La diversité de la sole de blé et ses conséquences sur la structure des populations de rouille brune

Mathilde Pichot Utrera

PICHOT UTRERA Mathilde

RAPPORT DU STAGE D'APPLICATION EN STATISTIQUE DE 2E ANNEE

STRUCTURE D'ACCUEIL : Institut National de la Recherche Agronomique (INRA) - Unité de Biostatistique et Processus Spatiaux (BioSP)

THEME DU STAGE : La diversité de la sole de blé et ses conséquences sur la structure des populations de rouille brune

LIEU DE STAGE : INRA Centre de recherche PACA
Domaine Saint-Paul – Site Agroparc

VILLE : AVIGNON PAYS : FRANCE

Promotion : 2017

Maître de stage & tuteur pédagogique : Julien PAPAÏX

Remerciements

Je tiens tout d'abord à remercier mon tuteur et maître de stage Julien Papaïx pour sa patience, sa gentillesse, sa disponibilité, et pour m'avoir fait découvrir son univers de travail et sa vision positive des choses. Un grand merci également à Emily Walker pour son aide précieuse et sa bonne humeur. Plus généralement, je tiens aussi à remercier pour leur accueil tous les membres de l'équipe BioSP, où j'ai pris beaucoup de plaisir à réaliser mon stage cet été. J'espère avoir le plaisir de travailler au sein d'un environnement aussi agréable plus tard. Merci enfin à William qui m'a accompagnée pendant un bon moment, à Catherine pour sa relecture, et Hervé pour m'avoir supportée durant la rédaction de ce rapport.

Table des matières

Inventaire des jeux de données (JDD) utilisés dans le cadre du stage

— JDD1. Source : ARVALIS – Institut du végétal

N° obs	Variété (V = 358)	Année (1999-2008)	N° d'essai (E = 200)	N° de départem^{t}	Note de rouille (de 0 à 10)	Organes mesurés	Date de notation
obs_i	ACCOR	2001	2	30	0	F1	20/06/2001

Table 1 – JDD1 - Notes de rouille

— JDD2. Source : ARVALIS – Institut du végétal

N° obs	Année (2008-2011)	Coordonnées (Lambert-93) de la station (710 réparties en France)	Note du risque de rouille (300<x<1800) (données centrées, réduites pour les calculs)
obs_i	2009	(553891,6888805)	478

Table 2 – JDD2 - Risque de rouille lié à la météo

— JDD3. Source : FranceAgriMer

N° obs	Année (1999-2011)	N° de départem^{t} (Fr. métropolitaine)	Variété de blé tendre (V>1000)	Proportion de la variété, cultivée dans le départem^{t}, cette année là
obsi	1999	72	SOISSONS	0.08

Table 3 – JDD3 - Proportion des variétés de blé tendre cultivées en France

— JDD4. Source : `stats.gouv.fr`, mesures : Agreste (2010), recensement agricole

N° obs	N° de départem^{t} (France)	Superficie de blé tendre cultivée dans le département (ha)
obsi	21	27608

TABLE 4 – JDD4 - Superficie de blé tendre cultivée par département

— JDD5. Carte des départements IGN, type SHP, sous forme d'objets de type polygones dans R
Source : `https://www.data.gouv.fr/fr/datasets/geofla-departements-30383060/`

Présentation de la structure d'accueil

L'Institut National de la Recherche Agronomique (INRA) est un Établissement Public à Caractère Scientifique et Technologique (EPST) placé sous la tutelle des Ministères chargés de l'Agriculture et de la Recherche. Avec un budget annuel d'environ 800 millions d'euros, il est le premier institut de recherche agronomique en Europe, et deuxième en sciences agricoles au niveau mondial. Parmi ses principales missions, on trouve l'amélioration de l'agriculture en termes de performance économique, sociale aussi bien qu'environnementale ; le développement de systèmes alimentaires sains et durables, la valorisation de la biomasse, ou encore l'atténuation et l'adaptation au changement climatique. L'INRA, qui compte plus de 8000 salariés répartis sur dix-sept centres de recherche en France, s'organise en treize départements de recherche, et est dirigé depuis peu par Philippe Mauguin (Annexe A).

Le centre de recherche Provence-Alpes-Côte d'Azur (PACA) rassemble 1000 agents, dont 700 agents permanents, répartis dans 26 unités, localisés sur 10 sites en PACA. La présidence du centre INRA PACA est assurée par Michel Bariteau (Annexe A). Les thématiques traitées dans ce centre sont souvent en relation avec l'environnement méditerranéen dans lequel il s'inscrit.

Plus spécifiquement, l'unité de Biostatistique et Processus Spatiaux - BioSP, du département Mathématique et Informatique Appliquées - MIA, dirigée par Étienne Klein, rassemble une vingtaine de chercheurs et ingénieurs. Ils conduisent des recherches en statistique et en modélisation spatiales et spatio-temporelles notamment applicables aux domaines de l'environnement, de l'écologie et de l'épidémiologie. Trois missions phares de l'unité sont : mener des recherches disciplinaires en mathématiques et en statistiques, étudier des phénomènes agronomiques, écologiques et biologiques spatialisés à l'occasion de recherches pluridisciplinaires, et promouvoir et faciliter l'usage des outils mathématiques et informatiques à l'INRA.

Le projet sur lequel j'ai travaillé est monté en partenariat avec l'institut technique ARVALIS - Institut du végétal qui est un organisme français de recherche appliquée en agriculture dont l'objectif majeur est d'aider les producteurs agricoles, leurs organisations et les entreprises des filières à résoudre tous les problèmes techniques, technico-économiques, sociétaux et environnementaux qui se posent à eux. La Recherche/Développement est le premier champ d'activités d'ARVALIS - Institut du végétal.

Introduction

Sur la période 1960 – 1990, les pays les plus développés connaissent la « révolution verte » qui transforme le paysage agricole. La formidable hausse de rendement qui en découle change le paysage mondial, notamment en permettant une croissance démographique phénoménale : + 4 milliards d'individus depuis les années 60. Les systèmes agricoles sont fondamentalement réformés et la production agricole mondiale a plus que doublé sur cette période. L'intensification de l'agriculture se fait via de nouvelles méthodes telles que l'utilisation d'engrais minéraux, de produits phytosanitaires, la mécanisation, l'irrigation, mais aussi la sélection variétale. Ces nouvelles conditions de production ont rendu les agrosystèmes hautement dépendants d'apports extérieurs. De plus, la simplification des paysages accompagnant cette révolution verte est à l'origine d'une réduction de la biodiversité dans les agro-écosystèmes, pourtant essentielle aussi bien à la productivité et la fertilité des sols qu'à la protection des plantes contre les maladies et les ravageurs. En effet, la biodiversité favorise le développement de systèmes naturels de régulation interne responsables de l'enrichissement, du renouvellement et de la protection des sols, notamment via le cycle des éléments nutritifs. Ces systèmes naturels de régulation peuvent aussi jouer un rôle dans la régulation des maladies et ravageurs des grandes cultures à travers la présence d'organismes auxiliaires, antagonistes aux nuisibles des cultures. De nombreuses études épidémiologiques menées sur les relations hôte(s) – pathogène illustrent comment la diversité d'une population hôte peut changer la prévalence d'une maladie. Dans cet article : [Pautasso 2005], les auteurs distinguent quatre cas établis à partir d'écosystèmes forestiers : (i) monocultures résistantes, (ii) cultures diversifiées et sensibles, (iii) monocultures sensibles et (iv) cultures diversifiées et résistantes. Les populations d'hôtes homogènes et peu ou pas malades (cas i) ne sont pas stables sur le long terme et résultent de l'absence de pathogène. Le cas (ii) résulte généralement de l'introduction d'un nouveau pathogène auquel la population hôte n'avait jamais été confronté. Dans ce cas là, la diversité spécifique ne reflète pas une diversité fonctionnelle en terme de résistance vis à vis du pathogène considéré. Les écosystèmes agricoles industrialisés sont typiquement des cas de monocultures sensibles (iii). Cette sensibilité provient de la facilité avec laquelle le pathogène peut se diffuser au sein d'une parcelle, se disperser entre parcelles et s'adapter à l'hôte. Il faudrait donc privilégier les cultures plus diversifiées (iv) qui sont régies par l' « hypothèse d'assurance ». Cette hypothèse vient de la constatation que la plupart des plantes sont sensibles à plus d'un

pathogène, mais les plantes ne sont pas toutes sensibles à tous les pathogènes. Dans un milieu diversifié, si une espèce ou variété est infectée, d'autres non sensibles au pathogène peuvent compenser et assurer la persistance de la population. En plus de ces effets purement démographiques, la diversité des plantes cultivées et plus largement des communautés végétales des espaces semi-naturels au sein des paysages agricoles joue un rôle décisif dans les dynamiques évolutives des populations pathogènes.

Bien que la conceptualisation des effets de la diversité sur l'écologie et l'évolution des populations pathogène soit bien développée, peu d'exemples l'illustrant sont reportés dans la littérature. La plupart de ceux-ci porte sur les mélanges variétaux, c'est à dire, sur la diversification au sein même d'une parcelle, et montre généralement une diminution de la maladie par rapport à des parcelles en monoculture. Une autre possibilité consiste à travailler à l'échelle des paysages et de construire des mosaïques de cultures ou de variétés permettant de contrôler le risque épidémiologique. A l'heure actuelle nous n'avons cependant pas d'exemples de ces effets en conditions de production et sur un large territoire. L'objectif de ce stage est donc de rechercher des traces de l'effet de la diversité sur le développement d'une maladie. Nous nous sommes intéressés au pathosystème blé tendre - rouille brune du blé. Le blé tendre (ou froment) représente environ 29% de la production mondiale de céréales et constitue la base de l'alimentation occidentale. En majorité destiné à la production de farine panifiable pour l'alimentation humaine (58%), et animale (38%), on l'utilise aussi pour son amidon dans les industries cosmétique, pharmaceutique et pour la production de papier ou encore bioéthanol. La France, qui lui consacre cinq millions d'hectares de surfaces agricoles (2014), est le premier producteur et exportateur de blé tendre au niveau européen et plus de la moitié de la production française est exportée, notamment vers l'Afrique du Nord et l'Afrique Centrale. La rouille brune du blé, est quant à elle, à l'origine de pertes de qualité et de rendement pouvant atteindre 60 q/ha avec une moyenne de 21,4 q/ha dans une parcelle infectée (source essais Syngenta et externes 2011-2014). Elle se manifeste sous la forme de pustules de couleur orange à brune disséminées de façon aléatoire sur les feuilles de la plante, voire jusqu'à l'épi en cas d'attaque sévère. Principalement de la mi-été à l'automne, la rouille brune profite de températures comprises entre 15°C et 22°C, et d'un fort taux d'humidité pour se développer.

Le projet dans lequel s'inscrit mon stage s'insère dans le cadre plus large d'une collaboration entre l'INRA et Arvalis – Institut du végétal qui vise à prendre en compte la diversité de l'assolement dans le conseil du choix variétal. La première approche du projet, qui constitue le matériel de ce stage, s'attache à déterminer si la diversité variétale limite (ou pas) le développement de la rouille à l'échelle de la France. Pour cela, après avoir cerné le contexte et les limites de notre étude, en grande partie dictées par les données, nous préciserons la notion de diversité variétale par laquelle nous voulons expliquer la rouille, pour enfin développer un/deux modèle(s) qui nous permettra(ont) d'observer dans quelle mesure notre indice de diversité du blé permet d'expliquer les attaques de rouille.

Chapitre 1

La rouille, de multiples paramètres d'influence

Nous verrons ici comment se présentent les données de rouille dont nous disposons et on s'attachera à cibler les variables explicatives pertinentes pour notre modèle. On finira par le cas de la diversité variétale, en définissant plus précisément ce que nous entendons par là.

1.1 La rouille brune du blé, un pathogène à l'échelle du territoire français

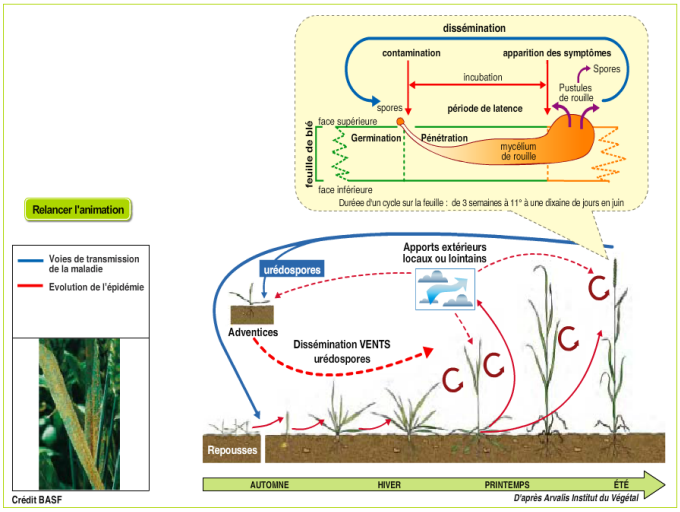


FIGURE 1.1 – Cycle de développement de la rouille

Des attaques de rouille brune ont été mesurées sur 358 variétés à travers près de 7000 notes dans [0,10] entre 1999 et 2008, sur 41 départements qui couvrent environ la moitié du territoire français métropolitain (plutôt dans le Nord et l'Ouest du pays) (JDD1).

Deux notions sont à dissocier pour bien caractériser l'importance de la rouille dans un secteur : l'incidence et la sévérité. L'incidence représente la probabilité d'observer la rouille : $\mathbb{P}(note > 0)$. La sévérité, elle, présuppose que la rouille a pu se développer, et reflète l'intensité de l'attaque.

On peut voir sur l'histogramme des notes (fig. 1.2) qu'elles suivent une distribution unimodale autour de 0, l'incidence sur nos données est de l'ordre de 63%.

