Burn-in *vs* Warmup

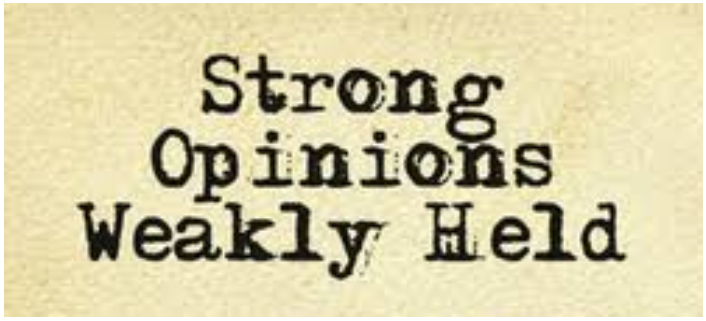
ou les divaguations d'un écologiste avec le posterior entre deux chaises

Matthieu Authier

Observatoire PELAGIS UMS-CNRS 3462

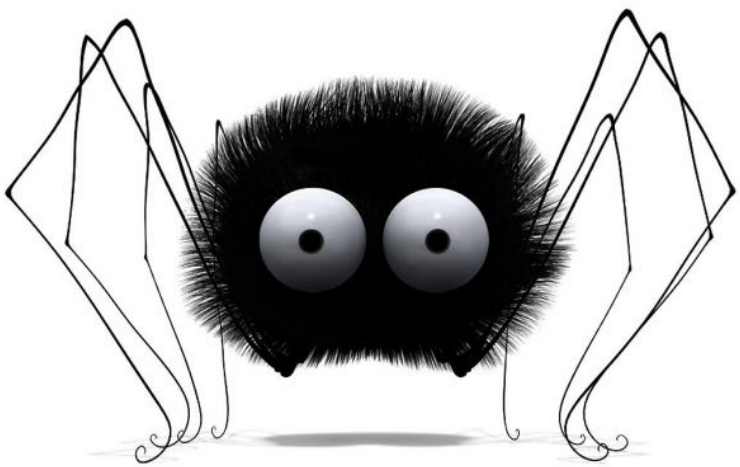
12 Juin 2020

Disclaimer



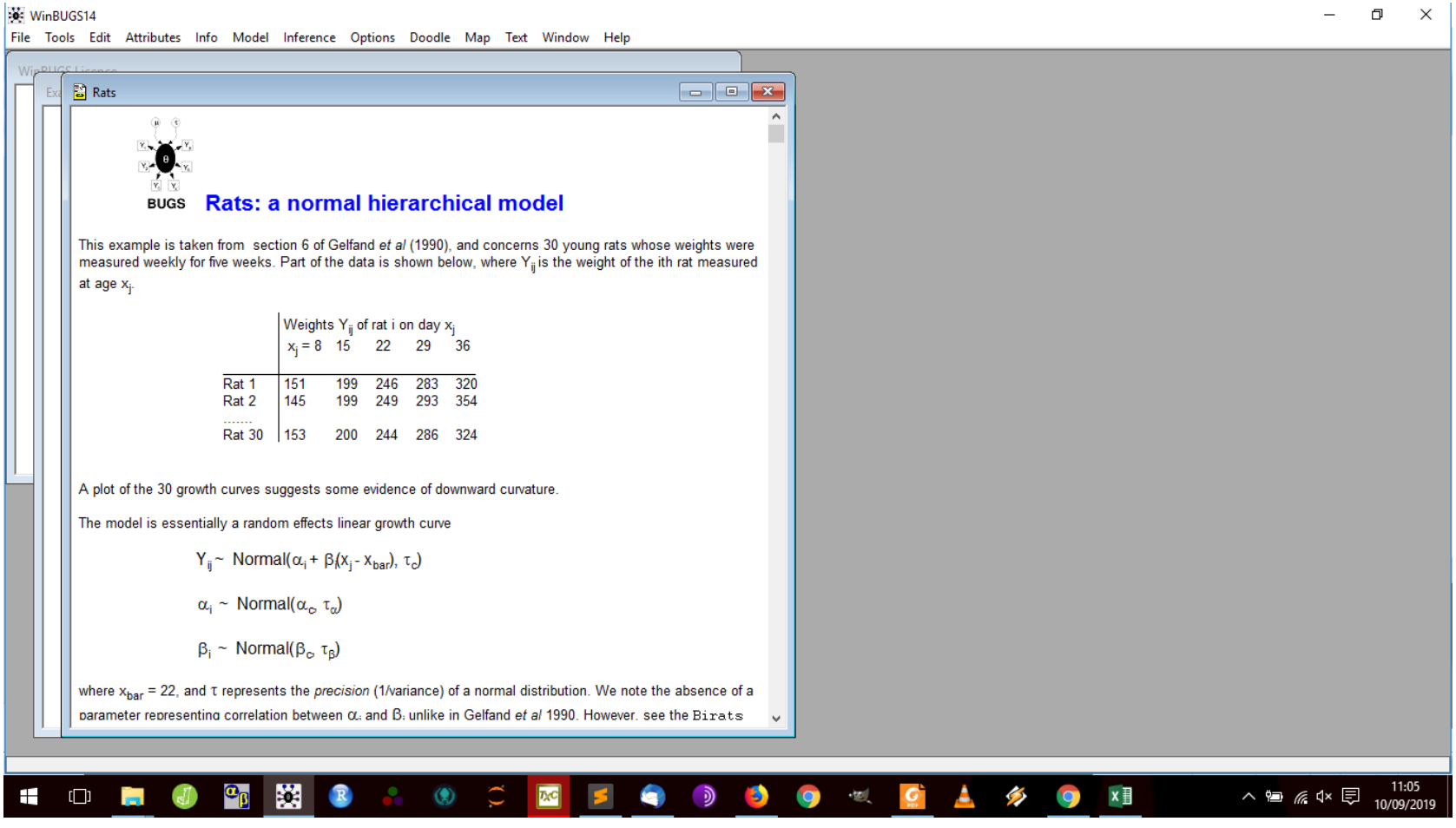
à l'origine...

BUGS (Bayesian inference Using Gibbs Sampling) project is concerned with flexible software for the Bayesian analysis of complex statistical models using Markov chain Monte Carlo (McMC) methods. The project began in 1989 in the MRC Biostatistics Unit, Cambridge.

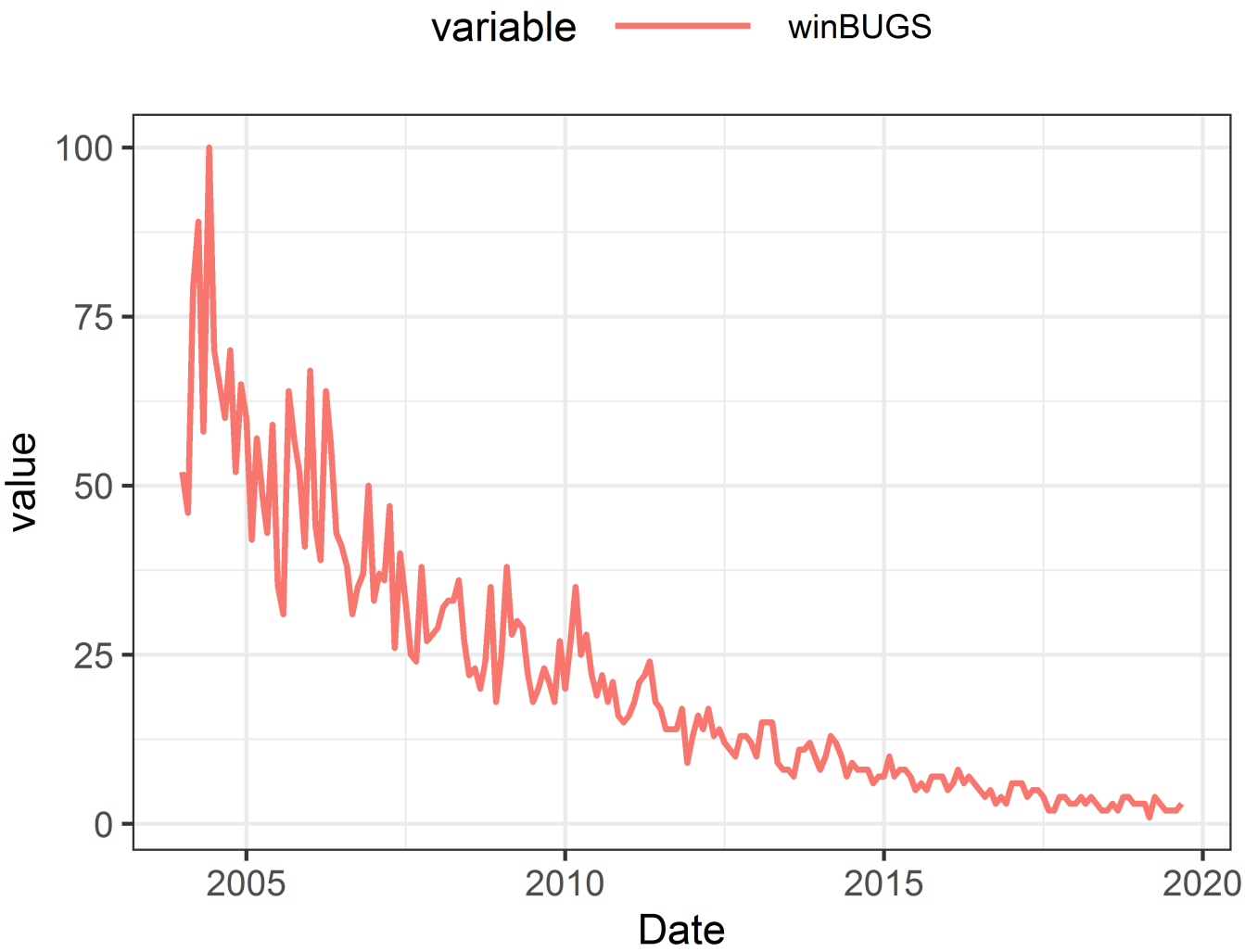


winBUGS

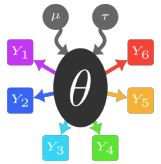
1989-2007: "Nous étions heureux" (E. Parent 2019 à l'EC BioBAYES)



Google trends



BUGS et ses rejetons



Nom	algorithme
`BUGS`	Gibbs, MH, slice
`JAGS`	Gibbs, MH, slice, etc.
`ADMB`	automatic differention model builder + Laplace approx.
`nimble`	McMC, SMC, etc.
`Stan`	HMC, etc.

