Analyse de l'impact de variables environnementales sur les réseaux plantes-pollinisateurs à l'aide d'auto-encodeurs variationnels pour graphes bipartites

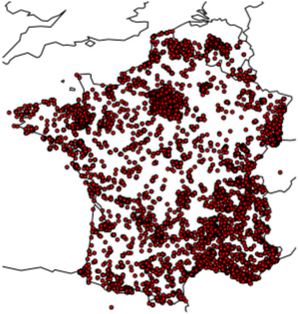
Emre Anakok[1]
supervised by Pierre Barbillon[1], Colin Fontaine[2] & Elisa Thebault[3]

[1] UMR MIA Paris-Saclay, Palaiseau
[2] Centre d'Écologie et des Sciences de la Conservation, MNHN, Paris
[3] Sorbonne Université, Institute of Ecology and Environmental Sciences, Paris

26/03/2024

Outline

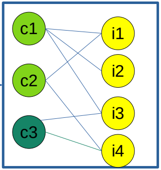
Context



- Citizen science program dataset from 2010 to today
- More than 580 000 entries of plant-pollinator interactions
- Timed observation sampling

From dataset to pollination network

Network from dataset



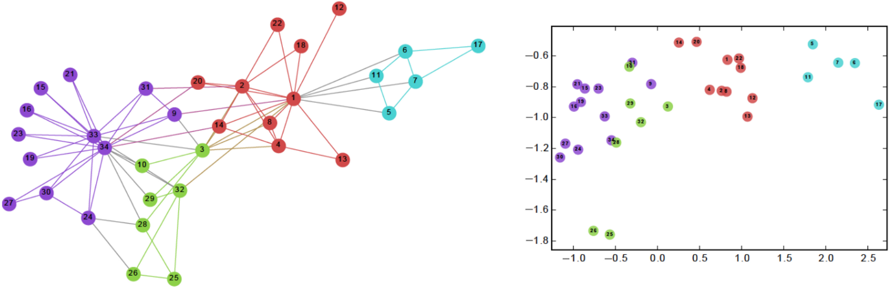
Collection	Plant species	Insect species	Temperature (T)	Date (D)
1	1	1	10°C	03/03/2020
1	1	2	10°C	03/03/2020
1	1	3	10°C	03/03/2020
2	1	1	20°C	08/08/2020
2	1	4	20°C	08/08/2020
3	2	3	10°C	04/04/2020
3	2	4	10°C	04/04/2020

Features from dataset

Features for collection nodes only.
T_{c1}= 10°C
T_{c2}= 20°C
D_{c1}= 03/03/2020
Plant species is also a feature encoded with {0,1}.

Double goal

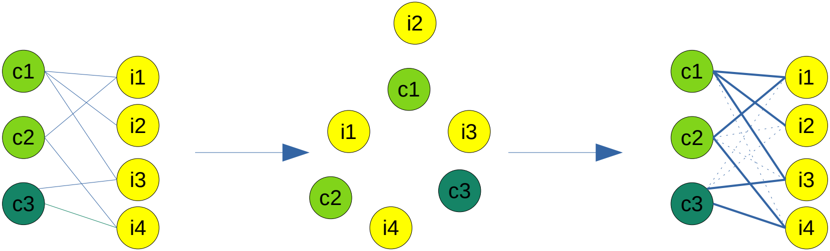
- Embedding : represent nodes and features with $Z \in \mathbb{R}^d$.
- Fairness : Z independent of S.



