



Les covariables non-observées dans les modèles de distribution d'espèce : conséquences inférentielles et solutions via les modèles de distribution conjointe

Arthur F. ROSSIGNOL & Frédéric GOSSELIN

INRAE, UR EFNO, Nogent-sur-Vernisson, France

Journée AppliBUGS • 19 juin 2026 • EDF Palaiseau

I

introduction

INTRODUCTION

un peu de contexte

species distribution models (modèles de distribution d'espèces) = outil majeur de modélisation en écologie pour étudier et prédire la distribution spatiale et temporelle des populations d'espèces

beaucoup d'applications : biologie de la conservation, gestion de la biodiversité, génie écologique, évaluation de l'impact des espèces exotiques envahissantes, évaluation des impacts du changement climatique

plusieurs implémentations existantes : GL(M)M, GA(M)M, MaxEnt, *machine learning*...

INTRODUCTION

un peu de contexte

MAIS il y a des hypothèses fortes derrière les SDM :

- distribution expliquée par rapport aux covariables mesurées qui sont le plus souvent des conditions abiotiques (e.g., conditions environnementales)
- interactions biotiques pas prises en compte (compétition, facilitation, mutualisme, prédation, dynamique hôte-parasitoïde)
- chaque espèce étudiée est considérée comme indépendante des autres espèces, donc les interactions interspécifiques sont considérées comme du bruit

INTRODUCTION

exhaustivité des variables

DONNÉES

sites $i = 1, \dots, n$

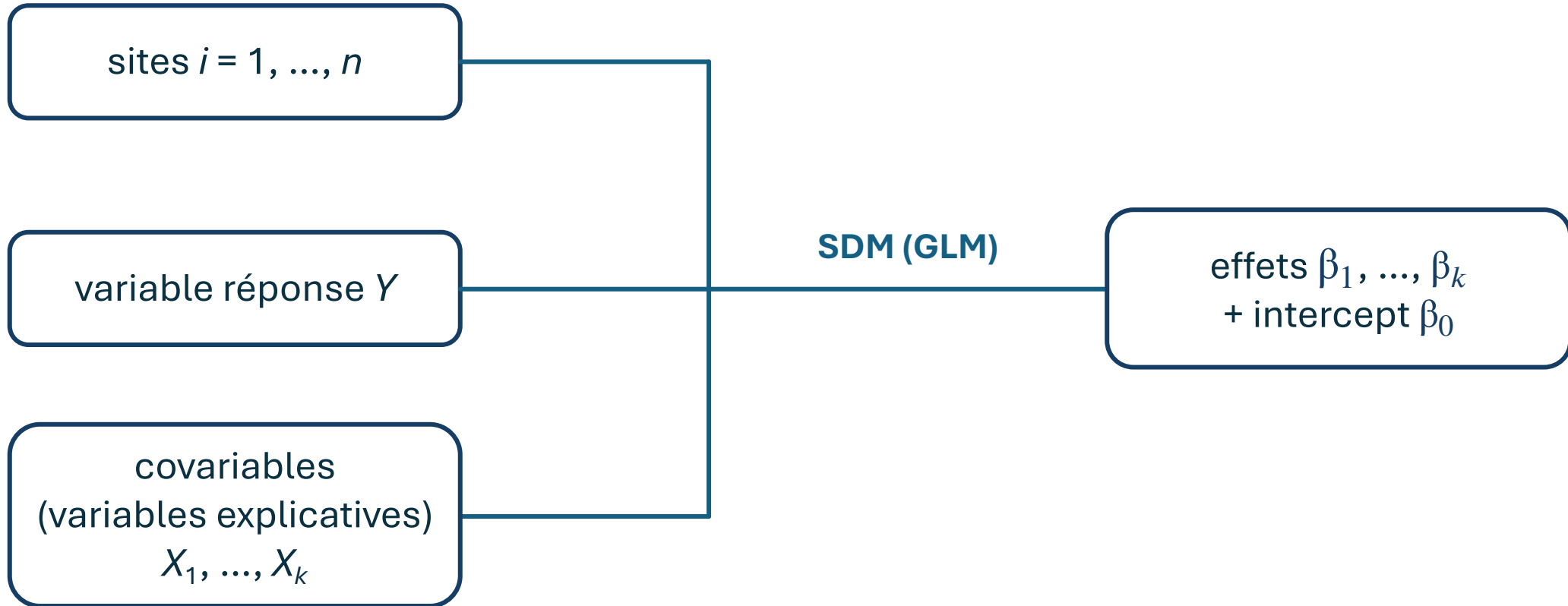
variable réponse Y

covariables
(variables explicatives)
 $X_1, \dots, X_k$

SDM (GLM)

PARAMÈTRES ESTIMÉS

effets $\beta_1, \dots, \beta_k$
+ intercept β_0



INTRODUCTION

exhaustivité des variables

DONNÉES

PARAMÈTRES ESTIMÉS

sites $i = 1, \dots, n$

mais en pratique dispose-t-on de toutes les variables explicatives !?

covariables
(variables explicatives)
 $X_1, \dots, X_k$

+ intercept β_0

INTRODUCTION

“omitted-variable bias”: un problème négligé en écologie

COVARIABLES:

$$X_1, X_2 \sim \text{Normal}(\mathbf{0}, \mathbf{I})$$

VRAI PROCESSUS:

$$Y \mid \{X_1, X_2\} \sim \mathcal{L}(f(\alpha + \beta X_1 + \gamma X_2))$$

MODÈLE DE REGRESSION:

$$Y \mid X_1 \sim \mathcal{L}(f(\alpha^* + \beta^* X_1))$$

→ A-t-on $\alpha^* = \alpha$ et $\beta^* = \beta$?

→ Je vous pose la question !

“omitted-variable bias”: un problème négligé en écologie

QUESTION 1

avec une régression gaussienne (variable réponse dans $\mathbb{R}$), a-t-on :

(1) $\alpha^* = \alpha$ **et** $\beta^* = \beta$?

(2) $\alpha^* = \alpha$ **et** $\beta^* \neq \beta$?

(3) $\alpha^* \neq \alpha$ **et** $\beta^* = \beta$?

(4) $\alpha^* \neq \alpha$ **et** $\beta^* \neq \beta$?

“omitted-variable bias”: un problème négligé en écologie

QUESTION 1

avec une régression gaussienne (variable réponse dans $\mathbb{R}$), a-t-on :

(1) $\alpha^* = \alpha$ **et** $\beta^* = \beta$?

(2) $\alpha^* = \alpha$ **et** $\beta^* \neq \beta$?

(3) $\alpha^* \neq \alpha$ **et** $\beta^* = \beta$?

(4) $\alpha^* \neq \alpha$ **et** $\beta^* \neq \beta$?

QUESTION 2

avec une régression de Poisson (variable réponse dans $\mathbb{N}$), a-t-on :

(1) $\alpha^* = \alpha$ **et** $\beta^* = \beta$?

(2) $\alpha^* = \alpha$ **et** $\beta^* \neq \beta$?

(3) $\alpha^* \neq \alpha$ **et** $\beta^* = \beta$?

(4) $\alpha^* \neq \alpha$ **et** $\beta^* \neq \beta$?

“omitted-variable bias”: un problème négligé en écologie

QUESTION 2

avec une régression de Poisson (variable réponse dans $\mathbb{N}$), a-t-on :

(1) $\alpha^* = \alpha$ **et** $\beta^* = \beta$?

(2) $\alpha^* = \alpha$ **et** $\beta^* \neq \beta$?

(3) $\alpha^* \neq \alpha$ **et** $\beta^* = \beta$?

(4) $\alpha^* \neq \alpha$ **et** $\beta^* \neq \beta$?

QUESTION 3

avec une régression de Bernoulli (variable réponse dans $\{0,1\}$), a-t-on :

(1) $\alpha^* = \alpha$ **et** $\beta^* = \beta$?

(2) $\alpha^* = \alpha$ **et** $\beta^* \neq \beta$?

(3) $\alpha^* \neq \alpha$ **et** $\beta^* = \beta$?

(4) $\alpha^* \neq \alpha$ **et** $\beta^* \neq \beta$?

QUESTION 3

avec une régression de Bernoulli (variable réponse dans $\{0,1\}$), a-t-on :

(1) $\alpha^* = \alpha$ **et** $\beta^* = \beta$?

(2) $\alpha^* = \alpha$ **et** $\beta^* \neq \beta$?

(3) $\alpha^* \neq \alpha$ **et** $\beta^* = \beta$?

(4) $\alpha^* \neq \alpha$ **et** $\beta^* \neq \beta$?

INTRODUCTION

“omitted-variable bias”: un problème négligé en écologie

en données de présence-absence, il y a
une atténuation de la magnitude de l'effet

OMITTED-VARIABLE BIAS

collapsibility

INTRODUCTION

“omitted-variable bias”: un problème négligé en écologie

distribution	fonction de lien	Biais[α^*]	Biais[β^*]
Normal	identité	$= 0$	$= 0$
Poisson	log	$\neq 0 (> 0)$	$= 0$
Bernoulli	probit	$\neq 0$ $\begin{cases} < 0 & \text{si } \alpha > 0 \\ > 0 & \text{si } \alpha < 0 \end{cases}$	$\neq 0$ $\begin{cases} < 0 & \text{si } \beta > 0 \\ > 0 & \text{si } \beta < 0 \end{cases}$
Bernoulli	logit	<i>idem</i>	<i>idem</i>

“omitted-variable bias”: un problème négligé en écologie

modèle de Poisson avec lien log

$$\text{Biais}[\alpha^*] = \gamma \left(\mu_{X_2} + \gamma \frac{\sigma_{X_2}^2}{2} \right)$$

$$\text{Biais}[\beta^*] = 0$$

INTRODUCTION

“omitted-variable bias”: un problème négligé en écologie

modèle de Poisson avec lien log

$$\text{Biais}[\alpha^*] = \frac{\gamma^2}{2}$$

(lorsque X_2 est centrée-réduite)

$$\text{Biais} [\beta^*] = 0$$

“omitted-variable bias”: un problème négligé en écologie

modèle de Bernoulli avec lien probit

$$\text{Biais}[\alpha^*] = \alpha \left(\frac{1}{\sqrt{1 + \gamma^2 \sigma_{X_2}^2}} - 1 \right) + \frac{\gamma \mu_{X_2}}{\sqrt{1 + \gamma^2 \sigma_{X_2}^2}}$$

$$\text{Biais}[\beta^*] = \beta \left(\frac{1}{\sqrt{1 + \gamma^2 \sigma_{X_2}^2}} - 1 \right)$$

