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] & Pierre Barbillon [2] & Colin Fontaine [3] & Elisa Thebault [4]

[1] *Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120, Palaiseau, France, emre.anakok@agroparistech.fr*
[2] *Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120, Palaiseau, France, pierre.barbillon@agroparistech.fr*
[3] *Centre d'Écologie et des Sciences de la Conservation, MNHN, CNRS, SU, 43 rue Buffon, 75005 Paris, France, colin.fontaine@mnhn.fr*
[4] *Sorbonne Université, CNRS, IRD, INRAE, Université Paris Est Créteil, Université Paris Cité, Institute of Ecology and Environmental Sciences (iEES-Paris), 75005 Paris, France, elisa.thebault@upmc.fr*

Résumé. Nous proposons une méthode de représentation des réseaux bipartites à l'aide d'auto-encodeurs de graphes adaptés à l'étude des réseaux écologiques issus de données de sciences participatives. Ceci représente un double défis, puisque l'on veut mettre en évidence les effets de nombreuses covariables d'intérêt écologique, comme par exemple la dégradation des habitats, tout en prenant en compte les effets d'échantillonnage, notamment l'effet observateur. Nous adaptons l'approche de l'auto-encodeur variationnel de graphes au cas bipartite pour générer des représentations dans un espace latent où les deux ensembles de nœuds sont positionnés en fonction de leur probabilité de connexion. En incorporant le critère d'indépendance de Hilbert-Schmidt (HSIC) comme un terme de pénalité supplémentaire dans la perte que nous optimisons, nous nous assurons que la structure de l'espace latent est indépendante des variables continues, qui sont liées au processus d'échantillonnage. Nous appliquons notre méthode à l'ensemble de données Spipoll, un programme d'observation participatif des interactions entre plantes et pollinisateurs à travers toute la France auquel contribuent de nombreux observateurs, ce qui le rend biaisé car les participants sont sujet à un phénomène d'apprentissage au fur et a mesure de leur participation. Enfin, nous prédisons les changement de structure du réseau de pollinisation en fonction de variations de composition du paysage, avec ou sans prise en compte de l'experience des observateurs. Les resultats mettent en lumière l'importance de la correction des biais d'echantillonnage, avec par exemple, une connectivité du réseau largement augmenté dans les paysages agricole dominés par de l'elevage lorsque les biais d'échantillonnage sont corrigés.

Mots-clés. Graphes et réseaux, Réseaux de neurones, Échantillonnage, Statistique appliquée à l'écologie

Abstract. We propose a method to represent bipartite networks using graph embeddings tailored to tackle the challenges of studying ecological networks derived from citizen science data. This represents a twofold challenge, since we want to highlight the effects of numerous

covariates of ecological interest, such as habitat degradation, while taking into account sampling effects, in particular the observer effect. We adapt the variational graph auto-encoder approach to the bipartite case, which enables us to generate embeddings in a latent space where the two sets of nodes are positioned based on their probability of connection. By incorporating the Hilbert-Schmidt independence criterion (HSIC) as an additional penalty term in the loss we optimize, we ensure that the structure of the latent space is independent of continuous variables which are related to the sampling process. We apply our method to the Spipoll dataset, a citizen science monitoring program of plant-pollinator interactions to which many observers contribute, making it prone to sampling bias because observers are subject to a learning phenomenon as they participate. Finally, we predict changes in the structure of the pollination network in response to variation in landscape composition, with or without taking into account the experience of the observers. The results highlight the importance of correcting for sampling bias, for example the network connectivity greatly increases in agricultural landscapes dominated by livestock when sampling bias is corrected.

Keywords. Graphs and networks, Neural networks, Sampling, Applied statistics in ecology

1 Introduction

Graph embedding regroups different methods, allowing to represent a network into a vector space in order to gain understanding of key network features. These methods are especially important in the context of large networks. Recently developed graph neural networks (GNNs) enable graph embedding with large-scale methods such as graph isomorphism network, graph attention network or the variational graph variational auto-encoder [Kipf and Welling, 2016]. All these methods can also handle numerous covariates on nodes. GNNs are currently growing in popularity in various domains such as bioinformatics, chemistry and geophysics, but they remain mostly unknown in other research fields.

