A JSDM with zero-inflation to increase the ecological relevance of analyses of species distribution data.

Jeanne Tous[1], Julien Chiquet[2], Amy E. Deacon[3], Ada Fontrodona-Eslava[4,6], Douglas F. Fraser[5], Anne E. Magurran[6]

[1]MIA 518, AgroParistech / INRAE, Universite Paris-Saclay, Palaiseau, France, jeanne.tous@inrae.fr - corresponding author.

[2]MIA 518, AgroParistech / INRAE, Universite Paris-Saclay, Palaiseau, France, julien.chiquet@inrae.fr.

[3]Department of Life Sciences, The University of the West Indies, St. Augustine, Trinidad and Tobago, amy.deacon@uwi.edu

[4]Cellular and Organismic Networks, Faculty of Biology, Ludwig-Maximilians-Universität München, Munich, Germany, ada.eslava@lmu.de

[5]Department of Biology, Siena College, Loudonville, NY 12211, USA fraser@siena.edu

[6]Centre for Biological Diversity, School of Biology, University of St. Andrews, St. Andrews, UK, aem1@st-andrews.ac.uk

Running headline: JSDM with zero-inflation - presentation & use case

Key words: Joint Species Distribution Models (JSDM) ; zero-inflation ; species association networks ; Poisson log-normal ; assemblage mechanisms ; Trinidad fish communities

Open research statement:
Data and code can be found in the public github repository
jeannetous/JSDM_with_zero_inflation_for_increased_ecological_relevance_in_species_distribution_analysis

Abstract

1. Deciphering the mechanisms that explain species distributions has been a long-term goal in ecology and is critical to predicting the responses of ecological communities to environmental change. To this end, Joint Species Distribution Models (JSDMs) assess the effects of abiotic factors on these communities and produce species association networks that describe statistical dependencies between species abundances. Species distribution data often contain numerous zeros. Not accounting for structural zeros (as opposed to zeros owing to species rarity or abiotic effects, modelled by JSDMs) can impair inferences of species association networks.

2. We propose a novel JSDM, the ZIPLN-network model, which directly models count data while including zero-inflation, thereby allowing it to account for structural zeros. Using simulated data, we compare the results obtained by this model to existing JSDMs in terms of association network inference from abundance data containing structural zeros. We illustrate the ZIPLN-network approach using data from tropical freshwater fish communities.

3. By integrating zero-inflation, our approach overcomes a previous shortcoming in analyses of species distribution data, and makes it possible to extract more of the ecological signal. A further advantage of this method is that it increases our understanding of ecological mechanisms that structure species distributions at the landscape scale ; this helps generate testable hypotheses. When the ZIPLN-network results are compared with those of its non-zero-inflated counterpart, a finer analysis of a species distribution dataset results.

4. The ZIPLN-network approach can be widely applied to species distribution datasets to produce more informed analyses and help distinguish different mechanisms so as to better understand community assembly rules. We provide guidance for getting started with our approach.

The data and code used for this paper can be found in the public github repository:

jeannetous/JSDM_with_zero_inflation_for_increased_ecological_relevance_in_species_distribution_analysis

Keywords: Joint Species Distribution Models (JSDM) ; zero-inflation ; species association networks ; Poisson log-normal ; assemblage mechanisms ; Trinidad fish communities

1. Introduction

Species distribution data describe the distribution of a set of species in a collection of sites, often based on counts of individuals at each site. Quantifying the patterns of these distributions, and deciphering the processes that underpin them has been a long-term goal in ecology [Connor and Simberloff, 1979, Diamond, 1975, Gotelli and Graves 1996, Ovaskainen et al. 2010]. Three main factors underlie these distributions [Guisans and Thuiller, 2005 ; Poggiato et al., 2021]: abiotic factors through environmental filtering (whether the environment is suitable for a given species [Cadotte and Tucker, 2017, Kraft et al., 2015]), biotic interactions (how species' presences influence one another) and dispersal (whether a species has been able to reach a site). Understanding how these mechanisms structure biodiversity patterns is critical for predicting ecosystem responses to environmental stressors and designing effective conservation strategies [Heinen et al. 2020]. However, the complicated ways in which these mechanisms operate, their intermingled influences, and the way their relative effects can depend on scale [Götzenberger et al. 2012, Viana and Chase 2019] make species distribution data both information rich and difficult to interpret.

Multiple approaches have been developed to explore these questions. One set of methods compares species distributions with those predicted by null models [Connor and Simberloff, 1979 ; Ulrich and Gotelli, 2010] to determine whether observed patterns can be attributed to chance. Species Distribution Models (SDM) [Elith and Leathwick, 2009 ; Miller, 2010] are a more recent approach. They model the relationship between the presence-absence or abundance of each species individually [Poggiato et al., 2021] to the environmental covariates, and potentially to biotic interactions [Poggiato et al., 2025, Staniczenko et al., 2017] or dispersal effects [Shipley et al., 2022]. They typically rely on Generalized Linear Models [Poggiato et al., 2021] but may also resort to Bayesian networks [Staniczenko et al., 2017] or machine-learning techniques [Burns et al., 2020].

Joint Species Distribution Models (JSDMs) [Chiquet et al., 2021 ; Pollock et al., 2014 ; Tikhonov et al., 2020] were later developed to simultaneously model all species in the distribution data, considering both the effects of environmental covariates and the residual correlations between species presences [Pollock et al., 2014] or abundances [Chiquet et al., 2021]. In the JSDMs framework, species distribution data is modeled by multivariate random variables following an appropriate distribution (e.g. Bernoulli for presence-absence, Poisson log-normal for species counts, log-normal for biomass). Additionally, residual correlations between species abundances / presences are often modeled by the covariance matrix of a latent variable [Chiquet et al., 2021 ; Pollock et al., 2014].

Together these residual correlations can be seen as a network, called a species association network. The network's nodes are the species and there is an edge between two nodes if and only if the species presences or abundances display a residual correlation. These associations are purely statistical relationships that may, for instance, be explained by biotic interactions, the influence of missing environmental covariates, the presence of species not accounted for in the data, or dispersal effects [Poggiato et al., 2021 ; Pollock et al., 2014 ; Tikhonov et al., 2020]. They can be used to disentangle the ecological mechanisms leading to species' co-occurrence patterns [Marjakangas et al., 2021 ; Ovaskainen et al., 2017 ; Pollock et al., 2014].

However, species distribution data often include numerous zeros that, under some circumstances, may hinder their study. Zeros in such data can be either false, random or structural [Blasco-Moreno et al., 2019 ; Martin et al., 2005]. False zeros result from errors in the experimental design, random zeros arise from sampling variability, and structural zeros are explained by ecological or evolutionary processes. Not properly accounting for these zeros in statistical models may result in impaired analyses [Blasco-Moreno et al., 2019]. While false zeros should be identified and removed when possible, specific methods are required for the other types of zeros [Blasco-Moreno et al., 2019]. Accounting for these zeros is a long-standing issue in the modelling of species abundances [Cunningham and Lindemayer, 2005 ; Nolan et al., 2021, Wenger and Freeman, 2008].

In the JSDMs frameworks, several strategies exist to account for the important presence of zeros in species distribution data. Zero-altered models, also known as hurdle models [Mullahy, 1986] separately model zero and non-zero counts by first fitting a presence-absence model and then fitting an abundance model conditionally on presence. This is the strategy used by a JSDM like HMSC to model extra zeros [Ovaskainen and Abrego, 2020]. A second strategy consists in directly modelling zero-inflation in an abundance model. Zero-inflation is defined as an excess of zeros compared to what a given distribution may produce [Barry and Welsh, 2002]. Zero-inflated models consist of a mixture of two distributions, a Dirac distribution in zero and a reference distribution to model abundance [Ovaskainen and Abrego, 2020]. In such a framework a zero in the abundance may arise from either the zero distribution or the abundance distribution. This is the strategy used for the zero-inflated Poisson log-normal model by Batardière et al. (2025). Brooks et al. (2017) implement both strategies in their **glmmTMB** R package, but they use univariate models and do not model inter-species statistical dependencies. While both approaches to model excess zeros are closely related [Ovaskainen and Abrego, 2020], zero-inflated models allow an integrated analysis and may seem more relevant when zeros may also arise from the abundance distribution [Blasco-Moreno et al., 2019] as is the case with a Poisson distribution. The SpiecEasi model [Kurtz et al., 2015], designed for the inference of microbial networks but that can be used as a JSDM, resorts to another strategy. While it does not directly account for zeros, it transforms abundance data into composition data and assumes that this allows the analysis of data following a zero-inflated negative-binomial distribution [Kurtz et al., 2015].

We introduce the ZIPLN-network (Zero-Inflated Poisson log-normal network) model as a JSDM that integrates zero-inflation. To our knowledge, there is no other JSDM designed for count data that can explicitly model zero-inflation. It is based on both the PLN-network [Chiquet et al., 2019] and the ZIPLN approaches [Batardière et al., 2024]. It is a generalised multivariate mixed graphical model that uses a Poisson distribution to directly model count data with a latent Gaussian layer to represent

the dependency structure corresponding to the association network (further explained in Section 2.1 below) and another latent variable to account for the zero-inflation in the data. It builds on the ZIPLN model and, as the PLN-network model does, adds an l_1-penalty on the latent variable precision matrix, in a LASSO-like approach [Friedman et al., 2008], to select statistically meaningful species associations.

We hypothesize that using a model that accounts for zero-inflation in species distribution data improves the inference of species association networks by avoiding the detection of spurious associations and refining the detection of true associations. To test this hypothesis, we simulate zero-inflated species distribution data and compare the quality of the association network inference by ZIPLN-network to that of PLN-network, its non zero-inflated counterpart and other reference methods, including SpiecEasi [Kurtz et al., 2015], said to be suited for zero-inflated data, Graphical-Lasso [Friedman et al., 2008], Neighborhood Selection [Meinshausen and Bühlmann, 2006] and SparCC [Friedman and Alm, 2012]. While the HMSC model can be included in a hurdle model, it does not model zero-inflation directly [Ovaskainen and Abrego, 2020] and is therefore not included in the comparison.

We further exemplify the use of the ZIPLN-network model with an application to a species distribution dataset of tropical freshwater fishes sampled in the Caribbean island of Trinidad. This dataset contains numerous zeros, including structural ones owing to natural barriers such as waterfalls. We show that modelling these zero-inflation patterns filters out numerous statistical associations and discuss how the resulting network can be interpreted.

Below, we first introduce the ZIPLN-network model and the methods we compare it to. We describe the simulation protocol used to test the hypothesis that modelling zero-inflation critically improves the inference of species association networks. We introduce the tropical fish dataset used to illustrate the

use of the ZIPLN-network model. We present the results of both the simulation study and the real data study before discussing the results and the ways the ZIPLN-network model can help improve the analysis of joint species distribution data.

2. Materials and methods

2.1 The ZIPLN-network model

In the model, n denotes the number of samples considered and p the number of species. Y_{ij} is the number of individual counts of species j found in sample i so that Y_i is a p-dimensional vector of integers. The Poisson log-normal (PLN) model [Aitchison and Ho, 1989 ; Chiquet et al., 2021] models observations Y using a latent variable Z with:

$$\text{Latent layer: } Z_i \sim \mathcal{N}(0, \Omega^{-1})$$

$$\text{Observed layer: } Y_{ij} \mid Z_{ij} \sim \mathcal{P}(exp(X_i^T B_j + Z_{ij}))$$

X_i denotes the covariates used for sample i , B_j the corresponding regression coefficients for species j. Just as in a regular generalised linear model, the $X_i^T B_j$ term accounts for abiotic effects on the abundances of species.