Andrew Gelman's Juju

`https://www.youtube.com/watch?v=pWow8Qe1snQ`

et les autres...

Nom	algorithme
`ELFI`	ABC, HMC
`Edwards`	many available
`Birch`	SMC
`INLA`	Laplace approx.
`Greta`	HMC, MH, Slice (utilise `Tensor flow`)

Ressources

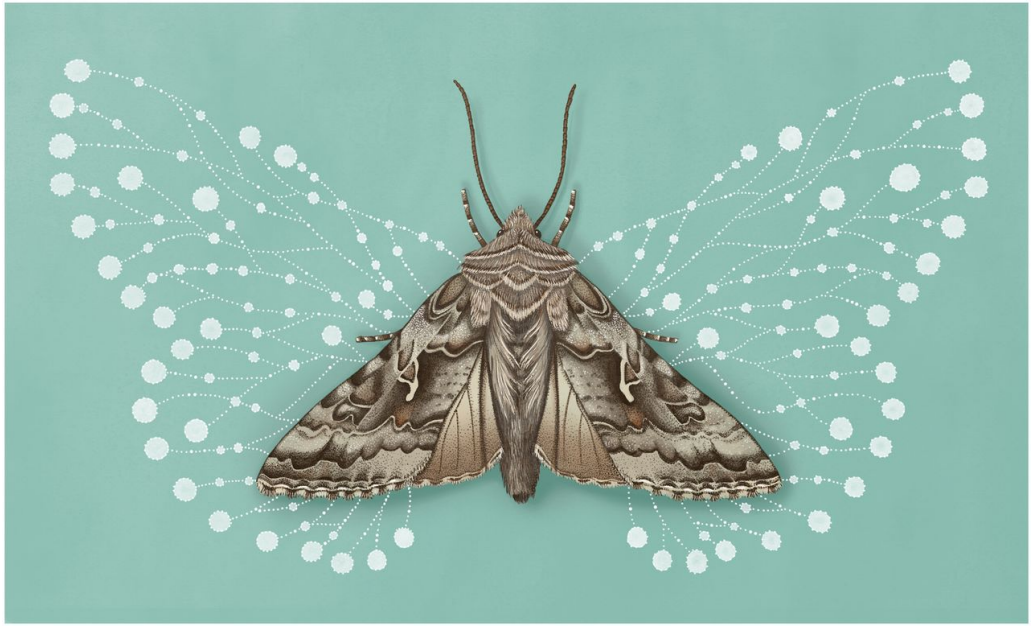
- BUGS: https://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/
- JAGS: http://mcmc-jags.sourceforge.net/
- AMBD: http://www.admb-project.org/
- INLA: http://www.r-inla.org/
- ELFI: https://elfi.readthedocs.io/en/latest/
- nimble: https://r-nimble.org/
- Stan: https://mc-stan.org/
- Edwards: http://edwardlib.org/
- Birch: https://birch-lang.org/
- Greta: https://greta-stats.org/

Que de choix ...

"Statistical Ecology Comes of Age" Gimenez et al. (2014) :
"we detect a recent rise in statistical awareness, manifested in various ways"

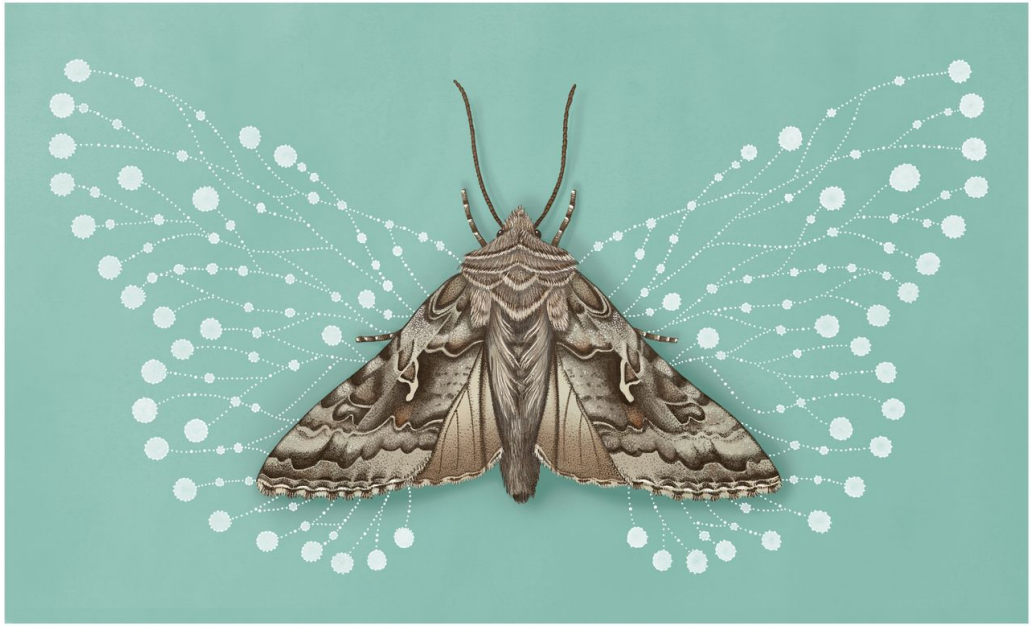
Que de choix ...

"Statistical Ecology Comes of Age" Gimenez et al. (2014) :
"we detect a recent rise in statistical awareness, manifested in various ways"



Que de choix ...

"Statistical Ecology Comes of Age" Gimenez et al. (2014) :
"we detect a recent rise in statistical awareness, manifested in various ways"



'burn-in' ou 'warmup' ?

'burn-in'

`http://users.stat.umn.edu/~geyer/mcmc/burn.html`
"Burn-in" is a colloquial term that describes the practice of throwing away some iterations at the beginning of an MCMC run

`http://www.catb.org/jargon/html/B/burn-in-period.html`

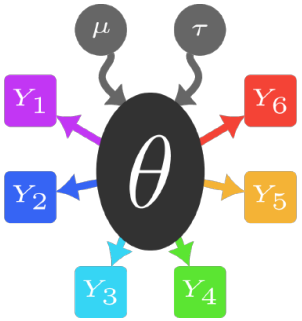
1. A factory test designed to catch systems with marginal components before they get out the door;
2. A period of indeterminate length in which a person using a computer is so intensely involved in his project that he forgets basic needs such as food, drink, sleep, etc.
   Warning: Excessive burn-in can lead to burn-out.

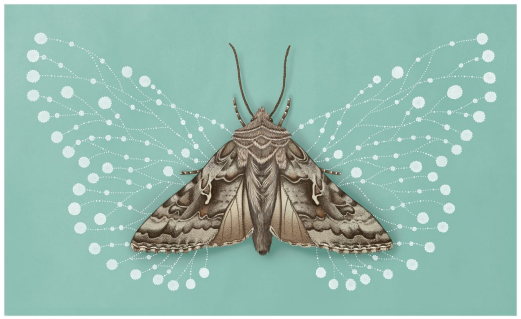
'warmup'

`https://www.merriam-webster.com/dictionary/warmup`

a preparatory activity or procedure

Pourquoi cette différence?





- 'burn-in' : on jette;
- 'warmup' : on ne jette pas vraiment $\longrightarrow$ phase adaptative des algorithmes modernes

Qu'est-ce qu'un algorithme?

Pour Wikipedia
"Un algorithme est une suite finie et non ambiguë d'opérations ou d'instructions permettant de résoudre une classe de problèmes."

Qu'est-ce qu'un algorithme?

Pour Wikipedia
"Un algorithme est une suite finie et non ambiguë d'opérations ou d'instructions permettant de résoudre une classe de problèmes."

Et pour un écologiste?

Qu'est-ce qu'un algorithme?

Pour Wikipedia
"Un algorithme est une suite finie et non ambiguë d'opérations ou d'instructions permettant de résoudre une classe de problèmes."

Et pour un écologiste?
`glm`, `gam` = algorithme (par exemple Hallgren et al., 2019)

Pourquoi en sommes-nous là?

Pourquoi en sommes-nous là?

La Guerre du 'SDM'

SDM - Species Distribution Models

- 'The goal of SDM is **to link** the location of species presences to some number p of environmental variables' (Renner & Warton, 2013) ;

SDM - Species Distribution Models

- 'The goal of SDM is **to link** the location of species presences to some number p of environmental variables' (Renner & Warton, 2013) ;
- 'SDMs relate observations of a species to environmental characteristics or spatial location **to better understand the processes** that determine where a species occurs' (Pacifici et al., 2017) ;

SDM - Species Distribution Models

- 'The goal of SDM is **to link** the location of species presences to some number p of environmental variables' (Renner & Warton, 2013) ;
- 'SDMs relate observations of a species to environmental characteristics or spatial location **to better understand the processes** that determine where a species occurs' (Pacifici et al., 2017) ;
- 'SDMs are **statistical methods** that relate species information (either presence-only or presence-absence) to environmental variables **to infer** spatially explicit habitat suitability (Zurell et al., 2020) .

