Estimation de paramètres démographiques à partir d'individus non marqués

Marc J. Mazerolle

CEF-UQAT, Institut de recherche sur les forêts,
Université du Québec en Abitibi-Témiscamingue, Rouyn-Noranda

Combien de personnes dans la salle portent un vêtement noir?

Jour 1 de la formation

Modèles d'occupation de site à une saison

- formulation du modèle
- suppositions
- type de données
- logiciel PRESENCE
- test d'ajustement
- extensions du modèle

Modèles d'occupation de sites

À quoi servent-ils?

Estimer les patrons de présence d'une espèce donnée à des sites

Ces modèles réussissent à répondre à des questions telles que:

Quelle est la probabilité de trouver un pluvier siffleur à une série de sites sur le littoral?

Comment la distribution de la limoselle australe répond-elle à l'augmentation de l'activité pédestre?

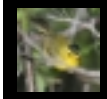
Modèles d'occupation de sites

À quoi servent-ils?

Estimer les patrons de présence d'une espèce donnée à des sites

Ces modèles réussissent à répondre à des questions telles que:

Est-ce que la présence de la martre d'Amérique dans une parcelle de forêt augmente avec la quantité de conifères?



Quelle est la probabilité d'utilisation des aulnaies par la paruline à calotte noire en période de nidification?

Modèles d'occupation de sites

Motivation derrière ces approches:

lors d'un inventaire on a la certitude que l'espèce est présente lorsqu'elle est détectée

par contre, une non-détection peut soit être due au fait que l'espèce soit réellement absente du site ou simplement qu'elle n'a pas été détéctée lors de l'inventaire

les deux possibilités ont des implications très différentes

D'où proviennent ces modèles?

Origine des modèles d'occupation de sites

Analyses de capture-marquage-recapture (*CMR*) pour populations fermées (Otis et al. 1978)

Grande famille de modèles développés pour estimer l'abondance

Pour mieux comprendre comment ça marche, un bref aperçu de l'estimation de l'abondance s'impose.

Estimation de l'abondance

Estimation de l'abondance

Problématique:

Ce qui nous intéresse, c'est la taille de la population (*N*).

Malheureusement, souvent impossible de recenser chaque individu dans une population

Stratégie normalement adoptée:

Dénombrer une partie des individus de la population

inventaires (pièges-fosses, points d'écoute, etc...)

Estimation de l'abondance

E(*C*): nombre d'individus observés lors d'inventaire

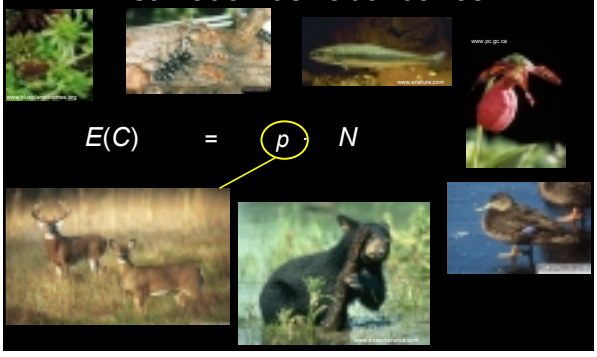
N: la taille réelle de la population

Relation entre les deux?

$$E(C) = p \cdot N$$

p: probabilité de détecter un individu dans la population

Estimation de l'abondance



$$E(C) = p \cdot N$$

Estimation de l'abondance

Ex. *E*(*C*) = 125 individus après un inventaire

E(*C*) = *pN*

Donc, *pN* = 125

N = ?

p peut varier entre 0 et 1

0.1 x 1250 = 125 $\hat{N} = 1250$

0.75 x 167 = 125 $\hat{N} = 167$

0.5 x 250 = 125 $\hat{N} = 250$

1 x 125 = 125 $\hat{N} = 125$

$E(C)$ = 125 individus après un inventaire

$E(C) = pN$

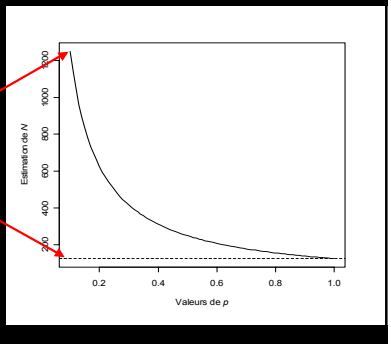
0.1 x 1250 = 125

$\hat{N} = 1250$

$E(C)$ = 125

Dès que $p < 1$, $E(C)$ sous estime la population réelle, N.

$\hat{N} \rightarrow \infty$ si $p \rightarrow 0$



Estimation de l'abondance

Ex. $E(C)$ = 125 individus après un inventaire

$E(C) = pN$

Donc, pN = 125

$\hat{N}$ = ?

En d'autres mots, il y a un nombre infini de combinaisons possibles de pN = 125.

On sait seulement que $\hat{N} \geq 125$.

Vous allez au garagiste pour réparer votre auto

Il vous répond:

« Il faut changer au moins 125 pièces, mais il pourrait y en avoir un nombre ∞ à changer. »

Serez-vous satisfait de son estimation?

Dans certains contextes, on n'accepte pas de réponses ambigües, pourquoi devrait-on l'accepter en écologie?

Estimation de l'abondance

Ça se complique lorsqu'on veut comparer deux ou plusieurs sites entre eux.

Ex.

Site 1 Site 2

$E(C_1) = 250$ vs $E(C_2) = 125$

$p_1N_1 = 250$ vs $p_2N_2 = 125$

Quel site a la plus grande abondance?

Scénario possible:

$p_1 = 0.5$ $p_2 = 0.125$

$\hat{N}_1 = 250/0.5 = 500$ $\hat{N}_2 = 125/0.125 = 1000$

$\hat{N}_1 < \hat{N}_2$

Estimation de l'abondance

Ça se complique lorsqu'on veut comparer deux ou plusieurs sites entre eux

Ex.

Site 1 Site 2

$E(C_1) = 250$ vs $E(C_2) = 125$

$p_1N_1 = 250$ vs $p_2N_2 = 125$

Autre scénario possible:

$p_1 = 0.5$ $p_2 = 0.25$

$\hat{N}_1 = 250/0.5 = 500$ $\hat{N}_2 = 125/0.25 = 500$

$\hat{N}_1 = \hat{N}_2$

Estimation de l'abondance

Ça se complique lorsqu'on veut comparer deux ou plusieurs sites entre eux

Ex.

Site 1 Site 2

$E(C_1) = 250$ vs $E(C_2) = 125$

$p_1N_1 = 250$ vs $p_2N_2 = 125$

Conclusion:

On ne peut rien dire à propos des deux sites sans SUPPOSER que $p_1 = p_2$.

Estimation de l'abondance

Suppositions communes (souvent non testées):

La probabilité de détection est constante à travers

- sites
- espèces
- observateurs (fatigue et expérience)
- type d'habitat
- technique d'échantillonnage
- conditions météo
- comportement
- sexe/âge
- coloration/morphologie
- temps de la journée
- saison
- année

Estimation de l'abondance

Impossible de contrôler tous ces facteurs

Il faut estimer *p*.

Comment estimer la probabilité de détection?

Estimation de l'abondance

Différentes avenues possibles pour estimer la probabilité de détection et l'abondance:

- Designs où les individus sont marqués individuellement (capture-marquage-recapture classique)
- Designs où les individus ne sont pas marqués (Distance sampling, *N*-mixture models)

Estimation de l'abondance

Une solution possible consiste à estimer directement la probabilité de détection et l'abondance à partir des données

Comment estimer *p* et *N* simultanément?

Étude de population

- individus identifiés
- plusieurs visites au site
- pour chaque individu capturé pendant l'étude, on détermine à chaque visite s'il a été capturé ou non

} Génère des histoires de captures

- période courte entre 1ière et dernière visite (population fermée aux changements démographiques)

Exemple:

Matrice d'histoires de capture

ID	Visite 1	Visite 2	Visite 3	Visite 4
#991	1	0	1	0
#992	1	1	1	0
#993	0	0	1	0
#994	1	1	1	0
#995	1	0	1	1
#996	1	1	1	1
...				

Application à d'autres problèmes

On peut appliquer la même approche à des sites visités pour estimer le nombre de sites où l'espèce d'intérêt est présente

À l'échelle des individus marqués à un même site:

N représente le nombre d'individus présents dans la population.

À l'échelle des sites visités:

N devient le nombre de sites occupés par l'espèce.

De retour aux modèles d'occupation de sites

Estimation de l'occupation

2 processus aléatoires influencent notre détection de l'espèce à un site lors d'un inventaire

I. Occupation
Site peut être occupé par l'espèce (avec probabilité ψ) ou non-occupé (avec probabilité 1-ψ).

II. Détection
Si le site est non-occupé: l'espèce ne peut être détectée.

Si le site est occupé: à chaque visite j, il y a une probabilité de détecter (p_j) ou non (1-p_j) l'espèce.

Estimation de l'occupation

2 processus aléatoires influencent notre détection de l'espèce à un site lors d'un inventaire

En d'autres mots:
La non-détection à un site n'implique pas que l'espèce y est absente.

Design

L'unité d'échantillonnage est le site
- le site est défini en fonction de l'écologie de l'espèce d'intérêt
- (e.g., parcelle de forêt, étang, île, quadrat, bûche)
- viser au moins une trentaine de sites (dépend de la complexité des modèles)

Collection de sites visités au moins deux fois
- à chaque visite de chaque site, on détermine si l'espèce est détectée ou non

Peut accomoder données manquantes
- (visites manquées à certains sites)

On veut estimer la probabilité de présence d'une espèce aux sites après avoir tenu compte de la détection.

On doit construire une matrice d'histoires de détection.

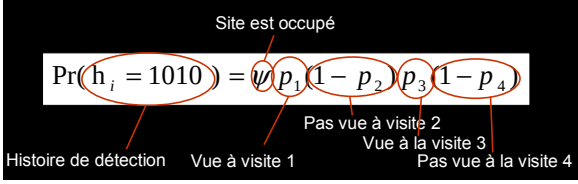
Matrice d'histoires de détection

Site ID	Visite 1	Visite 2	Visite 3	Visite 4
#991	1	0	1	0
#992	1	1	1	0
#993	0	0	1	0
#994	1	1	1	0
#995	1	0	1	1
#996	1	1	1	1
...				

Matrice d'histoires de détection

Site ID	Visite 1	Visite 2	Visite 3	Visite 4
#991	1	0	1	0

On peut exprimer formellement la probabilité d'observer cette histoire de détection (h_i = 1010) avec l'énoncé de probabilité suivant:



Construire des énoncés de probabilités

On procède de la même façon avec tous les sites où l'espèce a été détectée.

Mais qu'est-ce qu'on fait avec les sites où l'espèce n'a pas été détectée (i.e., h_i = 0000)?

