Formalisation mathématique et méthode "d'étalonnage probabiliste"

April 11, 2025

Redéfinition du cadre : processus ponctuel

Cadre mathématique

Soit $\mathcal{X}$ un domaine compact de R^2. Notons $\mathcal{B}_b$ l'ensemble des boréliens bornés de $\mathcal{X}$. Notons également $N_\mathcal{X}$ l'ensemble des motifs de points de $\mathcal{X}$ (i.e les ensembles de points de $\mathcal{X}$ localement finis, c'est-à-dire les ensembles finis sur tout borélien borné). Ses éléments sont appelés configurations.

Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé. On munit $N_\mathcal{X}$ de la tribu $\mathcal{N}_\mathcal{X}$ engendrée par les applications de comptage n_B où :

$$\begin{aligned} n_B \colon N_\mathcal{X} &\to \mathbb{N} \\ x &\mapsto |x \cap B| \end{aligned}$$

.

Soit λ une fonction d'intensité sur le domaine : $\lambda : \mathcal{X} \to \mathbb{R}^+$. De plus, on notera pour tout $B \in \mathcal{B}_b$: $\lambda(B) = \int_B \lambda(s)ds$.

Soit X le processus ponctuel de Poisson non-homogène sur $\mathcal{X}$ d'intensité λ. X est donc une variable aléatoire à valeurs dans $(N_\mathcal{X}, \mathcal{N}_\mathcal{X})$, de loi de probabilité induite par $\mathbb{P}$, indépendante sur des ensembles disjoints et telle que pour tout $B \in \mathcal{B}_b$, la variable aléatoire de comptage $N(B) = n_B(x)$ suit une loi de Poisson de paramètre $\lambda(B)|B|$ où $|B|$ est le volume de B :

$$\forall B \in \mathcal{B}_b, N(B) \sim \mathcal{P}(\lambda(B)|B|)$$

.

Estimation par maximum de vraisemblance

Soient $\phi_1, \phi_2, ..., \phi_m$ des covariables telles que $\forall i \in [\![1, m]\!]\, \phi_i : \mathcal{X} \to \mathbb{R}$. On écrit l'intensité comme une fonction dépendant des covariables :

$$\forall \theta \in \mathbb{R}^m, \lambda_\theta : x \mapsto \sum_{i=1}^{m} \exp(\phi_i(x)\theta_i)$$

.

L'objectif est alors d'estimer le paramètre θ de telle sorte à maximiser la log-vraisemblance :

$$\hat{\theta} = \text{argmax}_{\theta \in \mathbb{R}^m} l(\lambda_\theta | X)$$

.

où, en notant s_i les points de données :

$$l(\lambda_\theta | X) = \sum_{i=1}^{k} \log(\lambda_\theta(s_i)) - \int_{\mathcal{X}} \lambda_\theta(s) ds$$

Contexte de l'étude

Dans le cadre de cette étude, l'espace géographique $\mathcal{X}$ est le territoire métropolitain, défini par les axes de longitude et latitude).

On dispose d'un ensemble de $n \in \mathbb{N}^*$ stations $\mathcal{S} = \{s_1, s_2, ..., s_n\}$ de points de l'espace où la biodiversité a été échantillonnée, ainsi qu'un sous-ensemble $\mathcal{S}_1$ de $k \in [\![0, n]\!]$ points de $\mathcal{S}$ où a été relevée la présence de l'espèce.

De plus, on a connaissance de $m = 36$ covariables climatiques $\phi_1, ..., \phi_m$ que sont les températures du jour mensuelles (12 covariables), les températures de la nuit mensuelles (12 covariables) et les précipitations moyennes mensuelles (12 covariables). La valeur de ces covariables est connue sur tout le domaine $\mathcal{X}$, disponible en format *raster* et *vecteur*. En particulier, les valeurs sont connues aux stations (i.e aux points du processus).

On considère que **l'ensemble $\mathcal{S}_1$ des points de l'espèce est une réalisation d'un processus ponctuel de Poisson non-homogène (PPPN) d'intensité dépendant des covariables**. Ce choix de paradigme laisse de côté les relevés $\mathcal{S}_0$ où l'espèce est absente : il s'agit d'une perspective *presence-only data* [1].

