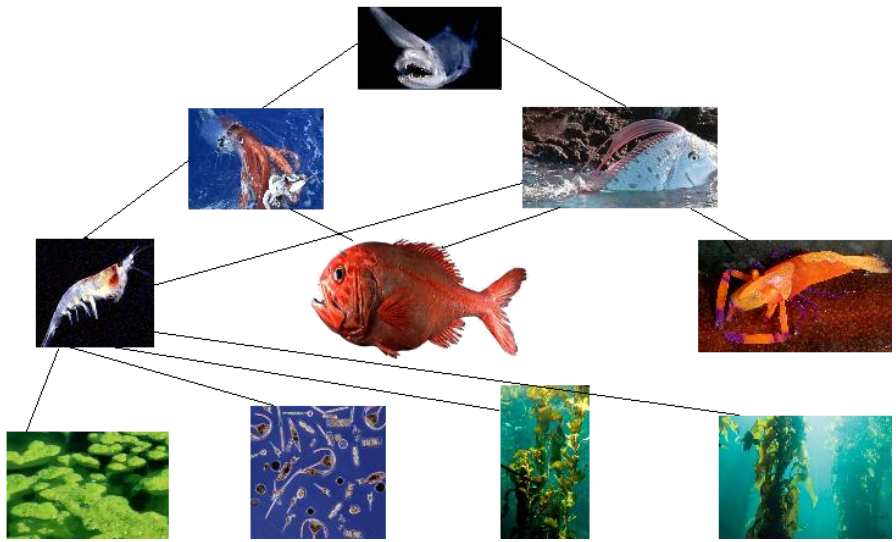
Apprentissage d'un réseau écologique marin à partir de données de comptage

Marie-Josée Cros, Le Toan Duong, Nathalie Peyrard

UR 875 MIAT, INRA Toulouse, France

Connaître les réseaux écologiques

- Protéger la biodiversité, mieux gérer les ressources, anticiper les conséquences du changement climatique... $\rightarrow$ mieux connaître les écosystèmes : les interactions entre espèces.



- Différentes observations peuvent être disponibles : présence/absence, classe d'abondance, comptage, densité de chaque espèce.

$\rightarrow$ Appréhender si les données de comptage [Kushnet et al., 2013] du projet PISCO [http://www.piscoweb.org] permettent d'apprendre des interactions entre espèces marines.

Les données

- Les 'Channel Islands'.
- Un protocole d'observation commencé en 1999, tous les ans.
- Certaines espèces observées rarement.
- Des comptages d'individus variants.
- Des observations groupées.
- Des localisations protégées différemment.

Filtrage des données et détermination des sites étudiés.

→ 12 sites avec environ 60 espèces observées 16 années.

Apprentissage de réseau

Un réseau peut être représenté sous forme de *graphe*.

Plusieurs approches probabilistes [Whittaker, 1990], [Lauritzen, 1996] existent pour apprendre une structure de graphe à partir d'observation de variables :

- **réseau bayésien**,
- **champ de Markov**,
- **modèle gaussien**...

Des hypothèses différentes (graphe orienté, observations discrètes...).
S'adaptent plus ou moins à la rareté des observations.

D'un point de vue statistique, les modèles graphiques gaussiens (GGM) sont les plus utilisés pour inférer la structure d'un modèle graphique sous-jacent non orienté.

Inférence de structure GGM

Identifier des relations d'indépendance conditionnelle entre les variables sous l'hypothèse d'une distribution gaussienne multivariée des données.

Généralement, les approches d'apprentissage dans le cadre GGM opèrent en 2 temps :

1. Transformation des comptages en observations pseudo-gaussiennes,
2. Estimation de la matrice de précision de la distribution (approche basée sur la corrélation).

Pénalisation par des méthodes de type Lasso ou Elastic Net.

Sélection de modèle (AIC, BIC, StARS).

Sélection de 3 approches s'appuyant sur GGM

- 'graphical lasso' : algorithme d'estimation de la matrice de précision en contrôlant le nombre de zéros avec la régularisation $\mathcal{L}1$ (glasso),
- inférence de la probabilité a posteriori de présence d'une arête à partir d'un mélange d'arbres (saturnin),
- inférence d'un modèle de Poisson Log-Normal qui traite directement les comptages qui sont conditionnés à un GGM caché (PLNmodels).