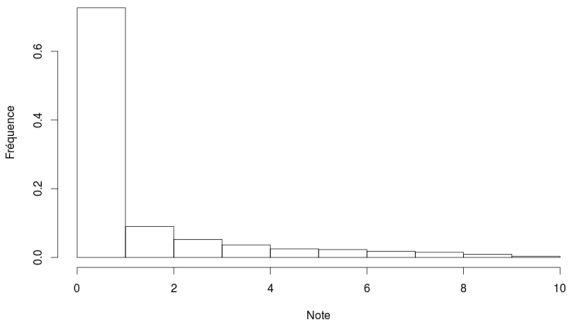
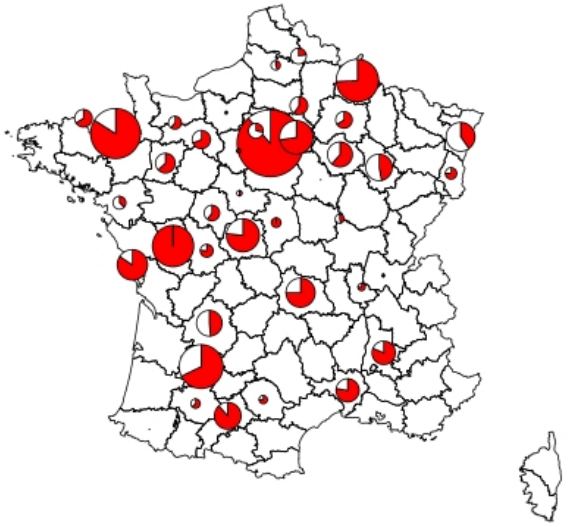


FIGURE 1.2 – Histogramme des notes de rouille



Note de lecture. Rayon des cercles proportionnel à la note moyenne de rouille dans le département. En rouge : incidence.

FIGURE 1.3 – Incidence et abondance moyennes de rouille sur 1999-2008 en France

De plus, on note que la rouille semble toucher l'ensemble du territoire (fig. 1.3), avec toutefois des moyennes plus élevées dans les départements du centre et de l'Ouest de la France, qui sont les régions où le blé est le plus cultivé.

1.2 Rouille et paramètres environnementaux

Les mesures de rouille ont été faites dans 200 localisations (essais), à des dates et par des personnes différentes. Les notes proviennent d'une échelle de notation discrète entre 0 et 10, mais pour certains essais, des moyennes préalables ont été réalisées, ce qui résulte

en des notes continues. De plus, selon l'essai, les organes mesurés sur la plante ne sont pas les mêmes. Pour ne pas négliger la variabilité liée à l'ensemble de ces différences dans le protocole d'observation, on considérera un effet aléatoire dû et propre à l'essai.

Les variétés de blé n'ont pas toutes les mêmes propriétés de résistance à la rouille, ni aux mêmes souches de rouille. Ainsi, la variété du blé sera comptée parmi les variables explicatives. L'estimation de l'effet variété sera aussi intéressant pour établir une classification des variétés selon leur sensibilité à la rouille.

La pression de rouille peut varier d'une année sur l'autre (par exemple à cause du climat), c'est pourquoi nous gardons la variable année dans la suite de l'analyse. De plus, la carte des climats est assez diverse en France. Il nous a donc semblé pertinent d'inclure une variable qui rendrait compte de cette variance climatique au sein d'une même année. Arvalis - Institut du végétal publie annuellement des cartes du risque de rouille brune qui se basent sur des relevés météorologiques qu'ils effectuent. Nous avons pu récupérer ces cartes pour les années 2008 à 2011 (JDD2). Les notes de rouille couvrant les années 1999 à 2008, nous avons décidé de considérer une note moyenne dans chaque département, toutes années confondues (voir Annexe B). Cette variable tiendra compte d'un effet météorologique uniquement lié à la géographie.

Des JDD1 et 2, nous gardons donc, pour expliquer la rouille, un facteur intrinsèque à la plante : sa variété, mais aussi des facteurs externes liés à l'environnement, à la géographique et la chronologie, à savoir le risque météorologique du département, l'année, ou encore un effet aléatoire lié à l'essai, auxquels s'ajoute notre indicateur de diversité variétale. Précisons donc ce concept de diversité variétale ainsi que l'indice que nous allons utiliser pour en rendre compte.

1.3 Rouille et diversité du blé (facteur intrinsèque)

La diversité peut se voir à plusieurs échelles : au niveau de l'espèce, de la variété, au niveau des gènes. La résistance d'une variété face à la rouille est de deux types : qualitative et quantitative. La première représente la capacité de la plante à éviter la contamination (incidence) et provient de gènes majeurs de résistance. Leur qualité, quantité, et selon quelles combinaisons ils sont présents dans le génome de la plante (pyramidage) déterminent sa sensibilité aux différentes souches de rouille. L'incidence étant facilement observable dans les cultures, la sélection variétale en agriculture se base principalement sur la résistance qualitative. D'autant plus que la résistance quantitatives des plantes est plus complexe à cerner car elle résulte de la combinaison de Loci de Caractères Quantitatifs (LCQ, ou QTL pour quantitative trait loci) de résistance qui influencent les différents traits d'histoire de vie du pathogène, notamment l'efficacité d'infection, le temps de latence, le nombre de spores disséminées... On dispose de deux jeux de données concernant la diversité du blé

en France. Le premier JDD (3) renseigne sur la fréquence des variétés de blé dans chaque département. Le deuxième répertorie les gènes majeurs de résistance de certaines variétés de blé tendre. On se concentrera sur la diversité variétale pour cette étude (JDD3), car elle englobe les différents points de vue de la diversité et constitue donc une première étape nécessaire (voir la discussion, partie 2.2.4 pour plus d'information).

Plus de 1000 variétés de blé tendre ont été recensées sur 69 départements entre 1999 et 2010 (JDD3). On s'attache ici à cerner la diversité des variétés de blé tendre en France métropolitaine, à l'échelle du département. Quelles sont les caractéristiques de l'assolementen blé tendre (homogénéité, structure dans l'espace,...) ?

1.3.1 Quelle diversité voulons-nous mesurer ?

La diversité du blé tendre au sein d'un département peut, simplement, être caractérisée par le nombre de variétés qui s'y trouvent. On appelle cet indice de diversité la richesse. Toutefois, ce premier indice ne nous permet pas d'utiliser toute l'information dont nous disposons. En effet, il ne prend pas en compte les proportions respectives de chaque variété au sein du département : l'équitabilité (*evenness*). Un département où deux variétés sont réparties à 50% de la surface cultivée chacune témoignera de plus de diversité qu'un autre avec une répartition de 90% et 10%. Nous voulons donc un indice de diversité qui rende compte de l'équirépartition, ou non, des variétés dans le département. Pour cela, nous allons utiliser les nombres de Hill [Marcon 2015]. Ce sont des indices de diversité qui dépendent d'un paramètre, q. Soit S le nombre de variétés dans un département, p_s les fréquences de ces variétés. Le nombre de Hill pour ce département vaut :

$$ {}^qD = \left(\sum_{s=1}^{S} p_s^q \right)^{\frac{1}{1-q}} $$

On voit que ${}^0D = S$ représente la richesse du département (nombre de variétés). Le nombre de Hill donne ici la même importance à toutes les variétés présentes dans le département. De plus, ${}^2D$ vaut l'inverse de l'indice de concentration de Gini-Simpson :

$$ \lambda = \sum_{s=1}^{S} p_s^2 $$

qui est égal à la probabilité que deux « individus » tirés aléatoirement dans la population soient de la même variété. Le nombre de Hill est alors équivalent à un « nombre de variétés abondantes » [Hill 1973] ; 2 est donc la valeur de q pour laquelle on donne le moins d'importance aux espèces rares.

Dans la suite des travaux et pour faciliter les analyses, on conservera comme indicateurs de diversité les nombres de Hill, aux valeurs extrêmes détaillées plus haut, ou par pas de 0,2 entre celles-ci (sauf 1, valeur pour laquelle le nombre de Hill n'est pas défini).

1.3.2 Un gradient de diversité important à l'échelle de la France

En cartographiant les nombres de Hill selon les années et pour les valeurs extrêmes de q (voir fig. 1.4, 1.5, 1.6, 1.7 et en annexe C), on voit un net gradient (croissant) Sud/Nord en ce qui concerne la diversité « due » aux variétés rares : on trouve plus de variétés rares différentes dans le Nord de la France. La hausse de diversité portée par les variétés abondantes suit, elle, plutôt un gradient Sud-Est/Nord-Ouest : on observe plus de variétés « abondantes » au Nord-Ouest du territoire français métropolitain. Enfin, on note un certain effet du temps : la diversité augmente avec les années. Il y aurait donc une tendance à la diversification des variétés du blé tendre.

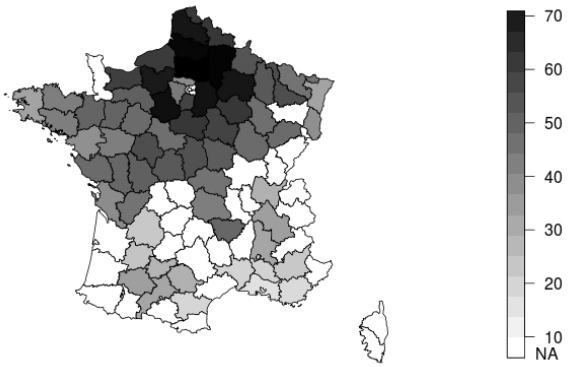


FIGURE 1.4 – Nombres de Hill par département en 1999 pour $q = 0$ ($\equiv$ richesse)

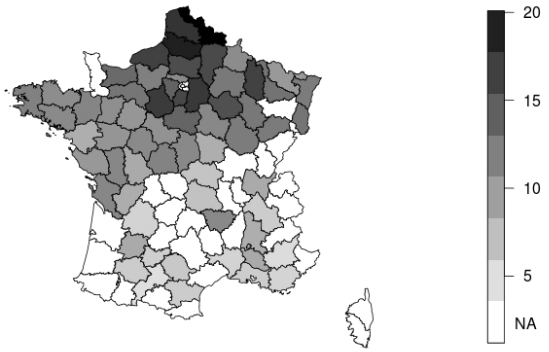


FIGURE 1.5 – Nombres de Hill par département en 1999 pour $q = 2$ ($\equiv$ nombre de variétés "abondantes")

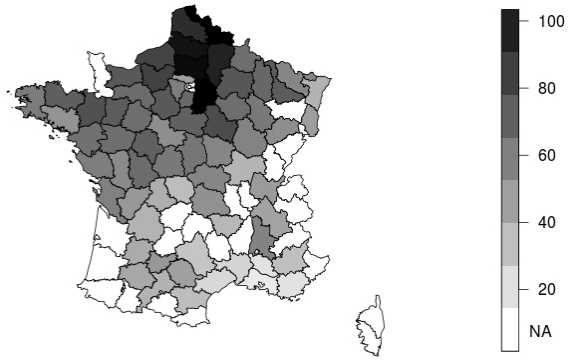


FIGURE 1.6 – Nombres de Hill par département en 2010 pour $q = 0$

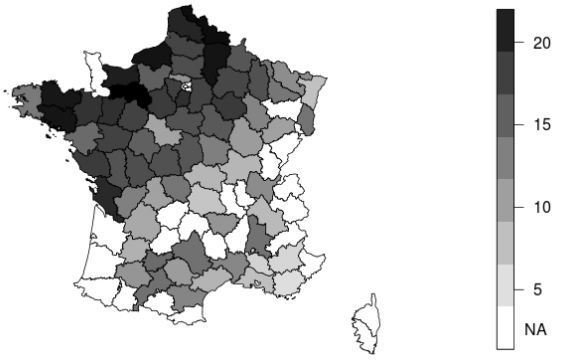


FIGURE 1.7 – Nombres de Hill par département en 2010 pour $q = 2$

De là, à titre indicatif et pour compléter notre étude de la diversité en lui donnant une dimension spatiale, on s'est demandé si l'on pouvait observer mathématiquement une

structure dans l'espace à notre indicateur de diversité variétale. Les résultats de ce travail sont présentés en annexe D.

On dispose de notes de rouille en France à l'échelle du département sur les années 1999 à 2008, ainsi que d'un indicateur de diversité et d'autres variables avec lesquelles nous voulons les expliquer (année, variété, essai : effet aléatoire, risque météo lié à la localisation : le département). Le chapitre 2 détaille maintenant un modèle pour expliquer les notes de rouille par ces variables que nous venons de cibler et définir.

Chapitre 2

Un modèle à effet de seuil

2.1 Données & notations

Variable(s) à expliquer :

Note de rouille : $y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}$, incidence : $inc = \begin{pmatrix} inc_1 \\ \vdots \\ inc_N \end{pmatrix}$, sévérité : $s = \begin{pmatrix} s_1 \\ \vdots \\ s_N \end{pmatrix}$

L'incidence et la sévérité sont directement tirées de la note de rouille et telles que : $\forall i \in \{1, ..., N\}\, inc_i = \mathbb{1}_{\{y_i > 0\}}$, et $s_i = \begin{cases} y_i & \text{si } y_i > 0 \\ NA & \text{sinon.} \end{cases}$

Variables explicatives :

On note $w = \left(\begin{array}{c|c|c|c|c} a & v & e & r & d \end{array} \right)$ la matrice contenant les variables explicatives, avec $a = \begin{pmatrix} a_1 \\ \vdots \\ a_N \end{pmatrix}$, $v = \begin{pmatrix} v_1 \\ \vdots \\ v_N \end{pmatrix}$, $e = \begin{pmatrix} e_1 \\ \vdots \\ e_N \end{pmatrix}$, $r = \begin{pmatrix} r_1 \\ \vdots \\ r_N \end{pmatrix}$ et $d = \begin{pmatrix} d_1 \\ \vdots \\ d_N \end{pmatrix}$, respectivement le vecteur des **années**, des **variétés**, des **essais**, du **risque** météo lié au département et celui des indices de **diversité** : les nombres de Hill (pour une valeur de q donnée), de chaque observation. De plus, $\mathscr{A} = \{1999, ..., 2008\}$, $\#\mathscr{A} = A = 10$; $\mathscr{V}$ l'ensemble des variétés, $\#\mathscr{V} = V = 358$ et $\mathscr{E} = \{1, ..., 200\}$ les essais, avec $E = 191$: le nombre d'essais que nous gardons.

Notes.
1. On utilisera $N = 6526$ observations sur les 6785 initiales car on enlève celles qui correspondent à un département pour lequel les données de fréquences des variétés ne sont pas disponibles l'année en question, nous empêchant de calculer l'indice de diversité. Ceci réduit donc les 200 essais à 191.

2. Dans la suite du document, les lois normales $\mathcal{N}(\mu, \tau)$ sont paramétrées par défaut par leur espérance μ et leur précision τ au lieu de leur variance (σ) par soucis de cohérence avec le programme utilisé ensuite : JAGS (voir partie 2.2.2), et on a $\tau = \frac{1}{\sigma^2}$.

2.2 Un modèle bayésien hiérarchique à effet de seuil

2.2.1 Inférence bayésienne, le modèle

On utilise une approche bayésienne dans notre analyse pour sa flexibilité par rapport à un modèle de régression fréquentiste, ce qui va nous permettre de modéliser conjointement l'incidence et la sévérité de la rouille. En effet, on voudrait mener une modélisation simultanée de l'incidence et de la sévérité de notre pathogène afin de voir si elle est pertinente.