Sites sans aucune détections

2 possibilités:

I. L'espèce était présente, mais n'a pas été détectée:

$$\psi (1 - p_1)(1 - p_2)(1 - p_3)(1 - p_4)$$

Sites sans aucune détections

2 possibilités:

I. L'espèce était présente, mais n'a pas été détectée:

$$\psi (1 - p_1)(1 - p_2)(1 - p_3)(1 - p_4) = \psi \prod_{j=1}^{4} (1 - p_j)$$

OU

II. L'espèce n'était pas présente sur le site.

$$1 - \psi$$

Sites sans aucune détections

Solution:

Inclure les deux possibilités dans l'énoncé de probabilité.

$$\Pr(\mathrm{h}_1 = 0000) = \psi \prod_{j=1}^{4} (1 - p_j) + (1 - \psi)$$

Construire le «likelihood » du modèle

$$L(\psi, p \mid \mathrm{h}_1, \mathrm{h}_2, \ldots \mathrm{h}_s) = \prod_{i=1}^{s} \Pr(\mathrm{h}_i)$$

La vraisemblance des probabilités de présence et de détection étant donné les histoires de détection.

Ce sont les énoncés de probabilité définis pour chaque site.

Construire le «likelihood » du modèle

$$L(\psi, p \mid \mathrm{h}_1, \mathrm{h}_2, \ldots \mathrm{h}_s) = \prod_{i=1}^{s} \Pr(\mathrm{h}_i)$$

On solutionne l'équation avec le maximum de vraisemblance:

Processus iteratif, trouve les valeurs des paramètres qui maximisent le « likelihood » de la fonction.

En d'autres mots:
Trouve les valeurs les plus vraisemblables pour les paramètres (ψ and p_j) à partir des données récoltées.

Solution par maximum de vraisemblance

Dans certains cas, on peut trouver la solution à un problème simplement en isolant la variable.

3X + 2 = 10

3X = 10 - 2

X = 8/3

La régression linéaire simple avec les moindres carrés.

Pour des situations plus complexes, comme dans des *GLM*'s ou modèles *CMR*, souvent impossible d'isoler variables.

Il faut opter pour d'autres stratégies.

Exemple de likelihood

On trouve une pièce de monnaie et on veut savoir quelle est la probabilité d'obtenir « face » avec cette pièce.

On la lance 20 fois et dénombre les « piles » et « faces »:

```
> flips
 [1] "pile" "face" "face" "face" "face" "face" "face" "face"
"pile" "pile" "face" "pile" "face" "face" "face" "face" "face"
"face" "face" "pile"
```

```
> table(flips)
flips
face pile
  15    5
```

Si quelqu'un nous demande la probabilité d'obtenir « face » avec cette pièce, on pourrait dire qu'elle est certainement plus grande que celle d'obtenir « pile ».

Exemple de pièce de monnaie

On a une expérience binomiale (succès vs échec)

à chaque lancer, on a deux réponses possibles (pile ou face)

On peut donc utiliser la distribution binomiale pour nous aider

Définition:

$$P(x \mid n, p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

Définie par deux paramètres

n = nombre total d'essais

p = probabilité d'observer 1 succès

x = nombre de succès pour *n* et *p* donnés

En mots:
la probabilité d'observer *x* succès, après *n* essais lorsqu'on a une probabilité *p* d'observer 1 succès

dans R:

```
dbinom(x= , size= , prob= )
```

Exemple de pièce de monnaie

Une propriété importante du likelihood:

$$L(\theta \mid x) = P(x \mid \theta)$$

le likelihood correspond à la fonction de densité de probabilité

(probabilité que *x* soit égal à une certaine valeur étant donné le ou les paramètres, θ)

On peut utiliser la fonction de densité (correspondant au type de variable aléatoire considérée) pour trouver le likelihood.

Pour notre exemple de pièce de monnaie on obtient:

$$L(n=20, p \mid x=15) = P(x=15 \mid n=20, p)$$

Exemple de pièce de monnaie

On utilise la distribution binomiale

$$L(\theta \mid x) = P(x=15 \mid n=20, p) = \frac{20!}{15!(20-15)!} p^{15} (1-p)^{20-15}$$

on sait que *p* peut varier entre 0 et 1

substituons différentes valeurs dans la fonction

p = 0.5 $= \frac{20!}{15!5!} 0.5^{15}(1-0.5)^5 = 0.0148$

p = 0.6 $= \frac{20!}{15!5!} 0.6^{15}(1-0.6)^5 = 0.0746$

p = 0.9 $= \frac{20!}{15!5!} 0.9^{15}(1-0.9)^5 = 0.0319$

De façon plus générale, on peut tracer la forme de la fonction du likelihood selon les valeurs possibles de *p*.

Exemple de pièce de monnaie

```
>theta<-seq(from=0, to=1, by=0.01)
```
on fait varier *p* de 0 à 1 à intervalle de 0.01

```
>Like<-dbinom(x=15, size=20,
prob=theta)
```
on calcule le likelihood pour chacune des valeurs de *p*

```
>plot(Like~theta, ...)
```
on fait le graphique de la courbe

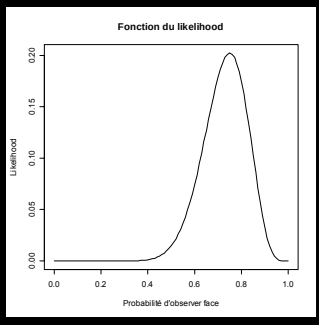
```
   >max(Like)
   [1] 0.2023312
```
on détermine la valeur maximale de likelihood obtenue

```
   >which(Like==max(Like))
   [1] 76
```
on trouve le numéro de l'observation de *p* qui correspond au maximum du likelihood

```
   >theta[76]
   [1] 0.75
```
on trouve la valeur de *p* qui correspond au maximum du likelihood

Exemple de pièce de monnaie

```
>theta<-seq(from=0, to=1, by=0.01)
>Like<-dbinom(x=15, size=20,
prob=theta)
>plot(Like~theta, ...)
   >max(Like)
   [1] 0.2023312
   >which(Like==max(Like))
   [1] 76
   >theta[76]
   [1] 0.75
```

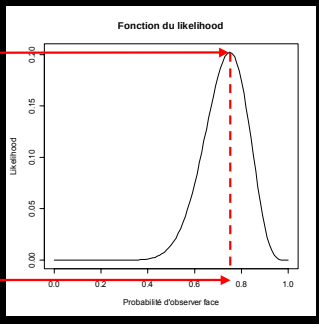


Exemple de pièce de monnaie

Valeur de la fonction au maximum de la courbe: 0.202

x / *n* est l'estimateur du maximum likelihood de *p*

Estimé de *p* au maximum de la courbe (maximum likelihood estimate, *MLE*): 0.75



Exemple de pièce de monnaie

Donc, si on observe *x* = 15 « faces » sur *n* = 20 lancer, on a une probabilité de 0.75 d'observer « face » avec cette pièce de monnaie.

0.75 est la valeur la plus probable selon la fonction du likelihood avec les données récoltées.

On peut étudier la fonction de likelihood pour différentes tailles d'échantillon

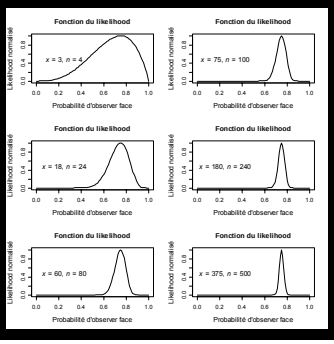
Forme de la fonction du likelihood

La forme de la fonction varie en fonction de la quantité d'information

$p = 0.75$

L'incertitude autour de l'estimé p diminue lorsque la taille d'échantillon est grande

le pic est bien défini avec n très grand
(pic peu évident avec $n = 4$)



Solution par maximum de vraisemblance

Même approche utilisée pour estimer les paramètres dans les *GLM*'s et la plupart de modèles de capture-recapture

Souvent plus efficace numériquement de travailler avec le log de $L(\theta \mid x)$, le log-likelihood.

la solution est invariable à la transformation
(on obtient le pic à la même valeur de θ)

ML génère des estimés distribués normalement (utilisation de test-*t*, *IC*)

On peut aussi utiliser le likelihood ou log-likelihood:

Pour construire un intervalle de confiance

Pour tester la différence entre deux modèles qui diffèrent seulement d'un paramètre (likelihood ratio test)

Suppositions du modèle de base (ψ constant et p constant)

1. L'état d'occupation à un site donné est constant entre la première et dernière visite (i.e., parcelle occupée reste occupée)

i.e., espèce ne s'éteint ni ne colonise de nouveaux sites.

2. Probabilité d'occupation est la même pour tous les sites.

3. Probabilité de détecter l'espèce pendant une visite, si elle est présente, est la même pour tous les sites.

4. Pour un site donné, la détection de l'espèce pendant une visite est indépendante de la détection aux autres visites.

Flexibilité des suppositions

Parfois ces suppositions ne peuvent être respectées

e.g., Occupation varie entre les sites dû au type d'habitat ou à la taille de la parcelle

e.g., Détection varie entre sites dû aux mauvaise conditions météo (comportement animal, effet d'observateurs) ou parce que l'espèce est plus évidente à certaines périodes

BONNE NOUVELLE:

Si covariables sont mesurées pendant l'étude, on peut les utiliser pour augmenter la plausibilité du modèle et expliquer les patrons qui nous intéressent vraiment.

Covariables

Sur l'occupation, ψ

- Covariables de site qui peuvent influencer la présence (type d'habitat, aire, structure de végétation, perturbations)
- Constantes à travers le temps

Sur la détection, *p*

- Variables qui peuvent influencer la détection (température, temps de la journée, effort d'échantillonnage, technique)
- Peuvent varier dans le temps (mesurées à chaque visite/site)

Comment incorporer les covariables dans le modèle?

Modéliser l'occupation et la détection

Ψ et *p* sont des probabilités et on veut qu'elles varient entre 0 et 1.

La fonction de lien logit (« logit link function ») impose une contrainte sur les paramètres afin qu'ils varient entre 0 et 1.

La fonction de lien logit:

$$\text{logit}(\psi) = \log\left(\frac{\psi}{1-\psi}\right)$$

On utilise le même lien dans une régression logistique.

$$\log\left(\frac{y}{1-y}\right) = \beta_0 + \beta_{Habitat} * Habitat + \beta_{Aire} * Aire$$

Modéliser l'occupation et la détection

Ψ et *p* sont des probabilités et on veut qu'elles varient entre 0 et 1.

La fonction de lien logit (« logit link function ») impose une contrainte sur les paramètres afin qu'ils varient entre 0 et 1.

