INITIATION AUX TESTS STATISTIQUES ET AUX MODÈLES LINÉAIRES

Corentin M. Barbu

INRAE et Ecole Doctorale ABIES (réseau ADUM)

10 février 2025

Section 1

Introduction

Subsection 1

PRÉSENTATION

Objectif du cours

Corentin M. Barbu

- 2006-2014, épidémiologie
- 2014- à l'INRAE, contrôle des maladies et ravageurs (GC)

Objectif du cours

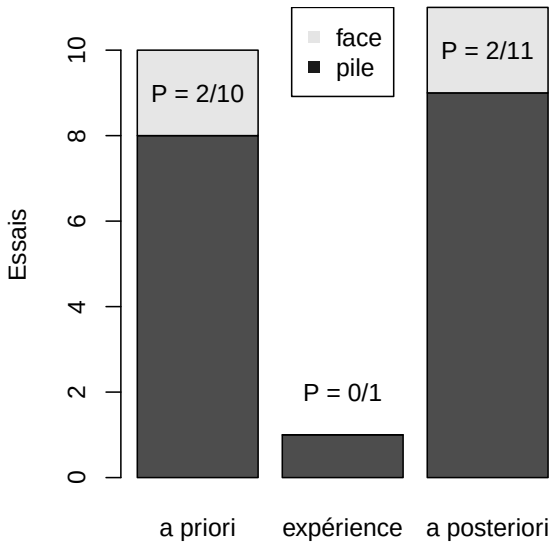
Donner les clés conceptuelles et méthodologiques pour réaliser des tests statistiques et modèles linéaires de manière rigoureuse (et lire des articles scientifiques de manière critique).

Section 2

Modélisation statistique

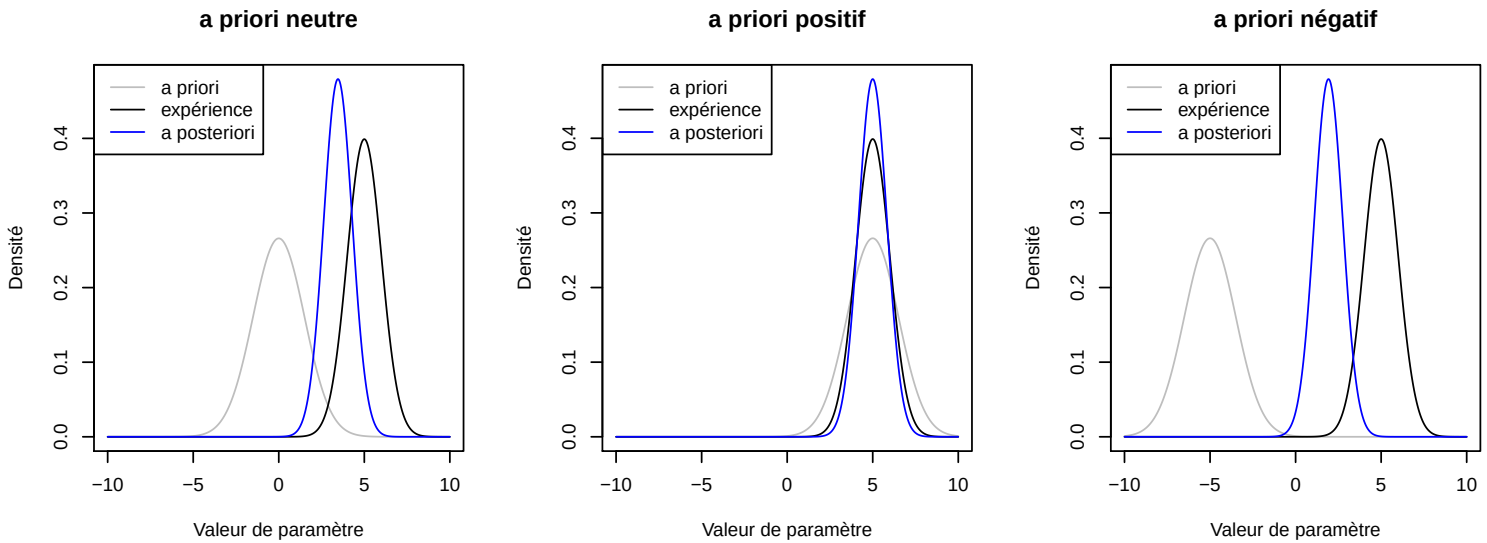
FORMALISMES STATISTIQUES

- fréquentiste (Fisher) : "pas" d'*a priori* sur les données (illusoire)
- bayésien : on spécicifie l'*a priori* sur les données