Pour correspondre au caractère continu des notes, il nous faut donc trouver une distribution continue, positive ou nulle, qui gère bien les 0 induits par l'incidence et assez flexible pour permettre un bon ajustement sur les données. Sur ces critères, on choisit d'essayer un premier modèle avec une loi normale rectifiée. Notée $\mathcal{N}^{\mathcal{R}}(\mu, \tau)$, cette loi diffère de la gaussienne classique en ce qu'elle est contrainte pour ne jamais prendre de valeurs négatives : ces dernières sont toutes ramenées à 0 (fig. 2.1). Ainsi, 0 représente le seuil en deçà duquel on ne détecte pas de maladie.

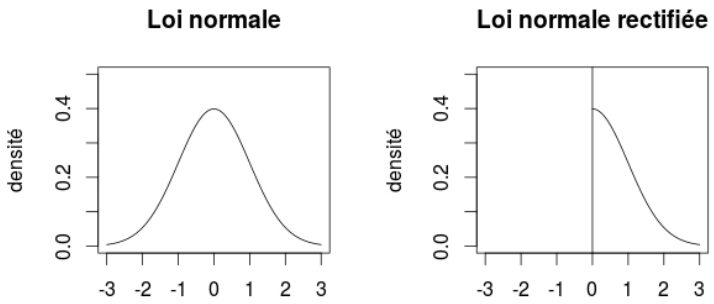


FIGURE 2.1 – Densité des lois $\mathcal{N}(0, 1)$ et $\mathcal{N}^{\mathcal{R}}(0, 1)$

Sa densité est la suivante :

$$\forall x \in \mathbb{R}^+, f_{\mu,\sigma}(x) = \Phi_{\mu,\sigma}(0)\delta(x) + \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \mathbb{1}_{\{x>0\}},$$

avec

$$\Phi_{\mu,\sigma}(0) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{0} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx, \tag{2.1}$$

fonction de répartition de la loi normale de paramètres μ et σ prise en 0, et $\delta(x)$ la distribution de Dirac.

On note Y la variable aléatoire dont nous supposons notre échantillon de notes $(y_1, ..., y_N)$ issu. On pose donc $Y \sim \mathcal{N}^{\mathcal{R}}(\mu, \tau)$.

On va ensuite faire intervenir nos variables explicatives dans l'espérance μ de Y par une régression. Pour chaque observation i, on a :

$$\mu_i = \mu^d * d_i + \mu^r * r_i + \sum_{j=1}^{A} \mu_j^a \mathbb{1}_{\{a_i=j\}} + \sum_{k=1}^{V} \mu_k^v \mathbb{1}_{\{v_i=k\}} + \sum_{l=1}^{E} \varepsilon_l \mathbb{1}_{\{e_i=l\}} \tag{2.2}$$

avec $\forall l \in \{1, ..., 200\}\, \varepsilon_l \sim \mathcal{N}(0, \tau_\varepsilon)$ où $\tau_\varepsilon = \frac{1}{\sigma_\varepsilon^2}$ car on veut que les paramètres ε_l traduisent un effet aléatoire propre à l'essai.

Le vecteur $\theta = (\mu^d, \mu^r, \mu_1^a, ..., \mu_A^a, \mu_1^v, ..., \mu_V^v, \varepsilon_1, ..., \varepsilon_E, \sigma_\varepsilon, \sigma)$ rassemble les paramètres du modèle, et est donc en partie défini (via la régression) comme une fonction des variables explicatives. De plus, quantifier les ε_l ne nous intéresse pas, on définit alors $\tilde{\theta} = (\mu^d, \mu^r, \mu_1^a, ..., \mu_A^a, \mu_1^v, ..., \mu_V^v, \sigma_\varepsilon, \sigma)$, vecteur des paramètres à estimer (moins les paramètres contraints, voir plus loin).

On peut alors écrire la vraisemblance du modèle :

$$\mathscr{L}(y|\theta) = f(y|\theta) = \prod_{i=1}^{N} f_i(y_i|\theta), \tag{2.3}$$

avec

$$f_i(y_i|\theta) = f_{\mu_i,\sigma}(y_i|\theta) = \Phi_{\mu_i,\sigma}(0)\delta(y_i) + \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(y_i-\mu_i)^2}{2\sigma^2}} \mathbb{1}_{\{y_i>0\}}. \tag{2.4}$$

On choisit comme lois a priori pour nos paramètres des lois peu informatives. Pour rendre le modèle identifiable, comme on a deux variables explicatives qualitatives et pas d'intercept, on doit annuler un des paramètres correspondant aux variables qualitatives, soit l'année ou la variété. On décide de fixer μ_1^a à 0 pour ne pas prendre de variété comme référence. Prendre une année en référence représente un plus petit parti pris. Les lois a priori de nos paramètres d'intérêt sont les suivantes :

$\sigma \sim \mathcal{U}[0, 10]$,
$\sigma_\varepsilon \sim \mathcal{U}[0, 50]$,
$\forall j \in \{2, ..., A\}, \mu_j^a \sim \mathcal{N}(0, 0.001)$,
$\forall k \in \mathscr{V}, \mu_k^v \sim \mathcal{N}(0, 0.001)$,
$\mu^d \sim \mathcal{N}(0, 0.001)$,
$\mu^r \sim \mathcal{N}(0, 0.001)$.

On note $\pi(\theta)$ la loi a priori du modèle et $\pi(y)$ la loi des données. D'après le théorème de Bayes, la loi cible du modèle sera :

$$\pi(\theta|y) = \frac{\pi(y, \theta)}{\pi(y)}$$

avec

$$\pi(y, \theta) = \mathscr{L}(y|\theta)\pi(\theta)$$

la loi a posteriori du modèle.

2.2.2 Méthode d'estimation : Échantillonneur de Gibbs (MCMC) avec JAGS

On utilise la méthode de Monte-Carlo par chaîne de Markov dont le principe est de construire une chaîne de Markov dont le noyau de transition converge vers la loi selon laquelle on veut simuler. Dans ce but, on se sert du programme JAGS (Just Another Gibbs Sampler) destiné à l'analyse de modèles bayésiens hiérarchiques, qui, comme son nom l'indique, implémente l'algorithme de Gibbs (échantillonneur). JAGS va donc nous permettre de simuler des données selon les distributions a priori de nos paramètres, ce qui nous permettra d'en tirer des estimateurs ou lois cibles des paramètres (pour plus de détails sur JAGS, voir annexe F).

Pour simplifier l'écriture du modèle, et comme la fonction de densité d'une $\mathcal{N}^{\mathcal{R}}(\mu, \tau)$ n'est pas précodée dans le programme, on décide de séparer les données d'incidence et de sévérité, sans pour autant les dissocier l'une de l'autre. Ainsi, on travaille sur les vecteurs de données *inc* et *s*. On note I et S les variables aléatoires qui modélisent respectivement l'incidence et la sévérité dans le modèle. On a alors $S \sim \mathcal{N}(\mu, \tau)$ et $I \sim \mathcal{B}ernoulli(p)$ avec $\forall i \in \{1, ..., N\}, p_i = \mathbb{P}(y_i > 0) = 1 - \mathbb{P}(y_i \leq 0) = 1 - \Phi_{\mu_i,\sigma}(0)$ (voir fig. 2.2). Le seuil à partir duquel on détecte de la maladie est fixé à 0.

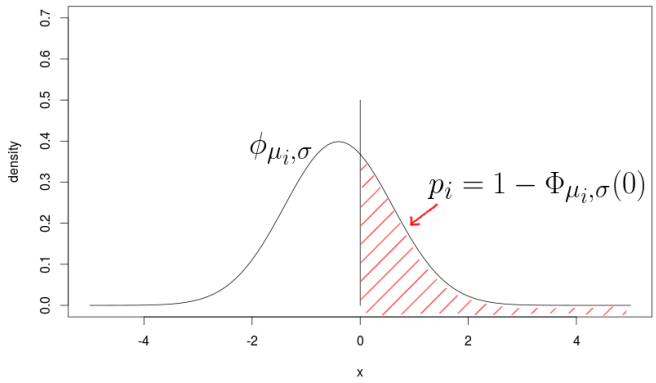


FIGURE 2.2 – Densité de la variable latente, lien avec l'incidence

Les lois a posteriori des paramètres μ_i et σ s'ajustent ainsi simultanément sur les données d'incidence et de sévérité, ce qui revient à la même modélisation que plus haut. On

ne retiendra que les valeurs positives ou nulles de la variable latente S pour les comparer avec nos données de rouille censurées.

Sont disponibles en annexe G un schéma de la structure du modèle hiérarchique à effet de seuil ainsi que le code pour JAGS.

On lance les simulations sur ce premier modèle sur 3 chaînes, 100 000 itérations (dont 50 000 de *burn in*), avec un *thinning interval* de 50 pour éviter les autocorrélations (voir annexe F, et diagnostic de convergence). On génère également des données répétées pour 10 valeurs de q comprises entre 0 et 2 inclus afin de mesurer la qualité de prédiction du modèle.

2.2.3 Diagnostic de convergence & qualité d'ajustement

On vérifie que les chaînes de Markov aient bien convergé pour tous les paramètres (autocorrélations). De plus, pour chaque paramètre, on vérifie que les trois chaînes ont convergé vers la même distribution a posteriori (voir graphiques en annexe G.3). Pour cela, on utilise le diagnostic de Gelman-Rubin qui se base sur le calcul d'un ratio des variances intra et inter-chaînes, de façon analogue à une analyse de variance (voir `https://cran.r-project.org/web/packages/coda/coda.pdf`). Sous R, le package 'coda' donne un "facteur de rétrécissement" (*shrink factor*), qui indique par combien les intervalles de crédibilité pourraient être divisés en continuant l'algorithme de convergence. Des valeurs très au dessus de 1 indiquent donc un manque de convergence. Les figures en annexe G.4 nous montrent le tracé de ce facteur au fil des itérations pour certains paramètres. Les chaînes pour tous les paramètres convergent assez bien.

Concernant les notes répétées, on voit (fig. 2.3) qu'on a tendance à surestimer les notes faibles et sous-estimer les plus grandes.

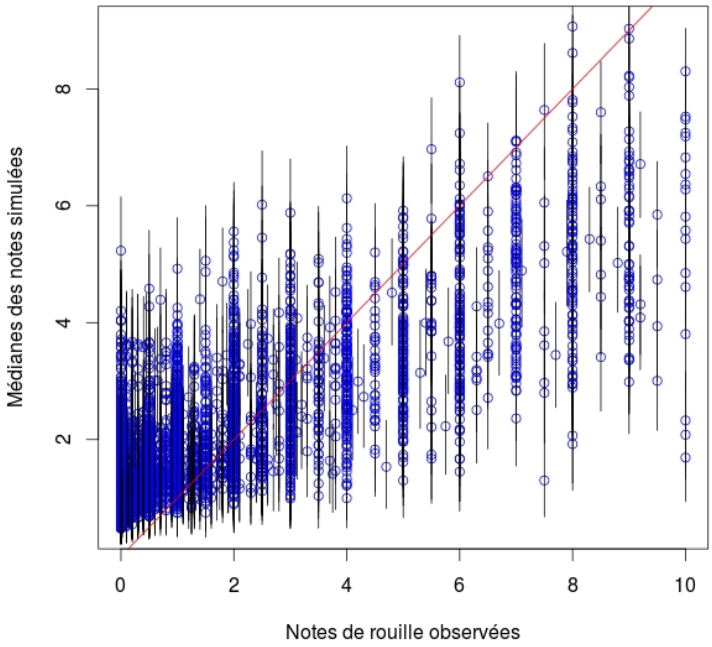


FIGURE 2.3 – Médianes (et écart interquartiles) des notes de rouille simulées en fonction des notes observées

De plus, l'incidence semble à première vue bien prédite, avec une incidence observée (sur les données utilisées pour l'ajustement du modèle seulement) de 0.6271836 (0.6299189 sur données complètes), contre une incidence prédite de 0.6196131. Cependant, plus en détails (tab. 2.1), on s'aperçoit que bien qu'on ait un peu plus de 57% de bonnes prédiction sur l'incidence, plus de la moitié des incidences prédites nulles sont observées à 1... Les prédictions sur l'incidence sont à améliorer.

Incidence prédite/observée	0	1
0	1040.3	1440.7
1	1352.7	2962.3

TABLE 2.1 – Moyenne des données d'incidence prédites x données d'incidence observées

Pour les données de sévérité, on voit (fig. 2.4 et 2.5) que l'ajustement reste à améliorer. La loi normale ne semble pas la plus appropriée pour s'ajuster aux données de sévérité.

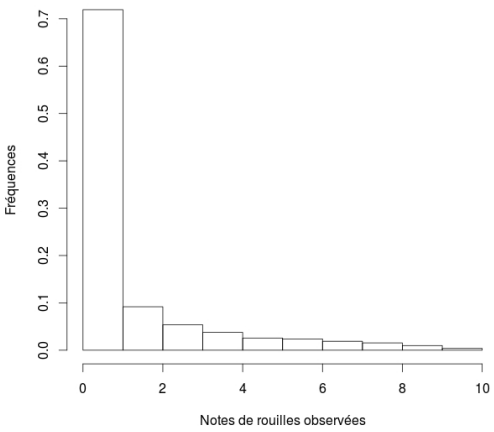


FIGURE 2.4 – Histogramme des notes de rouille observées

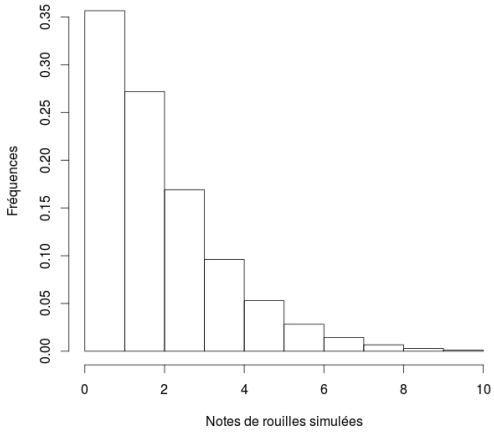


FIGURE 2.5 – Histogramme des notes de rouille simulées

2.2.4 Analyse des résultats

Pour chaque paramètre, si rien n'est précisé sur l'influence de la valeur q (degrés d'importance accordé aux variétés rares dans le calcul des nombres de Hill), c'est qu'aucun effet n'a été constaté. Dans ce cas, on utilise les données simulées avec $q = 2$.

Les variances de lois normales (du modèle : fig.2.6 et de l'effet essai : fig. 2.7) sont de l'ordre de 1,5. On voit que les densités sont bien dessinées, ce qui témoigne d'une bonne convergence.

