Apprentissage non supervisé de structures de graphes

Analyse de réseaux d'interactions pour comprendre l'organisation d'écosystème

Sophie Donnet[1] et François Massol[2]

[1] MIA Paris-Saclay, INRAE

[2] Center for Infection and Immunity of Lille, CNRS CIIL CENTER FOR INFECTION & IMMUNITY OF LILLE

1. A propos de la collaboration

2. De l'intérêt des réseaux en écologie

3. Unsupervised clustering of nodes by probabilistic models

   Matrix representation

   Descriptive statistics

   Probabilistic models for bipartite networks

   Variational inference

4. Application

Apprentissage et modélisation statistique - Ecologie

Sophie Donnet

- Chercheuse en modélisation et apprentissage statistique
- Spécialisée en modèles probabilistes pour les réseaux d'interactions

François Massol

- Chercheur en écologie
- Axes de recherche : écologie spatiale, écologie des communautés et écologie évolutive

Intérêt commun

Réseaux d'interactions

Projets communs

- MIRES: Méthodes Interdisciplinaires pour les Réseaux d'Échanges de Semences
  - Effet des échanges de graines sur la biodiversité cultivée
- GDR RESODIV :
  - Etude des réseaux de circulation d'objets biologiques (plantes et animaux), et des savoirs et savoir-faire qui leurs sont associés, dans les agricultures des pays des Nords et des Suds.
- ANR Econet
  - Advanced statistical modelling of ecological networks
- ANR NGB
  - Biosurveillance Next-Gen des changements dans la structure et le fonctionnement des écosystèmes

Projets communs

- Coencadrement de la thèse de Tam Le Minh (avec S. Robin) sur la comparaison de réseaux écologiques
- Dispense de formations à destination d'étudiants et chercheurs sur l'analyse de réseaux
- Copublication

A propos de la collaboration

De l'intérêt des réseaux en écologie

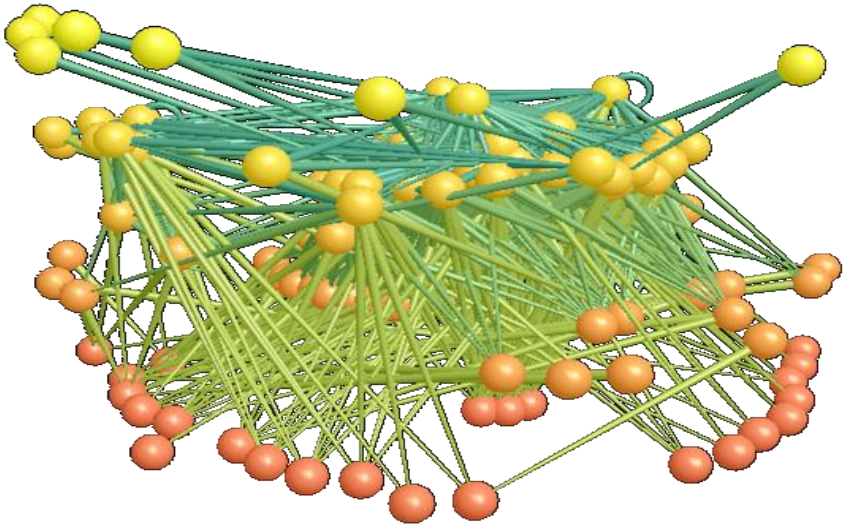
Unsupervised clustering of nodes by probabilistic models

Application

Question principale

Comment étudier les objets réseaux issus de l'écologie et de l'agronomie ?

Réseaux écologiques



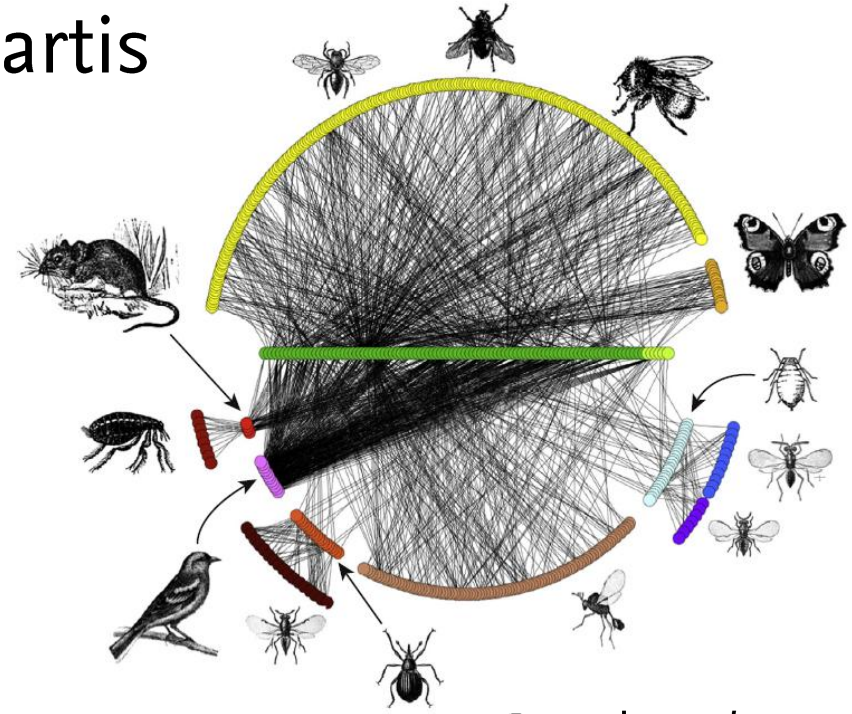
- Réseaux d'interactions, de cooccurrences, etc. entre espèces
- Réseaux de sites connectés dans l'espace

- De plus en plus de jeux de données (metabarcoding)

Réseaux écologiques

- Comprendre la dynamique des réseaux nécessite de comprendre leur structure

- Beaucoup de réseaux bipartis (2 niveaux d'espèces) ou multipartis (plus) = sans interactions intra-niveau



Pocock *et al.* 2012

Réseaux plantes-pollinisateurs

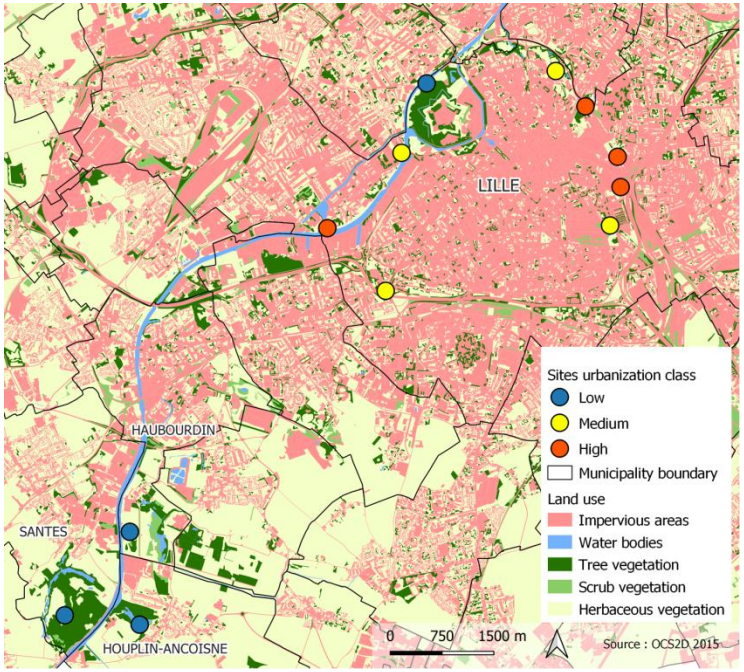
photo : N. de Manincor

- A la base du service de pollinisation (~ 87,5 % des angiospermes pollinisés par animaux, Ollerton et al. 2011)
- Dépendances réciproques entre plantes et pollinisateurs

  $\Rightarrow$ potentielles extinctions en cascade
- Diversité & complexité des réseaux liées à leur stabilité