(picture from https ://towardsdatascience.com/overview-of-deep-learning-on-graph-embeddings-4305c10ad4a4)

Bipartite adaptation of VGAE

$$\underset{\in(\mathcal{G},E,V)}{B, X_1, X_2} \xrightarrow[\text{encoder}(GCN)]{q(Z_1,Z_2|X_1,X_2,B)} \underset{\in\mathbb{R}^d}{Z_1, Z_2} \xrightarrow[\text{decoder}(distance)]{p(B|Z_1,Z_2)} \underset{\in\mathbb{R}^{n_1\times n_2}}{\widehat{B}}$$



Hilbert-Schmidt Independence Criterion

- $X \in \mathcal{X}$ compact with kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
- $Y \in \mathcal{Y}$ compact with kernel $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

$$\begin{aligned} HSIC(X,Y) = ||C_{X,Y}||^2 \\ &= \mathbb{E}_{XYX'Y'}[K(X,X')L(Y,Y')] + \mathbb{E}_{XX}[K(X,X')]\mathbb{E}_{YY'}[L(Y,Y')] \\ &- 2\mathbb{E}_{XY}[\mathbb{E}_{X'}[K(X,X')]\mathbb{E}_{Y'}[L(Y,Y')]]. \end{aligned}$$

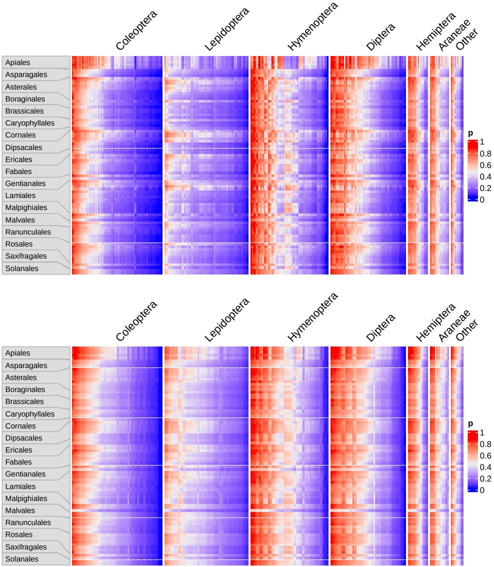
- $K(x_i,x_j) = e^{-\frac{||x_i-x_j||^2}{2\sigma^2}} \implies (HSIC(X,Y) = 0 \iff X \perp Y)$ (Gretton et al., 2005).
- HSIC test : if $X \perp Y$, $n \times \widehat{HSIC} \sim \frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\beta^\alpha \Gamma(\alpha)}$ (Gretton et al., 2007)

Loss with HSIC

- Reconstruction loss from (Kipf and Welling, 2016)
- HSIC (Gretton et al., 2005) as a penalty to have independence
- Variational lower bound $\mathcal{L}$:

$$\begin{aligned}\mathcal{L}(W_1, W_2) = \mathbb{E}_{q(Z_1,Z_2|X_1,X_2,B)}[\log p(B|Z_1, Z_2)] \\ -KL[q_1(Z_1|X_1, B)||p_1(Z_1)] \\ -KL[q_2(Z_2|X_2, B)||p_2(Z_2)] \\ +\delta RFF\ HSIC(\mu_1, S)\end{aligned}$$

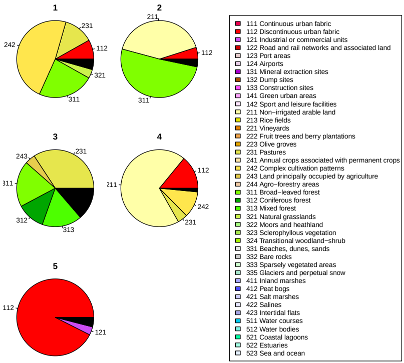
Estimated probabilities of connection between plants and insects on the Spipoll data set, BVGAE is on the top and the fair-BVGAE is on the bottom.



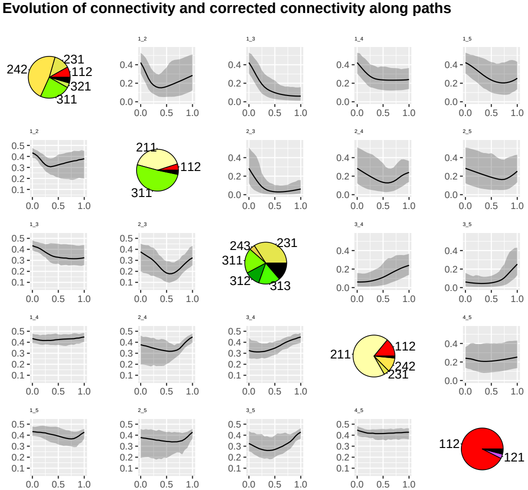
Outline

Work in progress

Typical landscape at places of sampling.