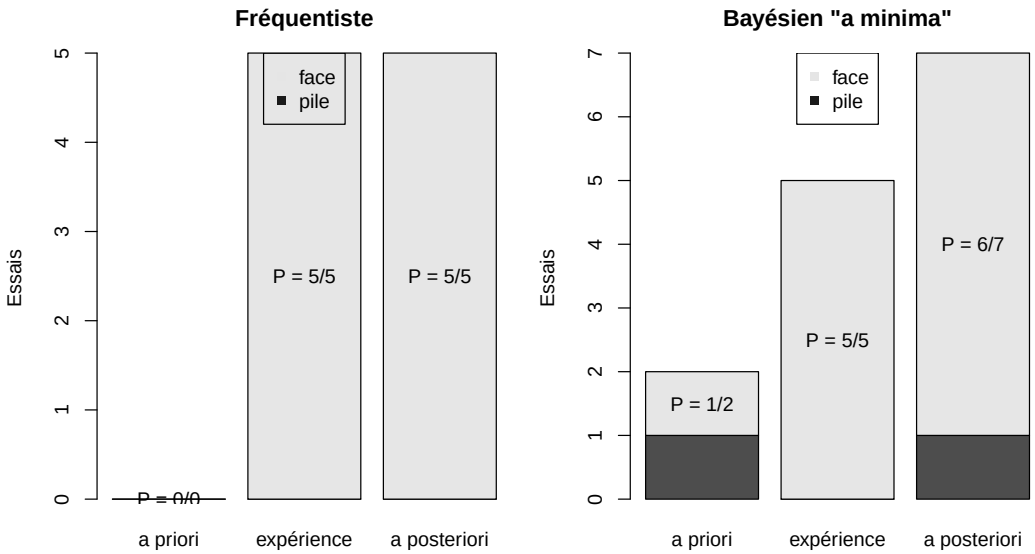
FORMALISMES STATISTIQUES

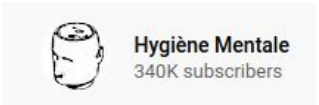
- fréquentiste (Fisher) : "pas" d'*a priori* sur les données (illusoire)
- bayésien : on spécicifie l'*a priori* sur les données



FORMALISMES STATISTIQUES

- fréquentiste (Fisher) : "pas" d'*a priori* sur les données (illusoire)
- bayésien : on spécicifie l'*a priori* sur les données



Voir aussi "Hygiène Mentale" 

(Ep 26 Bayesian Thinking) sur YouTube et séries de "science4all".

Subsection 1

BASES DE RÉGRESSIONS LINÉAIRES

Subsubsection 1

GLM et risque de la sélection de variable

Les bases de régression GLM (generalized linear model)

La régression linéaire de base

On considère que la valeur moyenne d'une variable de sortie y est fonction d'une combinaison linéaire de paramètres x:

$$\widehat{y} = \alpha + \beta \cdot x_1 + \gamma \cdot x_2 + \cdots$$

Souvent simplifié en utilisant les notations d'algèbres linéaires :

$$\widehat{y} = A + BX$$

Autour de cette valeur moyenne on considère dans le cas le plus simple que les valeurs observées ont une distribution Gaussienne (loi normale):