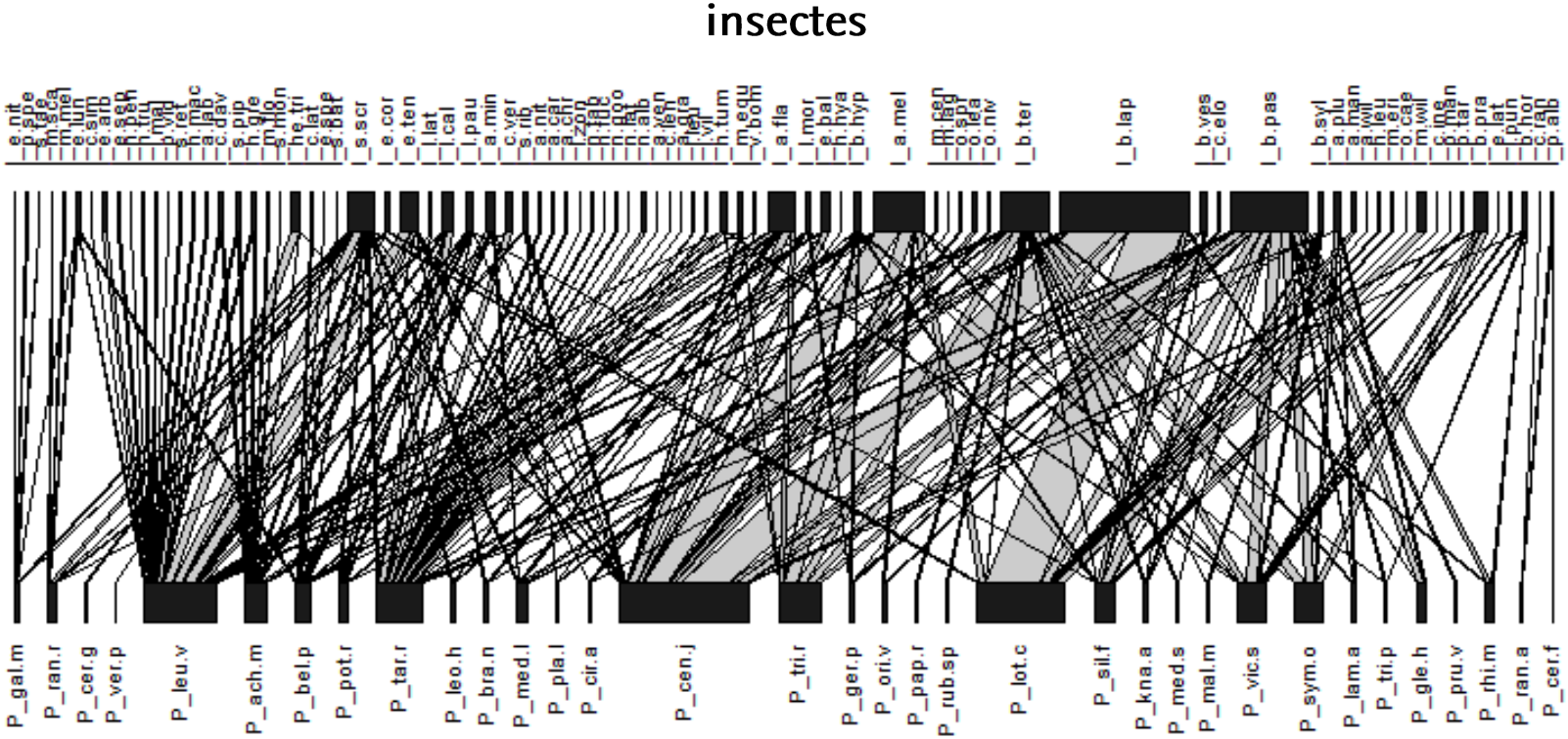
Jeu de données plantes-pollinisateurs

- Travail de post-doctorat d'Alessandro Fisogni
- Etude des réseaux plantes-pollinisateurs en ville
- 12 sites = 3 catégories d'urbanisation x 4 « réplicats » (sites semés par la MEL)



Jeu de données plantes-pollinisateurs

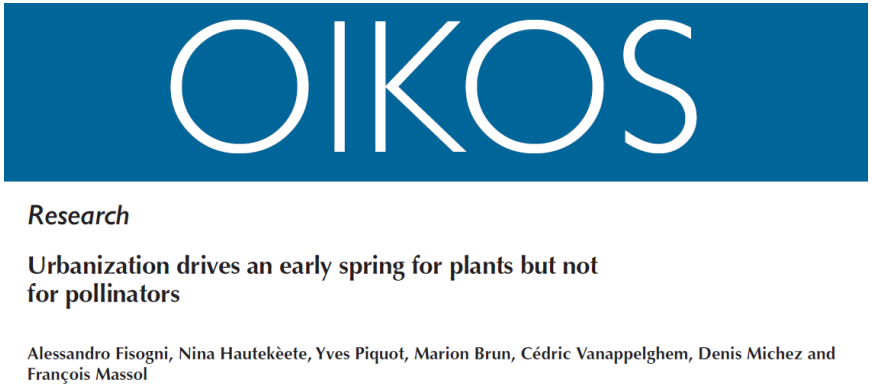
- Données SIG sur environnements
- Couverture florale



faible urbanisation, tous sites et dates confondus

Jeu de données plantes-pollinisateurs

- Premier travail publié sur la phénologie (dates de floraison et dates de sortie des insectes)



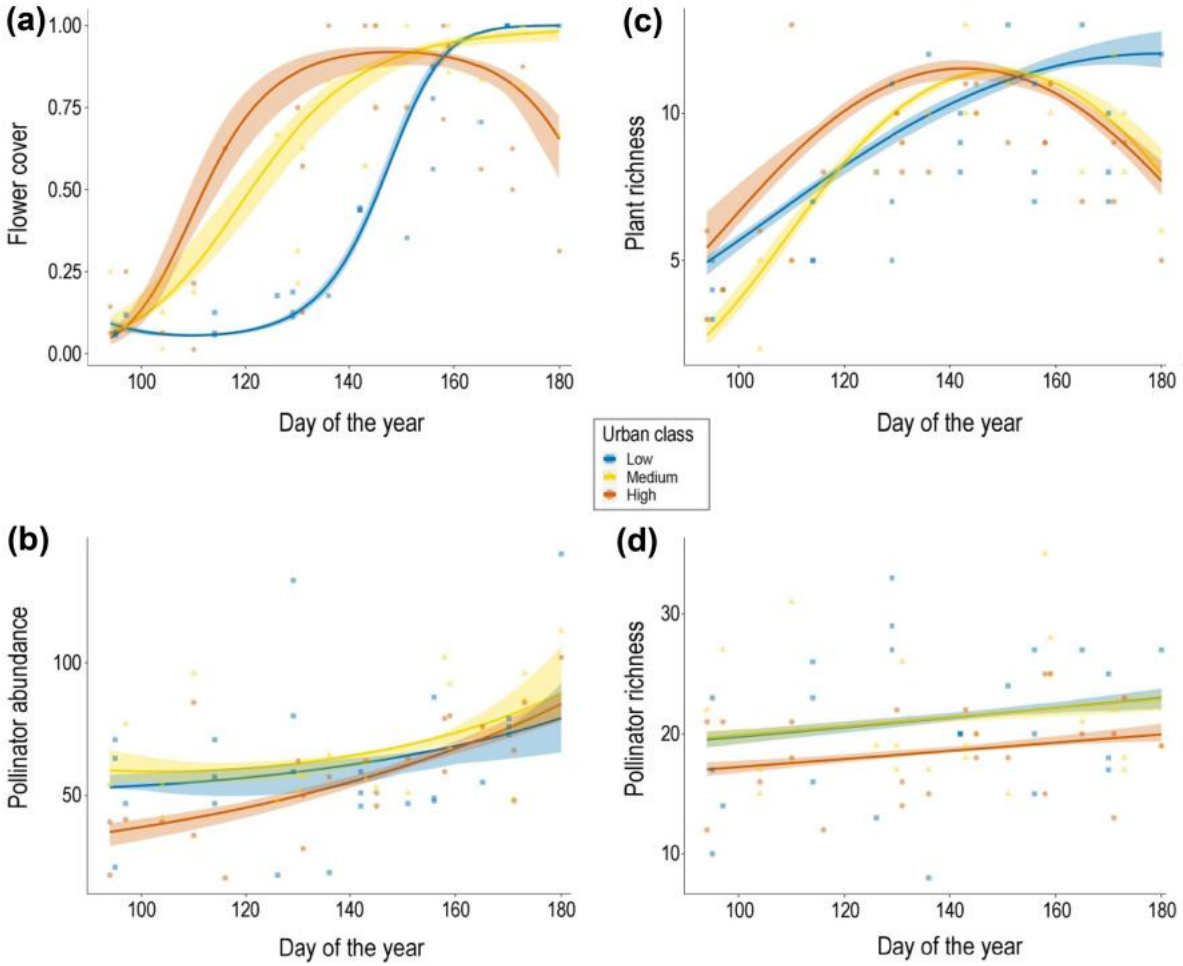
- Données réseaux = en révision…
- Données accessibles en ligne :

https://zenodo.org/record/5570297

Jeu de données plantes-pollinisateurs

Les plantes sont plus précoces en ville, mais pas les pollinisateurs

⇒ potentiel problème pour le réseau PP ?



Fisogni et al. 2020

A propos de la collaboration

De l'intérêt des réseaux en écologie

Unsupervised clustering of nodes by probabilistic models

Application

Objectif

- Highlight heterogeneity in the data (here nodes)
- Identify groups of nodes
  - Grouping nodes showing the same connection behavior
  - Each group represents a significantly different understanding from the other groups

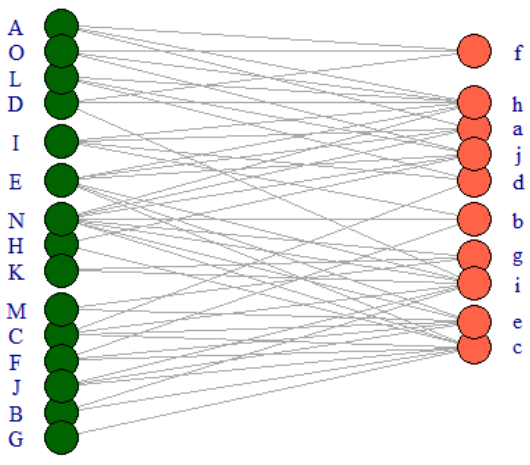
Matrix Representation

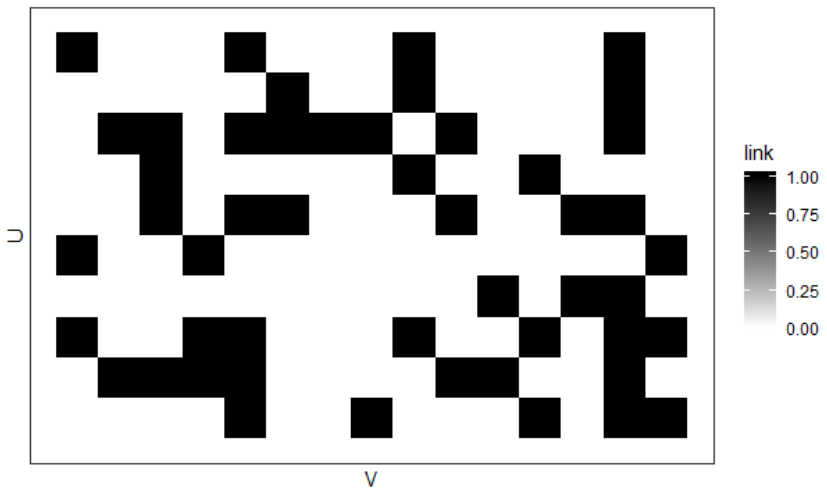
- Bipartite networks can be represented by an incidence matrix or bi-adjacency matrix
- For $i \in$  , $j \in$  ,

$$Y_{ij} = \begin{cases} 1 & \text{if there is an interaction between } i \text{ and } j \\ 0 & \text{otherwise.} \end{cases}$$