The latent variable Z allows the modelling of a statistical dependency structure between species. Indeed, Z_i follows a Gaussian distribution, with variance-covariance matrix Ω^{-1}. Ω is the precision matrix of the distribution, so that Z_{-j} and Z_{-k} (the $j - th$ and $k - th$ columns of Z) are conditionally independent given all other Z_{-l} if and only if $\Omega_{jk} = 0$. It ensues that Ω gives an association network between species. To infer a sparse network and avoid the detection of spurious associations, the PLN-network inference procedure [Chiquet et al, 2018] adds an ℓ_1 penalty on the off-diagonal values of Ω.

Selecting which penalty to apply can be done using different statistical criteria. In our study, we used the StARS method (Stability Approach to Regularisation Selection) [Liu et al., 2010]. This approach uses resampling to compute multiple association networks and keep only the associations that are present in a high enough proportion of the networks, a proportion that is fixed at 80 % here.

So far, we have described a model that assumes that all observations arise from a Poisson distribution (with an overdispersion arising from the latent variable). This means that any 0 observed in the data is explained by the effects contained in X or the correlations with other species through latent correlations. However, a particular species might be absent from a sample for structural reasons, not related to its rarity or abiotic effects, such as the existence of dispersal limits.

To account for such effects, we introduce the Zero-Inflated PLN-network (ZIPLN-network). Instead of assuming that the data are produced by a Poisson distribution, it assumes a mixture of a Dirac delta-distribution centered at zero to model zero-inflation and a Poisson distribution to model counts. To this end, it uses an additional binary latent variable W that follows a Bernoulli distribution. If $W_{ij} = 1$, Y_{ij} is automatically equal to 0 and if $W_{ij} = 0$, Y_{ij} is drawn from a Poisson distribution as with the PLN-network approach:

$$\text{Latent layer: } Z_i \sim \mathcal{N}(0, \Omega^{-1}), W_{ij} \sim B(\rho_{ij})$$

$$\text{Observed layer: } Y_{ij} | W_{ij}, Z_{ij} \sim \mathcal{P}(exp(X_i^T B_j + Z_{ij})) \text{ if } W_{ij} = 0$$

$$Y_{ij} | W_{ij}, Z_{ij} = 0 \text{ if } W_{ij} = 1$$

The parameters of W's distribution can depend on the species only ($\rho_{ij} = \rho_j$), the sample only($\rho_{ij} = \rho_i$), or the covariates, through a linear effect, with species-specific regression coefficient. In the latter case, we define $\rho_{ij} = logit^{-1}(X_i^0 B_j^0)$. ρ_{ij} then depends on both i and j but through covariates so that this does not lead to an overparametrization of the model. This means that it is possible to account for the

fact that some species are more likely to be absent from a sample than others, but also that the probability of having a zero for a given species can depend on specific covariates.

The ZIPLN-network model further adds an l_1-penalty on the off-diagonal value of precision matrix Ω, as the PLN-network model does.

The ZIPLN-network and other PLN models [Chiquet et al., 2021] are implemented in the publicly available R package PLNmodels [https://cran.r-project.org/package=PLNmodels], and in the python package pyPLN [Batardière et al., 2024]. The R implementation was used for this study.

In terms of notations, a PLN-network model run with covariates X_1 to X_d, with an intercept to analyse species distribution matrix Y is denoted PLN(Y ~ 1 + X_1 + ... + X_d). The "1" term corresponds to the intercept, meaning that the species mean abundance contains a species-specific term.

Similarly, a ZIPLN-network model run with covariates X_1 to X_d is denoted

- ZIPLN(Y ~ 1 + X_1 + ... + X_d | zi = species) for a species-dependent zero-inflation
- ZIPLN(Y ~ 1 + X_1 + ... + X_d | zi = site) for a site-dependent zero-inflation
- ZIPLN(Y ~ 1 + X_1 + ... + X_d | 1 + X_1^0 + ... $X_{d'}^0$) for a zero-inflation depending on an intercept and covariates X_1^0 to $X_{d'}^0$.

We provide guidance to get started with the ZIPLNnetwork model using R package **PLNmodels** in Appendix S3.

2.2 Reference methods

The performance of the ZIPLN-network model in terms of association network inference was compared to several other methods.

We first compared it to its non-zero-inflated counterpart, the PLN-network model [Chiquet et al., 2019]. In the case of a covariate-dependent zero-inflation, we tested the PLN-network model both

with and without including the zero-inflation covariate effect in the mean, this model is denoted PLN-ZIVAR (for PLN with zero-inflation covariate). Specifically, the ZIPLN-network model ZIPLN(Y ~ 1 + X_1+ ... + X_d | 1 + X_1^0+ ... X_d^0), is compared to both PLN(Y ~ 1 + X_1+ ... + X_d) and to PLN(Y ~ 1 + X_1+ ... + X_d + X_1^0 + ... X_d^0). We asked whether simply adding the ZIPLN-network zero-inflation covariates to the PLN-network mean effect, thereby allowing it to draw the Poisson distribution towards zero, would be as efficient as explicitly treating it as a zero-inflation effect.

We also compared our method to the SpiecEasi [Kurtz et al., 2015], Graphical-Lasso [Friedman et al., 2008], Neighborhood Selection [Meinshausen and Bühlmann, 2006] and SparCC [Friedman and Alm, 2012] methods.

SpiecEasi (SParse InversE Covariance Estimation for Ecological Association Inference) turns abundance data into compositional data before using a sparse graphical model inference framework. While it does not model zero-inflation, it turns count data into abundance data and is expected to be of use for data simulated under zero-inflated negative-binomial distribution [Kurtz et al., 2015].

SparCC (Sparse Correlations for Compositional data) proceeds similarly, log-transforming abundance data to turn it into compositional data before estimating linear Pearson correlations between variables [Friedman and Alm, 2012].

Graphical-Lasso [Friedman et al., 2008], also denoted GLASSO, is a reference l_1-penalized graphical inference method. It is used in both the (ZI)PLN-network approaches and the SpiecEasi method. We ran it directly on the residuals of a generalized linear model run on the log-transformed abundance matrix for comparison.

Finally, Neighborhood Selection [Meinshausen and Bühlmann, 2006] is another reference l_1 -penalized method. It consists in splitting the inference into several sub-tasks, using Lasso to infer the neighbourhood of each node in the graph. As for GLASSO, this method is run on the residuals of a generalized linear model run on the log-transformed abundance matrix.

2.3 Simulation protocol

We first compared the results of the ZIPLN-network model to those of the PLN-network model and the reference methods on simulated data.

We hypothesized that failing to account for the zero-inflation in abundance data impairs network inference. We further hypothesized that, the more important the zero-inflation, the more the network inference is impaired. To this end, we simulated abundance data with various types and strengths of zero-inflation and compared the networks inferred by each model.

Sparse networks were generated using an Erdös-Renyi model (no particular structure), a preferential attachment structure (edges are attributed progressively with a probability proportional to the number of edges each node is already involved in) or a community structure (in the Stochastic Block Models sense [Holland et al., 1983]). This allowed us to compare the results given different network structures. Using the **igraph** package [Csardi and Nepusz, 2006], we generated an adjacency matrix G corresponding to a given structure. Then a precision matrix Ω was created with the same sparsity pattern as G, as follows: $\Omega' = G \times v$, $\Omega = \Omega' + diag(|min(eig(\Omega))| + u)$ with $u, v > 0$ two scalars. A higher v implies stronger correlations whereas a higher u implies a better conditioning of Ω. Following [Chiquet et al., 2019], we fixed $v = 0.3$, $u = 0.4$ in the simulations.

We first simulated data under a Poisson log-normal model using the correlation network obtained from the method described above. To make the comparisons with reference methods more fair, we turned the abundances thereby obtained into proportions and resampled abundances from a multinomial distribution using these proportions. This already led to a number of random zeros in the abundance data, produced by the Poisson distribution. We ran simulations with $n = 25, 50, 100$ or 300 sites and $p = 25$ or 50 species. For reference, the dataset used for an illustration (see section

2.4) has $n = 304$ sites and $p = 20$ species. We used an intercept and a single continuous covariate (denoted X in the model) drawn between 0 and 1.

We then proceeded to add zero-inflation to the simulated data, emulating structural zeros. We tested all three types of zero-inflation: species-dependent, site-dependent and covariate-dependent. For each type, we tested four levels of zero-inflation. Reference zero-inflation levels for each zero-inflation types were designed to be similar to the zero-inflation levels observed in the illustration data we used (see section **2.4**), two of them were lower (respectively only about 10 % and 50 % of the reference level) and the last one was about 10 % more important. The covariate-dependent zero-inflation used a single categorical variable (to emulate similar patterns as those studied in the illustration data).

All the network inference methods used were penalized methods, meaning that a penalty is applied on the network to ensure it is sparse (the bigger the penalty, the sparser the network), different penalties being tested for each model. The different methods can therefore be compared in terms of AUC (Area Under the receiver operating characteristic Curve).

Code for the simulation study can be found in the public github repository

jeannetous/JSDM_with_zero_inflation_for_increased_ecological_relevance_in_species_distribution_analysis

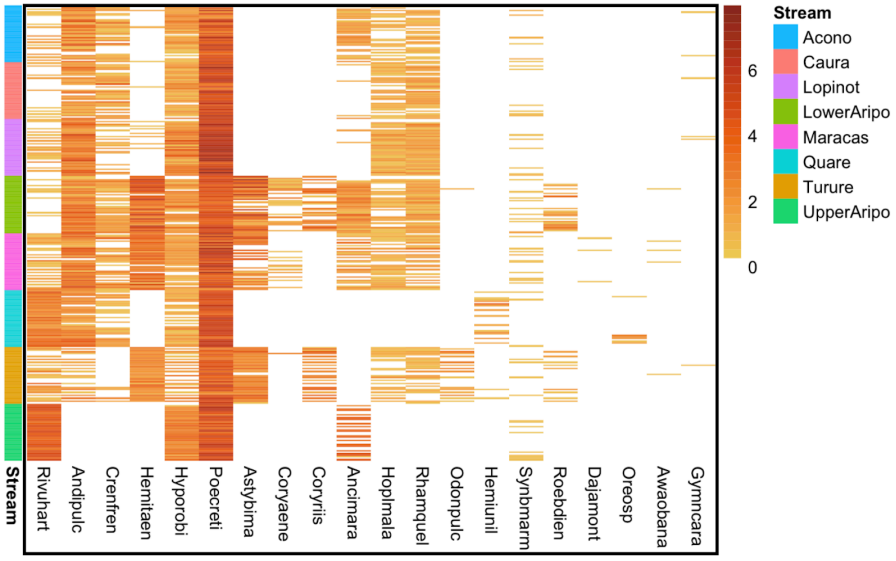
2.4 Tropical freshwater fish data

To illustrate how the ZIPLN-network approach can be used to study species distribution data, we used a freshwater fish survey from Trinidad. Trinidad is an island in the southern Caribbean, located 12 km from Venezuela. It is home to a rich freshwater fish fauna of both Antillean and South American origin that has been extensively investigated [Phillip et al., 2013].