In ecology, networks have been analyzed to study various types of ecological interactions among species, such as plant-pollinator, predator-prey or host-parasite interactions. The stochastic block model and the latent block model for bipartite graphs are notorious models using latent variable in ecology. While graph embedding methods start being used for ecological networks, GNNs have yet to be diffused in that research field. GNNs could be particularly relevant for ecological networks because very large data sets of interactions among species are now becoming available (e.g. through the development of citizen science programs) in addition to many covariates at the node level (e.g. species name and traits, environmental conditions at the time of interaction observation). An important issue with the analysis of ecological networks concerns the strong effects of sampling effort and methods on the observed network structure [Doré et al., 2021]. One could wish to have an embedding which is independent of a certain set of covariates linked to such sampling effects and related bias. This can be of particular interest for citizen science programs, where biases can arise from the large number observers involved with various experience levels [Jiguet, 2009, Deguines et al., 2018]. It is also known that the biotic and abiotic context of individual plants can

influence pollinator foraging behavior [Arroyo-Correa et al., 2021]. Especially, the land use plays a major role in the structure of the interaction network, as different pollinator groups, such as bees, flies or butterflies, have different affinities for different land-uses, e.g. urban or agricultural areas [Deguines et al., 2012]. For instance, diverse preferences among pollinators for various land-uses might result in reduced interconnectedness of plant-pollinator networks in urban environments compared to rural areas [Geslin et al., 2013, Cortina et al., 2022].

We aim to study the influence of environmental variables, such as the land use on the network structure using GNNs. To explore the impact of environmental conditions on the network structure, we first adapt the graph variational auto-encoder [Kipf and Welling, 2016] to the bipartite case, where the embedding should also be independent of covariates linked to sampling effect. After the learning phase, we analyze how the network embedding predicted by the GNN evolves with changes in input covariates, which should reflect realistic landscape composition.

In the following, a background on GNNs and the HSIC criterion is provided. Then, the model is introduced and applied to the Spipoll data set [Deguines et al., 2012], a citizen science program monitoring plant-pollinator interactions across France since 2010. Finally, we rely on the fitted model to assess how the land use impact the network connectivity.

2 Model

Embedding the nodes of a graph in a vector space using a variational graph auto-encoder (VGAE) [Kipf and Welling, 2016] would yield a Gaussian latent representation Z. We aim to have Z independent of a set of covariates linked to the sampling process S. As we cannot guarantee that Z and S are jointly Gaussian, we will use another criterion than the covariance to have independence between Z and S: the Hilbert-Schmidt independence criterion (HSIC), first proposed by Gretton et al. [2005] which is a metric testing for the independence of two variables. Compared to the other proposed methods of embedding, the probabilistic setting of the GVAE fits well with the use of the HSIC, and its generative aspect allows network generation for various ecological contexts.

2.1 Bipartite variational graph auto-encoder

We adapt the variational graph auto-encoder from Kipf and Welling [2016] to the bipartite case by considering two graph convolutional networks (GCN), one for each node types.

We consider a biadjacency matrix $B_{i,j}$ of size $n_1 \times n_2$ representing our graph. Let

$$D_1 = diag\left(\sum_{j=1}^{n_2} B_{i,j}\right) \; D_2 = diag\left(\sum_{i=1}^{n_1} B_{i,j}\right)$$

be respectively the row and the column degree matrices. For each i and each j we consider the stochastic latent variables Z_{1i} and Z_{2j} which are described respectively by a $n_1 \times D$ and a $n_2 \times D$ matrices (they share the same number of columns). X_1 is a $n_1 \times d_1$ matrix of node

features for the first category, and X_2 is a $n_2 \times d_2$ matrix of node features for the second. Finally, we consider the normalized biadjacency matrix $\tilde{B} = D_1^{-\frac{1}{2}} B D_2^{-\frac{1}{2}}$.

