Université de Rennes 1

Rapport de stage

M1 Modélisation en Ecologie

Apports du modèle Poisson log-normal dans l'analyse multivariée en écologie des communautés

---

Théo Fabien

Institut Agro Rennes-Angers	Maître de stage	Marie-Pierre Etienne
	Équipe encadrante	Jean-Louis Marchand Baptise Alglave Thomas Outrequin
Université de Rennes 1	Professeur référent	Andreas Prinzing

6 Avril - 10 Juin 2022

Soutenu à Rennes le 23 Juin 2022

Remerciements

Tout d'abord, je remercie Marie-Pierre Etienne, ma maître de stage pour avoir cru en mes capacités et pris le temps de me recevoir à plusieurs reprises avant même le début du stage. Sa patience et son suivi m'ont permis d'acquérir de nombreuses connaissances et de me dépasser dans les domaines des statistiques et de la rédaction scientifique. Sa confiance en moi a grandement participé à ma progression et m'a encouragé à prendre des initiatives.

Un grand merci également à Jean-Louis Marchand qui s'est toujours montré disponible pour m'aider. Ses conseils ont été autant d'éléments de réflexion pour faire avancer mon rapport, et son assistance sur certains points a été une aide indispensable. Je tiens aussi à le remercier pour l'attention qu'il a portée à la relecture de mon rapport.

Merci à Baptiste Alglave et Thomas Outrequin, étudiants en thèse au pôle Halieutique, Mer et Littoral qui m'ont fourni leurs jeux de données pour que j'y applique le modèle PLN, m'ont reçu régulièrement pour discuter de l'avancement de mon travail et de mes résultats, et ont également participé au travail de relecture.

Merci à Nathan, stagiaire de M2, pour avoir consacré du temps à m'expliquer le fonctionnement de l'algorithme EM. Je le remercie également pour m'avoir fait rencontrer Jean-Michel qui m'a apporté bonne humeur et soutien moral.

Merci à l'ensemble des stagiaires présents pour leur bonne humeur et l'atmosphère de travail incroyable qu'ils ont tous participé à créer au cours de cette période.

Enfin, merci beaucoup à toute l'équipe de l'unité pédagogique mathématiques appliquées de l'institut Agro, pour son accueil chaleureux et sa bienveillance.

Table des matières

1 Introduction

L'analyse statistique des communautés écologiques est devenu un domaine clé pour comprendre comment les espèces s'assemblent au sein des écosystèmes. À des fins de préservation, elle permet par exemple de comprendre quels facteurs environnementaux structurent les communautés d'espèces et en modifient la répartition [26]. Un autre exemple est son utilisation pour évaluer l'efficacité et les conséquences de mesures de conservation [22].

Dans ce contexte, le développement de méthodes quantitatives en écologie des communautés s'est accéléré au cours des dernières décennies. L'analyse multivariée de tableaux regroupant les abondances de plusieurs taxons sur de multiples échantillonnages est devenu un outil central [14, 36]. Ces données de comptage sont généralement caractérisées par une forte proportion de données nulles (l'espèce n'a pas été vue sur le site, on parle de forte présence de zéros) et des abondances d'ordres de grandeur différents. Pour les analyser, les approches d'ordination de type analyse en composantes principales (ACP, la première méthode d'ordination appliquée à des données écologiques [18]) et PCoA (pour Principal Coordinates Analysis) sont particulièrement populaires [14, 10]. Ces méthodes sont couramment utilisées pour résumer les données sur un espace réduit avant de procéder à une classification, dans le but de former des regroupements d'espèces assimilés à des communautés. L'ACP et la PCoA n'ont cependant pas été construites pour gérer des données de comptages. Ainsi, une transformation (généralement logarithmique) est couramment opérée pour réduire l'impact des espèces à forts effectifs. Sans cela, ces dernières peuvent nuire à la visualisation en impactant l'échelle des distances affichées sur l'espace d'ordination, ainsi qu'une éventuelle classification. Par ailleurs, d'autres travaux suggèrent que l'emploi de distances euclidiennes par l'ACP rend les performances de cette méthode très limitées en écologie des communautés [9, 27]. En effet, cette métrique est notamment connue pour mal gérer les données présentant une forte proportion de 0 [24] puisqu'elle traite comme une ressemblance entre deux sites le fait de présenter l'absence d'une même espèce. Si les données contiennent une forte proportion de 0, on observe donc couramment un regroupement de sites en un même point sur l'espace d'ordination, ce qui nuit à l'interprétation. Avec la PCoA, la métrique majoritairement utilisée est l'indice de dissimilarité de Bray-Curtis, qui varie de 0 à 100 et tient compte des effectifs totaux d'individus sur les sites, corrigeant le problème évoqué avec les distances euclidiennes.

Le modèle Poisson log-normal (PLN) a été introduit en 1989 par Aitchison et Ho [1] pour traiter des données de comptage et a reçu récemment une attention accrue notamment en biologie moléculaire. La composante Poisson du modèle est classique lorsqu'on cherche à analyser des données de comptage, par exemple dans le cas des régressions de Poisson. Pour apporter de la souplesse à ce modèle assez contraint, il est possible d'introduire des variables latentes, ce qui est aujourd'hui classique en écologie statistique [17]. Dans le cas du modèle PLN, la couche latente permet de représenter une dépendance entre les comptages de différentes espèces et de proposer assez naturellement une analyse multivariée de données de comptage. Il se décline

en plusieurs variantes selon les contraintes potentielles que l'on impose sur la couche cachée. Notamment, la variante PLNPCA permet de retrouver les aspects de réduction de dimension de l'ACP en imposant des contraintes de faible rang sur les variables latentes [7]. La variante PLN réseaux permet quant à elle d'identifier des formes de dépendance directes dans les abondances de chaque espèce [8] et offre une manière de revisiter la notion de communauté écologique, en adoptant un point de vue plus fonctionnel des relations entre espèces. Elle constitue une adaptation des modèles graphiques gaussiens à des données de comptage, ce qui rend accessible à l'étude des communautés ces méthodes classiques en génomique [37, 38].

Le travail présenté dans ce rapport explore les avantages et les limites de ces différentes approches, au travers de l'étude des communautés d'invertébrés benthiques dont les abondances sont mesurées par IFREMER au cours de campagnes scientifiques annuelles. Il permet également d'interroger la notion de communauté.

Dans un premier temps, le modèle PLNPCA, l'ACP et la PCoA ont été comparées. Nous nous sommes intéressés à leur capacité à retrouver des communautés connues et à la robustesse de leurs résultats. PLNPCA ouvre la possibilité d'un traitement de données de comptages obtenues avec des efforts d'échantillonnage variables. Nous nous attendions donc à ce qu'elle soit la plus performante.

Dans un second temps, la variante PLN réseaux (qui permet d'étudier l'interaction entre espèces conditionnellement aux autres espèces et à des covariables) a été mise en place pour adopter un point de vue différent des méthodes de réduction de dimension et évaluer le potentiel de cette approche sur ce type de données. Cela nous a aussi permis de questionner la notion de communauté.

2 Matériel et méthodes

2.1 Les données de campagne EVOHE

Le jeu de données utilisé est composé de macro-invertébrés benthiques. Il provient de la campagne EVOHE (Evaluation des ressources Halieutiques de l'Ouest de l'Europe) [20] réalisée chaque automne dans le golfe de Gascogne (France) entre 2008 et 2020. Chaque année, entre 80 et 90 sites sont choisis selon un échantillonnage stratifié et chalutés durant un temps cible de 30 minutes pouvant varier un peu selon les conditions de mer (temps moyen 29,91 min). Les abondances étudiées concernent 254 espèces sur 702 traits de chalut [1] (voir Figure 1).

Certaines analyses se sont focalisées sur un sous-ensemble d'espèces caractéristiques des grands types de milieux du golfe de Gascogne (tableau 1). Ces espèces ont une répartition et un mode de vie renseignés dans la bibliographie, pour faciliter l'interprétation et avoir une attente à priori sur les résultats. Elles ont été choisies principalement partir de l'étude réalisée par Brind'Amour et al. en 2014 [4], qui sépare 5 grands types de milieux du golfe de Gascogne

1. 86 taxons n'avaient pas été identifiés à l'espèce et n'ont donc pas été pris en compte dans nos analyses

et des espèces représentatives de ces milieux en confrontant des données observées avec la bibliographie préexistante. Nous avons sélectionné 13 de ces espèces, les autres étant absentes de notre jeu de données. Nous avons également ajouté *Adamsia palliata* dont le lien symbiotique avec *Pagurus prideaux* est déjà bien décrit [2].