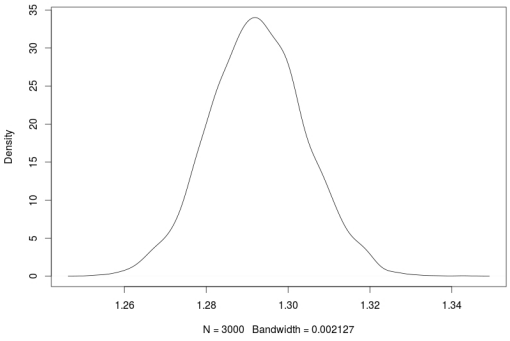


FIGURE 2.6 – Densité de la loi a posteriori de σ

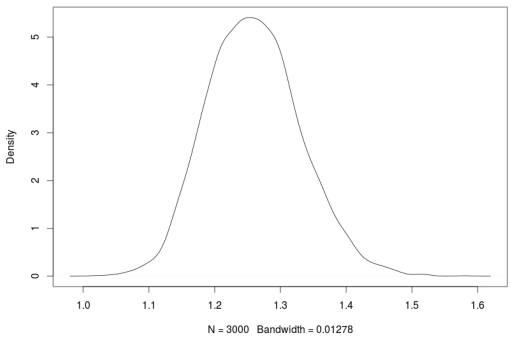


FIGURE 2.7 – Densité de la loi a posteriori de σ_ε

Risque : L'estimation du paramètre lié au risque est positive, on retrouve bien que le risque publié par ARVALIS-Institut du végétal est positivement corrélé avec les notes de rouille observées. Toutefois, le mode de μ^r (fig. 2.8) reste assez proche de 0. Cette covariable n'est pas très explicative. Pour améliorer cela on pourra relancer le modèle avec les notes de risque qui correspondent aux bonnes années lorsque qu'elles seront disponibles.

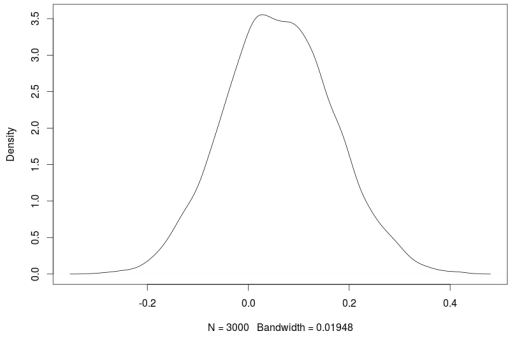


FIGURE 2.8 – Densité de la loi a posteriori de μ^r

Année : On voit sur la figure 2.9 que la valeur du coefficient lié à l'effet année augmente avec les années. On peut donc dire que les attaques de rouille deviennent de plus en plus fortes au fil des ans, avec un pic en 2007. Ce résultat est à mettre en relation avec le changement du climat, et spécialement en 2007, où on a connu des températures particulièrement favorables à l'apparition et à la prolifération de la rouille brune.

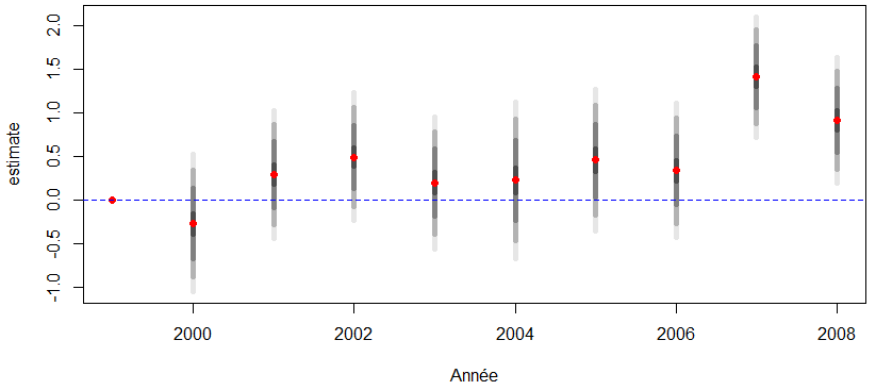


FIGURE 2.9 – Quantiles de la loi a posteriori des μ_j^a en fonction des années

Diversité : On voit d'après la figure 2.10 que moins on prend en compte les variétés rares (plus q augmente), plus le coefficient correspondant à la diversité (nombre de Hill) devient significativement non nul, et négatif. Ceci indique que la diversité variétale semble prévenir les attaques de rouille surtout lorsque que ce sont les variétés abondantes qui sont diversifiées. Pour limiter le développement de la rouille, il semblerait qu'il faille augmenter le nombre de variétés abondantes cultivées dans le département. Une diversité portée par les variétés rares n'aurait que peu voire pas d'impact sur la rouille.

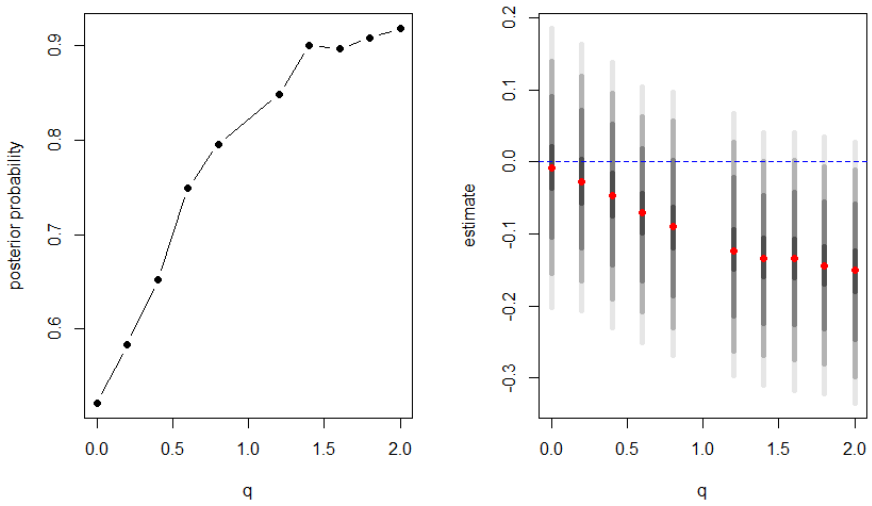


FIGURE 2.10 – Probabilité que la loi a posteriori de μ^d soit négative (à gauche), et ses quantiles (à droite) en fonction de q

Variété : Les paramètres liés aux variétés sont pour certains estimés avec les valeurs les plus significatives du modèle. On a des médianes de l'ordre de 6, donc très largement plus que pour les autres paramètres. La variété explique bien les différences d'attaques de rouille sur le blé tendre. On voit un gradient apparaître des variétés les moins malades à celles qui sont le plus atteintes par les attaques de rouille (fig. 2.11, 2.12).

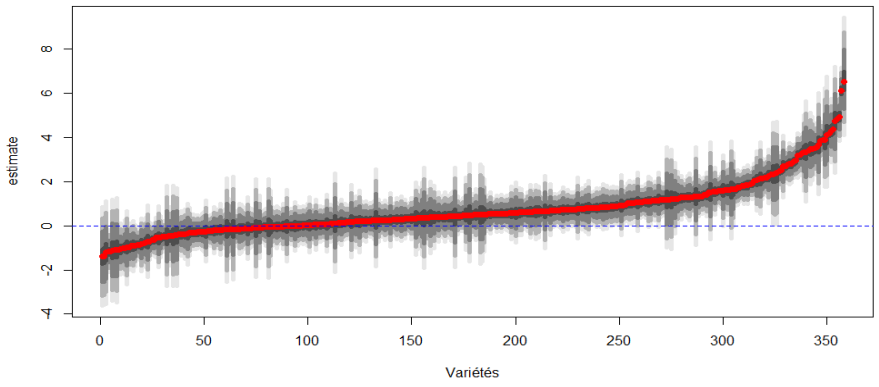


FIGURE 2.11 – Quantiles des lois a posteriori des μ_k^v

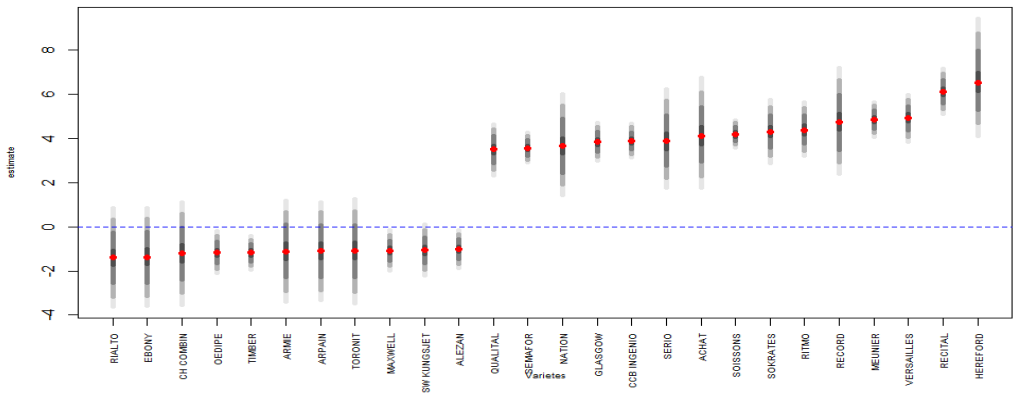


FIGURE 2.12 – Quantiles des lois a posteriori des μ_k^v correspondant aux variétés les plus et moins malades

Bilan épidémiologique du modèle : La variété semble bien expliquer la sensibilité du blé à la rouille, ce qui ne paraît pas surprenant car elle vient du génotype de la plante, identique au sein de chaque variété. Ensuite, l'effet année ressort assez bien. Sans doute cache-t-il l'effet des variations de climat d'une année sur l'autre. De plus, on a constaté un premier résultat intéressant au vu de notre problématique : la diversité des variétés abondantes semble limiter les attaques de rouille. Au contraire, la richesse ne semble pas intervenir. Le risque météorologique lié au département est peu significatif, mais on pourra préciser cette information avec de nouvelles données plus adéquates.

Bilan statistique : Le modèle à effet de seuil a bien convergé, les paramètres sont bien estimés. Toutefois, l'ajustement, notamment sur les données de sévérité, ainsi que les prédictions sont à améliorer. Pour cela, on va développer un second modèle dans lequel on ajustera une loi béta, très flexible, sur la sévérité. On séparera alors l'incidence et la sévérité pour bien appréhender les notes de rouille égales à zéro.

On a mis en oeuvre le modèle dissocié. On ne développe pas ici le détail de ces modèles mais les principaux résultats qui en ressortent sont les suivants. L'ajustement sur les données de sévérité est un peu meilleur, et on constate que les paramètres liés à la diversité, qu'elle soit portée par les variétés rares ou abondantes, n'interviennent pas : leurs lois a posteriori sont toutes estimées centrées autour de 0. La diversité variétale ne semble donc pas influer sur la sévérité des attaques de rouille. En revanche, le modèle sur les données d'incidence nous donne des posteriores pour ces mêmes coefficients centrées autour de -0.6, ou en deçà ; particulièrement lorsque l'indice de diversité se base sur les variétés abondantes. On en déduit que favoriser la diversité variétale diminue l'incidence de la rouille dans les départements français. Les résultats de ce modèle double, bien que non détaillés dans ce rapport, ont fait intégrante partie de mon travail durant ce stage, et seront donc pris en compte dans la discussion.

Discussion et conclusion

Le modèle à effet de seuil converge, mais l'ajustement n'est pas optimal, notamment concernant la sévérité de la rouille : on a tendance à surestimer les notes faibles. Le modèle dissocié nous a permis de réduire les variances et on diminuer la tendance à surestimer les notes faibles. Modéliser séparément incidence et sévérité de rouille semble donc être une meilleure idée qu'une modélisation jointe de ces deux phénomènes. Il reste tout de même un travail à faire pour arriver à mieux cerner la variabilité et réaliser de meilleures prédictions, en essayant de modifier certaines variables, par exemple mettre la variété en effet aléatoire, ou encore tester des modèles multiplicatifs vis-à-vis des erreurs (variance induite par les aléas de l'essai),...

Des pistes vers un modèle plus dynamique et plus précis :

Le risque météorologique (via notes de risque d'ARVALIS) apparaît dans les deux modèles, mais pas avec de grandes valeurs. Ceci est sans doute lié au fait qu'on a réalisé une moyenne du risque sur des années hors cadre de l'étude. Il sera intéressant de relancer les simulations avec les risques des années qui correspondent à celles de notre étude (demande envoyée à l'institut ARVALIS), afin de bien dissocier l'effet climatique global d'une année, présent sur l'ensemble du territoire (et que la variable année refléterait, entre autres), des variations climatiques locales engendrées par les particularités de chaque département (le risque publié par ARVALIS).

Dans notre modèle, l'année constituait la seule variable pour rendre compte de l'évolution de la résistance des variétés dans le temps et on voit qu'elle est significative, surtout pour expliquer l'incidence. Il serait sans doute pertinent d'ajouter aux covariables la date d'introduction de chaque variété sur le marché agricole. Ceci permettrait de prendre en compte l'évolution des populations de rouille via leur adaptation aux variétés de blé au cours du temps (La résistance d'une variété est-elle facilement contournée par le pathogène via des mutations ?).

En ce qui concerne la diversité, comme expliqué dans le premier chapitre, on s'est basé sur la fréquences des variétés dans les départements (JDD3) en supposant que la diversité variétale reflétait une certaine diversité fonctionnelle vis à vis de la résistance à la rouille. D'après les résultats il semblerait que cela soit le cas pour l'incidence mais pas pour la

sévérité. On est donc curieux maintenant d'ajouter une variable qui refléterait la diversité génétique des plantes. I. Bonnin, C. Bonneuil, R. Goffaux, P. Montalent et I. Goldringer ont réalisé une classification des départements français (pour lesquels ils disposaient d'assez de donnée) basée à la fois sur la régularité de la répartition spatiale des variétés de blé, sur leur diversité génétique, et sur l'évolution de ces deux paramètres au cours du XXième siècle [Bonnin I. 2014]. Remplacer l'indice de diversité variétale que nous avons utilisé par ces classes de départements nous donnerait deux résultats. Premièrement, cela nous indiquerait si l'hétérogénéité génotypique explique mieux l'incidence que la diversité variétale mais surtout si cette hétérogénéité intervient dans l'explication de la sévérité.