2.1.1 Encoder

The encoder is defined as

$$q(Z_1, Z_2 | X_1, X_2, B) = \prod_{i=1}^{n_1} q_1(z_{1i} | X_1, B) \prod_{j=1}^{n_2} q_2(z_{2j} | X_2, B)$$

with

$$q_v(z_{vi} | X_v, B) = \mathcal{N}(\mu_{v,i}, diag(\sigma_{v,i}^2)),\ v \in \{1, 2\}$$

with $\mu_v \in \mathbb{R}^d$ and $log(\sigma_v)$ obtained by the GCN_v defined similarly as in Kipf and Welling [2016]:

$$\text{GCN}_1(X_1, B) = \tilde{B}\text{ReLU}(\tilde{B}^\top X_1 W_{1,1}) W_{1,2}, \quad \text{GCN}_2(X_2, B) = \tilde{B}^\top \text{ReLU}(\tilde{B} X_2 W_{2,1}) W_{2,2}$$

with weight matrices W_v. $\text{GCN}_{\mu_v}(X_v, B)$ and $\text{GCN}_{\sigma_v}(X_v, B)$ share the first-layer parameters $W_{v,1}$ and $\text{ReLU}(x) = max(x, 0)$. The parameters $\mu_{1,i}$ and $\mu_{2,j}$ share the same dimension d.

2.1.2 Decoder

The decoder is defined as

$$p(B | Z_1, Z_2) = \prod_{i=1}^{n_1} \prod_{j=1}^{n_2} p(B_{i,j} | z_{1i}, z_{2j}), \qquad \text{with } p(B_{i,j} | z_{1i}, z_{2j}) = e^{-\frac{||z_{1i} - z_{2j}||^2}{2\sigma^2}}.$$

In the following, we fix $\sigma^2 = 1$.

The full auto-encoder can be summarized as

$$B, X_1, X_2 \xrightarrow[encoder]{q(Z_1, Z_2 | X_1, X_2, B)} Z_1, Z_2 \xrightarrow[decoder]{p(B | Z_1, Z_2)} \widehat{B}.$$

2.2 Bipartite and fair graph variational auto-encoder

Our goal is to obtain the latent representation Z_1 of the bipartite variational auto-encoder, independent of a protected variable denoted by S. To do so, we add the HSIC [Gretton et al.,

2005] computed between the posterior means μ_1 and the protected variable S as a penalty term in the loss:

$$L = \mathbb{E}_{q(Z_1,Z_2|X_1,X_2,B)}[\log p(B|Z_1,Z_2)] - KL[q_1(Z_1|X_1,B)||p_1(Z_1)] \\ - KL[q_2(Z_2|X_2,B)||p_2(Z_2)] + \delta RFF\ HSIC(\mu_1,S) \quad (1)$$

where p_1, p_2 are Gaussian priors for Z_1 and Z_2, KL is the Kullback-Leibler divergence, δ is a hyperparameter and RFF HSIC is the random Fourier feature [Rahimi and Recht, 2007, Zhang et al., 2018] estimation of the HSIC. At the end of the learning, we can compute the p-value of the HSIC test [Gretton et al., 2007] to check that the optimization of the loss has made the latent representation $Z_1 \sim \mathcal{N}(\mu_1, diag(\sigma_1^2))$ independent of S. Note that adding the HSIC in the loss may deteriorate the reconstruction of the original data through the decoder since an independence constraint has been added.Full description of the model is available in Anakok et al. [2024].

3 Application on Spipoll data set

3.1 Context

We apply the proposed method on the Spipoll [Deguines et al., 2012] data set, a French citizen science program which aims to monitor plant-pollinator interactions across metropolitan France since 2010. This monitoring follows a simple protocol: briefly, volunteers can choose a flowering plant where and when they like, and during 20 minutes, take pictures of all different insects that land on the flowers of the monitored plant. Then, using an online identification tool, they identify each different insect that has been photographed and upload their data on a dedicated website. Each participation is thus a set of insect interactions with a given plant species that have been observed at a given time and place, and by a given volunteer whose specific skills could affect the quality of the observation. The date and place of observations allowed us to extract corresponding climatic conditions as covariates, from the European Copernicus Climate data set, and the corresponding land use proportion with the Corine Land Cover (CLC).

A common practice in ecology to study plant-pollinator interactions is to consider plant and insect species as nodes of a bipartite network, with edges determined by the observations of insects pollinating plants. Since our data are at the participation level (a session of observation corresponds to a plant species with all the observed insects on that plant during 20 minutes), we consider a bipartite network where the first type of nodes are session of observations, and the second type are insect species observed during the session. Each session has the previously mentioned covariates and a one-hot encoding describing the plant genus. Link prediction task in this situation aims to predict which insect will be present during a given observation session. However, we still wish to ultimately obtain a bipartite plant-insect network, which is standard in this field of study. To ensure that the latent space could also be used to create a plant-insect network, we propose to draw for each taxon of plant one observation from the set of all the session where this plant was monitored.This would

generate another latent space corresponding to plant and insect species, and using the same decoder, would generate a plant-insect network. More details about this model are available in the full presentation of our model [Anakok et al., 2024].