Nous avons assimilé les groupes 1 à 5 à des représentants de 5 communautés. En retirant les traits de chalut ne contenant aucune espèce d'intérêt, ce jeu de données limité à ces 14 espèces a comporté 693 observations.

Des covariables environnementales ont été prises en compte pour chaque trait de chalut dans le cadre de l'analyse de réseaux. Ces covariables sont d'une part la profondeur mesurée et le type de substrat, et d'autre part les propriétés suivantes relevées en bas de la colonne d'eau : la concentration en oxygène, le pH, la quantité totale de carbone sous forme organique, la salinité et la température. Ces variables se sont déjà révélées pertinentes dans d'autres travaux pour caractériser les habitats marins [26, 32], en particulier dans l'étude des organismes benthiques.

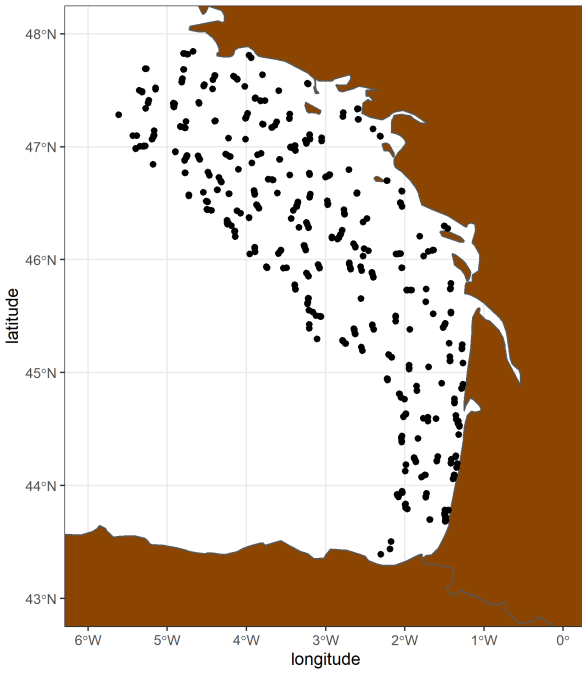

FIGURE 1 – Positions des traits de chaluts des campagnes EVHOE, entre 2008 et 2020.

TABLEAU 1 – Groupes formés avec des espèces caractéristiques de 5 grands types d'habitats du golfe de Gascogne [4]. Les espèces d'un même groupe sont assimilées à des représentantes d'une même communauté. Une brève description de chaque milieu est faite à titre indicatif.

Groupes	**1**	**2**	**3**	**4**	**5**
Milieu	150 à 350 mètres de profondeur, substrat mélange sable et vase	Milieu profond (500 mètres), vaseux	Milieu côtier, profondeur < 50 mètres	Substrat sableux, profondeur > 100 mètres	Grande vasière, profondeur de 100 à 120 mètres
Espèces	*Hyalinoecia turbicola, Leptometra celtica, Pagurus prideaux, Adamsia palliata*	*Pasiphaea sivado, Pagurus alatus*	*Ophiura ophiura, Asterias rubens*	*Astropecten irregularis, Macropipus tuberculatus*	*Nephrops norvegicus, Munida rugosa, Alpheus glaber, Munida intermedia*

2.2 Comparaison entre les méthodes d'identification des communautés

Des approches classiques de réduction de dimensions suivies d'une classification ascendante hiérarchique (CAH), utilisant Ward comme distance inter groupes, ont été mises en place pour identifier des communautés. La CAH a été faite sur un espace réduit à deux dimensions.

L'étape de réduction de dimension a été faite d'abord selon des méthodes classiquement utilisées ; une ACP centrée non réduite et une PCoA utilisant la dissimilarité de Bray-Curtis.

Cette dernière est la métrique la plus utilisée pour l'analyse de données d'assemblage [31]. Chacune de ces analyses a été testée selon les configurations suivantes : pas de manipulation des données (désignée dans la suite par ACP_x et $PCoA_x$), données log-transformées (ACP_t et $PCoA_t$), données divisées par l'aire chalutée (standardisation de l'effort d'échantillonnage) puis log-transformées (ACP_{st} et $PCoA_{st}$). Nous désignons par log-transformation le processus classiquement utilisé consistant à passer au logarithme après avoir ajouté une constante. La constante la plus naturelle pour des données de comptage est 1 ; pour des abondances nulles, la valeur obtenue après transformation est ln(0+1) = 0. Dans la suite, ACP et PCoA désigne ces méthodes dans toutes leurs configurations.

Le modèle PLNPCA permet également de procéder à une réduction de dimensions, mais offre aussi plus de souplesse de modélisation comme classiquement dans les régressions de Poisson. En effet, la structure latente du modèle PLN permet de rendre compte des dépendances. Plus précisément, on note $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ip})$ le vecteur des abondances observées pour le site i, p désignant le nombre total d'espèces considéré et Y_{ik} l'abondance pour l'espèce k sur le site i. On écrit comme dans un modèle linéaire généralisé que :

$$Y_{ij} \mid Z_{ij} \; i.i.d \sim \mathscr{P}(exp(Z_{ij}))$$
$$\log E(\boldsymbol{Y}_i) = \boldsymbol{Z}_i \sim \mathscr{N}_p(\mu_i, \Sigma)$$

en faisant ainsi apparaître le vecteur latent $\boldsymbol{Z}$ dont la matrice de covariance Σ de dimension $p \times p$ décrit les liens de dépendances entre les abondances moyennes de chaque espèce. On peut alors tirer partie de la grande flexibilité du modèle linéaire généralisé pour intégrer un effort d'échantillonnage variable d'un site à l'autre grâce à un offset et dans ce cas écrire

$$E(\boldsymbol{Z}_i) = o_i + \mu_i,$$

o_i étant connu et représentant la mesure d'effort pour le site i. Enfin, il est également très simple de lier l'abondance attendue à des covariables :

$$E(\boldsymbol{Z}_i) = o_i + \boldsymbol{x}_i \boldsymbol{B}.$$

où $\boldsymbol{x}_i$ est le vecteur des covariables pour le site i et $\boldsymbol{B}$ est la matrice des coefficients de régression.

La matrice de variance-covariance Σ et la matrice $\boldsymbol{B}$ sont inférées à partir des données à l'aide d'une version variationnelle de l'algorithme Expectation Maximisation [13] qui ne sera pas détaillée ici (voir [6]).

PLNPCA est un modèle PLN dans lequel on impose des contraintes de réduction de dimension (contraintes de rang) sur Σ, le nombre de dimensions pouvant être choisi à l'aide d'un critère statistique, classiquement BIC (critère d'information bayésien). Ainsi, PLNPCA revient à réaliser une ACP probabiliste [35] sur les vecteurs latents $\boldsymbol{Z}$. Aucune transformation manuelle des données n'est faite si ce n'est au travers du lien "ln()" qui lie espérance du modèle de Pois-

son et variables dans l'espace latent. Dans la suite, PLNPCA_x (respectivement PLNPCA_{off}) désigne la mise en oeuvre de PLNPCA sans (respectivement avec) prise en compte de l'offset calculé par le produit du temps de chalutage et de la largeur du chalut. PLNPCA désigne la méthode dans toutes ses configurations.

Pour comparer ces 3 grandes classes de méthodes de réduction de dimension dans leur capacité à retrouver des assemblages connus, elles ont d'abord été mises en place à partir des 14 espèces sélectionnées (tableau 1). Les covariables n'ont pas été prises en compte dans PLNPCA, pour qu'il reste comparable avec les autres méthodes. Pour estimer la capacité des méthodes a séparer les groupes attendus, le nombre de groupes faits par la CAH a été déterminé automatiquement en sélectionnant un nombre de groupes tel que la perte relative d'inertie soit maximale par rapport au nombre de groupes + 1.

