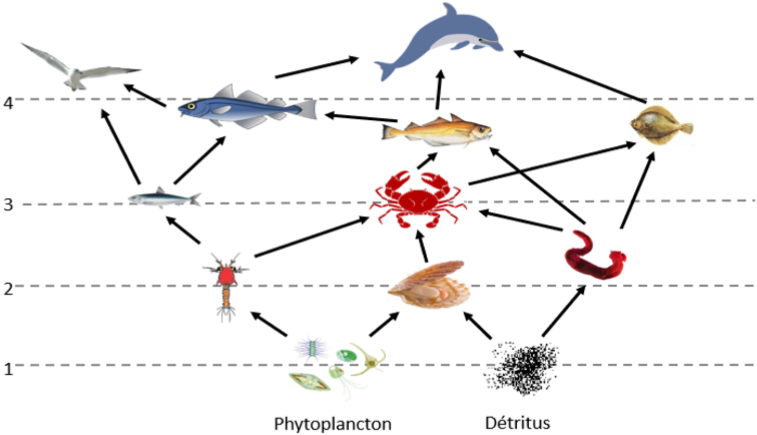
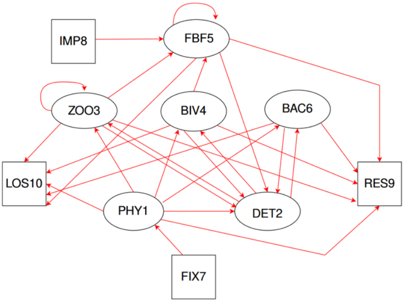
Étudier l'impact de la dégradation des données sur des indices issus de la théorie de l'information : le cas de l'approche LIM pour l'analyse des réseaux trophiques

Joint work with:
Jean-Guy Caputo, Arnaud Knippel, Hieu Nguyen, INSA ROUEN
Mathieu Dien, GREYC, UNICAEN
Valérie Girardin, LMNO, UNICAEN
Théo Grente, LMNO, UNICAEN & France Énergies Marines
Nathalie Niquil, Quentin Noguès, BOREA UNICAEN
Philippe Regnault, LMR, URCA
Jacques Bréhélin, LMNO, UNICAEN

Trophic networks: from species...



...to oriented graphs



Biochimical (metabolic process of a living body).
Urban (resources/materials ⟶ serviceable products and wastes).

Linear Inverse Modeling

From a mathematical point of view, a metabolic network is a valued oriented graph (V, E, F), where:
V is the set of all vertices of the network,
E is the set of all oriented edges ij from vertex i to vertex j,
$F = (F_{ij})_{ij \in E}$ is a positive vector, called flow.

- F satisfies the Mass Balance Equations:

$$\sum_{j \in V : ij \in E} F_{ij} - \sum_{j \in V : ji \in E} F_{ji} = 0, \quad i \in V,$$

  Additional knowledge on the network, comes from observation, literature, field measurements, laboratory experiments, yield additional linear equations on the flows , say $AF = b$.

- In biochimical and urban networks, the flows are known to remain between bounds. In trophic networks, they also satisfy more intricate linear inequalities.
  All in all, $GF \leq h$, where G is the identity for biochimical and urban, and has some other non null components for trophic.

The polytope of solutions

- Deterministic approach:

$$\mathcal{S} = \{F = (F_{ij})_{ij \in E} \in \mathbb{R}_+^n : \; AF = b, \; GF \leq h\},$$

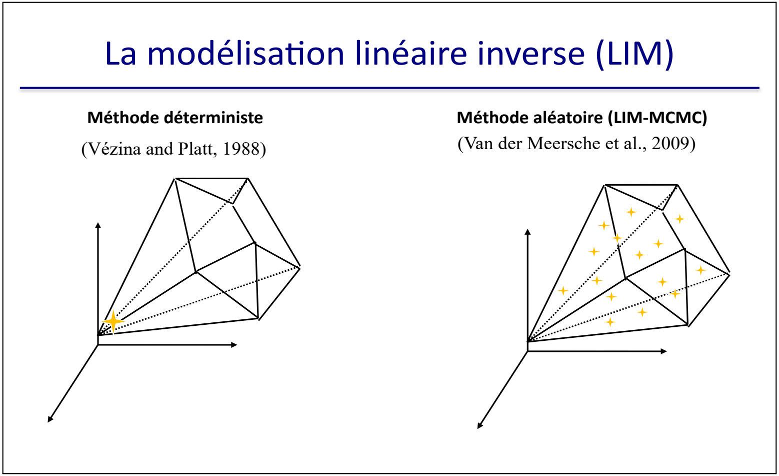
  where A and G are $m \times n$ and $k \times n$ matrices with $m < n$