SDM

Predire à partir de p covariables $(x_{k_{\in[1:p]}})$ la distribution ou l'abondance (Y) d'une espèce

$$\mathbb{E}[\text{Variable Réponse}] = f(\text{Variables Environnementales})$$

SDM

Predire à partir de p covariables $(x_{k_{\in[1:p]}})$ la distribution ou l'abondance (Y) d'une espèce

$\mathbb{E}[\text{Variable Réponse}] = f(\text{Variables Environnementales})$

Un SDM = formule mathématique (une "spécification") de la Fonction d'Espérance Conditionnelle:

SDM

Predire à partir de p covariables $(x_{k_{\in[1:p]}})$ la distribution ou l'abondance (Y) d'une espèce

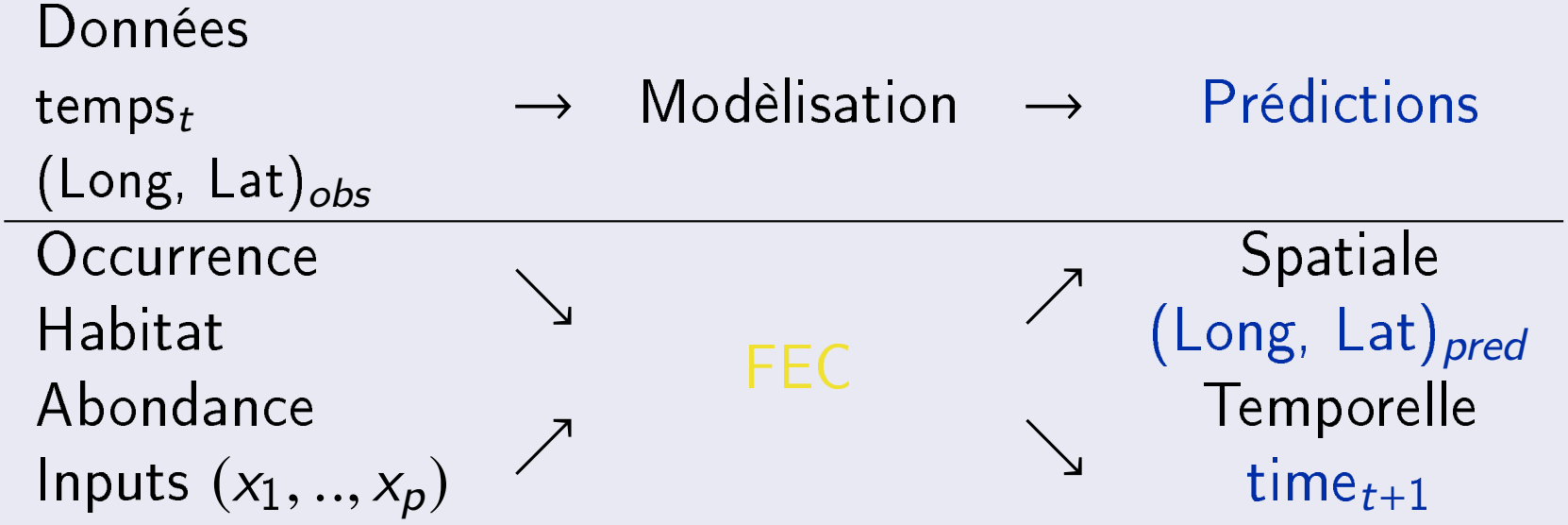
$\mathbb{E}[\text{Variable Réponse}] = f(\text{Variables Environnementales})$

Un SDM = formule mathématique (une "spécification") de la Fonction d'Espérance Conditionnelle:

1. reg. linéaire FEC: $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}] = \beta_0 + \sum_{k=1}^{p} \beta_k \times x_k$
2. reg. logistique FEC: $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}] = \dfrac{1}{1 + e^{-\beta_0 - \sum_{k=1}^{p} \beta_k \times x_k}}$
3. modèle additif généralisé FEC: $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}] = \beta_0 + \sum_{k=1}^{p} s_k(x_k)$
4. *etc...*

SDM

Flux de travail



SDM - Species Distribution Models

En 2006, 'publication' de `MaxEnt` ('Maximum Entropy') pour estimer des distributions d'espèces à partir de données de détections ($Y = \{1, 1, \dots\}$)
$\longrightarrow$ application java

SDM - Species Distribution Models

En 2006, 'publication' de `MaxEnt` ('Maximum Entropy') pour estimer des distributions d'espèces à partir de données de détections ($Y = \{1, 1, \dots\}$)
$\longrightarrow$ application java

Présentation de `MaxEnt` comme la méthode la plus performante et juste, en opposition à toutes les autres méthodes d'estimation;

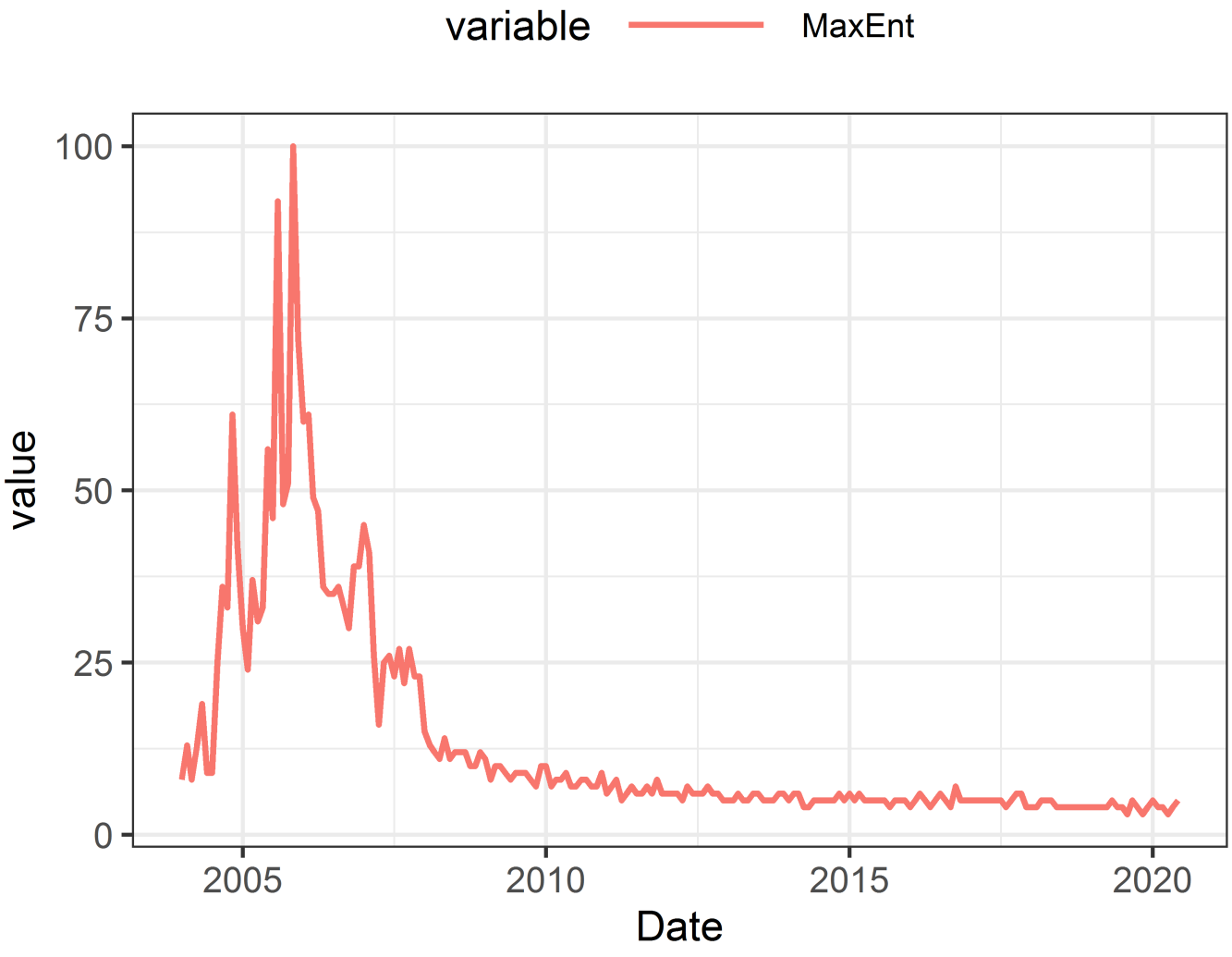
SDM - Species Distribution Models

En 2006, 'publication' de `MaxEnt` ('Maximum Entropy') pour estimer des distributions d'espèces à partir de données de détections ($Y = \{1, 1, \dots\}$)
$\longrightarrow$ application java

Présentation de `MaxEnt` comme la méthode la plus performante et juste, en opposition à toutes les autres méthodes d'estimation;

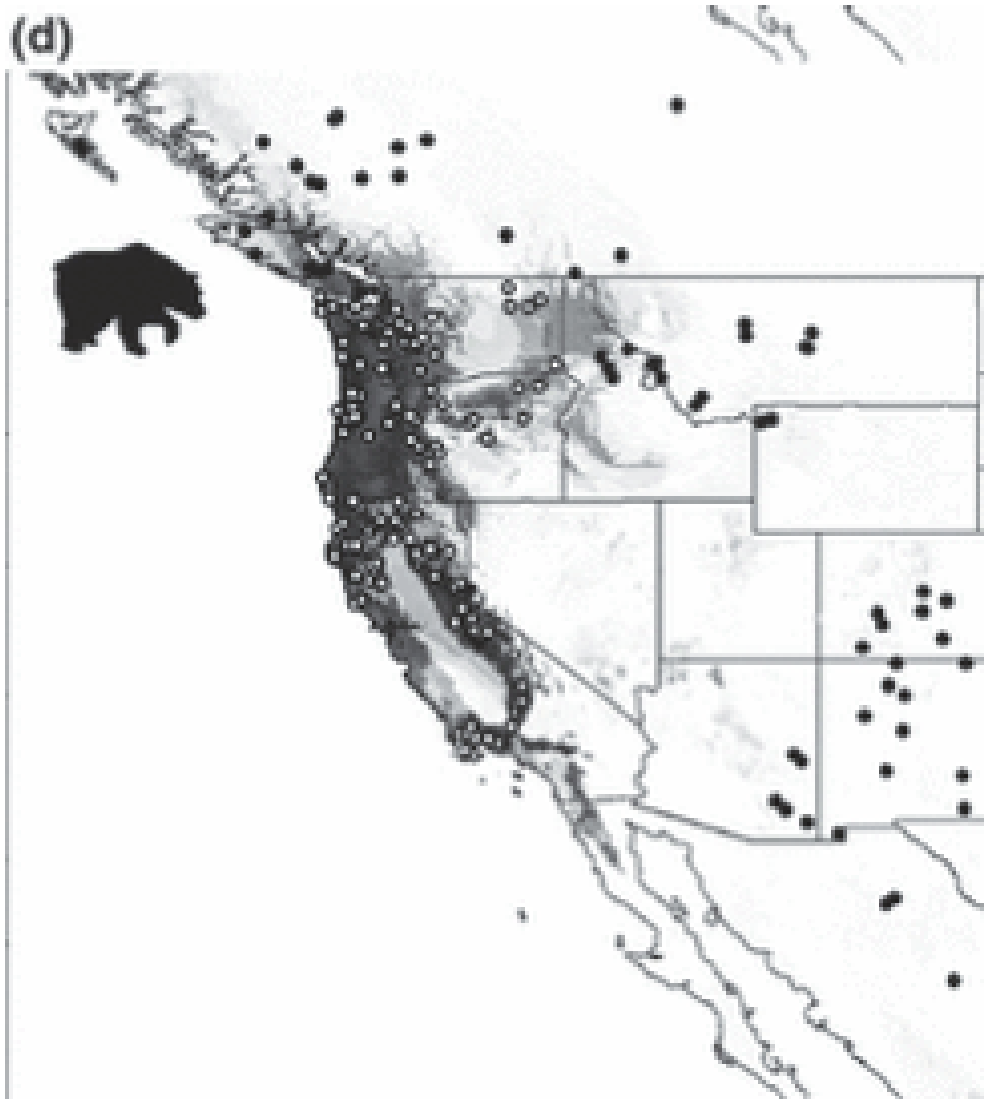
Adoption rapide et multiplication des cas d'études