Ensuite, la manière dont chaque méthode (sauf l'ACP_{st} et la PCoA_{st} qui sont exclues des prochaines analyses) répartit les espèces dans l'espace réduit a été évaluée selon une mesure de similarité par rapport aux répartitions attendues. Toujours sur les 14 espèces choisies, 5 groupes ont cette fois été imposés lors de l'étape de classification. Pour générer des réplicats et minimiser l'effet d'outliers potentiels parmi les sites, les regroupements ont été évalués à 30 reprises pour chaque méthode, en sélectionnant à chaque fois 400 sites au hasard. A chaque itération, cinq groupes ont été fixés (valeur attendue). L'accord entre les regroupements des espèces après CAH et ceux faits dans le tableau 1 a été évalué en calculant l'indice de Rand entre ces partitionnements. Il prend la valeur 1 si tous les paires d'éléments se retrouvent classées de la même manière dans les 2 classifications comparées. Il peut théoriquement valoir 0 si les paires d'éléments groupés et séparés de la première partition sont respectivement séparées et groupées dans la seconde. La distribution des indices de Rand entre deux regroupements indépendants de 14 individus en 5 classes a également été approchée en effectuant 1000 comparaisons de regroupements aléatoires. Le but a été d'estimer si les indices de Rand observés avec les données était significativement différent de ce qui aurait été obtenu aléatoirement.

L'impact du pourcentage de 0 sur chaque méthode a été illustré ensuite. Comme indiqué en introduction, l'ACP est connue pour gérer des données surchargées en 0 en regroupant l'essentiel des observations en un point. Nous avons illustré cet effet, puis observé si les autres méthodes se comportaient de manière similaire. Les 254 espèces ont été triées par ordre décroissant selon leurs pourcentages de présence parmi les traits de chalut. Les 160 premières ont été conservées et regroupées par paquets de 20, de manière à ce que la fraction de jeu de données formée des 20 premières ait le plus fort taux de présence (donc le plus faible pourcentage de 0), et que le taux de présence soit le plus faible pour la fraction formée des espèces 141 à 160, par exemple. Les ACP et PCoA appliquées sont l'ACP_t et la PCoA_t. Cette analyse a été réalisée avec une sélection automatique du nombre de groupe (voir le paragraphe précédent) puis avec 5 groupes imposés.

2.3 La communauté comme un réseau

La notion de communauté écologique qui est illustrée par les méthodes comparées précédemment a été questionnée par l'étude de réseaux. Au lieu de rapprocher les espèces trouvées fréquemment ensemble, l'inférence de réseaux adopte une autre vision qui consiste à établir des liens de dépendance directe entre espèces. Avec des données d'abondances, il est possible de rapprocher le modèle PLN des modèles graphique gaussiens, ce qui rend possible l'inférence de réseaux sur des comptages.

Le modèle PLN réseaux permet d'inférer des réseaux d'interaction en détectant les espèces en interactions directes, sans poser d'hypothèse sur les types de ces interactions (positives ou négatives). Deux espèces sont en interaction directe s'il existe une dépendance entre leurs abondances respectives conditionnellement à toutes les autres espèces ainsi qu'aux variables environnementales. Cette définition a été développée par Popovic et al. en 2019 [29]. Elle repose en statistique sur la notion de corrélation partielle et est également connue sous le terme de dépendance causale [5]. La prise en compte des covariables permet d'éviter qu'une interaction directe ne soit détectée entre deux espèces n'ayant pas d'interaction biotique entre elles mais partageant simplement des habitats similaires. Grâce à des techniques de régularisation statistique, on peut faire apparaître des indépendances conditionnelles entre espèces. Ceci peut se résumer à l'aide d'une matrice d'adjacence ou d'un graphe dans lequel les nœuds représentent des espèces et les arêtes des liens de dépendance directe. Le critère de régularisation choisi est le critère BIC.

La variante PLN réseaux a été mise en place d'abord sur les espèces du tableau 1, puis sur l'ensemble du jeu de données. Les réseaux ont été inférés d'abord sur le tableau de comptage uniquement, puis en ajoutant les covariables pour identifier les dépendances débarrassées des effets de l'environnement. Dans les deux cas, la prise en compte ou non des offsets n'impactait pas les réseaux.

Les analyses réalisés dans ce rapport ont été faites en utilisant R [30] sous `Rstudio` [33]. Les variantes du modèle PLN ont été générées grâce au package `PLNmodels` [6]. Le package `vegan` [21] a été utilisé pour la méthode PCoA. L'ACP et les CAH ont été réalisées avec `FactoMineR` [23]. Les indices de Rand ont été calculés avec le package `fossil` [25].

3 Résultats

3.1 Comparaison entre PLN, ACP et PCoA

Lecture qualitative des classifications obtenues

La méthode PCoA a fourni les résultats les plus proches des connaissances déjà établies (figure 2, d et f). Que ce soit pour des données transformées ou non, 3 groupes ont été obtenus avec cette méthode. Le groupe 1 du tableau 1 est intégralement retrouvé dans la PCoA,

comme c'est également le cas avec toutes les autres méthodes. *Pagurus alatus* et *Pasiphaea sivado*, qui forment le groupe 2, sont regroupées également avec cette méthode. *Ophiura ophiura* et *Asterias rubens*, qui forment le groupe 3, sont classées ensemble avec la $PCoA_t$. Avec la $PCoA_x$, elles sont dans des groupes différents, mais la distance qui sépare les deux points est la deuxième distance la plus faible pour des espèces de groupes différents. Les espèces du groupe 4, *Astropecten irregularis* et *Macropipus tuberculatus* sont bien regroupées dans la PCoA. Les espèces du groupe 5 sont partiellement séparées cependant, que ce soit avec des données non transformées (*Nephrops norvegicus* et *Munida rugosa* dans un groupe, les deux autres espèces dans un autre groupe) ou transformées (*Munida intermedia* séparée).