Modèle avec occupation constante (intercepte seulement)

$$\text{logit}(\psi) = \beta_0$$

Lorsqu'on connaît β_0, on peut trouver ψ

$$\psi = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}$$

Inclure des covariables

On procède de la même façon pour modéliser ψ (ou p) avec des covariables

$$\text{logit}(\psi) = \beta_0 + \beta_{Habitat} * Habitat + \beta_{Area} * Area$$

Donc, pour trouver ψ, on calcule:

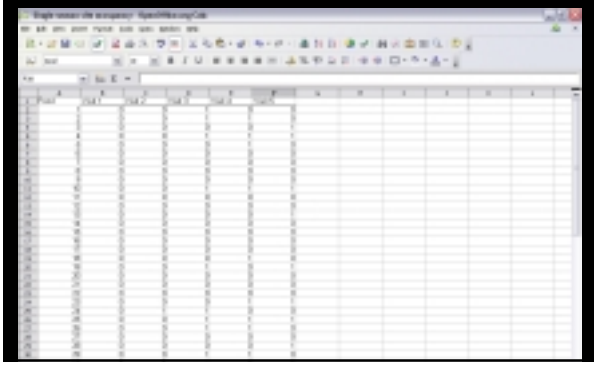
$$\psi = \frac{\exp(\beta_0 + \beta_{Habitat} * Habitat + \beta_{Area} * Area)}{1 + \exp(\beta_0 + \beta_{Habitat} * Habitat + \beta_{Area} * Area)}$$

Un exemple avec des grenouilles en étangs

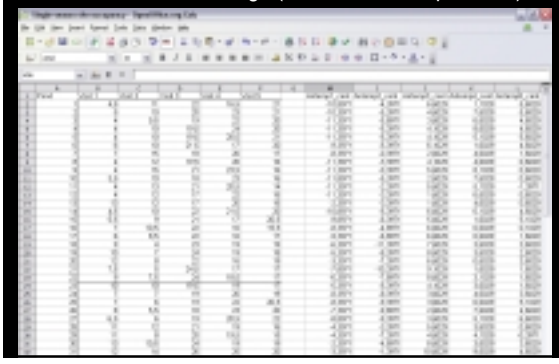
Adapté de Mazerolle et al. 2005 Ecol. Appl. **15**:824-834

34 étangs visités 5 fois pendant la saison de reproduction de la grenouille verte

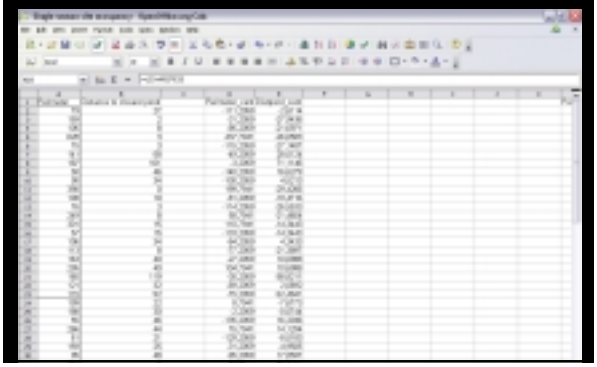
Les données



Variables d'échantillonnage (mesurées à chaque visite)



Variables de site



Et puis après?

Développer un programme pour maximiser le « likelihood » et trouver les estimés des paramètres dans le langage de programmation de notre choix (R, SAS, C++).

OU

Utiliser des logiciels qui permettent d'exécuter les analyses tels que MARK et PRESENCE.

MARK possède un très vaste éventail d'applications pour les problèmes dérivés des designs de marquage-recapture.

PRESENCE à été développé spécifiquement pour les analyses d'occupation de site.

R et certains packages.

Exemple avec R

Bien que des logiciels comme MARK et PRESENCE puissent ajuster les modèles de façon très efficace, c'est une bonne idée de se familiariser avec la mécanique des modèles avec un exemple simple: modèle $\psi(.)$ $p(.)$ - occupation constante et détection constante

On importe le jeu de données

```
> detect <- read.table(file = "Detect_frog.txt", header = TRUE)
> head(detect)
  Pond Visit1 Visit2 Visit3 Visit4 Visit5
1    1      0      0      1      0      0
2    2      0      0      1      1      0
3    3      0      0      0      0      1
4    4      0      0      1      1      1
5    5      0      0      0      1      0
6    6      0      0      0      0      0
```

Exemple avec R

On peut ensuite assembler les 0's et 1's en un vecteur d'histoires de détection à l'aide d'une boucle

```
> nsites <- length(detect[,1])

> hist<-rep(x = NA, times = nsites)
> for (i in 1:nsites) {
  hist[i]<-paste(detect[i, 2:6], sep="", collapse="")
  }

> hist
[1]  "00100" "00110" "00001" "00111" "00010" "00000" "00000" "00000"
[9]  "00000" "00111" "00000" "00000" "00001" "00000" "00000" "00000"
[17] "00000" "00010" "00101" "00000" "00000" "00000" "00011" "01100"
[25] "00111" "00110" "00000" "00001" "00110" "00100" "00010" "00000"
[33] "01110" "00000"
```

Exemple avec R

On détermine combien il y a de chaque type d'histoire de détection

```
> table(hist)
hist
00000 00001 00010 00011 00100 00101 00110 00111 01100 01110
   16     3     3     1     2     1     3     3     1     1
```

On traduit chaque type d'histoire de détection en énoncé de probabilité

00000: $\psi(1-p)(1-p)(1-p)(1-p)(1-p) + (1-\psi)$

00001: $\psi(1-p)(1-p)(1-p)(1-p)(p)$

01110: $\psi(1-p)(p)(p)(p)(1-p)$

etc...

Exemple avec R

Rappel:

on multiplie toutes les histoires de détections observées pour chaque site pour obtenir le likelihood

$$L(\psi, p \mid h_i) = \prod_{i=1}^{nsites} \Pr(h_i) = [(\psi(1-p)(1-p)(1-p)(1-p)p)^3] * [(\psi ppppp)^0] * ...$$

Pour le log-likelihood (numériquement efficace), on prend le log du produit des h_i

$$\log L(\psi, p \mid h_i) = [3*\log(\psi(1-p)(1-p)(1-p)(1-p)p)] + [0*\log(\psi ppppp)] + ...$$

(le log d'un produit donne une somme de log)

Exemple avec R

On assemble le tout dans une fonction de vraisemblance pour le modèle $\psi(.)$ $p(.)$

```
> site.occ <- function(p) {
  psi <- p[1]
  p <- p[2]
  (16*log((psi*(1-p)*(1-p)*(1-p)*(1-p)*(1-p) + (1 - psi)))) +
  (3*log((psi*(1-p)*(1-p)*(1-p)*(1-p)*(p)))) +
  (3*log((psi*(1-p)*(1-p)*(1-p)*(p)*(1-p)))) +
  (log((psi*(1-p)*(1-p)*(1-p)*(p)*(p)))) +
  (2*log((psi*(1-p)*(1-p)*(p)*(1-p)*(1-p)))) +
  (log((psi*(1-p)*(1-p)*(p)*(1-p)*(p)))) +
  (3*log((psi*(1-p)*(1-p)*(p)*(p)*(1-p)))) +
  (3*log((psi*(1-p)*(1-p)*(p)*(p)*(p)))) +
  (log((psi*(1-p)*(p)*(p)*(1-p)*(1-p)))) +
  (log((psi*(1-p)*(p)*(p)*(p)*(1-p))))
  }
```

Exemple avec R

On maximise le log-likelihood de cette fonction

```
> neg.LL <- function(p) {-1*site.occ(p)}
```

```
> hold <- nlm(f = neg.LL, p = c(0.5, 0.5), hessian = TRUE)
```

nlm( ) est une fonction d'optimisation qui minimise le *LL* (on multiplie par -1 pour maximiser)

```
> hold
$minimum
[1] 79.4022
```

log-likelihood = -1*minimum

```
$estimate
[1] 0.6434338 0.2925475
...
```

estimés des paramètres $\psi = 0.64$ $p = 0.29$

```
$hessian
          [,1]      [,2]
[1,] 92.39316  90.49483
[2,] 90.49483 381.56944
...
```

solve(hold$hessian) donne la matrice de variance-covariance des estimés

SE's = sqrt(diag(solve(hold$hessian)))

```
[1] 0.11873590 0.05842721
```

Exemple avec R

Plus efficace numériquement de travailler sur l'échelle logit dans la fonction de log-likelihood pour le modèle $\psi(.)$ $p(.)$

```
> site.occ <- function(p) {
  psi<-(exp(p[1]))/(1+exp(p[1]))
  p<-exp(p[2])/(1+exp(p[2]))
  (16*log((psi*(1-p)*(1-p)*(1-p)*(1-p)*(1-p) + (1 - psi)))) +
  (3*log((psi*(1-p)*(1-p)*(1-p)*(1-p)*(p)))) +
  (3*log((psi*(1-p)*(1-p)*(1-p)*(p)*(1-p)))) +
  (log((psi*(1-p)*(1-p)*(1-p)*(p)*(p)))) +
  (2*log((psi*(1-p)*(1-p)*(p)*(1-p)*(1-p)))) +
  (log((psi*(1-p)*(1-p)*(p)*(1-p)*(p)))) +
  (3*log((psi*(1-p)*(1-p)*(p)*(p)*(1-p)))) +
  (3*log((psi*(1-p)*(1-p)*(p)*(p)*(p)))) +
  (log((psi*(1-p)*(p)*(p)*(1-p)*(1-p)))) +
  (log((psi*(1-p)*(p)*(p)*(p)*(1-p))))
  }
```

Exemple avec R

On maximise le log-likelihood de cette fonction

```
> neg.LL <- function(p) {-1*site.occ(p)}

> hold <- nlm(f = neg.LL, p = c(0, 0), hessian = TRUE)

> hold
$minimum
[1] 79.4022                    log-likelihood = -1*minimum

$estimate                                          ψ              p
[1]  0.5903013 -0.8830409 estimés des paramètres   β0 = 0.59      β0 = -0.88
                                                   sur l'échelle du logit
...
```

logit(ψ) = 0.59 — logit(p) = -0.88

ψ = exp(0.59)/(1+exp(0.59)) — p = exp(-0.88)/(1+exp(-0.88))

ψ = 0.64 — p = 0.29

valeurs identiques à la diapo précédente (*ML* invariable à différentes paramétrisations)

Exemple avec R

```
...

$hessian
          [,1]      [,2]
[1,] 4.862703  4.296795
[2,] 4.296795 16.353601
```

on peut inverser la matrice hessienne pour obtenir la matrice de variance-covariance

```
> vars <- diag(solve(hold$hessian))
```

la variance des estimés des paramètres sur l'échelle du logit sont obtenues en prenant la diagonale de l'inverse de la matrice hessienne pour obtenir

```
> SE <- sqrt(vars)
> SE
[1] 0.5175203 0.2822017
```

la racine carrée des variances des estimés donnent les *SE*'s des estimés

Exemple avec R

On peut calculer un *IC* autour des estimés

```
> CL95 <- hold$estimate[1]+1.96*c(-SE[1], SE[1])
> CL95
[1] -0.4240386  1.6046411
```

sur l'échelle du logit

On peut faire la transformation inverse des bornes

```
> expit(CL95)
[1] 0.3955508 0.8326660

expit <- function(mu) {
exp(mu)/(1+exp(mu))
}
```

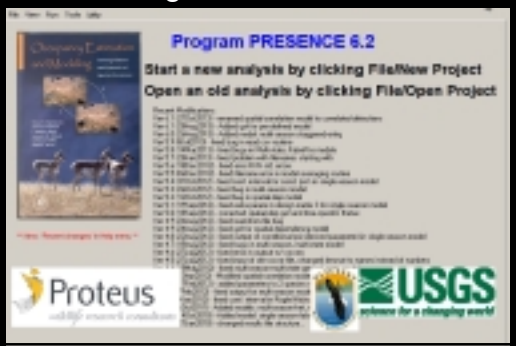
sur l'échelle originale

$$\hat{\psi}\,(IC95\%) = 0.64\,(0.39,\ 0.83)$$

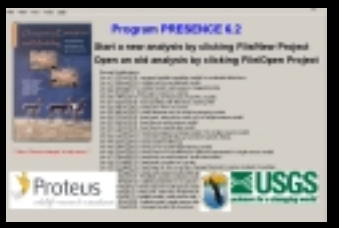
Cette méthode de calcul des *IC* performe mieux surtout lorsque le paramètre est près d'une borne (0 ou 1) – celle utilisée par PRESENCE.