Google trends



Lozier et al. (2009)

Lozier et al. (2009)



SDM - Species Distribution Models

Royle et al. (2012)

"We are critical of the routine application of the software package `MaxEnt` as applied to species distribution modelling. We specifically object to the pervasive views in the `MaxEnt` user community that one should avoid characterizing species distribution by occurrence probability, that occurrence probability is not identifiable and that one should instead obtain indices of species occurrence probability by using `MaxEnt`."

SDM - Species Distribution Models

Royle et al. (2012)

"We are critical of the routine application of the software package `MaxEnt` as applied to species distribution modelling. We specifically object to the pervasive views in the `MaxEnt` user community that one should avoid characterizing species distribution by occurrence probability, that occurrence probability is not identifiable and that one should instead obtain indices of species occurrence probability by using `MaxEnt`."

"In our view, poorly motivated and justified technical elements of `MaxEnt` distract from understanding the central inference problem of species distribution modelling."'

SDM - Species Distribution Models

Royle et al. (2012)

"We consider a formal model-based approach to analysis of presence-only data. We emphasize the critical assumption required for statistical inference about species occurrence probability from presence-only data, which is random sampling of space as a basis for accumulating presence-only observations. In addition, the estimator we devise here is most relevant only when species detection probability is constant."

SDM - Species Distribution Models

Royle et al. (2012)

"We consider a formal model-based approach to analysis of presence-only data. We emphasize the critical assumption required for statistical inference about species occurrence probability from presence-only data, which is random sampling of space as a basis for accumulating presence-only observations. In addition, the estimator we devise here is most relevant only when species detection probability is constant."

'Nouvelle' méthode : `MaxLike` ?

La guerre...

Fitzpatrick et al. (2013) : `MaxEnt` *vs* `MaxLike`

"the relative suitability index estimated by `MaxEnt` often was poorly correlated with the probability of occurrence estimated by `MaxLike`, suggesting that the two methods are estimating different quantities."

La guerre...

Fitzpatrick et al. (2013) : `MaxEnt` *vs* `MaxLike`

"the relative suitability index estimated by `MaxEnt` often was poorly correlated with the probability of occurrence estimated by `MaxLike`, suggesting that the two methods are estimating different quantities."

Renner & Warton (2013) établissent que ce que fait `MaxEnt` est d'estimer les paramètres d'un Processus (ponctuel) Poissonien non-homogène (IPP);

La guerre...

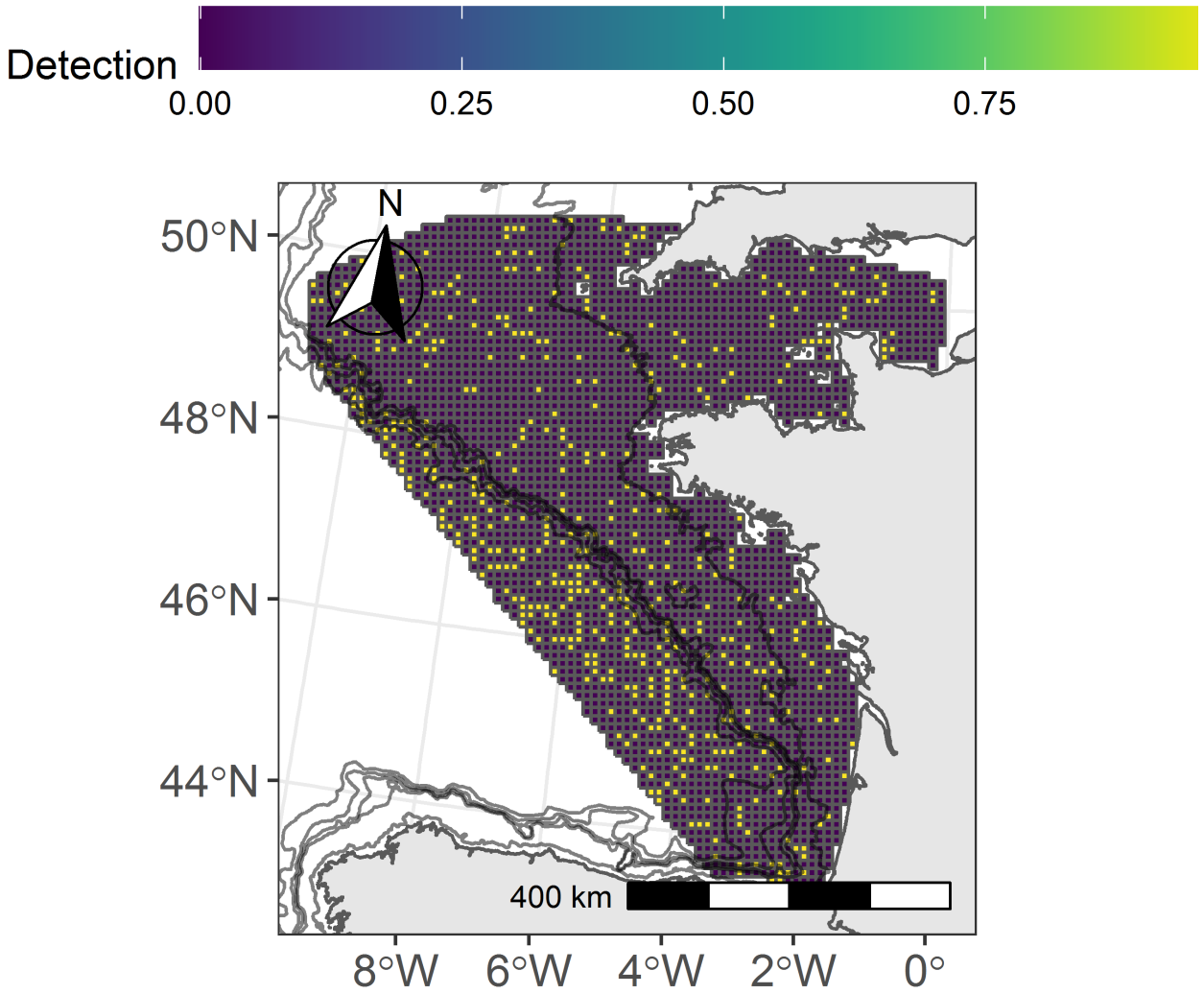
Fitzpatrick et al. (2013) : `MaxEnt` *vs* `MaxLike`

"the relative suitability index estimated by `MaxEnt` often was poorly correlated with the probability of occurrence estimated by `MaxLike`, suggesting that the two methods are estimating different quantities."

Renner & Warton (2013) établissent que ce que fait `MaxEnt` est d'estimer les paramètres d'un Processus (ponctuel) Poissonien non-homogène (IPP);

Gimenez et al. (2014) "Important innovations include the use of point processes to fit SDMs to presence-only data and the mathematical equivalence of `MaxEnt` to `glms`"

Un exemple: $n = 500,\ p = 4$



MaxLike

Vraisemblance Bernoulli, estimation avec `optim`

	β_1	β_2	β_3	β_4
vraie valeur	0.14	0.00	0.20	−0.35
estimation	0.16	0.04	0.20	−0.30