Approche	glasso	saturnin	PLNmodels
Données	continues		discrètes
Modèle	GGM		Poisson Log-Normal
Package R	huge	saturnin	PLNmodels
Référence	[Zhao and Liu, 2012]	[Schwaller et al., 2015]	[Chiquet et al., 2018]

Plus une méta-analyse

Pour synthétiser les 3 graphes, ajout d'une **méta-analyse** [Vignes et al., 2011] basée sur le méta-test Inverse Chi-Square.

Introduite pour combiner les p-valeurs de tests indépendants.

Paramètre de fiabilité de l'arête entre les noeuds i et j :

$$r_{ij} = 1 - exp(\sum_{m \in \mathcal{M}} log(1 - r_{ij}^m))$$

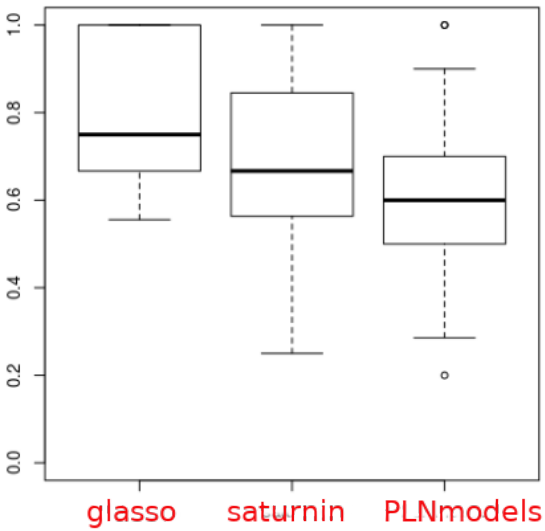
avec r_{ij}^m score de l'arête ij dans le modèle m.

Comparaison sur données simulées

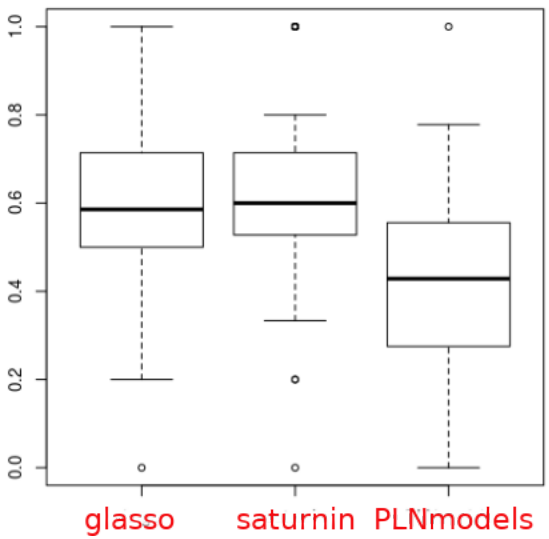
Simulations des observations selon la loi supposée des approches.

Choix du réseau ayant la densité du réseau sous-jacent.

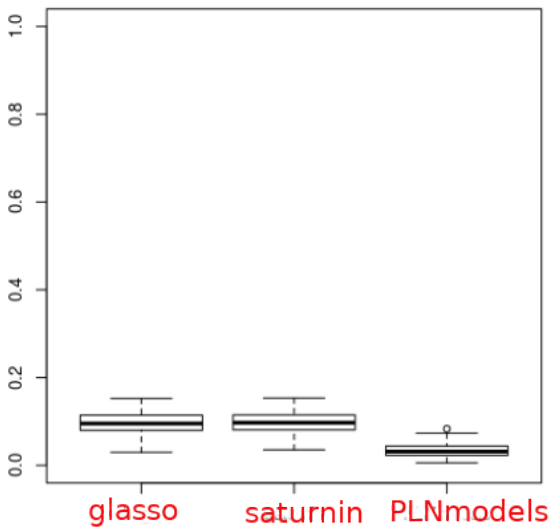
Mesure de similarité des arêtes de 2 réseaux G1=(V,E1) et G2=(V,E2) par indice de Jaccard $S_{k,l} = \frac{|E1 \cap E2|}{|E1 \cup E2|}$



5 variables, 200 échantillons



5 variables, 15 échantillons



60 variables, 15 échantillons

- Le réseau est mieux inféré si les observations sont nombreuses.
- huge et saturnin ont des résultats proches.
- PLNmodels infère moins bien le réseau.

