Analyse d’occupancy de Trithemis annulata

Arthur LE CAMUS & Félicia DOSSOU

2024-2025

3.1 Questions méthodologiques– données simulées

L’objectif de cette partie est d’évaluer la puissance statistique des modèles d’occupancy à détecter un effet d’artificialisation dans différents plans d’échantillonnage simulés.

Simulation 1 – Design homogène : 25 sites fixes, 5 ans, 3 visites

Dans cette première simulation, nous avons testé un design homogène basé sur 25 sites fixes observés pendant 5 années consécutives, avec 3 visites par an. L’objectif est d’évaluer l’effet de l’artificialisation sur la probabilité d’occupation (ψ) à l’aide d’un modèle d’occupancy avec détection constante. Les covariables du site incluent un gradient d’artificialisation simulé aléatoirement entre 0 et 1. La probabilité de détection est fixée à 0.6, et l’effet simulé de l’artificialisation sur ψ est négatif (β=−2). Les données de détection sont générées conditionnellement à l’occupation, puis analysées via la fonction occu() du package unmarked, en modélisant ψ en fonction de artif et une détection constante.

```
## 
## Call:
## occu(formula = ~1 ~ artif, data = umf)
## 
## Occupancy (logit-scale):
##             Estimate    SE     z P(>|z|)
## (Intercept)      1.0 0.973  1.03   0.303
## artif           -2.7 1.535 -1.76   0.079
## 
## Detection (logit-scale):
##  Estimate    SE    z P(>|z|)
##     0.283 0.174 1.63   0.103
## 
## AIC: 219.6765 
## Number of sites: 25
```

```
##                 0.025     0.975
## psi(Int)   -0.9054765 2.9077514
## psi(artif) -5.7056695 0.3123322
```

Le modèle estimé montre un effet négatif de l’artificialisation sur la probabilité d’occupation (estimate = -2.7), cohérent avec la valeur simulée (-2). Cependant, cet effet n’est pas significatif au seuil de 5 % (p = 0.079), bien qu’il le soit presque au seuil de 10 %, ce qui suggère une tendance. L’intervalle de confiance à 95 % pour cet effet est large, ce qui traduit une incertitude importante liée à la taille d’échantillon. Cela montre que, même avec un design de 25 sites sur 5 ans et 3 visites par an, la puissance statistique peut rester limitée pour détecter un effet modéré. Cela souligne l’importance de réaliser des analyses de puissance avant le terrain pour ajuster les paramètres de suivi.

Estimation de la puissance sur 100 simulations

La puissance d’un test statistique est définie comme :

$$\text{Puissance} = 1 - P(\text{Erreur de type II}) = P(ne pas rejeter H0|H1 vraie)$$

Dans notre modèle d’occupancy, la puissance mesure la **probabilité de détecter une différence de présence en fonction de l’artificialisation**. Elle dépend de plusieurs paramètres :

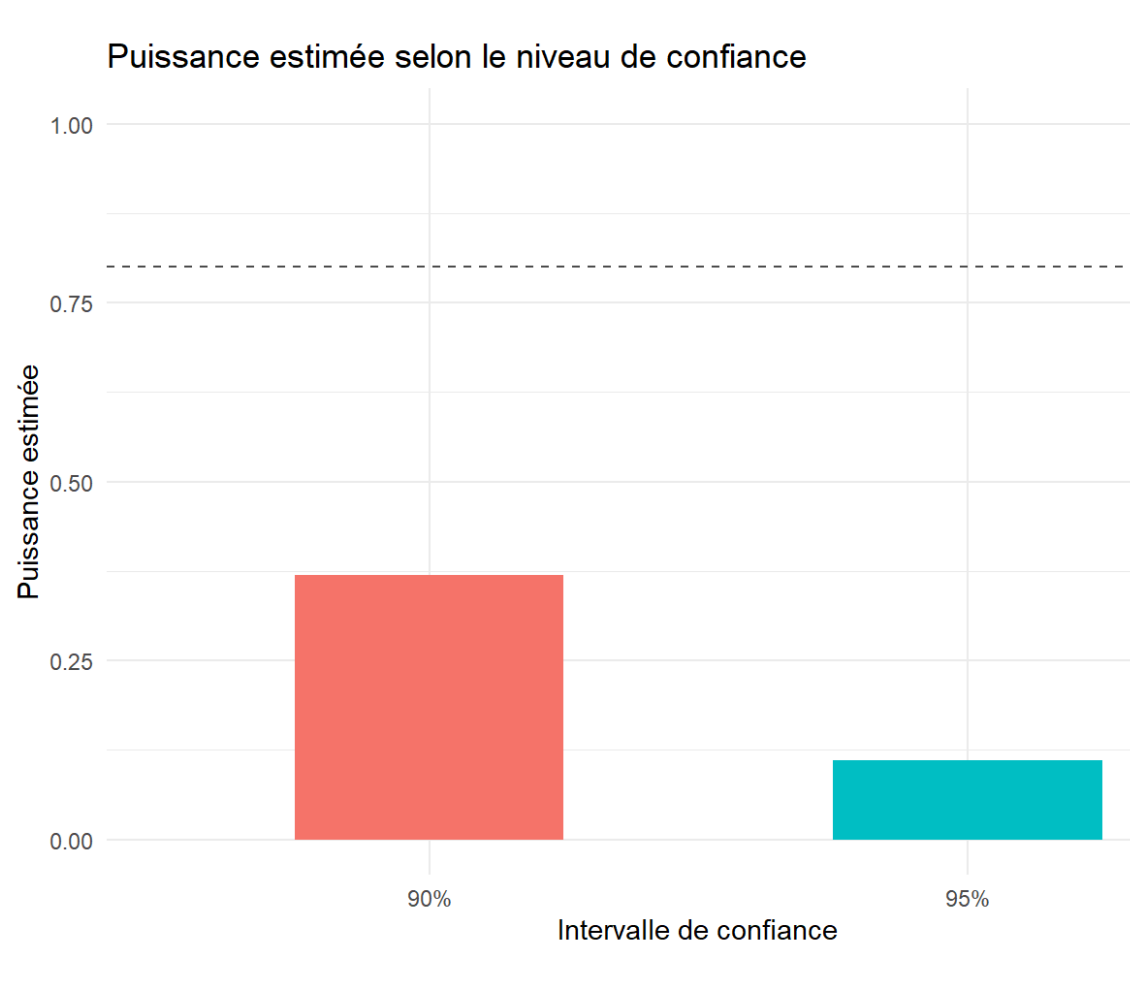
- **Nombre de sites échantillonnés** (n)
- **Nombre de répétitions (passages) par site**
- **Force de l’effet** (β_1) de l’artificialisation
- **Taux de détection** (p)
- **Taux de présence** (ψ)