Analyse à l'aide du logiciel PRESENCE

Program PRESENCE



Installation



Disponible gratuitement:
http://www.mbr-pwrc.usgs.gov/software.html

VÉRIFIER RÉGULIÈREMENT POUR UNE NOUVELLE VERSION

Installation

Fonctionne sur plate-forme `Windows`

pour d'autres OS (`Mac` ou `Linux`), on peut utiliser un émulateur `Windows` comme `Wine` ou carrément utiliser un programme de virtualisation comme `VirtualBox` ou `VMware Server`

L'archive `.zip` contient le fichier exécutable d'installation

IMPORTANT:
avant d'installer une nouvelle version, il est important de désinstaller l'ancienne version

Aide dans PRESENCE

Petit tutoriel avec des analyses de base

Un document en pdf sur l'installation et les analyses de base

Un hyperlien vers une série de documents pdf sur différents types de modèles à ajuster avec PRESENCE

Un hyperlien vers le forum d'aide sur les analyses CMR et occupation de site qui peut être interrogé

liste de trafic moyen avec une dizaine de messages par jour

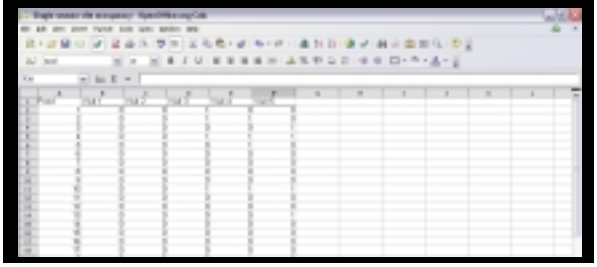
Aide dans PRESENCE

IMPORTANT:
Vérifier toutes les ressources disponibles dont les nombreux documents pdf, la litérature primaire et le forum d'aide, AVANT de lancer une question sur le forum.

Autrement, votre question risque de rester sans réponses ou pire, quelqu'un vous répondra RTFM...

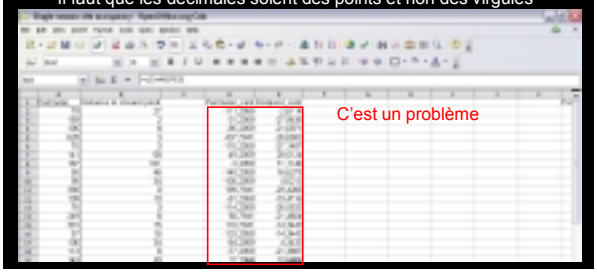
Préparer les données pour importation dans PRESENCE

Il faut au moins une matrice d'histoires de détection



Préparer les données pour importation dans PRESENCE

Il faut que les décimales soient des points et non des virgules



Préparer les données pour importation dans PRESENCE

Il faut que les décimales soient des points et non des virgules

On peut:

1) modifier les préférences linguistiques (options régionales) du système d'exploitation

2) modifier les préférences linguistiques du logiciel (`OpenOffice Calc`, `MS Word`)

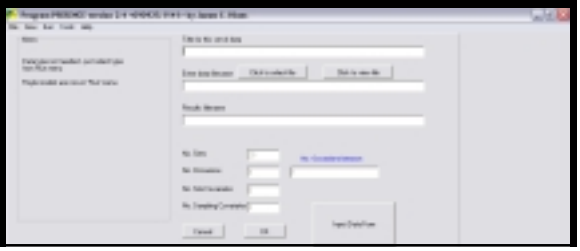
3) rechercher et remplacer les « , » par des « . » dans le logiciel

4) copier-coller dans un éditeur de texte (`Bloc notes`, `Tinn-R`) avant de faire un chercher et remplacer

Importation dans PRESENCE

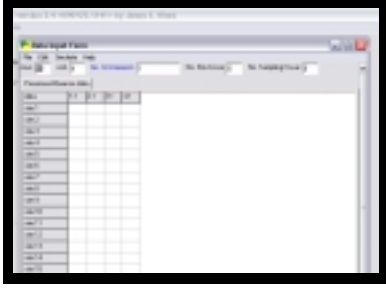
Pour l'instant, pas possible d'importer directement les données, il faut procéder par copié-collé.

On doit commencer par créer un nouveau projet (File –> New project).



Importation dans PRESENCE

On peut aller sur le chiffrier pour entrer les données en cliquant sur « Input Data Form »:



spécifier le nombre de sites (rows)

spécifier le nombre de total de visites (cols)

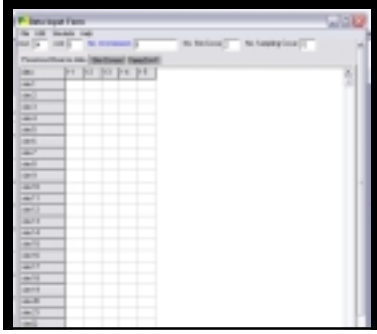
spécifier le nombre de visites par saison (No. Occ/season)

si plusieurs saisons, virgule sépare les visites/saison

spécifier le nombre de variables pour l'occupation (No. Site Covar)

spécifier le nombre de variables pour la détection (No. Sampling Covar)

Importation dans PRESENCE

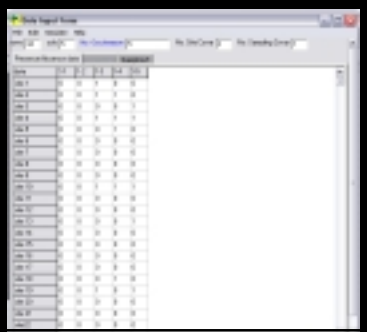


Vérifier que le nombre de lignes et colonnes est correct

avec des variables sur l'occupation ou la détection, il s'ajoute un chiffrier pour chacun.

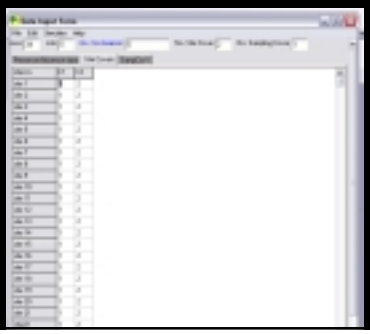
On peut ensuite copier-coller directement dans le chiffrier de PRESENCE à l'aide du menu (Edit -> Paste values)

Importation dans PRESENCE



un seul copié-collé pour les histoires de détection

Importation dans PRESENCE



Pour les variables de sites

Importation dans PRESENCE

S'il y a des variables continues, il faut s'assurer que la moyenne est près de 0 avant d'importer les données (évite problèmes d'estimation).

Stratégies possibles pour changer l'échelle d'une variable:

- Diviser les valeurs par une constante:
  e.g., 0.8 au lieu de 80%.
- Soustraire la moyenne de chaque valeur (centrer)
- Soustraire la moyenne de chaque valeur et diviser par l'écart-type (standardiser, normaliser)
- Transformation logarithmique

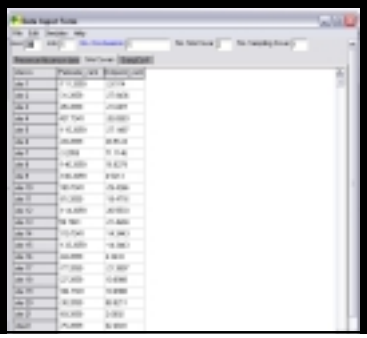
Importation dans PRESENCE

S'il y a des variables catégoriques, on doit les coder en variables binaires avant de les importer dans PRESENCE (pas de texte admis)

Ex. Type d'habitat - mixte, résineux ou feuillu

Site	Type		Mixte	Feuillu	
1	mixte	on doit créer une série de variables binaires	1	0	
2	resin		0	0	niveau de référence: resin
3	feuillu	avec *k* niveaux, on a besoin de *k* – 1 variables binaires	0	1	
4	feuillu		0	1	l'intercepte correspond au niveau de référence
5	mixte		1	0	

Importation dans PRESENCE



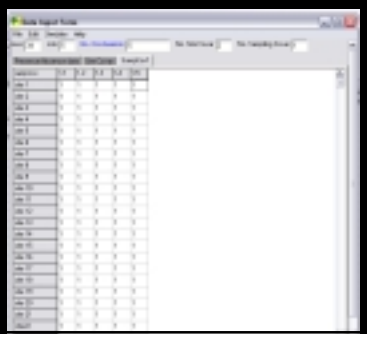
un seul copié-collé pour les variables sur l'occupation à l'aide du menu
(Edit -> Paste values)

ou

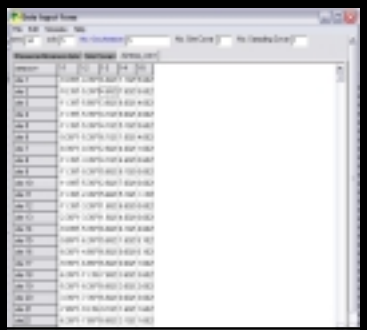
(Edit -> Paste w/sitenames)

on peut changer le nom d'une covariable
(Edit -> Rename covariate

Importation dans PRESENCE



Importation dans PRESENCE



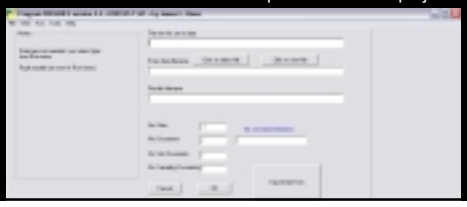
un seul copié-collé pour les variables sur la détection

Importation dans PRESENCE

On peut ensuite sauvegarder le fichier de données en `.pao`
(File -> Save as)

À la question « Use last col of data as frequency? », on répond « non »

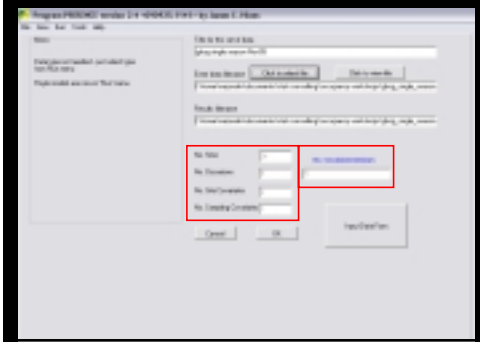
On retourne à la fenêtre de départ du nouveau projet:



On spécifie le fichier qu'on a créé

« Click to select file »

Importation dans PRESENCE



on vérifie que tous les champs sont corrects

finalement, on clique sur OK

Importation dans PRESENCE

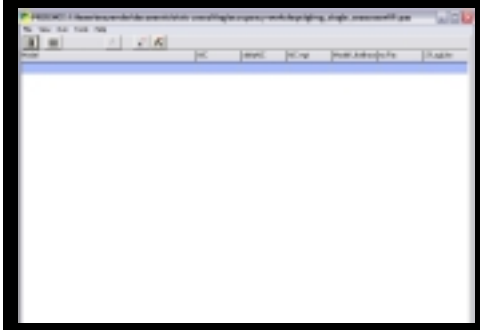


tableau de sélection de modèle prêt à recevoir des résultats

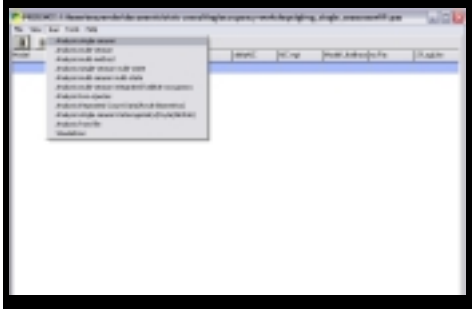
Importation dans PRESENCE

Une fois un projet créé, toutes les analyses effectuées dans ce projet seront stockées dans le même fichier `.pa2`.