“omitted-variable bias”: un problème négligé en écologie

modèle de Bernoulli avec lien probit

$$\text{Biais}[\alpha^*] = \alpha \left(\frac{1}{\sqrt{1 + \gamma^2}} - 1 \right)$$

(lorsque X_2 est centrée-réduite)

$$\text{Biais}[\beta^*] = \beta \left(\frac{1}{\sqrt{1 + \gamma^2}} - 1 \right)$$

INTRODUCTION

un peu de terminologie anglo-saxonne

dans la littérature (notamment en écologie), on trouve les termes suivants pour ce phénomène :

missing covariates

omitted covariates

unobserved covariates

unmeasured covariates

hidden covariates

...mais peu/pas de mentions du biais associé

INTRODUCTION

un nouvel outil en écologie des communautés : les JSDM

au lieu de considérer chaque espèce indépendamment,

on modélise simultanément l'ensemble des espèces de la communauté

ce sont les *joint species distribution models* (modèles de distribution conjointe d'espèces)

→ prise en compte des covariations entre espèces

différentes formulations existent → ici on s'intéressera aux JSDM à variable(s) latente(s)

INTRODUCTION

les objectifs

d'une part

**mettre en lumière un
"risque" important
mais négligé qui peut
impacter les SDM**

d'autre part

**contribuer au débat
actuel sur les JSdm
et leur potentiel pour
améliorer les SDM**

II

méthodes

MÉTHODES

trois grandes étapes

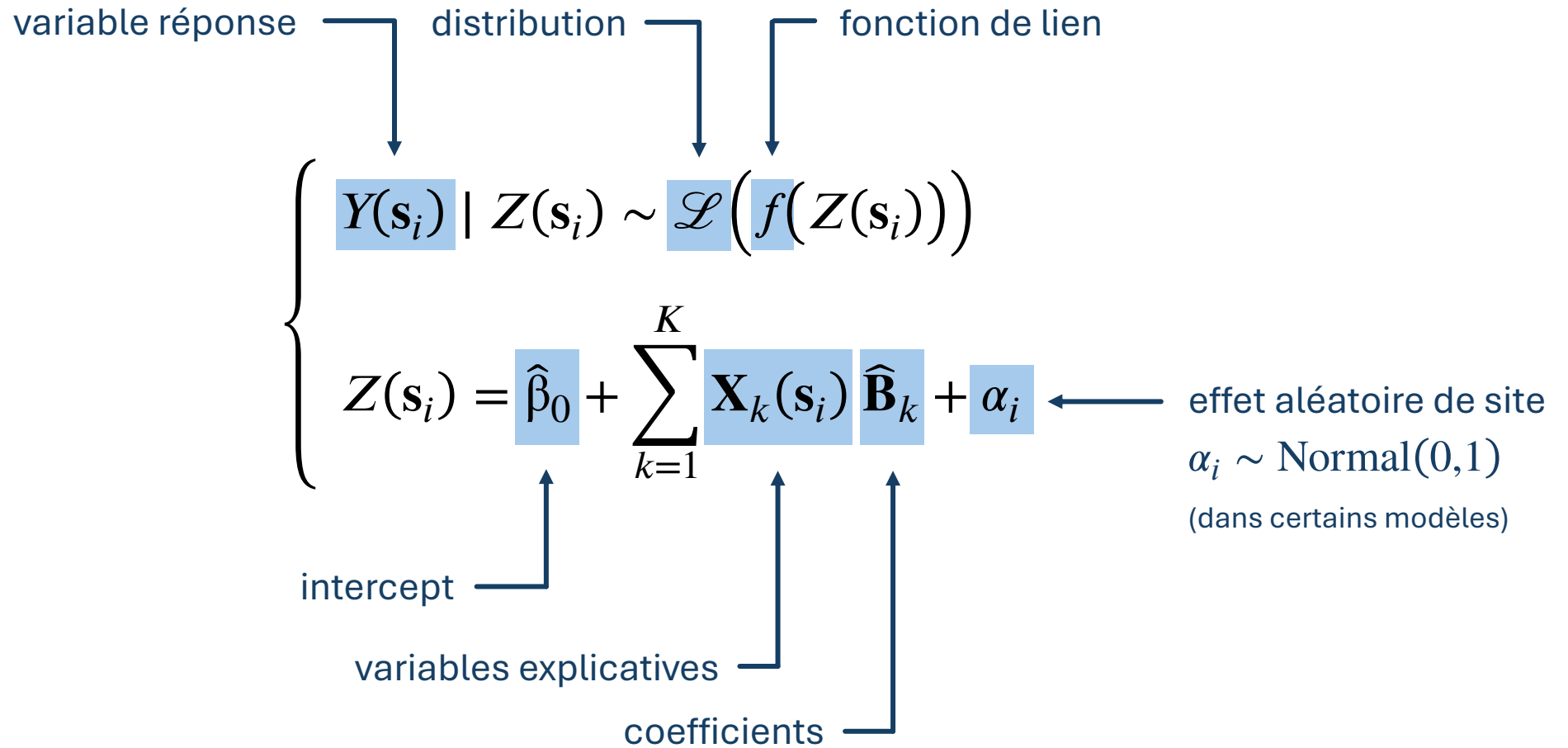
(i) étude par simulation des SDM

(ii) étude par simulation des JSDM

(ii) application à des données réelles

MÉTHODES

species distribution models (SDM)



MÉTHODES

species distribution models (SDM)

GAUSSIAN

continuous data

identity

POISSON

count data
(abundance)

log

BERNOULLI

binary data
(presence–absence)

logit
probit
cloglog

MÉTHODES

species distribution models (SDM)

formulations de SDM testées

modèle	function(s)	package(s) R	inférence	type de modèle
SDM.1	lm glm lmer glmer	stats lme4	fréquentiste	GL(M)M
SDM.2	glmmTMB	glmmTMB	fréquentiste	GL(M)M
SDM.3	fitme	spaMM	fréquentiste	GL(M)M
SDM.4	mixed_model	GLMMadaptive	fréquentiste	GLMM
SDM.5	brm	brms	Bayésien	GL(M)M
SDM.6	inla	INLA	Bayésien	GL(M)M
SDM.7	—	nimble	Bayésien	GL(M)M
SDM.8	—	runjags	Bayésien	GL(M)M
SDM.9	lmrob glmrob	robustbase	fréquentiste	robust GLM
SDM.10	glm	stats	fréquentiste	quasi-likelihood GLM
SDM.11	gam gamm	mgcv	fréquentiste	GA(M)M

MÉTHODES

joint species distribution models (JSDM)

variable réponse distribution fonction de lien

$$\left\{ \begin{array}{l} Y_j(\mathbf{s}_i) \mid Z_j(\mathbf{s}_i) \sim \mathcal{L}\left(f\left(Z_j(\mathbf{s}_i)\right)\right) \\ Z_j(\mathbf{s}_i) = \hat{\beta}_0^{[j]} + \sum_{k=1}^K \mathbf{X}_k(\mathbf{s}_i) \hat{\mathbf{B}}_k^{[j]} + \sum_{\ell=1}^L \widehat{\mathbf{W}}_{\ell}(\mathbf{s}_i) \hat{\Lambda}_{\ell}^{[j]} \end{array} \right.$$

intercept variables explicatives coefficients variables latentes loadings

The diagram illustrates the structure of a Joint Species Distribution Model (JSDM). It consists of two main equations grouped by a large left curly brace. The top equation is a conditional distribution: $Y_j(\mathbf{s}_i) \mid Z_j(\mathbf{s}_i) \sim \mathcal{L}\left(f\left(Z_j(\mathbf{s}_i)\right)\right)$. The bottom equation is a linear model for the latent variable $Z_j(\mathbf{s}_i)$: $Z_j(\mathbf{s}_i) = \hat{\beta}_0^{[j]} + \sum_{k=1}^K \mathbf{X}_k(\mathbf{s}_i) \hat{\mathbf{B}}_k^{[j]} + \sum_{\ell=1}^L \widehat{\mathbf{W}}_{\ell}(\mathbf{s}_i) \hat{\Lambda}_{\ell}^{[j]}$. Labels with arrows point to specific parts of these equations: 'variable réponse' points to $Y_j(\mathbf{s}_i)$; 'distribution' points to $\mathcal{L}$; 'fonction de lien' points to f ; 'intercept' points to $\hat{\beta}_0^{[j]}$; 'variables explicatives' points to $\mathbf{X}_k(\mathbf{s}_i)$; 'coefficients' points to $\hat{\mathbf{B}}_k^{[j]}$; 'variables latentes' points to $\widehat{\mathbf{W}}_{\ell}(\mathbf{s}_i)$; and 'loadings' points to $\hat{\Lambda}_{\ell}^{[j]}$.

MÉTHODES

joint species distribution models (JSDM)

formulations de JSDM testées

modèle	package R	inférence	coefficients de régression	contraintes sur les variables latentes	contraintes sur la matrice des loadings
JSDM.1	gllvm	fréquentiste	fixes	variance unitaire	diagonale positive & partie triangulaire sup. nulle
JSDM.2	jSDM	Bayésien	fixes	variance unitaire	diagonale positive & partie triangulaire sup. nulle
JSDM.3	Hmsc	Bayésien	aléatoires	variance unitaire	∅
JSDM.4	nimble	Bayésien	aléatoires	variance unitaire	∅
JSDM.5	runjags	Bayésien	fixes	variance unitaire	diagonale positive & partie triangulaire sup. nulle
JSDM.6	runjags	Bayésien	fixes	∅	diagonale unitaire & partie triangulaire sup. nulle

MÉTHODES

joint species distribution models (JSDM)

GAUSSIAN

continuous data

identity

POISSON

count data
(abundance)

log

BERNOULLI

binary data
(presence-absence)

logit
probit
cloglog

- comment interpréter la partie latente du modèle ?
 - reconstruire la matrice de variance–covariance interspécifique
 - reconstruire d'éventuelles covariables manquantes
- variable latente ~ covariable manquante mais importante pour la variable réponse ?
- loading ~ coefficient de régression associé à cette covariable manquante ?

