Modélisation dynamique des données d'abondance en écologie

Paul Doukhan & Lionel Truquet

May 6, 2021

1 Résumé du projet

Cette thèse sera co-encadrée par Lionel Truquet, Professeur Associé à l'ENSAI (CREST/ENSAI UMR 9194, rue Blaise Pascal, 35170 Bruz) et Paul Doukhan, Professeur à l'Université de Cergy (UMR 8088 Analyse, Géométrie et Modélisation, 2, avenue Adolphe Chauvin,95302 Cergy-Pontoise Cedex, France.), dans le contexte du projet ECODEP, http://doukhan.u-cergy.fr/EcoDep.html. Cette thèse est centrale dans le contexte du projet s http://doukhan.u-cergy.fr/ecodep_labs.html et la thèse sera au coeur des thématiques de ECODEP, avec un environnement de chercheurs aussi bien en France qu'à l'étranger (voir http://doukhan.u-cergy.fr/members.html). Dans ce travail, on se propose de définir et d'étudier des modèles mathématiques dynamiques ainsi des méthodes d'inférence statistique pour les données d'abondance en écologie, un problème au coeur du projet ECODEP, voir http://doukhan.u-cergy.fr/ecodep_abstract.html). L'étude des données d'abondance pour les espèces animales et végétales est un des problèmes majeurs en écologie. L'abondance désigne la quantité des espèces présentes dans un écosystème, quantité qui est impactée par différents facteurs tels que la prolifération d'espèces invasives, le pourcentage de biomasse disponible, la dynamique du changement climatique ainsi que les caractéristiques propres à l'écosystème (competition des espèces, phénomènes de prédation, type d'habitat, nutriments disponibles...). Voir par exemple McGill et al. (2007) pour une discussion sur les différentes approches utilisées pour modéliser la distribution d'abondance des espèces (SAD).

Il existe deux définitions principales pour la notion d'abondance des espèces en écologie. Une notion standard concerne le nombre d'individus pour chaque espèce présente dans l'écosystème. Il s'agit alors de données de comptages qui sont étudiées de façon statique dans de nombreux travaux de recherche en utiliant des modèles de type poissoninens ou poissoniens log-normaux. Voir par exemple Baldridge et al. (2016), Welsh et al. (1996), Shinen and Navarrete (2014) ou le récent travail de Chiquet et al. (2018). Lorsque on regarde les proportions de chaque espèce dans le système, on parle d'abondance relative. Voir par exemple MacArthur (1960) ou Volkov et al. (2003). L'abondance relative est une notion attractive en tant que mesure de biodiversité. Dans ce deuxième cas, une façon de modéliser la distribution probabiliste de ces proportions est de considérer des lois de probabilités sur le simplexe (lois beta/Dirichlet ou logit-normale par exemple). Du point de vue statistique, on parle de données compositionnelles. Voir par exemple Aitchison (1982) pour les distributions de probabilités dans ce contexte et Douma and Weedon (2019) pour des modèles de régression qui peuvent être utilisés dans le cas statique en écologie.

Les travaux traitant de l'abondance en écologie sont presque exclusivement consacrés à la modélisation statique et se concentrent pour la plupart sur la modélisation de l'abondance d'une seule espèce sans nécessairement prendre compte des interactions avec les autres espèces. Ce projet con-

sistera à modéliser de façon dynamique ces données d'abondance en écologie à l'aide de techniques de séries temporelles et à développer des techniques d'inférence statistique pour permettre identifier des interactions entre les espèces présentes dans l'écosystème ainsi qu'avec les facteurs environnementaux. Une étude de l'évolution temporelle de ces données pourrait ainsi fournir aux écologues de nouveaux indicateurs sur les propriétés des systèmes écologiques, l'évolution de la biodiversité et d'aider à quantifier l'impact des variables climatiques (variables exogènes) sur le devenir de ces systèmes. A cet effet, nous proposons d'étudier de nouveaux modèles de séries temporelles multivariés de type linéaire généralisés et dont l'inférence statistique devra prendre en compte la rareté de certaines espèces, la grande dimension potentielle des données, les covariables associées ainsi que la non-stationarité. Nous nous concentrerons en premier lieu sur la notion d'abondance relative mais l'utilisation de séries de comptage multivariées (voir Fokianos et al. (2020) pour des exemples de modèles récents) est aussi envisagée pour la notion d'abondance non renormalisée. Nous proposons aussi de tester nos modèles sur les données fournies par les chercheurs chiliens membres du projet (les écologues Pablo Marquet, Sergio Navarette et le mathématicien Rolando Reboledo) qui travaillent en parallèle sur la modélisation dynamique des données d'abondance à partir de processus de diffusion. Voir par exemple le travail Marquet et al. (2017) qui montre une possible distribution commune (loi beta) pour l'abondance relative d'une seule espèce dans divers écosystèmes à grande échelle qui est est compatible avec les observations empiriques de nombreuses espèces d'oiseaux, de papillons, d'arbres tropicaux ou de récifs coraux. Un travail en étroite collaboration avec les membres internationaux du projet sera donc au coeur de ce projet doctoral. Le site du projet ECODEP `https://retriever.readthedocs.io/en/latest/datasets_list.html` répertorie déjà des jeux de données accessibles pour le candidat afin de tester les modèles envisagés.