En général, on cherche une puissance **d’au moins 80%**.

On pose les hypothèses suivantes :
$H_0 : \beta \neq 0 \rightarrow$ l’artificialisation n’influe pas sur la présence des libellules

$H_1 : \beta = 0 \rightarrow$ l’artificialisation influe sur la présence des libellules

Nous avons simulé 100 jeux de données à partir des paramètres définis dans la première simulation (25 sites fixes, 5 ans, 3 visites, avec un effet simulé de β = -2). Pour chaque simulation, un modèle d’occupancy a été ajusté, et la significativité du coefficient associé à artif a été testée. Le taux de détection d’un effet significatif (p-value < α) représente alors la puissance empirique du design. Deux niveaux de confiance ont été testés : 95 % (α = 0.05) et 90 % (α = 0.10), afin de comparer leur impact sur la puissance.



L’analyse de puissance réalisée sur 100 simulations avec un effet réel de β = –2 montre une probabilité de détection faible : environ 10–15 % au seuil de 5 %, et 35–40 % au seuil de 10 %, soit bien en dessous du seuil recommandé de 80 %. Cela indique que le protocole actuel (25 sites, 5 ans, 3 visites par an) n’est pas suffisamment informatif pour détecter un effet modéré de l’artificialisation avec fiabilité. Cette faible puissance peut s’expliquer par un nombre de sites trop restreint (la covariable varie au niveau du site), une détection imparfaite (p = 0.6) qui introduit du bruit, et/ou une variabilité aléatoire importante dans les données simulées. Pour améliorer la puissance, il serait pertinent d’augmenter le nombre de sites, le nombre de visites annuelles, la probabilité de détection, ou de renforcer le contraste de la covariable simulée.

Simulation 2 – Design mixte : 20 sites fixes + 5 flottants par an

Dans cette variante, nous cherchons à évaluer la puissance d’un plan d’échantillonnage mixte : 20 sites permanents suivis pendant 5 ans avec 3 visites/an et 5 sites “flottants” différents chaque année, avec seulement 3 visites chacun (pas de suivi dans le temps).

Ce type de design reflète une structure réaliste dans les suivis de terrain, avec une partie du réseau fixe et une autre renouvelée chaque année. L’objectif est de tester si ce design permet de détecter un effet de l’artificialisation sur la probabilité d’occupation (ψ), avec un effet simulé de β = –2.

```
## 
## Call:
## occu(formula = ~1 ~ artif, data = umf2)
## 
## Occupancy (logit-scale):
##             Estimate    SE     z P(>|z|)
## (Intercept)     0.97 0.675  1.44  0.1505
## artif          -2.70 1.338 -2.02  0.0437
## 
## Detection (logit-scale):
##  Estimate   SE    z P(>|z|)
##     0.245 0.16 1.53   0.127
## 
## AIC: 286.3657 
## Number of sites: 45
```

```
##                 0.025       0.975
## psi(Int)   -0.3523767  2.29337145
## psi(artif) -5.3226165 -0.07610443
```

Le modèle ajusté sur les données simulées renvoie une estimation de l’effet artif à –2.699, avec un écart-type de 1.338, ce qui donne une statistique z ≈ –2.02 et une p-value de 0.0437. Cela signifie que l’effet est statistiquement significatif au seuil de 5 %, et donc détecté dans cette simulation. L’intercept est estimé à 0.97 (ψ moyen ≈ 0.73 pour artif = 0), avec une p-value non significative, ce qui est attendu car l’intérêt porte sur la variation de ψ en fonction de artif, et non sur sa valeur absolue.