MÉTHODES

d'une part, une étude par simulation

SIMULATION DES DONNÉES

deux covariables (X_1 et X_2)
indépendantes et
non-spatialement structurées

toutes les espèces répondent
aux deux covariables

ANALYSE DES DONNÉES

la covariable X_2
n'est jamais observée

les JSDM incluent une
unique variable latente
non-spatialement structurée

d'autre part, une application à des jeux de données écologiques réels

objectifs :

- comparer les estimations des SDM et des JSMD sur les mêmes données
- expliquer les différences dans le contexte théorique d'une covariable manquante

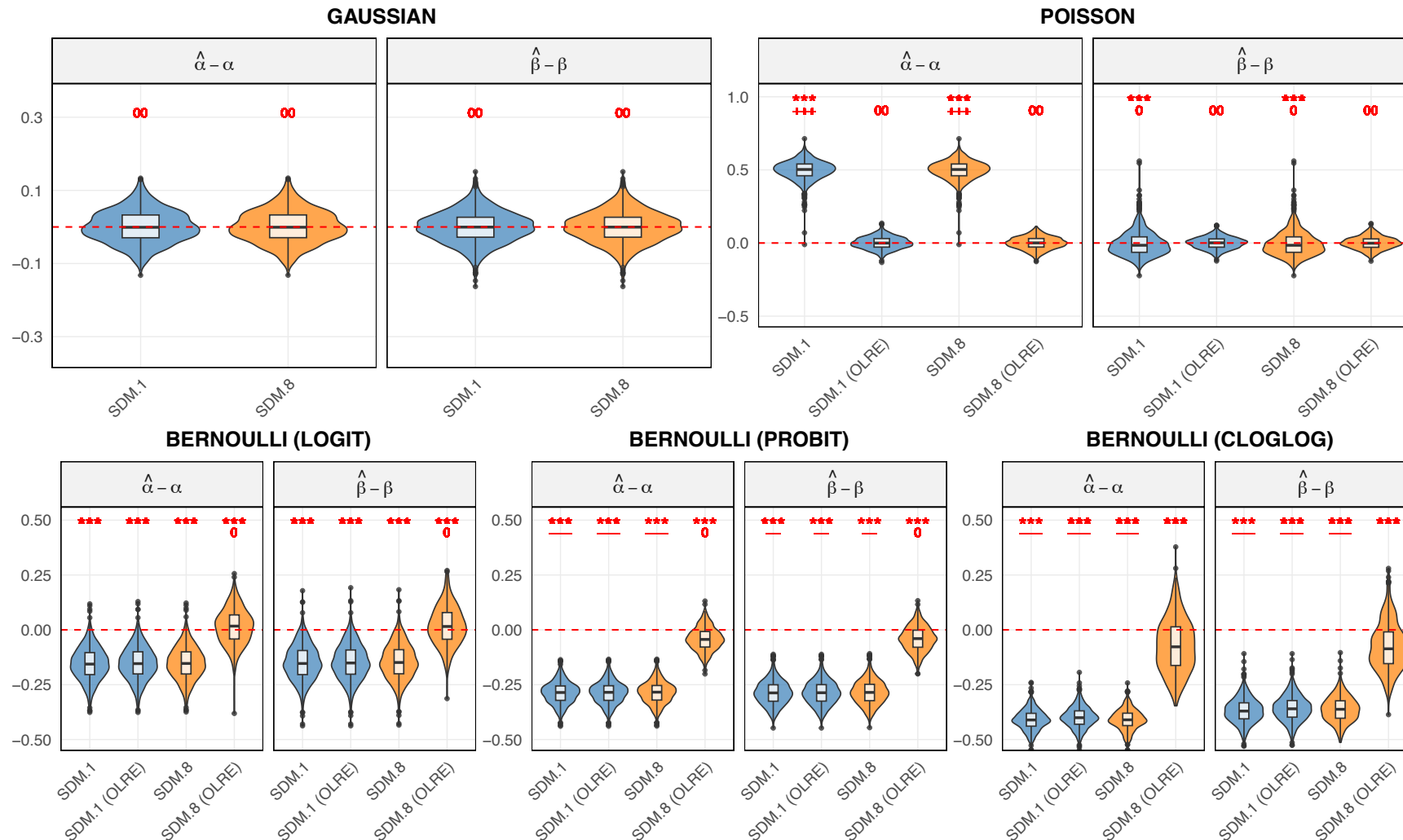
numéro	groupe taxonomique	nombre d'espèces	nombre de covariables	nombre de sites
#1	<i>tundra plants</i>	226	3	325
#2	<i>breeding birds</i>	370	8	2752

III

résultats

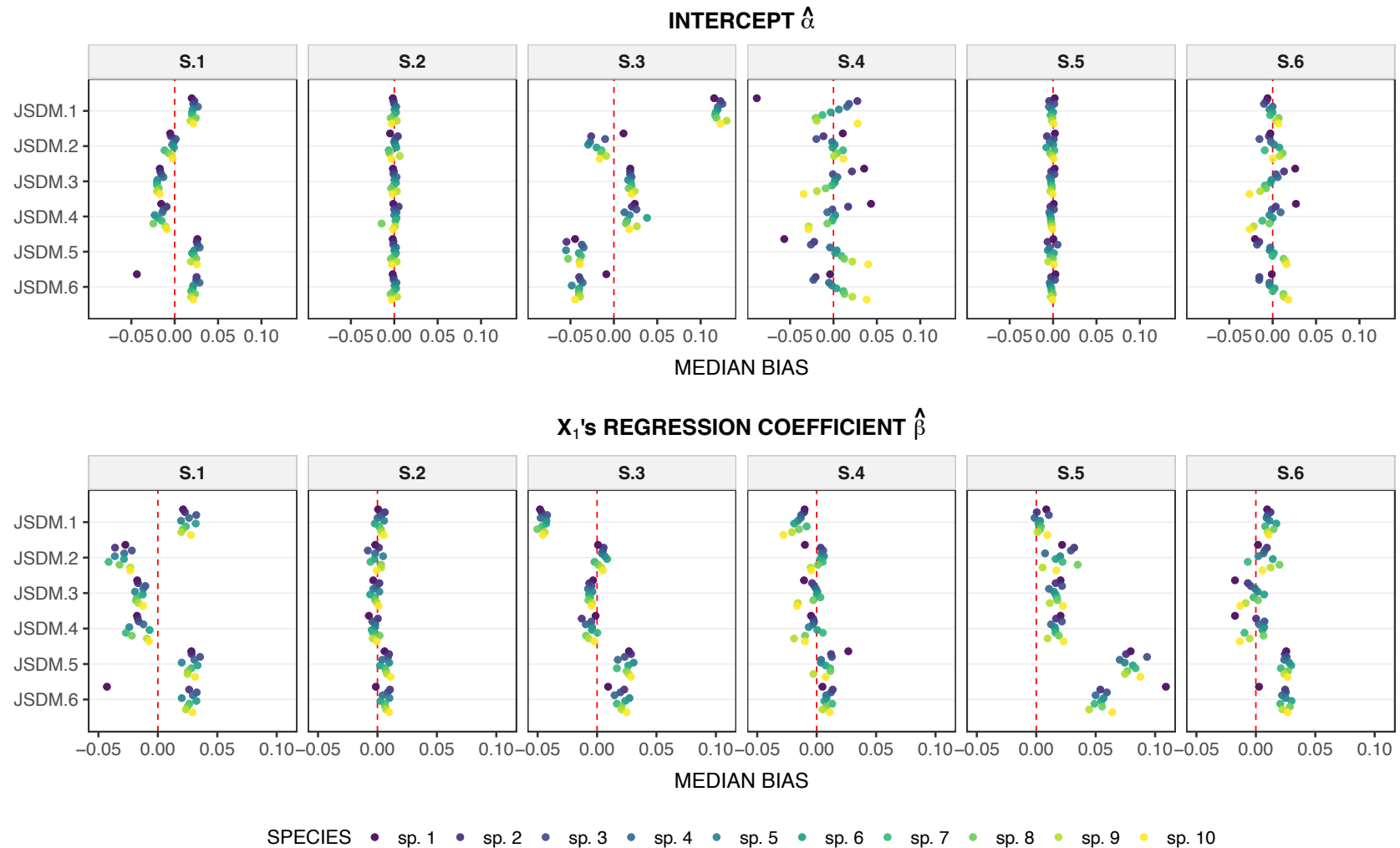
RÉSULTATS

évaluation des SDMs — observation du “omitted-variable bias”