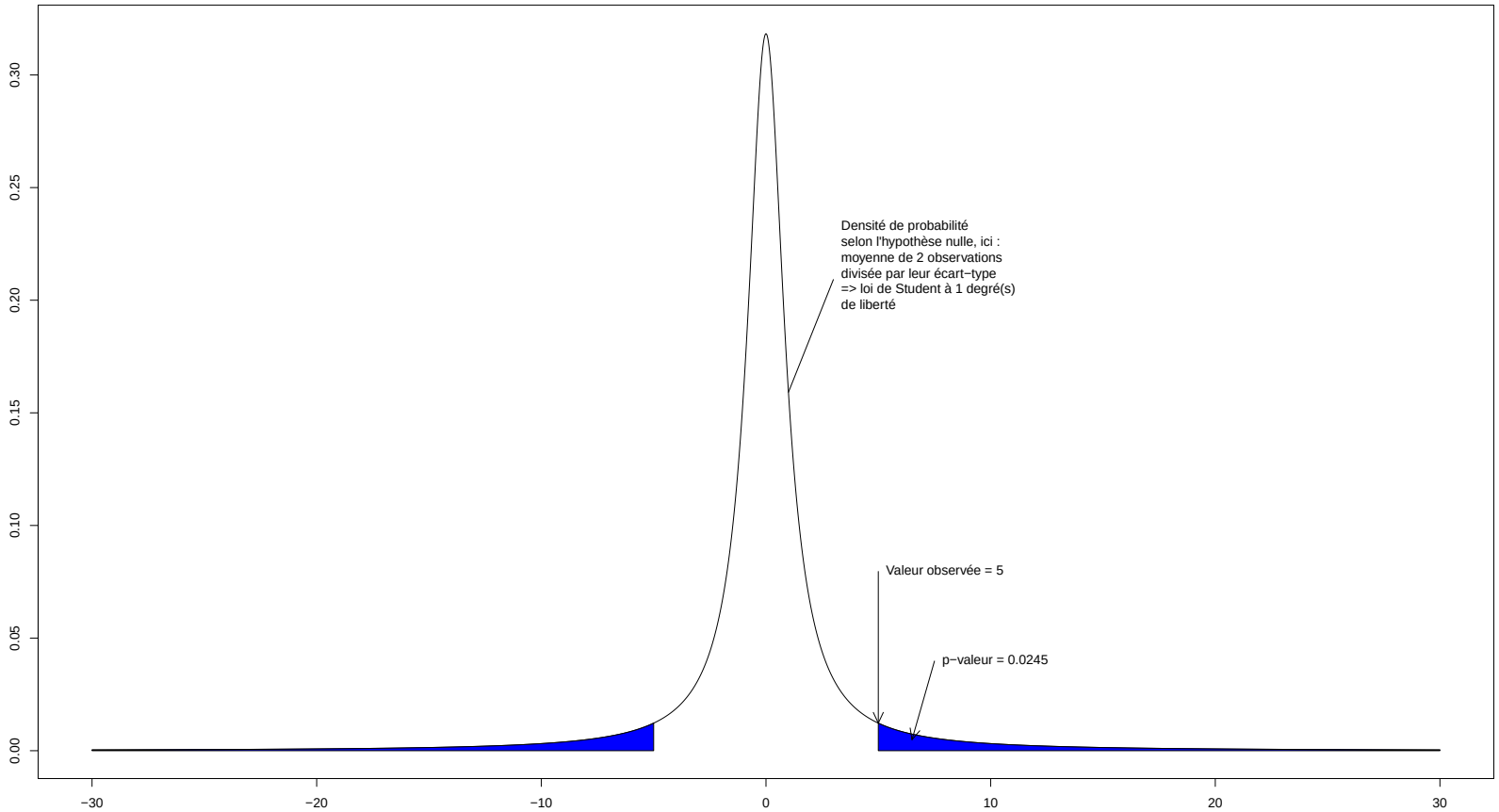
$$y_{obs} \sim N(\widehat{y}, \sigma)$$

Le paramètre σ correspondant à l'écart-type des observations autour de la valeur prédite.

La p-valeur (p-value en anglais)

Définition

Probabilité d'observer, par chance quelque chose d'aussi ou plus différent de 0 que ce qui est observé (voir aussi science4all "La plus grosse confusion des sciences, la p-value, Bayes 9").