2 Hypothèses, questions posées, identification des points de blocage

L'analyse des données d'abondance demandera d'abord de définir des modèles dynamiques sur des espaces moins standards que ceux étudiés la littérature (données entières multivariées ou données à valeurs dans le simplexe). Une des difficultés est le nombre relativement limité de distibutions de probabilités multivariées pour des données non gaussiennes.

Dans une première étape, le candidat pourra se consacrer à des modélisations stationnaires au cours du temps. Ce type de modélisation ne peut bien sûr être réaliste que sur le cours terme. Mais il peut déjà permettre de mettre en avant certaines corrélations (impact du nombre d'individus d'une espèce donnée sur les abondances futures des autres espèces, variables climatiques qui joue en faveur ou en défaveur des abondances futures...). L'étude de modèles non linéaires est inévitable dans ce domaine. Par exemple, la notion classique de "density dependence" vérifiée pour de nombreuses espèces, voir par exemple Brook and Bradshaw (2006), est basée sur l'hypothèse que la taille des populations a un effet positif ou négatif sur son taux de croissance. Toutefois, les modèles de séries temporelles utilisées dans la litérature en écologie sont essentiellement linéaires (ou linéaires après transformation logarithmique) ce qui ne peut répondre de manière satisfaisante à l'étude statistique de ces données. Une des questions proposées est de définir des modèles de type linéaires généralisés (GLM) innovants pour étudier l'évolution de l'abondance ou de l'abondance relative de plusieurs espèces en intéraction au cours du temps. Ce champ de recherche est relativement nouveau et peu de solutions ont été proposées jusqu'à présent dans la litérature des séries temporelles.

Pour ce qui concerne la modélisation des données entières multivariées (données d'abondance sous forme de comptages), nous proposons d'abord de développer la modélisation basée sur des

copules pour les les processus multivariés dans les lois conditionnelles marginales sont des lois de Poisson qui ont été introduits dans Fokianos et al. (2020) et étudiés également dans Debaly and Truquet (2019). A cet effet, on pourra d'abord étendre ces modèles en autorisant des covariables exogènes (variables climatiques, caractéristiques des systèmes) dans la dynamique. L'utilisation d'autres lois que la loi marginale de Poisson et qui puissent prendre en compte le phénomène de surdispersion des comptages (un problème assez classique dans ce domaine) pourra être envisagés. A cet effet, la loi binomiale négative pourrait être mieux adaptée que la loi de Poisson utilisée dans Fokianos et al. (2020). L'estimation des paramètres de copule pour des données de dimension importante est un point délicat car les données sont discrètes et la copule n'est pas unique. Une solution sera proposée ci-dessous. Pour ce qui est de la notion d'abondance relative, il est envisagé d'étudier des chaînes de Markov sur le simplexe et pour lesquelles l'opérateur de transition est construit à partir de la loi de Dirichlet. Cette approche dynamique demandera de généraliser les modèles de régression (donc qui ne peuvent que s'appliquer à des données statistiques) proposés par exemple dans Aitchison (1982), Douma and Weedon (2019) ou Tsagris and Stewart (2018). Il s'agira dans un premier temps de déterminer les conditions de stationarité pour ces modèles afin d'obtenir une estimation consistante de ses paramètres. Ces modèles, potentiellement de grande dimension si on considère l'abondance jointe de plusieurs espèces sur plusieurs sites d'observation, demanderont de considérer des estimateurs pénalisés afin de réduire cette dimension. Enfin, observer la dynamique de données sur plusieurs sites géographiques est souvent synonyme d'une sur-représentation des zeros (espèces non observées sur certains sites). Voir par exemple Wenger and Freeman (2008). Dans ce cas des versions "zero-inflated" des modèles susmentionnés devront être développés. Voir par exemple Zhu (2012) pour des exemples de modèles pour des séries temporelles univariées prenant en compte cet effet. Des extensions multivariées de ces distributions sont là aussi nouvelles et demandent un travail substantiel.