On peut ensuite y retourner pour faire d'autres modèles ou enlever certains modèles.

Ces fichiers ne sont pas supprimés si on désinstalle PRESENCE ou qu'on installe une version plus récente.

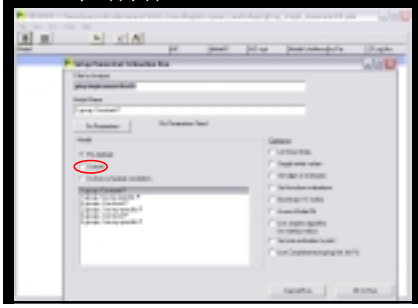
Préparation à l'analyse



Il faut choisir le bon type d'analyse

Faisons tourner un modèle simple

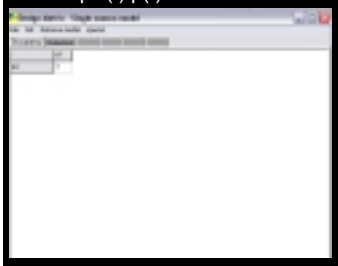
Modèle psi(.) p(.)



on choisit l'option « custom »

Faisons tourner un modèle simple

Modèle psi(.) p(.)



une matrice de design apparaît pour l'occupation

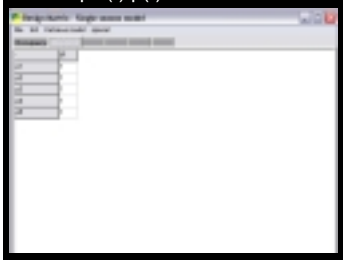
c'est à partir de la matrice de design qu'on spécifie les paramètres à estimer dans le modèle

par défaut, on obtient un modèle avec seulement l'intercepte sur l'occupation et l'intercepte sur la détection

le 1 indique l'intercepte

Faisons tourner un modèle simple

Modèle psi(.) p(.)

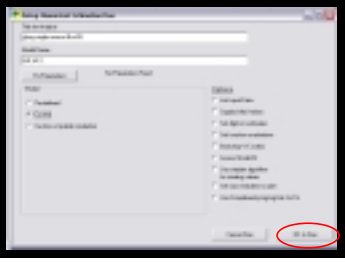


une matrice de design apparaît pour la détection

la colonne de 1 indique que la détection est constante pour toutes les visites

Faisons tourner un modèle simple

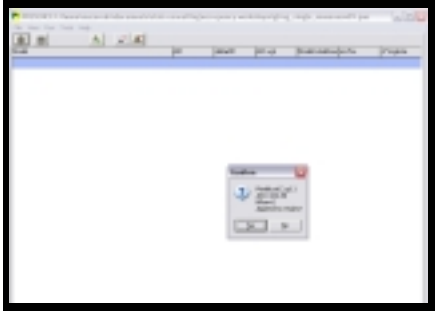
Après avoir modifié les matrices de design en conséquence, on peut faire rouler le modèle en retournant à la fenêtre « Setup Numerical Estimation Run ».



on donne un nom au modèle (ce nom sera utilisé dans le tableau de comparaison des modèles)

on clique sur « OK to Run »

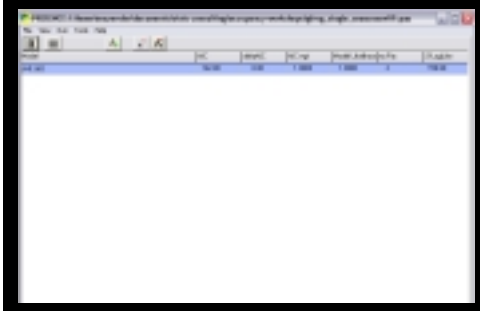
Faisons tourner un modèle simple



Lorsque l'algorithme a convergé, on confirme l'ajout des résultats au tableau de résultats

On clique « yes » pour ajouter les résultats du modèle au tableau général

Faisons tourner un modèle simple



Résultats intégrés au tableau d'*AIC*

Faisons tourner un modèle simple

Modèle psi(.) p(.)

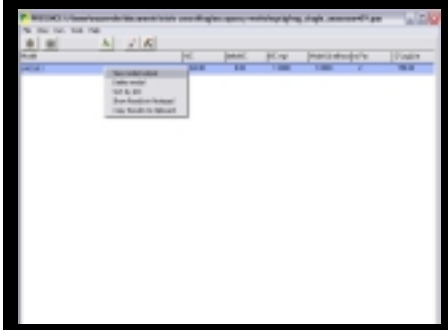
Output est automatiquement stocké dans le fichier de projet.

Ce qu'il faut vérifier:

- L'ajustement du modèle global (goodness of fit)
- Messages d'erreurs à propos de l'algorithme qui n'a pas trouvé de solution (convergence) ou calcul de la matrice de variance-covariance
- *SE*'s anormalement élevées pour les estimés des paramètres

Un ou plusieurs de ces éléments peuvent indiquer des problèmes et qu'on ne peut se fier aux résultats.

Visualiser le fichier d'output

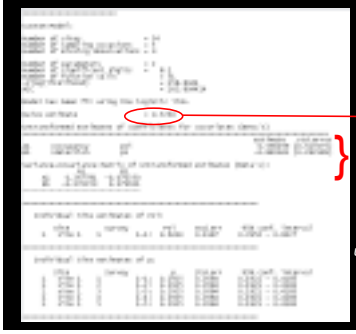


un clique du bouton droit de la souris permet d'ouvrir le fichier d'output

Visualiser le fichier d'output

détails sur le nom du modèle du fichier de jeu de données, type de modèle, matrices de design utilisées, etc...

Visualiser le fichier d'output



détails sur le nombre de sites, de visites, de données manquantes

nombre d'estimés, -2*LL, *AIC*, fonction de lien (logit(ψ) = β_0)

proportion de sites avec > 1 détection

estimés des paramètres sur l'échelle logit, *SE*'s, et matrice de variance-covariance

occupation prédite et *IC*

$\psi = e^{0.5903}/(1+e^{0.5903}) = 0.6434$

p de détection prédite et *IC*

Visualiser le fichier d'output

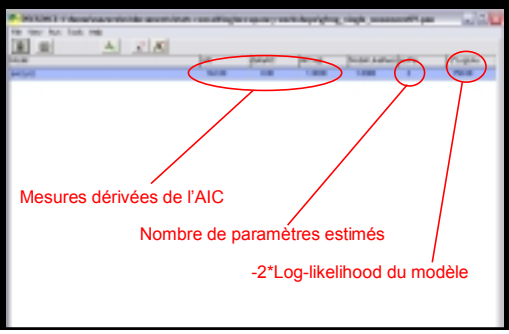
L'occupation correspond à la proportion des sites occupés après avoir corrigé pour la probabilité de détection

Le modèle donne une seule probabilité de présence et une seule probabilité de détection

pas surprenant, puisque le modèle qu'on a utilisé spécifiait que ψ était constant pour tous les sites, et que *p* était constant pour tous les sites et visites.

Le modèle $\psi(.)$ $p(.)$ n'est pas très intéressant en tant que tel, et est plutôt utilisé comme base de référence à des modèles plus complexes

La fenêtre du tableau des résultats



AIC et sélection de modèles

Le critère d'information d'Akaike (*AIC*)

«La beauté dans la simplicité»

En sciences, on essaie d'estimer la réalité à l'aide de modèles (e.g., déterminer les facteurs qui influencent l'occupation d'étangs)

Les bons modèles perdent peu d'information.
(bon modèle = bonne approximation de la réalité)

Akaike (1973): Trouve une relation entre le maximum de vraisemblance et l'information qui est perdue.

Développe un critère pour estimer la quantité d'information perdue (*AIC*).

AIC

Akaike Information Criterion (*AIC*)

$AIC = -2(\text{Log-likelihood}) + 2K$

Log-likelihood:
- mesure l'ajustement du modèle
- grande valeur indique modèle plus vraisemblable

K:
- nombre de paramètres estimés
- pénalise l'ajout de variables

AIC est un compromis entre l'ajustement et le nombre de paramètres (biais vs variance)

En d'autres mots: *AIC* trouve un modèle qui s'ajuste bien avec le moins de variables possible.

AIC pour petits échantillons

Dans les cas où (n/K < ~40), on utilise une version modifiée de l'*AIC* (AIC_c)

$$AIC_c = -2(\text{Log-likelihood}) + 2K + \frac{2K(K+1)}{(n-K-1)}$$

Facteur de correction

(tend vers 0 lorsque *n* augmente)

AIC pour petits échantillons

Dans les cas où (*n*/K < ~40), on utilise une version modifiée de l'*AIC* (AIC_c)

$$AIC_c = -2(\text{Log-likelihood}) + 2K + \frac{2K(K+1)}{(n-K-1)}$$

ON PEUT UTILISER *AICc* POUR TOUS LES CAS

AIC détails

Très utilisé dans la litérature de CMR mais s'applique à toute analyse standard (*GLM*'s, analyses temporelles)

AIC n'est pas un test *Ho*, n'est pas lié à alpha ou de notions de signification statistique (*P* < 0.05)

AIC se base sur le degré de preuve (strength of evidence)

AIC détails

AIC seul n'est pas très utile

C'est la différence entre modèles qui est intéressante.

Les modèles doivent être spécifiés a priori, et se basent sur la connaissance du système, publications, bon sens.

Modèle avec le plus petit *AIC* est le « meilleur » parmi les modèles candidats et pour les données utilisées

(Plus petit *AIC* = modèle qui perd le moins d'information)

Mesure du degré de preuve

Delta *AIC* (Δ_i) = AIC_i – min *AIC*

Delta *AIC*	Interprétation
< 2	modèle très probable
4 - 7	modèle moins probable
> 10	modèle très peu probable

Mesure du degré de preuve

Poids d'Akaike (Akaike weight, w_i)

Poids d'Akaike = w_i = $\frac{e^{\left(\frac{-\Delta_i}{2}\right)}}{\sum_{r=1}^{R} e^{\left(\frac{-\Delta_r}{2}\right)}}$

Correspondent au Delta *AIC* exprimés
sur une autre échelle pour que la somme des poids = 1.

Poids d'Akaike

Interprétation des poids d'Akaike:

la probabilité que le modèle soit le meilleur compte tenu des données et des modèles candidats

Ex. Un poids de 0.65 pour un modèle, indique qu'il a 65% de chance d'être le meilleur modèle parmi tous ceux que l'on a spécifié comme modèles candidats

Modèle	*K*	*AIC*	Delta *AIC*	Poids d'Akaike
modèle 1	2	12.3	0	0.65
modèle 2	4	14.45	2.15	0.22

on peut dire que le modèle 1 est 0.65/0.22 ≈ 3 fois meilleur que le deuxième (EVIDENCE RATIO)

Poids d'Akaike

Importance relative (w_{+j})

si on a le même nombre de modèles avec la variable vs sans variable, on peut faire la somme des poids d'Akaike des modèles qui incluent une variable d'intérêt

$w_{+\,var1}$ = 0.90 $w_{-\,var1}$ = 0.10

Inférence multimodèle

Model averaging (Inférence multimodèle)

Dans plusieurs cas, il y a plusieurs modèles qui peuvent « compétitionner pour la première place »

différents modèles peuvent bien expliquer les données

Si un modèle a un poids d'Akaike > 0.90, on peut utiliser ce modèle pour faire de l'inférence.