Apprentissage sur données PISCO

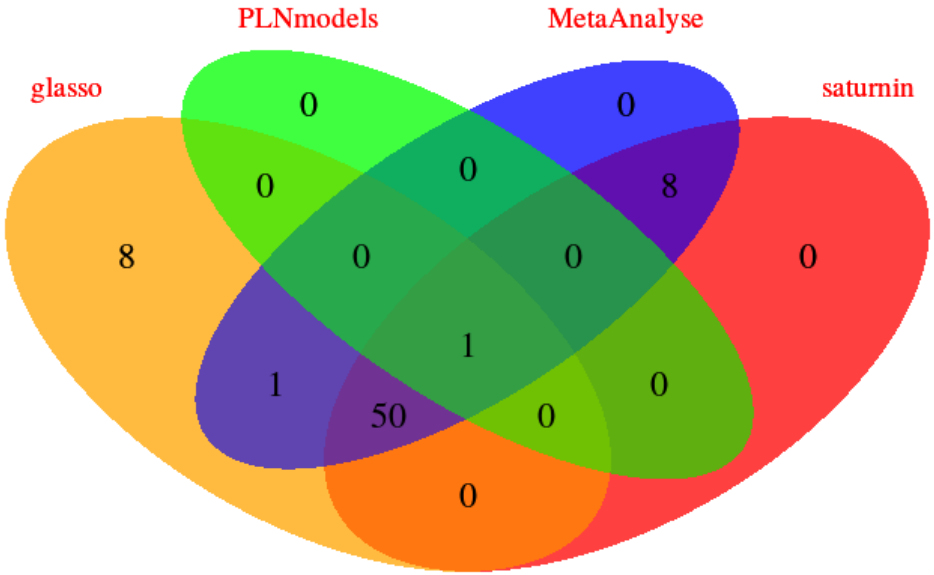
Transformation des comptages pour glasso et saturnin.

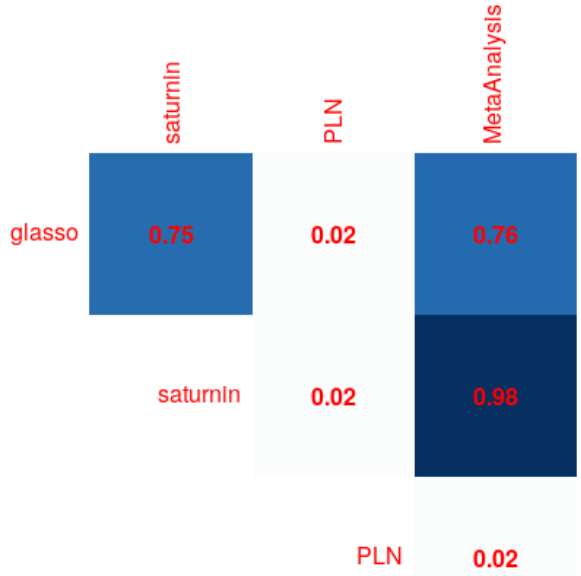
Sélection du réseau le plus stable (méthode StARS) pour glasso, PLNmodels.
Seuillage pour avoir la densité maximale de glasso et PLNmodels pour saturnin, Meta-analyse.