Dans un deuxième temps, des extensions non stationnaires des modèles envisagées ci-dessus seront étudiées. La présence de nonstationarité est classique pour les séries temporelles issues de l'écologie. Voir par exemple Turchin and Taylor (1992), Shannon et al. (2009) ou Litzow et al. (2019). Cette nonstationarité peut être causée par des effets saisonniers mais aussi par des perturbations de l'environnement et/ou de la biomasse qui affectent l'évolution ainsi que l'intéraction des espèces au cours du temps. L'hypothèse de stationnarité est alors peu réaliste pour expliquer les données d'abondance sur le long terme, en particulier pour les espèces menacées d'extinction. Disposer de modèles pour lesquels l'hypothèse de stationarité ne serait réaliste que locallement au cours du temps et pour lesquels l'inférence statistique des paramètres reste possible est un problème difficile. Les pistes envisagées seront présentées dans la section suivante.

3 Approches méthodologique et technique envisagées

L'approche méthodologique comportera aussi bien une partie probabiliste qu'une partie statistique.

La partie probabiliste sera consacrée dans un premier temps à la modélisation stationnaire en utilisant des techniques de chaines de Markov. Trouver une façon générique de construire des chaînes de Markov sur le simplexe est un des problèmes délicats. Nous pensons qu'il est possible de définir en première approche de tels modèles dynamiques à partir de la loi de Dirichlet, dans l'esprit des modèles déjà proposées en statique. La loi de Dirichlet est la distribution de probabilité de référence pour les données à valeurs dans le simplexe et est une extension naturelle de la loi beta univiariée. Définir de tels modèles en incorporant des covariables exogènes pourra être envisagé à partir de

techniques générales développées par les encadrants pour étudier la stabilité des chaînes de Markov en environnement aléatoire (l'environnement aléatoire correspond en particulier aux covariables exogènes qui influencent les systèmes écologiques). Voir en particulier Debaly and Truquet (2020) et Doukhan et al. (2020). Les mêmes procédés peuvent être utilisés pour inclure des variables exogènes dans les processus de comptage multivariés développés par Fokianos et al. (2020).

La partie estimation statistique en grande dimension pourra se baser sur des techniques de pénalisation déjà étudiés par Paul Doukhan pour des modèles de séries temporelles linéaires. Voir en particulier l'étude de l'estimateur LASSO dans Alquier and Doukhan (2011) et l'estimation de processus autorégressifs sous contrainte de faible rang dans Alquier et al. (2020). Il y aura toutefois un travail substantiel pour obtenir des guaranties statistiques pour ces méthodes dans le cadre de modèles linéaires généralisés puisque le critère classique des moindres carrés utilisé pour les modèles linéaires devra être remplacé par la log-vraisemblance. Pour les processus de comptage multivariés développés dans Fokianos et al. (2020), l'estimation de la copule pour modéliser la dépendance simultanée entre les séries de comptage reste un problème majeur non étudié. On pourra commencer par étudier le cas des copules gaussiennes pour lesquelles les paramètres de la matrice de corrélation restent identifiables. Des techniques de vraisemblance composite, qui ont montré de bonnes propriétés pour estimer les paramètres des copules gaussiennes sur données discrètes (voir par exemple Masarotto and Varin (2012)) sont envisagées.

Pour prendre en compte la non-stationarité des données d'abondance au cours du temps, nous proposons au candidat d'utiliser des versions localement stationnaires des modèles stationnaires définis dans un premier temps. L'hypothèse de stationarité locale consiste à autoriser les paramètres du modèle à varier avec le temps tout en étant approchable localement par la dynamique d'un processus stationnaire. Des techniques d'estimation non paramétrique permettent alors d'obtenir une estimation locale de ces paramètres. A cet effet, l'utilisation de techniques basées sur des propriétés de contraction des opérateurs markoviens développées dans Truquet (2019) et Truquet (2020) pourrait être adaptée à ce cadre pour étudier des chaînes de Markov en environnement aléatoire et inhomogènes. La modélisation de la non stationarité à partir de la stationarité locale a été abondamment utilisée pour les données financières, voir par exemple Truquet (2017) et peut aussi être adaptée aux données présentant des saisonnalités, voir Bardet and Doukhan (2018). L'utilisation de ces techniques dans le cadre des données en écologie serait assez nouvelle et permettrait d'identifier de nouvelles tendance qu'il est assez difficile d'identifier à partir de modèles paramétriques.