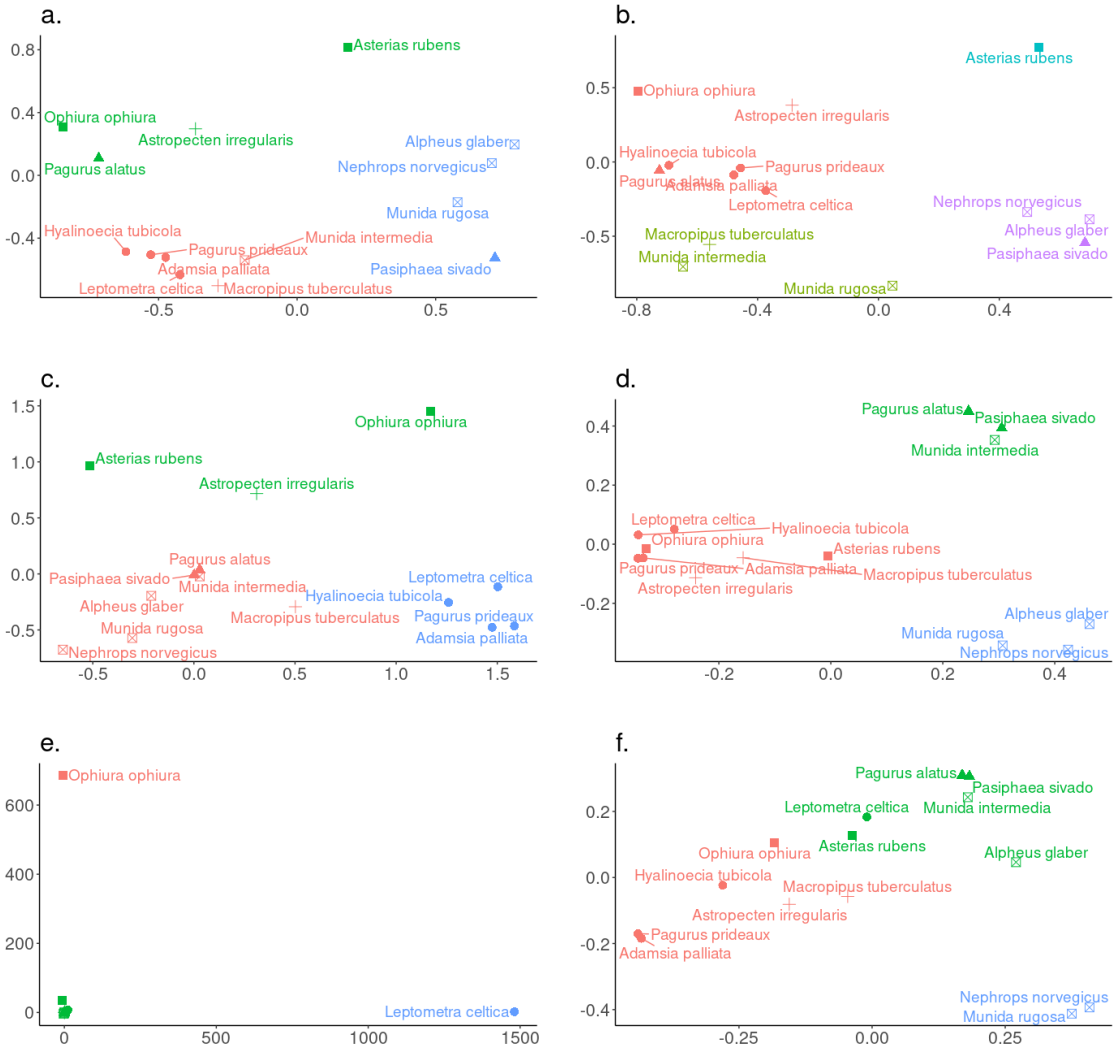

FIGURE 2 – Regroupements sélectionnés par l'algorithme de la classification ascendante hiérarchique, après avoir appliqué chacune des méthodes. a. et b. : modèles $PLNPCA_x$ et $PLNPCA_{off}$ (sans et avec offsets). c. et d. : ACP_t et $PCoA_t$ (données log-transformées). e. et f. : ACP_x et $PCoA_x$ (données non transformées). La forme des points correspond aux 5 groupes attendus.

Pour le modèle PLNPCA et l'ACP_t, on retrouve certains regroupements faits par les experts, mais le résultat est moins convaincant que pour la PCoA. De plus, l'offset, bien qu'il impacte visuellement le modèle PLNPCA (formation de 4 groupes), n'améliore pas son interprétation biologique (figure 2, b). La proximité des espèces du groupe 2 n'est pas retranscrite dans le

modèle PLNPCA, alors que c'est le cas pour l'ACP qui les place dans un même groupe. Les espèces du groupe 3 sont séparées pour l'ACP et le modèle $PLNPCA_{oft}$. Les espèces du groupe 4 sont séparées par le modèle PLN et l'ACP_x. Enfin, on retrouve le groupe 5 intégralement dans l'ACP, mais seulement en partie dans le modèle $PLNPCA_x$ (*Munida intermedia* séparée) et $PLNPCA_{oft}$ (*Alpheus glaber* et *Nephrops norvegicus* sont dans un groupe, les deux autres espèces dans un autre groupe).

L'ACP_x n'a pas fourni de résultats exploitables (figure 2, e). On constate une déformation importante, avec un regroupement de 12 espèces au niveau de l'origine du repère, et deux espèces qui s'en démarquent largement, *Ophiura ophiura* et *Leptometra celtica*, qui sont les deux espèces les plus abondantes parmi celles utilisées ici. L'interprétation biologique est compromise par le fait que la méthode échoue à retranscrire la plupart des distances.

La $PCoA_{st}$ et l'ACP_{st} n'ont pas été représentées, sachant que les points étaient répartis identiquement à ceux obtenus pour des données non transformées. Le tableau d'abondances correspondant contenait uniquement des valeurs inférieures à 0,001.

Evaluation des méthode par un critère statistique : l'indice de Rand

Les résultats obtenus avec l'indice de Rand sont consistants avec la partie précédente (figure 4). L'indice de Rand moyen le plus élevé a été observés pour la PcoA (0.8799 que ce soit avec ou sans transformation). Avec les méthodes ACP_x, ACP_t, $PLNPCA_x$ et $PLNPCA_{oft}$, nous avons obtenu respectivement des valeurs moyennes de 0.6934, 0.8223, 0.8124 et 0.8113.

Les indices de Rand obtenus à partir de regroupements aléatoires étaient condensées et n'exploitaient pas la gamme de variation théorique de cet indice (voir figure 3). En effet, 95% des valeurs étaient situées entre 0,6727 et 0,6909. Les valeurs moyennes obtenues pour les méthodes testées sont donc significativement plus élevées que la valeur obtenue avec une distribution aléatoire.

La transformation des données a un effet important sur l'ACP, mais n'affecte pas la PCoA qui a conservé des regroupements identiques. Graphiquement, les indices de Rand obtenus pour l'ACP_x sont nettement inférieurs à ceux obtenus pour les autres méthodes. De plus, les indices de Rand obtenus pour cette méthode varient beaucoup plus que pour les autres méthodes.

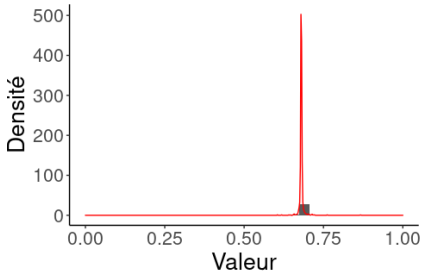

FIGURE 3 – Distribution de l'indice de Rand entre deux regroupements indépendants de 14 individus en 5 classes.

Impact du pourcentage de 0 dans le jeu de données

Quand la quantité de 0 augmente, un groupe unique tend à contenir l'ensemble des individus pour l'ACP, et dans une moindre mesure pour la PCoA (figure 5). Plus les espèces utilisées présentent des taux de présence faibles (c'est-à-dire des pourcentages de 0 élevés sur l'ensemble des sites), plus l'ACP tend à former un groupe qui contient l'ensemble des espèces, que ce soit

pour un nombre de groupes imposés ou non. Pour la PCoA, l'effet est très léger.

Le modèle PLNPCA semble plus robuste au pourcentage de 0 dans le jeu de données. Que le nombre de groupes soit imposé ou non, de faibles variations dans la taille du plus gros groupe formé avec le modèle PLN ont été observées. Cependant, aucune tendance n'a été observée.

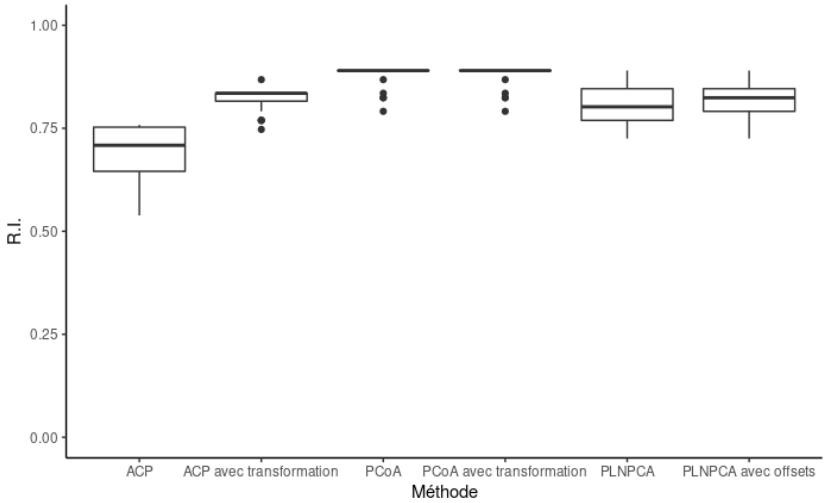

FIGURE 4 – Indices de Rand (R.I.) obtenus en comparant les résultats de la CAH pour 5 groupes imposés avec le regroupement théorique établi, pour chaque méthode étudiée. 30 jeux de données de 400 observations ont été créés aléatoirement à partir du jeu de données principal (693 observations). Les transformations réalisées sont de la forme log(abondances + 1).