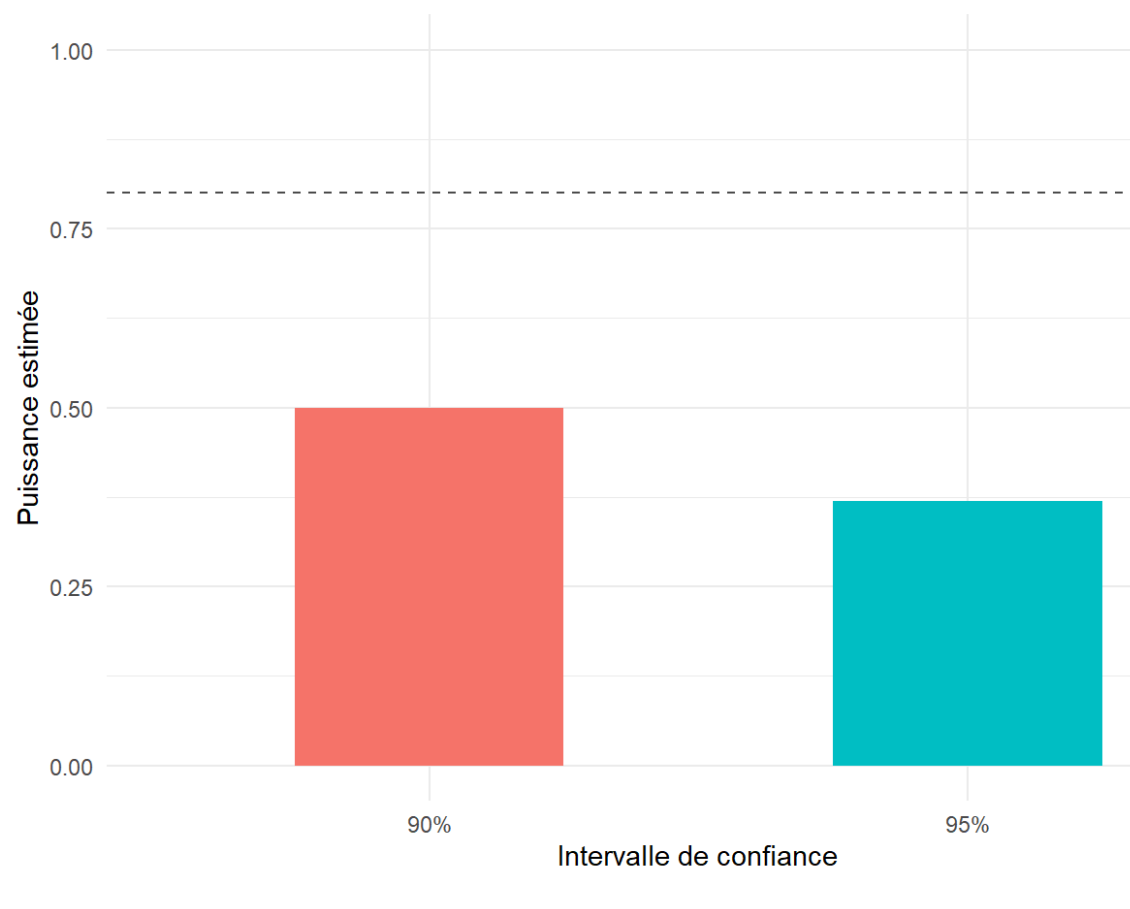
Dans cette simulation, le design mixte a permis de détecter l’effet de l’artificialisation avec succès. Le fait de combiner des sites suivis plusieurs années avec des sites nouveaux chaque année semble fournir une information suffisante pour capter l’effet de la covariable. Il serait toutefois nécessaire de répéter cette simulation (via une analyse de puissance empirique sur 100 itérations) pour confirmer que cette performance est robuste et non due au hasard d’un seul jeu de données.

Estimation de la puissance sur 100 simulations

Comme précédemment, nous avons simulé 100 jeux de données, cette fois selon un design mixte, combinant 20 sites fixes suivis pendant 5 ans avec 5 sites “flottants” différents chaque année (3 visites chacun). L’effet de l’artificialisation sur la probabilité d’occupation est également fixée à β = –2.

$H_0 : \beta \neq 0 \rightarrow$ l’artificialisation n’influe pas sur la présence des libellules

$H_1 : \beta = 0 \rightarrow$ l’artificialisation influe sur la présence des libellules



Le design mixte présente une puissance estimée à environ 50 % pour un seuil de 10 %, et 40 % pour un seuil de 5 %, ce qui représente une nette amélioration par rapport au design homogène précédent (≈35 % et 10–15 % respectivement). Cela suggère que l’ajout de sites flottants permet de mieux couvrir la variabilité spatiale de la covariable artif, ce qui renforce le signal et augmente la probabilité de détecter un effet. Toutefois, la puissance reste inférieure au seuil recommandé de 80 %, indiquant qu’un renforcement supplémentaire du protocole (plus de sites, meilleur taux de détection, plus de contrastes) serait nécessaire pour garantir une détection fiable de l’effet simulé.