- Rectangular matrix
- In most cases : $Y_{ij} \in \{0, 1\}$. However, sometimes $Y_{ij} \in \mathbb{R}$, weighted bipartite graph
- Directed bipartite graph : not classical. Proposition $Y_{ij} \in \{-1, 0, 1\}$

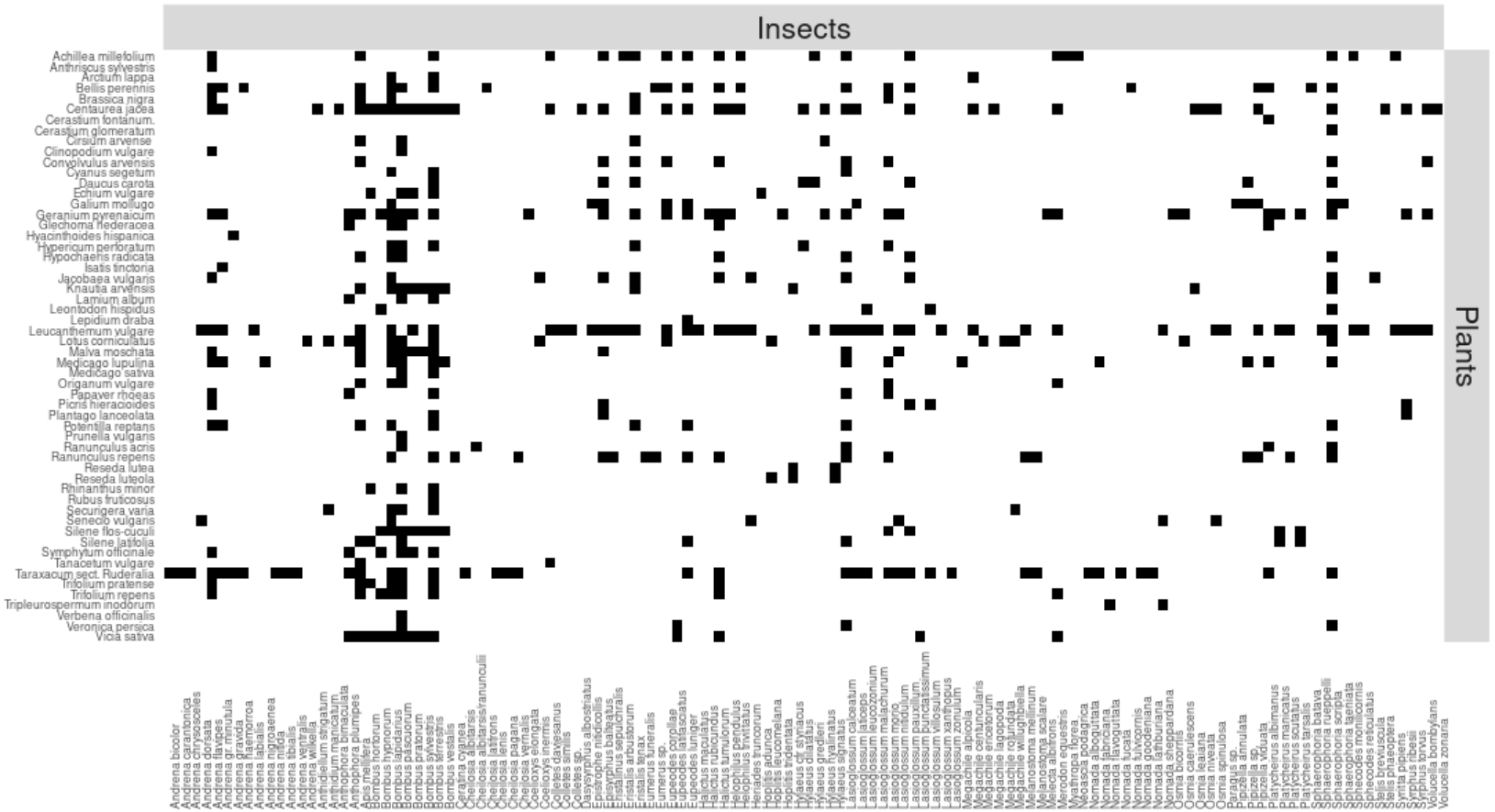
Matrix Representation





Matrix Representation

Ecological network of interest : i and j have been observed at least one time in interactions.



Metrics

Aim : give a short description of the network, give a hint about its structure, look for heterogeneity in the connections

- Many metrics supplied for simple networks
- Have been extended to bipartite networks

Libraries

R-packages

Name	Usage
Networksis	Tool to simulate bipartite networks
enaR	Provides algorithms for the analysis of ecological networks
Netpredictor	Prediction of missing links in any given bipartite network
biGRAPH	Extension of the igraph library for bipartite graphs
bipartite	Visualising Bipartite Networks and Calculating Some (Ecological) Indices

Degree

$$\begin{array}{rclcrcl} \deg(u) & = & \sum_{v \in \text{🐝}} (u \leftrightarrow v), & & \deg(v) & = & \sum_{u \in \text{🌱}} (u \leftrightarrow v) \\ \deg_i & = & \sum_{j=1}^{\text{🐝}} Y_{ij} & & \deg_j & = & \sum_{i=1}^{\text{🐝}} Y_{ij} \end{array}$$

- Nodes with high degree are hubs
- Nodes with null degree are isolated
- If edges are oriented : in- and out- degrees can be computed.

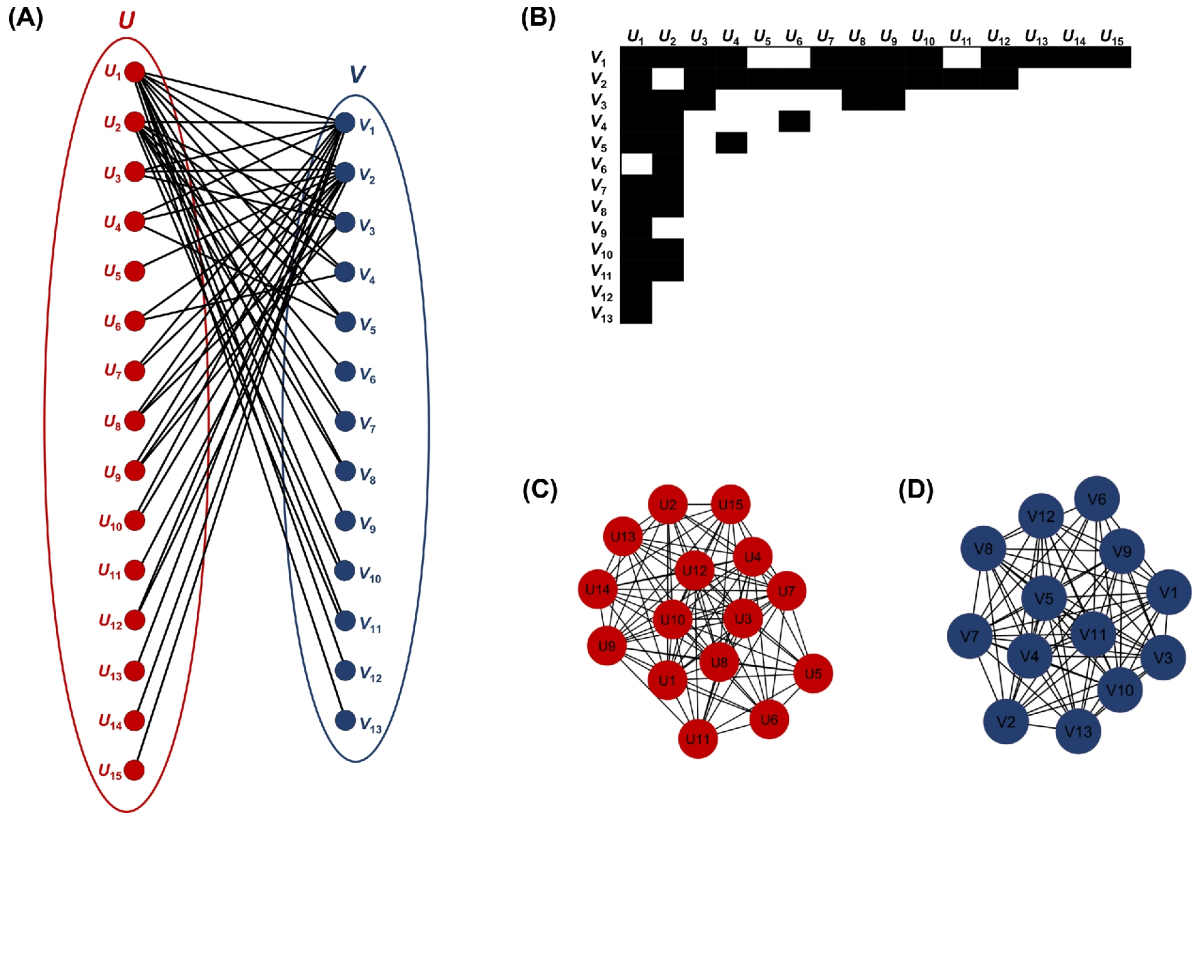
Nestedness

Property on the network

Definition

- Important property in ecology
- Defined as a pattern of interactions in which specialists (e.g. pollinators that visit few plant species) interact with plants that are visited by generalists.
- Mathematically, looking for a reordering of rows and columns such that Y is nested

