

# Stage ISIGE

## Concentration d'un taxon dans la gamme climatique

27 mars 2025

### 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  $v : \mathcal{S} \rightarrow \mathbb{R}$  la fonction définie sur l'ensemble des stations donnant la valeur de la variable climatique d'intérêt.

De plus, on note  $S \sim \mathcal{U}(\mathcal{S})$  et  $S_1 \sim \mathcal{U}(\mathcal{S}_1)$  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  $k \leq n$ ).

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

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

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:

$$\begin{aligned} PROX(w) &= \frac{1}{|\mathcal{S}_1|} \sum_{s_1 \in \mathcal{S}_1} \left( \mathbb{P}(v(S) \notin [\min(v(s_1), w), \max(v(s_1), w)]) \right. \\ &\quad \left. + \frac{1}{2} \mathbb{P}(\{v(S) = w\} \cup \{v(S) = v(s_1)\}) \right) \\ &= \sum_{s_1 \in \mathcal{S}_1} \mathbb{P}(S_1 = s_1) \left( \mathbb{P}(v(S) \notin [\min(v(s_1), w), \max(v(s_1), w)]) \right. \\ &\quad \left. + \frac{1}{2} \mathbb{P}(\{v(S) = w\} \cup \{v(S) = v(s_1)\}) \right) \end{aligned}$$

Donc, en utilisant la formule des probabilités totales, par indépendance de  $S$  et  $S_1$  :

$$PROX(w) = \mathbb{P}(v(S) \notin [\min(v(S_1), w), \max(v(S_1), w)]) + \frac{1}{2} \mathbb{P}(\{v(S) = w\} \cup \{v(S) = v(S_1)\})$$

Notons  $V = v(S)$  la variable aléatoire décrivant la valeur climatique sur les stations et  $V_1 = v(S_1)$  celle décrivant la valeur climatique sur les stations où le taxon est présent. Ainsi,  $V$  et  $V_1$  sont indépendantes et on peut écrire plus simplement :

$$PROX(w) = \mathbb{P}(V \notin [\min(V_1, w), \max(V_1, w)]) + \frac{1}{2} \mathbb{P}(\{V = w\} \cup \{V = V_1\})$$

Puis :

$$\begin{aligned} 1 - PROX(w) &= \mathbb{P}(V \in [\min(V_1, w), \max(V_1, w)]) - \frac{1}{2} \mathbb{P}(\{V = w\} \cup \{V = V_1\}) \\ &= \mathbb{P}(\{w \leq V \leq V_1\} \cup \{V_1 \leq V \leq w\}) - \frac{1}{2} \mathbb{P}(\{V = w\} \cup \{V = V_1\}) \end{aligned}$$

## 1.2 Simplification dans le cas continu

Dans le cas continu, les événements  $\{V = w\}$  et  $\{V = V_1\}$  sont de probabilité nulle, et la réunion de la première probabilité est disjointe. Ainsi :

$$\begin{aligned} 1 - PROX(w) &= \mathbb{P}(w < V < V_1) + \mathbb{P}(V_1 < V < w) \\ &= \mathbb{P}(\min(w, V_1) < V < \max(w, V_1)) \end{aligned}$$

Notons  $f_1$  la densité de  $V_1$  et  $F_V$  la fonction de répartition de  $V$ . Alors :

$$\begin{aligned} 1 - PROX(w) &= \int_{v(\mathcal{S}_1)} \mathbb{P}(\min(w, w_1) < V < \max(w, w_1)) f_1(w_1) dw_1 \\ &= \int_{v(\mathcal{S}_1)} \left( F_V(\max(w, w_1)) - F_V(\min(w, w_1)) \right) f_1(w_1) dw_1 \\ &= \mathbb{E} \left( F_V(\max(w, V_1)) - F_V(\min(w, V_1)) \right) \end{aligned}$$

Je ne sais pas trop comment interpréter cette expression..

## 2 Réécriture de la concentration

Notons  $V_u$  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  $V_U$  les valeurs prises par le taxon ubiquiste,  $V$  et  $V_U$  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 :

$$\begin{aligned} 1 - CTRA(w) &= \frac{\mathbb{P}(\min(w, V_1) < V < \max(w, V_1))}{\mathbb{P}(\min(w, V_U) < V < \max(w, V_U))} \\ &= \mathbb{P}(\min(w, V_1) < V < \max(w, V_1) | \min(w, V_U) < V < \max(w, V_U)) \end{aligned}$$