The data set used here is described in [Magurran et al., 2018]. The original fieldwork was undertaken as part of ERC AdG BioTIME (250189), and had received ethical clearance from the ERC.

The survey was carried out in the Northern Range of Trinidad, a mountain system in the north of Trinidad where a series of parallel rivers are found. The fish were surveyed at 16 sites in eight Northern Range rivers over a 5-year period from December 2010 to August 2015. In each river, two sites were surveyed, one considered as “disturbed” due to human recreational activities, while the other was “undisturbed” [Magurran et al., 2018]. Each site was surveyed four times a year, at the beginning and end of both the dry (January to May) and wet (June to December) season [Deacon et al., 2015, Magurran et al., 2018]; the first sample was removed due to differences in sampling effort [Magurran et al., 2018]. Hence, 19 surveys are considered for each site. The dataset contains 20 species, after filtering out those that had been observed less than twice in the dataset. For the sake of readability, we use shortened versions of the species names in the figures; their complete names can be found in Appendix S1.

Figure 1: Species log-abundances (log(1 + species count), in the dataset, sorted by stream. Blank cells indicate species absences. Each column corresponds to a species, each row corresponds to a combination (site, time step). The rows are grouped by stream (there are two sites per stream).



As shown in Figure 1, the dataset contains numerous zeros. We assume that it does not contain false zeros, given the rigorous sampling protocol used [Magurran et al., 2018] (while false zeros may remain, there would be no way to identify them). Part of the zeros may arise from sampling variability, owing to the overall rarity of several species, like *Synbranchus marmoratus* or *Awaous banana*. These zeros are expected to be accounted for by the Poisson part of the ZIPLN-network method, through the mean effect and more specifically in the intercept for those species that tend to be rare everywhere. Figure 1 also shows that some species, while relatively abundant in many sites, appear not to be found at all in several streams. For example *Andinoacara pulcher* is completely absent from the upper Aripo stream, and *Ancistrus maracasae* is not found in the Quare and Turure streams. These zeros seem to be structural zeros as defined by Blasco-Moreno et al., 2019. They are likely owing to natural limits in the dispersal of the species. Indeed in the context of these data, natural barriers, like waterfalls, can prevent a species from reaching a given site even if the abiotic and biotic contexts could have suited it.

We hypothesized that accounting for the zero-inflation in the analysis of this data set will remove several potentially spurious associations. It may also help conduct more precise analysis as it can help separate different types of zeros. We also hypothesized that some associations owing to the presence of structural zeros may appear in the PLN-inferred networks but not in the ZIPLN-inferred ones.

2.5 Choice of covariates

To account for abiotic effects, we ran a principal component analysis (PCA) on a set of descriptive variables. The variables used were: width, depth and volume of the stream, altitude, water conductivity, dissolved oxygen concentration, pH, temperature, and turbidity, as well as several quantitative descriptors of the substrates found in the river and a visual estimate of the canopy cover. We only kept the first two principal components, denoted *PC1* and *PC2*; using more did not significantly improve or modify the results. *PC1* mostly corresponded to water descriptors (depth, volume, temperature) whereas *PC2* corresponded to descriptors of the substrates in the river (see

Appendix S2). Including covariates that represent these abiotic covariates avoids finding statistical associations owing to their joint effect on two species.

We additionally accounted for two categorical variables: disturbance (binary variable, i.e., whether the site was a priori labeled as disturbed) and river identity. Together, these two variables characterise each geographical site. Accounting for the stream ID in this way allowed us to include a dispersal-related effect in the model: if a species was never found in a particular stream or in nearby sites, the model adjusted the coefficient of B corresponding to that stream, without attributing such an effect to the other abiotic covariates.

For the PLN model, we assumed that all the covariates we included influence the abundances, whereas for the ZIPLN-network model, we included the stream effect in the zero-inflation part of the model only, and used PC1, PC2, and the disturbance variable for the abundance.

In the end, we studied the species association networks obtained from six different models (three PLN-network and three ZIPLN-network models):

PLN(Y ~ 1 + PC1 + PC2)

PLN(Y ~ 1 + PC1 + PC2 + disturbance)

PLN(Y ~ 1 + PC1 + PC2 + disturbance + stream)

ZIPLN(Y ~ 1 + PC1 + PC2 | zi = species)

ZIPLN(Y ~ 1 + PC1 + PC2 + disturbance | zi = species)

ZIPLN(Y~ 1 + PC1 + PC2 + disturbance | 1 + stream)

3. Results

3.1 Simulation study

Figures 2 and 3 show how the different methods perform in terms of AUC in various configurations. When there is only a low level of zero-inflation (corresponding to level 1 on Figure 2, roughly 10 % of the zero-inflation typically observed in the tropical freshwater fish dataset we use for illustration), all

the methods display similar performance. This is also the case when $n = p$ (Figure 3) with all methods displaying rather low AUC and only slightly better results for the ZIPLN-network model.

In all the other configurations, the ZIPLN-network model consistently outperforms all the reference methods, including the PLN-network approach, even when the zero-inflation covariate is included in the Poisson mean effect for a covariate-dependent zero-inflation (PLN-ZIVAR method on Figure 2). As the zero-inflation level increases (Figure 2) the AUC values for the different methods tend to reach lower minima (around 0.4) than with the ZIPLN-network model.

Figure 2.1 shows that the covariates-dependent zero-inflation appears to be the one for which the ZIPLN-network model obtains the best AUC, whereas it obtains its worst AUC for the species-dependent zero-inflation, with values comparable to those obtained with the reference methods when the zero-inflation becomes very important. The network structure however does not seem to strongly influence the AUCs obtained by the different methods (Figure 4.2).

Figure 4 shows an example of how a species association network is inferred from zero-inflated abundance data by the ZIPLN-network and PLN-network methods. It illustrates how not accounting for zero-inflation may impair network inference. Figure 4.C shows that the PLN-network model infers positive associations that are absent from the ground-truth network (Figure 4.A) between species 4, 5 and 6. We note that these species share a similar zero-inflation pattern (see column 4 to 6 on the heatmap of Figure 4.B). When the zero-inflation covariate is accounted for as a mean effect in the PLN-network model (Figure 4.D), only the association between species 4 and 5 remain, but numerous other spurious associations persist, that are filtered out by the ZIPLN-network model (Figure 4.E).

Figure 2: Violin plot of the AUCs obtained with the different network inference methods tested, as the zero-inflation type and level vary (top) level 3 being the reference level ; and as the network structure and zero-inflation type vary (bottom) for a zero-inflation fixed at the reference level (**ER:** Erdös-Rényi, **PA** : Preferential Attachment, **C :** Community). The AUCs obtained by the ZIPLN model are consistently better than those obtained with the other models, this difference is more striking when the zero-inflation is more important (top) and when it is covariates or sites-dependent (top and bottom) but the network structure does not have a major impact on it (bottom).

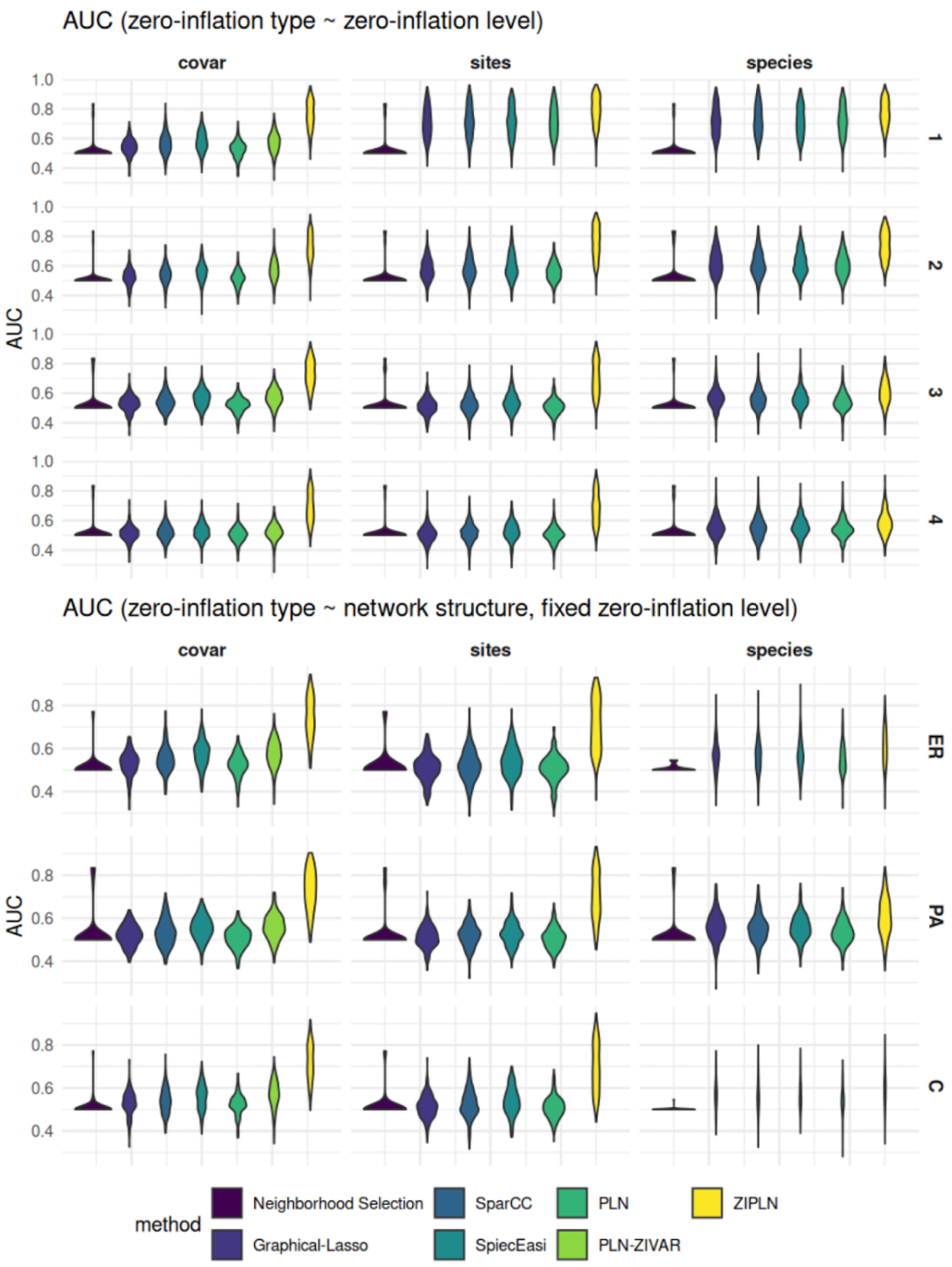


Figure 3: Violin plot of the AUCs obtained with the different network inference methods tested, as n and p vary for different zero-inflation types (top: covariate-dependent, middle: site-dependent, bottom: species-dependent) with the zero-inflation level fixed at the reference level (level 3 on Figure 2). The AUCs get worse when n is small compared to p, for all types of zero-inflation.