APPLICATIONS DU GLM

- exemple dans R: effetSelection.R
- p-value
- AIC
- taille d'effets

$\Rightarrow$ Attention, la stratégie "mini-monde" (on met toutes les variables et on vera bien) risque de faire choisir des choses qui n'ont aucun sens, surtout si on se concentre sur les p-values

Subsubsection 2

Le LASSO

LE LASSO

suite du fichier effetSelection.R

- on abandone la p-value
- on ne laisse s'exprimer que les variables très fortes: comme de la colle sur le zéro (lambda)
- validation croisée, choix du lambda 1se

Subsubsection 3

Fonction de lien

GLM : ATTENTION À LA FORME DE LA VARIABLE EXPLIQUÉE

LA FONCTION DE LIEN

Le paramètre “family” dans glm() de R :

- Si mesure continue, éventuellement négative : distribution normale ok (choix par défaut)
  - $y \sim N(\alpha + \beta x, \sigma)$
  - glm(y$\sim$ x, family=**“normale”**)
- Si mesure continue mais strictement positive : glm sur **log**
  - $log(y) \sim N(\alpha + \beta x, \sigma)$
  - **logy <- log(y)**; glm(logy$\sim$ x, family=“normale”)
- Si comptage, théoriquement loi de Poisson mais problématique $\Rightarrow$ soit **binomiale négative** soit quasi-poisson dans R.
  - **glm.nb**(y $\sim$ x)
  - Si plutôt “proportionalité” : glm.nb(y $\sim$ log(x))
- Si binaire (présence/absence): loi de distribution **binomial**

Subsection 2

Statistiques dans l'espace et auto-corrélation

Auto-corrélation : le problème

- Le problème: une source commune dont on se fiche est corrélée avec deux paramètres
- ex: position par rapport au flanc de coteaux, terre et exposition au vent
- si plein de points sur pleins de coteaux différents, il y aura des terres différentes et des expositions au vent différentes donc pas un problème
- Par contre, si plusieurs points sur le même coteaux, problème d'auto-corrélation spatiale : situation difficile à gérer statistiquement, à éviter dans le plan d'expérimentation.

Subsection 3

Cas d'étude statistique: Impact de la distance aux bois sur les méligèthes et sur leurs parasitoïdes

INTRODUCTION

Comment évolue l'abondance de méligèthe en fonction de la distance aux bois ?

- Exploration d'un système précis avec une forte connaissance à priori
- puis de voir où et comment taper
- description du dispositif

La simplification ici passe par un design de l'expérience pour simplifier les interactions

Subsection 4

Cas d'étude statistique: approche paysage multi-bio-agresseurs

Subsubsection 1

Introduction

Introduction

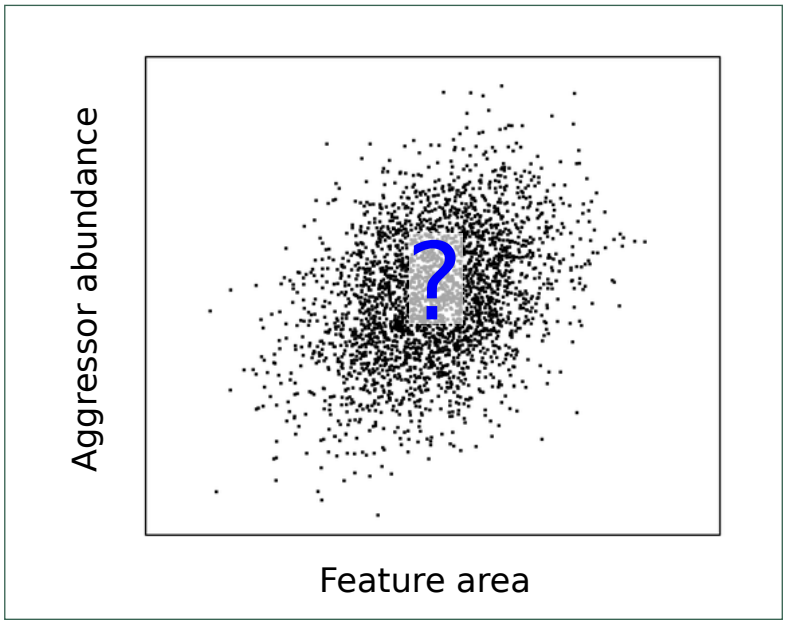
- Limiter l'utilisation de phyto
  - → Leviers paysagers ?
- Beaucoup d'études bio-agresseur par bio-agresseur
  - → Multi-bio-agresseurs ?
- Parapluie phyto-sanitaire
  - → C'est la réalité d'aujourd'hui, peut-on voir quelque chose malgré tout ?