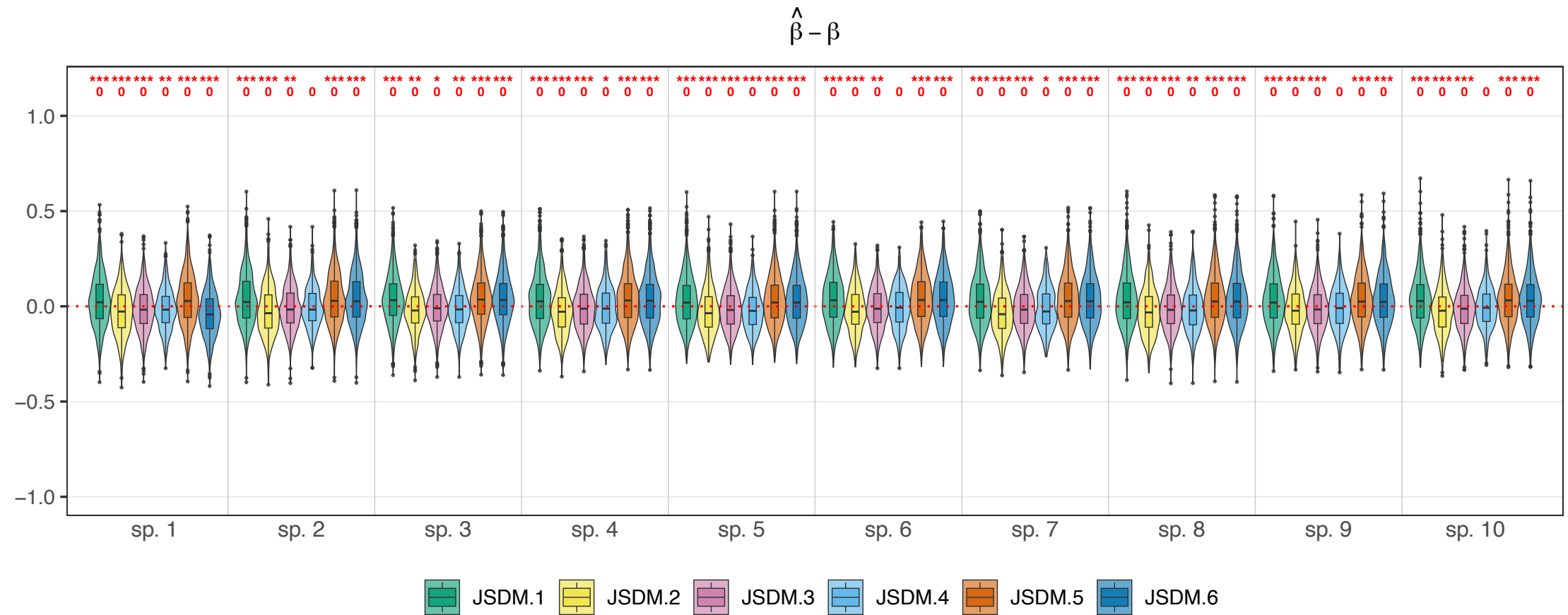
RÉSULTATS

évaluation des JSdMs — une covariable manquante et un facteur latent

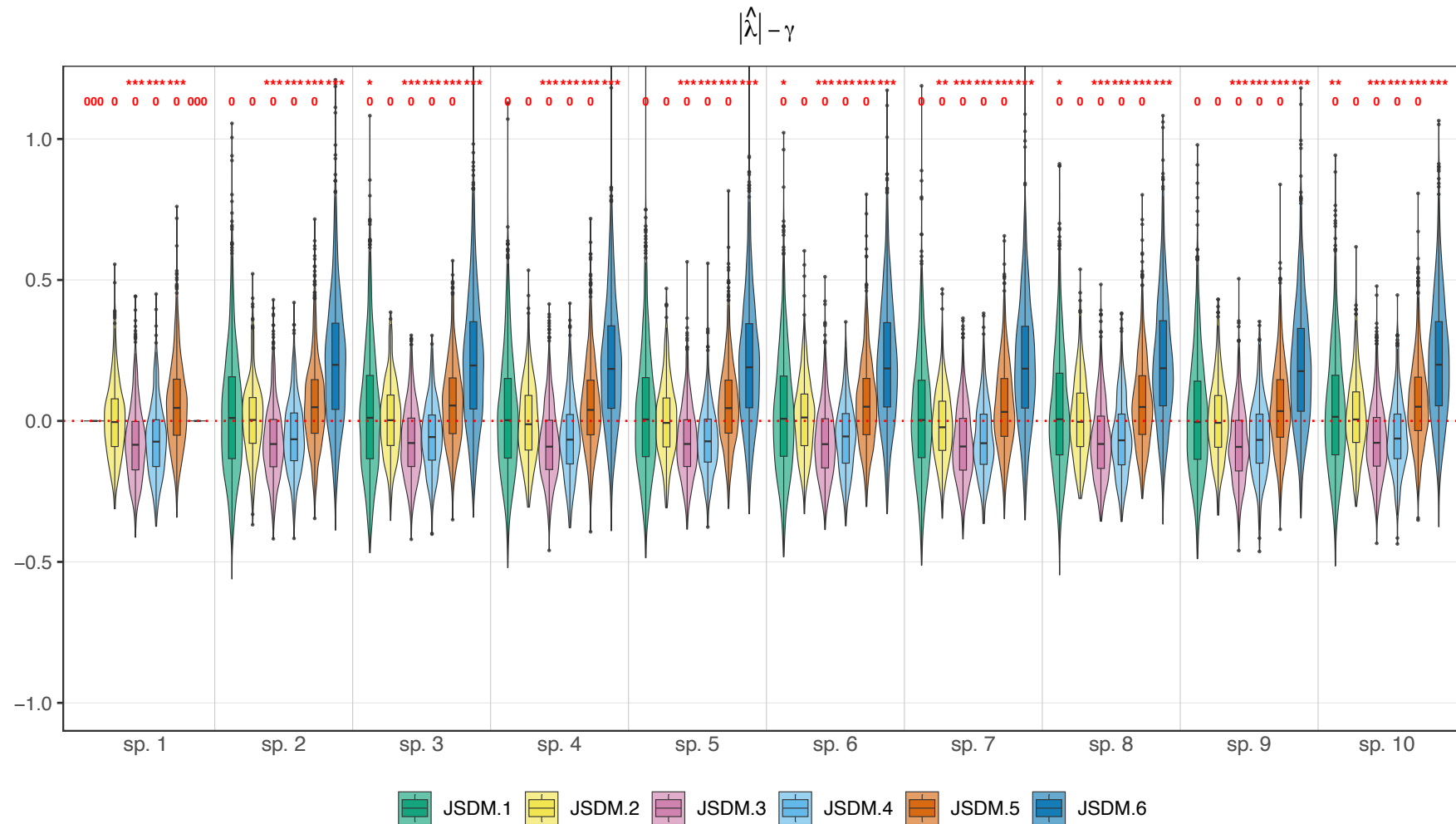


RÉSULTATS

évaluation des JSdMs — une covariable manquante et un facteur latent

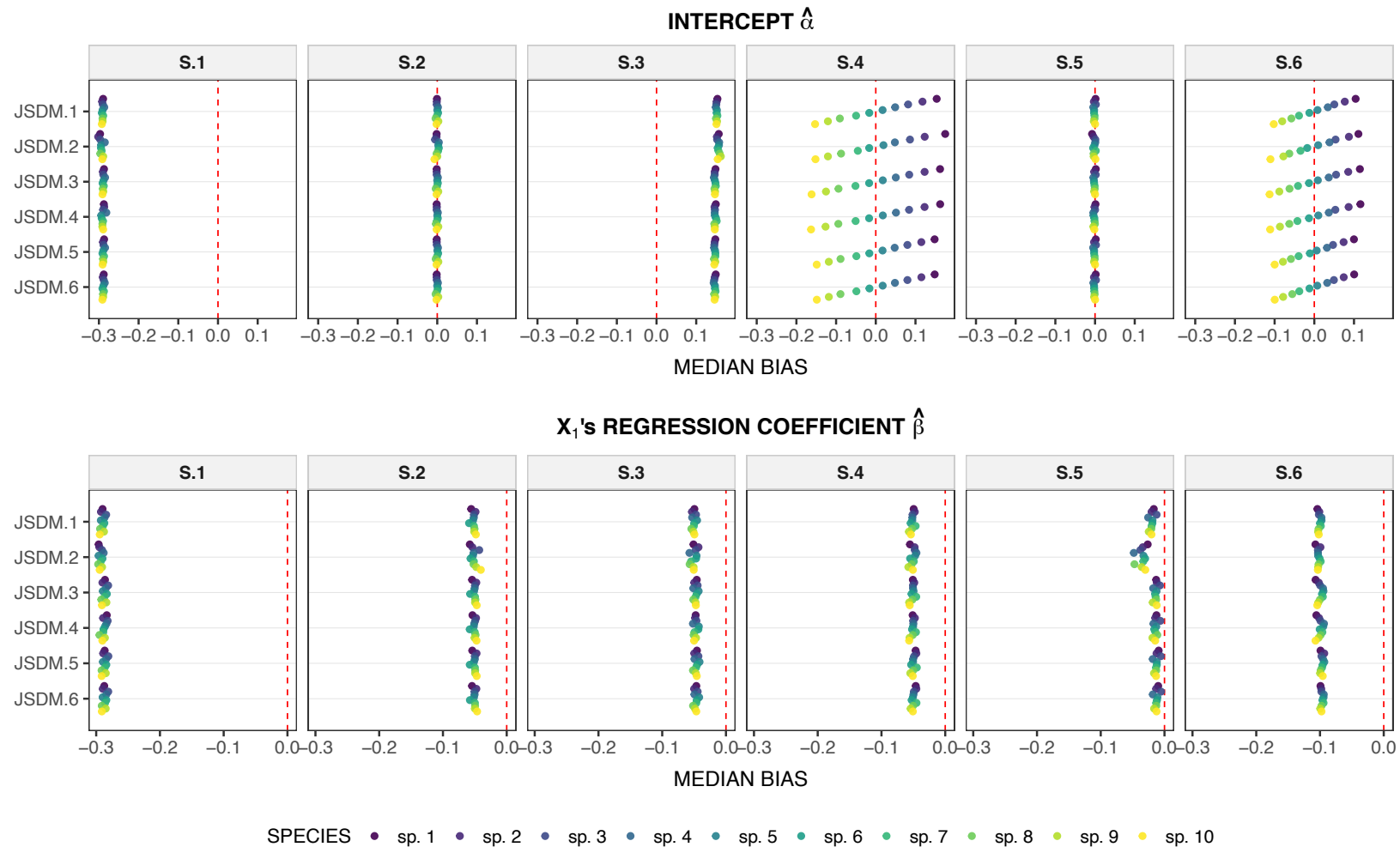


évaluation des JSDBs — une covariable manquante et un facteur latent