We fit our model, taking into account the specific requirements of the Spipoll dataset. We consider the observation period of the Spipoll dataset from 2010 to 2020 included in metropolitan France, on a set of 83 plant genus that have been monitored every year. This lead to 26267 observation sessions during which 306 insect taxa have been observed. The observation session-insect matrix has a total of 94 909 plant-pollinator interactions reported, and the plant-insect matrix has 9 754 different interactions. The covariates related to the observations sessions are $X_1 = (P, t, \Delta_T, CLC)$ where P is a binarized categorical variable (83 columns) giving the plant genus, t contains the day and the year of observation, Δ_T is the difference between the average temperature on the day of observation and the average of temperatures measured from 1950 to 2010 at the same observation location and CLC describes the proportion of land use with 44 categories (see fig. 1) in a 1000m radius around the observation location. We only consider covariates for the observation sessions, thus the covariates for insects are set as $X_2 = I_{n_2}$. While citizen science programs facilitate the accumulation of observed data, the sampled data may be biased by the participating observers. To take into account this bias, we propose in our model to define S, the protected variable, as the number of participation from the user at the time of observation. This number of participation would work as a proxy of the user's experience. By employing this measure, we aim to construct a latent space that remains unaffected by variations in observers' experience levels.

3.2 Results : Influence of the landscape on network connectivity

Once the model has been fit, we target the impact of the environmental variables on the network structure in particular we focus on the composition of the landscape described by the CLC. To do so, we could change the covariates input related to the landscape to create "pure" landscapes to assess its effect on the predicted network. However, setting a pure landscape with 100% for a specific type and the rest at 0% (e.g. 100% continuous urban fabric and the rest at 0%) would yield unrealistic landscapes. Moreover, the transition from one type of landscape to another is also not simple to simulate. In order to get more realistic landscape simulation, we can seek for typical landscapes by performing a clustering on the CLC indiced. Since the CLC are compositional data (proportions of land use), they are transformed through an isometric log ratio transformation (ILR) and a principal component analysis is computed as proposed by Aitchison [1983]. The typical landscapes are obtained as the centroids of a k-means clustering on the first component of the PCA. Such a clustering is displayed in fig. 2. The centroids of the k-means algorithm are represented with pie-charts. The pie charts composition are detailed in fig. 1, and we assume that they represent typical landscape at places of sampling. For example in fig. 1, landscape 5 is mostly composed of discontinuous urban fabric, which is typical of sampling performed in metropolitan area. Landscape 3 is mostly composed of pastures and forest, landscapes 1 and 2 are made of culture and forest, with a bit of urban fabric, and landscape 4 is mostly arable land. The ILR transform will also allow us to simulate realistic landscape proportion transition, from