3.2 Questions sur les données réelles

L’objectif de ce projet est d’analyser la distribution d’une espèce d’odonate, Trithemis annulata, sur le territoire de Bordeaux Métropole à l’aide de modèles d’occupancy. Cette espèce est suivie dans le cadre de programmes de sciences participatives, avec des données collectées entre 2018 et 2023. La problématique principale est de comprendre *dans quelles conditions cette espèce est présente : influence de l’urbanisation, des conditions météo ou du moment dans la saison ?*

Pour cela, nous structurons les données d’observation, construisons les covariables pertinentes, ajustons plusieurs modèles d’occupancy avec le package unmarked, et réalisons des analyses de puissance pour évaluer la robustesse des résultats obtenus.

Après avoir téléchargé les données, nous entamons une phase de prétraitement.

Nous transformons la date en jour julien, extrayons uniquement les valeurs numériques du vent, et filtrons les transects pour ne conserver que les plus fréquents (T1 à T4), en excluant les variantes telles que T4bis ou T4ter.

Puisque notre étude porte uniquement sur les sites où des odonates ont été observés, nous excluons tous les autres sites où aucune Odonate n’a jamais été observée.

L’espèce à laquelle nous nous interesserons est la Trithemis annulata (Palisot de Beauvois, 1807).

Il s’agit d’une espèce de libellule thermophile originaire d’Afrique, désormais bien implantée dans le sud de l’Europe. Elle est reconnaissable à sa coloration violette chez les mâles adultes et fréquente principalement les milieux aquatiques stagnants ou à faible courant. Son expansion récente en France en fait une espèce intéressante pour étudier l’influence des conditions environnementales, comme l’artificialisation des milieux, sur sa répartition. Elle est particulièrement sensible aux températures élevées, ce qui en fait un bon indicateur des effets du changement climatique.

Formatage des données pour le modèle d’occupancy

Pour utiliser la fonction occu() du package unmarked, il faut fournir trois objets :

- *Y* : une matrice d’observation (présence/absence) avec les sites en lignes et les visites en colonnes,
- *siteCovs* : un tableau des covariables mesurées au niveau du site (pour modéliser la probabilité d’occupation ψ),
- *obsCovs* : une liste de matrices contenant les covariables mesurées lors de chaque visite (pour modéliser la détection p).

Étape 1 – Construction de la matrice de détection Y

Nous filtrons d’abord le jeu de données pour ne conserver que les observations correspondant à l’espèce Trithemis annulata, notre espèce cible.

Ensuite, nous construisons une matrice Y dans laquelle chaque ligne représente un site (identifié par son Code_Maille) et chaque colonne une visite (correspondant à un passage annuel).

Les colonnes sont créées à partir des années de suivi (year2018, year2019, year2022, year2023) disponibles dans le jeu sites. Chaque année est dupliquée autant de fois qu’il y a eu de passages cette année-là (4 colonnes pour 2019 et 2022 , 3 pour 2018 et 2023).

Nous remplaçons les 1 (visites prévues) par des 0 (absence par défaut) et les 0 par des NA (absence de visite).

Enfin, nous parcourons les données d’observation filtrées pour repérer les passages où Trithemis annulata a été détectée, et nous remplaçons les 0 par des 1 dans la matrice Y aux endroits correspondants.

Ce format final permet d’obtenir une matrice binaire adaptée au modèle d’occupancy, indiquant pour chaque visite si l’espèce a été observée ou non.

Étape 2 – Préparation des covariables de site (siteCovs)

Pour modéliser la probabilité de présence en fonction de l’artificialisation, nous extrayons les informations pertinentes du fichier land_use.

Nous commençons par renommer la colonne ID en Code_Maille afin d’assurer la compatibilité avec les autres jeux de données.

Ensuite, nous filtrons uniquement les lignes correspondant à un buffer de 500 mètres autour des sites, puis nous sélectionnons les colonnes Code_Maille et MOS11.

La variable MOS11 représente le taux d’artificialisation autour de chaque maille, que nous utiliserons comme covariable explicative de la probabilité d’occupation dans le modèle.

Avant d’ajuster le modèle avec unmarked, il est essentiel de s’assurer que l’ordre des sites (Code_Maille) soit identique dans toutes les structures de données utilisées : Y, siteCovs et obsCovs.

On ne conserve que la variable MOS11 comme covariable pour la probabilité de présence.

Étape 3 – Préparation des covariables d’observation (obsCovs)

Pour modéliser correctement la probabilité de détection, nous avons sélectionné plusieurs covariables d’observation pertinentes : la température, la date (julian), le vent, ainsi que le recouvrement de végétation aquatique. Cette dernière peut jouer un rôle important car Trithemis annulata est une espèce thermophile associée à des plans d’eau spécifiques.