L'objectif est d'estimer les paramètres reliant les covariables climatiques à l'intensité de ce processus. A l'issue de cette estimation (par maximum de vraisemblance, inférence bayésienne), **on souhaite établir la probabilité de présence de l'espèce en tout point s du domaine** (i.e la probabilité de non-évitement en un point ?):

$$\mathbb{P}(N(s)) = e^{-\lambda(s)}$$

Questions pour Emilie, 14/04

1. Objectif : détermination de la fonction d'intensité ? → la considère-t-on comme aléatoire (processus de Cox) ?

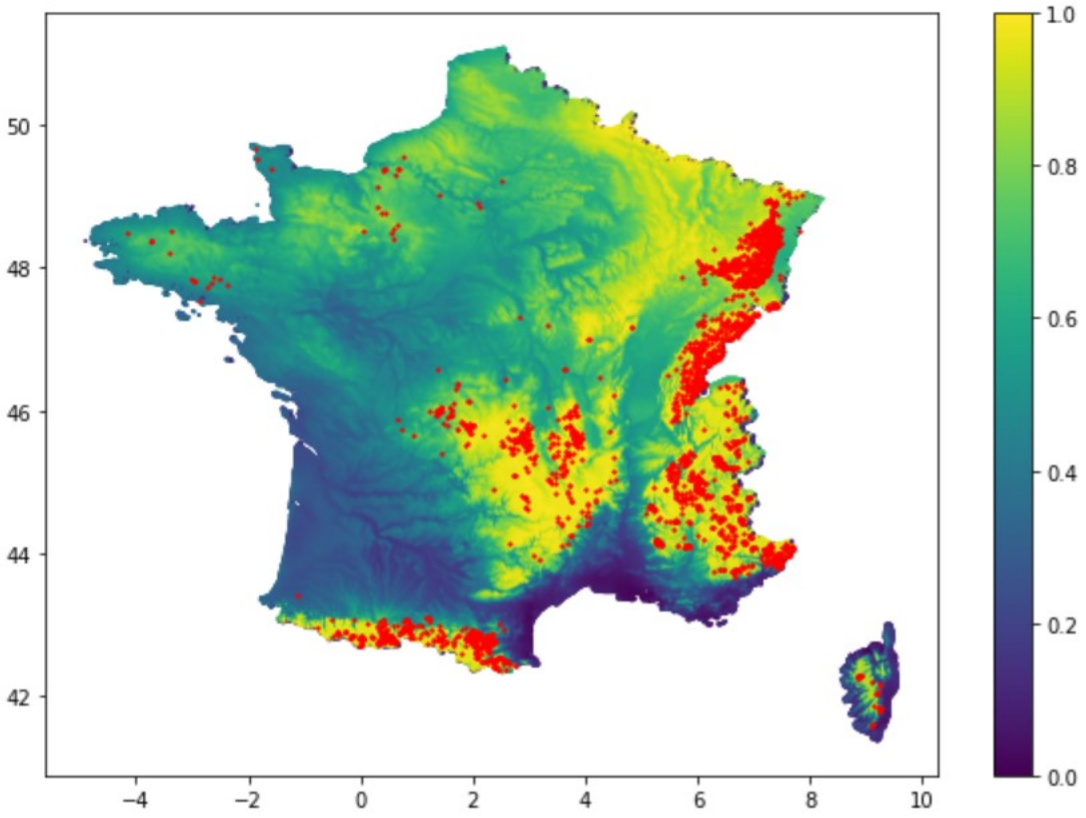

Figure 1: *Exemple-type de la méthode appliquée au sapin. En rouge, les points de données. En fond de couleur, la probabilité estimée de présence.*

2. Une fois la fonction d'intensité estimée, comment peut-on revenir à la probabilité de présence d'une espèce en un point ?

3. Processus de Poisson marqué pour tenir compte du niveau d'abondance ?

4. $\mathbb{R}^3$ pour tenir compte de l'aspect temporel du problème ?

Bio-indicateurs du climat : contexte et notations préalables

On dispose d'un ensemble de stations de mesures où sont mesurées une variable climatique ainsi que la présence, ou non, d'un taxon fixé. On va exprimer les concentrations du taxon dans la gamme climatique en passant du formalisme du dénombrement au formalisme des probabilités.

Soient $\mathcal{S}$ l'ensemble fini des stations climatiques, $\mathcal{S}_1$ l'ensemble des stations où le taxon est présent. Ainsi, $\mathcal{S}_1 \subset \mathcal{S}$. Notons $v : \mathcal{S} \to \mathbb{R}$ la fonction définie sur l'ensemble des stations donnant la valeur de la variable climatique d'intérêt.