A partir d'un jeu de données supplémentaire renseignant les gènes majeurs de résistance de certaines variétés de blé, on pourra être encore plus précis et prendre en compte directement les gènes mêmes responsables de la résistance qualitative de la plante. On pourra donc raffiner l'indice de diversité grâce à ces info, par ex en classant les variétés selon les deux facteurs qui interviennent ici : la diversité des gènes majeurs et les empilement (ou combinaisons) de ces gènes, à savoir le pyramidage présents chez la plante. On a vu via les nombres de Hill qu'une forte diversité des variétés majoritaires réduit la probabilité d'être infecté par la rouille (incidence). De plus, on sait que l'incidence est régulée par la résistance qualitative des plantes, portée par les « gènes majeurs ». Il paraît donc logique d'obtenir un bon résultat. De plus, malgré l'absence de lien direct des gènes majeurs avec la résistance quantitative, on pourrait également tester l'information des gènes majeurs pour modéliser la sévérité, car ils peuvent avoir un lien avec celle ci via des effets de dilution. L'effet de dilution désigne des situations où un pathogène risque moins d'être transmis à des hôtes qui y sont sensibles quand il est dans le même temps massivement acquis par des hôtes dans lesquels il ne peut se reproduire. Les gènes majeurs influencent donc directement l'incidence mais aussi indirectement la sévérité dans le cas où des hôtes sensibles et résistantes sont mélangés (donc en présence de diversité variétale, notamment des variétés abondantes).

On a montré au terme de ce stage l'influence de la diversité variétale à l'échelle du département sur la rouille. On peut affirmer que les départements où le nombre de variétés abondantes est assez élevé ont moins tendance à être touchés par la rouille. Les effets de la diversification sont donc bien soutenus théoriquement mais il n'existe que trop peu d'exemples pratiques, notamment à de grandes échelles de temps et d'espace, et en conditions de production. Nos résultats vont dans le sens de soutenir les concepts théoriques, ce qui pourra servir pour faire du conseil variétal en préconisant la culture de plusieurs génotypes au sein d'un même bassin de production.

Bilan personnel du stage :

Au cours de ce stage, j'ai réalisé un travail bibliographique afin de me familiariser avec le sujet, et de mieux en cerner les limites. Après avoir nettoyé et complété les données de sorte à ce qu'elles se recoupent assez pour permettre une étude statistique, j'ai réalisé une

première analyse descriptive du sujet. Ensuite, j'ai pu, à l'aide de mon tuteur, développer un modèle valable quoiqu'à parfaire, mais qui donne des résultats cohérents. En tenant compte des premiers résultats, j'ai également pu mettre en œuvre un second modèle à la fin de ma période de stage dont les résultats ne sont pas présentés dans ce rapport. Ils ont toutefois été présentés à mon tuteur.

Ce stage m'a été bénéfique à de nombreux niveaux. Tout d'abord, le fait de mettre en application des concepts étudiés en cours m'a permis de me les approprier et ainsi de mieux les assimiler. Ayant réalisé mon projet statistique de 2ième année en fréquentiste, j'ai trouvé très intéressant de développer un modèle bayésien au cours de ce stage. Cela m'a permis de me familiariser avec ce mode d'inférence statistique, dont mes connaissances restaient très théoriques jusque là.

J'ai également beaucoup appris sur l'épidémiologie, discipline qui passionnante qui m'était totalement étrangère jusqu'alors.

Le fait de réaliser ce stage au sein d'un laboratoire de recherche m'a permis de prendre conscience de la réalité du travail d'un chercheur, qui nécessite notamment de faire preuve de beaucoup de patience, et de persévérance dans les recherches bibliographiques, mais aussi de rigueur, de réflexion, d'esprit d'innovation, et encore de savoir jeter un regard critique sur son travail...

Annexes

Annexe A

Organisation INRA

CHEFS DE DÉPARTEMENT	PRÉSIDENTS DE CENTRE ET DÉLÉGUÉS RÉGIONAUX	DIRECTEURS DES SERVICES D'APPUI
Jean Dallongeville *Alimentation humaine - ALIMH*	**Henry Seegers** *Centre Angers-Nantes Pays de la Loire*	**David Moisan** *Centre Angers-Nantes Pays de la Loire*
Carole Caranta *Biologie et amélioration des plantes - BAP*	**Harry Ozier-Lafontaine** *Centre Antilles-Guyane*	**Patrick Labbé** *Centre Antilles-Guyane*
Michael O'Donohue *Caractérisation et élaboration des produits issus de l'agriculture - CEPIA*	**Jean-Baptiste Coulon** *Centre Auvergne-Rhône-Alpes*	**Rémy Beaufrère** *Centre Auvergne-Rhône-Alpes*
Thierry Caquet *Écologie des forêts, prairies et milieux aquatiques - EFPA*	**Hubert de Rochambeau** *Centre Bordeaux-Aquitaine*	**Lionel Roineau** *Centre Bordeaux-Aquitaine*
Guy Richard *Environnement et agronomie - EA*	**Frédérique Pelsy** *Centre de Colmar - Alsace*	**Évelyne Klotz** *(par interim)* *Centre de Colmar - Alsace*
Denis Milan *Génétique animale - GA*	**François Casabianca** *Centre de Corse*	**Dominique Ottomani** *Centre de Corse*
Frédérick Garcia *Mathématiques et informatique appliquées - MIA*	**Françoise Simon-Plas** *Centre de Dijon - Bourgogne - Franche-Comté*	**Christine Martinez** *Centre de Dijon - Bourgogne - Franche-Conté*
Emmanuelle Maguin *Microbiologie et chaîne alimentaire - MICA*	**Benoît Malpaux** *Centre de Jouy-en-Josas*	**Marie-Claude Paulien** *Centre de Jouy-en-Josas*
Françoise Médale *Physiologie animale et systèmes d'élevage - PHASE*	**Laurent Bruckler** *Centre de Montpellier*	**Dominique Ottomani** *Centre de Montpellier*
Thierry Pineau *Santé animale - SA*	**Erwin Dreyer** *Centre Nancy-Lorraine*	**Évelyne Klotz** *Centre Nancy-Lorraine*
Christian Lannou *Santé des plantes et environnement - SPE*	**Laurent Hémidy** *(par interim)* *Centre Nord-Picardie-Champagne*	**Gabrielle Inguscio** *Centre Nord-Picardie-Champagne*
Benoit Dedieu *Sciences pour l'action et le développement - SAD*	**Jean-Marc Chabosseau** *Centre Poitou-Charentes*	**Lilian Giry** *Centre Poitou-Charentes*
Alban Thomas *Sciences sociales, agriculture et alimentation, espace et environnement - SAE2*	**Michel Bariteau** *Centre Provence-Alpes-Côte d'Azur*	**Yves Foll** *Centre Provence-Alpes-Côte d'Azur*
	Patrick Herpin *Centre de Rennes - Bretagne - Basse-Normandie*	**Edwige Lassalas** *Centre de Rennes - Bretagne - Basse-Normandie*
	Michèle Marin *Centre Toulouse-Midi-Pyrénées*	**Stéphanie Bréhin** *Centre Toulouse-Midi-Pyrénées*
	Catherine Beaumont *Centre Val de Loire*	**Stéphane Cruzol** *Centre Val de Loire*
	Laurent Hémidy *Centre de Versailles-Grignon*	**Gabrielle Inguscio** *Centre de Versailles-Grignon*
	Benoît Malpaux *(par interim)* *Déléguée régionale Île-de-France*	**Karine Gueritat** *Administratrice du centre-siège*

FIGURE A.1 – Departements de recherche et Directeurs - INRA

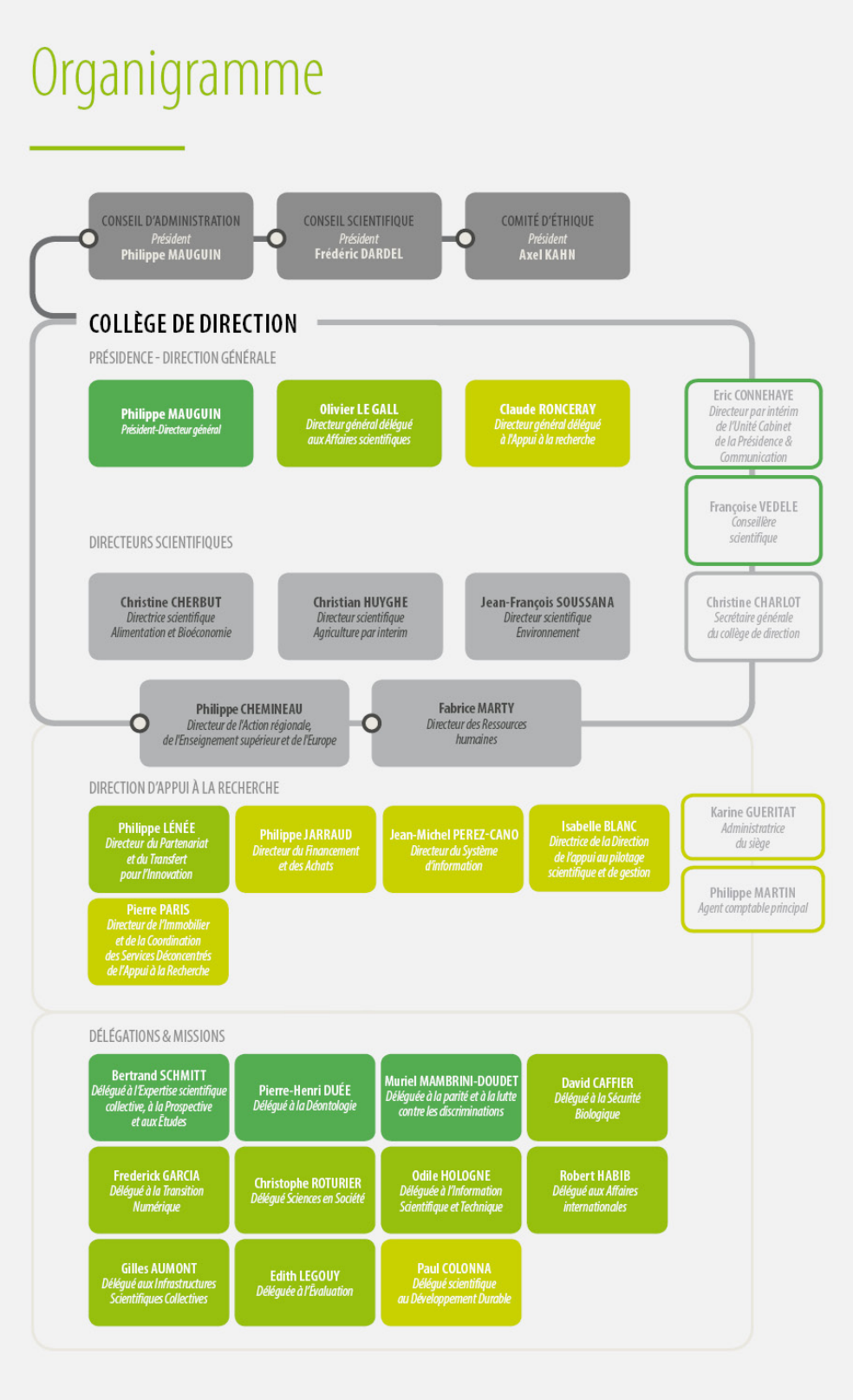


FIGURE A.2 – Organigramme INRA

Annexe B

Risque de rouille

Un indice de risque est établi chaque année, dans 710 stations sur la France (JDD2). Ses valeurs sont comprises entre environ 300 et 1800, et ont été divisées en trois degrés de risque : Faible, Moyen et Fort (voir fig. B.1).

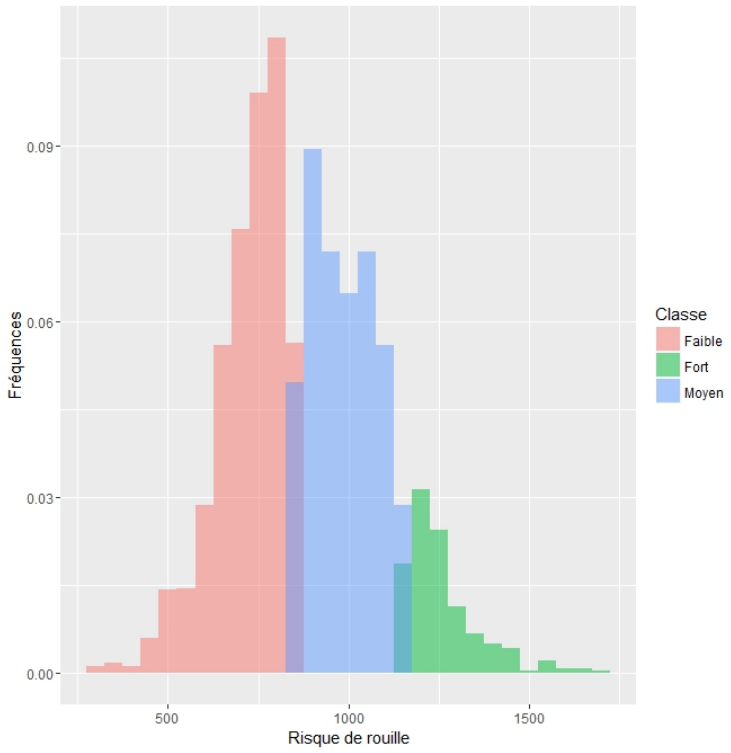


FIGURE B.1 – Histogramme des notes de risque

Nous allons travailler à l'échelle du département, il nous faut donc transformer les données dont nous disposons pour n'avoir qu'un risque pour chaque département. Toutes

années confondues, on compte environ 30 relevés par département en moyenne (min 4, max 84). Une solution simple consiste à calculer un risque moyen sur chaque département. Loin d'être optimale en fonction de la disposition des stations dans le département (bien répartis, ou pas) et de la variabilité des valeurs au sein d'un département, elle convient toutefois à nos données, assez peu variables dans chaque département. Une autre solution plus précise aurait été de réaliser un krigeage, qui consiste en une interpolation spatiale de nos données afin d'estimer la valeur du risque en tout point de l'espace. On décide de ne pas adopter cette solution pour le moment car plus coûteuse en temps et pas nécessaire au vu du peu de données de rouilles (qu'on ne pourra donc pas spatialiser).

Utiliserons-nous les valeurs continues ou les classes de risque ?
On remarque (fig. B.2) que la distribution des niveaux de risque est assez lisse, ce qui nous pousse donc plutôt à laisser de côté la classification de l'institut ARVALIS et à travailler sur les données continues. De plus, en traçant l'histogramme des risques moyens sur les départements (fig. B.3), on constate que les classes des risque faites par ARVALIS ne constituent pas un découpage très pertinent. On reste donc sur une variable quantitative du risque de rouille pour ne pas perdre d'information.