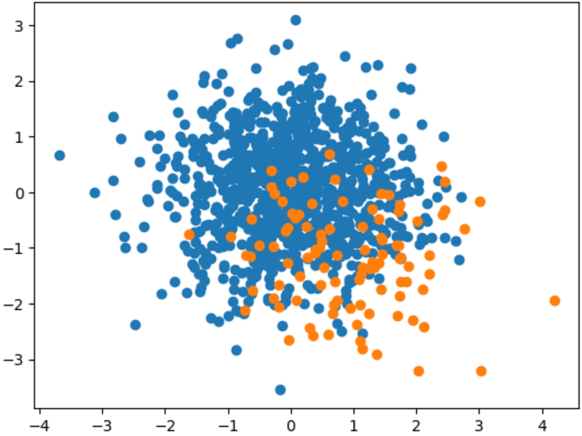
Work in progress



- We try to predict how changing the landscape changes the structure $f(X, B)$ of the network.
- This goal is achieved by changing the input value X.
- Accounting for sampling bias changes the results.

Is this truly interpretable ?

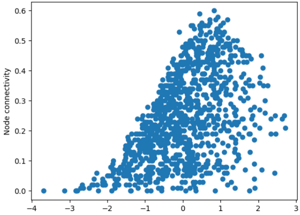
Simulation setting



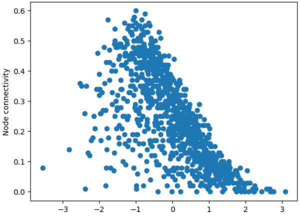
Simulated latent space

- We simulate $x_1, x_2, x_3 \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $y_1, y_2 \overset{\text{iid}}{\sim} \mathcal{N}(1,1)$.
- We generate a network depending on x_1, x_2 (in blue) and y_1, y_2 (in orange).
- Goal : observe the change of connectivity in the network.

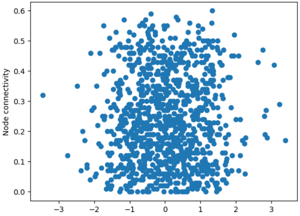
Simulation setting



x_1

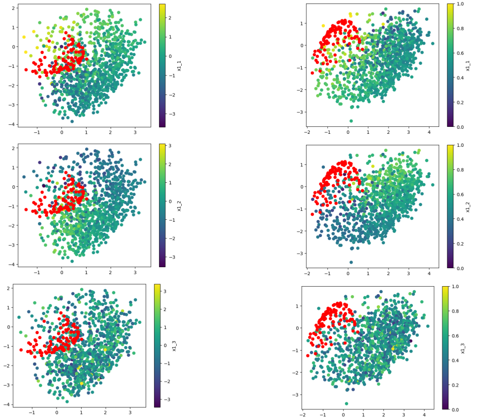


x_2



x_3

Were the covariates useful ?



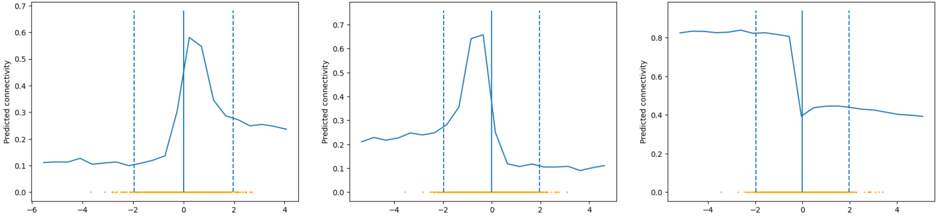
Latent space of f_0 Latent space of f

- Train $f_0(X, B)$ with $X = I_{n_1}$
- Train $f(X, B)$ with $X = [x_1, x_2, x_3]$

- f_0 AUC : 0.821 ± 0.012
- f AUC : 0.822 ± 0.013

Connectivity prediction

- Predict the connectivity : $f(X, B)$.
- Change the input X to see the effect.



x_1 x_2 x_3

Changing x_1 and x_2 yields the correct behavior, but the model is sensitive to x_3, which has no relationship with the data.