De plus, on note $\dot{s} \sim \mathcal{U}(\mathcal{S})$ et $\dot{s_1} \sim \mathcal{U}(\mathcal{S}_1)$ les variables aléatoires indépendantes décrivant respectivement les stations, et les stations avec présence du taxon (par exemple, on peut respectivement les indexer de 1 à n et de 1 à $k \leqslant n$).

Les cardinaux de $\mathcal{S}$ et de $\mathcal{S}_1$ seront notés respectivement $|\mathcal{S}|$ et $|\mathcal{S}_1|$.

1 Première expression probabiliste de la proximité

1.1 Cas général

Dans le livre, l'optimum climatique v^* est défini comme la valeur climatique maximisant la proximité ($PROX$) du taxon dans la gamme, et donc en maximisant le nombre de compositions favorables (NCF).

Ecrivons la proximité d'un taxon par rapport à la gamme climatique en traduisant de manière ensembliste la définition page 51, illustrée en page 52 avec plusieurs exemples sous forme de tableaux.

Soit $w \in v(\mathcal{S})$:

$$\begin{aligned}
PROX(w) &= \frac{NCF(w)}{|\mathcal{S}||\mathcal{S}_1|} \\
&= \frac{1}{|\mathcal{S}||\mathcal{S}_1|} \sum_{s_1 \in \mathcal{S}_1} \big(\text{Card}(\{s \in \mathcal{S}, v(s) \notin [\min(v(s_1), w), \max(v(s_1), w)]\}) \\
&+ \frac{1}{2}\text{Card}(\{s \in \mathcal{S}, v(s) = w \text{ ou } v(s) = v(s_1)\})\big) \\
&= \frac{1}{|\mathcal{S}||\mathcal{S}_1|} \sum_{\substack{s_1 \in \mathcal{S}_1 \\ s \in \mathcal{S}}} \big(\mathbb{1}_{v(s) \notin [\min(v(s_1), w), \max(v(s_1), w)]} + \frac{1}{2}\mathbb{1}_{\{v(s)=w\} \cup \{v(s)=v(s_1)\}}\big)
\end{aligned}$$

Dans le livre, cette dernière expression correspond au comptage des valeurs v_{ext} en dehors des intervalles de type $[v(s_1), w]$, les valeurs aux bornes $v(s_1)$ et w étant comptés comme à moitié dehors (d'où le coefficient $\frac{1}{2}$). On peut réécrire plus simplement la proximité en terme de probabilité:

$$\begin{aligned}
PROX(w) &= \frac{1}{|\mathcal{S}_1|} \sum_{s_1 \in \mathcal{S}_1} \Bigg(\mathbb{P}(v(\dot{s}) \notin [\min(v(s_1), w), \max(v(s_1), w)]) \\
&+ \frac{1}{2}\mathbb{P}(\{v(\dot{s}) = w\} \cup \{v(\dot{s}) = v(s_1)\})\Bigg) \\
&= \sum_{s_1 \in \mathcal{S}_1} \mathbb{P}(\dot{s_1} = s_1)\Bigg(\mathbb{P}(v(\dot{s}) \notin [\min(v(s_1), w), \max(v(s_1), w)]) \\
&+ \frac{1}{2}\mathbb{P}(\{v(\dot{s}) = w\} \cup \{v(\dot{s}) = v(s_1)\})\Bigg)
\end{aligned}$$

Donc, en utilisant la formule des probabilités totales, par indépendance de $\dot{s}$ et $\dot{s_1}$:

$$PROX(w) = \mathbb{P}(v(\dot{s}) \notin [\min(v(\dot{s_1}), w), \max(v(\dot{s_1}), w)]) + \frac{1}{2}\mathbb{P}(\{v(\dot{s}) = w\} \cup \{v(\dot{s}) = v(\dot{s_1})\})$$

Notons $\dot{v} = v(\dot{s})$ la variable aléatoire décrivant la valeur climatique sur les stations et $\dot{v_1} = v(\dot{s_1})$ celle décrivant la valeur climatique sur les stations où le taxon est présent. Ainsi, $\dot{v}$ et $\dot{v_1}$ sont indépendantes et on peut écrire plus simplement :

$$PROX(w) = \mathbb{P}(\dot{v} \notin [\min(\dot{v_1}, w), \max(\dot{v_1}, w)]) + \frac{1}{2}\mathbb{P}(\{\dot{v} = w\} \cup \{\dot{v} = \dot{v_1}\})$$