MaxEnt

Vraisemblance IPP, estimation avec `glm`

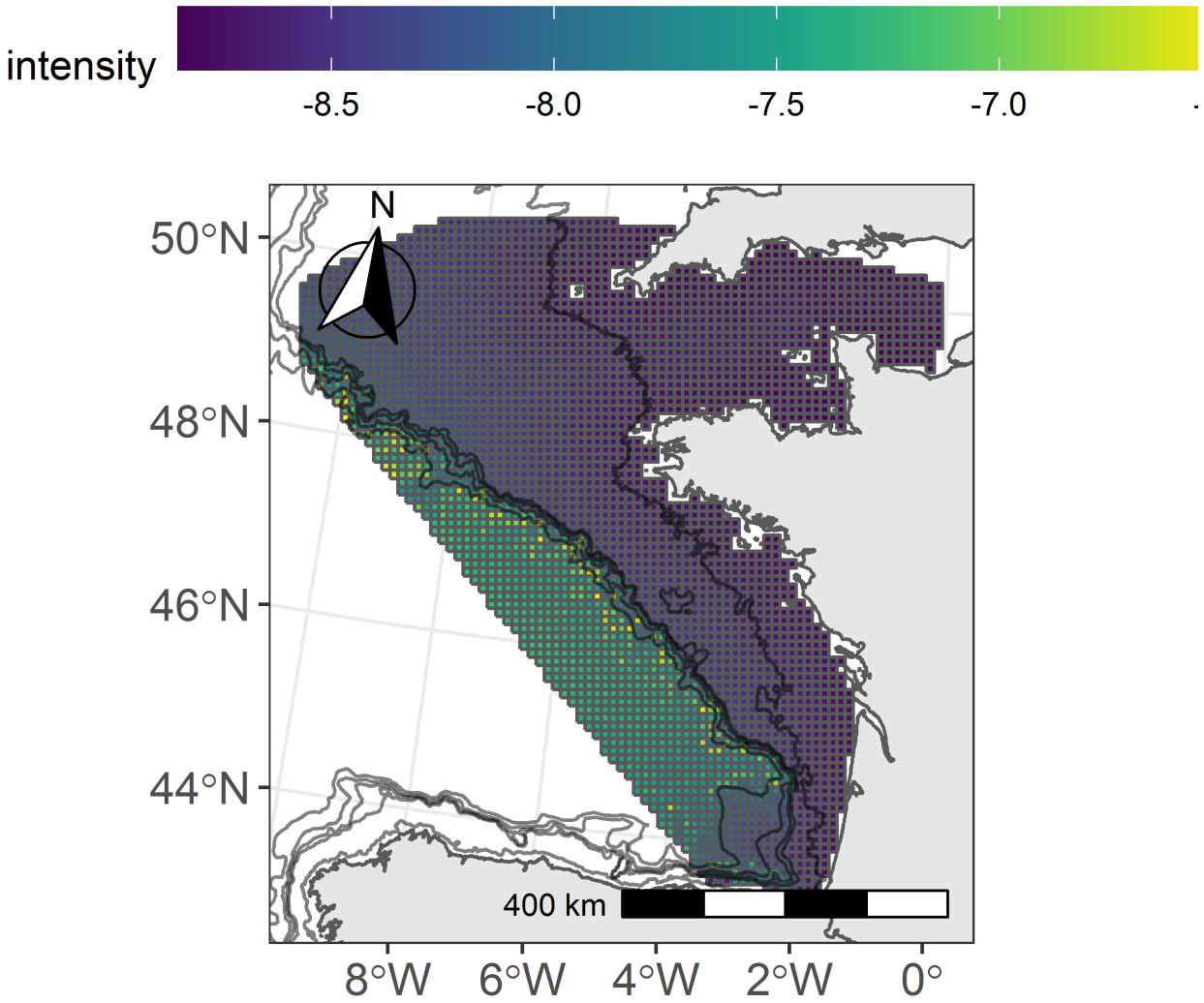
	β_1	β_2	β_3	β_4
vraie valeur	0.14	0.00	0.20	−0.35
estimation	0.17	−0.03	0.22	−0.33

MaxEnt

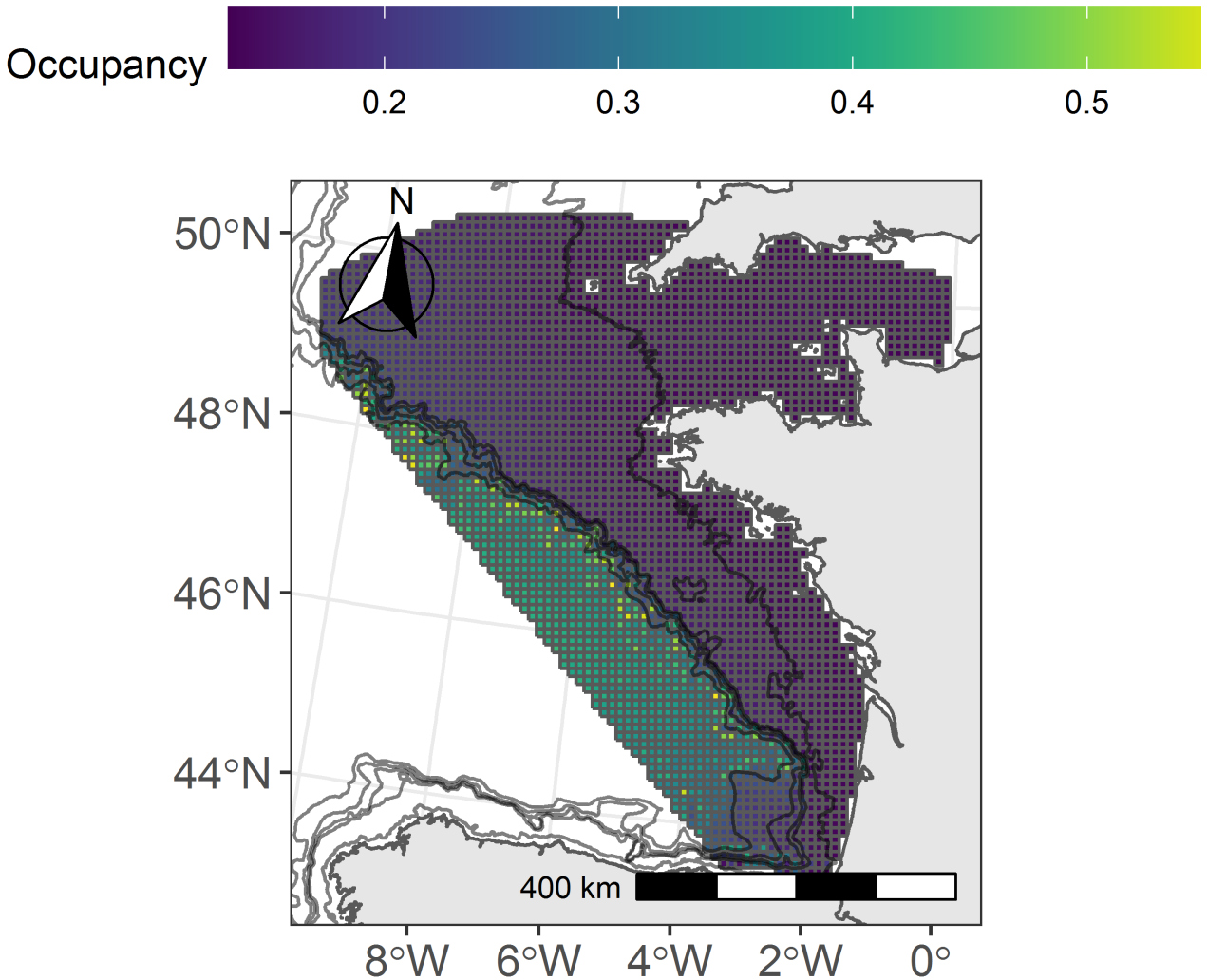
Vraisemblance IPP + regularisation ('horseshoe', Carvalho et al., 2010), estimation avec `Stan`

	β_1	β_2	β_3	β_4
vraie valeur	0.14	0.00	0.20	−0.35
estimation	0.16	−0.02	0.21	−0.32

IPP avec stan



Vraie carte



SDM - Species Distribution Models

Le programme de recherche en écologie sur les SDMs est confus

SDM - Species Distribution Models

Le programme de recherche en écologie sur les SDMs est confus

- `MaxEnt` est un modèle, pas un algorithme

SDM - Species Distribution Models

Le programme de recherche en écologie sur les SDMs est confus

- `MaxEnt` est un modèle, pas un algorithme
- `MaxEnt` est un modèle, avec des choix discutables mais obscurs pour la plupart des écologistes

SDM - Species Distribution Models

Le programme de recherche en écologie sur les SDMs est confus

- `MaxEnt` est un modèle, pas un algorithme
- `MaxEnt` est un modèle, avec des choix discutables mais obscurs pour la plupart des écologistes
- Fithian & Hastie (2013) "'infinitely weighted logistic regression': we can **use logistic regression as a device** for using `glm` software to maximize the IPP log-likelihood."

SDM - Species Distribution Models

Le programme de recherche en écologie sur les SDMs est confus

- `MaxEnt` est un modèle, pas un algorithme
- `MaxEnt` est un modèle, avec des choix discutables mais obscurs pour la plupart des écologistes
- Fithian & Hastie (2013) "'infinitely weighted logistic regression': we can **use logistic regression as a device** for using `glm` software to maximize the IPP log-likelihood."
- estimer le modèle supposé par `MaxEnt` est possible avec `glm`... ou `Stan`!

SDM - Species Distribution Models

Le programme de recherche en écologie sur les SDMs est confus

- Hallgren et al. (2019) (Ecological Modelling) 'comparent' `MaxEnt`, `glm`, `gam`, etc. et concluent 'SDMs can be highly sensitive to algorithm configuration'

SDM - Species Distribution Models

Le programme de recherche en écologie sur les SDMs est confus

- Hallgren et al. (2019) (Ecological Modelling) 'comparent' `MaxEnt`, `glm`, `gam`, etc. et concluent 'SDMs can be highly sensitive to algorithm configuration'
- '`MaxEnt` and `glm` have fewer configuration settings, but also showed less sensitivity to their settings. **This can be explained since these approaches are more model-based.**'

SDM - Species Distribution Models

Le programme de recherche en écologie sur les SDMs est confus

- Hallgren et al. (2019) (Ecological Modelling) 'comparent' `MaxEnt`, `glm`, `gam`, etc. et concluent 'SDMs can be highly sensitive to algorithm configuration'
- '`MaxEnt` and `glm` have fewer configuration settings, but also showed less sensitivity to their settings. **This can be explained since these approaches are more model-based.**'
- 'This sensitivity confirms that the samples are not exhaustive of the complex environment, and sampling effort was uneven'

Comment sortir de l'ornière?

- Présenter un modèle comme un algorithme ne rend pas service

Comment sortir de l'ornière?