- Stochastic approach:

$$\mathcal{S} = \{f = (f_{ij})_{ij \in E} \in [0,1]^n : \; Af = b/F_{..}, \; Gf \leq h/F_{..}\},$$

  where $f_{ij} = F_{ij}/F_{..}$ is the proportion of flows from vertex i to vertex j, with $F_{..} = \sum_{ij \in E} F_{ij}$.

- intersection of half-spaces with bounds on the flows
  $\rightarrow \mathcal{S}$ is a very anisotropic high dimensional polytope.
  $n \sim 100$, $\text{Range}_{\text{min}} \sim 10^{-6}$ and $\text{Range}_{\text{max}} \sim 10^{3}$ are usual.



La modélisation linéaire inverse (LIM)

Méthode déterministe

(Vézina and Platt, 1988)

Méthode aléatoire (LIM-MCMC)

(Van der Meersche et al., 2009)

[Saint Béat et al. 2015]

The optimization problem for a deterministic approach

[Caputo, Girardin, Knippel, Niquil, Nguyen, Noguès 2021] proposes to select a solution by minimizing different information theory tools, among which the classical Ecological Network Analysis indices in ecology.

$$\min_{\mathcal{S}} \; ENA(F) = \min \qquad \mathrm{ENA}(F)$$
$$AF = b$$
$$GF \leq h$$
$$F_{ij} \geq 0, \; \forall i, j$$

Since the biological constraints were given using a LIM file in R , we used that infrastructure for the optimization.
The method is the Sequential Quadratic Programming with an Augmented Lagrangian Solver, in the R library NlcOptim.

Ecological Network Analysis $\subset$ Information Theory

[Vezina & al. 1988] proposed to select the least square of flows:

$$Q(F) = \textstyle\sum_{ij\in E} F_{ij}^2.$$

- Mc Arthur index 1955 = Shannon entropy 1948

$$C(F) = \mathbb{S}(F) = -\sum_{ij\in E} (F_{ij}/F_{..}) \log \left(F_{ij}/F_{..}\right).$$

- Ascendency 1984 = Mutual information 1951

$$A(f) = \sum_{ij\in E} F_{ij}/F_{..} \log \left(F_{ij}F_{..}/F_{i.}F_{.j}\right).$$

- Overhead 1997 = Symmetrized conditional entropy 1959

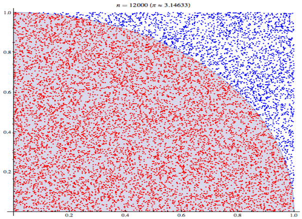
$$\Phi(f) = -\sum_{ij\in E} F_{ij}/F_{..} \log \left(F_{ij}^2/F_{i.}F_{.j}\right).$$

Stochastic approach: Uniform sampling of polytopes

- Basic Monte Carlo methods fail to generate efficiently uniform samples in high dimensional polytopes :

$$\frac{\text{Volume}(\mathcal{S})}{\prod_{ij\in E}[m_{ij}, M_{ij}]} \underset{n \nearrow}{\searrow\searrow}$$

makes the classical rejection method inefficient.



- As an alternative, numerous Monte Carlo Markov Chain methods have been introduced for generating samples drawn uniformly from a polytope.
  A Markov chain is designed, with the desired asymptotic uniform distribution over the polytope. A set of N points in the polytope is obtained by simulating N draws of the chain.

Two classical MCMC Algorithms

In the (Coordinated) Hit and Run [Turching 1971], [Smith 1984], flows $f^{(1)}, \ldots, f^{(N)}$, are iteratively built by repeating:

1. Choose a random direction $d_i =$ a realization of the uniform distribution on the unit sphere of $\mathbb{R}^n$ (among n coordinates);
2. Determine the two intersections points $I^{(i)+}$ and $I^{(i)-}$ between the between the line passing through $f^{(i-1)}$ and directed by d_i, and the borders of the polytope;
3. Sample uniformly a point in the segment $[I^{(i)+}, I^{(i)-}]$ and keep it as $f^{(i)}$.