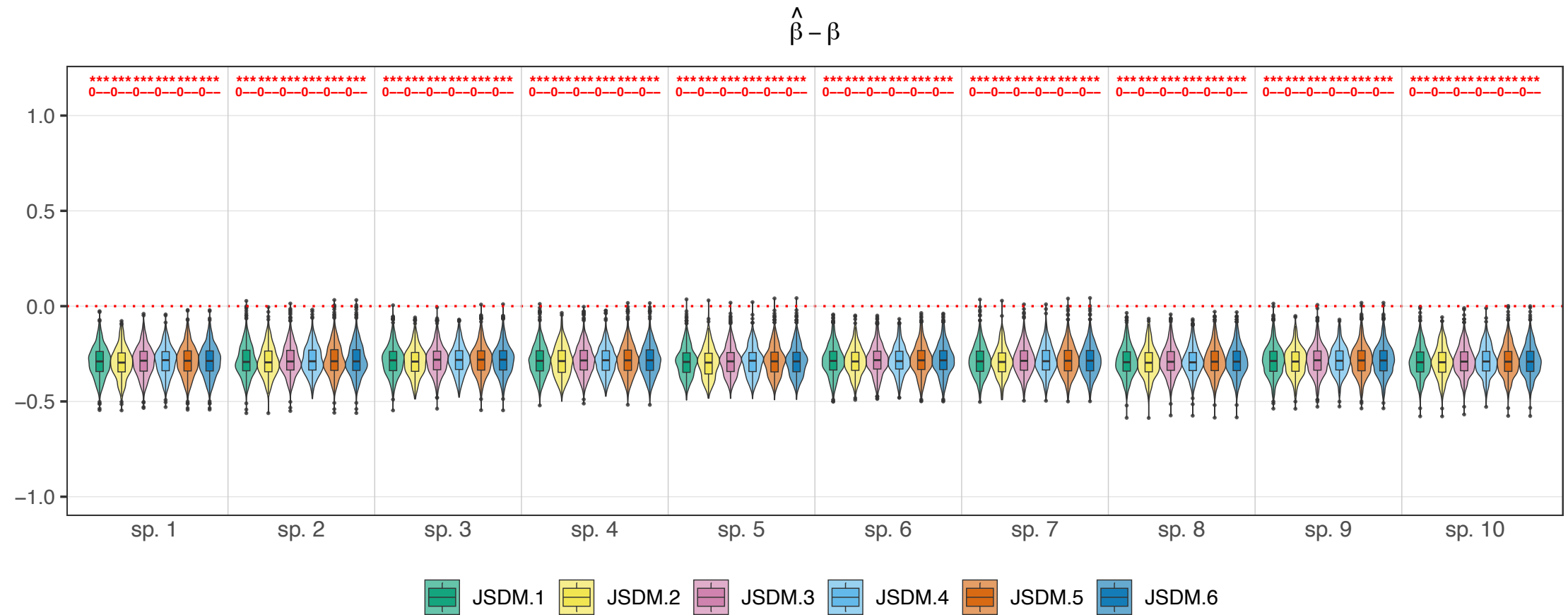
RÉSULTATS

évaluation des JSDMs — une covariable manquante mais aucun facteur latent



RÉSULTATS

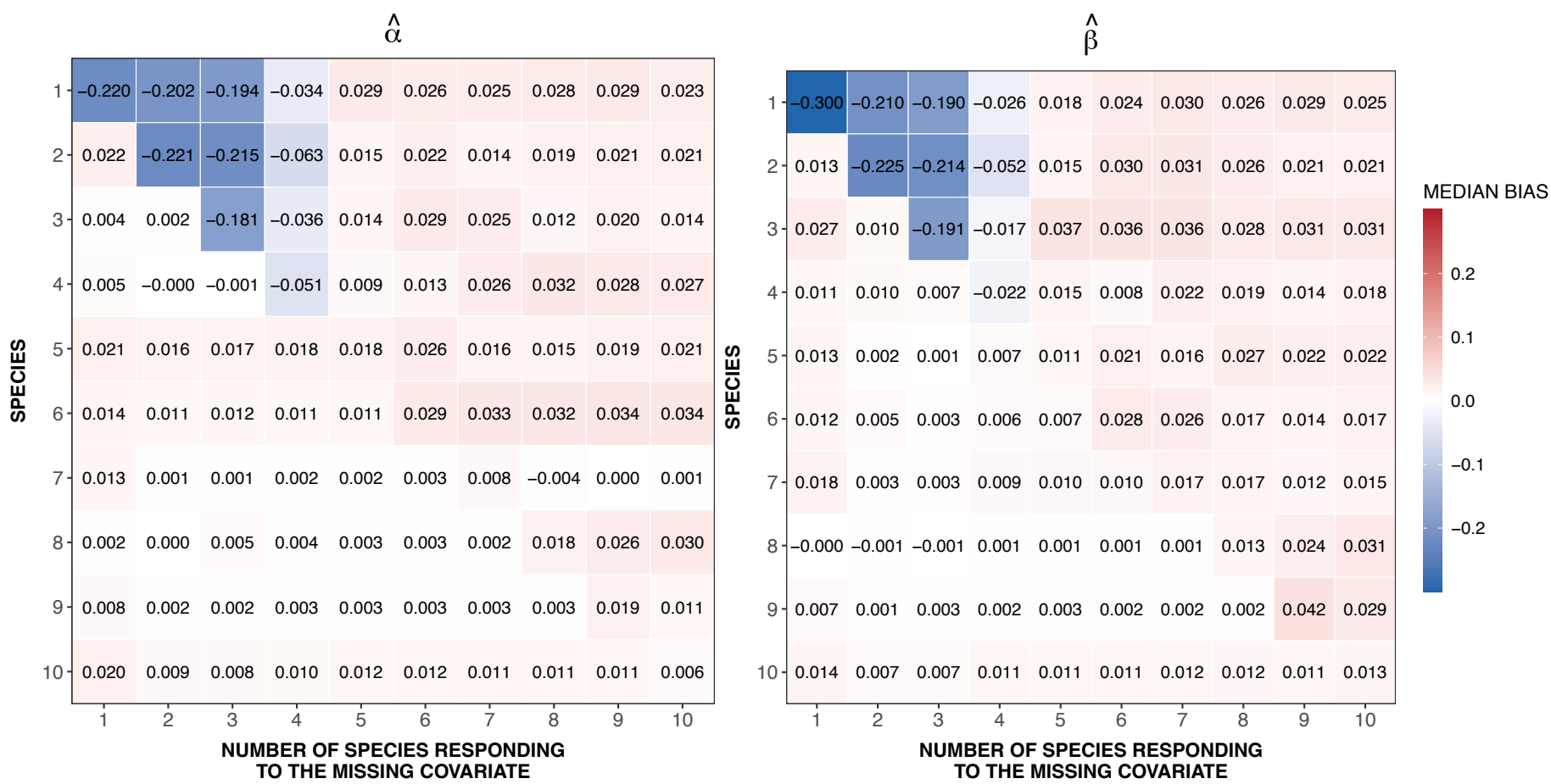
évaluation des JSdMs — une covariable manquante mais aucun facteur latent



RÉSULTATS

évaluation des JSDBs — influence du nombre d'espèces répondant à la covariable manquante

