Les modèles EDD : Une famille de modèles nuls génériques pour les réseaux écologiques

Tâm Le Minh, Sophie Donnet, François Massol, Stéphane Robin

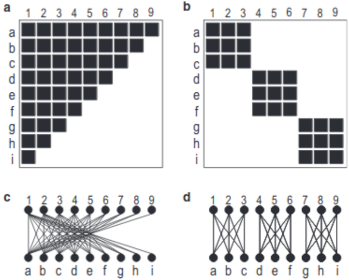
MIA-Paris, INRAE

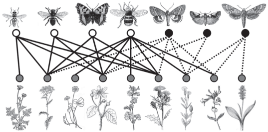
5 avril 2022

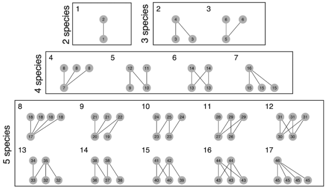
Journées du GdR Écologie Statistique

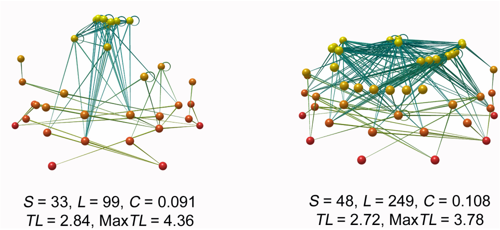
Calcul de métriques globales :

- Connectance, Nestedness, Modularité
- Fréquences de motifs
- etc.









Utilisation de modèles nuls

Test d'hypothèse avec une statistique :

- On décide d'un niveau de significativité α,
- On calcule la distribution $\mathcal{F}_0$ de la statistique associée à l'hypothèse nulle $\mathcal{H}_0$,
- On construit une zone de rejet en fonction de $\mathcal{F}_0$ et de α,
- Si la statistique observée est dans la zone de rejet, alors l'hypothèse $\mathcal{H}_0$ est rejetée.

Le modèle nul génère des réseaux aléatoires utilisés pour calculer la distribution nulle $\mathcal{F}_0$ de la statistique.
$\rightsquigarrow$ $\mathcal{H}_0$ est donc déterminée par les hypothèses du modèle nul.

Modèles nuls

Hypothèse écologique : l'hétérogénéité de spécialisation entre espèces crée le patron d'intérêt.

- Contraintes sur les degrés des lignes et des colonnes (Connor et Simberloff, 1979)

  $\rightsquigarrow$ Cependant, le support de $\mathcal{F}_0$ est restreint :
  - Pourquoi conserver exactement les degrés ?
  - Quelle distribution des réseaux générés ?

- Contraintes sur les espérances des degrés (Gilpin et Diamond, 1982, Gotelli et Graves, 1996)

  $\rightsquigarrow$ Par exemple : $p_{ij} = \lambda \times p_i^{(r)} \times p_j^{(c)}$

Les modèles EDD

Le réseau est organisé par les distributions de degrés (somme des arêtes d'un nœud) attendus.

Notations

- Matrice d'adjacence du réseau Y de taille $m \times n$
- Densité du réseau : λ
- "Degré attendu" d'une ligne : $f(U_i),\ U_i \sim \mathcal{U}[0,1],\ \int f = 1$
- "Degré attendu" d'une colonne : $g(V_j),\ V_j \sim \mathcal{U}[0,1],\ \int g = 1$

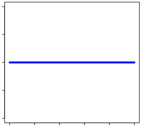
$$\begin{array}{rcl} U_i, V_j & \overset{iid}{\sim} & \mathcal{U}[0,1] \\ Y_{ij} \mid U_i, V_j & \sim & \mathcal{P}(\lambda f(U_i) g(V_j)) \end{array}$$

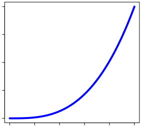
On peut remplacer la loi d'émission par la loi qu'on souhaite (par exemple, une loi de Bernoulli pour les réseaux binaires).

Les modèles EDD

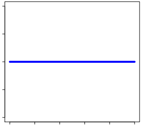
$g_0(v) =$ $g(v) =$

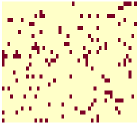
$$
\begin{array}{rcl}
U_i, V_j & \overset{iid}{\sim} & \mathcal{U}[0,1] \\
Y_{ij} \mid U_i, V_j & \sim & \mathcal{P}(\lambda f(U_i) g(V_j))
\end{array}
$$

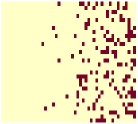




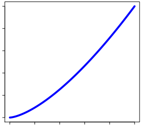
$f_0(u) =$

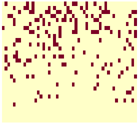


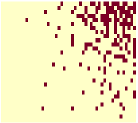




$f(u) =$







Propriété d'échangeabilité

Le modèle EDD est un modèle échangeable ligne-colonne : la loi jointe de la matrice est invariante par permutation des lignes ou des colonnes.

