariance matrix for realisations $U_i(x_1), \ldots, U_i(x_s)$ and $U_j(x_1), \ldots, U_j(x_s)$ has to be semi-positive definite.

Linear models of coreginalization (LMC) [14, 3, 10] provide valid cross-correlation models. Practical constraints regarding the implementation of these models, lead us to consider only the bivariate case. Hence, the spatial fields are linked by a linear transformation

$$\mathbf{U}(x) = [U_1(x), U_2(x)]^t = \mathbf{A}\mathbf{w}(s)$$

where $\mathbf{A}$ is a full-rank 2×2 matrix and $\mathbf{w} = (w_1, w_2)$ are two independent Gaussian random fields with mean 0 and the same correlation function $\rho_0(h)$[1] This model (3) is referred to as *separable* in the literature, as $\mathbf{C}(h) = \rho_0(h)\mathbf{A}\mathbf{A}^{\mathbf{t}}$, and hence the correlation is determined entirely by the coefficients a_{ij} of the matrix $\mathbf{A}$. In particular, we derive the cross-correlation between U_1 and U_2 at distance 0 as

$$\rho = \frac{a_{11}a_{21}}{\sqrt{a_{11}^2(a_{12}^2 + a_{22}^2)}}.$$

We estimate ρ_{ij} for every sequences pair (i, j) of the 601 plant MOTUs by implementing and fitting this model in a hierarchical Bayesian framework using the R library `R-INLA` [9]. This software package enables the user to fit latent Gaussian process models at a computational cost that is lower than the standard MCMC procedure. This was appealing given the many pair-wise models we had to calculate.

4 Results

For every sequence pair (i, j), the fitted model outputs the posterior mean $\hat{\rho}_{ij}$ of the cross-correlation along with its standard deviation and a 95%-credible intervals. Since our aim was to identify pairs with potential spatial interaction, we decided to keep all pairs for which the credible interval did not span 0. Depending on the sign of $\hat{\rho}_{ij}$ we speak of a positive or negative spatial cross-correlation.

Prior to further analysis, potential PCR errors are identified on the following thought: errors occur during PCR amplification, when small changes in the original DNA sequence are introduced. This leads to a similar sequence, but is likely identified as a different MOTU exhibiting positive spatial correlation with the original one. Assuming that the edit distance (total number of base differences) between the original and the erroneous sequence is small and that such a 'false' pair is not expected to be negatively correlated, we look at the paramter ρ_{ij} against the edit distances. Comparing the upper and lower half of Fig.1(a), corresponding to pairs with positive and negative spatial correlation, respectively, we discard pairs with a distance of less than 10 as likely PCR errors.

[1] Note that in general, LMCs can include an arbitrary number of Gaussian fields w_k with not necessarily identical correlation functions $\rho_k(h)$.

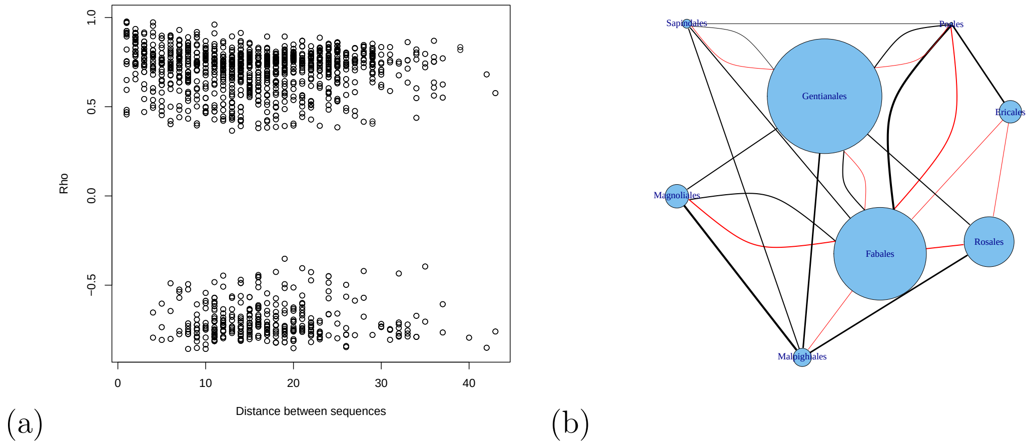

Figure 1: (a) cross-correlation for each pair of 601 tropical plant MOTUs plotted against the edit distances between the two sequences. (b) Subset of an interaction graph at the order level. The size of a vertex corresponds to the number of sequences within the order, edges show positive (black) and negative (red) spatial cross-correlation between pairs of MOTUs in the respective orders.