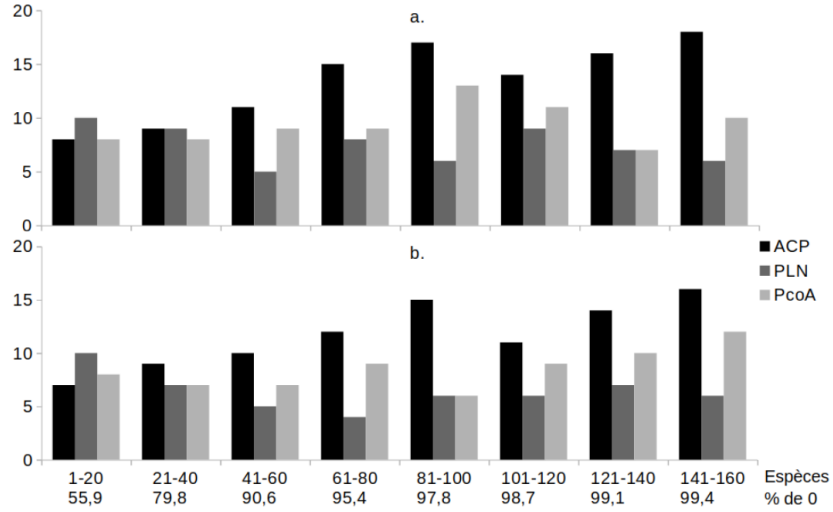

FIGURE 5 – Taille du plus gros cluster observé pour chaque méthode suite à une CAH, pour des jeux de données avec différents pourcentages de 0. Toutes les espèces ont été classées par ordre croissant de proportion de 0, et des groupes de 20 espèces ont été formés. a : nombre de groupes libre. b : 5 groupes imposés lors de la classification. L'ACP et la PCoA ont été faites sur des données log-transformées.

3.2 Exploration des interactions directes avec le modèle PLN réseaux

Dans le modèle PLN réseaux, la prise en compte des covariables permet d'abord de retirer un grand nombre d'arêtes dans le réseau, que ce soit pour les 14 espèces sélectionnées précédemment (figure 6) ou avec toutes les espèces du jeu de données (figure 7). Avec 14 espèces, 19

arêtes ont été observées dans le réseau établi sans prendre en compte les covariables. Après les avoir pris en compte, le nombre d'arêtes s'est abaissé à 11. Par exemple, les liens qui liaient *Leptometra celtica* à beaucoup d'autres espèces dans le premier réseau ont pour la plupart disparu. Avec toutes les espèces du jeu de données, le réseau est passé de 501 arêtes à 33 (diminution de 93,4%).

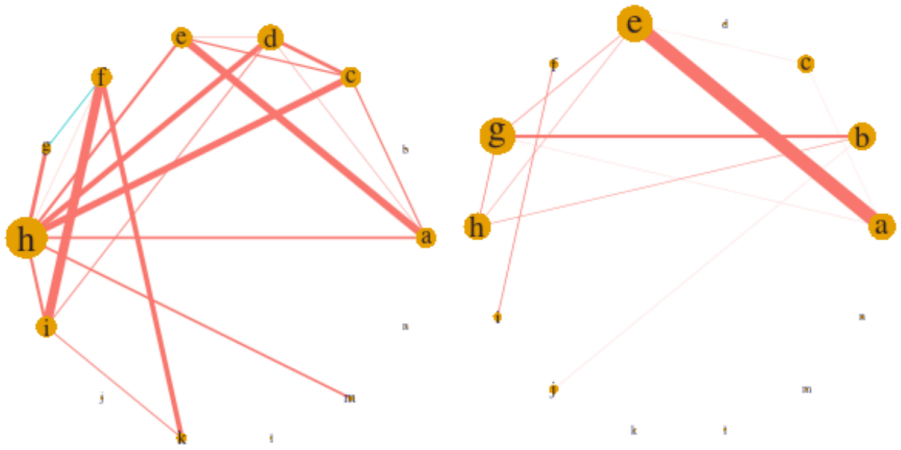

FIGURE 6 – Réseaux d'interactions inférés pour les invertébrés benthiques du golfe de Gascogne, d'abord sans covariables (gauche), puis avec les covariables et les offsets (droite). La largeur des segments est proportionnelle à la valeur absolue des corrélations partielles. La couleur des segments indique le signe de ces dernières : bleue si la corrélation est négative, rouge sinon. La taille d'un nœud est proportionnelle à son nombre de connexions. Correspondances lettres - espèces ; a : *Adamsia palliata*, b : *Astropecten irregularis*, c : *Hyalinoecia tubicola*, d : *Macropipus tuberculatus*, e : *Pagurus prideaux*, f : *Nephrops norvegicus*, g : *Ophiura ophiura*, h : *Leptometra celtica*, i : *Munida rugosa*, j : *Asterias rubens*, k : *Alpheus glaber*, l : *Pagurus alatus*, m : *Munida intermedia*, n : *Pasiphaea sivado*

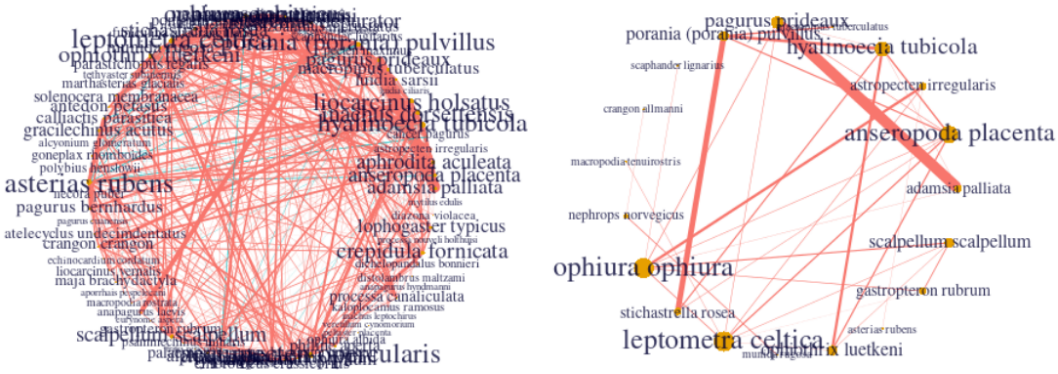

FIGURE 7 – Réseaux d'interactions inférés pour l'ensemble des invertébrés benthiques du golfe de Gascogne, d'abord sans covariables (gauche), puis avec les covariables et les offsets (droite). Les espèces n'étant reliées à aucune autre n'ont pas été représentées. La largeur des segments est proportionnelle à la valeur absolue des corrélations partielles. La couleur des segments indique le signe de ces dernières : bleue si la corrélation est négative, rouge sinon. La taille d'un nœud est proportionnelle à son nombre de connexions.

L'intégration des covariables dans le modèle a également révélé des liens entre espèces qui n'était pas visibles autrement. C'est le cas par exemple dans le réseau de 14 espèces, où l'interaction entre *Ophiura ophiura* et *Astropecten irregularis* n'est apparue qu'après avoir intégré les covariables. Avec toutes les espèces, des liens forts sont révélés entre *Pagurus prideaux* et

Adamsia palliata, et entre *Stichastrella rosea* et *Porania (porania) pulvillus*. Des hubs sont également apparus, c'est le cas de *Ophiura ophiura* qui s'est trouvée connectée à 4 autres espèces dans la figure 6 et à 8 autres espèces dans la figure 7.

Certains liens correspondent bien à ce que l'on peut lire dans la bibliographie. Que ce soit avec 14 espèces ou toutes les espèces du jeu de données, le lien symbiotique entre *Adamsia palliata* et *Pagurus prideaux* est très visible une fois les covariables prises en compte.

4 Discussion

4.1 Comparaison entre PLN, ACP et PCoA

Lecture qualitative des classifications obtenues

Aucune méthode n'a retrouvé les 5 groupes attendus. Les groupes du tableau 1 n'ont été que partiellement retrouvés par la plupart des approches, à l'exception du groupe 1 dont les espèces ont été regroupées dans tous les cas. Sur l'ensemble des méthodes, les groupes ont tous été retrouvés plus ou moins précisément, ce qui suggère que la sélection effectuée sur la base de la bibliographie est cohérente.

L'ACP_x n'a pas fourni de résultats interprétables : toutes les espèces étaient agrégées à l'exception des deux espèces les plus abondantes. Ces premiers résultats sont en accord avec nos attentes. L'ACP utilise des distances euclidiennes, qui ont tendance à trop mettre en valeur les espèces abondantes pour des données non transformées. Plus précisément, cela vient sans doute du passage au carré des différences entre coordonnées opéré dans la formule de la distance euclidienne. La transformation des données semble néanmoins régler ce problème.