C'est un modèle raisonnable pour la plupart des problèmes car en général, on omet la taxomonie :

- La plupart des statistiques étudiées donnent des informations sur la topologie globale du réseau (nestedness, modularité, fréquences de motifs).
- Si on permute les lignes ou les colonnes de la matrice d'adjacence, elle représente toujours le même réseau.

U-statistiques

Hoeffding (1948) : $(X_1, ..., X_n)$ variables i.i.d.

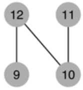
$$U = r! \binom{n}{r}^{-1} \sum_{1 \leq i_1 \neq ... \neq i_r \leq n} h(X_{i_1}, ..., X_{i_r}).$$

U-statistique sur une matrice $m \times n$

$$U = m!n! \left[\binom{m}{2} \binom{n}{2} \right]^{-1} \sum_{i_1 \neq i_2}^{m} \sum_{j_1 \neq j_2}^{n} h(Y_{\{i_1, i_2; j_1, j_2\}})$$

Exemple : fréquences de motifs

$$h(Y_{\{1,2;1,2\}}) = Y_{11} Y_{12} Y_{21} (1 - Y_{22})$$



Estimation du modèle EDD

Rappel : Modèle EDD Poisson

$$
\begin{array}{rcl}
U_i, V_j & \overset{iid}{\sim} & \mathcal{U}[0,1] \\
Y_{ij} \mid U_i, V_j & \sim & \mathcal{P}(\lambda f(U_i) g(V_j))
\end{array}
$$

où :

- $\lambda = \mathbb{E}[Y_{ij}]$
- $\int f = \int g = 1$, $\int f^k = F_k$, $\int g^k = G_k$.

Quelques propriétés :

$\rightarrow \mathbb{E}[Y_{i_1 j_1}^2 - Y_{i_1 j_1}] = \lambda^2 F_2 G_2$

$\rightarrow \mathbb{E}[Y_{i_1 j_1} Y_{i_1 j_2}] = \lambda^2 F_2$

$\rightarrow \mathbb{E}[Y_{i_1 j_1} Y_{i_2 j_1}] = \lambda^2 G_2$

U-statistiques

Estimateur $\widehat{\theta}_N := U^h_{cN,(1-c)N} \rightsquigarrow \mathbb{E}[\widehat{\theta}_N] = \mathbb{E}[h(Y_{\{i_1,i_2;j_1,j_2\}})] = \theta$

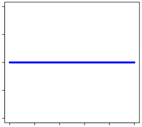
TCL pour les modèles échangeables ligne-colonne (LM, 2021)

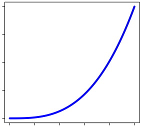
$$\sqrt{\frac{N}{V}}(\widehat{\theta}_N - \theta) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Les U-statistiques permettent de faire de l'inférence statistique avec un minimum d'hypothèses

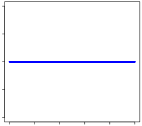
- Estimation
- Intervalles de confiance
- Tests de comparaison

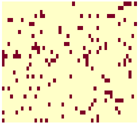
Exemple : test sur f

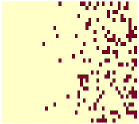
$g_0(v) =$ 

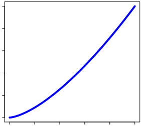
$g(v) =$ 

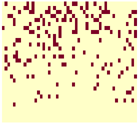
$\mathcal{H}_0 : f \equiv 1$ contre $\mathcal{H}_1 : f \not\equiv 1$
$(F_2 = 1$ contre $F_2 > 1)$

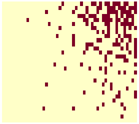
$f_0(u) =$ 





$f(u) =$ 





Bilan

Les modèles EDD sont :

- des modèles nuls probabilistes génératifs
  $\rightsquigarrow$ $\mathcal{H}_0$ correspond à une hypothèse écologique,
- des modèles échangeables ligne-colonne
  $\rightsquigarrow$ résultats de convergence des U-statistiques,
- des modèles semi-paramétriques mais les U-statistiques ne nécessitent pas de connaître les distributions de degrés
  $\rightsquigarrow$ possibilité de faire de l'inférence avec le minimum d'hypothèses.

Bibliographie

Le Minh, T. (2021). *U*-statistics on bipartite exchangeable networks. *arXiv preprint*, arXiv :2103.12597.

Merci de votre attention !