⇒ Quantification des (dis)services des éléments paysagers

Subsubsection 2

Méthodes

Comment



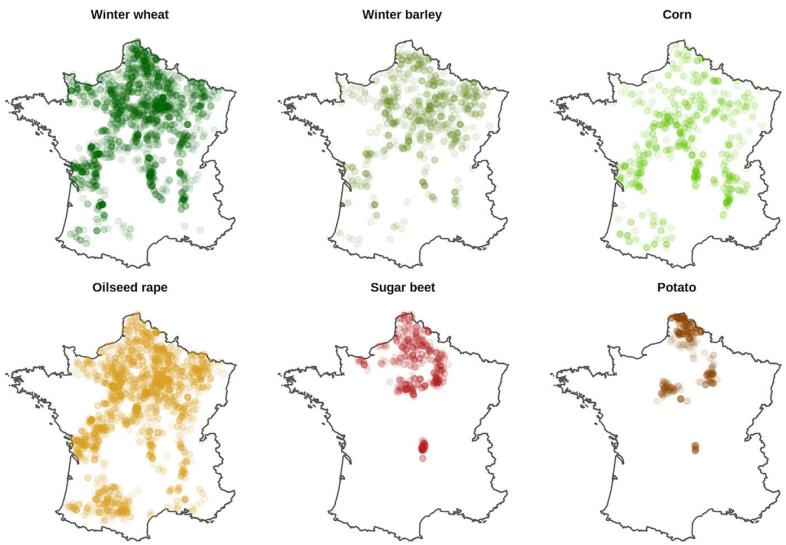
Object

- France continentale
- Corrélations entre éléments paysagers and abondances des agresseurs
- quatre échelles: 200m, 1000m, 5km, 10km

Approches

- Statistiques sur des données nationales
- Formalisation de la connaissance existente (élicitation)
  - Experts
  - Bibliographie

Données nationales



Points de relevés (2009-2017)

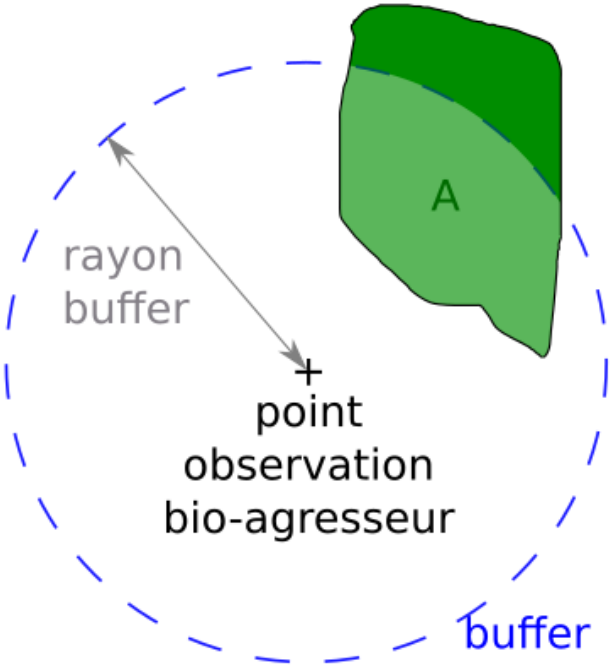
Paysage

- RPG (PAC) → grandes cultures & prairies
- couche végétation BDTOPO (IGN) → bois, vergers, haies, landes