Nestedness



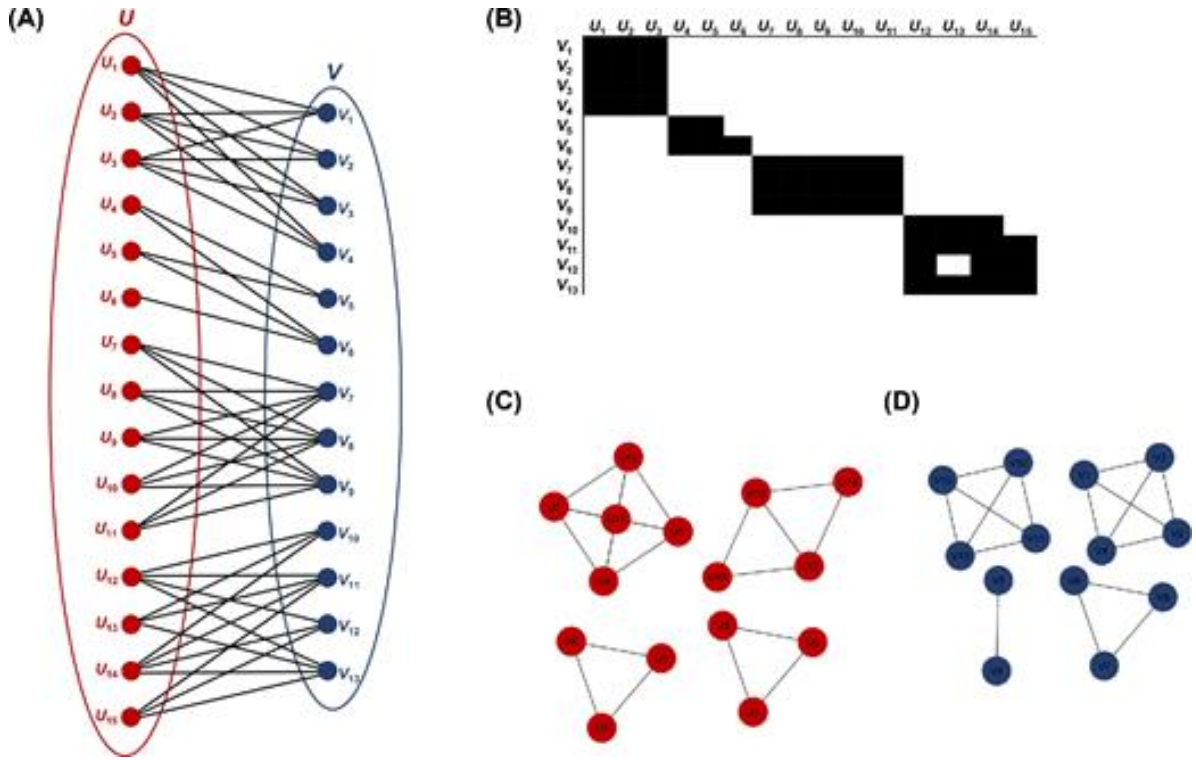
[PKP+18]

Modularity

Property on the network

Definition

Existence of clusters (blocks, module, communities) where nodes are much more connected than with other clusters



[PKP+18]

A first probabilistic model

- Context: our incidence matrix Y is the realization of a stochastic process.
- Aim: Propose a stochastic process is able to mimic heterogeneity in the connections.

Naive model

$$\forall (i,j) \in \text{  } \times \text{  }, \qquad Y_{ij} \sim \mathcal{B}\text{ern}(p)$$

- Homogeneity of the connections
- No hubs, no community, no nestedness

Latent Block Model

- Aim : introduce heterogeneity in the connections
- Tool : introduce blocks of nodes gathering entities that interact roughly similarly in the network

Latent Block Model with equations : latent variables i

- Each group of nodes (and) is divided into blocks / clusters
- K number of blocks in and L number of blocks in
- For any $i \in \{1, \ldots, n\}$, let Z_i be such that

$$Z_i = k \quad \text{if} \quad i \text{ belongs to cluster } k$$

- For any $j \in \{1, \ldots, n\}$, let W_j be such that

$$W_j = \ell \quad \text{if entity } j \text{ of group belongs to cluster } \ell$$

Latent Block Model with equations : latent variables ii

Random latent variables

$(Z_i)_{i=1\dots n}$ and $(W_j)_{j=1\dots n}$ independent random variables, such that,

$$\begin{aligned} \mathbb{P}(Z_i = k) &= \pi_k, \\ \mathbb{P}(W_j = \ell) &= \pi_\ell \end{aligned}$$

with

$$\sum_{k=1}^{K} \pi_k = 1 \quad \text{and} \quad \sum_{\ell=1}^{L} \pi_\ell = 1$$

Latent Block Model with equations : connection probability

Conditionally to the latent variables...

$$C = (Z, W) = \{Z_i, i = 1 \dots n, W_j, j = 1 \dots n\} :$$

$$\mathbb{P}(Y_{ij} = 1 | Z_i = k, W_j = \ell) = \alpha_{k\ell} .$$

Other emission distributions

- Previous model adapted to 0-1 network
- If Y_{ij} is a count

$$Y_{ij} | Z_i = k, W_j = \ell \sim \mathcal{P}(\alpha_{k\ell})$$

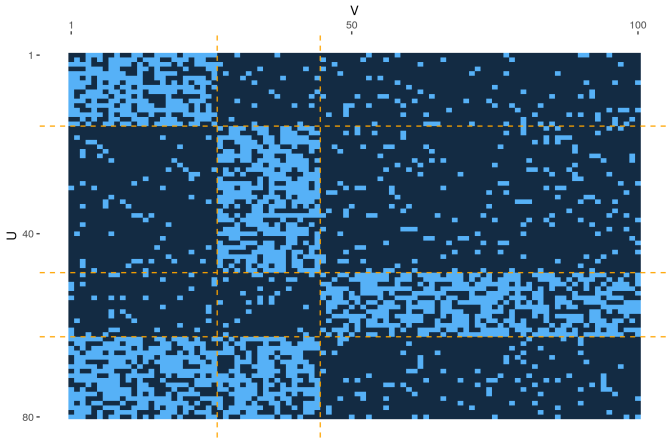
- If $Y_{ij} \in \mathbb{R}$

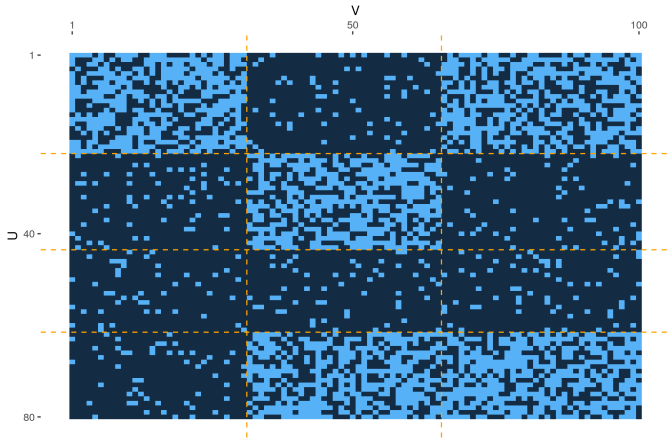
$$Y_{ij} | Z_i = k, W_j = \ell \sim \mathcal{N}(\alpha_{k\ell}, \sigma_{k\ell})$$

[GN08]

A very flexible model i

LBM able to generate communities...

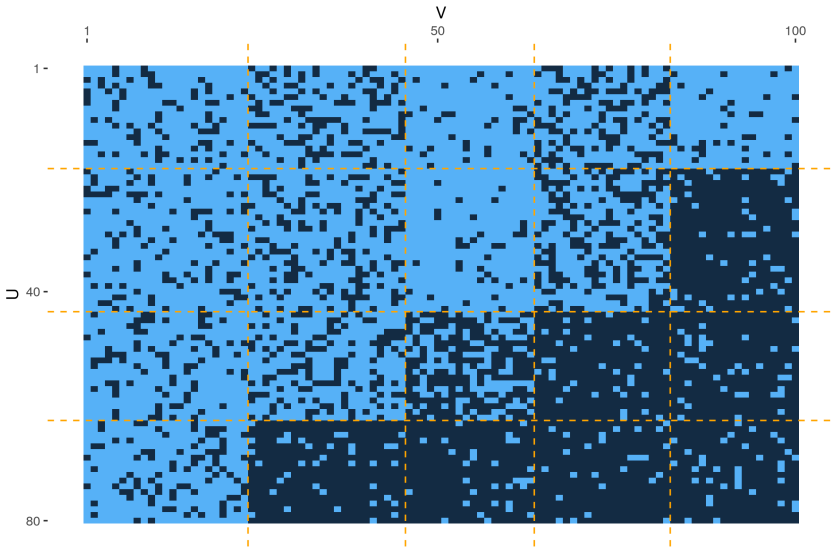




$$\alpha = \begin{pmatrix} 0.60 & 0.09 & 0.09 \\ 0.09 & 0.60 & 0.09 \\ 0.09 & 0.09 & 0.60 \\ 0.60 & 0.60 & 0.09 \end{pmatrix} \qquad \alpha = \begin{pmatrix} 0.60 & 0.09 & 0.60 \\ 0.09 & 0.60 & 0.09 \\ 0.09 & 0.09 & 0.09 \\ 0.09 & 0.60 & 0.60 \end{pmatrix}$$