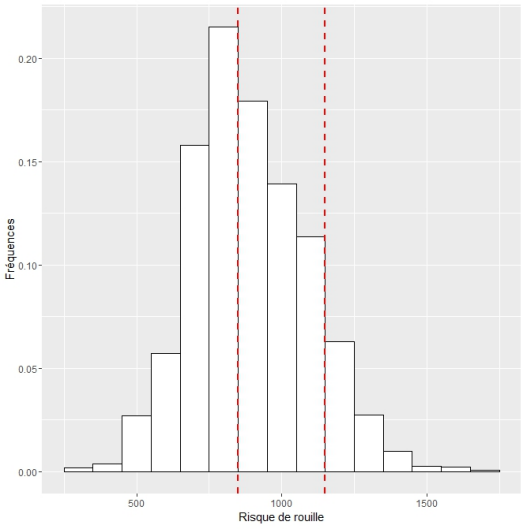


FIGURE B.2 – Histogramme des notes de risque

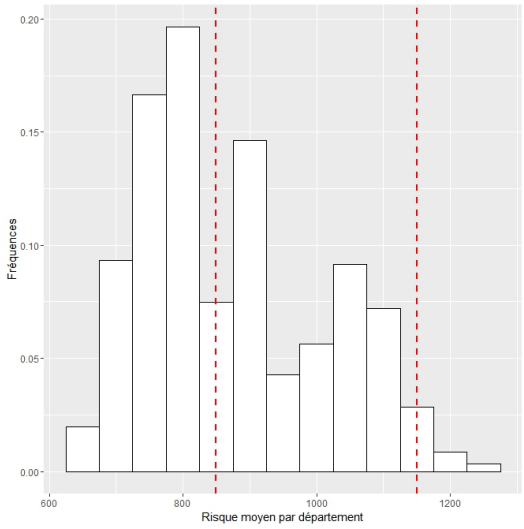


FIGURE B.3 – Histogramme des notes moyennes de risque par département

Annexe C

Cartes des nombres de Hill

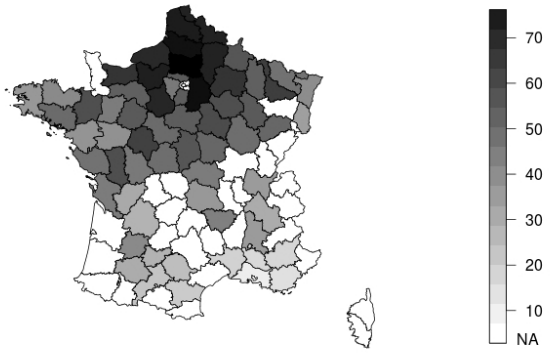


FIGURE C.1 – Nombres de Hill par département en 2000 pour $q = 0$

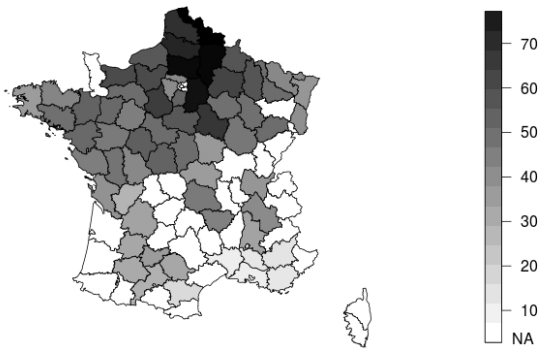


FIGURE C.2 – Nombres de Hill par département en 2002 pour $q = 0$

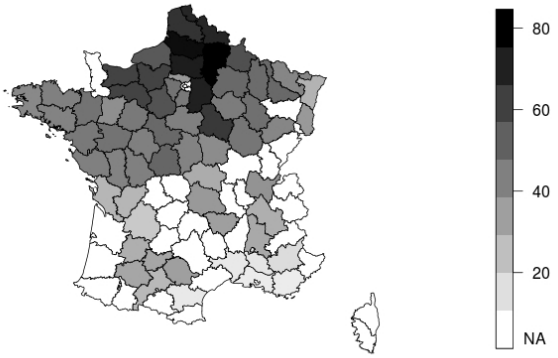


FIGURE C.3 – Nombres de Hill par département en 2004 pour $q = 0$

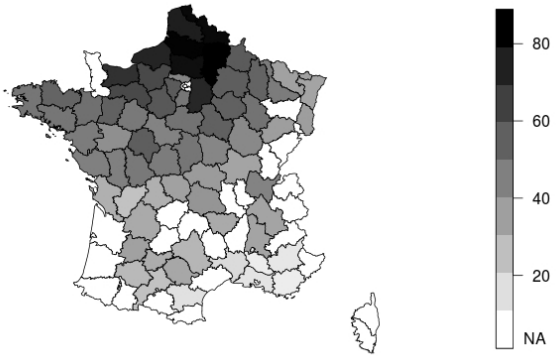


FIGURE C.4 – Nombres de Hill par département en 2006 pour $q = 0$

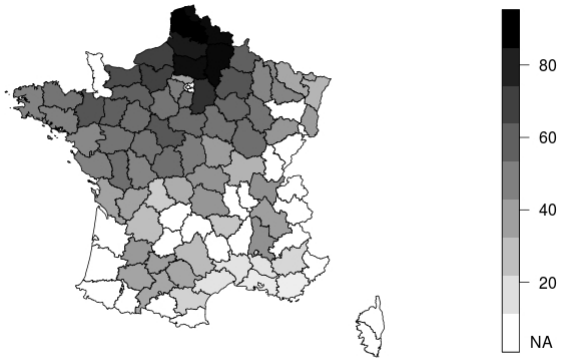


FIGURE C.5 – Nombres de Hill par département en 2008 pour $q = 0$

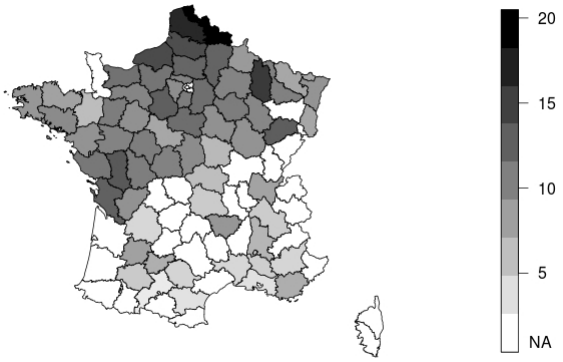


FIGURE C.6 – Nombres de Hill par département en 2000 pour $q = 2$

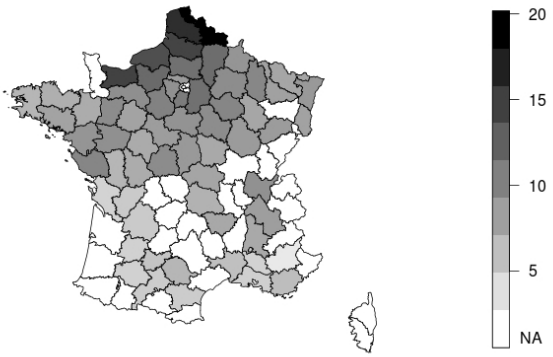


FIGURE C.7 – Nombres de Hill par département en 2002 pour $q = 2$

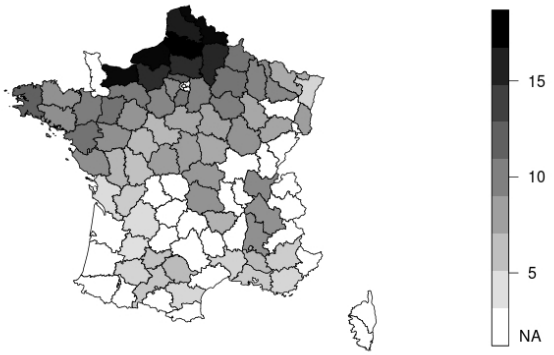


FIGURE C.8 – Nombres de Hill par département en 2004 pour $q = 2$

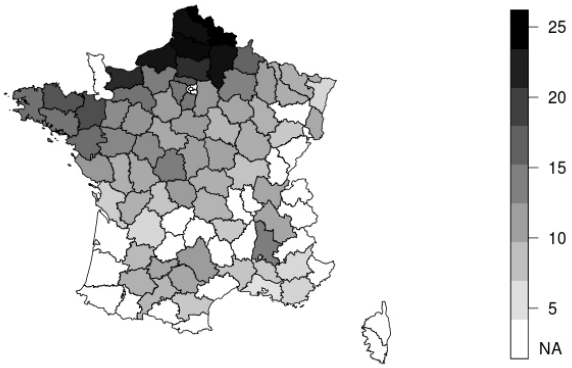


FIGURE C.9 – Nombres de Hill par département en 2006 pour $q = 2$

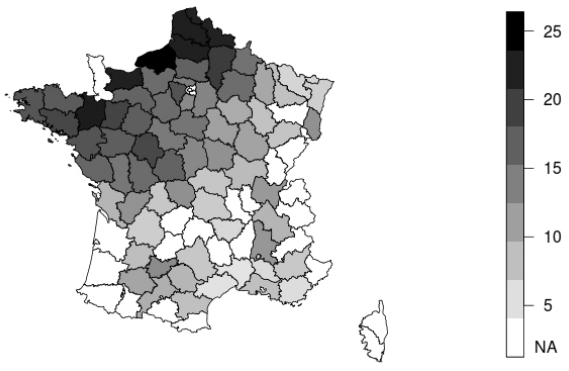


FIGURE C.10 – Nombres de Hill par département en 2008 pour $q = 2$

Annexe D

Indices de diversité spatialisés

Afin de visualiser si les valeurs prises par le nb de Hill dans un département sont corrélées dans l'espace, et si oui, avec quelle portée ? Une méthode répandue en biostatistiques consiste à tracer les variogramme et bétagramme empiriques.

D.1 Variogrammes

D.1.1 Variogramme théorique

Le variogramme est une fonction mathématique utilisée en géostatistique. L'analyse variographique est l'estimation et l'étude d'un variogramme sur une variable aléatoire. Soit Z une variable aléatoire de la variable d'espace x. Supposons la stationnaire, c'est-à-dire que la moyenne et la variance de Z(x) sont indépendantes de x. On pose la grandeur :

$$\gamma(x, y) = \frac{1}{2} E\left[|Z(x) - Z(y)|^2\right]$$

. Comme Z est stationnaire, le membre de droite dépend uniquement de la distance entre les points x et y. Le variogramme à une distance h est alors la demi moyenne des carrés des différences des réalisations de Z sur les points espacés de h.

$$\gamma(h) = \frac{1}{2} \mathbb{E}_{|y-x|=h}\left[|Z(x) - Z(y)|^2\right]$$

Le variogramme est défini pour toute fonction aléatoire intrinsèque et dépendant uniquement de l'interdistance h, alors que la fonction de covariance ne l'est que pour le cas d'une fonction aléatoire stationnaire d'ordre 2. De plus, l'estimation du variogramme n'est pas biaisée par la moyenne, au contraire de la covariance. Si la covariance de Z tend vers 0 à l'infini, le variogramme présente un palier $\gamma(\infty) = \mathbb{V}[Z]$. On nomme portée la distance à partir de laquelle le variogramme atteint son palier.

D.1.2 Variogrammes empiriques

Le variogramme expérimental ou variogramme empirique est un estimateur du variogramme théorique à partir des données. On somme – sur des classes de distances entre départements – les écarts au carré entre les valeurs de notre indice de diversité selon la formule suivante :

$$\forall h \in]0, dist_{max}], \gamma(h) = \frac{1}{2N_c} \sum_{(d,d') \in \mathbb{C}_h} (Z(d) - Z(d'))^2$$

où $dist_{max}$ représente la distace maximale entre deux départements, $Z(d)$ le nb de Hill du département d, $\mathbb{C}_h$: la classe de distances à laquelle appartient h, $N_{\mathbb{C}}$ le nombre de couples de départements dans la classe de distances $\mathbb{C}$ et (d, d') un couple de départements.

Il faut noter que le variogramme expérimental n'est pas très fiable lorsque la distance dépasse la moitié de la distance maximale, notamment à cause du nombre de couples qui diminue, faisant augmenter la variance de la valeur estimée [Arnaud M. 2000]. De ce fait, on a designé nos classes de façon à homogénéiser au mieux le nombre de couples de départements dans chacune. Classes de découpe distances (abscisses) : de 0 à $dist_{max}/5$: tous les 100 km, de $dist_{max}/5$ à $dist_{max}/2$, tous les 60 km, de $dist_{max}/2$ à $dist_{max}$, tous les 200 km.

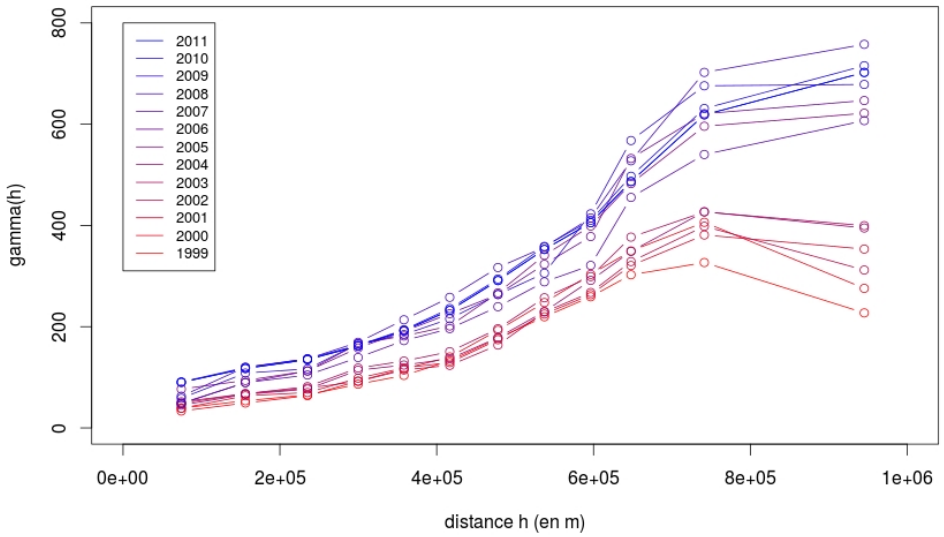


FIGURE D.1 – Variogramme empirique sur nombres de Hill, pour $q = 0.1$

Lorsque l'indicateur de diversité est proche de la richesse du département en blé tendre (fig .D.1), on n'observe a priori pas spécialement de structure, mais on note une tendance

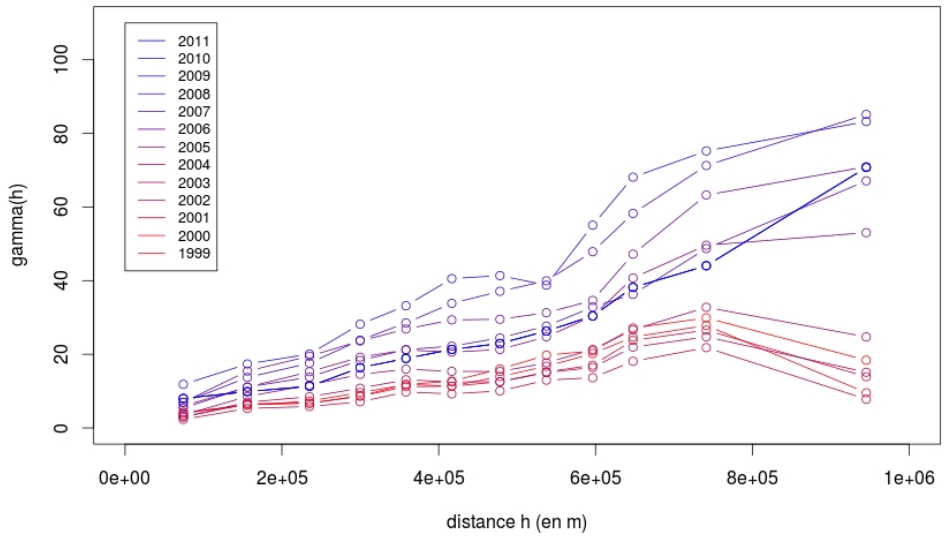


FIGURE D.2 – Variogramme empirique sur nombres de Hill, pour $q = 2$

élevée. En regardant les courbes qui reflètent la diversité portée par les variétés abondantes au contraire (fig .D.2), on note un premier palier avec une portée d'environ 400 km. Dans les deux cas, on observe l'effet de diversification identifié plus haut.