Puis :

$$\begin{aligned} 1 - PROX(w) &= \mathbb{P}(\dot{v} \in [\min(\dot{v_1}, w), \max(\dot{v_1}, w)]) - \frac{1}{2}\mathbb{P}(\{\dot{v} = w\} \cup \{\dot{v} = \dot{v_1}\}) \\ &= \mathbb{P}(\{w \leqslant \dot{v} \leqslant \dot{v_1}\} \cup \{\dot{v_1} \leqslant \dot{v} \leqslant w\}) - \frac{1}{2}\mathbb{P}(\{\dot{v} = w\} \cup \{\dot{v} = \dot{v_1}\}) \end{aligned}$$

1.2 Simplification dans le cas continu

Dans le cas continu, les événements $\{\dot{v} = w\}$ et $\{\dot{v} = \dot{v_1}\}$ sont de probabilité nulle, et la réunion de la première probabilité ci-dessus est disjointe. Ainsi :

$$\begin{aligned} 1 - PROX(w) &= \mathbb{P}(w < \dot{v} < \dot{v_1}) + \mathbb{P}(\dot{v_1} < \dot{v} < w) \\ &= \mathbb{P}(\min(w, \dot{v_1}) < \dot{v} < \max(w, \dot{v_1})) \end{aligned}$$

Notons f_1 la densité de $\dot{v_1}$ et F la fonction de répartition de $\dot{v}$. Alors :

$$\begin{aligned} 1 - PROX(w) &= \int_{v(\mathcal{S}_1)} \mathbb{P}(\min(w, w_1) < \dot{v} < \max(w, w_1)) f_1(w_1)\, dw_1 \\ &= \int_{v(\mathcal{S}_1)} \Big(F(\max(w, w_1)) - F(\min(w, w_1)) \Big) f_1(w_1)\, dw_1 \\ &= \mathbb{E}\Big(F(\max(w, \dot{v_1})) - F(\min(w, \dot{v_1})) \Big) \end{aligned}$$

Je ne sais pas trop comment interpréter cette expression..

2 Réécriture de la concentration

Notons $\dot{v_U}$ la variable aléatoire décrivant les valeurs climatiques prises par le taxon ubiquiste. Par définition, le taxon ubiquiste ne dépend pas des facteurs climatiques : sa distribution dans la gamme climatique est donc la même que celle des stations quelconques. Autrement dit, en notant $\dot{v_U}$ les valeurs prises par le taxon ubiquiste, $\dot{v}$ et $\dot{v_U}$ sont indépendantes et identiquement distribuées (iid).

Soit $w \in v(\mathcal{S})$. D'après la page 54 du livre, l'expression de la concentration est la suivante :

$$\begin{aligned} CTRA(w) &= \frac{PROX(w) - UPROX(w)}{1 - UPROX(w)} \\ &= 1 - \frac{1 - PROX(w)}{1 - UPROX(w)} \end{aligned}$$

D'où, en utilisant le résultat de la section précédente dans le cas continu :

$$1 - CTRA(w) = \frac{\mathbb{P}(\min(w, \dot{v_1}) < \dot{v} < \max(w, \dot{v_1}))}{\mathbb{P}(\min(w, \dot{v_U}) < \dot{v} < \max(w, \dot{v_U}))}$$

3 Optimisation de la complexité

3.1 Cas général : relation de récurrence sur les proximités

Pour appliquer la méthode, on va devoir calculer les proximités pour $w \in v(\mathcal{S})$ parcourant toute la gamme climatique, en cherchant à minimiser le nombre d'opértions nécessaires.

On rappelle que pour tout $w \in v(\mathcal{S})$, en notant $g(w) = 1 - PROX(w)$:

$$\begin{aligned} g(w) &= (1 - PROX(w)) \\ &= \mathbb{P}(\{w \leqslant \dot{v} \leqslant \dot{v_1}\} \cup \{\dot{v_1} \leqslant \dot{v} \leqslant w\}) - \frac{1}{2}\mathbb{P}(\{\dot{v} = w\} \cup \{\dot{v} = \dot{v_1}\}) \\ &= \mathbb{P}(w \leqslant \dot{v} \leqslant \dot{v_1}) + \mathbb{P}(\dot{v_1} \leqslant \dot{v} \leqslant w) - \mathbb{P}(\dot{v} = w, \dot{v_1} = w) - \frac{1}{2}\mathbb{P}(\{\dot{v} = w\} \cup \{\dot{v} = \dot{v_1}\}) \end{aligned}$$