When the polytope is very anisotropic, CHR needs many iterations to achieve uniformity: the convergence time is

$$O(n^2 \mathrm{Range}^2_{\mathrm{max}} / \mathrm{Range}^2_{\mathrm{min}}).$$

For instance, $R_{\mathrm{min}} \sim 10^{-6}$ and $R_{\mathrm{max}} \sim 10^3$, $n = 100$ yield $O(10^{19})$ or $O(10^{22})$.

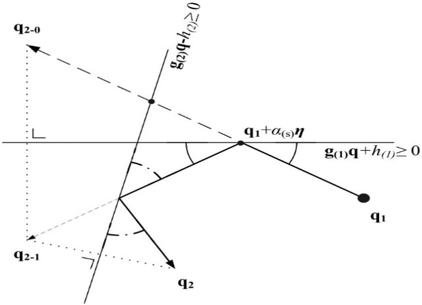
Reflective Hamiltonian MCMC Algorithms

Two similar variants of HR, called Mirror Walk and Billard Walk, have been independently designed in [Van den Meersche 2010] and [Polyak 2014]:

if the trajectory reaches a border of the polytope before the random distance is achieved, then it is reflected on the border.

Also, the direction and distance are drawn simultaneously as the hyperspheric coordinates of a Gaussian vector $\mathcal{N}(0, \sigma^2 \mathrm{Id})$, where $\sigma^2 > 0$ is called the jump.

In other words, $L = \sqrt{-\sigma^2 \log U}$, where $U \sim \mathcal{U}[0,1]$.


$\mathbf{q}_{2\text{-}0}$
$\mathbf{g}_{(2)}\mathbf{q}-h_{(2)}\geq 0$
$\mathbf{q}_1+\alpha_{(s)}\boldsymbol{\eta}$
$\mathbf{g}_{(1)}\mathbf{q}+h_{(1)}\geq 0$
$\mathbf{q}_1$
$\mathbf{q}_{2\text{-}1}$
$\mathbf{q}_2$

xsolve/xsample() ⟶ samplelim/rlim()

- [Van den Meersche et al. 2010] introduced Mirror Walk in limsolve, an R package dedicated to sample LIM in trophic systems in ecology.
- **+** :
  - Good quality of the obtained samples: very low correlation between two consecutive points.
  - Adaptive exploration of the polytope: length of trajectories proportional to the range of flows, choice of the jump length;
  - Annex functions computing quantities of interest for ecology.
- **−** :
  - For highly anisotropic polytopes with high dimensions, xsample() is VERY slow [Fallahi et al. 2020];
  - Presence of several bugs;
  - Code poorly written in R.
- Conclusion: limsolve was developed 15 years ago ⟶ Need for a substantial update

samplelim/rlim()

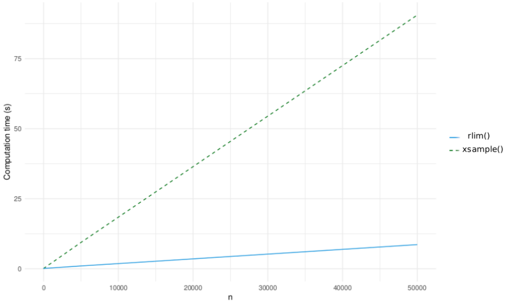
[Girardin, Grente, Niquil, Noguès, Regnault 2024]

- Based on the R package volesti designed for computing the volume of polytopes [Chalkis et al. 2021];
- rlim() uses the same exploration method (MCMC mirror) as xsample() in limsolve;
- Fully coded in C++ with an R interface: invisible for users but MUCH faster computation time;
- Corrects annex functions in limsolve that are of interest for ecology.