A titre indicatif, voici le tableau répertoriant le nombre de couples de départements de chaque classe de distance des variogrammes effectués sur les nombres de Hill.

Classe de distance	Nb de couples
(0,1e+05]	424
(1e+05,2e+05]	1086
(2e+05,2.68e+05]	980
(2.68e+05,3.28e+05]	930
(3.28e+05,3.88e+05]	986
(3.88e+05,4.48e+05]	936
(4.48e+05,5.08e+05]	920
(5.08e+05,5.68e+05]	772
(5.68e+05,6.28e+05]	662
(6.28e+05,6.69e+05]	376
(6.69e+05,8.69e+05]	858
(8.69e+05,1.07e+06]	152

TABLE D.1 – Nombre de couples de départements par classe de distance

Par ailleurs, on met en avant dans l'annexe E le lien qui semble exister entre notre indice de diversité et la superficie cultivée en blé tendre dans le département. Alors, afin de neutraliser la tendance portée par le paramètre superficie, on réalise les variogrammes sur les résidus issus de la régression de nos indices de diversité par la superficie en blé tendre cultivée dans les départements où ils sont calculés.

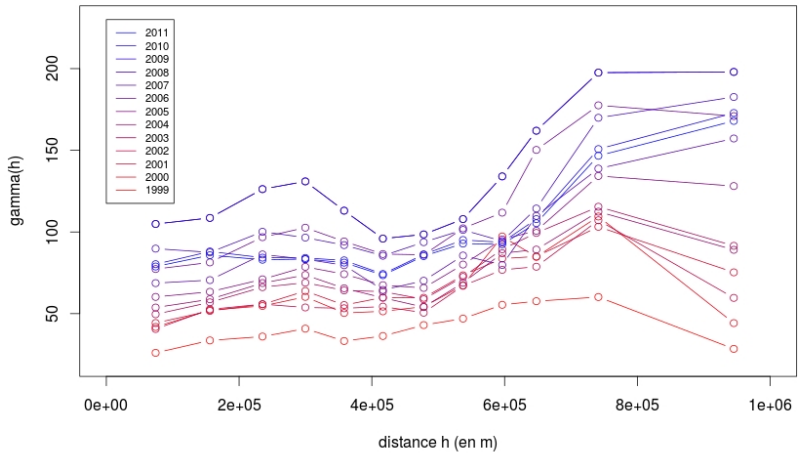


FIGURE D.3 – Variogramme empirique sur résidus des nombres de Hill après régression par superficie, pour $q = 0.1$

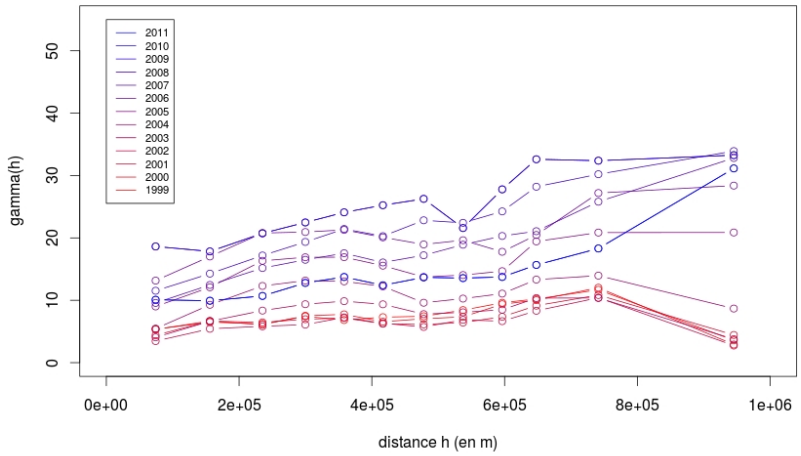


FIGURE D.4 – Variogramme empirique sur résidus des nombres de Hill après régression par superficie, pour $q = 2$

Les figurent D.3 et D.4 semblent confirmer la tendance à la diversification avec les années déjà identifiée plus tôt qui est d'ailleurs plus visible en ce qui concerne les variétés abondantes. La diversité portée par les variétés rares connaît elle effectivement un palier de variance autour des 400 km de portée. Ceci témoigne d'une certaine structure spatiale : la France semble être composée de groupes de départements (en zones d'environ 400 km de large) ayant une diversité en variété rares relativement homogène.

D.2 Betagrammes empiriques

Jusqu'ici, on considérait la diversité de chaque département vis-à-vis de la France. On peut également calculer la diversité entre deux départements (aussi appelée diversité beta), que l'on peut assimiler à un indicateur de dissimilarité :

$$\forall\,(d,d')\,\text{couple de départements},\ \beta(d,d') = \sum_{s\in V_{d,d'}} \left(p_s^d - p_s^{d'}\right)^2,$$

avec p_s^d la fréquence de la variété de blé s dans le département d.

Le betagramme empirique (tracé sur le même principe que le variogramme pour les nombres de Hill) nous indique si la dissimilarité entre deux départements dépend de la distance qui les sépare.

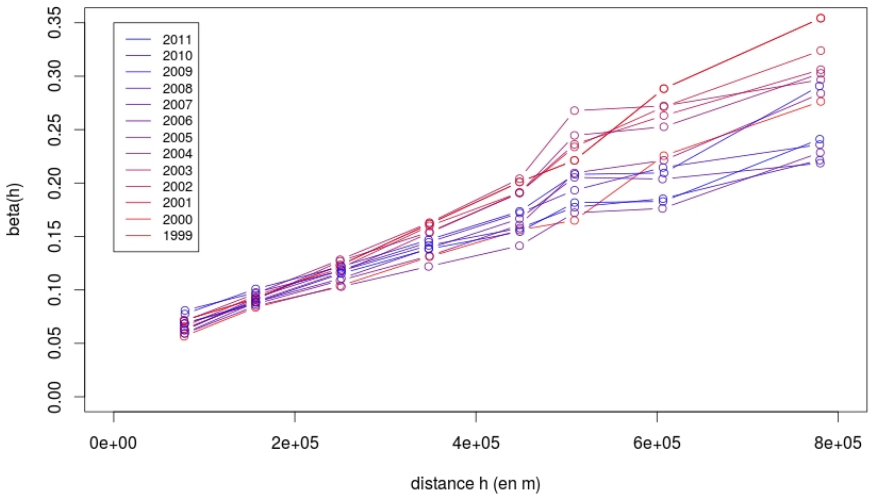


FIGURE D.5 – Betagramme empirique, avec h distance entre les départements d et d'

On obtient un tracé relativement similaire avec celui du variogramme sur les nombres de Hill : on retrouve la structure en deux paliers, ce qui semble confirmer l'hypothèse selon

laquelle la France est organisée en groupes de départements voisins assez homogènes en terme de diversité variétale. Toutefois, la diversité beta a diminué à partir du milieu des années 2000, donc on peut penser que les variétés cultivées deviennent de plus en plus les mêmes d'un département à l'autre.

On peut donc tirer de ces graphiques un phénomène chronologique : il semble y avoir une tendance à la diversification des variétés rares au sein des départements, qui s'accompagne d'une uniformisation entre départements des variétés cultivées. De plus, on note que la France semble organisée en groupes de départements voisins assez homogènes en terme de diversité variétale.

Annexe E

Nombres de Hill et superficie de blé tendre

Intuitivement, il paraît naturel de penser que l'indice de diversité d'un département est positivement corrélé avec la superficie de blé tendre qui y est cultivée.

L'idée serait donc d'arriver à corriger nos nombres de Hill afin de les affranchir du biais engendré par la superficie. Considérer le ratio de notre indice de diversité sur la superficie n'est pas une solution viable : cela produit des distorsions : "[...] mistake of trying to 'standardize' the richness of different samples by dividing the species counts by the area, the number of individuals sampled, or any other measure of effort. As we have repeatedly emphasized, this rescaling produces serious distortions : extrapolations from small sample ratios of species density inevitably lead to gross over-estimates of the number of species expected in larger sample areas . », [Gotelli Nicholas J. 2011]. Une alternative serait de faire un rééchantillonnage afin de ramener tous les départements à une taille standard, ce qui nous ferait construire des échantillons contenant des individus de blé fictifs... On décide ici de garder en tête cette idée, et de ne pas l'inclure pour le moment dans le modèle. Si nous observons que le modèle ne s'ajuste pas bien, on ajoutera ce paramètre.

Annexe F

JAGS

JAGS (`http://mcmc-jags.sourceforge.net/`) est un logiciel qui, à l'aide de simulations MCMC, permet d'inférer dans le cadre de modèles bayésiens. Il est pratique car facilement utilisable depuis R. Processus en 5 étapes :

1. **Définition du modèle**
   On définit les nœuds entre variables, en mode hiérarchique, les paramètres, les données et les loi a priori.
2. **Compilation**
   Elle n'opère pas si le modèle est cyclique ou s'il manque des données ou que des lois a priori n'ont pas été renseignées.
3. **Initialisation**
   On donne des valeurs initiales aux paramètres. Par défaut, une « valeur typique » de la priore (moyenne, mode, médiane, ...) est utilisée. Attention, la même valeur initiale est utilisée si les simulations sont lancées sur plusieurs chaînes. On ne peut donc pas voir si les chaînes ont bien convergé et "oublié" leur valeur initiale. On précise donc des valeurs initiales dispersées pour pouvoir utiliser le test de Gelman et Rubin ensuite. Puis, un générateur de nombres pseudo-aléatoires est choisi pour chaque chaîne, et un échantillonneur est défini pour chaque paramètre.
4. **Adaptation et burn-in**
   Les chaînes de Markov en sortie sont séparées en deux : le burn-in (période d'ajustement, rodage, non prise en compte) et le reste de la chaîne, qui est supposé être stable. On échantillonne dans la deuxième partie de la chaîne.
5. **Monitoring**
   Enregistrement des valeurs échantillonnées. Peut être paramétré pour tirer dans certaines chaînes, et ou toutes les n itérations (*thinning*) pour réduire les autocorrélations dans les chaînes.

Annexe G

Modèle hiérarchique à effet de seuil

G.1 Schéma du modèle - DAG (Directed Acyclic Graph)

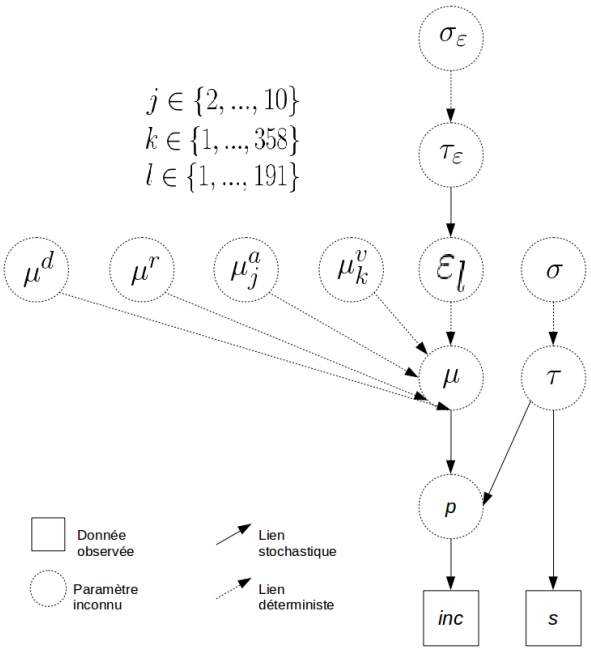


FIGURE G.1 – Schéma du modèle à effet de seuil

G.2 Code JAGS

```
model{
  #### A PRIORI ####
  sigma ~ dunif(0,10)
  tau <- 1 / pow(sigma,2)
  #essai (effet aléatoire)
  for(e in 1:E){
    epsilon[e] ~ dnorm(0,taueps)
  }
  sigmaeps ~ dunif(0,50)
  taueps <- 1 / pow(sigmaeps,2)
  #année
  man[1] <- 0
  for(a in 2:A){
    man[a] ~ dnorm(0,0.001)
  }
  #diversité
  mdiv ~ dnorm(0,0.001)
  #risque de rouille (météo)
  mrisque ~ dnorm(0,0.001)
  #variété
  for(v in 1:V){
    mvar[v] ~ dnorm(0,0.001)
  }
  #### VRAISEMBLANCE ####
  for(i in 1:N){
    mu[i] <- mvar[variete[i]] + man[annee[i]] + mdiv*hill[i] + mrisque*risque[i] + epsilon[essai[i]]
    severite[i] ~ dnorm(mu[i],tau)
    proba[i] <- 1-pnorm(0,mu[i],tau)
    incidence[i] ~ dbern(proba[i])
    #données répétées
    severiteRep[i] ~ dnorm(mu[i],tau)
    incidenceRep[i] ~ dbern(proba[i])
  }
}
```