Si le poids d'Akaike du modèle au premier rang < 0.9, il est préférable de faire l'inférence à partir de tous les modèles

Inférence multimodèle

Principe:

- L'inférence repose sur tous les modèles, et non un seul meilleur.
- L'estimé d'une variable est pondérée par le poids d'Akaike.
- *SE* se calcule d'une façon similaire.

Approche très robuste, précise et élégante

AIC vs *Ho*

Ho

Stepwise, backward, etc… peuvent mener à différentes conclusions.

- Dépendent de l'ordre dans lequel les modèles sont roulés.

Modèles doivent être nichés.

AIC est insensible à ces problèmes.

AIC vs *Ho*

AIC

Avantages principaux:

- L'expérimentateur décide lui-même des modèles pertinents et ce choix ne dépend pas d'un algorithme.
- On peut comparer des modèles entre eux et déterminer lesquels sont les plus pertinents avec les poids d'Akaike.
  - Ex. Modèle A est 120 fois meilleur que le modèle B.
    - Modèle A nettement supérieur

AIC vs *Ho*

AIC

Avantages principaux:

- L'expérimentateur décide lui-même des modèles pertinents et ce choix ne dépend pas d'un algorithme.
- On peut comparer des modèles entre eux et déterminer lesquels sont les plus pertinents avec les poids d'Akaike.
  - Ex. Modèle A est 1.1 fois meilleur que le modèle B.
    - Modèles équivalents (inférence multimodèle)

De retour à notre exemple de grenouilles en étangs de tourbières

Les modèles candidats

Spécifions 4 modèles candidats:

psi(Distpond) p(Airtemp)	modèle le plus complexe
psi(.) p(Airtemp)	ici, le modèle le plus complexe est un modèle global (pas toujours le cas)
psi(Distpond) p(.)	les 3 autres modèles sont des versions plus simples du modèle plus complexe (nichés dans modèle complexe)
psi(.) p(.)	

important d'avoir un modèle global pour tester l'ajustement

Un modèle avec covariable

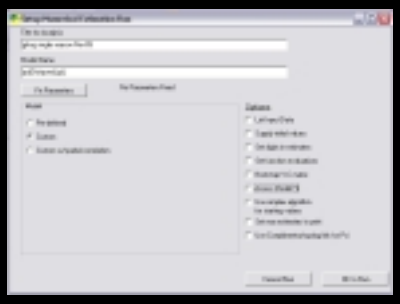
Modèle psi(Distpond) p(.)

l'occupation varie en fonction de la proximité d'étangs voisins

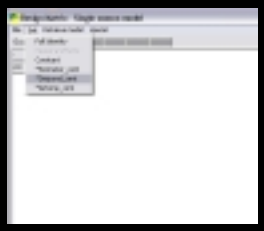
la probabilité de détection à un site donnée et une visite donnée est constante

Il faut aller dans la fenêtre « Setup Numerical Estimation Run »

Un modèle avec covariable



Un modèle avec covariables

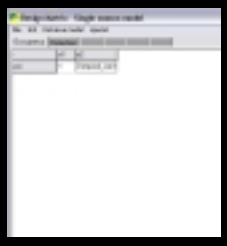


on modifie la matrice de design de psi

on doit ajouter une colonne après l'intercepte (clique du bouton droit de la souris)

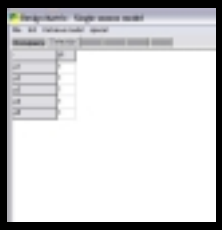
on doit insérer la bonne variable

Un modèle avec covariables



important de garder l'intercepte dans le modèle

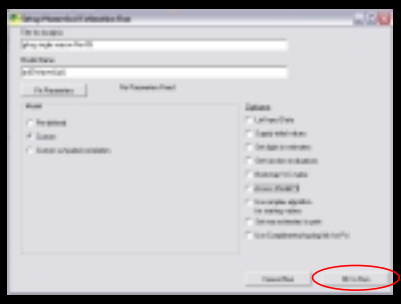
Un modèle avec covariables



pas besoin de modifier la matrice de design pour *p* puisque détection constante dans le modèle psi(Distpond)p(.)

on retourne dans la fenêtre « Setup Numerical Estimation Run »

Un modèle avec covariables



on peut ensuite faire rouler le modèle

Visualiser le fichier d'output

détails sur le nom du modèle du fichier de jeu de données, type de modèle, matrices de design utilisées etc...

Visualiser le fichier d'output

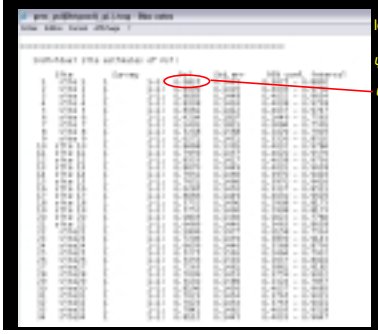
Estimés sur échelle logistique

Rappel:
logit(ψ) = β_0 + $\beta_{Distpond}$*Distpond

Insérer différentes valeurs de *Distpond* pour trouver la valeur à chaque site.

logit(ψ_1) = 0.6782 + (-0.0351*-2.6114) = 0.7699

Visualiser le fichier d'output



logit(ψ_1) = 0.7699

ψ_1 = exp(0.7699)/(1+exp(0.7699))

ψ_1 = 0.6835

Visualiser le fichier d'output

Distribution des valeurs de ψ

logit(p) = -0.9060

p = exp(-0.9060)/(1+exp(-0.9060))

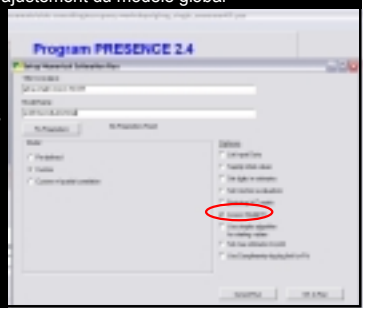
p = 0.2878

Ajustement (GOF)

Très important de vérifier l'ajustement du modèle global

Option
« Assess model fit »

Ne fonctionne qu'avec les modèles à une saison



Ajustement

Modèles nichés et non nichés

Modèles

Var1 Var2 Var3 Var4 Var5 — Modèle global

Var1 Var2
Var3 Var4 Var5
Var1 Var3 Var5
} Modèles plus simples nichés

Var7 Var8 Var9 Var10
Var1 Var2 Var3 Var4 Var1*Var3
Var1 Var2 Var3 Var1*Var3 Var2*Var3
} Modèles non nichés

On peut avoir plusieurs modèles globaux: routes différentes

Ajustement

Test basé sur χ^2 entre fréquence de sites observées pour chaque histoire de détection et celles prédites par le modèle.

Pour déterminer fréquences théoriques d'histoire 10101:

$$E_{10101} = s \times P(h_i = 10101)$$
$$= s \times \psi(p_1)(1-p_2)(p_3)(1-p_4)(p_5)$$

on connaît s, le nombre de sites

on connaît ψ, la probabilité de présence estimée

on connaît p, la probabilité de détection

Ajustement

Plusieurs fréquences d'histoires de captures sont faibles (< 2) , particulièrement lorsque plusieurs visites

distribution du χ^2 calculé ne suit pas la distribution théorique.

Solution:

Utilisation du bootstrap paramétrique pour déterminer la distribution théorique du χ^2

générer des données simulées à partir du modèle testé et effectuer un χ^2 à chaque fois

(on crée notre propre distribution)

Au moins 10 000 échantillons devraient être générés pour tester l'ajustement du modèle global.

Ajustement

On détermine si χ^2 observé est une observation commune lorqu'on génère des données qui suivent le modèle.

Si la valeur est observée moins de 5% des fois parmi la distribution des χ^2 simulés, cela indique potentiellement un mauvais ajustement.

On peut estimer $\hat{c}$

un indice d'ajustement du modèle concernant la variabilité des données selon un modèle testé (CMR, *GLM*).

$$\hat{c} = \frac{\chi^2_{observé}}{\chi^2_{bootsrap}}$$

Surdispersion

$\hat{c}$ = calculé par bootstrap paramétrique

si = 1, pas de surdispersion
si > 1, surdispersion

Assez commun de rencontrer de la surdipersion ($\sigma^2 > \mu$)

(overdispersion)

Si $\hat{c}$ << 1 ou $\hat{c}$ >> 4, modèle probablement pas adéquat

Surdispersion

On incorpore $\hat{c}$ dans le calcul de l'*AIC*

$$QAIC = \frac{-2(\text{log-likelihood})}{\hat{c}} + 2K$$

$$QAICc = QAIC + \frac{2K(K+1)}{n-K-1}$$

À noter qu'on inclut $\hat{c}$ dans le nombre de paramètres (*K*)

Surdispersion

On doit utiliser le $\hat{c}$ calculé pour le modèle le plus complexe dans tous les autres modèles subséquents

À NOTER: Il faut ajuster les *SE* des estimés par $\sqrt{\hat{c}}$

Ajustons le modèle global

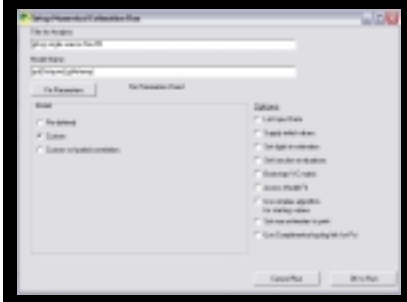
Modèle psi(Distpond)p(Airtemp)

l'occupation varie en fonction de la proximité d'étangs voisins

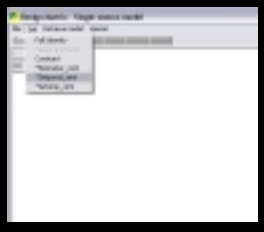
la probabilité de détection à un site donnée et une visite donnée varie en fonction de la température de l'air

Il faut aller dans la fenêtre « Setup Numerical Estimation Run »

Le modèle global



Le modèle global

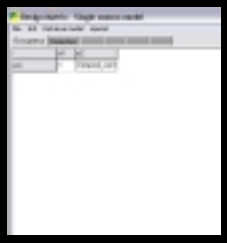


on modifie la matrice de design de psi

on doit ajouter une colonne après l'intercepte (clique du bouton droit de la souris)

on doit insérer la bonne variable

Le modèle global



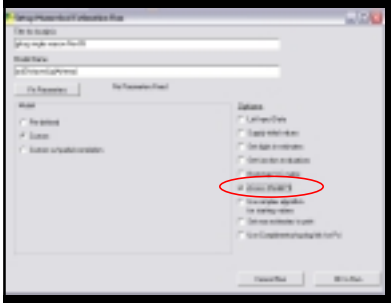
important de garder l'intercepte dans le modèle