For the remaining pairs, a 'spatial interaction graph' is drawn, depending on a threshold on ρ (Fig.1(b)). This graph visualises spatial correlation between taxonomic classes (here at the order level), where edge width represents the strength of positive and negative correlation between the remaining sequence pairs in the respective classes. Vertex size corresponds to the size of the order as it is represented in the data set (total number of sequences belonging to this order).

5 Discussion

We present results from a first statistical analysis of the spatial correlations in tropical plants using data from environmental DNA sequencing. In a 'brute force' approach a model based on cross-correlated Gaussian spatial fields was fitted to all MOTU pairs and the estimated pairwise cross-correlation can be investigated. Even with a very simple model (separable bivariate LMC without covariate data), this already allows us to visualise and reveal some co-occurrence pattern demonstrating the interest to further analyse the co-occurrence network exhibited by this approach. The interaction graph as it is shown appears to be at a too coarse taxonomic level for a more detailed ecological interpretation and requires additional tuning of the taxonomic resolution at this stage as well as connection to relevant ecological pattern. However, given the large number of variables (601 MOTUs), computational costs are still high, even when a computationally efficient algorithm is applied. The computational effort that is necessary to fit even a very simple spatial model, such as the bivariate, separable LMC, seems unreasonable and possible infeasible once we pass onto the final, considerably larger data set. Hence, future work will consider alternatives , such as dimensional reduction based on model-based subspace clustering including sparsity constraints [1] and to extend those to take into account spatial dependencies, along the lines of [13].

References

[1] C. Bouveyron and C. Brunet. Discriminative variable selection for clustering with the sparse Fisher-EM algorithm. *Computational Statistics*, 2013, In press.

[2] L.S. Epp, S. Bossenkool, E.P. Bellemain, J. Haile, A. Esposito, T. Riaz, C. Erséus, V.I. Gusarov, M.E. Edwards, A. Johnsen, H.K. Stenoien, K. Hassel, H. Kauserud, N.G. Yoccoz, K.A. Brathen, E. Willerslev, P. Taberlet, E. Coissac, and C. Brochmann. New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems. *Molecular Ecology*, 2012.

[3] Alan E. Gelfand, Alexandra M. Schmidt, Sudipto Banerjee, and C.F. Sirmans. Nonstationary multivariate process modeling through spatially varying coregionalization. *Test*, 13:263–312, 2004.

[4] T. Hastie and R. Tibshirani. Generalized additive models (with discussion). *Statistical Science*, 1986.

[5] P.D.N. Herbert, S. Ratnasingham, and J.R. DeWaard. Barcoding animal life: cytochrome *c* oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society, Series B*, 2003.

[6] J.B. Illian, S.H. Sorbye, and H. Rue. A toolbox for fitting complex spatial point process models using integrated nested laplace approximation (INLA). *Annals of Applied Statistics*, 2012.

[7] T. Kneib and L. Fahrmeirr. Structured additive regression for multicategorical space-time data: a mixed model approach. *Biometrics*, 2006.

[8] T. Rajala, J.B. Illian, and D. Simpson. Practical inference for inhomogeneous multivariate log-Gaussian Cox processes, In preparation.

[9] H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society B*, 71:319–392, 2009.

[10] Alexandra M. Schmidt and Marco A. Rodriguez. Modelling multivariate counts varying continuously in space. In *Bayesian Statistics*.

[11] P. Taberlet, E. Coissac, F. Pompanon, L. Gielly, C. Miquel, A. Valentini, T. Vermat, G. Corthier, C. Brochmann, and E. Willerslev. Power and limitations of the chloroplast *trnl*(uaa) intron for plant DNA barcoding. *Nucleic Acids Research*, 2077.

[12] P. Taberlet, S.M. Prud'homme, E. Campione, J. Roy, C. Miquel, W. Shehzad, L. Gielly, D. Rioux, P. Choler, J.-C. Clément, C. Melodelima, F. Pompanon, and E. Coissac. Soil sampling and isolation of extracellular DNA from large amount of starting material suitable for metabarcoding studies. *Molecular Ecology*, 2012.

[13] M. Vignes, J. Blanchet, D. Leroux, and F. Forbes. Clustering of incomplete, high dimensional and dependent biological data with spaCEM3, 2010.

[14] H. Wackernagel. *Multivariate Geostatistics.* Springer, 1998.

[15] N.G. Yoccoz. The future of environmental DNA in ecology. *Molecular Ecology*, 2012.