Afin d’intégrer ces covariables dans le modèle, nous reprenons la même structure que la matrice Y (c’est-à-dire mêmes sites en ligne et mêmes visites en colonnes). Pour chaque covariable, une matrice est créée avec uniquement des valeurs NA dans un premier temps. Elle sera ensuite remplie avec les moyennes par site et passage extraites des données brutes.

Ensuite, pour chaque covariable d’observation, on calcule la moyenne des valeurs par site, année et passage. Cela permet de remplir les matrices créées précédemment selon le bon format.

On commence par la température : pour chaque combinaison de site, année et numéro de passage, on extrait les températures observées et on en prend la moyenne (en ignorant les valeurs manquantes), qu’on insère à la bonne position dans la matrice Temperature.

Une fois toutes les matrices de covariables d’observation remplies (Température, Julian, Vent, etc.), on les regroupe dans une liste de matrices. Cette structure est requise pour les passer à la fonction unmarkedFrameOccu() via l’argument obsCovs.

Modèle d’occupation

Les équations du modèle sont :

$$\text{Logit}(\psi) = \beta_0 + \beta_1 \times \text{Artificialisation (MOS11)}$$

$$\text{Logit}(p) = \alpha_0 + \alpha_1 \times \text{Julian} + \alpha_2 \times \text{Température} + \alpha_3 \times \text{Vent} + \alpha_4 \times \text{Recouvrement aquatique}$$

où :

- β_0, β_1 sont les coefficients de la probabilité de présence (liés à l’artificialisation).
- $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ sont les coefficients de la probabilité de détection (liés aux conditions).

Étape 4 – Création unmarked

On commence par créer l’objet unmarkedFrameOccu à l’aide de la matrice de détections Y, des covariables de site siteCovs (ici : artificialisation MOS11) et des covariables d’observation obsCovs (Température, Julian, Vent, Recouvrement aquatique).

Ensuite, on ajuste le modèle d’occupancy avec la fonction occu() en spécifiant les équations de la probabilité d’occupation (ψ) et de détection (p) à l’aide des covariables choisies.

Enfin, on utilise summary() pour afficher les résultats du modèle, incluant les estimations des coefficients, leurs erreurs standard, les statistiques z, et les valeurs de p.

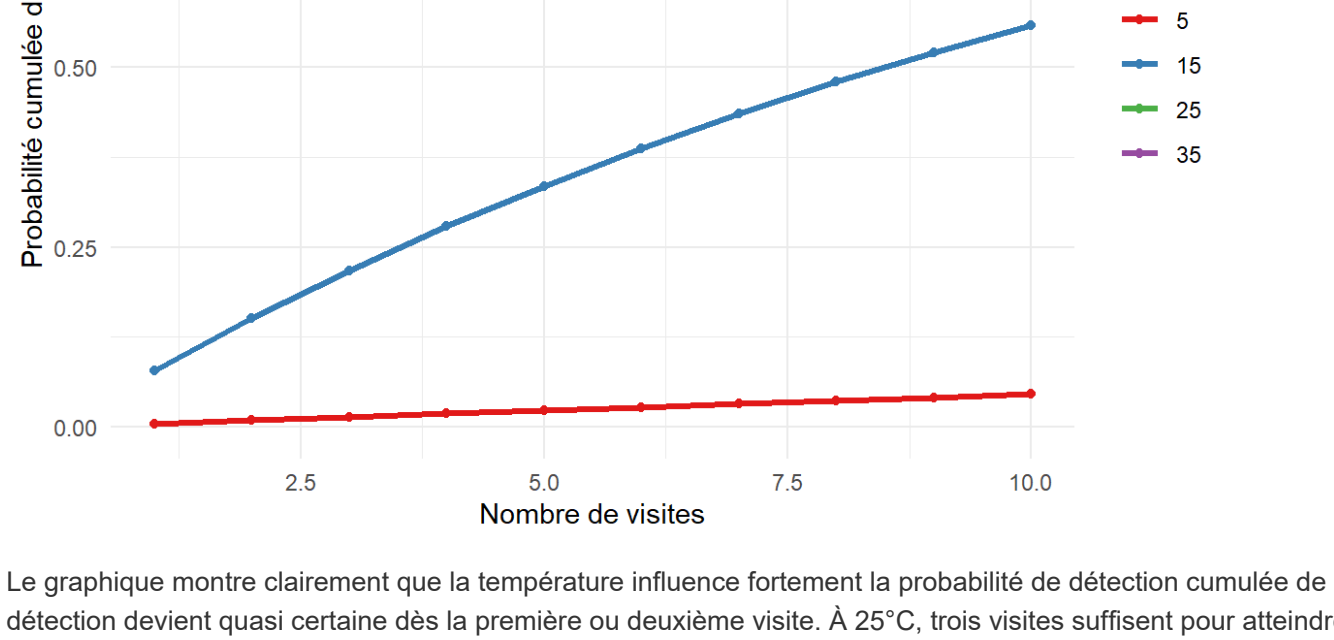
```
## 
## Call:
## occu(formula = ~Temperature + Julian + Vent + Recouvrement_Vegetation_Aquatique ~ 
##     MOS11, data = umf)
## 
## Occupancy (logit-scale):
##             Estimate    SE      z P(>|z|)
## (Intercept)   -0.225 0.496 -0.453   0.650
## MOS11          0.824 2.227  0.370   0.712
## 
## Detection (logit-scale):
##                                   Estimate      SE      z  P(>|z|)
## (Intercept)                       -7.25955 1.92264 -3.776 1.59e-04
## Temperature                        0.29105 0.06728  4.326 1.52e-05
## Julian                             0.00394 0.00617  0.639 5.23e-01
## Vent                               0.01197 0.31549  0.038 9.70e-01
## Recouvrement_Vegetation_Aquatique -0.01897 0.01355 -1.400 1.61e-01
## 
## AIC: 206.2504 
## Number of sites: 40
## ID of sites removed due to NA: 3 4 10 26 45
```