- Présenter un modèle comme un algorithme ne rend pas service
- confusion entre le modèle et comment en estimer les paramètres

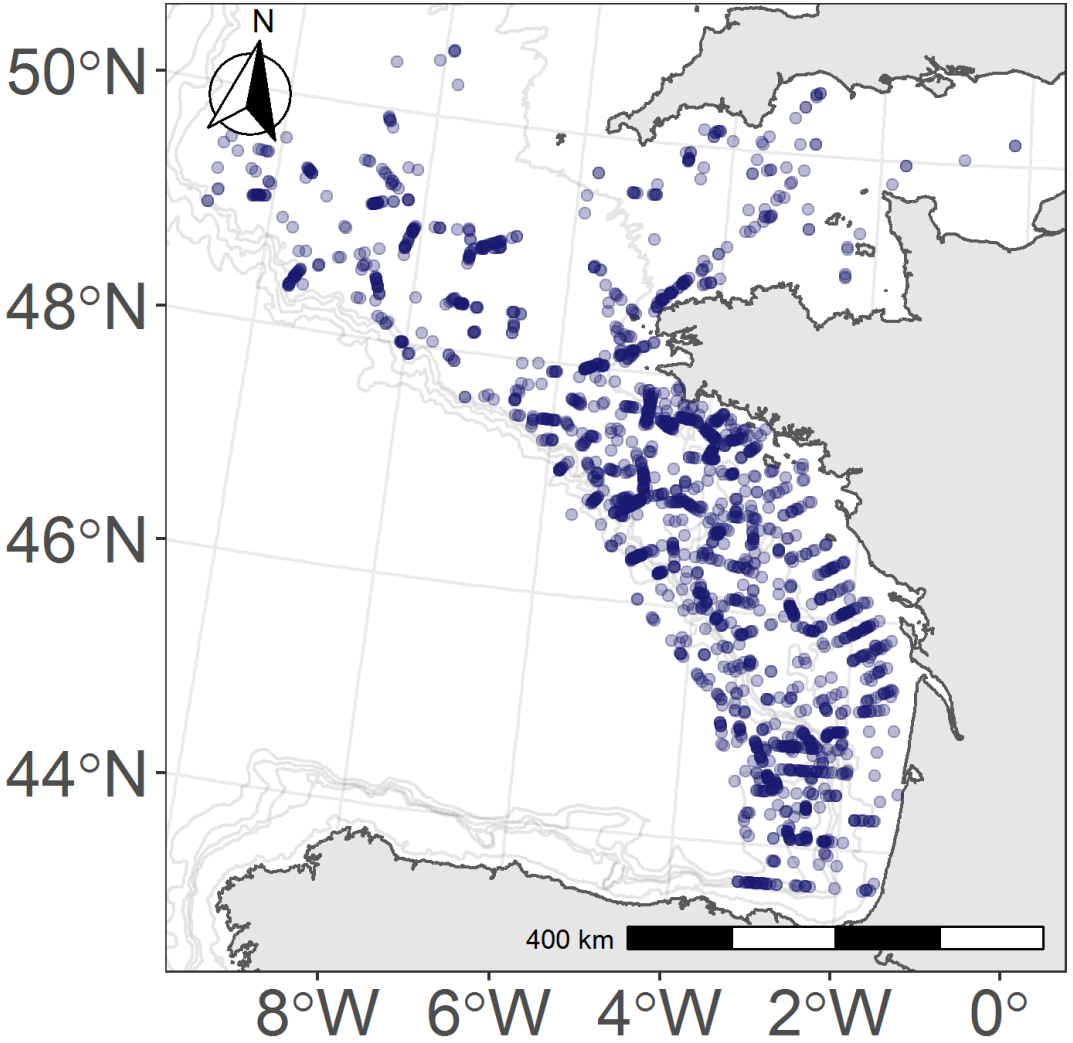
Comment sortir de l'ornière?

- Présenter un modèle comme un algorithme ne rend pas service
- confusion entre le modèle et comment en estimer les paramètres
- $\longrightarrow$ empêche de situer correctement les problèmes

Comment sortir de l'ornière?

- Présenter un modèle comme un algorithme ne rend pas service
- confusion entre le modèle et comment en estimer les paramètres
- $\longrightarrow$ empêche de situer correctement les problèmes
- est-ce le modèle? sont-ce les données?

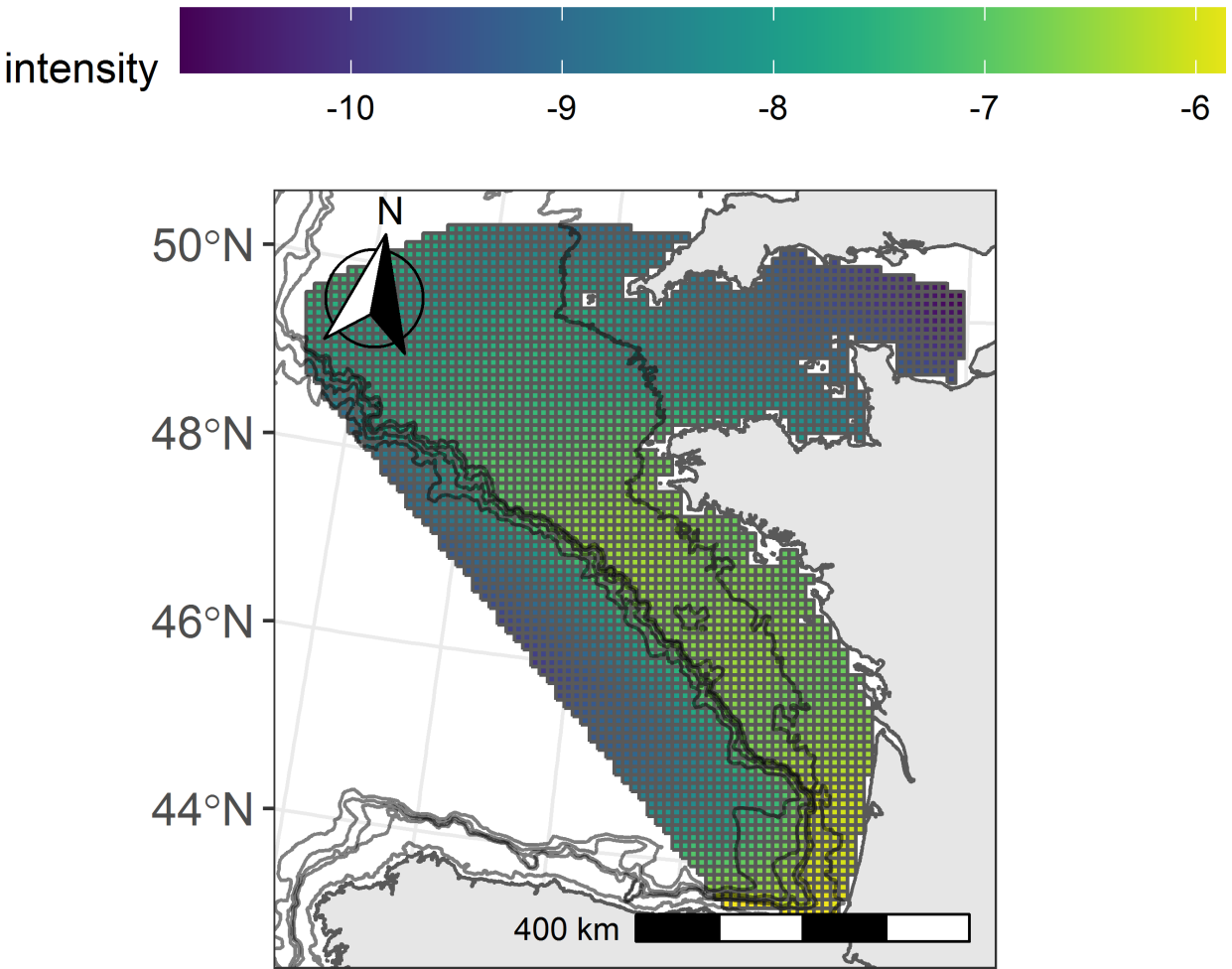
Un exemple: $n \approx 2,500$, $p = 4$



Où voir des dauphins dans le Golfe de Gascogne?

	Côte β_1	Talus β_2	Inclinaison β_3	Profondeur β_4
`MaxLike`	−1.03	−1.57	0.12	0.62
`MaxEnt + glm`	−0.56	−1.01	0.01	0.34
`MaxEnt + Stan`	−0.56	−1.01	0.01	0.33

Où (ne pas) voir des dauphins dans le Golfe de Gascogne?



Où voir le(s) problème(s)?

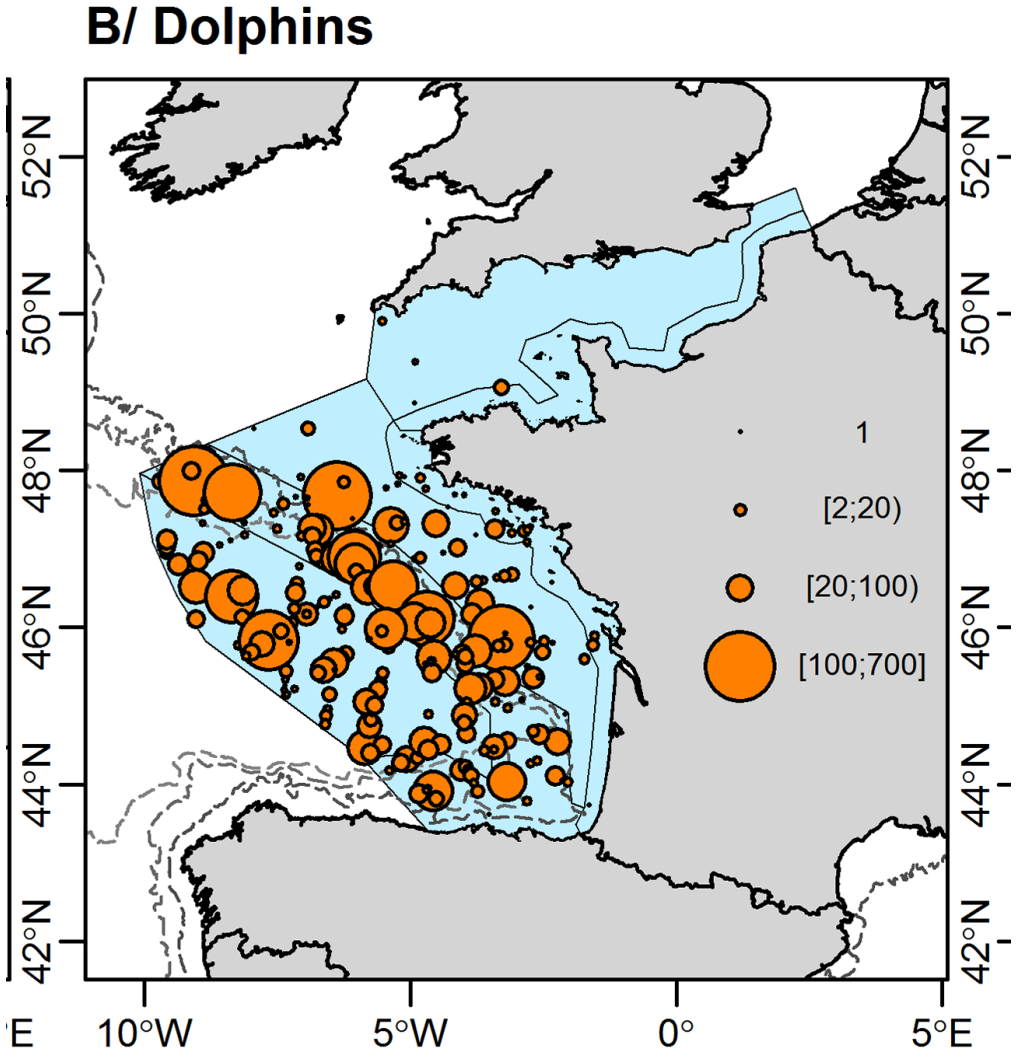
Les modèles sont cohérents entre eux: même signe et même rang pour les différents effets...