Comparaison time

- On the true model of [Nogues et al. 2021] with 144 flows:
  (sample size 500 000 and jump=0.05)
  `xsample` = 5 days ⟶ `rlim` = 3 hours
- On a reduced model of [Caputo et al. 2021] with 28 flows :

n	`rlim()`	`xsample()`
50	0.240	0.226
100	0.137	0.291
500	0.21	1.106
1 000	0.304	1.964
5 000	1.012	9.398
10 000	1.852	18.421
50 000	8.629	90.563



Work in Progress: Studying the Impacts of Degradation

$$\mathcal{S} = \{f = (f_{ij})_{ij \in E} \in \mathbb{R}^n : \ f_{ij} \geq \varepsilon_{ij}, \ Af = b, \ Gf \leq h\}$$

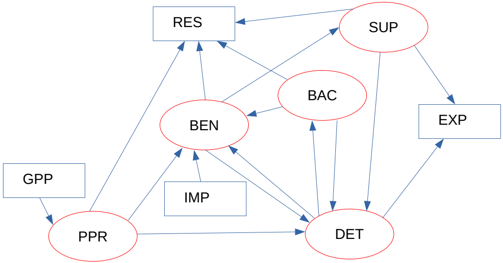
- Degrading Equation s, of the form $\sum a_{ij}^s f_{ij} = b^s$, means: replacing this equation, that is removing the s-th row of matrix A, and adding two new inequations $\sum a_{ij}^s \leq (1+\delta)b^s$ and $\sum -a_{ij}^s \leq (\delta - 1)b^s$ in matrix G.
- The so-called relaxing coefficient is usually $\delta = 0.1$ or 0.3 in ecology.
- The set $\mathcal{S}_{\overline{P}} \subset \mathcal{S}$ with $P \subseteq \{1, \ldots, m\}$ and $\overline{P} = E \setminus P$, is the set of flows where all equations with index $s \in P$ have been degraded. Such a degradation will be said to be of level $d = |P|$.

Work in Progress: Studying the Impacts of Degradation

A global method for assessing the impact of degradation on ENA:

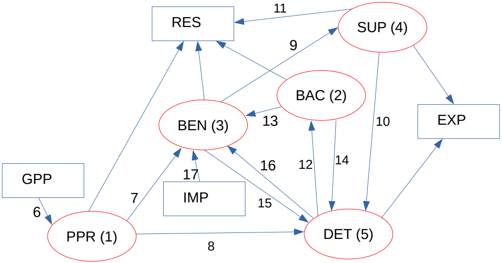
1. start from a solution set of flows $\mathcal{S}$, and compute a reference value for the ENA;
   The reference value for a given ENA is computed as a function (mean, minimum, maximum, or range of all values) of its values for each solution of a representative sample of $\mathcal{S}$.
2. degrade the equations either one by one, or in all possible ways ($2^m - 1$ possibilities), or according to expert's advice;
3. compute the new value of the indexes for each of these degraded systems, as in step 1.
4. compare these new values and the reference values computed in step 1;
5. aggregate the results by level of degradation.

A fully determined trophic network



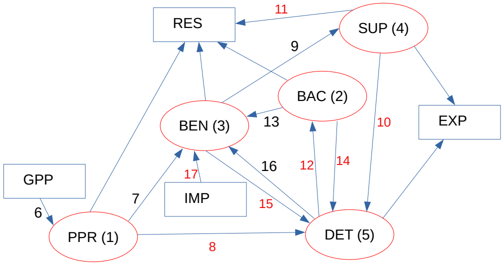
- Trophic network with 5 components and 17 flows

A fully determined trophic network



- Trophic network with 5 components and 17 flows
- Matrix A of dimension $17 \times 17 \Rightarrow$ fully determined problem

A fully determined trophic network



- Trophic network with 5 components and 17 flows
- Matrix A of dimension $17 \times 17 \Rightarrow$ fully determined problem
- 7 equalities have been selected to be degraded with $\delta = 0.3$

Which ENA ?