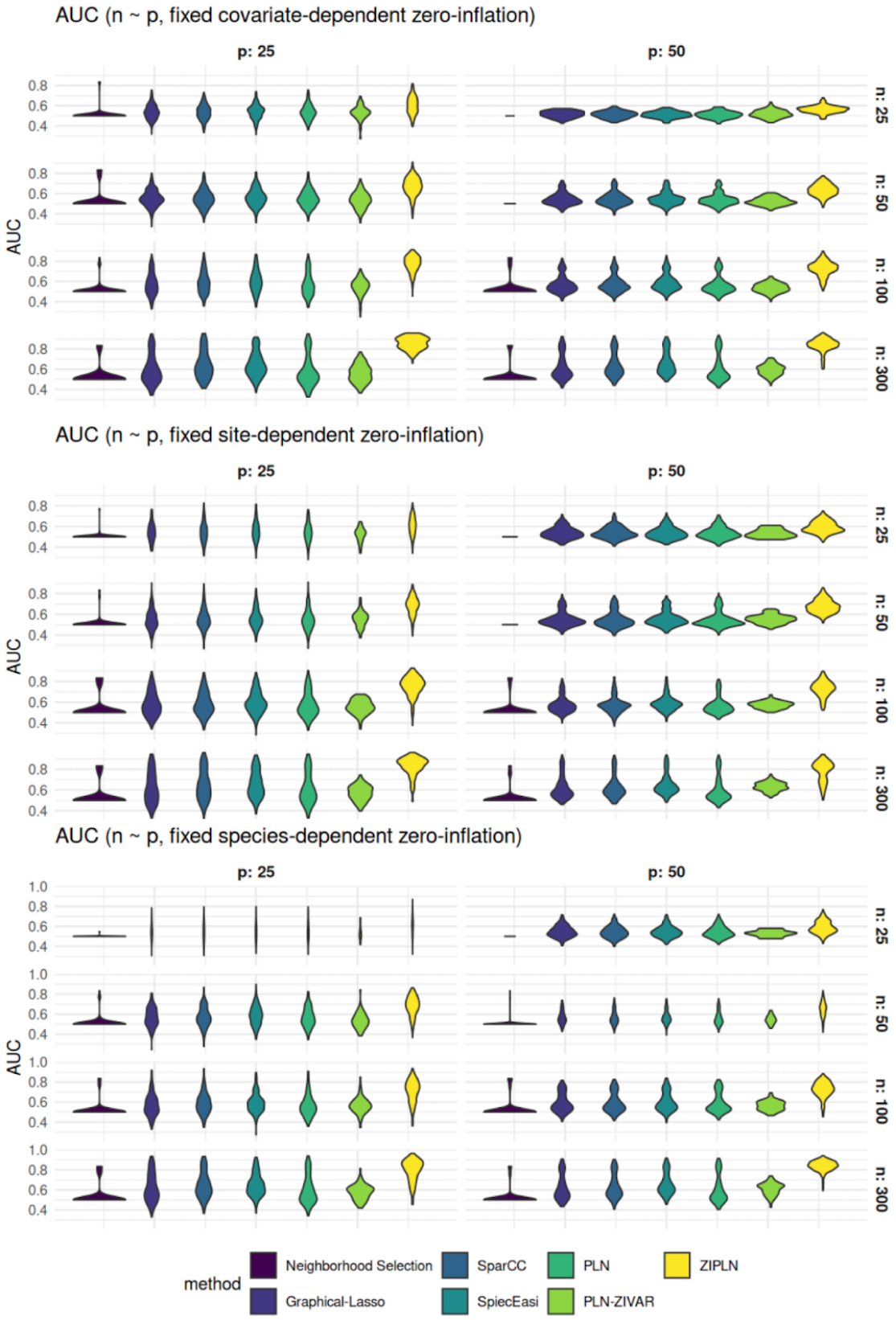
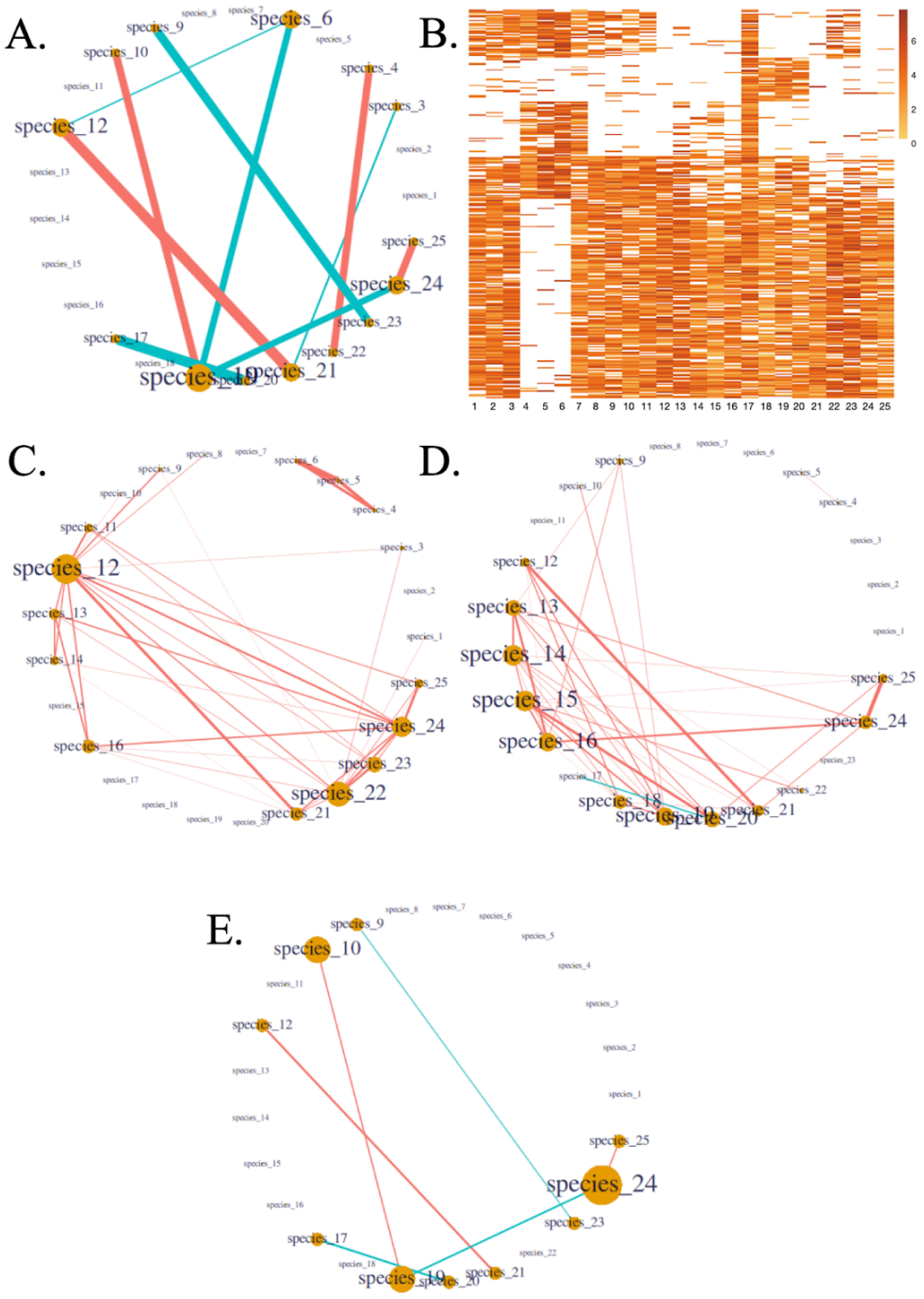


Figure 4: Illustrating how an association network is retrieved from zero-inflated abundance data with and without accounting for zero-inflation. **A.** Ground-truth network used for the simulation. **B.** Heatmap of the log-abundance generated for this simulation, with covariate-dependent zero-inflation, zero-inflation level being at the reference level (level 3 on Figure 2). **C.** Species association network inferred by the PLN-network model. **D.** Species association network inferred by the PLN-network model with the zero-inflation covariate included in the mean of the model. **E.** Species association network inferred by the ZIPLN-network model.



3.2 Tropical freshwater fish data analysis

We compare the association networks (Figure 7) inferred from the tropical freshwater fish data by the PLN-network and ZIPLN-network models with different combinations of covariates (see section 2.5). For a finer interpretation, we include in the analysis the coefficients describing the effects of abiotic covariates on each species that the models infer (Figure 5), corresponding to matrix B in the model's notations (section 2.1) and the zero-inflation probabilities inferred by the ZIPLN-network model (Figure 6). Both give important elements of context to better analyse the species association networks.

Figure 5 shows that the effects of the environment-related principal components display little variance, with a stronger intercept effect, whereas Figure 5.3 shows an important variance in the stream effect (for the ZIPLN-network model, the stream effect is only included in the zero-inflation coefficients, see Figure 6).

The ZIPLN-network model infers a zero-inflation probability for each pair $(species, sample)$ depending on the covariates used in the model. Here, for the first two ZIPLN-network models, this probability is only species-dependent whereas for the third one it also depends on the stream considered. Figure 6 shows the resulting probabilities. We see that when the probabilities are made stream-dependent, for a given species, they can vary a lot from one stream to another.

Figure 5: Regression coefficients for abiotic effects on the abundances (log-scale).