Où voir le(s) problème(s)?

Les modèles sont cohérents entre eux: même signe et même rang pour les différents effets...

inférences fausses!!

Effort homogène



Où voir le(s) problème(s)?

Les modèles sont cohérents entre eux: même signe et même rang pour les différents effets...

inférences fausses!!

Le problème vient de l'échantillonnage (des données, donc):
Botella et al. (2020)
"We modelled species occurrences and observation process as a thinned Poisson point process.
We conclude that none of the methods are immune to estimation bias."

A standard protocol for reporting SDMs

Zurell et al. (2020)

A standard protocol for reporting SDMs

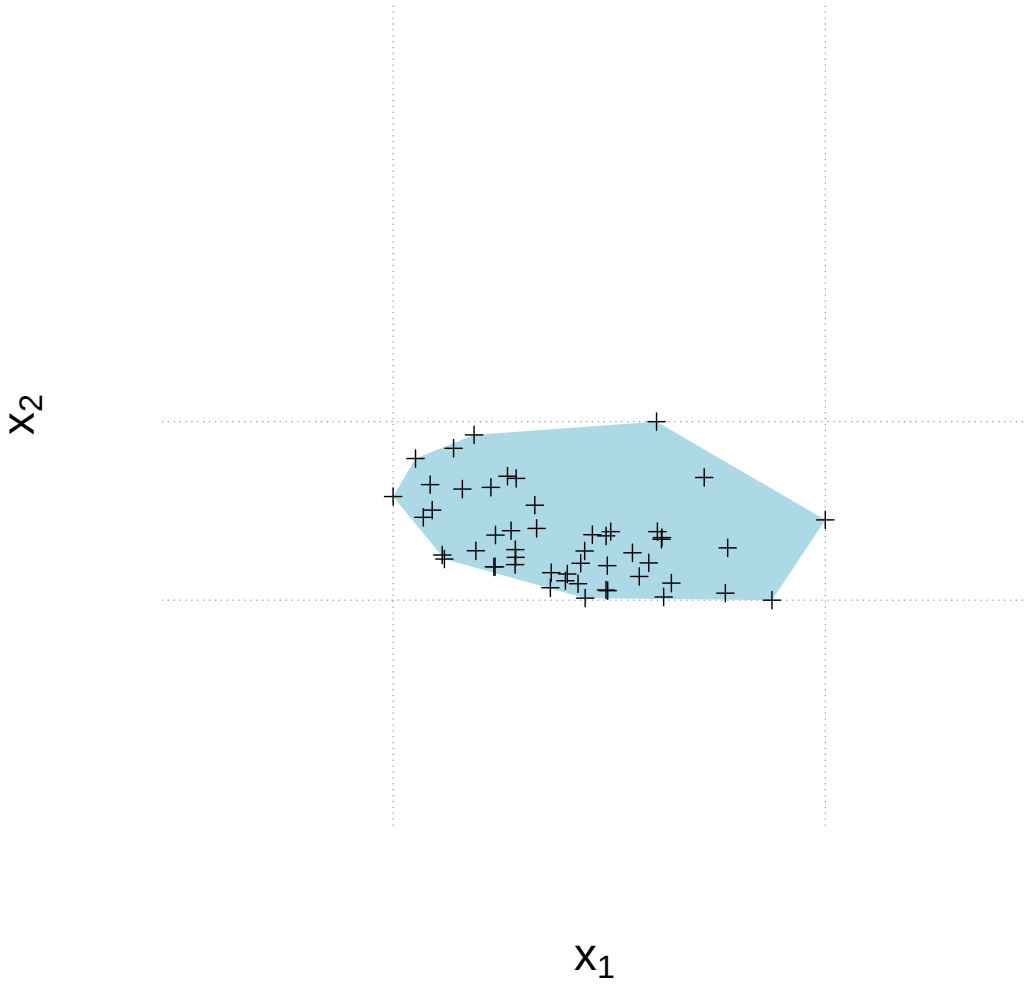
Zurell et al. (2020)

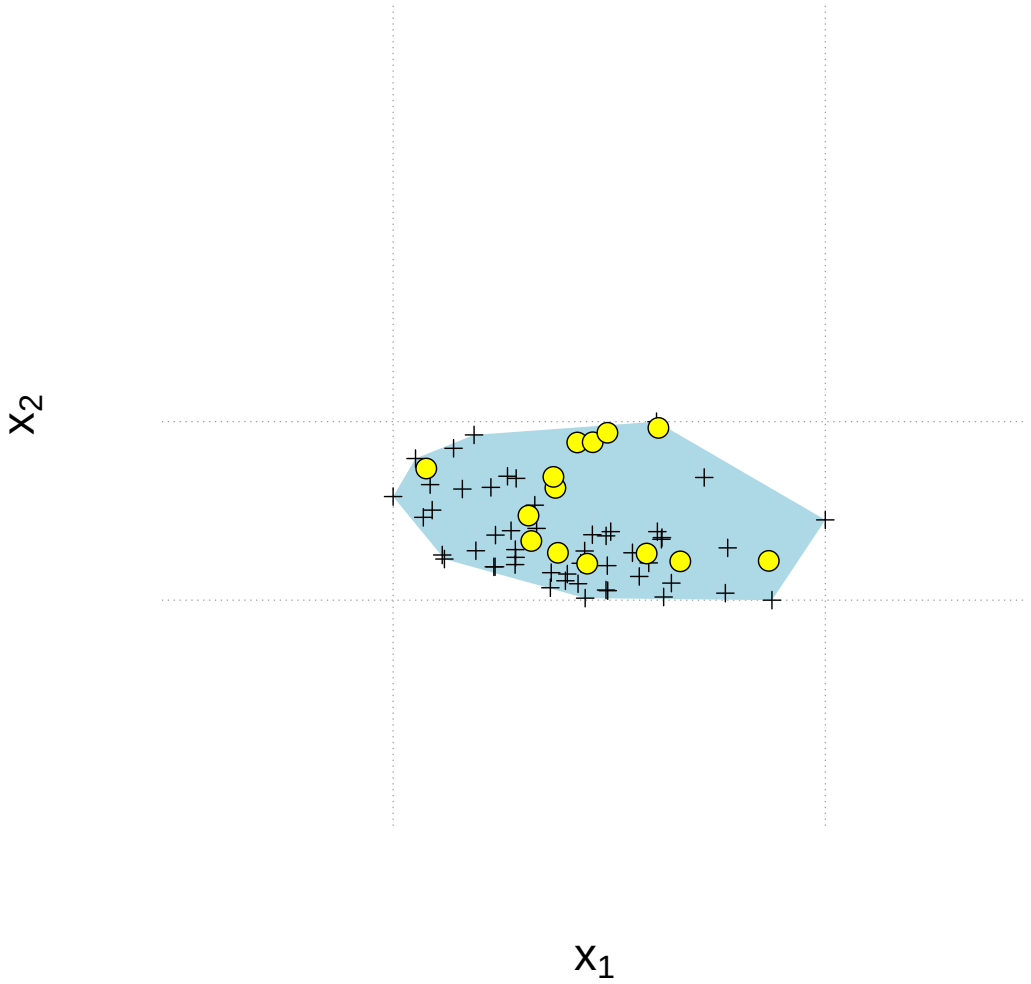
- 'Explanation' (also termed inference) regards detailed analyses of species-environment relationships and aims to provide or test specific hypotheses.