4 Environnement scientifique et positionnement dans le contexte régional, national et international

Le candidat bénéficiera des expertises variées des différents membres du projet ECODEP. Tout d'abord, les co-encadrants Lionel Truquet et Paul Doukhan ont une expertise certaine dans la modélisation des séries temprorelles à valeurs dicrètes. Paul Doukhan est le PI du projet Ecodep et a une expertise reconnue au niveau international pour la modélisation des données dépendantes. Son ouvrage "Mixing: Properties and Examples" est d'ailleurs une référence majeure sur le sujet. Les travaux méthodologiques de Lionel Truquet sur la non-stationarité des séries temporelles ont été publiés dans les meilleurs revues internationales en statistique. L'expertise des deux co-encadrants sera déterminante pour lever les points de blocage méthodologiques de ce projet (définition des modèles, validité des méthodes d'inférence statistique).

Lionel Truquet travaille au CREST, une UMR bi-localisée sur la région rennaise et parisienne

et dont l'expertise est reconnue au niveau de la statistique des données dépendantes et des données de grande dimension.

Au niveau de la région Bretagne, Gilles Durrieu, membre du projet et Professeur à l'Université de Bretagne sud, travaille également sur les séries temporelles en écologie (voir cours dans le cadre du projet sur https://doukhan.u-cergy.fr/education.html), en particulier pour développer des indicateurs de la qualité de l'eau et quantifier l'effet du changement climatique sur les espèces à partir des données issues de la valmométrie. Des collaborations pour étudier les données d'abondance seront naturellement envisagées.

Au niveau national, Thierry Huillet, chercheur CNRS à l'université de Cergy, travaille également sur les données d'abondance en écologie. En particulier, des notions de copule pour modéliser l'abondance de plusieurs espèces sont présentées dans Huillet (2018). Des discussions autour du présent projet sont aussi envisagées.

Enfin, au niveau international, le candidat aura l'occasion d'interagir avec des écologues de premier plan tels que Pablo Marquet et Sergio Navarette. En particulier faire le lien entre les modèles envisagées et les déterminants à prendre en compte pour l'étude des espèces sera au coeur de ce projet.

5 Contexte scientifique et partenarial: éléments généraux

Cette thèse sera menée au sein de l'UMR 8088 AGM à CYU et aura pour objectif de travailler sur un des axes majeurs du projet ECODEP, projet inclus dans ce même laboratoire. Elle sera cofinancée par le projet ECODEP sous réserve de l'approbation des instances du centre d'Excellence. Cette thèse a pour objectif de produire des avancées sur les deux axes "Population growth models and Ecological networks" et "Non-stationary time series and impact on the ecology and climate", voir http://doukhan.u-cergy.fr/ecodep_abstract.html et de pouvoir valoriser AGM et le Labex MME-DII à travers sa contribution par le biais de ses composante incluses dans CYU.

Le doctorant aura la possibilité et sera encouragé à échanger avec les membres du projet (en particulier, Paul Doukhan et Thierry Huillet de l'Université de Cergy, Gilles Durrieu de l'Université de Bretagne Sud ainsi que Pablo Marquet et Sergio Navarette de Pontificia Universidad Católica de Chile) qui sont spécialistes des séries temporelles et des données d'abondance en écologie.

Dans le contexte du projet, d'autres acteurs académiques travaillant à l'ISFA de Lyon, à University of Rockefeller et à University of Columbia aux Etats-Unis sont impliqués sur d'autres axes de travail tels que la loi de Taylor en écologie ou les problèmes de causalité dans les systèmes écologiques. Les problèmes liés à la loi de Taylor ont par ailleurs amené au recrutement récent d'un post-doctorant, Benjamin Bobbia.

Le projet ECODEP rassemble aussi d'autres acteurs associés, tels que l'Agrocampus Ouest (Gabriel Lang) ainsi que l'entreprise Biotope qui s'intéresse par exemple à la réduction des risques de collision entre chiroptères et éoliennes et qui sollicite des stages de M1/M2 au niveau de l'Ensai et de l'Université Bretagne Sud.

References

John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

Pierre Alquier and Paul Doukhan. Sparsity considerations for dependent variables. *Electronic journal of statistics*, 5:750–774, 2011.

Pierre Alquier, Karine Bertin, Paul Doukhan, and Rémy Garnier. High-dimensional var with low-rank transition. *Statistics and Computing*, pages 1–15, 2020.

Elita Baldridge, David J Harris, Xiao Xiao, and Ethan P White. An extensive comparison of species-abundance distribution models. *PeerJ*, 4:e2823, 2016.