Les points obtenus avec l'ACP_{st} et la $PCoA_{st}$ étaient répartis de la même manière que pour l'ACP_x et la $PCoA_x$. Le tableau des abondances divisées par les aires contenait des valeurs toutes inférieures à 0,001. Pour de telles valeurs, la transformation logarithmique utilisée a un effet négligeable et tend vers la fonction identité. On pourrait chercher à changer l'unité de l'aire échantillonnée pour réduire cet effet, mais dans tous les cas, le fait que prendre en compte l'effort d'échantillonnage par une division a des conséquences que l'on ne contrôle pas sur la répartition des points. Ce procédé impacte notamment l'effet cherché par une transformation. Il n'a donc rien d'anodin sur la visualisation des données. Le modèle $PLNPCA_{oft}$ a été la seule méthode où 4 groupes ont été obtenus après la CAH. Aucun de ces groupes ne correspondait précisément à l'un des 5 groupes établis à priori. Bien que cette approche permette d'obtenir le nombre de groupes le plus proche du nombre attendu, un examen plus précis laisse à penser que les offsets n'améliorent pas l'interprétation biologique. Cela peut être dû à un mauvais calcul des offsets, où à une sensibilité trop importante du modèle quand le nombre d'espèces est faible. Le fait que le protocole d'échantillonnage ait été standardisé suggère un impact moindre des offsets, mais les modifications observées suggèrent que même dans cette situation leur impact est observable. Un calcul soigneux de ces valeurs pourrait donc être indispensable pour des

résultats fiables. D'autres études sont nécessaires pour étudier l'apport des offsets sur des jeux de données de plus grande taille.

Mis à part les cas de l'ACP_x, de l'ACP_{st} et de la $PCoA_{st}$, les méthodes semblent complémentaires. Pour un nombre de groupes non choisi à l'avance, les approches PLN, PCoA et ACP donnent des résultats différents. La PCoA se démarque très légèrement, mais elle ne capte pas certaines proximités entre espèces qui sont détectées par les autres méthodes. L'analyse la plus précise d'un jeu de données de comptages pourrait donc être faite en comparant les résultats de ces méthodes. Par ailleurs, le développement d'un outil statistique permettant de combiner les résultats de plusieurs méthodes d'ordination pourrait être d'un grand intérêt.

Evaluation des méthode par un critère statistique : l'indice de Rand

Les méthodes ont montré des performances similaires en comparant les 5 groupes imposés au partitionnement établi à priori, avec des résultats plus proches de ce partitionnement que ce qui aurait été obtenu aléatoirement (ce qui nous conforte dans l'idée que nos attentes à priori étaient cohérentes). Cependant, l'ACP_x a encore une fois montré une performance inférieure aux autres méthodes, ainsi qu'une plus faible robustesse aux modifications du jeu de données. Les fortes variations de l'indice de Rand avec cette approche peuvent s'expliquer par l'impact, variable selon les sites, d'espèces dominantes dans le jeu de données que cette méthode ne gère pas correctement. La PCoA a donné les meilleurs résultats. Le fait que la PcoA ait fourni les meilleurs résultats va dans le sens d'études précédentes. En effet, des travaux faits sur des données simulées ont déjà montré que la dissimilarité de Bray-Curtis constituent une métrique fiable pour l'analyse de données de comptages [15]. Ces mêmes travaux suggèrent par ailleurs que les distances euclidiennes sont moins performantes dans ce domaine. Cependant nous ne nous attendions pas à ce que la PCoA soit plus puissante que l'approche PLNPCA. Un travail complémentaire doit être fait pour déterminer si ce constat est reproduit sur d'autres données.

Aucune différence n'a été détectée entre $PCoA_x$ et $PCoA_t$. Cela est consistant avec la première analyse, où l'interprétation biologique de la PCoA n'était pas améliorée par une transformation. Pourtant, les études publiées effectuent couramment une transformation logarithmique avant de procéder à une PCoA avec les dissimilarités de Bray-Curtis [34].

La distribution de l'indice de Rand entre deux regroupements indépendants de 14 individus en 5 classes était fortement condensée autour du maximum. Cela peut être dû au faible nombre d'espèces, qui génère nécessairement une interaction entre les groupes. Si 14 espèces sont regroupées en 5 groupes, il sera impossible de séparer toutes les espèces en théorie regroupées ensemble, donc d'obtenir un indice de Rand de 0. Par sa faible gamme de variations forcée par notre situation, l'indice de Rand constitue donc un critère limité. Des analyses futures gagneraient à le remplacer par un critère moins dépendant du nombre d'espèces et de groupes.

Impact du pourcentage de 0 dans le jeu de données

Quand la quantité de 0 augmente, l'ACP tend à former un groupe unique contenant l'essentiel des espèces. Cela est cohérent avec les travaux précédents qui mettent en cause l'utilisation des distances euclidiennes [9]. Cette métrique rapproche systématiquement des observations qui ont en commun des absences d'espèces. Cela n'a pas de sens écologique, car ces observations ont uniquement en commun le fait de contenir essentiellement des 0. Nos résultats suggèrent un phénomène similaire, mais plus léger, pour la PCoA. D'autres travaux qui ont suggéré que la dissimilarité de Bray-Curtis devient instable pour des données surchargées en 0 [11]. Par exemple, si deux observations contiennent chacune un individu, cette dissimilarité prend la valeur 0 si les individus sont d'espèces différentes et 100 si ils sont de la même espèce [11]. Cela pourrait expliquer ce que nous avons observé. Pour l'ACP comme pour la PCoA, avec un nombre important de 0 les groupes obtenus ne traduisent plus une réalité biologique. Nos résultats nous encouragent à penser que l'approche PLNPCA serait donc la méthode la plus stable pour des données surchargées en 0. D'autres études sont nécessaires pour montrer qu'il s'agit bien d'une robustesse réelle et pas d'une déformation différente des données.

Les approches classiques et la notion de communauté

Dans une première partie, nous avons adopté une approche classique. Les communautés ont été constituées en regroupant des espèces retrouvées fréquemment ensembles dans les traits de chalut. Le modèle PLN permet d'adopter un autre point de vue en adaptant l'inférence de réseaux aux données de comptage. Cette approche modélise des liens qui relient les espèces, qui peuvent être liés à l'habitat comme à des interactions directes. Cela ouvre donc la possibilité de définir la notion de communautés autrement à travers nos analyses statistiques.

4.2 Le modèle PLN réseaux

Les covariables structurent fortement les corrélations entre espèces. La disparition d'un grand nombre d'arêtes après avoir pris en compte les covariables montre la prédominance des "fausses interactions" dans le réseau sur le tableau de comptage uniquement. Un deuxième effet de cet ajout est l'apparition de nouvelles arêtes. Cela suggère que le modèle détecte mieux certains liens si les covariables sont prises en compte.

L'analyse de la pertinence biologique du modèle PLN réseaux est difficile, sachant que les interactions entre les espèces d'intérêt ont été peu étudiées. Il est probable que certaines interactions détectées ne soient pas encore connues, ce qui implique une investigation pour les valider ou non. La bibliographie existante nous a permis de valider la consistance écologique de certains liens identifiés. Dans les réseaux où les covariables sont prises en compte, la symbiose entre *Adamsia palliata* et *Pagurus prideaux* a bien été retrouvée. A notre connaissance, parmi le reste de nos données, aucune autre relation de symbiose ou de co-évolution n'est connue.

Le réseau nous a donc permis d'identifier un lien fort entre ces deux espèces, qui est cohérent avec les connaissances que nous avions précédemment [19, 2]. Que ce soit avec 14 ou 254 espèces, *Ophiura ophiura* figure parmi les espèces ayant le plus grand nombre de connexions. *Ophiura ophiura* est considérée comme un élément important des réseaux trophiques [3], étant donné qu'elle est à la fois un prédateur généraliste et une proie pour une grande diversité d'organismes (notamment crustacés et échinodermes). Avec les covariables choisies, il semble donc que les réseaux que nous avons obtenus sont suffisamment puissants pour identifier des liens trophiques. Cependant, il est possible que ces liens soient toujours parasités ou masqués par des covariables ou espèces non prises en compte. La plupart des autres liens restent difficiles à interpréter. Par exemple, la figure 7 suggère une interaction forte entre *Porania (porania) pulvillus* et *Stichastrella rosea*. A notre connaissance cette interaction n'est à ce jour pas connue, et il est possible qu'ici des variables n'aient pas été prises en compte. D'autres travaux ont notamment suggéré que ces deux espèces pouvaient être attirées par des sources de nourriture similaires [16]. Il est possible que ces organismes n'interagissent pas, mais que nous n'ayons simplement pas pris en compte les bonnes covariables.