A very flexible model ii

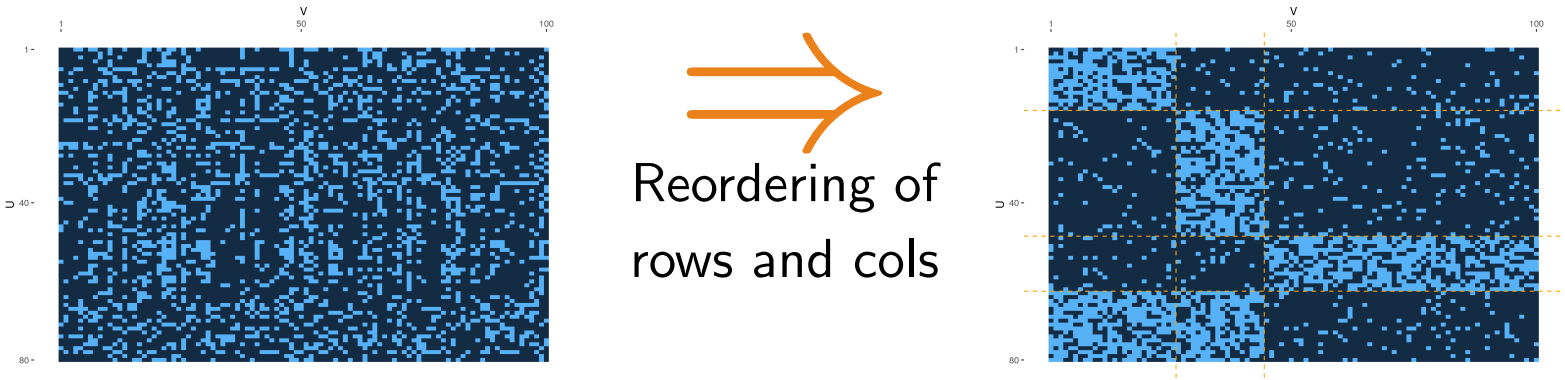
... or nested networks



$$\alpha = \begin{pmatrix} 0.80 & 0.70 & 0.90 & 0.60 & 0.90 \\ 0.80 & 0.70 & 0.90 & 0.60 & 0.09 \\ 0.80 & 0.70 & 0.40 & 0.09 & 0.09 \\ 0.80 & 0.09 & 0.09 & 0.09 & 0.09 \end{pmatrix}$$

Inference for LBM

Aim : From an incidence matrix, discovering the clusters



Remarks

- Looking for the blocks such that, under the assumption that my data come from the LBM model, the observed data Y is most probable (= most likely to occur)
- No specific prior structure
- Entities gathered because they have similar behavior in the network

Maximum likelihood inference

$$\begin{aligned}\widehat{\theta} &= \operatorname*{argmax}_{\theta} p(Y;\theta) = \operatorname*{argmax}_{\theta} \sum_{C} p_\theta(Y, C) \\ &= \operatorname*{argmax}_{\theta} \sum_{Z^{\text{plant}}, W^{\text{bee}}} p_\theta(Y, Z^{\text{plant}}, W^{\text{bee}})\end{aligned}$$

- Complete likelihood: $p_\theta(Y, Z^{\text{plant}}, W^{\text{bee}})$ easy to compute
- Likelihood $p(Y;\theta)$: integration over all the possible row and column clusterings $(Z^{\text{plant}}, W^{\text{bee}})$ (sum of $K^{n}_{\text{plant}} \times L^{n}_{\text{bee}}$ terms)
- Latent variables $\Rightarrow$ Expectation-Maximization
  - Requires to evaluate $p(Z^{\text{plant}}, W^{\text{bee}} \mid Y;\theta)$
  - No independence in this distribution
  - Complicated distribution

Variational Inference

- Use a variational version of the Expectation-Maximization algorithm [DPR08, BKM17, MRV10]
- Penalized criterion to select the numbers of blocks K and L.
- R-package sbm + blockmodels

⏩ Skip details

Variational inference

Principle [WJ08, BKM17].

- Choose a divergence measure $D(q \,||\, p)$
- Choose a class of distributions $\mathcal{Q}$
- Maximize w.r.t. θ and $q \in \mathcal{Q}$ the lower bound

$$J(Y; \theta, q) = \log p_\theta(Y) - D(q(C) \,||\, p_\theta(C \mid Y)) \leq \log p_\theta(Y)$$

Popular choice for SBMs. [GN08, DPR08, Leg16, MM15]

- $D = KL$:

$$\begin{aligned} J(Y; \theta, q) &= \log p_\theta(Y) - KL(q(C) \,||\, p_\theta(C \mid Y)) \\ &= \mathbb{E}_q \left(\log p_\theta(Y, C) \right) + \mathcal{H}\left(q(C) \right) \end{aligned}$$

- q factorizable: $\mathcal{Q} = \left\{ q(C) : q(C) = \prod_i q_i(Z_i^{\text{🌱}}) \prod_j q_j(W_i^{\text{🐝}}) \right\}$

$$\tau_{ik} = \mathbb{P}_q(Z_i = k)$$

$\rightarrow$ mean field approximation

Variational EM

Algorithm

At iteration (t), given $(\theta^{(t-1)}, q_{\tau^{(t-1)}})$,

- **Step 1** Maximization w.r.t. τ

$$\begin{aligned}
\tau^{(t)} &= \arg\max_{\tau \in \mathcal{T}} J(Y; \theta^{(t-1)}, q_\tau) \\
&= \arg\max_{\tau \in \mathcal{T}} \mathbb{E}_{q_\tau} \left[\log p_{\theta^{(t-1)}}(Y, C)\right] + \mathcal{H}\left(q_\tau(C)\right) \\
&= \arg\min_{\tau \in \mathcal{T}} \mathrm{KL}[q_\tau, p(\cdot | Y; \theta^{(t-1)})]
\end{aligned}$$

- **Step 2** Maximization w.r.t. θ

$$\begin{aligned}
\theta^{(t)} &= \arg\max_{\theta} J(Y; \theta, q_{\tau^{(t)}}) \\
&= \arg\max_{\theta} \mathbb{E}_{q_{\tau^{(t)}}} \left[\log p_\theta(Y, C)\right]
\end{aligned}$$

Outputs

For a given $(K_{\text{🌱}}, L_{\text{🐝}})$

- Parameter estimates $\widehat{\theta}$: $\widehat{\alpha}$ and $\widehat{\pi}^{\text{🌱}}, \widehat{\pi}^{\text{🐝}}$
- The best clustering of nodes

$$\widehat{C} = \underset{C}{\operatorname{argmax}}\, p(C|Y; \widehat{\theta}, \mathcal{M}) \approx \underset{C}{\operatorname{argmax}}\, q_{\widehat{\tau}}(C|Y; \widehat{\theta}, \mathcal{M})$$

Choice of the numbers of blocks

ICL : penalized criterion that relies on the complete likelihood

For any model $\mathcal{M}$ defined by $(K_{🌱}, L_{🐝})$

$$ICL(\mathcal{M}) = \log p_\theta(Y, \widehat{C}; \mathcal{M}) - \frac{1}{2} pen_{\mathcal{M}}$$

with

$$pen_{\mathcal{M}} = \underbrace{(K_{🌱} - 1) \log n_{🌱} + (L_{🐝} - 1) \log n_{🐝}}_{\text{Clustering}} + \underbrace{K_{🌱} L_{🐝} \log (n_{🌱} n_{🐝})}_{\text{Connection}}$$

A propos de la collaboration

De l'intérêt des réseaux en écologie

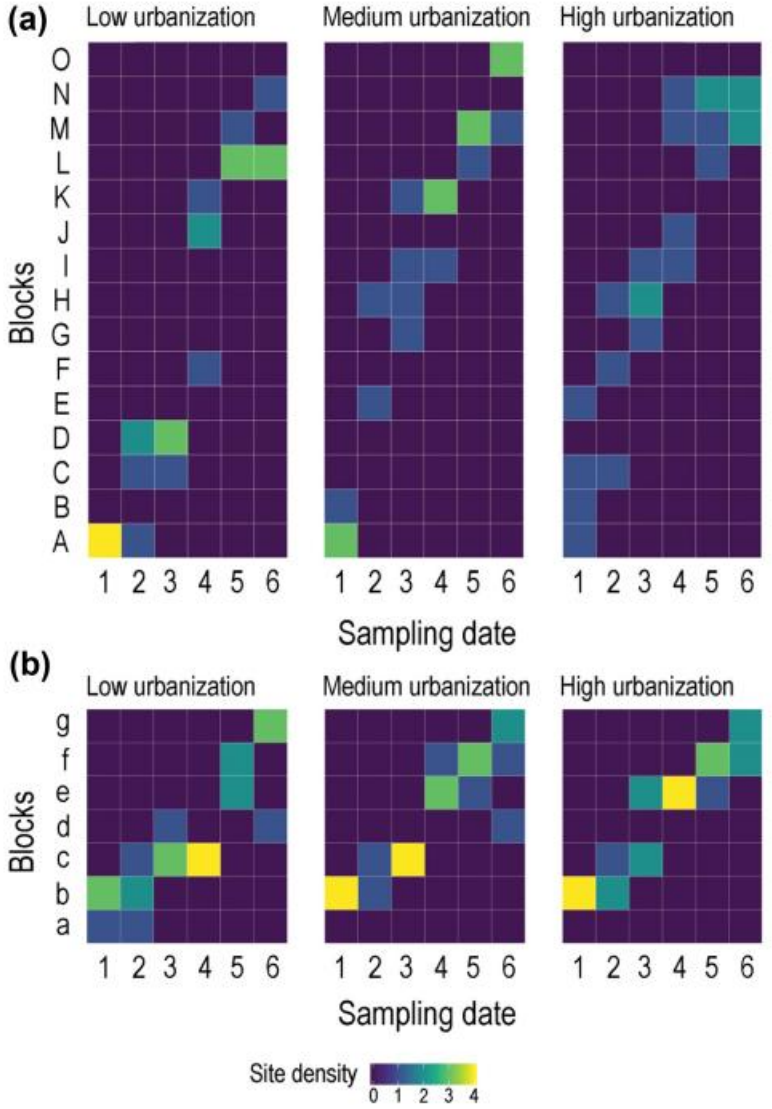
Unsupervised clustering of nodes by probabilistic models