Pour calculer toutes les valeurs de proximité, on répète $n = |\mathcal{S}|$ fois le comptage ci-dessus, de complexité $|\mathcal{S}| \times |\mathcal{S}_1|$. En considérant que $|\mathcal{S}_1| \approx |\mathcal{S}|$ et en appliquant directement la formule dans le livre, la complexité est donc de l'ordre de n^3 (Emmanuel m'a dit que cela prenait beaucoup de temps de calcul).

Peut-on optimiser le comptage de telle sorte à en réutiliser une partie des valeurs selon la valeur w considérée ? Intuitivement, les comptages pour déterminer la proximité de deux valeurs consécutives dans la gamme climatique ne diffèrent que de quelques valeurs.

Notons donc $v(\mathcal{S}) = \{w_k, k \in [\![1, n]\!]\}$ l'ensemble des valeurs de la gamme, telles que $w_1 < w_2 < ... < w_n$. Notons aussi F_1 la fonction de répartition (discrète) de $\dot{v_1}$, pour tout $i \in [\![1, n]\!]$ $p_i = \mathbb{P}(\dot{v} = w_i)$ et $q_i = \mathbb{P}(\dot{v_1} = w_i)$.

On cherche donc à établir une relation de récurrence sur la suite finie $(g(w_k))_{k \in [\![1,n]\!]}$.

Soit $k \in [\![1, n]\!]$. Alors :

$$\begin{aligned} \mathbb{P}(w_{k+1} \leqslant \dot{v} \leqslant \dot{v_1}) &= \mathbb{P}(w_k \leqslant \dot{v} \leqslant \dot{v_1}) - \mathbb{P}(\dot{v} = w_k, \dot{v_1} \geqslant w_k) \\ &= \mathbb{P}(w_k \leqslant \dot{v} \leqslant \dot{v_1}) - \mathbb{P}(\dot{v} = w_k)\mathbb{P}(\dot{v_1} \geqslant w_k)) \\ &= \mathbb{P}(w_k \leqslant \dot{v} \leqslant \dot{v_1}) - p_k(1 - F_1(w_{k-1})) \end{aligned}$$

où par convention, si $k = 1$, alors $F_1(w_{k-1}) = 0$. D'autre part :

$$\begin{aligned} \mathbb{P}(\dot{v_1} \leqslant \dot{v} \leqslant w_{k+1}) &= \mathbb{P}(\dot{v_1} \leqslant \dot{v} \leqslant w_k) + \mathbb{P}(\dot{v} = w_{k+1}, \dot{v_1} \geqslant w_{k+1}) \\ &= \mathbb{P}(\dot{v_1} \leqslant \dot{v} \leqslant w_k) + \mathbb{P}(\dot{v} = w_{k+1})\mathbb{P}(\dot{v_1} \geqslant w_{k+1}) \\ &= \mathbb{P}(\dot{v_1} \leqslant \dot{v} \leqslant w_k) + p_{k+1}F_1(w_{k+1}) \end{aligned}$$

Et par ailleurs :

$$\begin{aligned}\mathbb{P}(\{\dot{v} = w_{k+1}\} \cup \{\dot{v} = \dot{v_1}\}) &= \mathbb{P}(\dot{v} = w_{K+1}) + \sum_{i=1}^{n} \mathbb{P}(\dot{v} = w_i)\mathbb{P}(\dot{v_1} = w_i) \\ &\quad - \mathbb{P}(\dot{v} = w_{k+1}, \dot{v_1} = w_{k+1}) \\ &= p_{k+1} + \sum_{i=1}^{n} p_i q_i - p_{k+1} q_{k+1}\end{aligned}$$

D'où, en réunissant les trois égalités précédentes :

$$g(w_{k+1}) = g(w_k) - p_k(1 - F_1(w_{k-1}) + p_{k+1}F_1(w_{k+1}) - p_{k+1}q_{k+1} + p_k q_k - \frac{1}{2}(p_{k+1} - p_{k+1} - p_k + p_k q_k)$$

donc :

$$\begin{aligned}g(w_{k+1}) - g(w_k) &= p_{k+1}\big(F_1(w_{k+1}) - \frac{1}{2}q_{k+1} - \frac{1}{2}\big) + p_k\big(F_1(w_{k-1}) + \frac{1}{2}q_k - \frac{1}{2}\big) \\ &= p_{k+1}\big(F_1(w_{k+1}) - \frac{1}{2}q_{k+1} - \frac{1}{2}\big) + p_k\big(F_1(w_k) - \frac{1}{2}q_k - \frac{1}{2}\big)\end{aligned}$$