Bio-agresseurs: Vigiculture®

- engagement annuel, surveillance hebdomadaire
- vigilance sur tous les bio-agresseurs → BSV
- souvent plusieurs mesures par bio-agresseur
- géolocalisés à la parcelle

MÉTRIQUES ET MODÈLES



MÉTRIQUES PAR POINT ANNÉE

- abondance bio-agresseurs : $n_{obs} > seuil$
- métriques paysagères : aires dans 4 buffer

MODÈLES DE POISSON OU BINOMIAL NÉGATIF

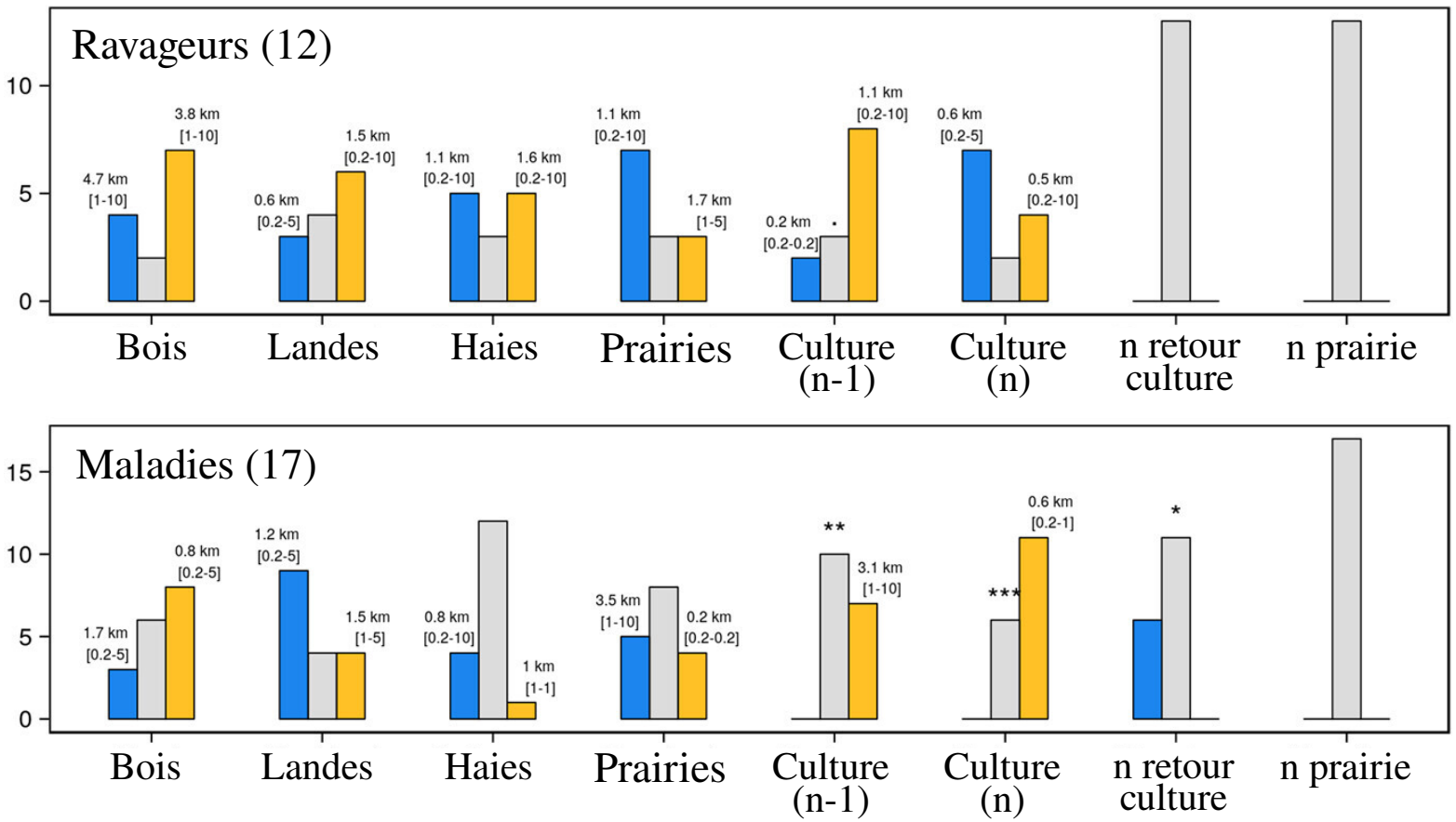
$$\log(Y) = \alpha \cdot \text{région agro-climatique} + \sum_{e} \beta_e \cdot \log(Area_e)$$