Le modèle ajusté sur les données réelles n’indique aucun effet significatif de l’artificialisation (MOS11) sur la probabilité d’occupation de Trithemis annulata (p = 0.712), bien que le coefficient soit positif. Cela suggère que, dans ce jeu de données, l’urbanisation autour des sites n’explique pas de manière claire la présence de l’espèce. Concernant la probabilité de détection, seule la température ressort comme hautement significative (p < 0.001), ce qui est cohérent avec le caractère thermophile de l’espèce. Les autres covariables (Julian, Vent, Recouvrement_Vegetation_Aquatique) ne sont pas significatives. Ces résultats doivent toutefois être interprétés avec prudence : les analyses de puissance précédentes ont montré que, dans un design similaire, la probabilité de détecter un effet réel était faible. Il est donc possible qu’un effet existe, mais que le design de l’étude ne permette pas de le mettre en évidence de manière robuste.

Puisque la température a un effet, interessons-en nous de plus près.

Combien de passages sont nécessaires pour obtenir une détection fiable ?

Afin d’estimer le nombre de visites nécessaires à une détection fiable, nous avons simulé la probabilité cumulée de détection en fonction du nombre de passages, pour différentes températures (5°C, 15°C, 25°C, 35°C). Ces valeurs représentent un gradient de conditions météorologiques, la température ayant été identifiée par le modèle comme le seul facteur significatif influençant la détection (p < 0.001). Pour chaque température, nous avons fixé les autres covariables à leur valeur moyenne, afin d’isoler l’effet de ce prédicteur. Cette analyse permet de déterminer à partir de combien de visites la probabilité de détection dépasse 80%, un seuil couramment utilisé pour juger de la fiabilité des observations.



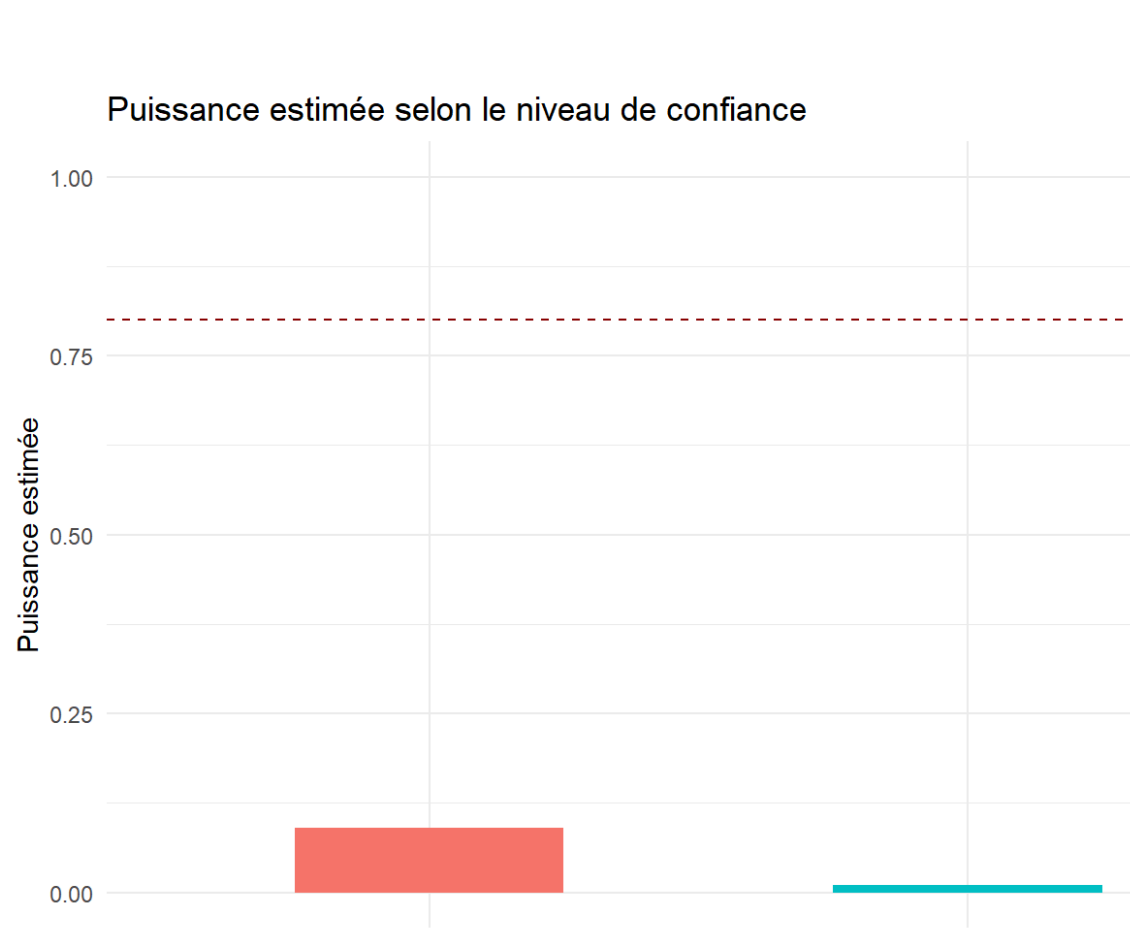
Le graphique montre clairement que la température influence fortement la probabilité de détection cumulée de Trithemis annulata. À 35°C, la détection devient quasi certaine dès la première ou deuxième visite. À 25°C, trois visites suffisent pour atteindre une probabilité de détection supérieure à 0.8, le seuil couramment utilisé pour juger une détection fiable. En revanche, à 15°C, il faut environ 10 visites pour approcher ce seuil, ce qui montre une détection beaucoup moins efficace. Enfin, à 5°C, la probabilité de détection reste très faible même après 10 passages, suggérant que l’espèce est peu active, voire absente, dans ces conditions. Ces résultats confirment le caractère thermophile de l’espèce, déjà identifié par le modèle, et soulignent la nécessité d’adapter l’effort d’échantillonnage aux conditions climatiques pour maximiser la détection.