Enfin, déterminons une expression simplifiée de $g(w_1)$:

$$\begin{aligned}g(w_1) &= \mathbb{P}(\dot{v} \leqslant \dot{v_1}) - \frac{1}{2}\mathbb{P}(\{\dot{v} = w_1\} \cup \{\dot{v} = \dot{v_1}\}) \\ &= \sum_{1 \leqslant i \leqslant k \leqslant n} p_i q_k - \frac{1}{2}p_1(1 - q_1) - \frac{1}{2}\sum_{i=0}^{n} p_i q_i\end{aligned}$$

Concrètement, on pourra utiliser les histogrammes de $\dot{v}$ et $\dot{v_1}$ pour passer de la proximité de w_k à celle de w_{k+1}. Notons $g_k = g(w_k) = 1 - PROX(w_k)$, $g_{U.k} = 1 - UPROX(w_k)$ la proximité du taxon ubiquiste, et $u_k = p_k\big(F_1(w_k) - \frac{1}{2}q_k - \frac{1}{2}\big)$. Ainsi :

$$\begin{cases} g_1 = \sum_{1 \leqslant i \leqslant k \leqslant n} p_i q_k - \frac{1}{2}p_1(1 - q_1) - \frac{1}{2}\sum_{i=0}^{n} p_i q_i \\ g_{k+1} - g_k = u_{k+1} + u_k \\ CTRA(w_k) = 1 - \frac{g_k}{g_{U.k}} \end{cases}$$

où les u_k se déterminent par un nombre fini d'opérations élémentaires (une dizaine de sommes et de produits issus des histogrammes de $\dot{v}$ et $\dot{v_1}$).

3.2 Cas particulier : le taxon ubiquiste

Dans le cas du taxon ubiquiste, $\dot{v}$ et $\dot{v_U}$ sont iid. donc pour tout $i \in [\![1, n]\!], p_i = q_i$.

3.3 Algorithme

1. Calcul des histogrammes p, q et de l'histogramme cumulé F_1 (complexité linéaire).
2. Calcul de g_1 et $g_{U.k}$ (complexité quadratique). Il s'agit d'un simple comptage conditionnel : $g_1 = \sum_{1 \leqslant i \leqslant k \leqslant n} p_i q_k - \frac{1}{2} p_1 (1 - q_1) - \frac{1}{2} \sum_{i=0}^{n} p_i q_i$.
3. Calcul par récurrence des suites finies $(g_k)_{k \in [\![1,n]\!]}$ et $(g_{U.k})_{k \in [\![1,n]\!]}$ et des concentrations (complexité linéaire) en utilisant les simples relations : $g_{k+1} - g_k = u_{k+1} + u_k$ et $CTRA(w_k) = 1 - \frac{g_k}{g_{U.k}}$.
4. Recherche de l'optimum (i.e la concentration maximale) par changement de signe de la différence finie des concentrations (complexité logarithmique).

4 Interprétation

Rappellons l'expression de la concentration en fonction des proximités du taxon et du taxon ubiquiste.

$$1 - CTRA(w) = \frac{\mathbb{P}(\min(w, \dot{v_1}) < \dot{v} < \max(w, \dot{v_1}))}{\mathbb{P}(\min(w, \dot{v_U}) < \dot{v} < \max(w, \dot{v_U}))}$$

Pourquoi considérer le rapport des deux probabilités ?

Utiliser la concentration comme une probabilité (p 109) est incorrect car la concentration peut être négative... Utiliser la proximité du taxon comme une probabilité ? Pourquoi ? Comment le justifier ?

References

[1] Trevor J. Hefley and Mevin B. Hooten. "Hierarchical Species Distribution Models". en. In: *Current Landscape Ecology Reports* 1.2 (June 2016), pp. 87–97. ISSN: 2364-494X. DOI: 10.1007/s40823-016-0008-7. URL: http://link.springer.com/10.1007/s40823-016-0008-7 (visited on 04/10/2025).