Avec 14 espèces, le réseau inféré sans covariables a mis en valeur un hub constitué par *Leptometra celtica*. Cette espèce est un crinoïde suspensivore caractéristique de zones hautement productives [12, 28]. Il n'est donc pas surprenant de détecter des relations de co-occurence avec la plupart des autres espèces. Pour autant, cette espèce n'interagit pas directement avec les autres espèces, ce qui est cohérent avec le fait que les liens observés tendent à disparaître en ajoutant les covariables au modèle. Même si la première version du réseau contient de "fausses interactions", elle reste donc un sujet d'intérêt pour l'interprétation, notamment lorsqu'on la compare au réseau avec covariables.

PLN réseaux permet donc d'identifier des interactions directes, mais une interprétation poussée peut être délicate. D'abord, le signe des interactions est ambigu. Une interaction proie-prédateur s'interprète intuitivement comme une interaction négative. Mais si on considère que la prédateur réside là où les proies sont abondantes, il s'agit d'une interaction positive. Prendre en compte les dynamiques temporelles pourrait être une perspective d'évolution du modèle palliant ce problème. Une autre limite est l'intensité de l'interaction. Pour certaines interactions, il pourrait être difficile de prédire si elles seront effectivement visibles. Nous avons identifié clairement une relation symbiotique entre deux espèces, mais il est fort probable qu'une relation proie-prédateur soit nettement moins visible, surtout si le prédateur est généraliste, auquel cas la dépendance à la proie est moindre.

Beaucoup de liens restent inexpliqués dans les réseaux observés. Cela peut venir d'un manque de connaissances sur le milieu d'étude. Plus vraisemblablement, il est probable que certaines covariables ou espèces non capturées influencent les interactions inférées entre les espèces du jeu de données. Le travail à effectuer pour choisir les covariables environnementales serait donc une étape primordiale pour que cette approche fonctionne. Le choix des espèces n'est pas anodin non plus et influence le résultat. Étudier les communautés écologiques revient

à s'intéresser à des systèmes ouverts et complexes. Les utilisateurs de la méthode PLN réseaux seront donc toujours exposés à de fausses interactions entre espèces, qui n'ont en commun qu'une covariable ou une espèce tierce. Pour autant, cette méthode met bien en valeur des liens forts tels que des relations symbiotiques.

5 Conclusion

Le modèle PLN constitue une approche prometteuse pour l'analyse des communautés écologiques. En tant que méthode d'ordination, la puissance de l'approche PLNPCA est comparable à celle d'une PCoA ou d'une ACP pour laquelle les données auraient été transformées. Pour des données présentant un nombre important de 0, elle serait la méthode la plus robuste. PLNPCA pourrait donc être privilégiée en particulier pour ce type de données rencontrées fréquemment par les écologues. De plus cette approche permet de se soustraire aux choix arbitraires de la métrique utilisée (choix fait lors de la PCoA), de la transformation des données et de la prise en compte de l'effort d'échantillonnage. La variante PLN réseaux, bien que limitée dans son efficacité et impliquant un choix rigoureux des variables, permet d'approcher les communautés sous un angle plus fonctionnel et nous a fait nous questionner sur notre définition de communauté. Les méthodes de réduction de dimension forcent l'utilisateur à assimiler une communauté à un groupe d'individus trouvés fréquemment ensemble. Le modèle PLN est plus flexible et permet éventuellement de modéliser des dépendances conditionnellement aux facteurs biotiques et abiotiques.

L'intérêt principal de ce modèle réside dans le panel de variantes qu'il présente, à partir d'une base commune. Des travaux supplémentaires pourraient analyser d'autres variantes telles que PLNLDA ou PLN mixture. Ce modèle pourrait faciliter l'accès aux analyses statistiques pour des écologues dont ce n'est pas la spécialité. La compréhension de cette méthode uniquement pourrait permettre de mettre en place rapidement une large gamme d'analyses complexes sur des tableaux de comptage.

Pour conclure, ce stage a constitué une première exploration de l'utilisation du modèle PLN en écologie des communautés. Sachant que nos analyses ont porté sur des données réelles, les reprendre sur d'autres jeux de données est important. Comparer des approches multivariées sur des données réelles est un travail que nous considérons comme primordial. Les données simulées sont populaires [14] car elles permettent un meilleur contrôle et une évaluation plus précise des méthodes. Cependant elles se limitent toujours à des jeux de données simplistes. Utiliser des données réelles permet une mise à l'épreuve des méthode dans des cas "non idéaux". La production de travaux similaires avec d'autres jeux de données est donc nécessaire pour confirmer ou infirmer les résultats obtenus.

Bibliographie

[1] J. AITCHISON et C. H. HO. "The multivariate Poisson-log normal distribution". In : *Biometrika* 76.4 (1989), p. 643-653.

[2] R. M. L. ATES. "*Pagurus prideaux* and *Adamsia palliata* are not obligate symbionts". In : *Crustaceana* 68.4 (1995), p. 522-524.

[3] K. BOOS, L. GUTOW, R. MUNDRY et H. D. FRANKE. "Sediment preference and burrowing behaviour in the sympatric brittlestars *Ophiura albida* Forbes, 1839 and *Ophiura ophiura* (Linnaeus, 1758) (Ophiuroidea, Echinodermata)". In : *Journal of Experimental Marine Biology and Ecology* 393 (2010), p. 176-181.

[4] A. BRIND'AMOUR, P. LAFFARGUE, J. MORIN, S. VAZ, A. FOVEAU et H. LE BRIS. "Morphospecies and taxonomic sufficiency of benthic megafauna in scientific bottom trawl surveys". In : *Continental Shelf Research* 72 (2014), p. 1-9.

[5] J. CHIQUET. "Contributions to sparse methods for complex data analysis". Thèse de doct. Université d'Évry-val-d'Essonne, 2015.

[6] J. CHIQUET, M. MARIADASSOU et S. ROBIN. "The Poisson-Lognormal Model as a Versatile Framework for the Joint Analysis of Species Abundances." In : *Front Ecol Evol* 9 (2021), p. 188.

[7] J. CHIQUET, M. MARIADASSOU et S. ROBIN. "Variational inference for probabilistic Poisson PCA". In : *The Annals of Applied Statistics* 12.4 (2018), p. 2674-2698.

[8] J. CHIQUET, S. ROBIN et M. MARIADASSOU. "Variational inference for sparse network reconstruction from count data". In : *International Conference on Machine Learning*. PMLR. 2019, p. 1162-1171.

[9] M. E. CLAPHAM. "Ordination methods and the evaluation of Ediacaran communities". In : *Quantifying the Evolution of Early life*. Springer, 2011, p. 3-21.

[10] K. R. CLARKE et M. AINSWORTH. "A method of linking multivariate community structure to environmental variables". In : *Marine Ecology-Progress Series* 92 (1993), p. 205.

[11] K. R. CLARKE, P. J. SOMERFIELD et M. G. CHAPMAN. "On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray–Curtis coefficient for denuded assemblages". In : *Journal of experimental marine biology and ecology* 330.1 (2006), p. 55-80.

[12] F. COLLOCA, P. CARPENTIERI, E. BALESTRI et G. D. ARDIZZONE. "A critical habitat for Mediterranean fish resources: shelf-break areas with *Leptometra phalangium* (Echinodermata: Crinoidea)". In : *Marine Biology* 145.6 (2004), p. 1129-1142.

[13] A. P. DEMPSTER, N. M. LAIRD et D. B. RUBIN. "Maximum likelihood from incomplete data via the EM algorithm". In : *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), p. 1-22.

[14] S. DRAY, R. PÉLISSIER, P. COUTERON, M. J. FORTIN, P. LEGENDRE, P. R. PERES-NETO, E. BELLIER, R. BIVAND, F. G. BLANCHET, M. DE CÁCERES, A. B. DUFOUR, E. HEEGAARD, T. JOMBART, F. MUNOZ, J. OKSANEN, J. THIOULOUSE et H. H. WAGNER. "Community ecology in the age of multivariate multiscale spatial analysis". In : *Ecological Monographs* 82.3 (2012), p. 257-275.