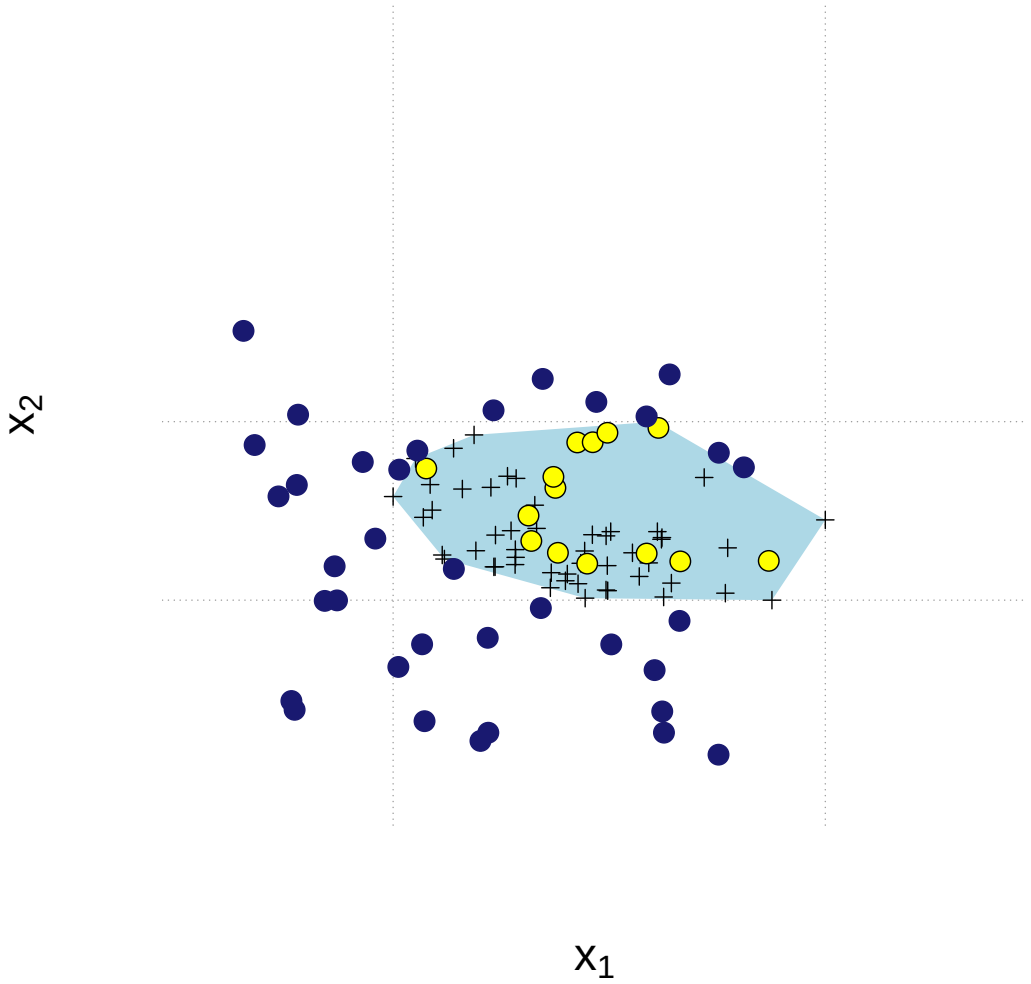
A standard protocol for reporting SDMs

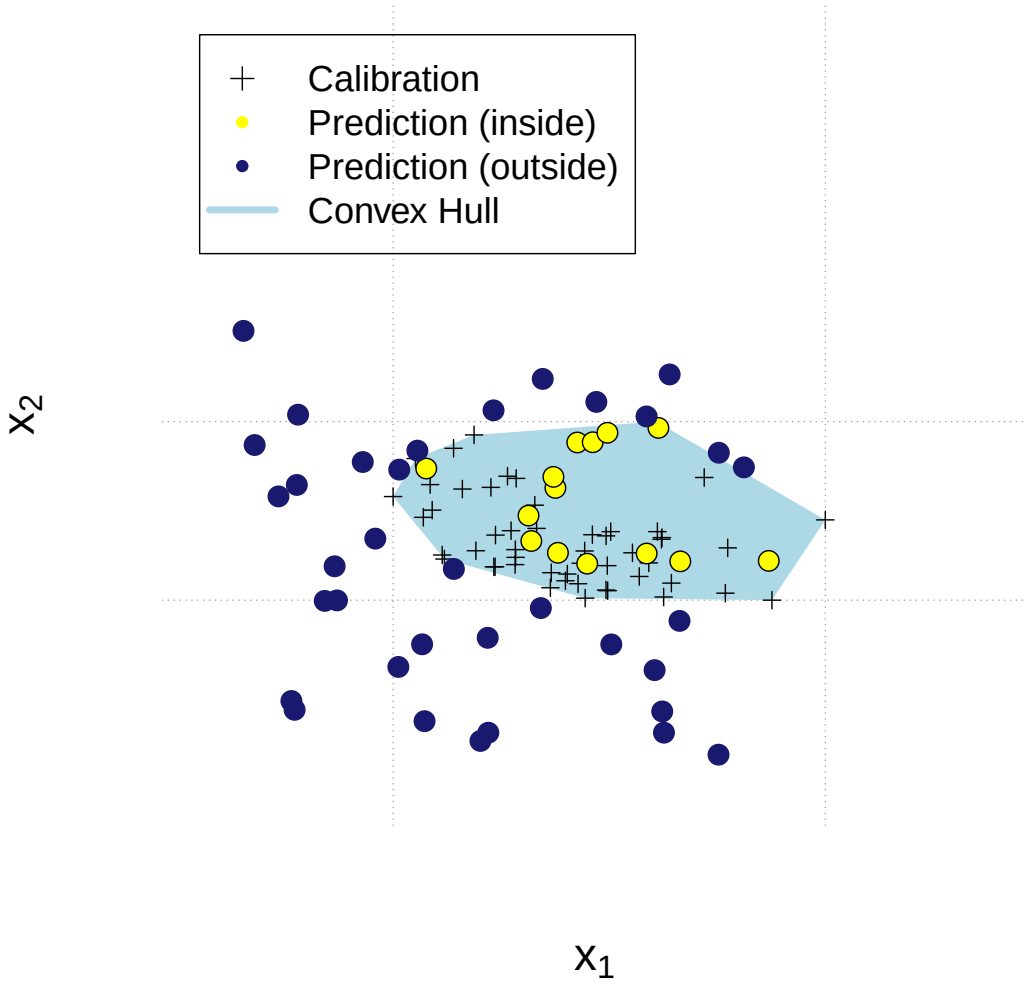
Zurell et al. (2020)

- 'Explanation' (also termed inference) regards detailed analyses of species-environment relationships and aims to provide or test specific hypotheses.
- 'Mapping' (also termed interpolation) means that the estimated species-environment relationships are used to map (or interpolate) the species distributions in the same geographic area and time period in which the model was calibrated.


x_2
x_1


x_2
x_1


x_2
x_1



King & Zeng (2007)

A standard protocol for reporting SDMs

Zurell et al. (2020)

- 'Transfer' (also termed forecast or projection; but these terms are less precise)

A standard protocol for reporting SDMs

Zurell et al. (2020)

- 'Transfer' (also termed forecast or projection; but these terms are less precise)
- 'algorithmic uncertainty'

A standard protocol for reporting SDMs

Zurell et al. (2020)

- 'Transfer' (also termed forecast or projection; but these terms are less precise)
- 'algorithmic uncertainty'
- 'We define model complexity as the flexibility of the fitted biodiversity-environment relationship'

A standard protocol for reporting SDMs

Zurell et al. (2020)

- 'Transfer' (also termed forecast or projection; but these terms are less precise)
- 'algorithmic uncertainty'
- 'We define model complexity as the flexibility of the fitted biodiversity-environment relationship'
- Trop de focus sur les 'algorithmes', trop peu sur les données et leur collecte...

Coming of age...?

Divaguations

- bannir la 'pornographie algorithmique ؟'

Divaguations

- bannir la 'pornographie algorithmique ?'
- qu'apprend-t-on en soumettant un jeu de données empiriques à une batterie de méthodes/modèles/algorithmes en l'absence d'un standard ?

Divaguations

- bannir la 'pornographie algorithmique ?'
- qu'apprend-t-on en soumettant un jeu de données empiriques à une batterie de méthodes/modèles/algorithmes en l'absence d'un standard ?
- quel progrès depuis 2006 si chaque nouvel article en écologie récapitule une comparaison de tout ce qui est disponible sur le marché?

Divaguations

- bannir la 'pornographie algorithmique ?'
- qu'apprend-t-on en soumettant un jeu de données empiriques à une batterie de méthodes/modèles/algorithmes en l'absence d'un standard ?
- quel progrès depuis 2006 si chaque nouvel article en écologie récapitule une comparaison de tout ce qui est disponible sur le marché?
- quel dialogue avec les statisticiens quand on utilise les mêmes mots pour recouvrir des concepts différents?

Merci de votre attention

References I

Botella, C., Joly, A., Monestiez, P., Bonnet, P. & Munoz, F. (2020). Bias in Presence-Only Niche Models Related to Sampling Effort and Species Niches: Lessons for Background Point Selection. *PLoS One* 15 e0232078. URL `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0232078`.

Carvalho, C., Polson, N. & Scott, J. (2010). The Horseshoe Estimator for Sparse Signals. *Biometrika* 97 465–480.

Fithian, W. & Hastie, T. (2013). Finite-Sample Equivalence in Statistical Models for Presence-Only Data. *The Annals of Applied Statistics* 7 1917–1939.

Fitzpatrick, M. C., Gotelli, N. J. & Ellison, A. M. (2013). MaxEnt versus MaxLike: Empirical Comparisons with Ant Species Distributions. *Ecosphere* 4 55. URL `https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/ES13-00066.1`.

References II

Gimenez, O., Buckland, S. T., Morgan, B. J., Bez, N., Bertrand, S., Choquet, R., Dray, S., Etienne, M. P., Fewster, R., Gosselin, F., Mérigot, B., Monestiez, P., Morales, J. M., Mortier, F., Munoz, F., Ovaskainen, O., Pavoine, S., Pradel, R., Schurr, F. M., Thomas, L., Thuiller, W., Trenkel, V., de Valpine, P. & E., R. (2014). Statistical Ecology Comes of Age. *Biology Letters* 10 1–4.

Hallgren, W., Santana, F., Low-Choy, S., Zhao, Y. & Mackey, B. (2019). Species Distribution Models Can Be Highly Sensitive to Algorithm Configuration. *Ecological Modelling* 408 108719.

King, G. & Zeng, L. (2007). When Can History Be Our Guide? The Pitfalls of Counterfactual Inference. *International Studies Quaterly* 51 183–210.

Lozier, J., Aniello, P. & Hickerson, M. (2009). Predicting the Distribution of Sasquatch in Western North America: Anything Goes with Ecological Niche Modelling. *Journal of Biogeography* 36 1623–1627.

References III

Pacifici, K., Reich, B. J., Miller, D. A., Gardner, B., Stauffer, G., Singh, S., McKerrow, A. & Collazo, J. A. (2017). Integrating Multiple Data Sources in Species Distribution Modeling: a Framework for Data Fusion. *Ecology* 98 840–850.

Renner, I. W. & Warton, D. I. (2013). Equivalence of MAXENT and Poisson Point Process Models for Species Distribution Modeling in Ecology. *Biometrics* 69 274–281.

Royle, J., Chandler, R., Yackulic, C. & Nichols, J. (2012). Likelihood Analysis of Species Occurrence Probability from Presence-Only Data for Modelling Species Distributions. *Methods in Ecology and Evolution* 3 545–554.

References IV

Zurell, D., Franklin, J., König, C., Bouchet, P. J., Dormann, C. F., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J. J., Leitão, P. J., Park, D. S., Peterson, A. T., Rapacciuolo, G., Schmatz, D. R., Schröder, B., Serra-Diaz, J. M., Thuiller, T., Yates, K. L., Zimmermann, N. E. & Merow, C. (2020). A Standard Protocol for Reporting Species Distribution Models. *Ecography* 43 1–17.