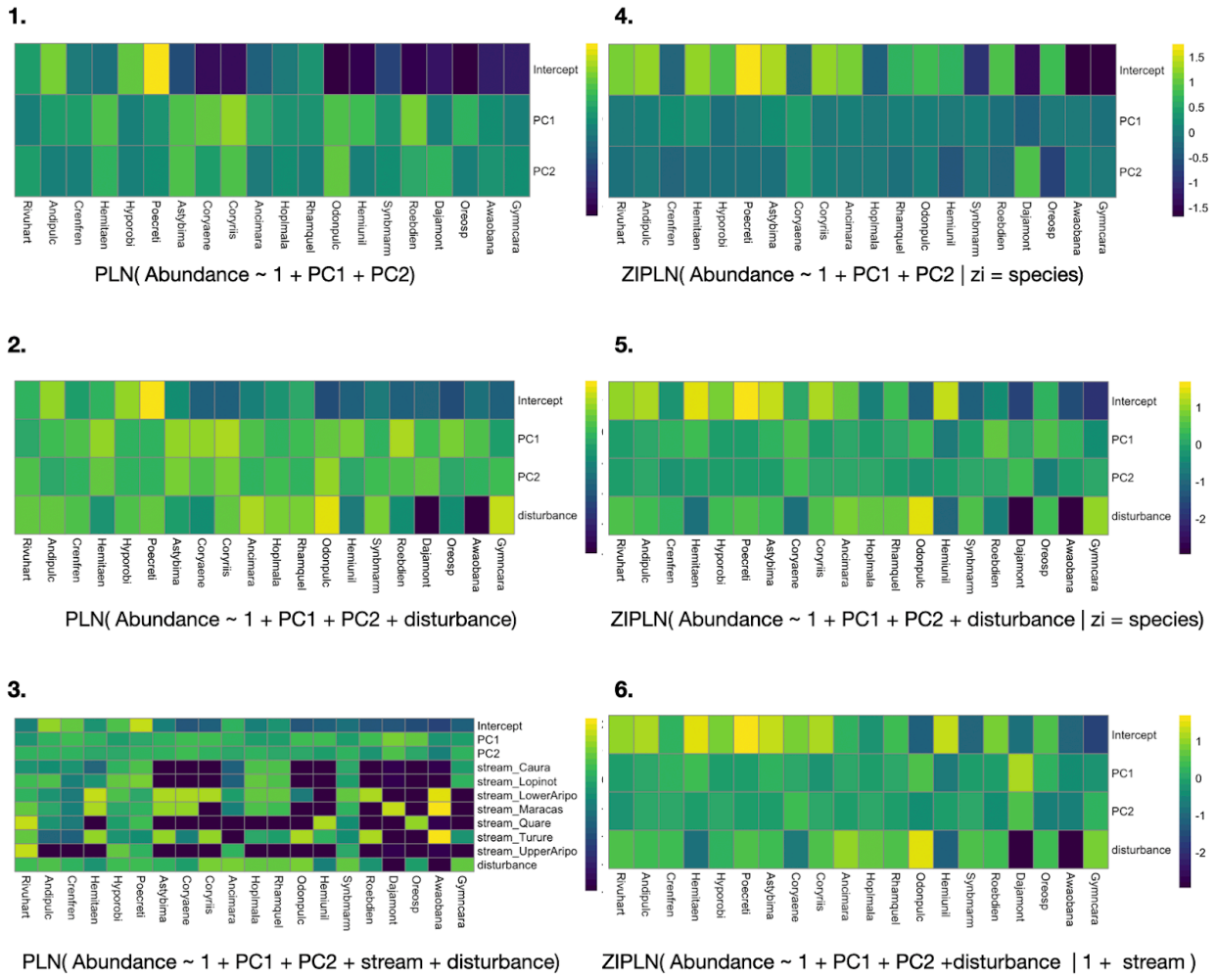


Figure 6: zero-inflation probabilities inferred by the ZIPLN-network approach with the different combinations of covariates.

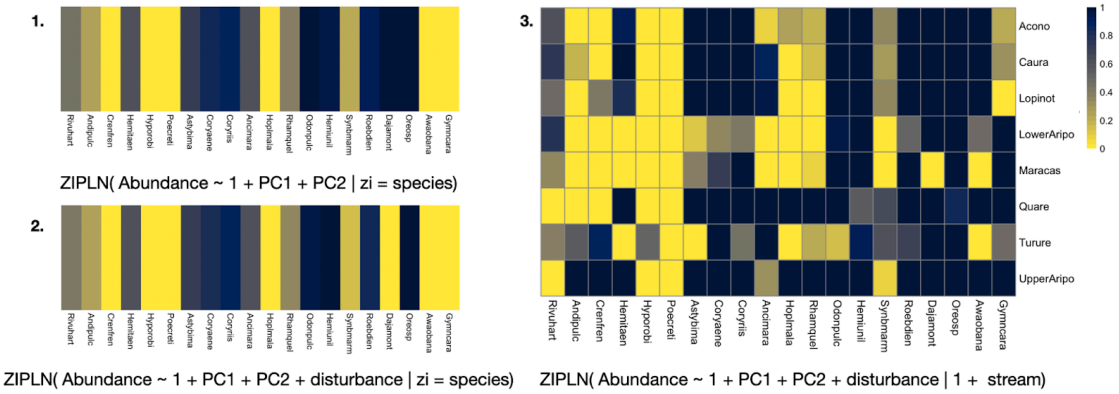


Figure 7: Association networks obtained with PLN-network and ZIPLN-network models for different combinations of covariates. Blue and pink edges represent negative and positive associations, respectively. The thicker the edge, the stronger the association. The size of each species' name is proportional to the sum of absolute values of associations it is involved in. License details for the species silhouettes are given in the references.

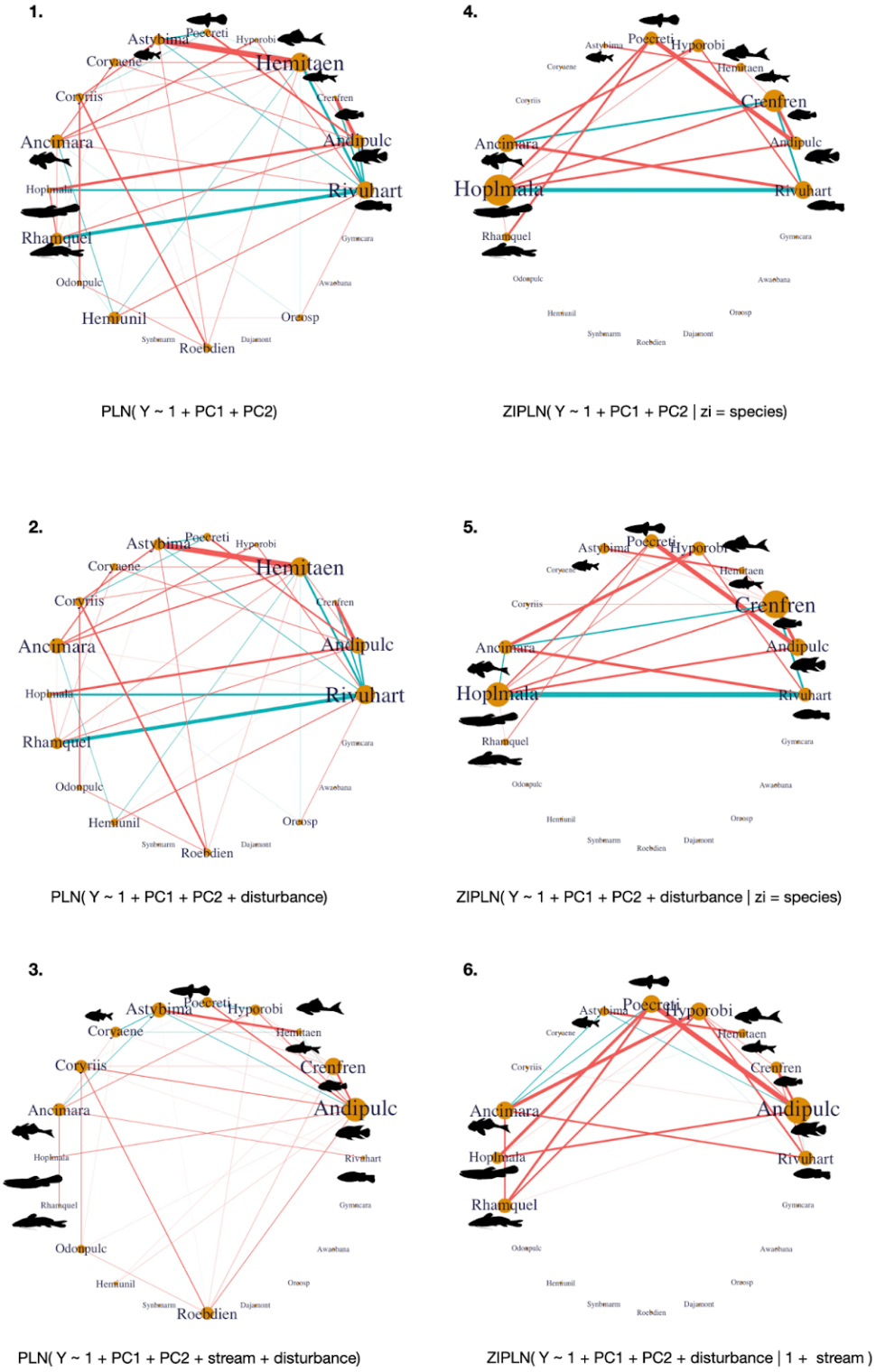


Figure 7 shows the association networks inferred by the PLN-network and ZIPLN-network models using various combinations of covariates. Accounting for a zero-inflation effect significantly modifies the association networks by removing numerous weak associations. For both models, adding a disturbance effect to the abundance only marginally modifies the network. However, accounting for a stream effect, be it in the abundance for the PLN model or in the zero-inflation for the ZIPLN model removes a certain number of associations. When the stream is accounted for, the PLN and ZIPLN association networks are quite similar, but the associations in the former appear to be much weaker than in the latter. Some associations are only present in the non-zero-inflated model, such as the positive *Roeboides dientonito* - *Corynopoma riisei* association, the negative associations of *Poecilia reticulata* with both *Astyanax bimaculatus* and *Hypostomus robinii*. Other associations are only visible in the ZIPLN network, such as the positive associations of *Poecilia reticulata* with *Hoplias malabaricus* and *Rhamdia quelen* or the negative one with *Ancistrus maracasae*.

Several associations of interest can also be noted. We find a strong negative association between *Hoplias malabaricus* and *Rivulus hartii* (also known as *Anablepsoides hartii*), present with both approaches, which disappears when the stream variable is taken into account. Positive associations between *Astyanax bimaculatus* and *Hemibrycon taenerius* on the one hand and *Poecilia reticulata* and *Andinoacara pulcher* on the other hand persist in all networks.

4. Discussion

We have introduced the new ZIPLN-network approach to analyse species distribution data. To our knowledge there exists no other JSDM that models count data while explicitly accounting for zero-inflation. While zeros in species distribution data can have different origins, this approach can prove particularly useful to model structural zeros owing to ecological processes, like the dispersal limits that can be observed in the freshwater fish species dataset we use as an illustration.

In this section we first analyse how the ZIPLN-network model performs compared to reference methods on simulated data and discuss whether it improves the inference of association networks from abundance data as we hypothesized.

Then, we analyse the results obtained on the illustration dataset. We discuss whether the ZIPLN-network model can help remove spurious associations in real data too. We consider how the model can be used to build ecological hypotheses regarding the mechanisms underlying the observed distribution. We then discuss the choice of covariates and the consequences of this choice for the interpretation of the model results. Finally, we provide recommendations for using the ZIPLN-network approach to analyse species distribution data.

4.1 Simulations show that the ZIPLN-network model improves the inference of species associations networks in data that contains structural zeros

In the simulation study, zeros are added to simulated abundance data, following either species-dependent, site-dependent or covariate-dependent zero-inflation patterns. Such zeros are added independently of a species' rarity when present, thereby emulating structural zeros [Blasco-Moreno et al., 2019].

Figures 2 and 3 show that the ZIPLN-network model consistently outperforms all the other reference models tested for network inference in terms of AUC, even when the zero-inflation covariate is included in the mean effect (PLN-ZIVAR approach) or when using SpiecEasi, a model said to be suited for zero-inflated data [Kurtz et al., 2015]. As the zero-inflation becomes more important, the difference between the ZIPLN-network model and the reference models becomes more striking. Logically, it also appears that a low zero-inflation makes this difference less striking and that increasing the number of sites n for a fixed number of species p makes the network inference task easier for all the methods, with a low n value (compared to p) making it really hard.

Figure 4 helps understand why explicitly modeling structural zeros through zero-inflation is critical in avoiding the detection of spurious associations. Indeed, it shows that shared zero-inflation patterns may be falsely interpreted as statistical correlations between species abundances. While including the zero-inflation covariate in the PLN-network model seems to mitigate this effect in the example displayed on Figure 4, Figures 2 and 3 show that this approach performs slightly worse than the regular PLN-network model in general. This is likely owing to the fact that including this effect in the mean implies the zero-inflation covariate to be considered as also having an effect on the abundance of a given species even when present, whereas the ZIPLN-network model clearly distinguishes these two aspects.