Application

Applications

Mise en évidence du décalage phénologique en utilisant un clustering des (sites x dates) x espèces

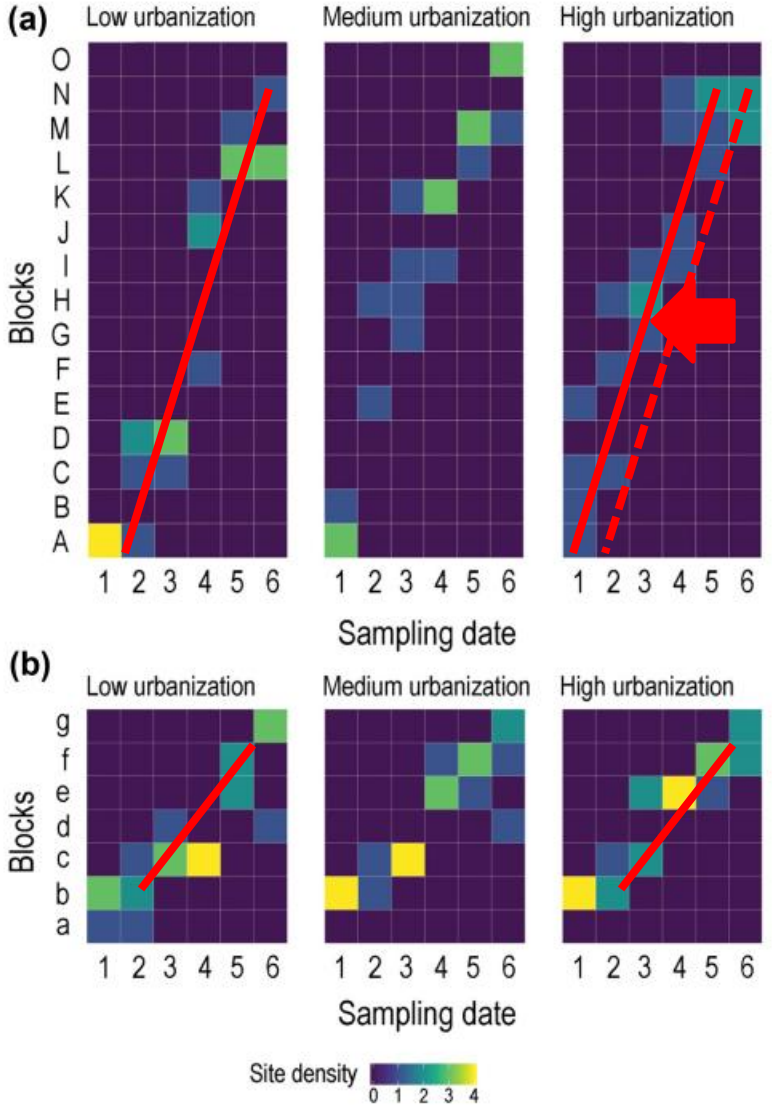
Représentation des densités de site par date et par bloc



Applications

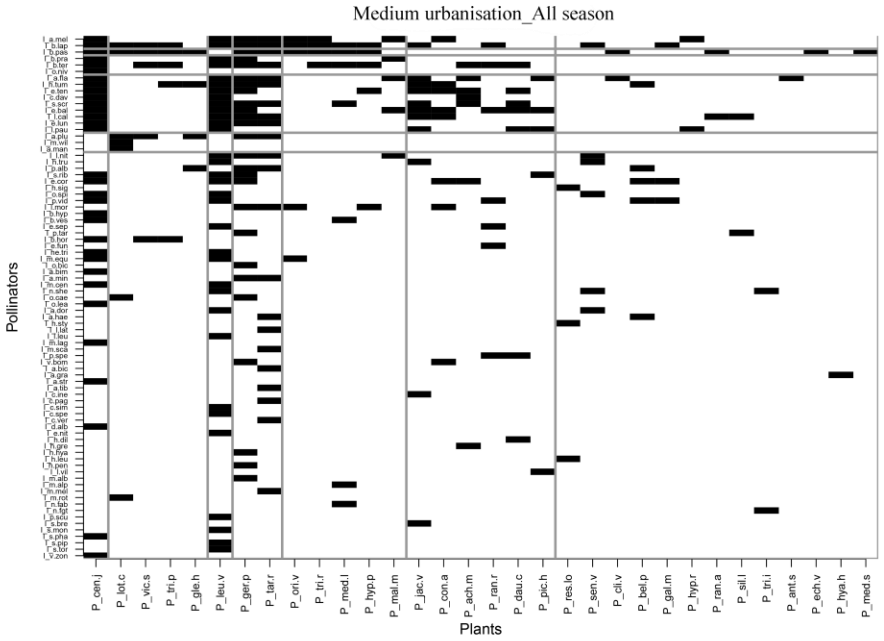
Mise en évidence du décalage phénologique en utilisant un clustering des (sites x dates) x espèces

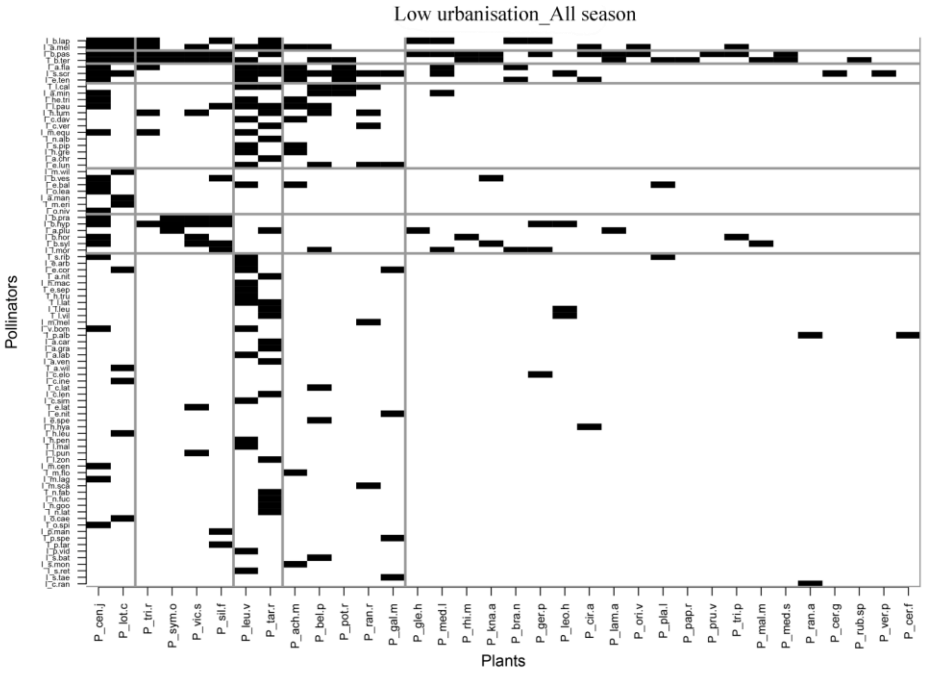
Représentation des densités de site par date et par bloc

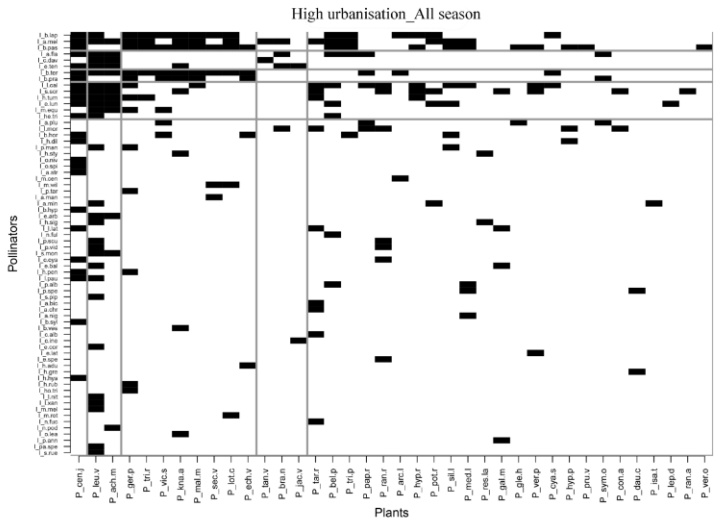


Applications

Réseaux « totaux »