[15] D. P. FAITH, P. R. MINCHIN et L. BELBIN. "Compositional dissimilarity as a robust measure of ecological distance". In : *Vegetatio* 69.1 (1987), p. 57-68.

[16] A. R. GATES, T. HORTON, A. SERPELL-STEVENS, C. CHANDLER, L. J. GRANGE, K. ROBERT, A. BEVAN et D. O. B. JONES. "Ecological role of an offshore industry artificial structure". In : *Frontiers in Marine Science* 6 (2019), p. 675.

[17] O. GIMENEZ et N. PEYRARD. *Statistical Approaches for Hidden Variables in Ecology*. 1. ISTE, 2022.

[18] D. W. GOODALL. "Objective methods for the classification of vegetation. III. An essay in the use of factor analysis". In : *Australian Journal of Botany* 2.3 (1954), p. 304-324.

[19] B. A. HAZLETT. "Agonistic Behavior in *Pagurus prideaux* Leach, 1815 (Decapoda, Anomura)". In : *Crustaceana* 41.3 (1981), p. 307-310.

[20] ICES. "International Bottom Trawl Survey Working Group (IBTSWG)". In : *ICES Scientific Reports* 2 (2020).

[21] O. JARI, F. GUILLAUME BLANCHET, M. FRIENDLY, R. KINDT, P. LEGENDRE, D. MCGLINN, P. R. MINCHIN, R. B. O'HARA, G. L. SIMPSON, P. SOLYMOS, M. H. M. STEVENS, E. SZOECS et H. WAGNER. *vegan: Community Ecology Package*. 2020. URL : `https://CRAN.R-project.org/package=vegan`.

[22] S. JENNINGS, S. S. MARSHALL et N. V. C. POLUNIN. "Seychelles' marine protected areas: comparative structure and status of reef fish communities". In : *Biological Conservation* 75.3 (1996), p. 201-209.

[23] S. LÊ, J. JOSSE et F. HUSSON. "FactoMineR: A Package for Multivariate Analysis". In : *Journal of Statistical Software* 25.1 (2008), p. 1-18.

[24] P. LEGENDRE et E. D. GALLAGHER. "Ecologically meaningful transformations for ordination of species data". In : *Oecologia* 129.2 (2001), p. 271-280.

[25] J. V. MATTHEW. "fossil: palaeoecological and palaeogeographical analysis tools". In : *Palaeontologia Electronica* 14.1 (2011). R package version 0.4.0, 1T.

[26] L. MÉRILLET, D. KOPP, M. ROBERT, M. MOUCHET et S. PAVOINE. "Environment outweighs the effects of fishing in regulating demersal community structure in an exploited marine ecosystem". In : *Global Change Biology* 26.4 (2020), p. 2106-2119.

[27] P. R. MINCHIN. "An evaluation of the relative robustness of techniques for ecological ordination". In : *Theory and models in vegetation science*. Springer, 1987, p. 89-107.

[28] I. M. NESTOROWICZ, F. OLIVEIRA, P. MONTEIRO, L. BENTES, N. S. HENRIQUES, R. AGUILAR, B. HORTA E COSTA et J. GONÇALVES. “Identifying Habitats of Conservation Priority in the São Vicente Submarine Canyon in Southwestern Portugal”. In : *Frontiers in Marine Science* (2021), p. 1313.

[29] G. C. POPOVIC, D. I. WARTON, F. J. THOMSON, Francis K. C. HUI et A. T. MOLES. “Untangling direct species associations from indirect mediator species effects with graphical models”. In : *Methods in Ecology and Evolution* 10.9 (2019), p. 1571-1583.

[30] R CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2022. URL : `https://www.R-project.org/`.

[31] C. RICOTTA et S. PAVOINE. “A new parametric measure of functional dissimilarity: Bridging the gap between the Bray-Curtis dissimilarity and the Euclidean distance”. In : *Ecological Modelling* 466.109880 (2022).

[32] A. ROBERT. “Effets combinés des facteurs naturels et anthropiques sur les communautés d’invertébrés benthiques des vasières à langoustines (*Nephrops Norvegicus*) du golfe de Gascogne”. Thèse de doct. Agrocampus Ouest, 2017.

[33] RSTUDIO TEAM. *RStudio: Integrated Development Environment for R*. RStudio, PBC. Boston, MA, 2020. URL : `http://www.rstudio.com/`.

[34] C. TEBBY, S. JOACHIM, P. J. VAN DEN BRINK, J. M. PORCHER et R. BEAUDOUIN. “Analysis of community-level mesocosm data based on ecologically meaningful dissimilarity measures and data transformation”. In : *Environmental Toxicology and Chemistry* 36.6 (2017), p. 1667-1679.

[35] M. E. TIPPING et C. M. BISHOP. “Probabilistic principal component analysis”. In : *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.3 (1999), p. 611-622.

[36] D. I. WARTON, S. D. FOSTER, G. DE’ATH, J. STOKLOSA et P. K. DUNSTAN. “Model-based thinking for community ecology”. In : *Plant Ecology* 216.5 (2015), p. 669-682.

[37] J. YIN et H. LI. “A sparse conditional Gaussian graphical model for analysis of genetical genomics data”. In : *The annals of applied statistics* 5.4 (2011), p. 2630.

[38] H. ZHAO et Z. H. DUAN. “Cancer genetic network inference using gaussian graphical models”. In : *Bioinformatics and biology insights* 13.1177932219839402 (2019).

Résumé — Le modèle Poisson lognormal (PLN) s'est développé récemment comme un cadre permettant l'analyse multivariée de données de comptage. Il a reçue une attention accrue en biologie moléculaire, mais son utilisation en écologie des communautés reste anecdotique. Dans cette étude, les variantes PLNPCA et PLN réseaux ont été appliquées à des données de comptages de macroinvertébrés benthiques pour cerner des apports du modèle PLN à ce type de données. La variante PLNPCA a d'abord été comparée aux approches ACP et PCoA. En s'intéressant à des espèces connues dont nous avions établi une répartition théorique, nous avons trouvé qu'à la suite d'une classification, la PCoA donnait les résultats les plus fidèles à ceux attendus, que ce soit pour un nombre de groupes imposés ou libre. Le modèle PLNPCA serait cependant le plus robuste pour des données avec une forte proportion de 0. Pour toutes les analyses, les moins bons résultats ont été obtenus pour une ACP sur des données non transformées, qui donne trop d'importance aux espèces les plus abondantes. Le modèle PLN réseaux a ensuite été appliqué pour détecter des corrélations entre espèces conditionnellement aux autres espèces et à des covariables, ce qui nous a permis de questionner la notion de communauté établie dans les analyses faites précédemment. L'apport des covariables s'est révélé déterminant pour l'interprétation. Certaines interactions biotiques classiques ont été retrouvées. Plusieurs liens n'ont pas été expliqués et pourraient constituer soit des corrélations dues à des variables non prises en compte, soit des interactions directes inconnues à ce jour.

Mots clés — analyse multivariée, invertébrés benthiques, variables latentes, données de comptage, réseau

CONTRIBUTIONS OF THE POISSON-LOGNORMAL MODEL IN MULTIVARIATE ANALYSIS FOR COMMUNITY ECOLOGY

Abstract — The Poisson lognormal model (PLN) has recently developed as a framework for multivariate analysis of count data. It has received increased attention in molecular biology, but its use in community ecology remains anecdotal. In this study, the PLNPCA and PLN network variants were applied to benthic macroinvertebrate count data to identify potential contributions of the PLN model to this type of data. First, the PLNPCA variant was compared to the PCA and PCoA approaches. By focusing on known species for which we had established a theoretical distribution, we found that following clustering, PCoA gave the most faithful results to those expected, whether for an imposed or free number of clusters. However, the PLNPCA model would be the most robust for 0-inflated data. For all analyses, the worst results were obtained for the PCA on untransformed data, which gives too much importance to the most abundant species. The PLN model was then applied to detect correlations between species conditionally to other species and covariates, which allowed us to question the notion of community established in the analyses done previously. The contribution of covariates proved to be decisive for the interpretation. Some classic interactions were found. Several links were not explained and could be either correlations due to variables not taken into account or direct interactions unknown to this day.

Key words — multivariate analysis, benthic invertebrates, latent variables, count data, network