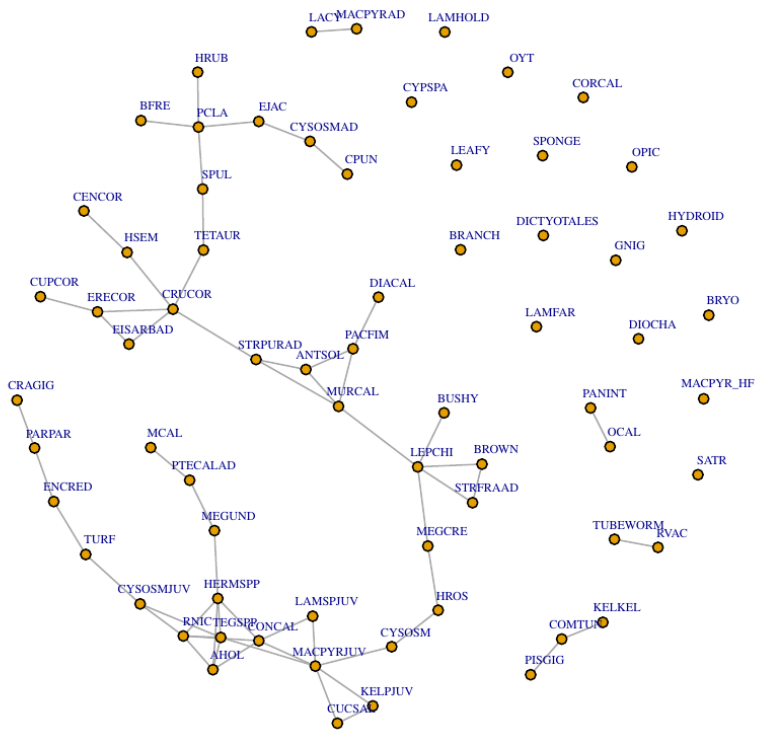
$\rightarrow$ 4 réseaux non vides appris.
Ces réseaux vont être prochainement expertisés par les écologues marins.

- PLNmodels n'apprend pratiquement pas d'arête.
- Les réseaux appris par saturnin et glasso sont assez similaires.
- La méta-analyse fait bien une agrégation des approches.

Réseau site ile Anacapa est, côté est [68 variables, 16 échantillons]

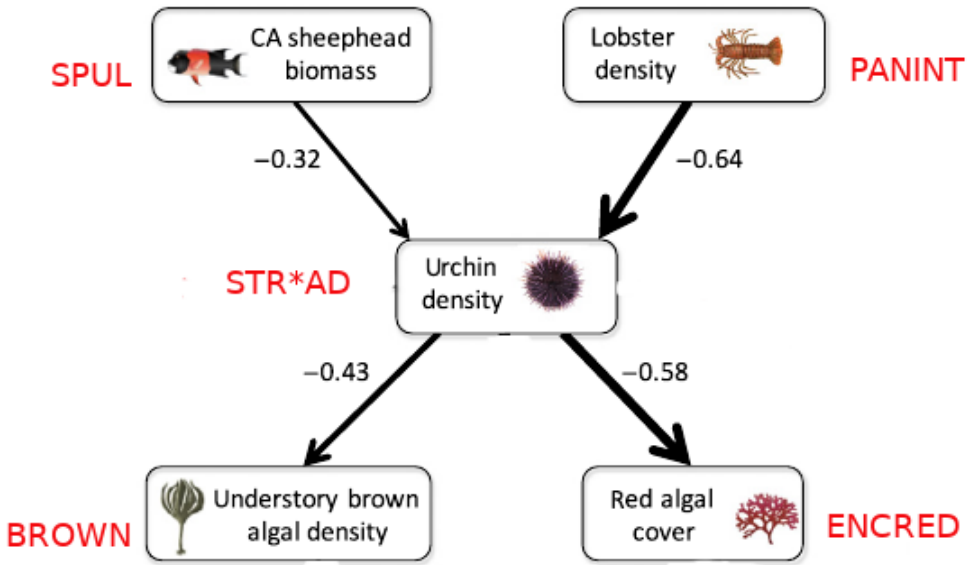


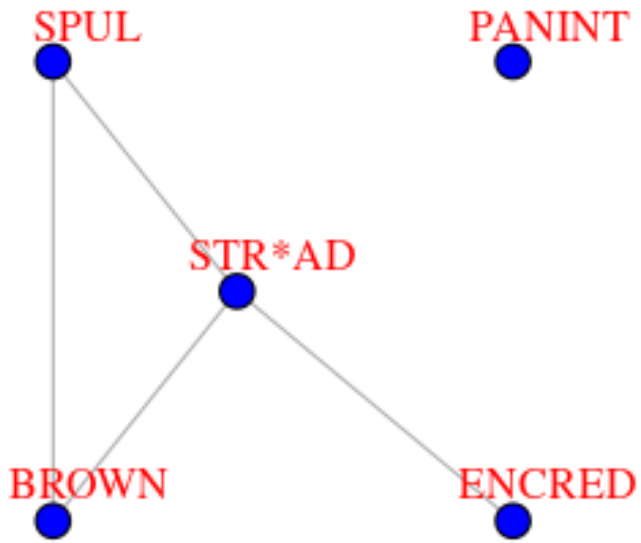




Apprentissage d'un petit sous-réseau expert

Sur le même site, réseau de 5 espèces [Caselle et al., 2018].
Densité recherchée correspondant au sous-réseau.