Estimation de la puissance sur 100 simulations

Afin d’évaluer la capacité du protocole réel à détecter un effet de l’artificialisation, nous avons simulé 100 jeux de données en conservant la structure réelle d’échantillonnage et de covariables. Un effet simulé positif de l’artificialisation (β = 0.8) a été injecté dans le modèle d’occupation, puis testé à deux seuils de confiance : 90% (α = 0.10) et 95% (α = 0.05).

$H_0 : \beta \neq 0 \rightarrow$ pas d’effet

$H_1 : \beta = 0 \rightarrow$ effet significatif



Les résultats montrent une puissance statistique très insuffisante, avec un taux de détection significative d’environ 17% au seuil de 10%, et moins de 5% au seuil de 5%. Ces valeurs sont très en dessous du seuil communément attendu de 80%, indiqué par la ligne pointillée. Cela suggère que, même si un effet réel de l’artificialisation existait, le protocole actuel — incluant le nombre de sites, le nombre de visites et le taux de détection — ne permettrait pas de le mettre en évidence avec fiabilité. Ces résultats confirment l’importance d’ajuster la stratégie de suivi pour augmenter la puissance : en augmentant notamment le nombre de sites ou la fréquence des visites, ou en ciblant les conditions optimales de détection (par exemple, des températures élevées).