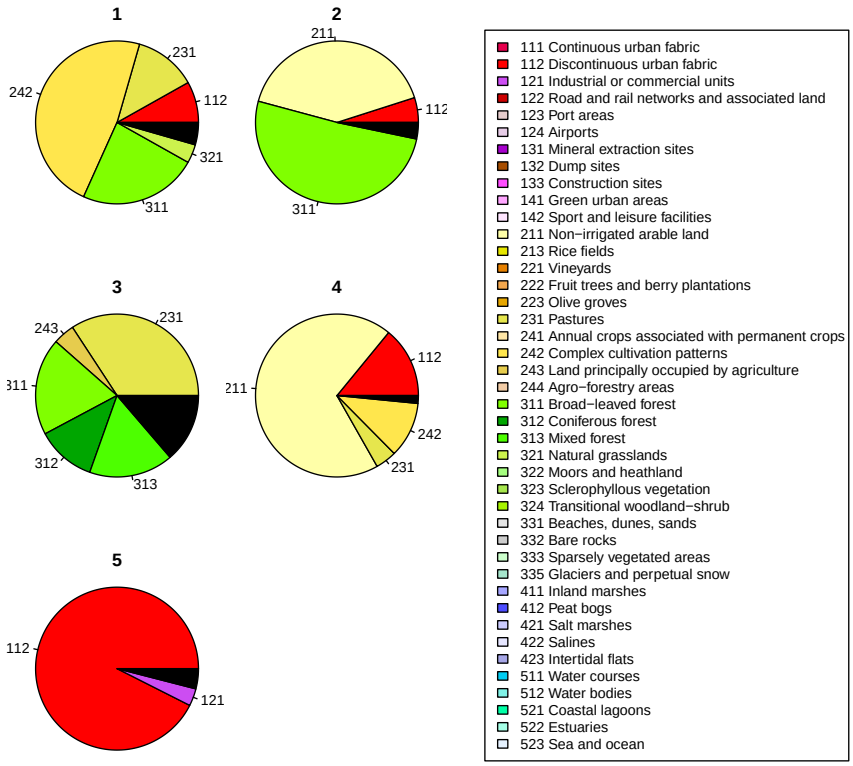

Figure 1: Typical landscapes at sampling localization and CLC legends. All proportions smaller than 4% are regrouped and colored in black for better visibility.

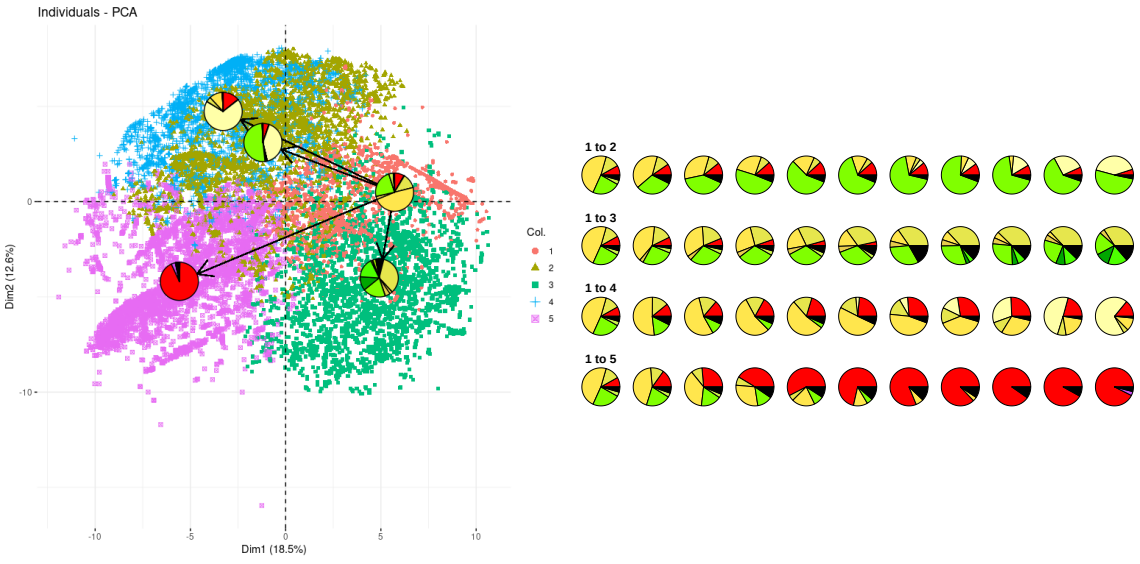

Figure 2: On the left, the first two principal components of the PCA performed on the ILR of the land use proportion of each in observation session in the Spipoll dataset. The arrows represent some simulated trajectories of landscape change from the first typical landscape to the others. On the right, we see how these trajectories change the simulated landscape.

one landscape to the other (fig. 2) by using convex combination of the ILR transform of the typical landscapes.

We fit the model 100 times on the Spipoll dataset,with different initializations and different train-test splits, with and without taking into account the observers' experience levels to compare the predictions. After the learning phase, we input the simulated transition of landscape in the fitted algorithm and observe how it changes the network. The result for connectivity prediction can be seen in fig. 3. Taking into account the observers' experience level increases the probability of connection. Going from landscape 1, which is mostly composed of complex cultivation patterns and broad-leaved forest, to landscape 3, which is mostly made of pasture and forest, drastically decreases the connectivity of the plant-pollinator network from 0.4 to almost 0 (trajectory 1_3), unless you take into account the observers' experience levels (trajectory F 1_3), where we see that the decreasing is less pronounced. The uncorrected connectivity seems to be the highest in landscape 1 but with the correction, the connectivity in landscape 4 is on par with the one in landscape 1. The decline in complex cultivation patterns from landscape 1 to other landscapes also correlates with a decline in connectivity in the network when we do not take into account the observers' experience levels, but this effect seems to be less visible in the debiased estimation. These observations provide insights into the relationship between land use and the network structure, and how taking into account the sampling bias could change our understanding of the network.