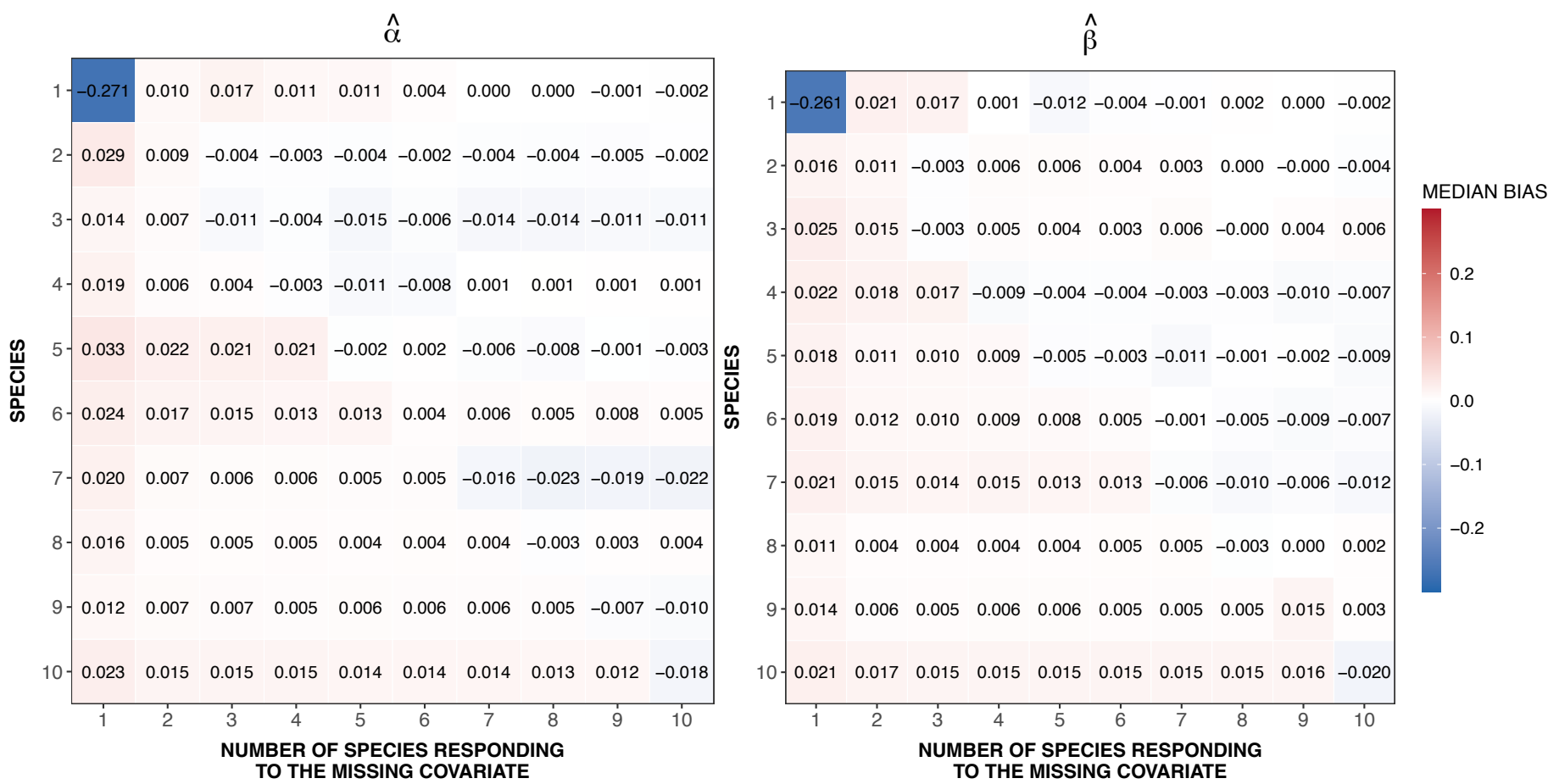
JSDB.1



RÉSULTATS

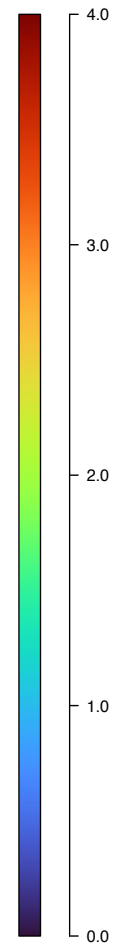
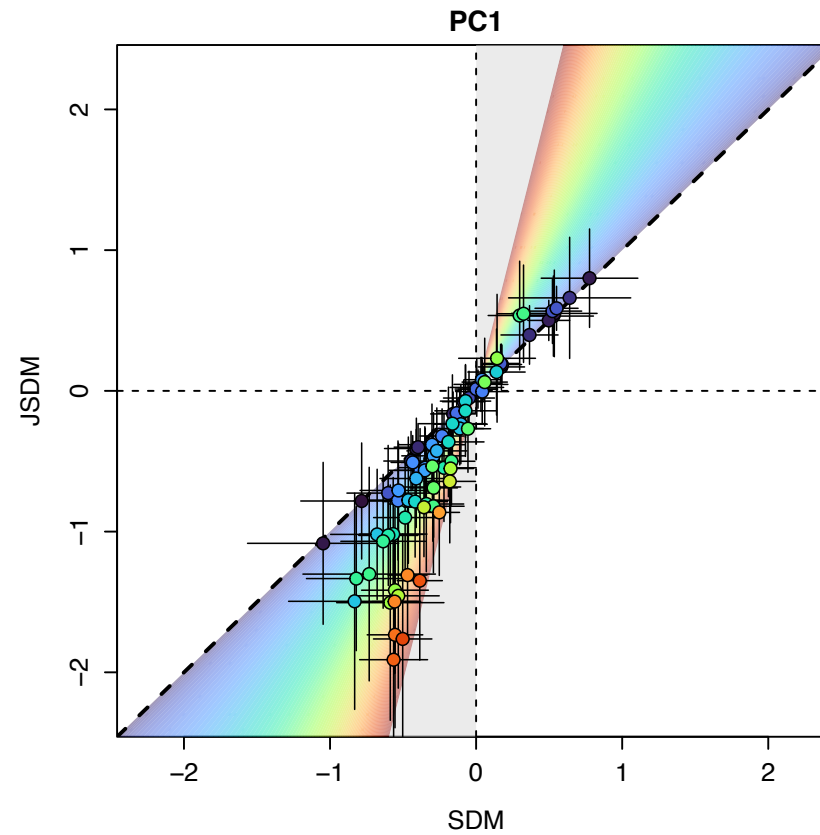
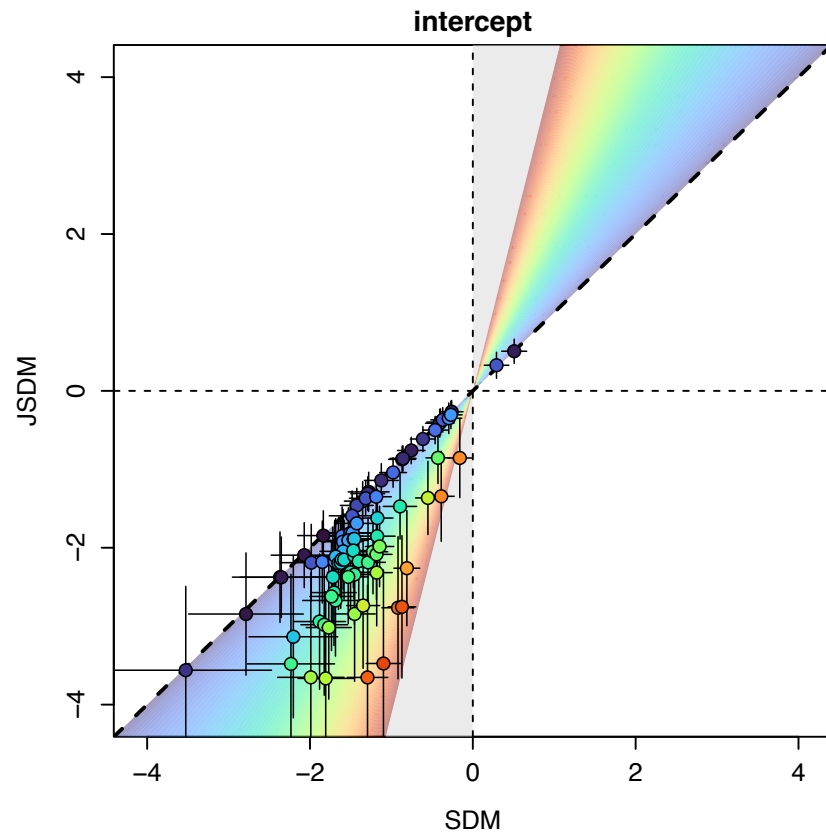
évaluation des JSDBs — influence du nombre d'espèces répondant à la covariable manquante

JSDB.3



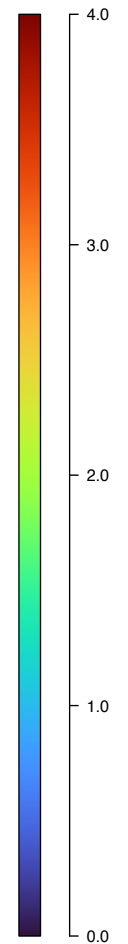
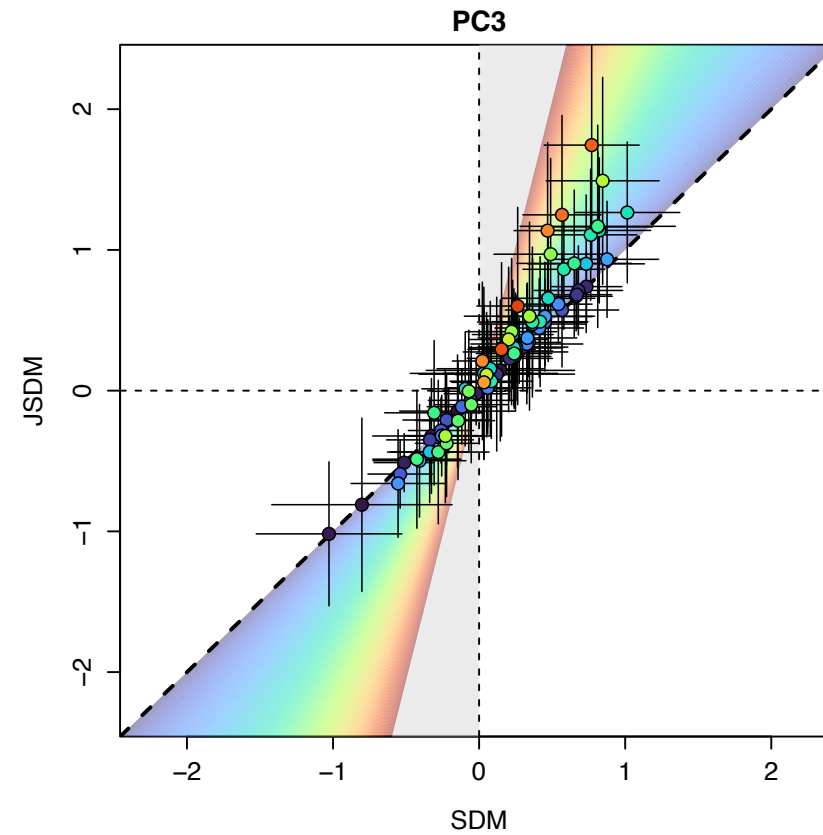
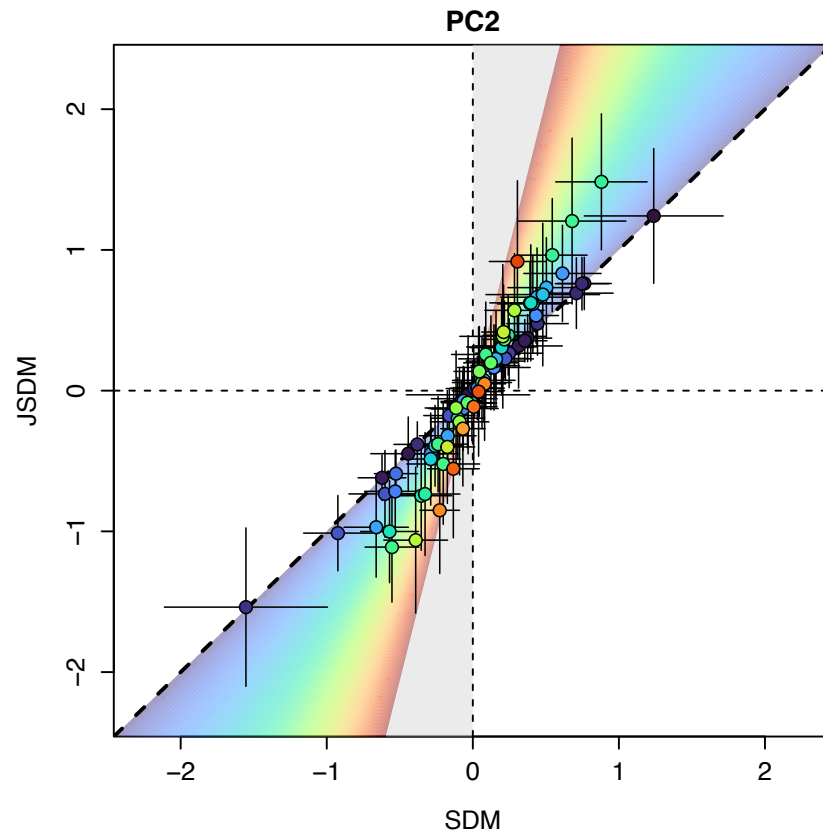
RÉSULTATS