Le modèle global

on modifie la matrice de design de *p*

on ajoute une colonne pour la variable

Le modèle global



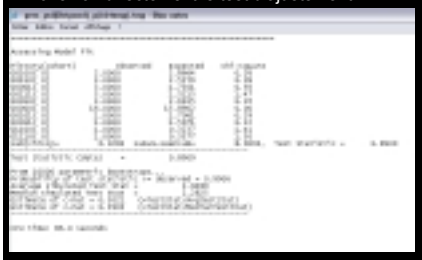
étant donné que c'est le modèle global, on va ajouter un test d'ajustement

il faut un minimum de 10 000 échantillons de bootstrap pour vérifier l'ajustement

avec de gros jeux de données et des modèles complexes, peut prendre du temps

Le modèle global

Allons voir directement le test d'ajustement



$$\hat{c} = \frac{\chi^2_{obs}}{\overline{\chi}^2_{boot}} = 0.30$$

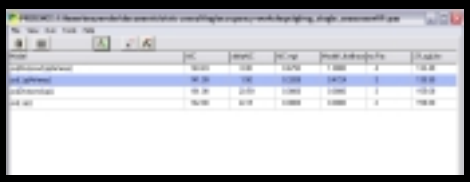
Estimé à 0.3 ne suggère pas de surdispersion: on utilise 1

On peut ajouter le dernier modèle

Modèle

psi(.) p(Airtemp)

On obtient ensuite



pour changer à *AICc*: « Tools –> Change effective sample size (0) » et spécifier nombre de sites

Sélection de modèle

Rang des modèles d'occupation d'étangs selon *AICc*

Modèle	Nombre de paramètres	Delta *AICc*	Akaike weight
psi(Distpond) p(Airtemp)	4	0	0.61
psi(.) p(Airtemp)	3	0.92	0.39
psi(Distpond) p(.)	3	20.01	0.00
psi(.) p(.)	2	21.32	0.00

Sélection de modèle

Rang des modèles d'occupation d'étangs selon *AICc*

Modèle	Nombre de paramètres	Delta *AICc*	Akaike weight
psi(Distpond) p(Airtemp)	4	0	0.61
psi(.) p(Airtemp)	3	0.92	0.39
psi(Distpond) p(.)	3	20.01	0.00
psi(.) p(.)	2	21.32	0.00

Modèle psi(Distpond) p(Airtemp) a 61% de chance d'être le meilleur modèle (plus parcimonieux).

Néanmoins, il est seulement ca. 1.6 fois (0.61/0.39) meilleur que le modèle au deuxième rang.

Sélection de modèle

Rang des modèles d'occupation d'étangs selon *AICc*

Modèle	Nombre de paramètres	Delta *AICc*	Akaike weight
psi(Distpond) p(Airtemp)	4	0	0.61
psi(.) p(Airtemp)	3	0.92	0.39
psi(Distpond) p(.)	3	20.01	0.00
psi(.) p(.)	2	21.32	0.00

On constate que les modèles avec Airtemp sur la détection sont nettement meilleurs que les autres qui ne l'ont pas ($w_+ = 1$).

De plus, Distpond n'a potentiellement pas beaucoup d'effet puisque le modèle avec Distpond seul n'a aucun poids.

Résultats

Deux modèles sont considérés équivalents.

Serait préférable de baser les conclusions sur l'ensemble des modèles avec l'inférence multimodèle « model averaging ».

En d'autres mots:

Au lieu de dépendre des résultats d'un seul modèle, on base nos conclusions sur l'information contenue dans l'ensemble des modèles (conclusions tiennent ainsi compte de l'incertitude dans la sélection de modèles).

Allons voir un exemple pour l'occupation prédite à un site qui est à 27.1 m d'un étang voisin.

Au lieu de baser nos conclusions sur un seul « meilleur modèle », nous allons baser la prédiction sur l'ensemble des modèles.

« Model averaging » pour l'estimé de ψ

(chaque modèle nous donne une prédiction)

Model averaging pour l'estimé de ψ.

Model	Delta AICc	Akaike weight	ψ	*SE*
psi(Distpond) p(Airtemp)	0	0.61	0.639	0.120
psi(.) p(Airtemp)	0.92	0.39	0.617	0.109
psi(Distpond) p(.)	20.01	0.00	0.684	0.145
psi(.) p(.)	21.32	0.00	0.643	0.119

Estimé pondéré de ψ:

Calcul de la moyenne de modèles (Model averaging, inférence multi-modèle)

Au lieu de se fier seulement aux estimés provenant du meilleur modèle, on peut baser nos calculs sur une moyenne pondérée de tous les estimés à travers tous les modèles

Approche robuste et précise

$$\bar{\hat{\theta}} = \sum_{i=1}^{R} w_i \hat{\theta}_i$$

w_i = poids d'Akaike

$\hat{\theta}_i$ = estimé du modèle *i*

« Model averaging » pour l'estimé de ψ

Model averaging pour l'estimé de ψ

Model	Delta *AICc*	Akaike weight	ψ	*SE*
psi(Distpond) p(Airtemp)	0	0.61	0.639	0.120
psi(.) p(Airtemp)	0.92	0.39	0.617	0.109
psi(Distpond) p(.)	20.01	0.00	0.684	0.145
psi(.) p(.)	21.32	0.00	0.643	0.119

Estimé pondéré de ψ: 0.630

Précision de l'estimé pondéré

On peut également calculer la *SE* des estimés de façon analogue

plus récente version (équation 6.12, Burnham and Anderson 2002):

SE inconditionnelle $= \sqrt{\sum_{i=1}^{R} w_i \, \mathrm{var}(\hat{\theta}_i \mid g_i) + (\hat{\theta}_i - \bar{\hat{\theta}})^2}$

une mesure du biais

« inconditionnelle » car elle n'est pas conditionnelle à un seul modèle, mais utilise l'information de tous les modèles

ancienne version (équation 4.9, Burnham et Anderson 2002):

SE inconditionnelle $= \sum_{i=1}^{R} w_i \sqrt{\mathrm{var}(\hat{\theta}_i \mid g_i) + (\hat{\theta}_i - \bar{\hat{\theta}})^2}$

« Model averaging » pour l'estimé de ψ

Model averaging pour l'estimé de ψ

Model	Delta *AICc*	Akaike weight	ψ	*SE*
psi(Distpond) p(Airtemp)	0	0.61	0.639	0.120
psi(.) p(Airtemp)	0.92	0.39	0.617	0.109
psi(Distpond) p(.)	20.01	0.00	0.684	0.145
psi(.) p(.)	21.32	0.00	0.643	0.119

Estimé pondéré de ψ: 0.630

SE inconditionnelle: 0.129

« Model averaging » pour l'estimé de ψ

Model averaging pour l'estimé de ψ

Model	Delta *AICc*	Akaike weight	ψ	*SE*
psi(Distpond) p(Airtemp)	0	0.61	0.639	0.120
psi(.) p(Airtemp)	0.92	0.39	0.617	0.109
psi(Distpond) p(.)	20.01	0.00	0.684	0.145
psi(.) p(.)	21.32	0.00	0.643	0.119

On peut procéder de la même façon pour les autres paramètres d'intérêt, comme l'effet de la distance à l'étang voisin (Distpond) sur l'occupation ou de la température de l'air sur la probabilité de détection.

« Model averaging » pour l'estimé β

On peut faire de l'inférence multimodèle pour les estimés de paramètres apparaissant dans les meilleurs modèles.

Pour ce faire, il faut faire un nouveau tableau d'*AICc* avec seulement les modèles incluant le paramètre d'intérêt:
il faut recalculer les poids d'Akaike.

MISE EN GARDE:
le paramètre doit avoir la même interprétation pour tous les modèles à partir desquels on veut calculer un estimé pondéré.

- un paramètre impliqué dans une interaction (ou effet quadratique ou polynôme supérieur), n'a pas la même interprétation dans un modèle avec interaction vs sans interaction
- il faut donc exclure les modèles qui contiennent une interaction impliquant l'effet d'intérêt

« Model averaging » pour l'estimé β

On peut calculer une moyenne pondérée pour l'estimé de Airtemp:

Model averaging pour l'estimé de Airtemp

Model	Delta *AICc*	Akaike weight	$\beta_{Airtemp}$	*SE*
psi(Distpond) p(Airtemp)	0	0.61	0.213	0.056
psi(.) p(Airtemp)	0.92	0.39	0.215	0.056

ici, les modèles avec Airtemp avaient tout le poids, donc les poids d'Akaike n'ont pas changé (mais ce n'est pas toujours le cas)

« Model averaging » pour l'estimé β

On peut calculer une moyenne pondérée pour l'estimé de Airtemp:

Model averaging pour l'estimé de Airtemp

Model	Delta *AICc*	Akaike weight	$\beta_{Airtemp}$	*SE*
psi(Distpond) p(Airtemp)	0	0.61	0.213	0.056
psi(.) p(Airtemp)	0.92	0.39	0.215	0.056

Estimé pondéré de $\beta_{Airtemp}$: 0.214

SE inconditionnelle: 0.056

IC inconditionnel à 95%: 0.104, 0.324

On conclut que la température de l'air a un effet positif sur la détection des grenouilles vertes.

Détection différente par visite

Si aucune variable de détection n'a été mesurée, on peut estimer une détection différente pour chaque visite (e.g., p(t)).

- implique que la détection est la même pour tous les sites à une visite donnée (différente détection à chaque visite)
- bonne idée d'en inclure parmi les modèles candidats

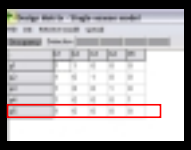
En présence de variables numériques, la détection peut différer d'un site à un autre à chaque visite (si chaque site a une valeur différente à chaque visite)

- souhaitable puisque l'hétérogénéité dans la détection est un problème

Détection différente par visite

Pour spécifier une détection différente par visite, il y a plusieurs stratégies de codage équivalentes:

Contrastes de traitement (comparaison à un niveau de référence)



ici, β_1 correspond à la détection pour le niveau de référence (T5)

β_2 correspond à la différence entre la détection à T1 et T5

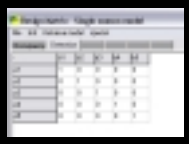
$\beta_3$correspond à la différence entre la détection à T2 et T5

β_4 correspond à la différence entre la détection à T3 et T5

β_5 correspond à la différence entre la détection à T4 et T5

Détection différente par visite

Matrice d'identité (diagonale de 1's - aucun intercepte):



ici, β_1 correspond à la détection à T1

β_2 correspond à la détection à T2

β_3 correspond à la détection à T3

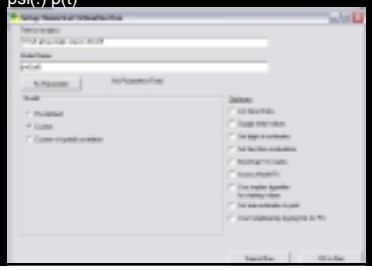
β_4 correspond à la détection à T4

β_5 correspond à la détection à T5

Détection différente par visite

Peu importe la stratégie de codage, l'important c'est de comprendre celle qu'on a utilisée pour bien interpréter les résultats