- Univarié + régions
- Lasso ($\lambda = \lambda_{1se}$)

Subsubsection 3

Résultats

STATISTIQUES PAYSAGÈRES



N bioagresseurs ayant une corrélation négative (bleu), absente (gris) ou positive (orange) avec chaque facteur paysager ou temporel

- statistiques: il y a de la puissance prédictive, accord pas fantastique mais incertitude expert énorme
- l'effet du paysage varie fortement entre bio-agresseurs
- En tout cas des cas avec service d'autre avec disservice

Subsubsection 4

Discussion

Conclusions principales

Ecosystem (dis)service des éléments paysagers

- Service/disservice dépend du bio-agresseurs

$\Rightarrow$ supposer par défaut un servide de régulation écosystèmique paraît déraisonnable

Subsection 5

APPLICATION

THIES ET TSCHARNTKE 1999, SCIENCE

Landscape structure and biological control in agroecosystems

RÉSULTATS

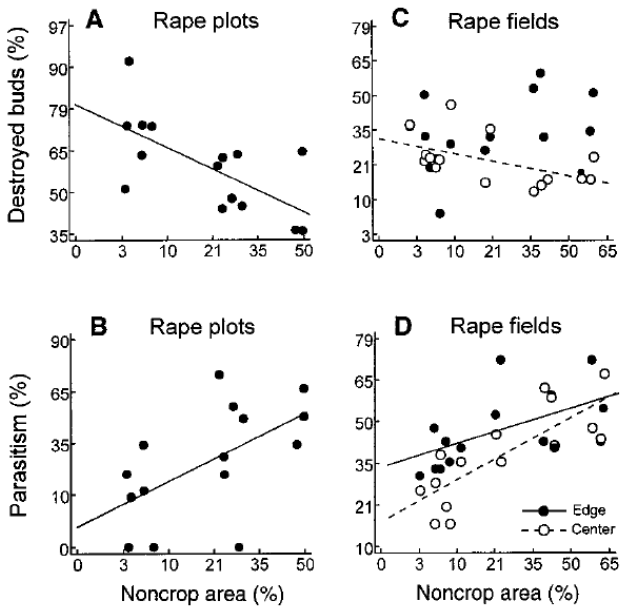


Fig. 2. Plant damage caused by rape pollen beetle (*M. aeneus*) and parasitism of rape pollen beetle by *T. heterocerus*, *P. interstitialis*, and *P. morionellus* on oilseed rape (*B. napus*) in relation to the percentage of noncrop area in agricultural landscapes {arcsine[(–transformed percentages)$^{1/2}$]}. (**A**) Percent destroyed buds in experimental summer rape plots in relation to the percentage of noncrop area. $Y = 70.8–0.51X$, $F = 10.06$, $P = 0.007$, $R^2 = 0.44$, $N = 15$. (**B**) Percent parasitism in experimental summer rape plots ($N = 15$) in relation to the percentage of noncrop area. $Y = 7.75 + 0.87X$; $F = 6.51$, $P = 0.02$, $R^2 = 0.33$, $N = 15$. (**C**) Percent destroyed buds in winter rape fields in relation to the percentage of noncrop area. Mean percentages per landscape ($N = 15$) from 26 winter rape fields are given. Edge of the fields: P = not significant; center of the fields: $Y = 37.2–0.21X$; $F = 4.85$, $P = 0.046$, $R^2 = 0.27$, $N = 15$. (**D**) Percentage parasitism in winter rape fields in relation to the percentage of noncrop area. Mean percentages per landscape ($N = 15$) from 26 winter rape crop fields are given. Edge of the fields: $Y = 39.4 + 0.28X$; $F = 5.86$, $P = 0.03$, $R^2 = 0.31$, $N = 15$; center of the fields: $Y = 27.1 + 0.5X$; $F = 21.1$, $P = 0.0005$, $R^2 = 0.62$, $N = 15$. The intercepts of these regression lines are significantly different ($F = 4.47$, $P = 0.04$), whereas the slopes do not differ ($F = 1.89$, $P = 0.18$).