## 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'opérations nécessaires.

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

$$\begin{aligned} g(w) &= (1 - PROX(w)) \\ &= \mathbb{P}(\{w \leq V \leq V_1\} \cup \{V_1 \leq V \leq w\}) - \frac{1}{2}\mathbb{P}(\{V = w\} \cup \{V = V_1\}) \\ &= \mathbb{P}(w \leq V \leq V_1) + \mathbb{P}(V_1 \leq V \leq w) - \mathbb{P}(V = w, V_1 = w) - \frac{1}{2}\mathbb{P}(\{V = w\} \cup \{V = V_1\}) \end{aligned}$$

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  $v(\mathcal{S}) = \{w_k, k \in \llbracket 1, n \rrbracket\}$  l'ensemble des valeurs de la gamme, telles que  $w_1 < w_2 < \dots < w_n$ . Notons aussi  $F_1$  la fonction de répartition (discrète) de  $V_1$ , pour tout  $i \in \llbracket 1, n \rrbracket$   $p_i = \mathbb{P}(V = w_i)$  et  $q_i = \mathbb{P}(V_1 = w_i)$ .

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

Soit  $k \in \llbracket 1, n \rrbracket$ . Alors :

$$\begin{aligned}
\mathbb{P}(w_{k+1} \leq V \leq V_1) &= \mathbb{P}(w_k \leq V \leq V_1) - \mathbb{P}(V = w_k, V_1 \geq w_k) \\
&= \mathbb{P}(w_k \leq V \leq V_1) - \mathbb{P}(V = w_k)\mathbb{P}(V_1 \geq w_k) \\
&= \mathbb{P}(w_k \leq V \leq V_1) - p_k(1 - F_1(w_{k-1}))
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{P}(V_1 \leq V \leq w_{k+1}) &= \mathbb{P}(V_1 \leq V \leq w_k) + \mathbb{P}(V = w_{k+1}, V_1 \geq w_{k+1}) \\
&= \mathbb{P}(V_1 \leq V \leq w_k) + \mathbb{P}(V = w_{k+1})\mathbb{P}(V_1 \geq w_{k+1}) \\
&= \mathbb{P}(V_1 \leq V \leq w_k) + p_{k+1}F_1(w_{k+1})
\end{aligned}$$

Et par ailleurs :

$$\begin{aligned}
\mathbb{P}(\{V = w_{k+1}\} \cup \{V = V_1\}) &= \mathbb{P}(V = w_{k+1}) + \sum_{i=1}^n \mathbb{P}(V = w_i)\mathbb{P}(V_1 = w_i) \\
&\quad - \mathbb{P}(V = w_{k+1}, V_1 = w_{k+1}) \\
&= p_{k+1} + \sum_{i=1}^n p_i q_i - p_{k+1} q_{k+1}
\end{aligned}$$

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 :

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

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

$$\begin{aligned}
g(w_1) &= \mathbb{P}(V \leq V_1) - \frac{1}{2}\mathbb{P}(\{V = w_1\} \cup \{V = V_1\}) \\
&= \sum_{1 \leq i \leq k \leq n} p_i q_k - \frac{1}{2}p_1(1 - q_1) - \frac{1}{2}\sum_{i=0}^n p_i q_i
\end{aligned}$$

Concrètement, on pourra utiliser les histogrammes de  $V$  et  $V_1$  pour passer de la proximité de  $w_k$  à celle de  $w_{k+1}$ . Notons  $g_k = g(w_k) = 1 - PROX(w_k)$ ,  $g_{U.k} = UPROX(w_k)$  la proximité du taxon ubiquiste, et  $u_k = p_k(F_1(w_k) - \frac{1}{2}q_k - \frac{1}{2})$ . Ainsi :

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

où les  $u_k$  se déterminent par un nombre fini d'opérations élémentaires (une dizaine de sommes et de produits issus des histogrammes de  $V$  et  $V_1$ ).

### 3.2 Cas particulier : le taxon ubiquiste

Dans le cas du taxon ubiquiste, pour tout  $i \in \llbracket 1, n \rrbracket$ ,  $p_i = q_i$ . En particulier :

$$g_{U,1} = \sum_{1 \leq i \leq k \leq n} p_i p_k - \frac{1}{2} p_1 (1 - p_1) - \frac{1}{2} \sum_{i=0}^n p_i^2$$

### 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 :  $g_1 = \sum_{1 \leq i \leq k \leq n} p_i q_k - \frac{1}{2} p_1 (1 - q_1) - \frac{1}{2} \sum_{i=0}^n p_i q_i$ .
3. Calcul par récurrence des suites finies  $(g_k)_{k \in \llbracket 1, n \rrbracket}$  et  $(g_{U,k})_{k \in \llbracket 1, n \rrbracket}$  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).