Jean-Marc Bardet and Paul Doukhan. Non-parametric estimation of time varying ar (1)–processes with local stationarity and periodicity. *Electronic Journal of Statistics*, 12(2):2323–2354, 2018.

Barry W Brook and Corey JA Bradshaw. Strength of evidence for density dependence in abundance time series of 1198 species. *Ecology*, 87(6):1445–1451, 2006.

Julien Chiquet, Mahendra Mariadassou, Stéphane Robin, et al. Variational inference for probabilistic poisson pca. *The Annals of Applied Statistics*, 12(4):2674–2698, 2018.

Max Zinsou Debaly and Lionel Truquet. Iterations of dependent random maps and exogeneity in nonlinear dynamics. *To appear in Econometric Theory*, 2020.

Zinsou Max Debaly and Lionel Truquet. Stationarity and moment properties of some multivariate count autoregressions. *arXiv preprint arXiv:1909.11392*, 2019.

Paul Doukhan, Michael H Neumann, and Lionel Truquet. Stationarity and ergodic properties for some observation-driven models in random environments. *arXiv preprint arXiv:2007.07623*, 2020.

Jacob C Douma and James T Weedon. Analysing continuous proportions in ecology and evolution: A practical introduction to beta and dirichlet regression. *Methods in Ecology and Evolution*, 10 (9):1412–1430, 2019.

Konstantinos Fokianos, Bård Støve, Dag Tjøstheim, Paul Doukhan, et al. Multivariate count autoregression. *Bernoulli*, 26(1):471–499, 2020.

Thierry E Huillet. Stochastic species abundance models involving special copulas. *Physica A: Statistical Mechanics and its Applications*, 490:77–91, 2018.

Michael A Litzow, Lorenzo Ciannelli, Patricia Puerta, Justin J Wettstein, Ryan R Rykaczewski, and Michael Opiekun. Nonstationary environmental and community relationships in the north pacific ocean, 2019.

Robert MacArthur. On the relative abundance of species. *The American Naturalist*, 94(874):25–36, 1960.

Pablo A Marquet, Guillermo Espinoza, Sebastian R Abades, Angela Ganz, and Rolando Rebolledo. On the proportional abundance of species: Integrating population genetics and community ecology. *Scientific reports*, 7(1):1–10, 2017.

Guido Masarotto and Cristiano Varin. Gaussian copula marginal regression. *Electronic Journal of Statistics*, 6:1517–1549, 2012.

Brian J McGill, Rampal S Etienne, John S Gray, David Alonso, Marti J Anderson, Habtamu Kassa Benecha, Maria Dornelas, Brian J Enquist, Jessica L Green, Fangliang He, et al. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology letters*, 10(10):995–1015, 2007.

Lynne J Shannon, Marta Coll, and Sergio Neira. Exploring the dynamics of ecological indicators using food web models fitted to time series of abundance and catch data. *Ecological Indicators*, 9(6):1078–1095, 2009.

Jennifer L Shinen and Sergio A Navarrete. Lottery coexistence on rocky shores: weak niche differentiation or equal competitors engaged in neutral dynamics? *The American Naturalist*, 183(3): 342–362, 2014.

Lionel Truquet. Parameter stability and semiparametric inference in time varying auto-regressive conditional heteroscedasticity models. *Journal of the Royal Statistical Society Series B*, 79(5): 1391–1414, 2017.

Lionel Truquet. Local stationarity and time-inhomogeneous markov chains. *Annals of Statistics*, 47 (4):2023–2050, 2019.

Lionel Truquet. A perturbation analysis of markov chains models with time-varying parameters. *Bernoulli*, 26(4):2876–2906, 2020.

Michail Tsagris and Connie Stewart. A dirichlet regression model for compositional data with zeros. *Lobachevskii Journal of Mathematics*, 39(3):398–412, 2018.

Peter Turchin and Andrew D Taylor. Complex dynamics in ecological time series. *Ecology*, 73(1): 289–305, 1992.

Igor Volkov, Jayanth R Banavar, Stephen P Hubbell, and Amos Maritan. Neutral theory and relative species abundance in ecology. *Nature*, 424(6952):1035–1037, 2003.

Alan H Welsh, Ross B Cunningham, CF Donnelly, and David B Lindenmayer. Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling*, 88(1-3):297–308, 1996.

Seth J Wenger and Mary C Freeman. Estimating species occurrence, abundance, and detection probability using zero-inflated distributions. *Ecology*, 89(10):2953–2959, 2008.

Fukang Zhu. Zero-inflated poisson and negative binomial integer-valued garch models. *Journal of Statistical Planning and Inference*, 142(4):826–839, 2012.