CONCLUSION:

"These results provide evidence that complex landscapes with a high density and connectivity of uncultivated, perennial habitats may enhance populations of natural enemies, which immigrate into neighboring annual crop fields, attack pest insects, and contribute significantly to the reduction of pest populations below an economic threshold." ?!?!

Tscharntke 2016 et al.

[. . .] natural habitat provides important ecosystem services including pest control (. . .) Vraiment?

- Landis et al. 2000: Habitat management to conserve natural enemies of arthropod pests in agriculture
  - dizaines d'articles $\nearrow$ ennemis naturels.
  - $\sim$ rien sur succès de contrôle
  - quelques exemples de favorisation des ravageurs. . .
- Bianchi et al. 2006: Sustainable pest regulation in agricultural landscapes: a review on landscape composition, biodiversity and natural pest control
  - 24 articles sur ennemis naturels. ($\nearrow$ 18/ = 5/ $\searrow$1)
  - 10 articles sur contrôle ($\searrow$4.5/=4/$\nearrow$1.5)
- Karp et al. 2013: Forest bolsters bird abundance, pest control and coffee yield
  - oiseaux $\searrow$ ravageurs (exclusion)
  - oiseaux mangent ravageurs et forêts $\nearrow$ ravageurs
  - forêts $\searrow$ ravageurs : significatif à 125m (quelques buissons)
- Shackelford et al. 2013: Comparison of pollinators and natural enemies: A meta- analysis of landscape and local effects on abundance and richness in crops.
  - Rien sur les paysage/bioagresseurs, tout sur ennemis naturels
  - Effet espaces semi-naturels sur ennemis naturels $<$ effet sur pollinisateurs
- Milligan et al., 2016: Quantifying pest control services by birds and ants in Kenyan coffee farms
  - Predation sur cartes décroit avec distance à la forêts (oiseaux) (12 points AIC)
  - prédit: décroit de 16% à 5% de prédation sur 100 m

Tscharntke et al. (21) 2016

When natural habitat fails to enhance biological pest control – Five hypotheses

Abstract

Ecologists and farmers often have contrasting perceptions about the value of natural habitat in agricultural production landscapes, which so far has been little acknowledged in ecology and conservation. Ecologists and conservationists often appreciate the contribution of natural habitat to biodiversity and potential ecosystem services such as biological pest control, whereas many farmers see habitat remnants as a waste of cropland or source of pests. While natural habitat has been shown to increase pest control in many systems, we here identify five hypotheses for when and why natural habitat can fail to support biological pest control, and illustrate each with case studies from the literature: (1) pest populations have no effective natural enemies in the region, (2) natural habitat is a greater source of pests than natural enemies, (3) crops provide more resources for natural enemies than does natural habitat, (4) natural habitat is insufficient in amount, proximity, composition, or configuration to provide large enough enemy populations needed for pest control, and (5) agricultural practices counteract enemy establishment and biocontrol provided by natural habitat. In conclusion, we show that the relative importance of natural habitat for biocontrol can vary dramatically depending on type of crop, pest, predator, land management, and landscape structure. This variation needs to be considered when designing measures aimed at enhancing biocontrol services through restoring or maintaining natural habitat.

- (6) et si les ennemis naturels étaient tout simplement proportionels aux ravageurs ?
- (7) et si le service écosystémique était l'exception plutôt que la règle ?

Attention à l'objectivité de ceux qui écrivent…

- envoi des codes?