These results confirm our hypotheses that modeling zero-inflation is critical for correctly inferring species association networks in species abundance data with structural zeros and that it becomes more critical as the zero-inflation gets more important.

4.2 The ZIPLN-network model filters out associations in the freshwater fish distribution data analysis

Similarly to what is observed in the simulation study, Figure 7 shows that modelling zero-inflation filters out a number of associations, especially weak ones. In that case however, one cannot directly tell whether the removed associations were spurious ones or not. Analysing the association networks in light of the zero-inflation probabilities can help answer this question.

For instance, the ZIPLN-network model filters out the positive association between *Roeboides dientonito* (hunchback sardine) and *Corynopoma riisei* (swordtail sardine) (Figure 7). Figure 6 shows that these species share similar zero-inflation patterns, explaining the presence of a positive association in the PLN-inferred network and its absence in the ZIPLN-inferred networks. Therefore it could be that the species are simply influenced by the same dispersal limits. However, as opposed to

the simulation analysis, one cannot rule out that the two species' zero-inflation patterns are not independent. Indeed, in general they may also be explained by the fact that the species need one another to be present at detectable levels. The model alone cannot answer that question in the absence of ecological knowledge of the two species.

This confirms the hypothesis we made that modelling zero-inflation removes multiple associations inferred from species distribution data. The joint analysis of the zero-inflation probabilities and the association networks inferred by the model allows the user to build more precise hypotheses about the reasons why a given association is absent in the ZIPLN-network analysis, with the caveat that the underlying ecological explanation for the result cannot be deduced from the model results alone.

4.3 Interpreting the ZIPLN-network inferred associations

While species association networks may reflect various explanations like biotic interactions or the influence of missing abiotic covariates [Pollock et al., 2014 ; Tikhonov et al., 2020] there is no general correspondence between a given ecological mechanism and a corresponding statistical association [Poggiato et al., 2021]. One can however build on existing ecological knowledge to make hypotheses about the ecological mechanisms behind specific associations. Here, we study the association network inferred with the ZIPLN-network model using *PC1*, *PC2* (the first two principal components off a PCA run on the covariates that describe each site) and the disturbance level as covariates for abundance, with the stream as a zero-inflation covariate (Figure 7.6).

Several positive associations are found between putative prey and predator species. For instance, *Poecilia reticulata* (guppies) are positively associated with *Andinoacara pulcher* (blue coscorob) and *Hoplias malabaricus* (guabine). Several hypotheses could explain this situation. As guabines are primarily nocturnal predators [Fraser et al., 2006], their coexistence with guppies could reflect temporal niche sharing [Fraser et al., 2004]. Additionally, the behavioural response of guppies to the

threat of predation is less in the form of habitat shifts and more likely to be shoaling and avoidance during the most vulnerable crepuscular periods [Fraser et al., 2004]. Guppies and blue coscorobs co-occur in many stream habitats in the Northern Range; the latter is an invertivore, rather than a piscivore, as was sometimes assumed in the earlier literature.

The positive association between *Astyanax bimaculatus* (two-spot sardine) and *Hemibrycon taenerius* (mountain stream sardine) could be explained by the fact that these species can benefit from adaptive advantages due to heterospecific schooling behaviour [Krause and Ruxton, 2002]. Interestingly, while there is a strong association between these species, there are no others within the group of shoaling characids (locally referred to as 'sardines') that also includes *Corynopoma riisei* (swordtail sardine) and *Roeboides dientonito* (hunchback sardine). It would be interesting to further explore this pattern to study whether it is explained by different schooling preferences, or by other mechanisms.

However, although our analysis uncovers statistically robust associations, not all of these necessarily correspond to ecological interactions, nor have all potential interactions been identified. For instance, the predator-prey interaction between *Crenicichla frenata* (pike cichlids) and guppies is not detected in the network, although a large literature documents the evolutionary interactions between these species. An explanation for this could be that behavioural (e.g., shoaling [Magurran and Seghers 1991, Seghers, 1974]) and morphological (e.g., color patterns [Endler 1987]) adaptations in guppy populations, which covary with predation risk, enable coexistence in localities where the two species naturally co-occur.

There were also examples of unexpected associations in the network. In some cases, mechanisms other than interactions, such as similar habitat preference, could be responsible. For example, *Rhamdia quelen* (silver catfish) and *Hypostomus robinii* (teta) (significant positive association) are both benthic dwellers (they live near river beds) [Phillip and Ramnarine, 2001] and may prefer sites

with overhangs along the banks. Pike cichlids and blue coscorobs (another positive association) both tend to prefer clearer and deeper waters [Phillip and Ramnarine, 2001].

In the network, a third of the species do not show any association. This might be partly a statistical artefact, a phenomenon that can be explained by the small populations and limited distributions of many species. This means that there is less signal for these species, which makes it harder to identify patterns in their distributions that could reveal underlying mechanisms. Some species may also be very localised and therefore have little chance of displaying statistical correlations with more widespread species. This relates to the difficulties that SDMs and JSDMs can have more generally in modelling rare species [Pollock et al., 2014 ; Zhang et al., 2019].

Thus, we have shown how, in this example, the ZIPLN-network analysis allows one to propose several hypotheses regarding the biotic and abiotic mechanisms that shape community structure. It can also be used to understand how pre-identified biotic interactions show in species distribution data.

4.4 Influence of covariates

How does the choice of covariates influence the results? In JSDMs, covariates are generally used to account for abiotic effects. In our case, including streams as covariates also makes it possible to add a dispersal effect since the model then accounts for the fact that some species may never be found in a given stream.

As expected, Figures 5 and 7 demonstrate that changing the model covariates means that the signal is captured differently. For example, Figures 5.2 and 5.3 show that the variance of the regression coefficients variance for *PC1* and *PC2* decreases when adding a stream covariate (by 30 % for PC1 and 34 % for PC2). Thus, in this system, the signal is clearer when the stream effect is accounted for, than with just the abiotic parameters.

Similarly, Figure 7 shows that the association networks change when the set of covariates is modified. However, most of the strongest associations seem robust to the addition of new covariates. This stability confirms that the PLN and ZIPLN models are robust; it supports the fact that these associations correspond to strong patterns in the observations.

One striking change is the disappearance of the strong negative association between the apex predator guabine and its prey *Rivulus hartii* (jumping guabine) when a stream covariate is added. This association probably reflects strong predator-avoidance behaviour. Indeed, Fraser and Gilliam (1992) analysed the predator-prey relationship between both species and found that the presence of guabines suppressed jumping guabines egg production by 50 %. Jumping guabines have excellent jumping abilities (hence their common name) and can survive out of water for extended periods, owing to cutaneous breathing [Fraser, Pers. obs.]. They can easily disperse overland on wet nights to colonise remote bodies of water free of piscivorous fish [Fraser et al. 1995]. These strategies allow them to concentrate at sites that are inaccessible to piscivorous fish. A hypothesis that explains the absence of this association when accounting for a stream effect is the fact that the predator-prey dynamics induce a concentration of jumping guabines in sites where guabines cannot be found, a pattern that is caught in the zero-inflation, stream-related part of the model. Looking back at the abundance heatmap (Figure 1), it indeed appears that jumping guabines seem to thrive more at sites where the guabines are completely absent. This example shows that the same effect can be represented in different ways by the model, depending on which covariates are included.

Therefore, while statistical criteria can be used to define a “best” choice of covariates, ecological knowledge plays a crucial role in the interpretation of the resulting networks. Taking into account both abiotic effects and association networks strengthens ecological understanding and helps formulate ecologically relevant hypotheses to explain distribution patterns.

5. Conclusion

We conclude that the ZIPLN-network model is a useful tool for analysing species distribution data as it allows the modelling of structural zeros with zero-inflation, along with species abundances, with the Poisson distribution modelling discrete counts. Our simulation study proves that accounting for the presence of structural zeros is critical to correctly infer species association networks.

We have also shown that comparing the results obtained with the PLN-network and ZIPLN-network models helps better disentangle the assemblage mechanisms at stake. Finally, it can help distinguish dispersal effects from other mechanisms by accounting for the absence of species at sites they did not reach.

Finally, we have demonstrated how the ZIPLN-network approach can be used to produce more informed analyses of species distribution data. Accounting for zero-inflation and using both presence-absence and count data can filter out spurious associations and help distinguish different mechanisms. We argue that such a model is better used in combination with its non-zero-inflated counterpart so as to better identify putative ecological mechanisms. Similarly, comparing the results obtained with different combinations of covariates allows for a more nuanced analysis.

Acknowledgements:
This work was supported by public grants from the Fondation Mathématique Jacques Hadamard, AgroParisTech, the DESSE INRAE and the Erasmus+ program. We thank Stéphane Robin and Wilfried Thuiller for the fruitful discussions about this work.

Author contributions:
JT led the investigation
JT, JC & AEM designed the study
JC designed the ZIPLN-network model and provided technical expertise on its use.
JT ran the analyses
AEM, AED, A F-E & DFF provided ecological expertise to support the interpretation of the results
JT wrote the initial draft
All authors reviewed and approved the final manuscript

Conflicts of interest:
The authors declare that they have no conflict of interest.

References

Aitchison, J., & Ho, C. H. (1989). The multivariate Poisson-log normal distribution. Biometrika, 76(4), 643-653.

Batardière, B., Chiquet, J., Gindraud, F., & Mariadassou, M. (2025). Zero-inflation in the multivariate poisson lognormal family: B. Batardière et al. Statistics and Computing, 35(6), 196.

Batardiere, B., Kwon, J., & Chiquet, J. (2024). pyPLNmodels: A Python package to analyze multivariate high-dimensional count data. Journal of Open Source Software, 9(104), 6969.

Blasco-Moreno, A., Pérez-Casany, M., Puig, P., Morante, M., & Castells, E. (2019). What does a zero mean? Understanding false, random and structural zeros in ecology. Methods in Ecology and Evolution, 10(7), 949-959.

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

Botham, M. S., Kerfoot, C. J., Louca, V., & Krause, J. (2006). The effects of different predator species on antipredator behavior in the Trinidadian guppy, Poecilia reticulata. Naturwissenschaften, 93(9), 431-439.

Brooks, M. E., Kristensen, K., Van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., ... & Bolker, B. M. (2017). Modeling zero-inflated count data with glmmTMB. BioRxiv, 132753.

Burns, P. A., Clemann, N., & White, M. (2020). Testing the utility of species distribution modelling using Random Forests for a species in decline. Austral Ecology, 45(6), 706-716.

Cadotte, M. W., & Tucker, C. M. (2017). Should environmental filtering be abandoned?. Trends in ecology & evolution, 32(6), 429-437.

Chiquet, J., Robin, S., & Mariadassou, M. (2019, May). Variational inference for sparse network reconstruction from count data. In International Conference on Machine Learning (pp. 1162-1171). PMLR.

Chiquet, J., Mariadassou, M., & Robin, S. (2021). The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances. Frontiers in Ecology and Evolution, 9, 588292.

Connor, E. F., & Simberloff, D. (1979). The assembly of species communities: chance or competition?. Ecology, 60(6), 1132-1140.