Feature importance

- Feature importance : score ϕ that indicates how much a feature j contributes to a prediction $f(X)$ of the model.

Possible feature importance estimation :

- Shapley values
- Gradient
- Integrated Gradient

Shapley values

- $S \in \{0,1\}^F$ represent a coalition of selected features.
- $X_S = \{X_{i,k} | k \in S\}$ is the submatrix of selected columns.
- $val(S) = \mathbb{E}[f(X)|X_S = x_s] - \mathbb{E}[f(X)]$ captures the marginal contribution of the coalition S

$$\phi_j(val) = \sum_{S \subseteq 1,\dots,F \setminus \{j\}} \frac{|S|!(F-|S|-1)!}{F!}(val(S \cup \{j\}) - val(S))$$

Estimating Shapley values for graphs

Duval and Malliaros (2021)

- let $\mu = (\mathbb{E}[X_1], ..., \mathbb{E}[X_F])$.
- Pick $z \in \{0, 1\}^F$ uniformly at random.
- Let X' such as $X'_j = \begin{cases} X_j & \text{if } z_j = 1 \\ \mu_k & \text{otherwise} \end{cases}$ and predict $f(X')$.
- Construct data set $\mathcal{D} := \{(z, f(X'))\}$.
- Apply weighted linear regression on $\mathcal{D}$, each coefficient corresponds to an estimated Shapley value ϕ_j.

Gradient

For each node i and for each feature j :

$$\phi_{i,j} = \nabla[f(X)]_{i,j}$$

$$\phi_j = \frac{1}{n_1} \sum_{i=1}^{n_1} \phi_{i,j}$$

Integrated gradient 1

- Chose a baseline X'
- For each node i and for each feature j calculate

$$IG_j(x_i) = (x_{i,j} - x'_{i,j}) \times \int_{\alpha=0}^{1} \nabla[f(X' + \alpha(X - X'))]_{i,j} d\alpha$$

- for each feature j fit linear regression on $\{(x_{i,j}, IG_j(x_i))\}$, each slope corresponds to an estimation of ϕ_j

Integrated gradient 2

- Let $g : \mathbb{R}^F \mapsto \mathbb{R}$ such as $g(y_1, ..., y_F) = f(\mathbf{1}.(y_1, ..., y_F)^\top)$
- Chose a baseline y'
- For each feature j calculate

$$IG_j(y) = (y_{i,j} - y'_{i,j}) \times \int_{\alpha=0}^{1} \nabla f(y' + \alpha(y - y'))_j d\alpha$$

Result

Sign : $\phi_1 > 0$ and $\phi_2 < 0$
Magnitude : $|\phi_1| > |\phi_3|$ and $|\phi_2| > |\phi_3|$

	Shapley	Grad	IG1	IG2
sign	0.4	1	1	1
magnitude	0.93	0.86	0.83	0.90

Duval, A. and Malliaros, F. D. (2021). Graphsvx : Shapley value explanations for graph neural networks. CoRR, abs/2104.10482.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring Statistical Dependence with Hilbert-Schmidt Norms. In Jain, S., Simon, H. U., and Tomita, E., editors, Algorithmic Learning Theory, Lecture Notes in Computer Science, pages 63–77, Berlin, Heidelberg. Springer.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. (2007). A kernel statistical test of independence. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, Advances in Neural Information Processing Systems, volume 20. Curran Associates, Inc.

Kipf, T. N. and Welling, M. (2016). Variational graph auto-encoders.