- glasso et PNLmodels n'apprennent plus d'arêtes.
- Lorsque l'on diminue le nombre de variables, l'apprentissage de saturnin semble assez pertinent.

Conclusion

3 approches basées sur des GGM appliquées à des données de comptages.

- ✓ glasso + StARS (package huge) apprend des réseaux peu denses.
- ✓ saturnin semble intéressant, mais demande de fixer une densité voulue.
- ✗ PLNmodels + StARS n'a pas pu s'adapter à la rareté des observations.

La méta-analyse permet de synthétiser les différentes approches.

Intérêt des interactions trouvées ?

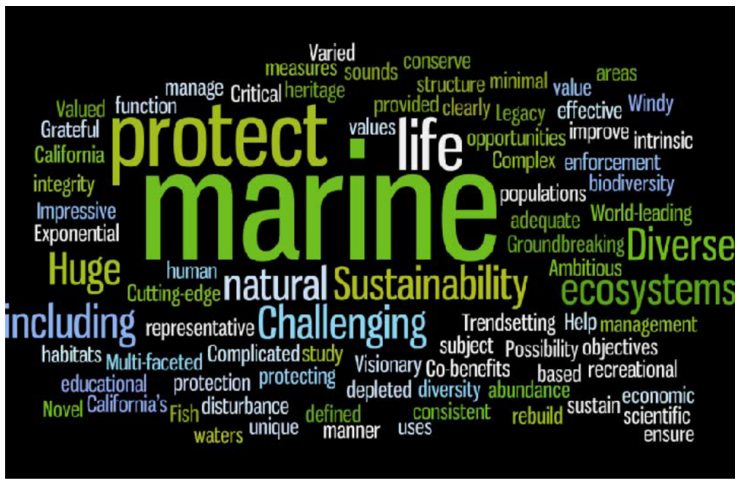
Difficultés de mise en oeuvre : filtrage des données, paramétrage des approches.

Perspectives

Incorporer de la connaissance :

- grouper des noeuds (espèces proches, stades de développement...),
- prendre en compte des interactions connues,
- transformer les comptages en biomasse...
- visualiser le groupe fonctionnel des noeuds,

Rechercher si le mode de gestion du site a une influence sur les réseaux écologiques.



References I

Caselle, J., Davis, K., and Mark, L. (2018).

Marine management affects the invasion success of a non-native species in a temperate reef in California, USA.

Ecology Letters, 21:43-53.

Chiquet, J., Mariadassou, M., and Robin, S. (2018).

Variational inference for sparse network reconstruction from count data.

ArXiv:1806,03120v1.

Gilarranz, L., Hastings, A., and Bascompte, J. (2015).

Inferring topology from dynamics in spatial networks.

Theoretical Ecology.

Hedges, L. and Olkin, I. (1985).

Statistical methods for meta-analysis.

Academic Press.

Kushnet, D., Rassweiler, A., and Lafferty, K. (2013).

A multi-decade time series of kelp forest community structure at the California Channet Islands.

Ecology, 94:2655.

Lauritzen, S. (1996).

Graphical models.

Clarendon Press.

Schwaller, L., Robin, S., and Stumpf, M. (2015).

Bayesian inference of graphical model structures using trees.

ArXiv:1504.02723.

References II

Vignes, M., Vandel, J., D, A., Ramadan-Alban, N., Cierco-Ayrolles, C., Schiex, T., Mangin, B., and de Givry, S. (2011).
Gene regulatory network reconstruction using bayesian networks, the dantzig selector, the lasso and their meta-analysis.
PLoS ONE, 6:12.

Whittaker, J. (1990).
Graphical models in applied multuvariate statistics.
Wiley.

Zhao, T. and Liu, H. (2012).
The huge package for high-dimensional undirected graph estimation in R.
Journal of Machine Learning Research.