Csardi, G., & Nepusz, T. (2006). The igraph software. Complex syst, 1695, 1-9 <https://igraph.org>.

Cunningham, R. B., & Lindenmayer, D. B. (2005). Modeling count data of rare species: some statistical issues. Ecology, 86(5), 1135-1142.

D'Amen, M., Mod, H. K., Gotelli, N. J., & Guisan, A. (2018). Disentangling biotic interactions, environmental filters, and dispersal limitation as drivers of species co-occurrence. Ecography, 41(8), 1233-1244.

Deacon, A. E., Jones, F. A., & Magurran, A. E. (2018). Gradients in predation risk in a tropical river system. Current Zoology, 64(2), 213-221.

Deacon, A. E., Shimadzu, H., Dornelas, M., Ramnarine, I. W., & Magurran, A. E. (2015). From species to communities: the signature of recreational use on a tropical river ecosystem. Ecology and Evolution, 5(23), 5561-5572.

Diamond, J. M. (1975). Assembly of species communities. Ecology and evolution of communities, Harvard Univ. Press, pp. 342-444.

Elith, J., & Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. Annual review of ecology, evolution, and systematics, 40(1), 677-697.

Fox, J. Mathematical constraints in ecology and evolution, part 1: species can't all covary negatively (updated).
https://dynamicecology.wordpress.com/2016/10/05/mathematical-constraints-in-ecology-part-1-species-cant-all-covary-really-negatively/,2016. Accessed: 2024-11-13.

Fraser, D. F., & Gilliam, J. F. (1992). Nonlethal impacts of predator invasion: facultative suppression of growth and reproduction. Ecology, 73(3), 959-970.

Fraser, D. F., Gilliam, J. F., & Yip-Hoi, T. (1995). Predation as an agent of population fragmentation in a tropical watershed. Ecology, 76(5), 1461-1472.

Fraser, D. F., Gilliam, J. F., Akkara, J. T., Albanese, B. W., & Snider, S. B. (2004). Night feeding by guppies under predator release: effects on growth and daytime courtship. Ecology, 85(2), 312-319.

Fraser, D. F., Gilliam, J. F., Albanese, B. W., & Snider, S. B. (2006). Effects of temporal patterning of predation threat on movement of a stream fish: evaluating an intermediate threat hypothesis. Environmental Biology of Fishes, 76(1), 25-35.

Fraser, D. F., & Lamphere, B. A. (2013). Experimental evaluation of predation as a facilitator of invasion success in a stream fish. Ecology, 94(3), 640-649.

Friedman, J., & Alm, E. J. (2012). Inferring correlation networks from genomic survey data.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 9(3), 432-441.

Gilliam, J. F., Fraser, D. F., & Alkins-Koo, M. (1993). Structure of a tropical stream fish community: a role for biotic interactions. Ecology, 74(6), 1856-1870.

Goodale, E., Beauchamp, G., Magrath, R. D., Nieh, J. C., & Ruxton, G. D. (2010). Interspecific information transfer influences animal community structure. Trends in ecology & evolution, 25(6), 354-361.

Götzenberger, L., de Bello, F., Bråthen, K. A., Davison, J., Dubuis, A., Guisan, A., ... & Zobel, M. (2012). Ecological assembly rules in plant communities—approaches, patterns and prospects. Biological reviews, 87(1), 111-127.

Gotelli, N. J. and Graves, G. R. Null models in ecology. – Smithsonian Inst., Washington, 1996.

Guisan, A., & Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. Ecology letters, 8(9), 993-1009.

Haskins, C. P., Haskins, E. F., McLaughlin, J. J. A., & Hewitt, R. E. (1961). Polymorphism and population structure in Lebistes reticulatus, an ecological study. Vertebrate Speciation. Vertebrate Speciation. University of Texas Press, Austin, 320-395.

Heinen, J. H., Rahbek, C., & Borregaard, M. K. (2020). Conservation of species interactions to achieve self-sustaining ecosystems. Ecography, 43(11), 1603-1611.

Holland, P. W., Laskey, K. B., & Leinhardt, S. (1983). Stochastic blockmodels: First steps. Social networks, 5(2), 109-137.

Johansson, J., Turesson, H., & Persson, A. (2004). Active selection for large guppies, Poecilia reticulata, by the pike cichlid, Crenicichla saxatilis. Oikos, 105(3), 595-605.

Jones, F. A., Rutherford, M. G., Deacon, A. E., Phillip, D. A., & Magurran, A. E. (2019). Quantifying regional biodiversity in the tropics: A case study of freshwater fish in Trinidad and Tobago. Biotropica, 51(5), 700-708.

Kraft, N. J., Adler, P. B., Godoy, O., James, E. C., Fuller, S., & Levine, J. M. (2015). Community assembly, coexistence and the environmental filtering metaphor. Functional ecology, 29(5), 592-599.

Jens Krause and P. Krause. Living in Groups, 2002.

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., & Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. PLoS computational biology, 11(5), e1004226.

Lau, M. K., Borrett, S. R., Baiser, B., Gotelli, N. J., & Ellison, A. M. (2017). Ecological network metrics: opportunities for synthesis. Ecosphere, 8(8), e01900.

Lauritzen, S. L. (1996). Graphical models (Vol. 17). Clarendon Press.

Liu, H., Roeder, K., & Wasserman, L. (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. Advances in neural information processing systems, 23.

Magurran, A.E. (2005). Evolutionary ecology: the Trinidadian guppy. Oxford University Press, USA

Magurran, A. E., Deacon, A. E., Moyes, F., Shimadzu, H., Dornelas, M., Phillip, D. A., & Ramnarine, I. W. (2018). Divergent biodiversity change within ecosystems. Proceedings of the National Academy of Sciences, 115(8), 1843-1847.

Marjakangas, E. L., Ovaskainen, O., Abrego, N., Grøtan, V., de Oliveira, A. A., Prado, P. I., & de Lima, R. A. (2021). Co-occurrences of tropical trees in eastern South America: disentangling abiotic and biotic forces. Plant Ecology, 222(7), 791-806.

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., ... & Possingham, H. P. (2005). Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecology letters, 8(11), 1235-1246.

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso.

Miller, J. (2010). Species distribution modeling. Geography Compass, 4(6), 490-509.

Momal, R., Robin, S., & Ambroise, C. (2021). Accounting for missing actors in interaction network inference from abundance data. Journal of the Royal Statistical Society Series C: Applied Statistics, 70(5), 1230-1258.

Mullahy, J. (1986). Specification and testing of some modified count data models. Journal of econometrics, 33(3), 341-365.

Nolan, V., Gilbert, F., & Reader, T. (2022). Solving sampling bias problems in presence–absence or presence-only species data using zero-inflated models. Journal of Biogeography, 49(1), 215-232.

O'Hara, R., & Kotze, J. (2010). Do not log-transform count data. Nature Precedings, 1-1.

Ovaskainen, O., Hottola, J., & Siitonen, J. (2010). Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. Ecology, 91(9), 2514-2521.

Ovaskainen, O., & Abrego, N. (2020). Joint species distribution modelling: With applications in R. Cambridge University Press.

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., ... & Abrego, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. Ecology letters, 20(5), 561-576.

Phillip, D. A. T., & Ramnarine, I. W. (2001). An illustrated guide to the freshwater fishes of Trinidad and Tobago. University of the West Indies, St. Augustine, Trinidad and Tobago.

Phillip, D. A. T. (1998). Biodiversity of freshwater fishes of Trinidad and Tobago, West Indies (Doctoral dissertation, University of St Andrews).

Phillip, D. A., Taphorn, D. C., Holm, E., Gilliam, J. F., Lamphere, B. A., & Lopez-Fernandez, H. (2013). Annotated list and key to the stream fishes of Trinidad & Tobago. Zootaxa, 3711(1), 1-64.

Poggiato, G., Andréoletti, J., Pollock, L. J., & Thuiller, W. (2025). Integrating food webs in species distribution models can improve ecological niche estimation and predictions. Ecography, 2025(5), e07546.

Poggiato, G., Münkemüller, T., Bystrova, D., Arbel, J., Clark, J. S., & Thuiller, W. (2021). On the interpretations of joint modeling in community ecology. Trends in ecology & evolution, 36(5), 391-401.

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., ... & McCarthy, M. A. (2014). Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). Methods in Ecology and Evolution, 5(5), 397-406.

Shipley, B. R., Bach, R., Do, Y., Strathearn, H., McGuire, J. L., & Dilkina, B. (2022). megaSDM: integrating dispersal and time-step analyses into species distribution models. Ecography, 2022(1).

Staniczenko, P. P., Sivasubramaniam, P., Suttle, K. B., & Pearson, R. G. (2017). Linking macroecology and community ecology: refining predictions of species distributions using biotic interaction networks. Ecology letters, 20(6), 693-707.

Tikhonov, G., Opedal, Ø. H., Abrego, N., Lehikoinen, A., de Jonge, M. M., Oksanen, J., & Ovaskainen, O. (2020). Joint species distribution modelling with the R-package Hmsc. Methods in ecology and evolution, 11(3), 442-447.

Tran, S. M., Howell, K. J., & Walsh, M. R. (2024). Increased eye size is favoured in Trinidadian killifish experimentally transplanted into low light, high competition environments. Journal of Evolutionary Biology, 37(8), 960-966.

Ulrich, W., & Gotelli, N. J. (2010). Null model analysis of species associations using abundance data. Ecology, 91(11), 3384-3397.

Valiente-Banuet, A., Aizen, M. A., Alcántara, J. M., Arroyo, J., Cocucci, A., Galetti, M., ... & Zamora, R. (2015). Beyond species loss: the extinction of ecological interactions in a changing world. Functional ecology, 29(3), 299-307.

Viana, D. S., & Chase, J. M. (2019). Spatial scale modulates the inference of metacommunity assembly processes. Ecology, 100(2), e02576.

Wainwright, M. J., & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. Foundations and Trends® in Machine Learning, 1(1–2), 1-305.

Walsh, M. R., & Reznick, D. N. (2009). Phenotypic diversification across an environmental gradient: a role for predators and resource availability on the evolution of life histories. Evolution, 63(12), 3201-3213.

Walsh, M. R., Fraser, D. F., Bassar, R. D., & Reznick, D. N. (2011). The direct and indirect effects of guppies: implications for life-history evolution in Rivulus hartii. Functional Ecology, 25(1), 227-237.

Warton, D. I., Blanchet, F. G., O'Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C., & Hui, F. K. (2015). So many variables: joint modeling in community ecology. Trends in ecology & evolution, 30(12), 766-779.

Barry, S. C., & Welsh, A. H. (2002). Generalized additive modelling and zero inflated count data. Ecological modelling, 157(2-3), 179-188.

Wenger, S. J., & Freeman, M. C. (2008). Estimating species occurrence, abundance, and detection probability using zero-inflated distributions. Ecology, 89(10), 2953-2959.

Wilkinson, D. P., Golding, N., Guillera-Arroita, G., Tingley, R., & McCarthy, M. A. (2019). A comparison of joint species distribution models for presence–absence data. Methods in Ecology and Evolution, 10(2), 198-211.

Winemiller, K. O. (1989). Ontogenetic diet shifts and resource partitioning among piscivorous fishes in the Venezuelan ilanos. Environmental Biology of fishes, 26(3), 177-199.