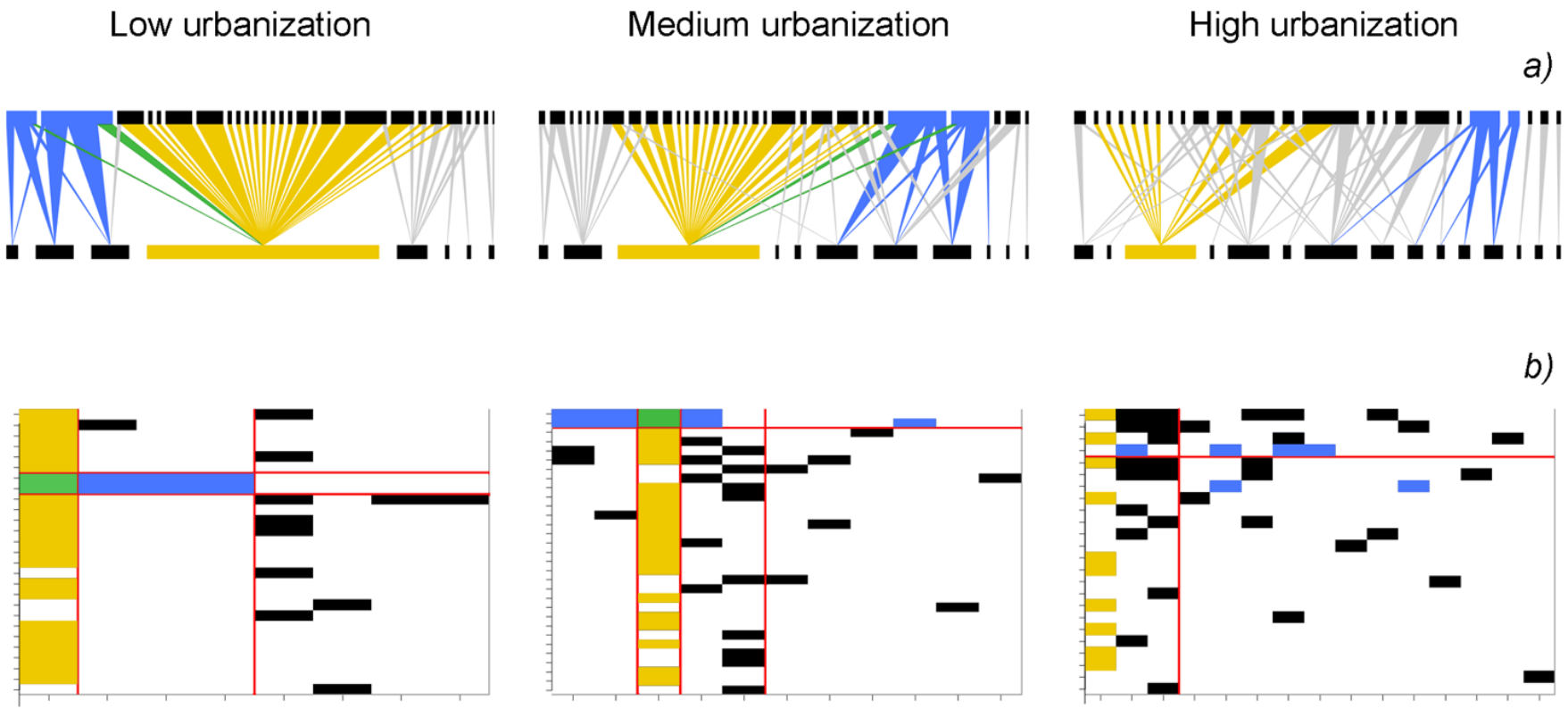






Applications

Mois d'avril seulement (plus clair)

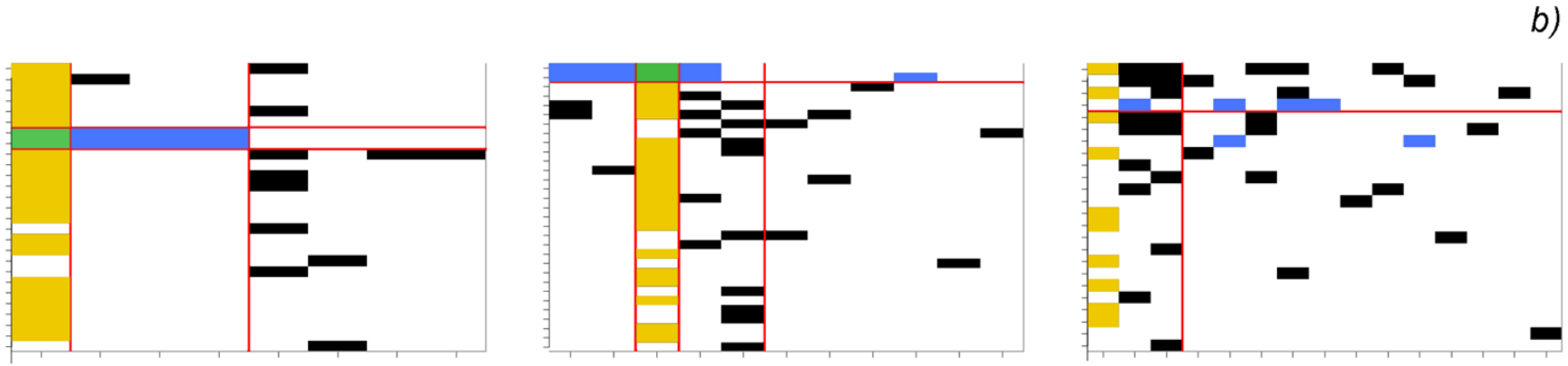


jaune = *Taraxacum* section *ruderalia*
bleu = *Anthophora plumipes* & *Bombus pascuorum*

Applications

Mois d'avril seulement (plus clair)

Permet de mettre en évidence une réorganisation des blocs de plantes et d'insectes entre types de site



jaune = *Taraxacum* section *ruderalia*
bleu = *Anthophora plumipes* & *Bombus pascuorum*

Conclusions

- Sur la saison entière, les espèces super-généralistes forment un bloc par espèce
- La phénologie semble un bon indicateur des blocs
- Les blocs sont liés aux degrés
- Réorganisation des blocs entre types de site

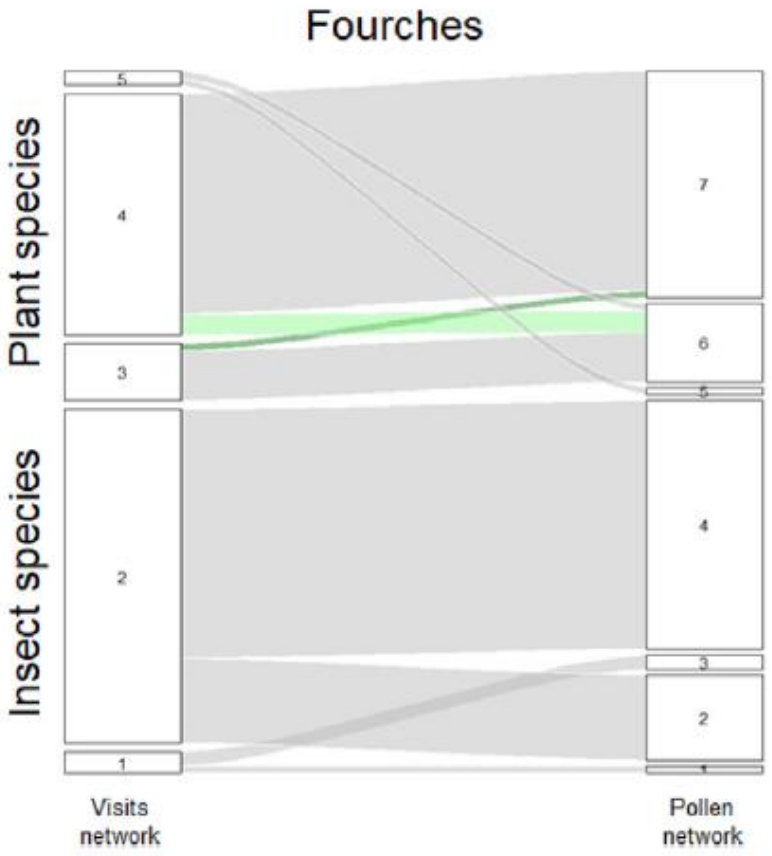
Perspectives

1. Variation et comparaisons des réseaux

Utilisation des LBMs pour comparer la congruence générale des réseaux

Comparaison des blocs obtenus par LBM pour le même « réseau », observé de deux manières différentes (observation des insectes sur fleurs vs. pollens sur insectes)

De Manincor et al. 2020



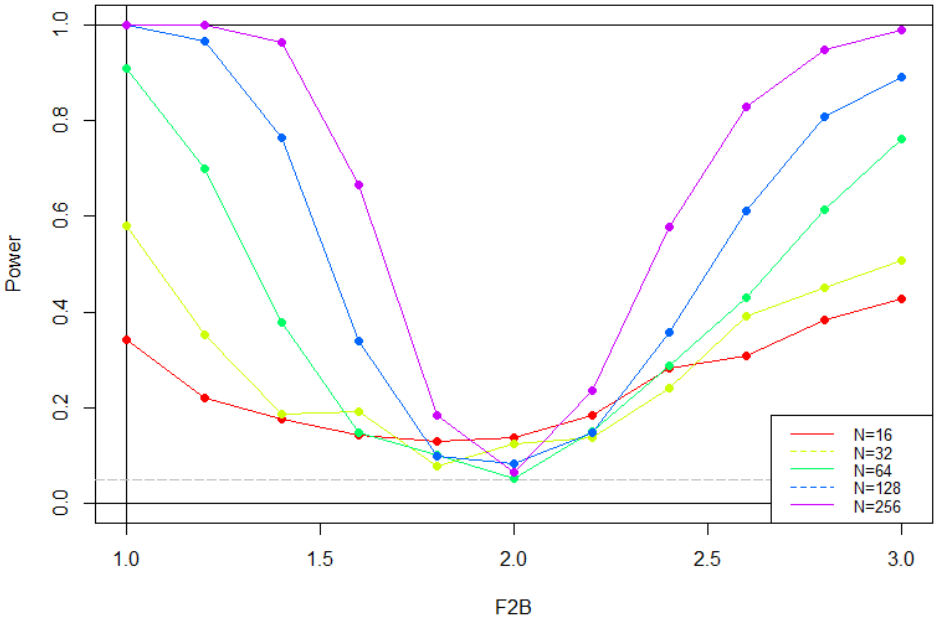
Perspectives

1. Variation et comparaisons des réseaux

Utilisation des U-statistiques pour mettre en évidence les différences de propriétés des réseaux