4 Perspectives

In this exploratory work, we demonstrate how our approach can take into account the bias induced by the participating observers, alongside an examination of how varying covariates induces changes in the plant-pollinator network structure. However, there remains further exploration into the impact of these environmental covariates. Additional metrics, such as modularity, nestedness or robustness can be studied to enhance the understanding of the network evolution. Categorizing organisms by order of insect or plants could reveal instances where overall connectivity decreases for certain order while increasing for others. Expanding the study to the effect of the temperature, or other environmental covariates could also provide valuable insights on various ecological questions. Other sampling bias, such as the uneven distribution of the sampling location on the territory, could be taken into account. All of these aspects will be taken into consideration in future work.

References

J. Aitchison. Principal component analysis of compositional data. *Biometrika*, 70(1):57–65, 1983. ISSN 00063444. URL http://www.jstor.org/stable/2335943.

E. Anakok, P. Barbillon, C. Fontaine, and E. Thebault. Bipartite graph variational auto-encoder with fair latent representation to account for sampling bias in ecological networks, 2024.

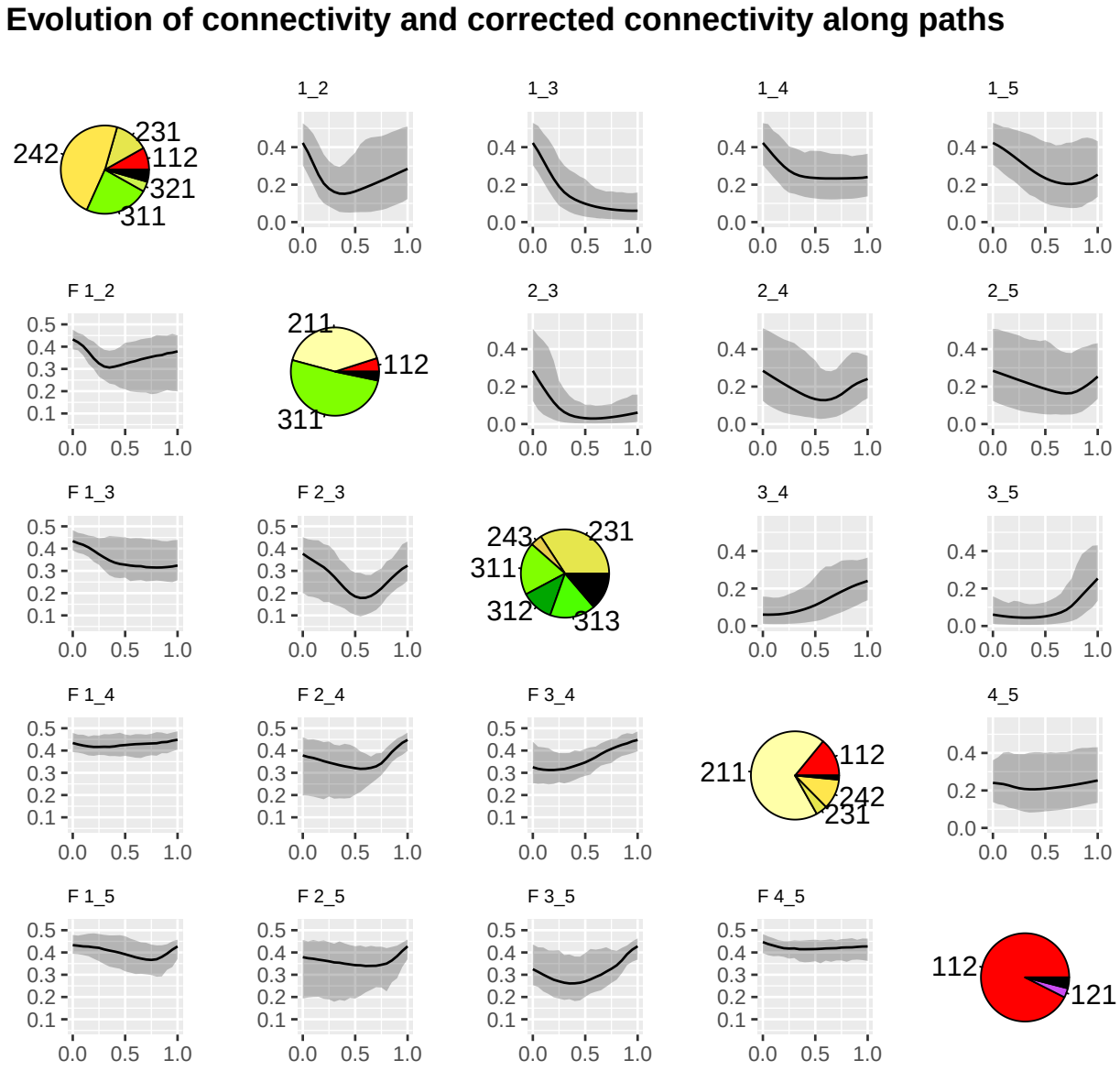

Figure 3: Evolution of network connectivity along the trajectories of landscape. The trajectory is simulated from typical landscape i to j (written as "i_j"). Above the diagonal, we can see the evolution of the predicted network connectivity without taking account of the observers' experience levels (uncorrected connectivity). Below the diagonal, the prediction takes into account the observers' experience levels (corrected connectivity, noted with a "F"). The simulated trajectories of the first row, and by symmetry the first column, are represented in fig. 2. The mean of the results is represented with a black curve, and 95% of the predictions are in the gray area.

B. Arroyo-Correa, I. Bartomeus, and P. Jordano. Individual-based plant–pollinator networks are structured by phenotypic and microsite plant traits. *Journal of Ecology*, 109(8):2832–2844, 2021. ISSN 1365-2745. doi: 10.1111/1365-2745.13694.