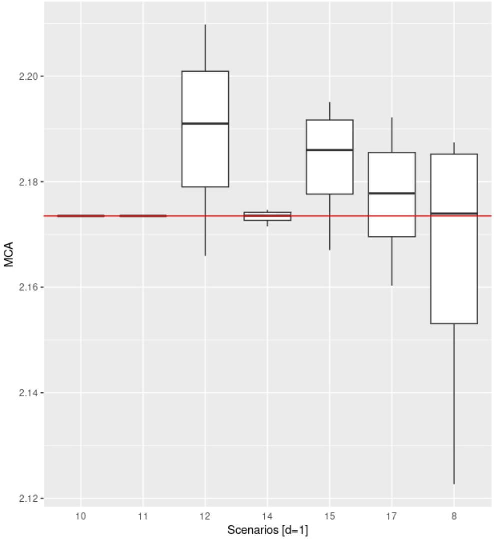
- Quadratic Energy

$$QE(F) = \textstyle\sum_{ij \in E} F_{ij}^2 .$$

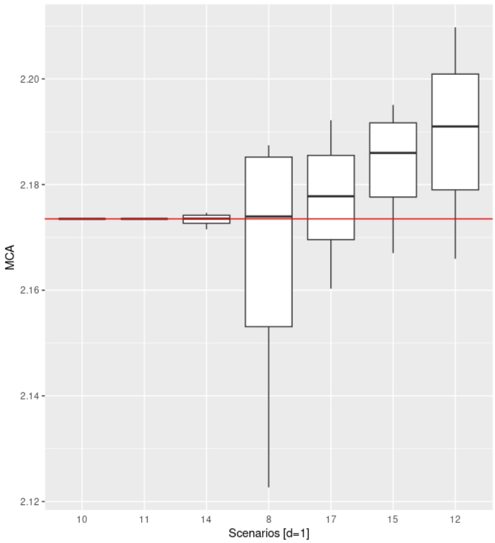
- Mc Arthur index

$$MCA(F) = -\sum_{ij \in E} (F_{ij}/F_{..}) \log (F_{ij}/F_{..}) .$$

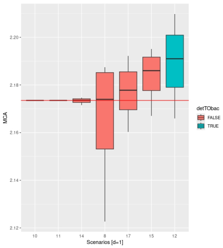
MCA and degradation of level 1

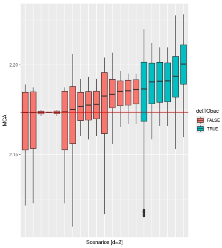


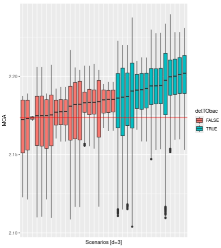
MCA and degradation of level 1

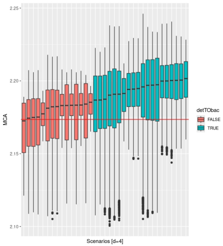


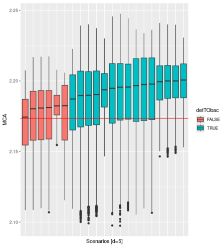
Most impactful flow on MCA

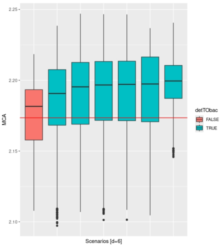




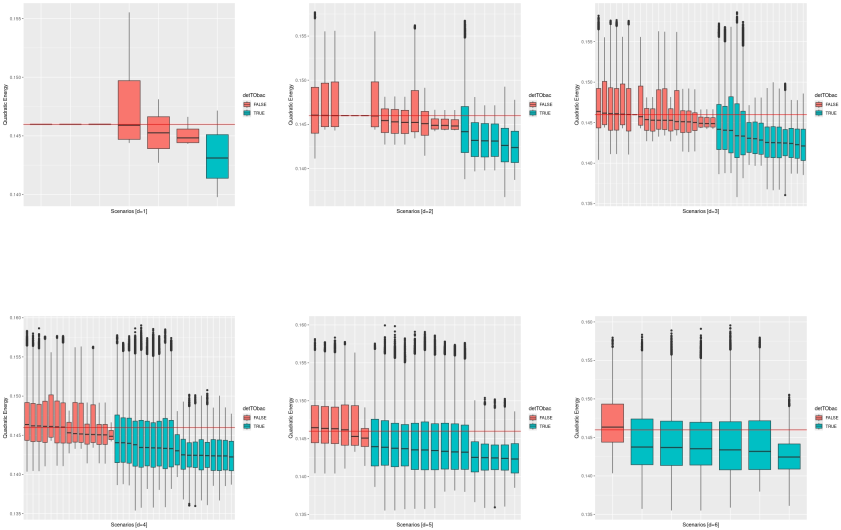




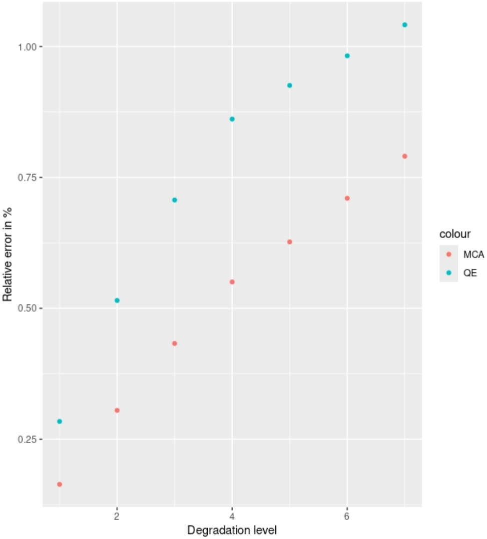




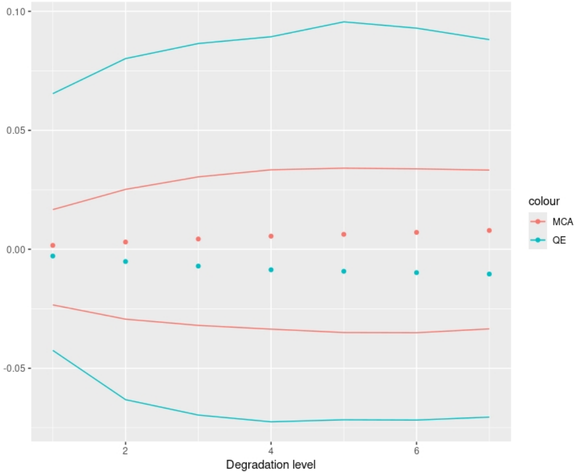
Most impactful flow on QE



Comparing ENA



Comparing ENA



Work in Progress: Divergence-like Goal Functions

- ENA indices or **entropic functions**: classicaly used to bring more information on a model or as a tool for comparing two of them;
- Adding to the collection **divergence-like functions**, that take into account reference pdfs, should lead to a better fit to a priori information on the ecosystem.
- These reference pdf f^* can be solutions to the problem obtained in a previous study or a previous year (not solution but reasonable reference), the middle of the constraint intervals, etc.