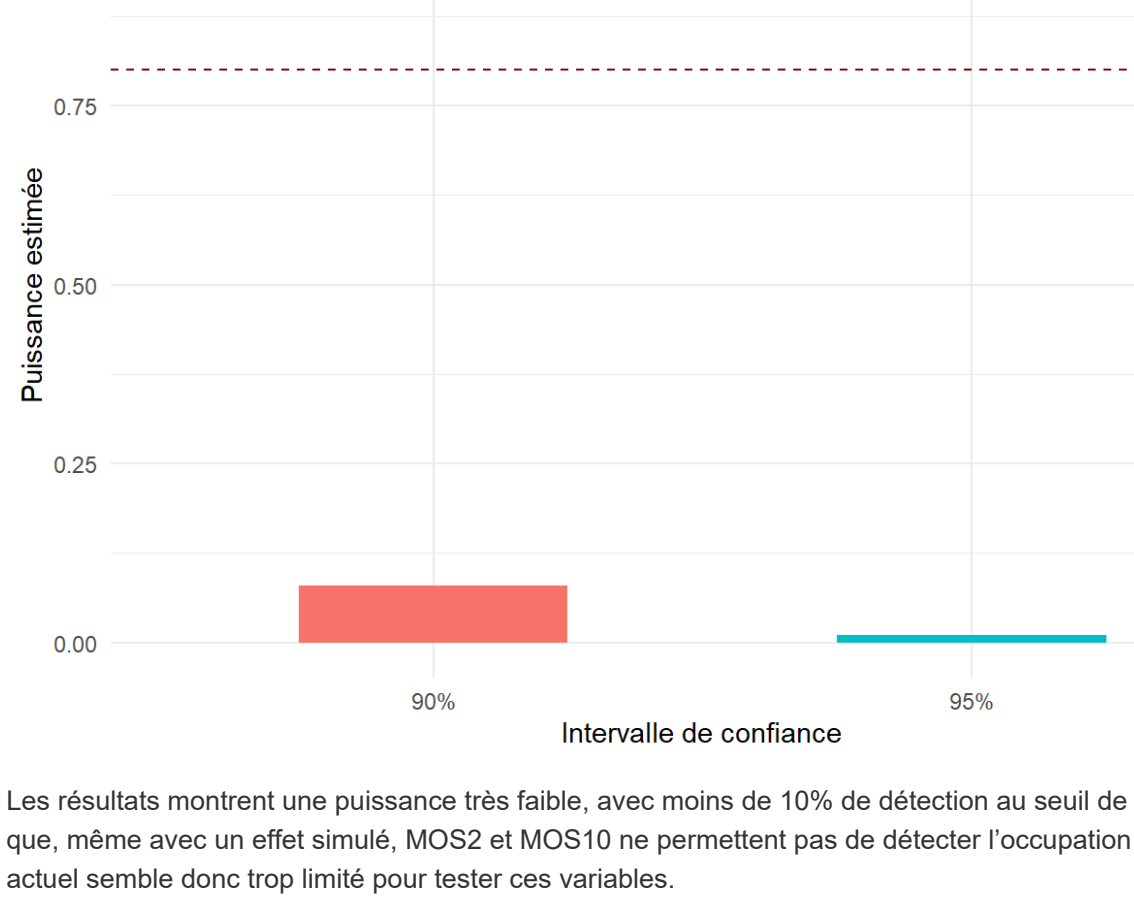
Test avec autres variables

` Dans la continuité de l’analyse précédente, nous avons testé l’influence d’autres variables paysagères issues du MOS pour expliquer la probabilité de présence de Trithemis annulata. Après avoir observé l’absence d’effet significatif pour MOS12 et MOS14, nous avons cette fois exploré deux nouvelles composantes : MOS2 (végétation arborée) et MOS10 (reseau hydrique). L’objectif est de vérifier si des habitats plus ouverts ou semi-naturels, comme les zones boisées (MOS2) ou les milieux en lien avec l’eau (MOS10), peuvent jouer un rôle complémentaire à l’artificialisation (MOS11) dans l’explication de la distribution de Trithemis annulata.

```
## 
## Call:
## occu(formula = ~Temperature + Julian + Vent + Recouvrement_Vegetation_Aquatique ~ 
##     MOS2 + MOS10, data = umf_multi)
## 
## Occupancy (logit-scale):
##             Estimate    SE      z P(>|z|)
## (Intercept)   -0.477  0.64 -0.745   0.456
## MOS2          -3.096  2.80 -1.104   0.270
## MOS10         38.245 25.08  1.525   0.127
## 
## Detection (logit-scale):
##                                   Estimate      SE       z  P(>|z|)
## (Intercept)                        -7.2244 1.92354 -3.7558 1.73e-04
## Temperature                         0.2882 0.06685  4.3118 1.62e-05
## Julian                              0.0041 0.00618  0.6640 5.07e-01
## Vent                                0.0216 0.31278  0.0692 9.45e-01
## Recouvrement_Vegetation_Aquatique  -0.0189 0.01342 -1.4059 1.60e-01
## 
## AIC: 194.6315 
## Number of sites: 40
## ID of sites removed due to NA: 3 4 10 26 45
```

Les résultats du modèle montrent que ni MOS2 (végétation arborée), ni MOS10 (réseau hydrique) n’ont d’effet significatif sur la probabilité d’occupation de Trithemis annulata (p > 0.1), bien que le coefficient de MOS10 soit relativement élevé. Cette absence d’effet rejoint les résultats précédents obtenus pour MOS12 et MOS14, confirmant que, dans le jeu de données actuel, aucune variable paysagère extraite du MOS ne permet d’expliquer significativement la distribution de l’espèce. En revanche, la température reste le seul prédicteur significatif de la détection, ce qui souligne une fois de plus le rôle central des conditions climatiques dans la détection et peut-être l’activité de cette espèce thermophile.

Estimation de la puissance sur 100 simulations



Les résultats montrent une puissance très faible, avec moins de 10% de détection au seuil de 10% et proche de 0% au seuil de 5%. Cela confirme que, même avec un effet simulé, MOS2 et MOS10 ne permettent pas de détecter l’occupation de manière fiable dans ce protocole. Le design actuel semble donc trop limité pour tester ces variables.

Conclusion

Ce travail a permis d’évaluer la capacité des modèles d’occupancy à détecter un effet de l’artificialisation sur la présence de Trithemis annulata, à partir de simulations et de données réelles. Les simulations ont montré que, même avec un effet réel modéré, la puissance statistique restait très faible dans le design actuel, en particulier au seuil de 5 %. L’analyse des données réelles confirme cette difficulté : aucun effet significatif de l’artificialisation (MOS11) ou des autres variables paysagères (MOS2, MOS10) n’a été mis en évidence. En revanche, la température apparaît de manière constante comme un facteur fortement lié à la probabilité de détection, soulignant le comportement thermophile de l’espèce. La simulation de détection cumulée montre qu’un effort d’échantillonnage adapté aux conditions climatiques est crucial. Même lorsque des effets simulés sont introduits sur MOS2 et MOS10, la puissance reste très faible, confirmant que le protocole actuel ne permet pas de détecter ces effets paysagers avec fiabilité.