Zhang, C., Chen, Y., Xu, B., Xue, Y., & Ren, Y. (2019). How to predict biodiversity in space? An evaluation of modelling approaches in marine ecosystems. Diversity and Distributions, 25(11), 1697-1708.

Zurell, D., Pollock, L. J., & Thuiller, W. (2018). Do joint species distribution models reliably detect interspecific interactions from co-occurrence data in homogenous environments?. Ecography, 41(11), 1812-1819.

License details for species silhouettes that are not in the public domain:

For Ancistrus maracasae, Andinoacara pulcher, Hemibrycon taeniurus, Hoplias malabaricus, Hypostomus robinii, Roeboides dientonito,: credits to Grupo de Ictiología de la Universidad de Antioquia. https://creativecommons.org/licenses/by/4.0/

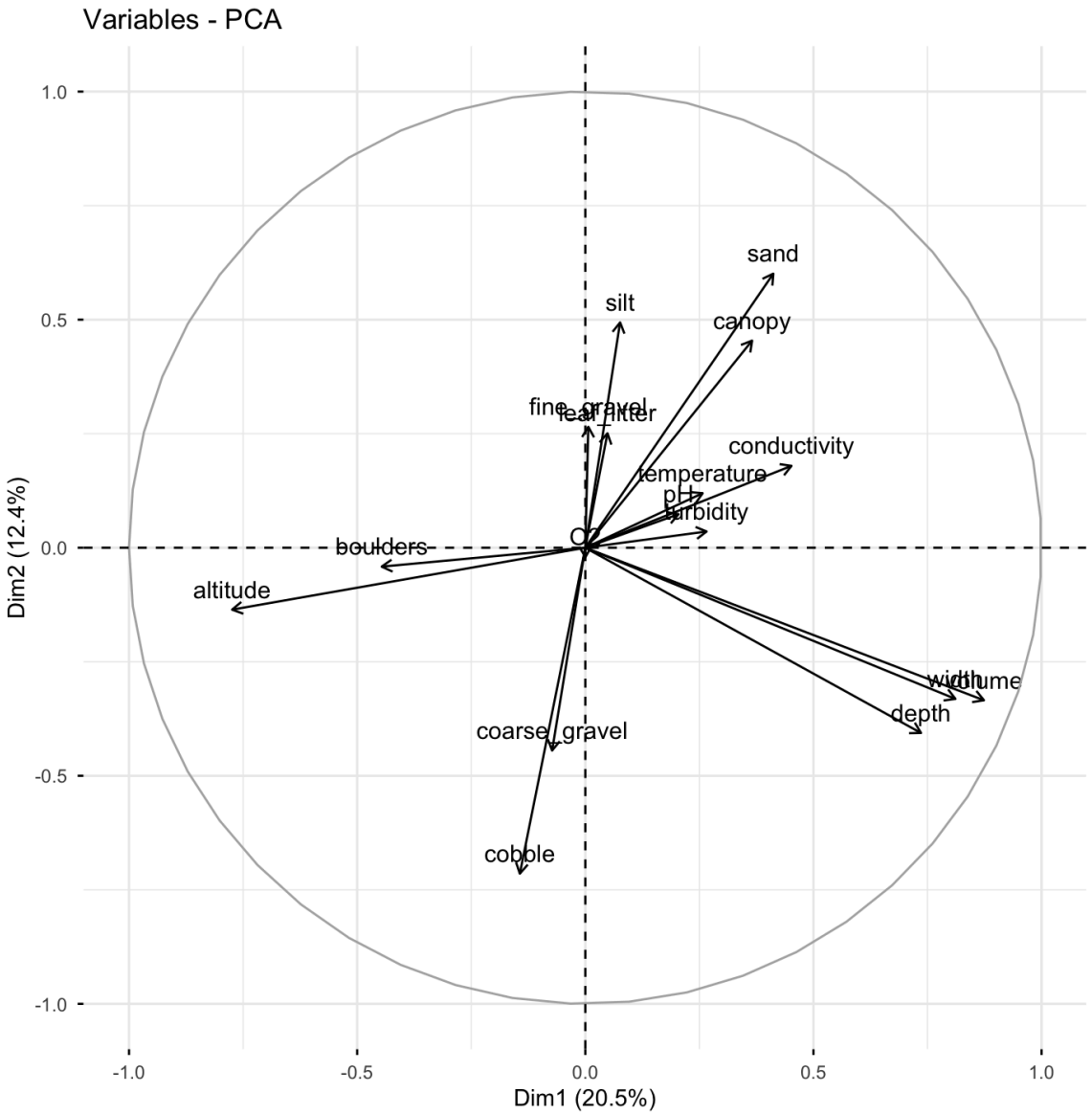
For Rhamdia quelen, Crenicichla frenata: credits to Cesar Julian.
https://creativecommons.org/licenses/by/3.0/

Appendices

Appendix S1: Correspondence between full species names and their shortened versions.

Shortened name	Full name	Common name
Ancimara	*Ancistrus maracasae*	Jumbie teta
Andipulc	*Andinoacara pulcher*	Blue coscorob
Angurost	*Anguilla rostrata*	American eel
Arconudu	*Arcos nudus*	Padded clingfish
Astybima	*Astyanax bimaculatus*	Two spot sardine
Awaobana	*Awaous banana*	River goby
Cichtaen	*Cichlasoma taenia*	Brown coscarob
Coryaene	*Corydoras aeneus*	Bronze cory
Coryriis	*Corynopoma riisei*	Swordtail sardine
Crenfren	*Crenicichla frenata*	Pike cichlid
Dajamont	*Dajaus monticola*	Mountain mullet
Eleopiso	*Eleotris pisonis*	Spinycheek sleeper
Gastster	*Gasteropelecus sternicla*	Silver hatchet fish
Gobidorm	*Gobiomorus dormitor*	Bigmouth sleeper
Gymncara	*Gymnotus carapo*	Banded knifefish
Hemitaen	*Hemibrycon taeniurus*	Mountain stream sardine
Hemiunil	*Hemigrammus unilineatus*	Featherfin tetra
Hopllitt	*Hoplosternum littorale*	Cascadu
Hoplmala	*Hoplias malabaricus*	Guabine
Hyporobi	*Hypostomus robinii*	Teta
Odonpulc	*Odontostilbe pulchra*	Tetra
Oreosp	*Oreochromis* species (information given at the genus level)	Tilapia
Poecboes	*Poecilia boesemani*	Maraval toothcarp
Poecreti	*Poecilia reticulata*	Guppy
Poecvivi	*Poecilia vivipara*	Eye spot toothcarp
Polyscho	*Polycentrus schomburgkii*	Guyana leaffish
Rhamquel	*Rhamdia quelen*	Silver catfish
Rivuhart	*Rivulus hartii (= Anablepsoides hartii)*	Jumping guabine
Roebdien	*Roeboides dientonito*	Hunchback sardine
Sicypunc	*Sicydium punctatum*	Tri Tri
Steiarge	*Steindachnerina argentea*	Stout sardine
Synbmarm	*Synbranchus marmoratus*	Zangee

Appendix S2: Results of the Principal Component Analysis on the environmental covariates.



This figure describes the weight of each covariate in each of the first two principal components resulting from the PCA run on the tropical freshwater fish species distribution dataset. The variables used include each site's altitude, the depth, width, volume, temperature, turbidity, conductivity, oxygen concentration and pH of the water, descriptors of the canopy and of the soil substrates (boulder, fine gravel, leaf litter, sand, leaf litter, coarse gravel and cobble).

Appendix S3: A quick guide to using the ZIPLN-network model for the analysis of species abundance data containing structural zeros.

The code used for the study of the analysis of the tropical freshwater fish distribution data is provided in the github repository:
jeannetous/JSDM_with_zero_inflation_for_increased_ecological_relevance_in_species_distribution_analysis

For more details about the R package **PLNmodels** and its features, the reader may refer to [Chiquet et al., 2021] or to https://pln-team.github.io/PLNmodels/articles/.

We provide a quick step-by-step guide to analyse species distribution data with structural zeros using the R package **PLNmodels**. This guide does not include all the refinements detailed in the **PLNmodels** documentation. The user should first install and load the **PLNmodels** package (version 1.2.2 or higher).

1. Data formatting:

- Species distribution data must be turned into a matrix Y with n rows (the number of sites) and p columns (the number of species), with Y_{ij} an integer giving the number of individuals from species j in sites i. If the data are contained in a csv, this can be done using the **read.csv** R function.

- The covariates must be turned into a matrix X with n rows (the number of sites) and d columns (the number of covariates). These covariates can be the results of a PCA or other transformations of the original covariates; we used a PCA (section 2.5 and appendix S2).

- A data object **df** specifically designed for the **PLNmodels** functions is then created using the function **prepare_data** of **PLNmodels**: **df ← prepare_data(Y, X)**.

2. Running the ZIPLN-network model:

- The user must first decide which covariates should be included in the model. This can be done *a priori* using ecological knowledge of the data or *a posteriori* comparing statistical criteria describing the model fitting quality obtained with different covariates combinations. Then denote X_1 to X_d the selected covariates. We consider that an intercept is also used. In our study we selected as covariates the first two components of a PCA run on a number of environmental covariates (see section 2.5).

- The user should also decide what type of zero-inflation is required: sites, species or covariates-dependent and, in the latter case, decide what covariates should be used. We denote them X_1^0 to $X_{d'}^0$. In our study, after analysing the data (Figure 1) we tested both a

species-dependent zero-inflation and a covariate-dependent zero-inflation, using the stream covariate (see section 2.5).

- For a sites-dependent zero-inflation, the user should run:

  models ← ZIPLNnetwork(Abundance ~ 1 + X_1 **+ … +** X_d**, data = df, zi = "row",**
  control = ZIPLNnetwork_param(penalize_diagonal = FALSE))

- For a species-dependent zero-inflation, the user should run:

  models ← ZIPLNnetwork(Abundance ~ 1 + X_1 **+ … +** X_d**, data = df, zi = "col",**
  control = ZIPLNnetwork_param(penalize_diagonal = FALSE))

- For a covariates-dependent zero-inflation (including an intercept), the user should run:

  models ← ZIPLNnetwork(Abundance ~ 1 + X_1 **+ … +** X_d **| 1 +** X_1^0 to $X_{d'}^0$ **,**
  data = df,
  control = ZIPLNnetwork_param(penalize_diagonal = FALSE))

3. Analyzing the results:

- The object **models** contains several ZIPLN-network models, one per ℓ_1 penalty used (see section 2.1 and [Chiquet et al., 2019]). The user must first select one model, either picking a specific penalty λ running **model ← models$getModel** (λ) or by selecting the penalty that gives the best result for a specific criterion (BIC, EBIC or StARS) **model ← models$getBestModel(*criterion*)**. In our study, we used the StARS selection criterion with a stability selection value of 0.8 (see section 2.1).

- The user may then access several attributes of the object **model** to study the results. Details are given in the **PLNmodels** documentation but the user may for instance visualize the inferred species association network running **model$plot_network()** (Figure 7) or extract the regression coefficients using **model$model_par$B** (Figure 5) or the zero-inflation probabilities with **model$model_par$Pi** (Figure 6).