jeux de données réel — le 1^{er} exemple



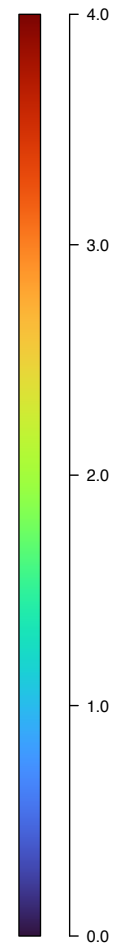
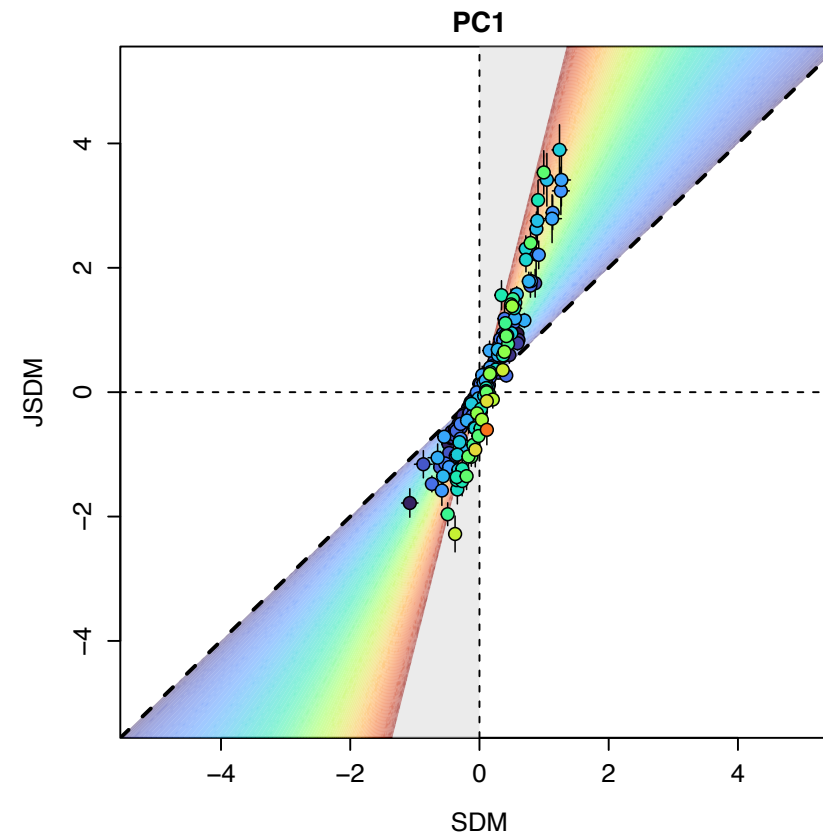
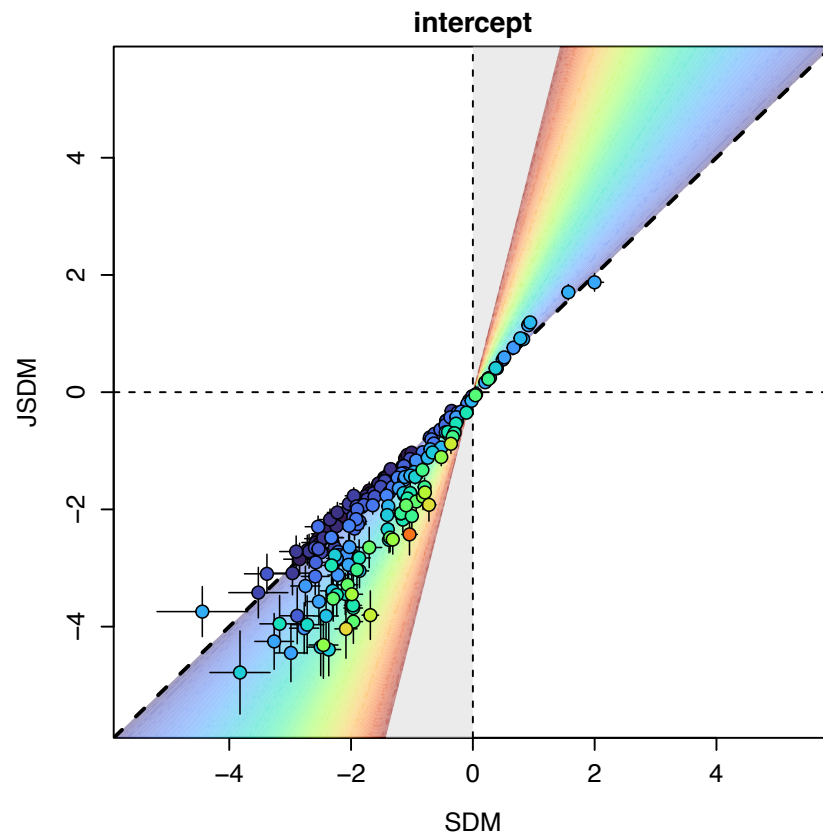
RÉSULTATS

jeux de données réel — le 1^{er} exemple



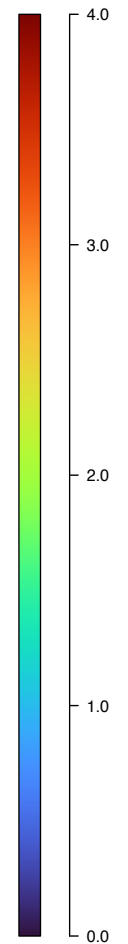
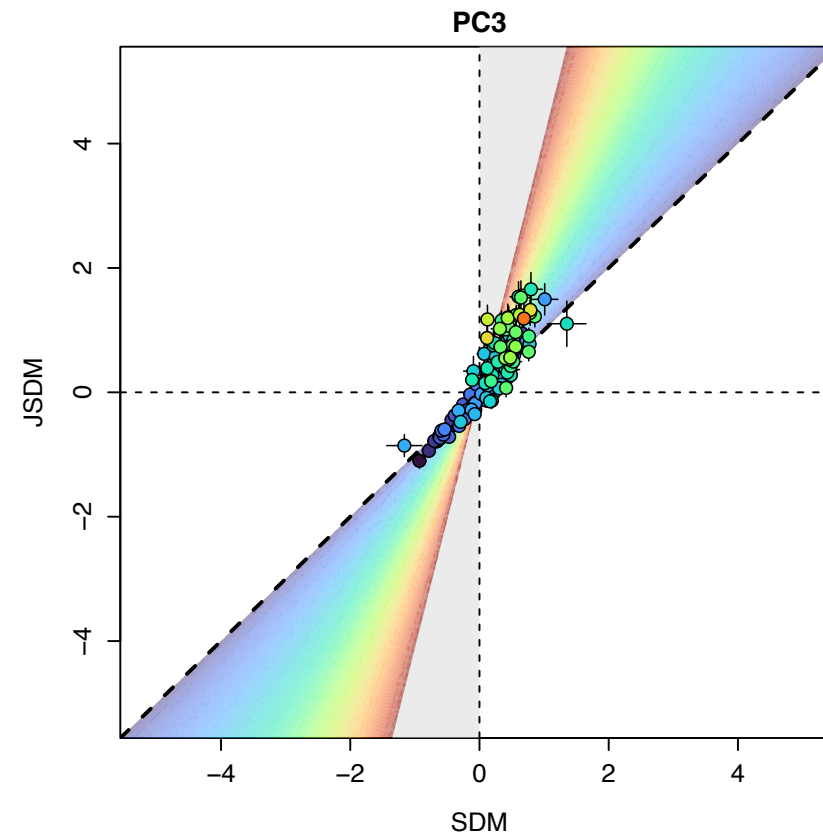
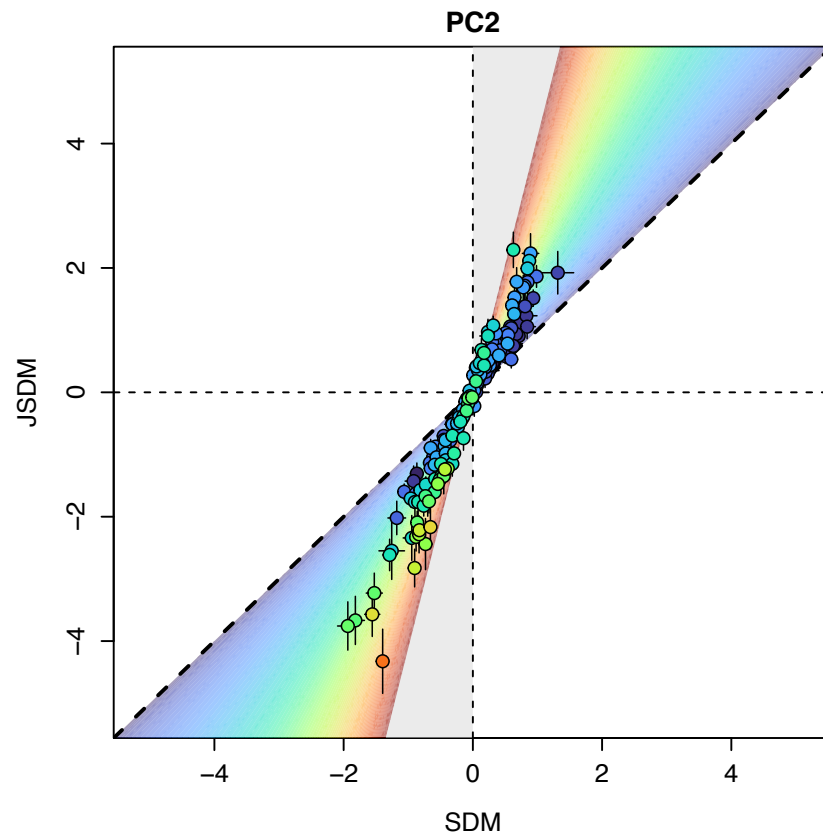
RÉSULTATS

jeux de données réel — le 2nd exemple



RÉSULTATS

jeux de données réel — le 2nd exemple



IV

discussion

DISCUSSION

principaux résultats issus des simulations

SDM

- modèle gaussien : pas de biais !
- modèle de Poisson : attention à l'intercept mais un effet aléatoire par observation solutionne
- modèle de Bernoulli : biais pour l'intercept et les coefficients de régression (atténuation), un effet aléatoire par observation ne résout pas toujours/pas bien

JSDM

- pour des données de présence-absence : biais et *coverage rates* corrigés
- différents JSDMs fournissent des performances similaires sur un même jeu de données
- la variable latente “reconstruit” la covariable manquante et le loading absorbe le coefficient associé

DISCUSSION

que tirer des cas réels ?

différences entre SDM et JSMD cohérentes avec ce que prédit une variable manquante et l'OVB