Puissance d'un test de comparaison (F2) entre réseaux fondé sur le modèle WBEDD (thèse de Tâm Le Minh)

Dubart et al. 2022



Perspectives

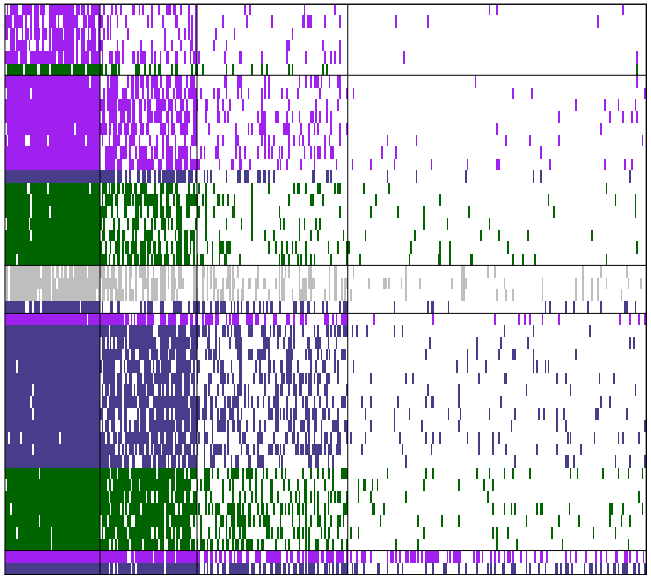
2. LBMs sur réseaux plus grands

Problème potentiel = taille des réseaux, en particulier gros réseaux (ex. microbiote)

- solution 1 : travailler à l'échelle de la famille

LBM sur réseaux de microbiote (espèces de moules x OTU obtenues sur séquençage 16S)
couleurs = espèces de *Mytilus*
colonnes = familles de bactéries

Ben Cheikh & Massol, in prep.



Perspectives

3. Metabarcoding et reconstruction de réseaux

Echantillonnage coûteux en temps, dépendant d'une expertise taxonomique rare

$\Rightarrow$ passage au metabarcoding ?

Pour les espèces microbiennes, pas moyen de faire autrement...

Perspectives

3. Metabarcoding et reconstruction de réseaux

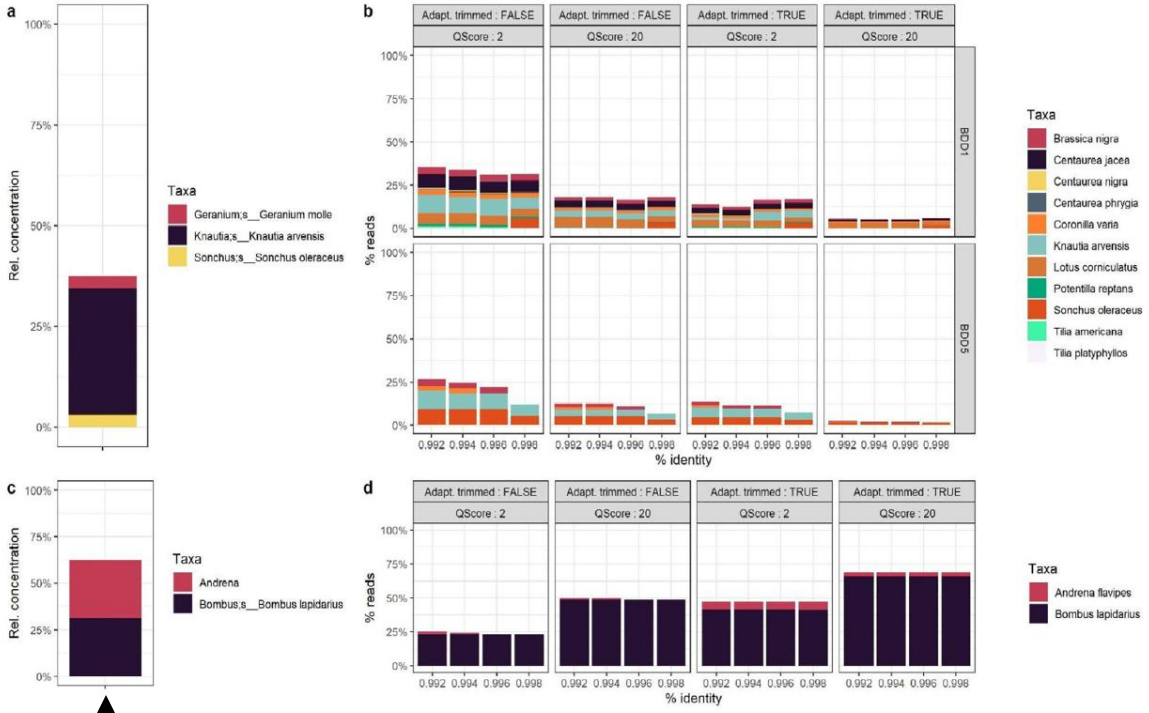
Difficultés multiples :

- abondances des espèces à partir d'ADN ?
- taux d'erreur d'assignation ?
- taux d'espèces non perçues ?

- quels pipelines bioinformatiques ?
- quelles séquences utiliser ?
- quel coût ?

Perspectives

3. Metabarcoding et reconstruction de réseaux



Tests sur communautés « mocks » pour retrouver plantes et pollinisateurs selon différentes méthodes bioinfo

Dubart et al. 2022

ce qu'on retrouve après bioinfo…

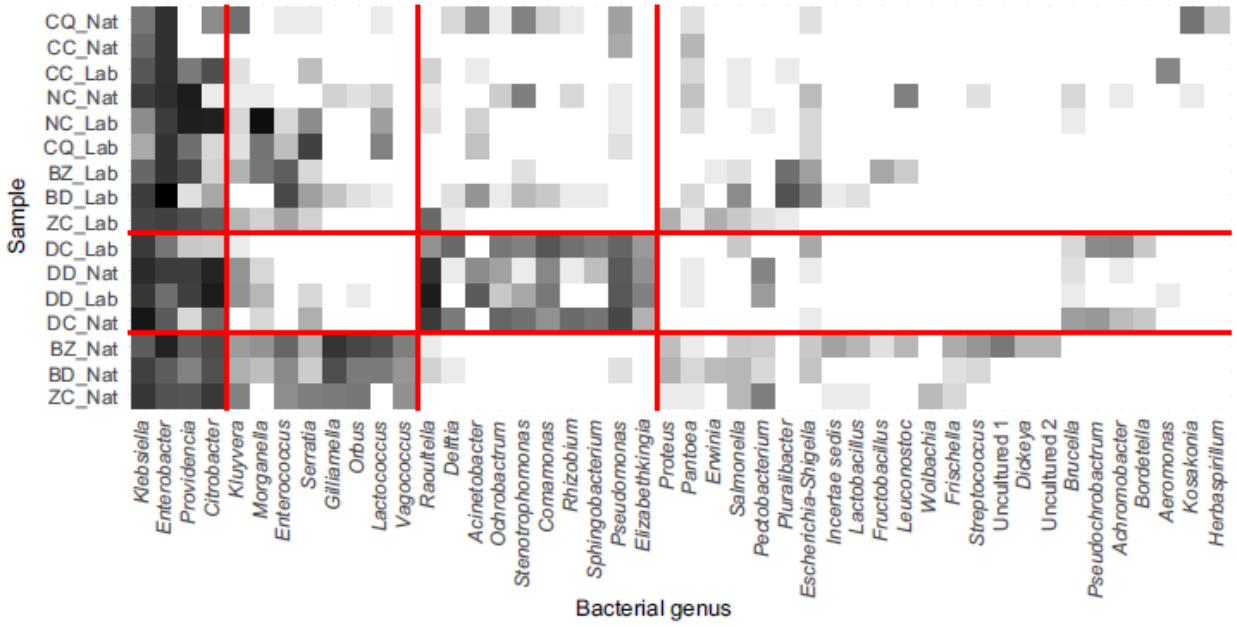
ce qu'on met dans le tube

Perspectives

3. Metabarcoding et reconstruction de réseaux

LBM entre espèces de mouches, élevées au labo ou échantillonnées sur le terrain et leur microbiote

Ravigné et al. 2022



References


D. M. Blei, A. Kucukelbir, and J. D. McAuliffe.
Variational inference: A review for statisticians.
Journal of the American Statistical Association, 112(518):859–877, 2017.


J.-J. Daudin, F. Picard, and S. Robin.
A mixture model for random graphs.
Stat. Comput., 18(2):173–83, 2008.


Gérard Govaert and Mohamed Nadif.
Block clustering with bernoulli mixture models: Comparison of different approaches.
Comput. Stat. Data Anal., 52(6):3233–3245, February 2008.


Jean-Benoist Leger.
Blockmodels: A R-package for estimating in latent block model and stochastic block model, with various probability functions, with or without covariates.
Technical report, arXiv:1602.07587, 2016.

Remerciements

A. Fisogni, N. de Manincor, M. Dubart, T. Le Minh, S.-C. Chabert-Liddell

ARSENIC, NGB & EcoNet ANR consortiums

Groupe MIRES, GDR Resodiv

cnrs dépasser les frontières

anr agence nationale de la recherche

INRAE la science pour la vie, l'humain, la terre