In this aim, the Kullback-Leibler divergence is the most classical tool in information theory:

$$\mathbb{K}(f|f^*) = \sum_{ij \in E} f_{ij} \log\left(\frac{f_{ij}}{f^*_{ij}}\right),$$

Rényi entropy and divergences

They have not yet been used in ecological networks, but should give a better fit of the goal function to the problem through the choice of a positive parameter $s \neq 1$.

- Rényi entropy

$$R_s(f) = \frac{1}{1-s} \log \left[\sum_{ij \in E} (f_{ij})^s \right] ,$$

- Divergence associated to Rényi entropy

$$R_s(f|f^*) = \frac{1}{1-s} \log \left[\sum_{ij \in E} \frac{(f_{ij})^s}{(f^*_{ij})^{s-1}} \right] .$$

- Rényi mutual information:

$$A_s(f) = R_s(f|(f_{i.}) \otimes (f_{.j})) = \frac{1}{1-s} \log \left[\sum_{ij \in E} \frac{(f_{ij})^s}{(f_{i.} f_{.j})^{s-1}} \right] .$$

References

- Analysis of trophic networks: an optimisation approach, *Journal of Mathematical Biology* 2021
  (ECUME RIN)

  Jean-Guy Caputo, Arnaud Knippel, Hieu Nguyen, INSA Rouen,
  Valérie Girardin, LMNO, UNICAEN,
  Nathalie Niquil, Quentin Noguès, BOREA UNICAEN
- Comparing and updating R packages of MCMC Algorithms for Linear Inverse Modeling of Metabolic Networks *HAL* 2024
  (NESTORE CoRed ANR/FEM 2022)

  Valérie Girardin, LMNO, UNICAEN
  Théo Grente, LMNO, UNICAEN & France Énergies Marines
  Nathalie Niquil, Quentin Noguès, BOREA UNICAEN
  Philippe Regnault, LMR, URCA