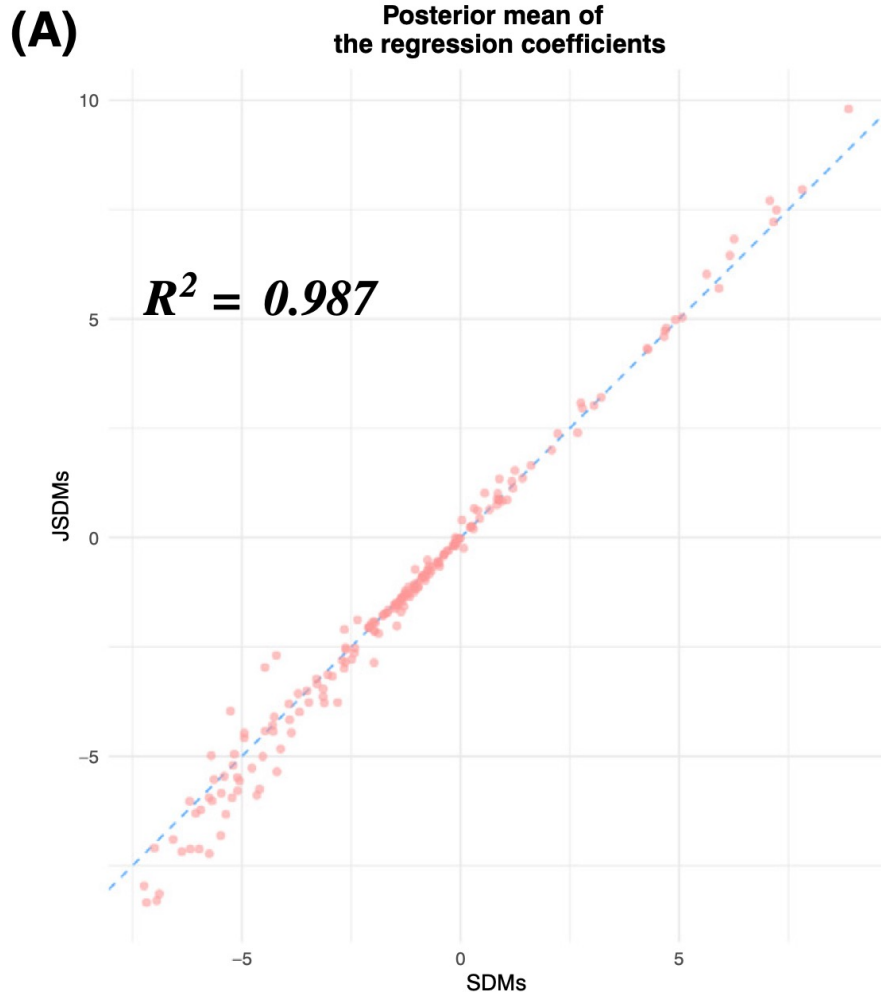
les JSMD augmentent la magnitude de l'effet des covariables observées

→ la “covariable manquante” = voie pertinente d'interprétation des différences entre paramètres estimés d'un JSMD et des SDM mono-espèce correspondants

/!\ la cause réelle n'est pas forcément une covariable manquante

DISCUSSION

insertion dans le débat scientifique des JSdM



un article de 2021 dans *Trends in Eco. & Evo.* :

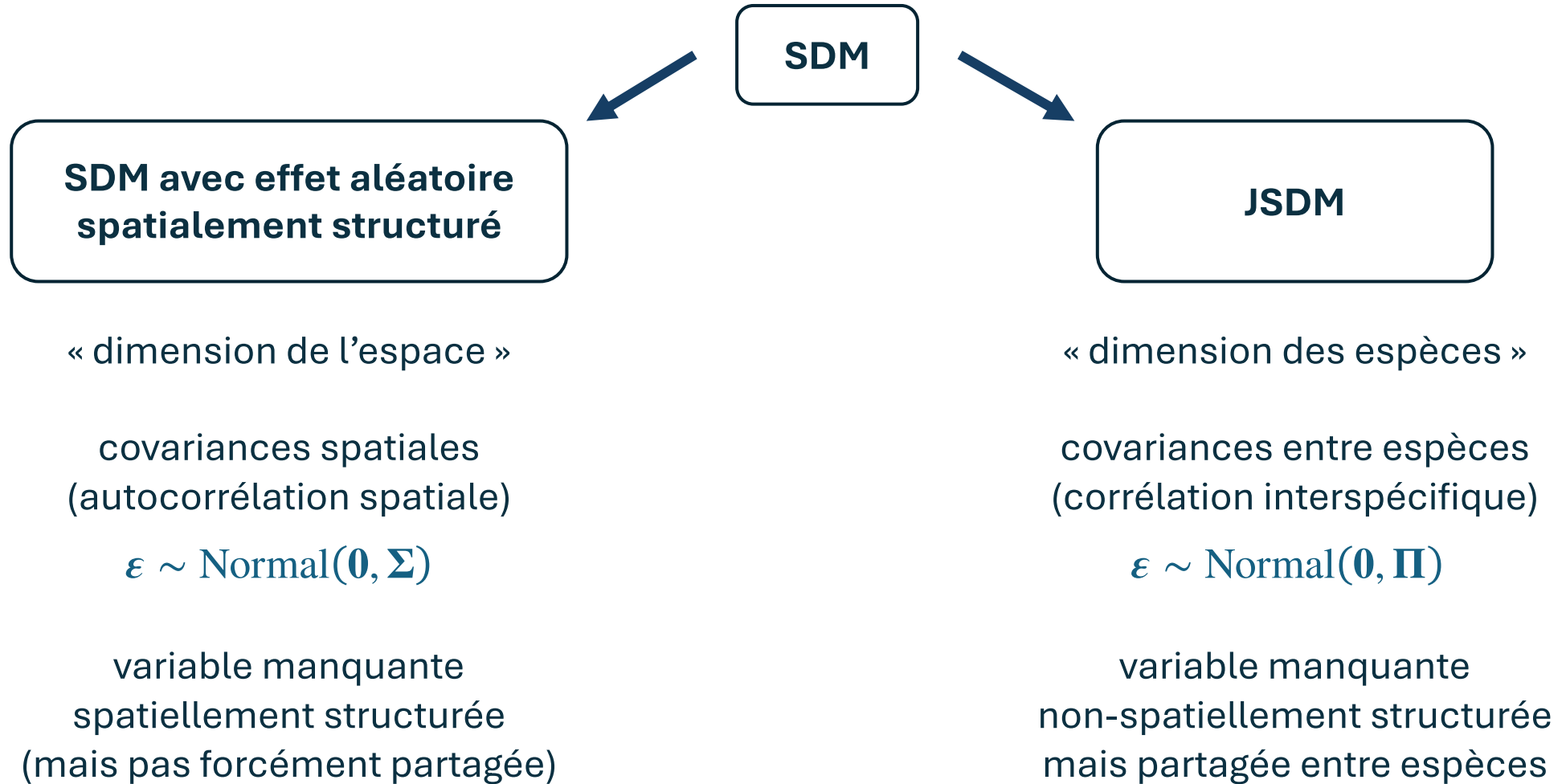
« non, les JSdM ne fournissent pas des estimateurs significativement différents de ceux des SDM »

→ les JSdM auraient un avantage limité

dans cette étude : oui, il y a des différences significatives, et elles peuvent s'expliquer par une (des) covariable(s) manquante(s)

DISCUSSION

comparaison heuristique entre SDM spatial et JSMD



DISCUSSION

il faut l'admettre : les limites de cette étude

- (1) d'autres types de modèles utilisés dans la littérature des SDM
 - modèles semi-paramétrique (GAMM) ou non-paramétriques (*machine/deep learning*)
- (2) une covariable manquante toujours non-spatialement structurée et indépendante des covariables incluses
 - en réalité, il y a de l'autocorrélation spatiale et du *spatial confounding*
- (3) la détection est supposée parfaite bien que l'on puisse modéliser également un processus d'observation

DISCUSSION

il faut l'admettre : les limites de cette étude (encore !)

- (4) les JSMD sont aussi connus pour « capturer » les interactions biotiques symétriques
 - conséquences de relations biotiques sur l'inférence des JSMD pas claires
- (5) pas d'outils à l'heure actuelle pour détecter à l'avance des covariables manquantes
 - piste des *goodness-of-fit* à investiguer
- (6) un coût en données et un coût computationnel significatifs
 - des pistes pour s'affranchir du coût computationnel avec les GPU

DISCUSSION

par-delà l'écologie et la conservation...

JSDM, c'est un nom qu'on donne en écologie

→ plus généralement : *generalized linear latent variable models* (GLLVM)

/!\ les présents résultats s'étendent immédiatement à ce type de modèle dans d'autres disciplines

imaginons :

- en médecine, une étude avec plusieurs patients affectés par différentes maladies, avec des sous-groupes de patients influencés par des caractéristiques inconnues
- en sciences politiques, les partis ont obtenu des voix dans différentes communes sous l'influence non mesurée de la stratification sociale
- mais aussi en économétrie, en sociologie, en génomique...

V

conclusion

CONCLUSION

qu'en est-il de la suite?

plusieurs extensions plus ou moins immédiates de ce travail :

- analyse systématique de la prévalence de décalages entre SDM et JSMD sur de nombreux jeux de données réels publiés
- évaluation de la capacité prédictive des JSMD par rapport aux SDM (à la fois théorique et sur des jeux de données réels)
- développement de *goodness-of-fit* pour les JSMD

CONCLUSION

quelques remerciements chaleureux



**MINISTÈRE
DE L'AGRICULTURE
ET DE LA SOUVERAINETÉ
ALIMENTAIRE**

*Liberté
Égalité
Fraternité*



misole

4
ABIMS

CONCLUSION

éléments de bibliographie

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.

GUISAN, A., THUILLER, W., ZIMMERMANN, N.E. (2017). *Habitat Suitability and Distribution Models: With Applications in R* (1st ed.). Cambridge University Press.

AUSTIN, M.P. (2007). Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling*, 200(1), 1–19.

LEE-YAW, J.A., L. McCUNE, J., PIRONON, S., N. SHETH, S. (2022). Species distribution models rarely predict the biology of real populations. *Ecography*, 2022(6), e05877.

GAIL, M.H., WIEAND, S., & PIANTADOSI, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71(3), 431–444.

MOOD, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review*, 26(1), 67–82.

YATCHEW, A., GRILICHES, Z. (1985). Specification error in probit models. *The Review of Economics and Statistics*, 67(1), 134.

ISHII, R., MARUO, K., GOSHO, M. (2022). Effect of Covariate Omission in Randomised Controlled Trials: A Review and Simulation Study. *International Statistical Review*, 90(1), 100–117.

OVASKAINEN, O., & ABREGO, N. (2020). *Joint Species Distribution Modelling: With Application in R*. Cambridge University Press.

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.

ABREGO, N., NIITYNEN, P., KEMPPINEN, J., OVASKAINEN, O. (2025). Joint species-trait distribution modeling: The role of intraspecific trait variation in community assembly. *Ecology*, 106(9), e70174.

HARRIS, D.J. (2015). Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution*, 6(4), 465–473.

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.

ROSSIGNOL, A., GOSSELIN, F. (2026). Addressing Missing Covariates in Species Distribution Models: Inferential Impacts and Mitigation via Joint Species Distribution Models. *EcoEvoRxiv*, preprint.

Merci pour votre attention !

Avez-vous des questions ?