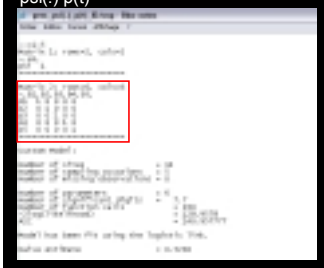
psi(.) p(t)



ici, matrice d'identité utilisée

Détection différente par visite

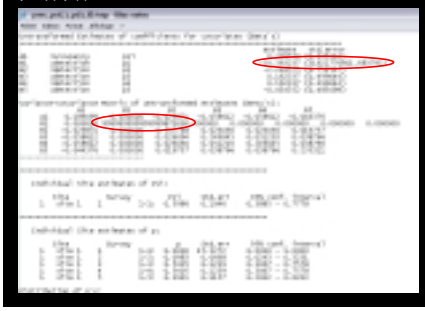
psi(.) p(t)



on confirme qu'une matrice d'identité a bien été utilisée (aucun intercepte – pas de colonne de 1's)

Détection différente par visite

psi(.) p(t)



indication qu'il y a un problème d'estimation quelque part

Détection différente par visite

on remarque dans les histoires de détection, qu'il n'y a jamais eu de détection à T1:

00100
00110
00001
00111
00010
00000
00101
00011
01100
01110

l'estimé β associé à T1 est très petit (-20.36)

```
> exp(-19.66)/(1+exp(-19.66))
[1] 2.895813e-09
```

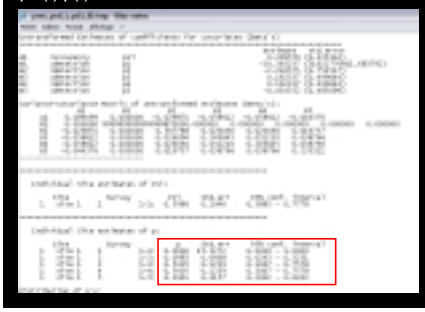
donc p à T1 = 0

c'est normal d'observer une telle valeur dans notre cas (pas eu de détection à la 1ière visite)

puisque p à T1 = 0, on pourrait fixer ce paramètre à 0 afin d'éviter d'entraîner des problèmes dans la matrice de variance-covariance

Détection différente par visite

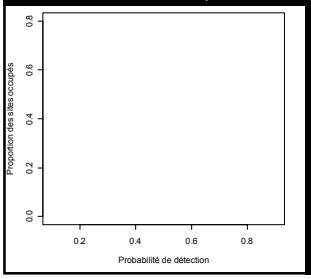
psi(.) p(t)



une détection
estimée par visite

Coûts de ne pas tenir compte de la probabilité de détection

Estimés naïfs vs d'occupation de site: Simulations avec 500 réplicats

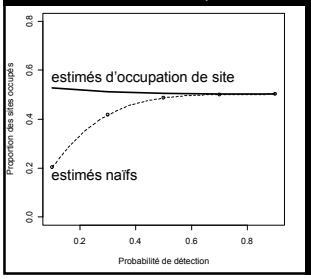


70 sites $\psi = 0.5$

5 visites $p = 0.1, 0.3, 0.5, 0.7, 0.9$

Coûts de ne pas tenir compte de la probabilité de détection

Estimés naïfs vs d'occupation de site: Simulations avec 500 réplicats



70 sites $\psi = 0.5$

5 visites $p = 0.1, 0.3, 0.5, 0.7, 0.9$

Estimés naïfs sous-estiment l'occupation lorsque la probabilité de détection < 1.

Estimation moins biaisée avec occupation de site

biais = $\theta - \hat{\theta}$

Coûts de ne pas tenir compte de la probabilité de détection

Estimés naïfs vs occupation:

Estimés naïfs sous-estiment l'occupation dès que la probabilité de détection < 1.

Plus sérieux encore:

On ne peut comparer les estimés naïfs provenant de différents habitats, période, etc... sans supposer que p est constant pour tous les inventaires et les habitats.

Meilleure approche consiste à estimer directement la probabilité de détection avec des modèles d'occupation de sites pour obtenir de meilleures estimations de la présence.

Suppositions non respectées

Fermeture démographique:

- Si l'espèce se déplace aléatoirement à l'intérieur et extérieur des sites, les estimés devraient être OK (estimés biaisés si mouvements non-aléatoires).

Hétérogénéité dans la probabilité de détection (i.e., différentes probabilités de détections pour chaque site):

- Problème important mais peu étudié.
- Mène à sous-estimer l'occupation, mais moins que les estimés naïfs (Mazerolle, Nichols et Hines en prép.).
- Présence de covariables peuvent régler le problème.
- En absence de covariables, on peut utiliser d'autres types de modèles.

Conseils de design

Efficacité augmente avec le nombre de sites et le nombre de visites

- Meilleurs estimés obtenus avec p élevées.
- Généralement, si $p > 0.3$, les estimés ne sont pas biaisés lorsqu'on a au moins 5 visites.
- Avec seulement 2 visites, p doit être > 0.5 pour donner de bons estimés (certains suggèrent au moins 3 visites).
- Si l'espèce est rare, investir en plus de sites.
- Si l'espèce est commune, investir en plus de visites.

Conseils de design

Autre scénario possible (removal design)

- On visite les sites un maximum de T visites.
- Après une détection, on ne visite plus le site, mais on continue à visiter les sites où l'espèce n'a pas été détectée.
- Dans ce cas particulier, on ne peut pas estimer une probabilité de détection différente pour chaque visite.

Incorporer covariables lorsque possible (hétérogénéité)

- Amène plus de flexibilité au modèle et peut régler plusieurs problèmes.

Applications

Espèces animales et végétales difficiles à détecter

Monitoring d'espèces sur plusieurs années et plusieurs sites (modèles multi-saisons)

Études en biologie de la conservation ou écologie du paysage à de grandes échelles spatiales

Dynamique des métapopulations

Applications moins conventionnelles

Estimation de richesse en espèces

- pour un site donné, on établit une liste d'espèces que l'on peut potentiellement détecter
- on note la détection/non-détection de chaque espèce lors de visites répétées au même site
- chaque espèce est considérée comme un « site » et on estime la proportion d'espèces (ψ) sur notre liste qui serait présente au site (on peut ajouter groupes taxonomiques sur ψ)
- la richesse en espèce est la somme des probabilités pour chacune des espèces

Estimation de la richesse en espèces

Estimation de la proportion d'espèces potentiellement présentes à un site visité à plusieurs reprises

ici, les espèces deviennent les « sites »

	T1	T2	T3	T4
sp 1	0	1	1	0
sp 2	0	0	0	0
sp 3	1	0	1	0
sp 4	0	1	1	0
...				

permet d'estimer *p* de détection pour groupes d'espèces (reproduction, taille)

et

la proportion des espèces dans la région présentes au site

Extensions au modèle à une saison

Modèle à multiples saisons

Modèles d'occupation à plusieurs saisons

(MacKenzie et al. 2003) - *dynamic occupancy model*

- Sites visités plusieurs fois pendant une saison
- Sites visités plusieurs années
- On permet à l'espèce de s'éteindre à un site occupé l'année précédente, ou de coloniser un site qui n'était pas occupé.
- On peut estimer la probabilité de colonisation (γ) et d'extinction (ε), pour modéliser l'état d'occupation (occupé une année, pas l'autre).
- Application directe aux modèles de métapopulation.

Modèle à multiples saisons

L'état du site (occupé ou non occupé) ne change pas à l'intérieur d'une saison.

Fonction de vraisemblance beaucoup plus complexe, mais semblable à celle des modèles à une saison.

Différente paramétrisation possible, pour estimer colonisation et extinction, ou occupation à chaque année.

Modèle à multiples saisons

On peut déduire un terme si on connaît les deux autres:

$$\psi_t = \psi_{t-1}(1 - \varepsilon_t) + (1 - \psi_{t-1})\gamma_t$$

P d'être occupé à *t*

P qu'un site occupé à *t* - 1 le soit encore à *t*

P qu'un site non occupé à *t* - 1 soit colonisé à *t*

Important: les dynamiques (extinction, colonisation) s'effectuent entre les deux saisons.

Autres types de modèles

Modèles d'hétérogénéité (Pledger 2000)

Estiment l'hétérogénéité dans *p* et groupent les sites ensemble selon ce qu'ils ont une haute ou faible détection.

Modèles d'hétérogénéité d'abondance (Royle-Nichols 2003)

Heterogeneité dans *p* est due à l'abondance.

Modèles d'indices d'abondance (Royle 2004a, Royle et Link 2005)

Utilisent des données récoltées sur une échelle ordinale (e.g., aucun, rare, abondant) au lieu de détecté ou non.

Modèles d'abondance (Royle 2004b)

Utilisent les données brutes d'abondance lorsque les individus ne sont pas identifiés individuellement.

Autres types de modèles

Modèles d'occupation multi-états (Nichols et al. 2007)

Estiment plusieurs états d'occupation et l'erreur de classification:

e.g., espèce présente, couple reproducteur, jeunes atteints l'envol

Modèles d'occupation de co-occurrence d'espèces (MacKenzie et al. 2004)

Estiment la probabilité de présence de deux espèces en tenant compte de la détection afin de déterminer si la présence de l'une est indépendante de la présence de l'autre

Modèles d'occupation avec détections corrélées (Hines et al. 2010)

Estiment la probabilité d'occupation avec des données le long de transects

Livres très utiles

MacKenzie et al. 2006

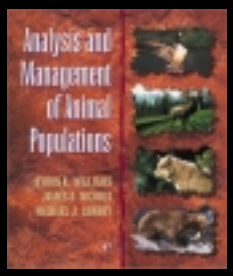
plusieurs saveurs de modèles d'occupation de sites

Livres très utiles

Royle et Dorazio 2008

construction flexible à partir de `R` et `WinBUGS` pour ajuster certains types de modèles incluant plusieurs types de modèles d'occupation de site

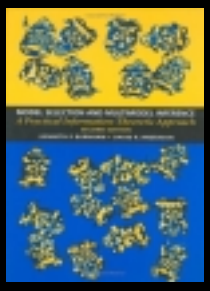
Livres très utiles



Williams, Nichols et Conroy 2002

la bible des techniques permettant l'estimation de la probabilité de détection (*CMR*, échantillonnage à distance)

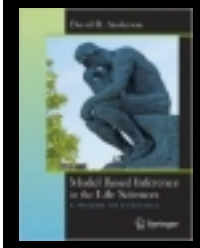
Livres très utiles



Burnham et Anderson 2002

l'évangile de la sélection de modèle et inférence multimodèle

Livres très utiles



Anderson 2008

une version abrégée et mise à jour
du matériel présenté dans Burnham et
Anderson 2002

Articles importants

Hines et al. 2010. Ecology 20:1456-1466.
Otis et al. 1978. Wildl. Monogr. **62**:1-135.
MacKenzie et al. 2002. Ecology **83**:2248-2255.
MacKenzie et al. 2003. Ecology **84**:2200-2207.
MacKenzie et al. 2004. J. Anim. Ecol. **73**:546-555.
MacKenzie et Royle 2005. J. Appl. Ecol. **42**:1105-1114.
Nichols et al. 2007. Ecology **88**:1395-1400.
Pledger 2000. Biometrics **56**:434-442.
Royle 2004a. Conserv. Biol. **18**:1378-1385.
Royle 2004b. Biometrics **60**:108-115.
Royle et Nichols 2003. Ecology **84**:777-790.
Royle et Link 2005. Ecology **86**:2505-2512.