FIGURE G.2 – Code JAGS du premier modèle

G.3 Chaînes de Markov issues de l'éch. de Gibbs

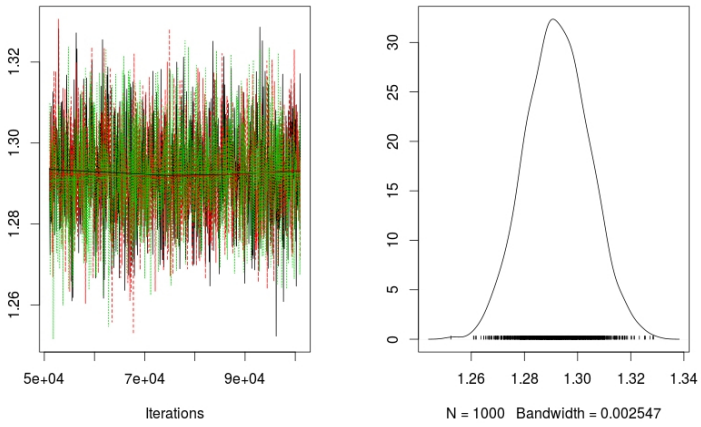


FIGURE G.3 – Paramètre σ

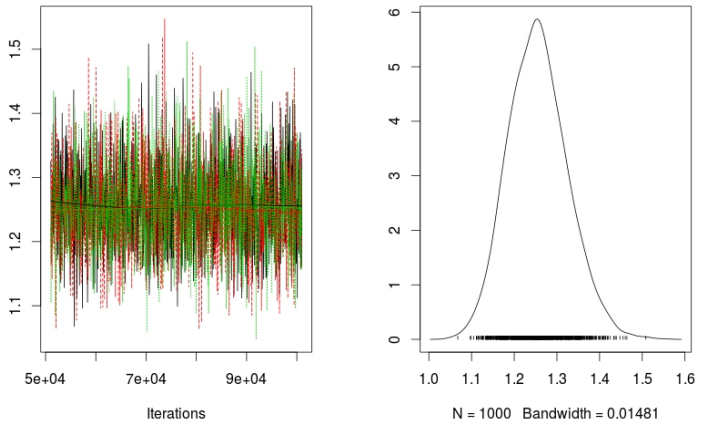


FIGURE G.4 – Paramètre σ_ε

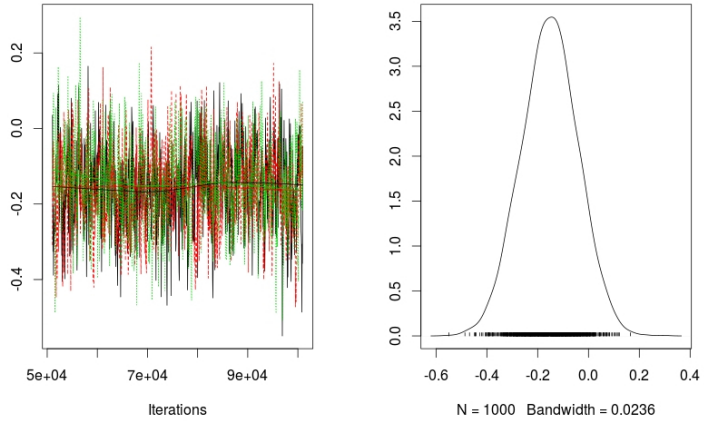


FIGURE G.5 – Paramètre μ^d

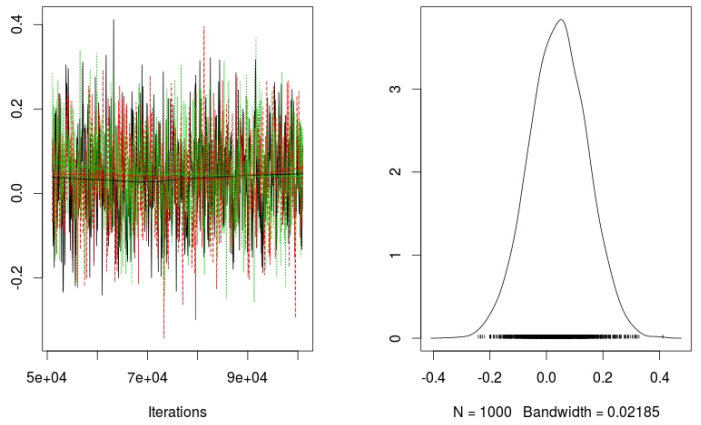


FIGURE G.6 – Paramètre μ^r

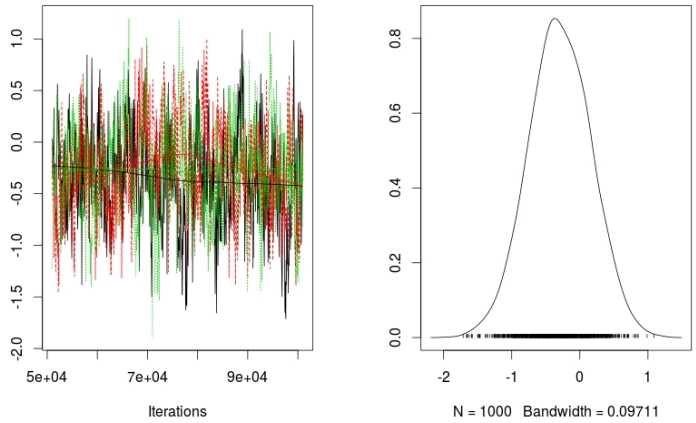


FIGURE G.7 – Paramètre μ_2^a

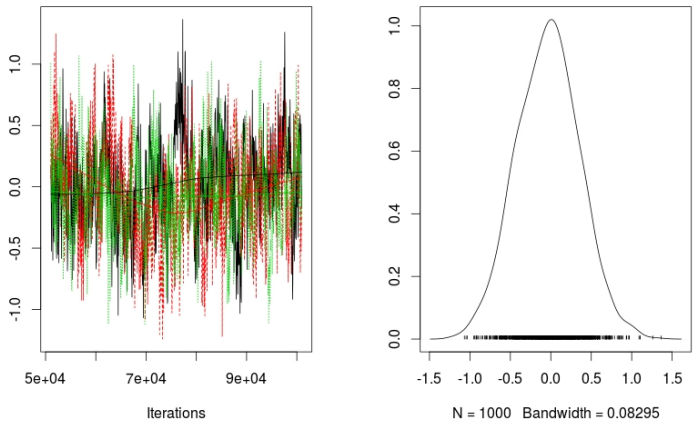


FIGURE G.8 – Paramètre μ_{100}^v

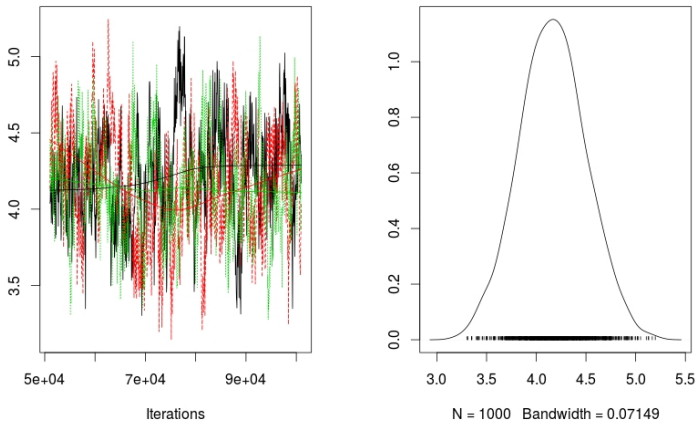


FIGURE G.9 – Paramètre μ_{314}^v

G.4 Diagnostic de convergence : Gelman-Rubin plot

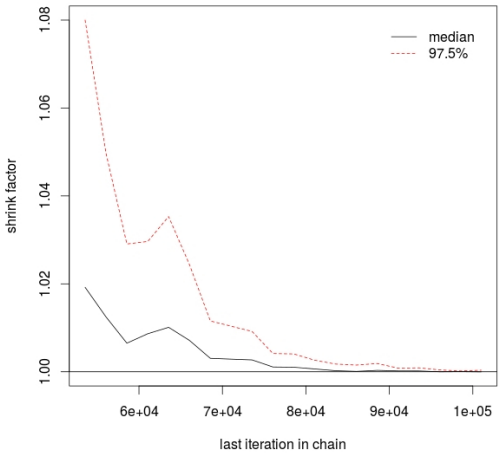


FIGURE G.10 – Shrink factor de σ

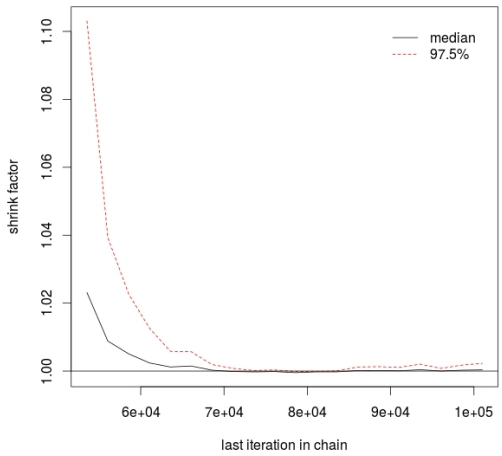


FIGURE G.11 – Shrink factor de σ_ε

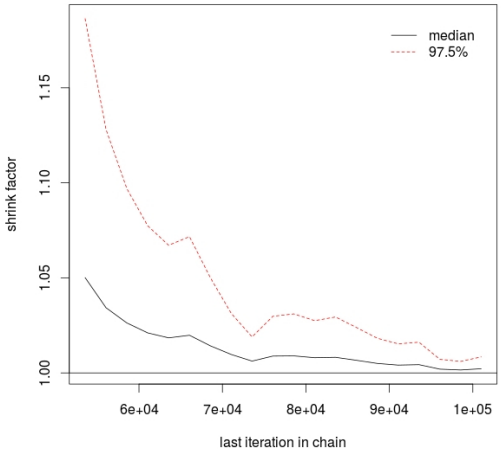


FIGURE G.12 – Shrink factor de μ^d

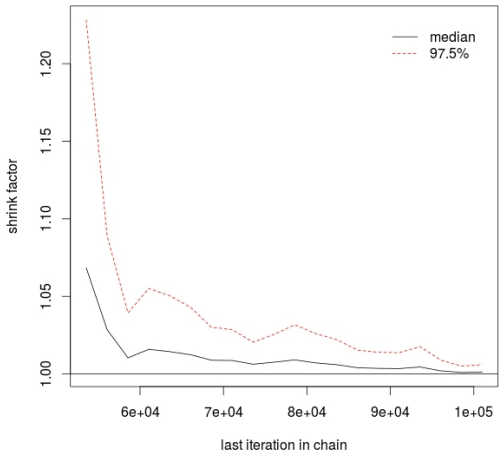


FIGURE G.13 – Shrink factor de μ^r

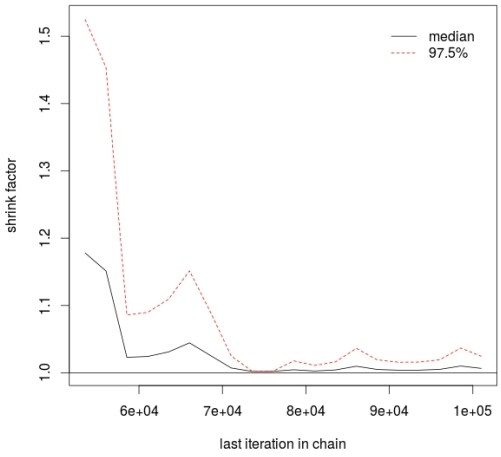


FIGURE G.14 – Shrink factor de μ_2^a

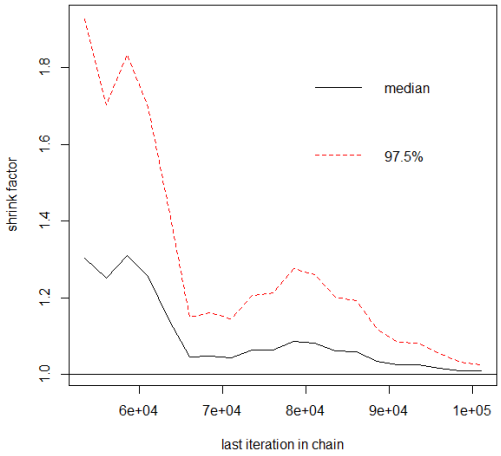


FIGURE G.15 – Shrink factor de μ^v_{100}

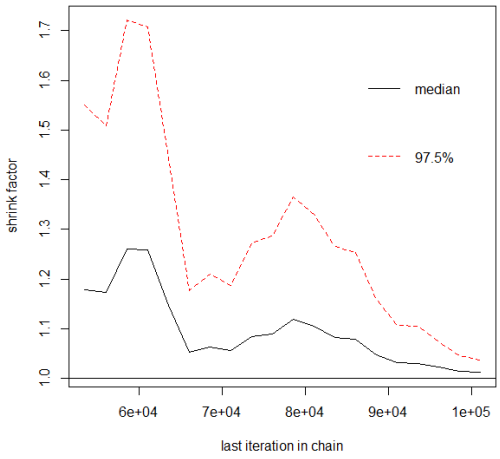


FIGURE G.16 – Shrink factor de μ^v_{314}

Bibliographie

[Arnaud M. 2000] Emery X. Arnaud M. *Estimation et interpolation spatiale : méthodes déterministes et méthodes géostatistiques.* page 126, 2000.

[Bonnin I. 2014] Goffaux R. Montalent P. Goldringer I. Bonnin I. Bonneuil C. *Explaining the decrease in the genetic diversity of wheat in France over the 20th century.* Agriculture, Ecosystems Environment, pages 183–192, 2014.

[Gotelli Nicholas J. 2011] Colwell Robert K. Gotelli Nicholas J. *Estimating species richness.* Biological diversity : frontiers in measurement and assessment 12 : 39-54., 2011.

[Hill 1973] M. O. Hill. *Diversity and Evenness : A Unifying Notation and Its Consequences.* Ecology, 1973.

[Marcon 2015] E. Marcon. *Mesures de la Biodiversité. Master. Kourou, France. <cel-01205813v2>.* 2015.

[Pautasso 2005] M. Pautasso, O. Holdenrieder et J. Stenlid. *13 Susceptibility to Fungal Pathogens of Forests Differing in Tree Diversity.* Ecological Studies, vol. 176, pages 263–289, 2005.

Autres sources

Présentation structure d'accueil

http://www.paca.inra.fr/

http://institut.inra.fr/

plaquette BioSP – Communication INRA PACA, Mai 2015

http://www.arvalisinstitutduvegetal.fr/index.html

https://fr.wikipedia.org/wiki/ARVALIS_-_Institut_du_végétal

Introduction

http://www.agro.basf.fr/agroportal/fr/fr/cultures/les_cereales/la_protection_phyto_du_ble/les_maladies_ravageurs_et_adventices/les_maladies/Rouille_brune.html

Annexe D.1.1

https://fr.wikipedia.org/wiki/Variogramme

Partie 2.2.4

http://www.terre-net.fr/observatoire-technique-culturale/strategie-technique-cultu
article/la-rouille-brune-au-coeur-du-rechauffement-217-97740.html