C. A. Cortina, J. L. Neff, and S. Jha. Historic and Contemporary Land Use Shape Plant-Pollinator Networks and Community Composition. *Frontiers in Ecology and Evolution*, 10, 2022. ISSN 2296-701X.

N. Deguines, R. Julliard, M. de Flores, and C. Fontaine. The Whereabouts of Flower Visitors: Contrasting Land-Use Preferences Revealed by a Country-Wide Survey Based on Citizen Science. *PLOS ONE*, 7(9):e45822, Sept. 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0045822.

N. Deguines, M. De Flores, G. Loïs, R. Julliard, and C. Fontaine. Fostering close encounters of the entomological kind. *Frontiers in Ecology and the Environment*, 16(4):202–203, May 2018. ISSN 15409295. doi: 10.1002/fee.1795.

M. Doré, C. Fontaine, and E. Thébault. Relative effects of anthropogenic pressures, climate, and sampling design on the structure of pollination networks at the global scale. *Global Change Biology*, 27(6):1266–1280, 2021. ISSN 1365-2486. doi: 10.1111/gcb.15474. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.15474.

B. Geslin, B. Gauzens, E. Thébault, and I. Dajoz. Plant Pollinator Networks along a Gradient of Urbanisation. *PLOS ONE*, 8(5):e63421, May 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0063421.

A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring Statistical Dependence with Hilbert-Schmidt Norms. In S. Jain, H. U. Simon, and E. Tomita, editors, *Algorithmic Learning Theory*, Lecture Notes in Computer Science, pages 63–77, Berlin, Heidelberg, 2005. Springer. ISBN 978-3-540-31696-1. doi: 10.1007/11564089_7.

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

F. Jiguet. Method learning caused a first-time observer effect in a newly started breeding bird survey. *Bird Study*, 56(2):253–258, July 2009. ISSN 0006-3657, 1944-6705. doi: 10.1080/00063650902791991.

T. N. Kipf and M. Welling. Variational graph auto-encoders. 2016. doi: 10.48550/ARXIV.1611.07308. URL https://arxiv.org/abs/1611.07308.

A. Rahimi and B. Recht. Random Features for Large-Scale Kernel Machines. In *Advances in Neural Information Processing Systems*, volume 20, Jan. 2007.

Q. Zhang, S. Filippi, A. Gretton, and D. Sejdinovic. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28(1):113–130, Jan. 2018. ISSN 1573-1375. doi: 10.1007/s11222-016-9721-7.